Question #1 *Topic 1*

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You are designing an HDInsight/Hadoop cluster solution that uses Azure Data Lake Gen1 Storage.
The solution requires POSIX permissions and enables diagnostics logging for auditing.
You need to recommend solutions that optimize storage.
Proposed Solution: Ensure that files stored are larger than 250MB.
Does the solution meet the goal?

A. Yes

B. No

**Suggested Answer:** *A*
Depending on what services and workloads are using the data, a good size to consider for files is 256 MB or greater. If the file sizes cannot be batched when landing in Data Lake Storage Gen1, you can have a separate compaction job that combines these files into larger ones.
Note: POSIX permissions and auditing in Data Lake Storage Gen1 comes with an overhead that becomes apparent when working with numerous small files. As a best practice, you must batch your data into larger files versus writing thousands or millions of small files to Data Lake Storage Gen1. Avoiding small file sizes can have multiple benefits, such as:
✑ Lowering the authentication checks across multiple files
✑ Reduced open file connections
✑ Faster copying/replication
✑ Fewer files to process when updating Data Lake Storage Gen1 POSIX permissions
Reference:
https://docs.microsoft.com/en-us/azure/data-lake-store/data-lake-store-best-practices

**Piiri565** Highly Voted 4 years, 7 months ago
POSIX permissions and auditing in Data Lake Storage Gen1 comes with an overhead that becomes apparent when working with numerous small files. As a best practice, you must batch your data into larger files versus writing thousands or millions of small files to Data Lake Storage Gen1.
according to this docs resource, I think the given answer is correct
upvoted 17 times

**arpit_dataguy** Highly Voted 4 years ago
We can ignore questions where we see GEN1 as it is out of scope now.
upvoted 5 times

**Ambujinee** Most Recent 4 years ago
File size is accepted within 256MB to 2GB
upvoted 1 times

**cadio30** 4 years, 1 month ago
Referencing the provided link the minimum acceptable file size is 256MB whereas the propose solution started at 250MB. I would say the answer is 'NO'

Reference: https://docs.microsoft.com/en-us/azure/data-lake-store/data-lake-store-best-practices
upvoted 2 times

**ZodiacC** 4 years ago
That makes not really sense for the this question
upvoted 1 times

**baobabko** 4 years ago
250 MB vs 256 MB gives less than 3% waste in worst-case. So it is acceptable. Answer should be YES
upvoted 1 times

**SplMonk** 4 years, 2 months ago
So is this a trap question? as the guidance is 256MB and they are saying larger than 250MB... a small difference but below we recommended size
upvoted 1 times

**Deepu1987** 4 years, 4 months ago

The given solution is correct

Typically, analytics engines such as HDInsight and Azure Data Lake Analytics have a per-file overhead. If you store your data as many small files, this can negatively affect performance.

pls refer this link https://docs.microsoft.com/en-us/azure/data-lake-store/data-lake-store-performance-tuning-guidance#structure-your-data-set

In general, organize your data into larger sized files for better performance. As a rule of thumb, organize data sets in files of 256 MB or larger

upvoted 2 times

**chaoxes** 4 years, 6 months ago

Given answer B. No is correct.

In POSIX-style model it is recommended to avoid small size files, due to following considerations:

-Lowering the authentication checks across multiple files

-Reduced open file connections

-Faster copying/replication

-Fewer files to process when updating Data Lake Storage Gen1 POSIX permissions

upvoted 1 times

**SudhakarMani** 4 years, 6 months ago

Is it correct answer?

upvoted 1 times

**syu31svc** 4 years, 6 months ago

Provided link says at least 265 MB but greater than 250 MB seems good enough.

I would agree with the answer

upvoted 2 times

**Torent2005** 4 years, 7 months ago

Not really, it's a trap. Files should be grater than 256 mb regarding to best practises. So bigger file thant 250 like 251 it's not a solution.

upvoted 1 times

**BaisArun** 4 years, 7 months ago

Agree with @Piiri565

upvoted 2 times

Question #2 *Topic 1*

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.

After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You are designing an HDInsight/Hadoop cluster solution that uses Azure Data Lake Gen1 Storage.

The solution requires POSIX permissions and enables diagnostics logging for auditing.

You need to recommend solutions that optimize storage.

Proposed Solution: Implement compaction jobs to combine small files into larger files.

Does the solution meet the goal?

A. Yes

B. No

**Suggested Answer:** *A*

Depending on what services and workloads are using the data, a good size to consider for files is 256 MB or greater. If the file sizes cannot be batched when landing in Data Lake Storage Gen1, you can have a separate compaction job that combines these files into larger ones.

Note: POSIX permissions and auditing in Data Lake Storage Gen1 comes with an overhead that becomes apparent when working with numerous small files. As a best practice, you must batch your data into larger files versus writing thousands or millions of small files to Data Lake Storage Gen1. Avoiding small file sizes can have multiple benefits, such as:

✑ Lowering the authentication checks across multiple files

✑ Reduced open file connections

✑ Faster copying/replication

✑ Fewer files to process when updating Data Lake Storage Gen1 POSIX permissions

Reference:

https://docs.microsoft.com/en-us/azure/data-lake-store/data-lake-store-best-practices

**chaoxes** Highly Voted 👍 4 years, 6 months ago

Correct answer

upvoted 7 times

**Deepu1987** Most Recent ⊙ 4 years, 4 months ago

Somewhat similar to above qn

https://docs.microsoft.com/en-us/azure/data-lake-store/data-lake-store-performance-tuning-guidance#structure-your-data-set

upvoted 2 times

Question #3 *Topic 1*

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.

After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You are designing an HDInsight/Hadoop cluster solution that uses Azure Data Lake Gen1 Storage.

The solution requires POSIX permissions and enables diagnostics logging for auditing.

You need to recommend solutions that optimize storage.

Proposed Solution: Ensure that files stored are smaller than 250MB.

Does the solution meet the goal?

A. Yes

B. No

**Suggested Answer:** *B*

Ensure that files stored are larger, not smaller than 250MB.

You can have a separate compaction job that combines these files into larger ones.

Note: The file POSIX permissions and auditing in Data Lake Storage Gen1 comes with an overhead that becomes apparent when working with numerous small files. As a best practice, you must batch your data into larger files versus writing thousands or millions of small files to Data Lake Storage Gen1. Avoiding small file sizes can have multiple benefits, such as:

✑ Lowering the authentication checks across multiple files

✑ Reduced open file connections

✑ Faster copying/replication

✑ Fewer files to process when updating Data Lake Storage Gen1 POSIX permissions

Reference:

https://docs.microsoft.com/en-us/azure/data-lake-store/data-lake-store-best-practices

**chaoxes** Highly Voted 4 years, 6 months ago

Given answer B. No is correct.

In POSIX-style model it is recommended to avoid small size files, due to following considerations:

-Lowering the authentication checks across multiple files

-Reduced open file connections

-Faster copying/replication

-Fewer files to process when updating Data Lake Storage Gen1 POSIX permissions

It is recommended that size of files are at least 256 MB.

Source: https://docs.microsoft.com/pl-pl/azure/data-lake-store/data-lake-store-best-practices

upvoted 8 times

**Deepu1987** Most Recent 4 years, 4 months ago

Please check this in the mentioned link

https://docs.microsoft.com/en-us/azure/data-lake-store/data-lake-store-best-practices

It's under "Improve throughput with parallelism"

Avoid Small file sizes. in this heading you need to look out for the below line

"you can have a separate compaction job that combines these files into larger ones." which states the given stmt is false

upvoted 2 times

## Question #4 *Topic 1*

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.

After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You are designing an Azure SQL Database that will use elastic pools. You plan to store data about customers in a table. Each record uses a value for

CustomerID.

You need to recommend a strategy to partition data based on values in CustomerID.

Proposed Solution: Separate data into customer regions by using vertical partitioning.

Does the solution meet the goal?

A. Yes

B. No

**Suggested Answer:** *B*

Vertical partitioning is used for cross-database queries. Instead we should use Horizontal Partitioning, which also is called charding.

Reference:

https://docs.microsoft.com/en-us/azure/sql-database/sql-database-elastic-query-overview

**kate208** Highly Voted 5 years ago

Sharding, not charding haha

upvoted 32 times

**Treadmill** Highly Voted 4 years, 10 months ago

Customer scenarios for elastic query are characterized by the following topologies:

• Vertical partitioning - Cross-database queries (Topology 1): The data is partitioned vertically between a number of databases in a data tier. Typically, different sets of tables reside on different databases. That means that the schema is different on different databases. For instance, all tables for inventory are on one database while all accounting-related tables are on a second database. Common use cases with this topology require one to query across or to compile reports across tables in several databases.

Horizontal Partitioning - Sharding (Topology 2): Data is partitioned horizontally to distribute rows across a scaled out data tier. With this approach, the schema is identical on all participating databases. This approach is also called “sharding”. Sharding can be performed and managed using (1) the elastic database tools libraries or (2) self-sharding. An elastic query is used to query or compile reports across many shards. Shards are typically databases within an elastic pool. You can think of elastic query as an efficient way for querying all databases of elastic pool at once, as long as databases share the common schema.

upvoted 27 times

**ShauryaRana** Most Recent 3 years, 12 months ago

'Avoid creating "hot" partitions that can affect performance and availability. For example, using the first letter of a customer's name causes an unbalanced distribution, because some letters are more common. Instead, use a hash of a customer identifier to distribute data more evenly across partitions.' - From MS documentation for Horizontal Partitioning.

upvoted 1 times

**Deepu1987** 4 years, 4 months ago

Here we're storing Customers data in a table and now we want to partition cust region so we need to use sharding as per the right concept as they are performed as long as DBs share common schema as per defn.

upvoted 1 times

**Shanmahi** 4 years, 5 months ago

Answer : No

Applicable solution : Horizontal partitioning

Reference : https://docs.microsoft.com/en-us/azure/architecture/best-practices/data-partitioning

upvoted 1 times

**redalarm2000** 4 years, 5 months ago

Ok i am confused as to the difference between question 4 and question 5 on this site. Question 4 says to use horizontal partitioning but Question 5 says it recommends to use horizontal partition and the wording is the same but they say that answer should be No still why?

upvoted 1 times

⊟ **mojedapr** 4 years, 9 months ago

Still don't know why horizontal and not vertical !

upvoted 1 times

⊟ **clownfishman** 4 years, 9 months ago

it is because it is to partition customers IDs, so it means it is just 1 database.

upvoted 3 times

⊟ **chaoxes** 4 years, 6 months ago

Vertical partitioning is to reduce the I/O and performance costs associated with fetching items that are frequently accessed. Vertical partitioning splits table and in the result we have more partitions with different schema instead of 1 big table. This is not what is expected in this scenario.

Horizontal partitioning using sharding is expected. Horizontal sharding will split table row-wise, so we have multiple partitions with the same schema, but based on region in that case.

For instance split table containing all customers world-wise into multiple partitions based on the regions (customer from Europe, customers from USA etc)

upvoted 1 times

Question #5 *Topic 1*

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You are designing an Azure SQL Database that will use elastic pools. You plan to store data about customers in a table. Each record uses a value for
CustomerID.
You need to recommend a strategy to partition data based on values in CustomerID.
Proposed Solution: Separate data into customer regions by using horizontal partitioning.
Does the solution meet the goal?

A. Yes

B. No

**Suggested Answer:** *B*
We should use Horizontal Partitioning through Sharding, not divide through regions.
Note: Horizontal Partitioning - Sharding: Data is partitioned horizontally to distribute rows across a scaled out data tier. With this approach, the schema is identical on all participating databases. This approach is also called ג€shardingג€. Sharding can be performed and managed using
(1) the elastic database tools libraries or
(2) self-sharding. An elastic query is used to query or compile reports across many shards.
Reference:
https://docs.microsoft.com/en-us/azure/sql-database/sql-database-elastic-query-overview

**Ankush1994** 3 years, 10 months ago
Correct answer is
No
becuase
Separate data into shards by using horizontal partitioning
upvoted 1 times

**osoroshi** 4 years, 3 months ago
harding can be performed and managed using (1) the elastic database tools libraries or (2) self-sharding. An elastic query is used to query or compile reports across many shards. Shards are typically databases within an elastic pool. You can think of elastic query as an efficient way for querying all databases of elastic pool at once, as long as databases share the common schema.
upvoted 1 times

**redalarm2000** 4 years, 5 months ago
Ok i am confused as to the difference between question 4 and question 5 on this site. Question 4 says to use horizontal partitioning but Question 5 says it recommends to use horizontal partition and the wording is the same but they say that answer should be No still why?
upvoted 2 times

**Shanmahi** 4 years, 5 months ago
Answer : No
Applicable solution : Horizontal partitioning (based on customerID not region i.e. using sharding concept)
Reference : https://docs.microsoft.com/en-us/azure/architecture/best-practices/data-partitioning
upvoted 6 times

**fmunozse** 4 years, 10 months ago
I don't understand why is not recommend horizontal ... Each shard could be the region, no?
upvoted 2 times

**stijn5454** 4 years, 10 months ago
Horizontal partitioning splits one or more tables by row, usually within a single instance of a schema and a database server.

Sharding goes beyond this: it partitions the problematic table(s) in the same way, but it does this across potentially multiple instances of the schema. The obvious advantage would be that search load for the large partitioned table can now be split across multiple servers (logical or physical), not just multiple indexes on the same logical server.

Ref: https://en.wikipedia.org/wiki/Shard_(database_architecture)#Shards_compared_to_horizontal_partitioning

upvoted 9 times

Question #6 *Topic 1*

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You are designing an Azure SQL Database that will use elastic pools. You plan to store data about customers in a table. Each record uses a value for
CustomerID.
You need to recommend a strategy to partition data based on values in CustomerID.
Proposed Solution: Separate data into shards by using horizontal partitioning.
Does the solution meet the goal?

A. Yes

B. No

**Suggested Answer:** *A*
Horizontal Partitioning - Sharding: Data is partitioned horizontally to distribute rows across a scaled out data tier. With this approach, the schema is identical on all participating databases. This approach is also called ג€shardingג€. Sharding can be performed and managed using (1) the elastic database tools libraries or (2) self- sharding. An elastic query is used to query or compile reports across many shards.
Reference:
https://docs.microsoft.com/en-us/azure/sql-database/sql-database-elastic-query-overview

**sjain91** Highly Voted 4 years, 2 months ago
Yes is the answer
upvoted 6 times

**Pdpj** Most Recent 4 years ago
No is the answer
This solution does not meet the requirements. You need to use sharding, which is partitioning data horizontally to distribute data across multiple databases in a scaled-out design, but CustomerID is not the best choice in this scenario. Sharding by RegionalID will make sorting by geographic location more efficient.

Sharding requires that the schema is the same on all of the databases involved. Sharding helps to minimize the size of individual databases, which in turn helps to improve transactional process performance. Hardware support requirements are minimized, which helps to reduce related costs. Elastic queries let you run queries across multiple shards. You can configure and manage sharding through the elastic database tools libraries or through self-sharding
upvoted 3 times

**syu31svc** 4 years, 6 months ago
This is the correct solution
upvoted 4 times

**JohnCrawford** 4 years, 2 months ago
I disagree. CustomerID will be unique and that means you would have as many shards as you have customers. This would be a poor design. The question is poorly worded and given the wording the answer might be correct, but it is lousy design.
upvoted 2 times

**Wisenut** 4 years, 1 month ago
I agree sharding based on region would be a better fit
upvoted 1 times

**cadio30** 4 years ago
this would only be correct if compound shard is created for customerid and region
upvoted 1 times

**suvenk** 4 years, 1 month ago
Well, the more the shards, the lesser is the likelihood of you facing the hot partition problem. The region will create a hot partition problem.
upvoted 3 times

## Question #7 *Topic 1*

HOTSPOT -

You are designing a data processing solution that will run as a Spark job on an HDInsight cluster. The solution will be used to provide near real-time information about online ordering for a retailer.

The solution must include a page on the company intranet that displays summary information.

The summary information page must meet the following requirements:

✑ Display a summary of sales to date grouped by product categories, price range, and review scope.

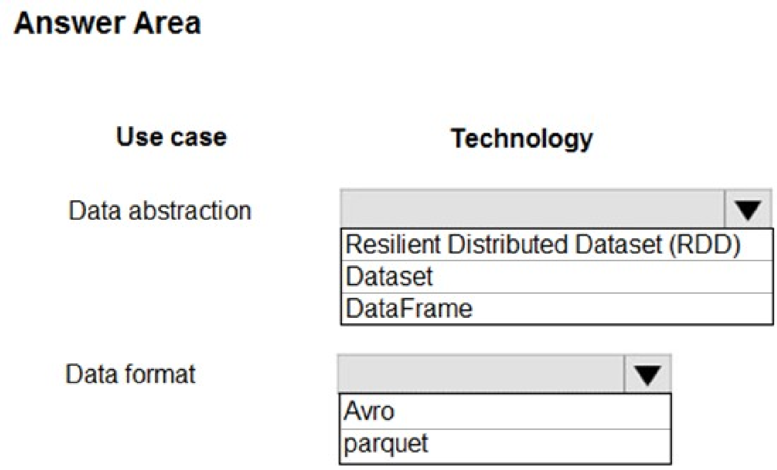
✑ Display sales summary information including total sales, sales as compared to one day ago and sales as compared to one year ago.

✑ Reflect information for new orders as quickly as possible.

You need to recommend a design for the solution.

What should you recommend? To answer, select the appropriate configuration in the answer area.

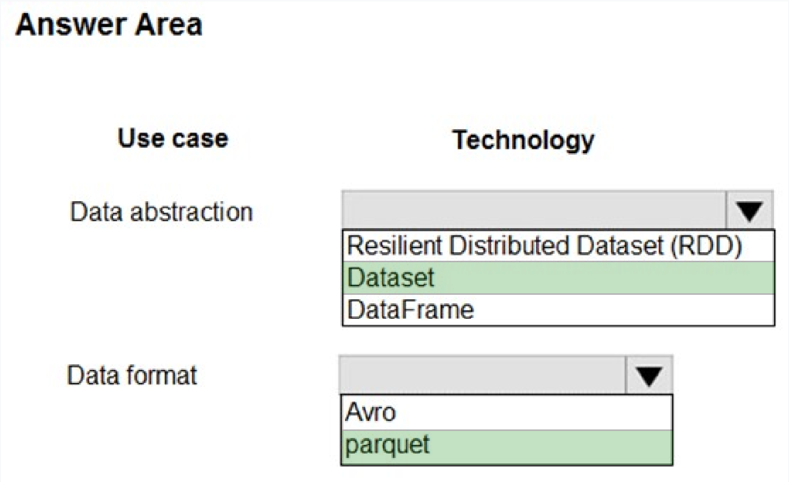
Hot Area:



**Answer Area**

| Use case | Technology |
|---|---|
| Data abstraction | ▼<br>Resilient Distributed Dataset (RDD)<br>Dataset<br>DataFrame |
| Data format | ▼<br>Avro<br>parquet |

**Suggested Answer:**



**Answer Area**

| Use case | Technology |
|---|---|
| Data abstraction | ▼<br>Resilient Distributed Dataset (RDD)<br>[✓ Dataset]<br>DataFrame |
| Data format | ▼<br>Avro<br>[✓ parquet] |

Box 1: DataFrame -

DataFrames -

Best choice in most situations.

Provides query optimization through Catalyst.

Whole-stage code generation.

Direct memory access.

Low garbage collection (GC) overhead.

Not as developer-friendly as DataSets, as there are no compile-time checks or domain object programming.

Box 2: parquet -

The best format for performance is parquet with snappy compression, which is the default in Spark 2.x. Parquet stores data in columnar format, and is highly optimized in Spark.

Incorrect Answers:

DataSets -

Good in complex ETL pipelines where the performance impact is acceptable.

Not good in aggregations where the performance impact can be considerable.

RDDs -
You do not need to use RDDs, unless you need to build a new custom RDD.
No query optimization through Catalyst.
No whole-stage code generation.
High GC overhead.
Reference:
https://docs.microsoft.com/en-us/azure/hdinsight/spark/apache-spark-perf

**kempstonjoystick** Highly Voted 5 years, 2 months ago

The highighted answer and the explanation differ. Should be dataframe I believe.

upvoted 50 times

**apz333** Highly Voted 5 years, 2 months ago

I think it should be dataframe as well. In most cases parquet and dataframe are the best choice.

upvoted 22 times

**frakcha** 5 years, 1 month ago

They say Dataset is good for complex ETL situations
https://docs.microsoft.com/en-us/azure/hdinsight/spark/apache-spark-perf

upvoted 1 times

**hsetin** Most Recent 3 years, 10 months ago

1. Dataframe
2. Parquet
Confirmed

upvoted 1 times

**mchatrvd** 3 years, 10 months ago

Anyone knows why Exam Topics have taken AWS certification questions offline? There is nothing related to AWS certifications which used to be there earlier.

upvoted 1 times

**victor90** 3 years, 7 months ago

Hi, I found the link to the associate SA exam.
https://www.examtopics.com/exams/amazon/aws-certified-solutions-architect-associate-saa-c02/view/

upvoted 1 times

**satyamkishoresingh** 3 years, 10 months ago

The practical combination is Dataframe + Parquet . Here answer clarification is ambiguous.

upvoted 1 times

**HichemZe** 3 years, 11 months ago

1- DATFRAME
2 - Data Format = Avro
Because only Avro support Streaming (Against Parquet)

upvoted 1 times

**ismaelrihawi** 4 years, 1 month ago

Data abstraction = Dataframe

upvoted 5 times

**BobFar** 4 years, 1 month ago

Dataframe is correct ,
https://docs.microsoft.com/en-us/azure/hdinsight/spark/optimize-data-storage

upvoted 2 times

**Deepu1987** 4 years, 4 months ago

It's wrong selection shown in the display. It's actually
- Data Frame [Reason for elimination Not as developer-friendly as DataSets, as there are no compile-time checks or domain object programming,don't need to use RDDs, unless you need to build a new custom RDD]
Anyhow "Parquet" is selected

upvoted 1 times

**syu31svc** 4 years, 6 months ago

https://docs.microsoft.com/en-us/azure/hdinsight/spark/optimize-data-storage:
"Parquet stores data in columnar format, and is highly optimized in Spark."
"DataFrames
Best choice in most situations."

upvoted 2 times

**BaisArun** 4 years, 7 months ago

Dataset is not good for Aggregation, Should be dataframe.

upvoted 3 times

**Nihar258255** 4 years, 7 months ago

Can some correct the answers??

upvoted 1 times

**AhmedReda** 5 years ago

The question need quick processing but Dataset add overhead, also the query is aggregation and Dataset not good at that

DataSets : Adds serialization/deserialization overhead, High GC overhead, Not good in aggregations where the performance impact can be considerable.

DataFrames : Best choice in most situations, Direct memory access.

upvoted 11 times

**Runi** 5 years ago

Data set is Not good in aggregations where the performance impact can be considerable.So. I think dataframe should be correct one. Can anyone confirm. Please Thanks.

upvoted 5 times

**serger** 5 years, 1 month ago

dataframe for sure

upvoted 4 times

**Tombarc** 5 years, 2 months ago

I think it's dataframe too.

upvoted 7 times

Question #8 *Topic 1*

You are evaluating data storage solutions to support a new application.
You need to recommend a data storage solution that represents data by using nodes and relationships in graph structures.
Which data storage solution should you recommend?

A. Blob Storage

B. Azure Cosmos DB

C. Azure Data Lake Store

D. HDInsight

**Suggested Answer:** *B*
For large graphs with lots of entities and relationships, you can perform very complex analyses very quickly. Many graph databases provide a query language that you can use to traverse a network of relationships efficiently.
Relevant Azure service: Cosmos DB
Reference:
https://docs.microsoft.com/en-us/azure/architecture/guide/technology-choices/data-store-overview

**ismaelrihawi** Highly Voted 4 years, 1 month ago
CosmosDB with Gremlin API
upvoted 7 times

**cadio30** Most Recent 4 years, 1 month ago
the proposed solution is correct as the azure cosmos db has gremlin api that can support the graph requirement.

reference: https://docs.microsoft.com/en-us/azure/cosmos-db/graph-introduction
upvoted 4 times

**IAMKPR** 4 years, 1 month ago
No other option supports Graph Structures. So it should be only Azure Cosmos DB.
upvoted 1 times

**NamishBansal** 4 years, 1 month ago
101% correct
upvoted 1 times

**syu31svc** 4 years, 6 months ago
It can only be B
upvoted 2 times

**BaisArun** 4 years, 7 months ago
yes agree with Cosmos DB
upvoted 2 times

**Rajdeep_Chakraborty** 4 years, 7 months ago
Gremlin - API
upvoted 3 times

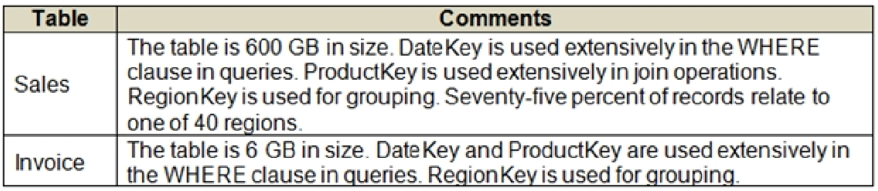
**Andrexx** 4 years, 8 months ago
I agree with the answer - Cosmos DB.
upvoted 4 times

Question #9 *Topic 1*

HOTSPOT -

You have an on-premises data warehouse that includes the following fact tables. Both tables have the following columns: DataKey, ProductKey, RegionKey.

There are 120 unique product keys and 65 unique region keys.



| Table | Comments |
|---|---|
| Sales | The table is 600 GB in size. DateKey is used extensively in the WHERE clause in queries. ProductKey is used extensively in join operations. RegionKey is used for grouping. Seventy-five percent of records relate to one of 40 regions. |
| Invoice | The table is 6 GB in size. DateKey and ProductKey are used extensively in the WHERE clause in queries. RegionKey is used for grouping. |

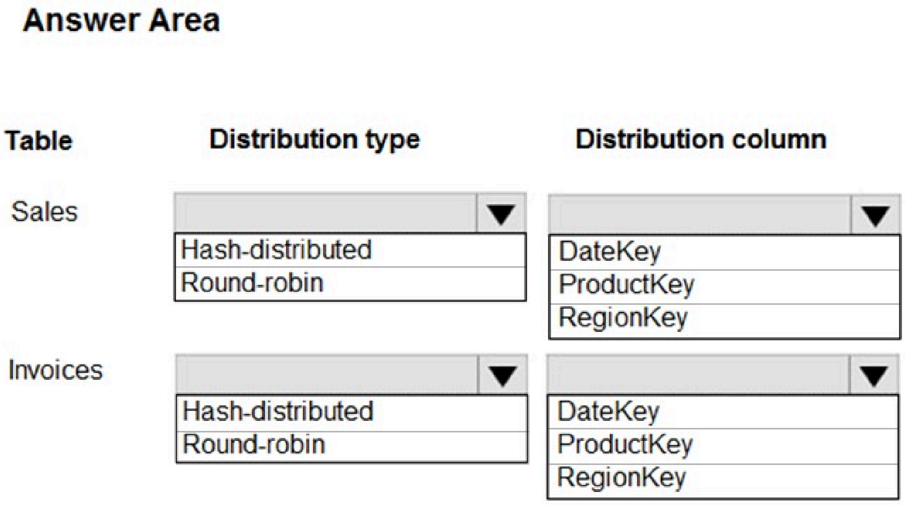
Queries that use the data warehouse take a long time to complete.

You plan to migrate the solution to use Azure Synapse Analytics. You need to ensure that the Azure-based solution optimizes query performance and minimizes processing skew.

What should you recommend? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

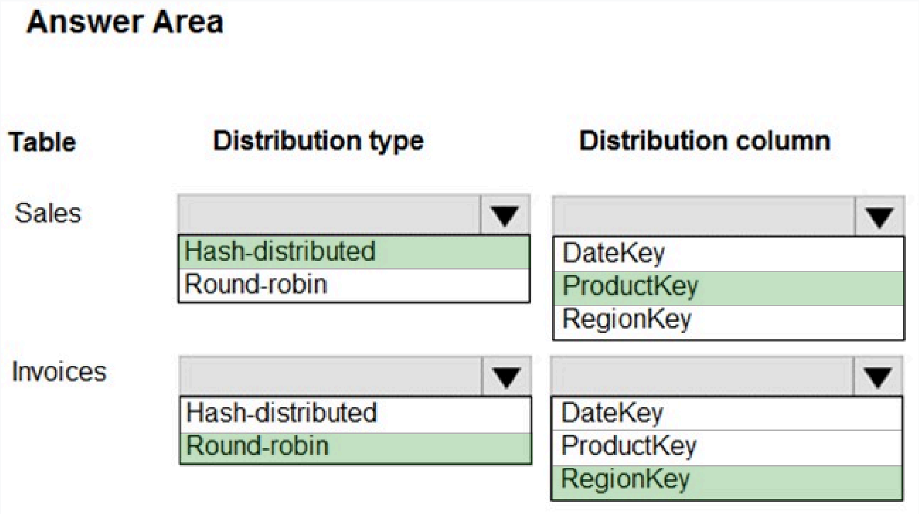
Hot Area:



Answer Area

| Table | Distribution type | Distribution column |
|---|---|---|
| Sales | Hash-distributed<br>Round-robin | DateKey<br>ProductKey<br>RegionKey |
| Invoices | Hash-distributed<br>Round-robin | DateKey<br>ProductKey<br>RegionKey |

Suggested Answer:



Answer Area

| Table | Distribution type | Distribution column |
|---|---|---|
| Sales | [✓ Hash-distributed]<br>Round-robin | DateKey<br>[✓ ProductKey]<br>RegionKey |
| Invoices | Hash-distributed<br>[✓ Round-robin] | DateKey<br>ProductKey<br>[✓ RegionKey] |

Box 1: Hash-distributed -

Box 2: ProductKey -

ProductKey is used extensively in joins.

Hash-distributed tables improve query performance on large fact tables.

Box 3: Round-robin -

Box 4: RegionKey -

Round-robin tables are useful for improving loading speed.

Consider using the round-robin distribution for your table in the following scenarios:

✑ When getting started as a simple starting point since it is the default
✑ If there is no obvious joining key
✑ If there is not good candidate column for hash distributing the table
✑ If the table does not share a common join key with other tables
✑ If the join is less significant than other joins in the query
✑ When the table is a temporary staging table
Note: A distributed table appears as a single table, but the rows are actually stored across 60 distributions. The rows are distributed with a hash or round-robin algorithm.
Reference:
https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-tables-distribute

**H_S** Highly Voted 4 years, 3 months ago
Table sales:
**Distribution type: Hash-Distributed
For 2 Reasons: the table is 600GB and we want to optimize queries
**Distribution column:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute
=> Product key is the only possible correct choice
Table Invoices:
**Distribution type: Hash-Distributed
The table size is more than 2GB and probably growing up.
Consider using a hash-distributed table when:The table size on disk is more than 2 GB. And The table has frequent insert, update, and delete operations.
**Distribution column: for sure it's regionkey To minimize data movement, select a distribution column that:… same link
upvoted 31 times

**cadio30** Highly Voted 4 years, 1 month ago
Distribution for both should be "hash-distributed" as we are talking about fact tables while round-robin is mostly use in staging tables. As a rule of the thumb when using hash-distributed it should be applied in the columns that uses JOIN, GROUP BY, DISTINCT, OVER, and HAVING and one shouldn't apply it in WHERE and DATE columns.

Sales: Hash-distributed, ProductKey
Invoices: Hash-distributed, RegiongKey

Reference: https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute
upvoted 9 times

**Qrm_1972** Most Recent 4 years ago
The correct answer is: Sales : Hash-Distributed>>>> Product Key
Invoices: Hash-Distributed>>>> Region Key
upvoted 2 times

**SrinivasR** 4 years, 1 month ago
i think Distribution should be : "HASH-DISTRIBUTION" as both are Fact tables and ProductKey for sales and Region Key for Invoices .
upvoted 2 times

**NarenG1** 4 years, 1 month ago
I don't think there is a distribution column option for Round Robin. The distribution column is available only for Hash Partitioning. So it must be Hash Partitioning & Region Key for Invoice table.
upvoted 2 times

**DataDani** 4 years, 1 month ago
As there are some different answers for table invoices.
For sure hash-distributed, as the table size is more than 2 GB.
Explanation for RegionKey:
To minimize data movement, select a distribution column that:
Is used in JOIN, GROUP BY, DISTINCT, OVER, and HAVING clauses. When two large fact tables have frequent joins, query performance improves when you distribute both tables on one of the join columns. When a table is not used in joins, consider distributing the table on a column that is frequently in the GROUP BY clause.

Is not used in WHERE clauses. This could narrow the query to not run on all the distributions.
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute
upvoted 2 times

**Geo_Barros** 4 years, 3 months ago
I think that the right answer for the ditribution type at the invoice table would be hash-distributed with regionkey as the distributed key as it is used for grouping.
upvoted 4 times

**Mariekumi** 4 years, 3 months ago
I would say Hash distributed and Date key for both tables because date key is used extensively in queries in both tables, region key will result in skewed partitioning as 75% of data falls in one region. Also Hash is best for both because we are optimizing query performance and not loading which Round-Robin is best suited for
upvoted 5 times

**JohnCrawford** 4 years, 2 months ago
From the provided link we learn that generally we should not use date values as the partitioning key. As noted by H_S the Invoices table is large enough to warrant being hash distributed as well and as noted by you, Mariekumi, RegionKey would result in hot spots/skew. I think hash distributed on product key for both tables makes the most sense.
upvoted 4 times

**akram786** 4 years, 3 months ago
why round robin for invoices.
upvoted 1 times

Question #10 *Topic 1*

You are designing a data processing solution that will implement the lambda architecture pattern. The solution will use Spark running on HDInsight for data processing.

You need to recommend a data storage technology for the solution.

Which two technologies should you recommend? Each correct answer presents a complete solution.

NOTE: Each correct selection is worth one point.

A. Azure Cosmos DB

B. Azure Service Bus

C. Azure Storage Queue

D. Apache Cassandra

E. Kafka HDInsight

**Suggested Answer:** *AE*

To implement a lambda architecture on Azure, you can combine the following technologies to accelerate real-time big data analytics:

✑ Azure Cosmos DB, the industry's first globally distributed, multi-model database service.

✑ Apache Spark for Azure HDInsight, a processing framework that runs large-scale data analytics applications

Azure Cosmos DB change feed, which streams new data to the batch layer for HDInsight to process

▪

✑ The Spark to Azure Cosmos DB Connector

E: You can use Apache Spark to stream data into or out of Apache Kafka on HDInsight using DStreams.

Reference:

https://docs.microsoft.com/en-us/azure/cosmos-db/lambda-architecture

**Abhilvs** Highly Voted 👍 5 years ago

for batch processing - cosmos DB ,

for Stream processing - Kafka HDinsight

upvoted 26 times

**mclawson1966** Highly Voted 👍 5 years, 3 months ago

Is Kafka considered a data storage solution? I thought it was a streaming technology.

upvoted 11 times

**JamesCho** 5 years, 1 month ago

https://www.confluent.io/blog/okay-store-data-apache-kafka/ [ it states something like this - "It is much closer in architecture to a distributed filesystem or database then to traditional message queue." ]

upvoted 2 times

**Wendy_DK** Most Recent ⊙ 4 years, 1 month ago

Question here is :You need to recommend a data storage technology for the solution.

Answer: cosmos DB and Blob blob. Yet Azure Kafka is for stream processing

upvoted 1 times

**sjain91** 4 years, 2 months ago

for batch: Cosmos DB

for stream: Kafka HD insight

upvoted 2 times

**davita8** 4 years, 2 months ago

A. Azure Cosmos DB

D. Apache Cassandra

upvoted 3 times

**Deepu1987** 4 years, 4 months ago

Given solution is right & pls go through this link

https://www.bluegranite.com/blog/exploring-the-lambda-architecture-in-azure

Kafka hdsight is for ingestion

Cosmos DB for processing

upvoted 3 times

**syu31svc** 4 years, 6 months ago

https://www.bluegranite.com/blog/exploring-the-lambda-architecture-in-azure

Kafka for ingestion

As for processing, Cosmos DB would be it

upvoted 1 times

**Tombarc** 5 years, 2 months ago

Lambda architecture is usually built with Cassandra as a storage solution and Kafka as a Data stream technology, so Cosmos DB is the correct answer. There is no such thing as Apache Cassandra.

upvoted 8 times

**chaoxes** 4 years, 6 months ago

What do you mean? There is Apache Cassandra - a distributed, wide column storage on Apache license.

However, Cosmos DB & HDI Kafka are the answers for this question.

upvoted 2 times

## Question #11 *Topic 1*

A company manufactures automobile parts. The company installs IoT sensors on manufacturing machinery.

You must design a solution that analyzes data from the sensors.

You need to recommend a solution that meets the following requirements:

✑ Data must be analyzed in real-time.

✑ Data queries must be deployed using continuous integration.

✑ Data must be visualized by using charts and graphs.

✑ Data must be available for ETL operations in the future.

✑ The solution must support high-volume data ingestion.

Which three actions should you recommend? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

A. Use Azure Analysis Services to query the data. Output query results to Power BI.

B. Configure an Azure Event Hub to capture data to Azure Data Lake Storage.

C. Develop an Azure Stream Analytics application that queries the data and outputs to Power BI. Use Azure Data Factory to deploy the Azure Stream Analytics application.

D. Develop an application that sends the IoT data to an Azure Event Hub.

E. Develop an Azure Stream Analytics application that queries the data and outputs to Power BI. Use Azure Pipelines to deploy the Azure Stream Analytics application.

F. Develop an application that sends the IoT data to an Azure Data Lake Storage container.

**Suggested Answer:** *BCD*

*Community vote distribution*

BDE (100%)

**uge** Highly Voted 5 years, 8 months ago

Reading"Data queries must be deployed using continuous integration", i think than correct answer it´s BDE and not BCD.

upvoted 120 times

**Paakofi** Highly Voted 4 years, 9 months ago

Pipelines are subset activity in Azure factory, so C is correct which makes the answer BCD correct.

upvoted 14 times

**Ashtrixx** 4 years ago

ADF pipelines are used for ETL jobs and all, not for CI/CD, for that we need to use pipelines in azure devops

upvoted 1 times

**nefarious_smalls** Most Recent 3 years, 1 month ago

Selected Answer: BDE

I believe it is BDE

upvoted 1 times

**ismaelrihawi** 4 years, 1 month ago

CI = Azure Pipeline!

upvoted 1 times

**cadio30** 4 years, 1 month ago

B,D,E are correct

Azure pipeline is related to CI/CD process and one should differentiate the "pipeline" of Azure Data Factory. Also the azure stream analytics can output the data into PowerBI dataset. Hence, the Azure Analysis Service is not needed in the solution.

upvoted 2 times

**davita8** 4 years, 2 months ago

BDE is the correct answer.

upvoted 4 times

**chirag1234** 4 years, 2 months ago

Answer will be BDE

upvoted 2 times

**syu31svc** 4 years, 6 months ago

https://docs.microsoft.com/en-us/azure/data-factory/continuous-integration-deployment:

"There are two suggested methods to promote a data factory to another environment:

Automated deployment using Data Factory's integration with Azure Pipelines"

BDE it is

upvoted 2 times

**sandGrain** 4 years, 7 months ago

BDE is the correct answer.

upvoted 2 times

**Arsa** 4 years, 10 months ago

The correct answer should be BDE

Automate continuous integration by using Azure Pipelines releases

upvoted 3 times

**Taddi10** 4 years, 10 months ago

In the same time they said :Data must be available for ETL operations in the future.

So the response is OK

upvoted 1 times

**envy** 4 years, 11 months ago

Tutorial: Deploy an Azure Stream Analytics job with CI/CD using Azure Pipelines https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-tools-visual-studio-cicd-vsts

upvoted 2 times

> **envy** 4 years, 11 months ago
>
> https://docs.microsoft.com/en-us/azure/data-factory/continuous-integration-deployment
>
> upvoted 1 times

**bob_** 5 years ago

Continuous integration and deployment using Azure Data Factory | Azure Friday

https://www.youtube.com/watch?v=WhUAX8YxxLk

upvoted 3 times

> **Polash** 4 years, 12 months ago
>
> Right bob
>
> upvoted 2 times

**Runi** 5 years ago

I believe BDE is the right onec.

https://azure.microsoft.com/en-us/blog/refreshing-reference-data-with-azure-data-factory-for-azure-stream-analytics-jobs-3/

upvoted 3 times

**Abhitm** 5 years, 1 month ago

I think the correct answer is ABD

https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-power-bi-dashboard

upvoted 1 times

**zenomas** 5 years, 2 months ago

Should be E instead of C.

Data queries must be deployed using continuous integration.

Any thoughts?

upvoted 3 times

**Nehuuu** 5 years, 3 months ago

Answer - BDE

upvoted 7 times

## Question #12 *Topic 1*

You are designing an Azure Databricks interactive cluster.
You need to ensure that the cluster meets the following requirements:
✑ Enable auto-termination
✑ Retain cluster configuration indefinitely after cluster termination.
What should you recommend?

A. Start the cluster after it is terminated.

B. Pin the cluster

C. Clone the cluster after it is terminated.

D. Terminate the cluster manually at process completion.

**Suggested Answer:** *B*
To keep an interactive cluster configuration even after it has been terminated for more than 30 days, an administrator can pin a cluster to the cluster list.
Reference:
https://docs.azuredatabricks.net/user-guide/clusters/terminate.html

**Arsa** Highly Voted 4 years, 10 months ago
Pin a cluster
30 days after a cluster is terminated, it is permanently deleted. To keep an interactive cluster configuration even after a cluster has been terminated for more than 30 days, an administrator can pin the cluster. Up to 20 clusters can be pinned.
upvoted 24 times

**awron_durat** Highly Voted 5 years ago
You're trying to keep the configuration, not keep the cluster running. According to Satabricks, the answer is to pin the cluster.
https://docs.databricks.com/clusters/clusters-manage.html#pin-a-cluster
upvoted 10 times

**cadio30** Most Recent 4 years, 1 month ago
B. Pin the cluster is the appropriate answer for the requirement

Reference: https://docs.databricks.com/clusters/clusters-manage.html#pin-a-cluste
upvoted 3 times

**chaoxes** 4 years, 6 months ago
B. Pin the cluster is an answer.

To keep a cluster configuration even after a cluster has been terminated (which is after 30 days) administrator must pin the cluster.

Source: https://docs.microsoft.com/en-us/azure/databricks/clusters/clusters-manage
upvoted 2 times

**Wirehinge** 4 years, 12 months ago
Databricks documentation points to the use of pinning as the way to keep configurations: https://docs.databricks.com/clusters/index.html - 'Important!' section.
upvoted 4 times

**shampoolegend** 5 years ago
according to the instructor at one of the training sessions provided by MS, the answer is A.
Here is a screenshot from the session:
https://ibb.co/LgL5gCz
upvoted 6 times

> **mojedapr** 4 years, 9 months ago
> just watched the session and you are right !
> upvoted 3 times

Question #13 *Topic 1*

You are designing a solution for a company. The solution will use model training for objective classification.
You need to design the solution.
What should you recommend?

A. an Azure Cognitive Services application

B. a Spark Streaming job

C. interactive Spark queries

D. Power BI models

E. a Spark application that uses Spark MLib.

**Suggested Answer:** *E*
Spark in SQL Server big data cluster enables AI and machine learning.
You can use Apache Spark MLlib to create a machine learning application to do simple predictive analysis on an open dataset.
MLlib is a core Spark library that provides many utilities useful for machine learning tasks, including utilities that are suitable for:
✑ Classification
✑ Regression
✑ Clustering
✑ Topic modeling
✑ Singular value decomposition (SVD) and principal component analysis (PCA)
✑ Hypothesis testing and calculating sample statistics
Reference:
https://docs.microsoft.com/en-us/azure/hdinsight/spark/apache-spark-machine-learning-mllib-ipython

*Community vote distribution*

E (100%)

**agnaldo** Highly Voted 5 years, 3 months ago
ooops... the correct is Spark ML Lib
upvoted 18 times

**agnaldo** Highly Voted 5 years, 3 months ago
one observation: the correct is Apache ML Lib
upvoted 12 times

**brieucboonen1** Most Recent 3 years, 4 months ago
Selected Answer: E
one observation: the correct is Apache ML Lib
upvoted 1 times

**bdsrca** 3 years, 10 months ago
A. an Azure Cognitive Services application
upvoted 1 times

**cadio30** 4 years, 1 month ago
appropriate answer is E

Reference: https://docs.microsoft.com/en-us/azure/hdinsight/spark/apache-spark-machine-learning-mllib-ipython
upvoted 4 times

**sjain91** 4 years, 2 months ago
Spark Mlib should be the correct answer
upvoted 3 times

**Deepu1987** 4 years, 4 months ago
the keyword is Machine Learning "objective classification" to choose the ans choice Spark ML Lib
upvoted 1 times

**chaoxes** 4 years, 6 months ago

E. a Spark application that uses Spark MLib. is correct answer

upvoted 3 times

⊟ **syu31svc** 4 years, 6 months ago

Model training so Machine Learning is what should come into mind

E is the answer

upvoted 2 times

⊟ **Gluckos** 4 years, 9 months ago

Cognitive services.. it's possibile train model with Custom Vision API

upvoted 2 times

⊟ **Abhitm** 5 years, 1 month ago

It is for model training hence Spark ML Lib

upvoted 3 times

⊟ **serger** 5 years, 1 month ago

For training model this is Spark MLlib that contains ML models for spark. Cognitive services is not for training a new model but to use some existing pretrained models.

upvoted 4 times

⊟ **DrC** 5 years, 6 months ago

The Computer Vision API in Cognitive Services can do this too, but it's out of scope for this exam.

upvoted 6 times

⊟ **STH** 5 years, 7 months ago

Why not use Cognitive Services ? it is built to achieve such classification tasks, is'nt it ?

upvoted 4 times

⊟ **josecipiace** 5 years, 2 months ago

It says model training, if you need to do model training cognitive services are not the correct solutions. They are already trained. The question refers to a custom scenario.

upvoted 17 times

⊟ **Gluckos** 4 years, 9 months ago

Cognitive services.. it's possibile train model with Custom Vision API

upvoted 4 times

Question #14 *Topic 1*

A company stores data in multiple types of cloud-based databases.
You need to design a solution to consolidate data into a single relational database. Ingestion of data will occur at set times each day.
What should you recommend?

A. SQL Server Migration Assistant

B. SQL Data Sync

C. Azure Data Factory

D. Azure Database Migration Service

E. Data Migration Assistant

**Suggested Answer:** *C*
Incorrect Answers:
D: Azure Database Migration Service is used to migrate on-premises SQL Server databases to the cloud.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/introduction https://azure.microsoft.com/en-us/blog/operationalize-azure-databricks-notebooks-using-data-factory/ https://azure.microsoft.com/en-us/blog/data-ingestion-into-azure-at-scale-made-easier-with-latest-enhancements-to-adf-copy-data-tool/

**SidN** Highly Voted 5 years ago
Source data is stored on different cloud storage and you need migrate into relational database, so only Azure Data Factory can do this task
upvoted 36 times

**Israel2** 4 years, 11 months ago
The member databases can be either databases in Azure SQL Database or in instances of SQL Server. See https://docs.microsoft.com/en-us/azure/azure-sql/database/sql-data-sync-data-sql-server-sql-database. So it's probably not Data Sync.
upvoted 2 times

**Israel2** 4 years, 11 months ago
Quote from the website: "The member databases can be either databases in Azure SQL Database or in instances of SQL Server."
upvoted 1 times

**D_Duke** Highly Voted 4 years, 8 months ago
Here are my thoughts: we know that A and E are obviously incorrect, and since this is a single-selection question, we've got to select the most appropriate answer from the rest choices. SQL Data Sync can only be used to sync data between a Hub Database and a Member Database, and the Hub Database must be an Azure SQL Database. Azure Database Migration Service is mainly used to migrate data from on-prem RDMS to Azure Database or from MongoDB to Azure Cosmos DB (you can still do cloud-to-cloud migrations with it but there are strict network topology requirements applied). Considering that we have various types of databases in Azure and the consolidation requirements are not clear, Azure Data Factory is the most universal solution, so the answer is C.
upvoted 8 times

**sjain91** Most Recent 4 years, 2 months ago
Azure data factory can be used to connect to multiple databases, hence the ideal solution for sourcing data on multiple cloud databases
upvoted 2 times

**Deepu1987** 4 years, 4 months ago
Normally ADF is recommended for ingestion when when mult.. cluod dbs are stored
upvoted 1 times

**syu31svc** 4 years, 6 months ago
https://docs.microsoft.com/en-us/azure/data-factory/copy-activity-overview#supported-data-stores-and-formats
Data Factory is correct
upvoted 2 times

**Abhilvs** 5 years ago
The data is already hosted in cloud data stores, SQL data sync appeal to customers who are considering moving to the cloud and would like to put some of their application in Azure, So Azure Datafactory is appropriate here
upvoted 6 times

⊟ **Runi** 5 years ago

why not Data sync?

https://docs.microsoft.com/en-us/azure/azure-sql/database/sql-data-sync-data-sql-server-sql-database

upvoted 2 times

⊟ **drdean** 5 years ago

I think you might be right. https://www.bdo.com/digital/insights/cloud/azure-sql-data-sync-is-now-generally-available

This article highlights the possibility

upvoted 2 times

⊟ **Jzerpa_ccs** 4 years, 8 months ago

A company stores data in multiple types of cloud-based databases

"Oracle", "MySQL", "Postgres"

upvoted 1 times

⊟ **Yaswant** 4 years, 10 months ago

Though we can achieve this using data sync but it isn't recommended for this scenario. As data sync is mostly used for synchronizing but not transfer related activities.

upvoted 6 times

Question #15 Topic 1

HOTSPOT -

You manage an on-premises server named Server1 that has a database named Database1. The company purchases a new application that can access data from

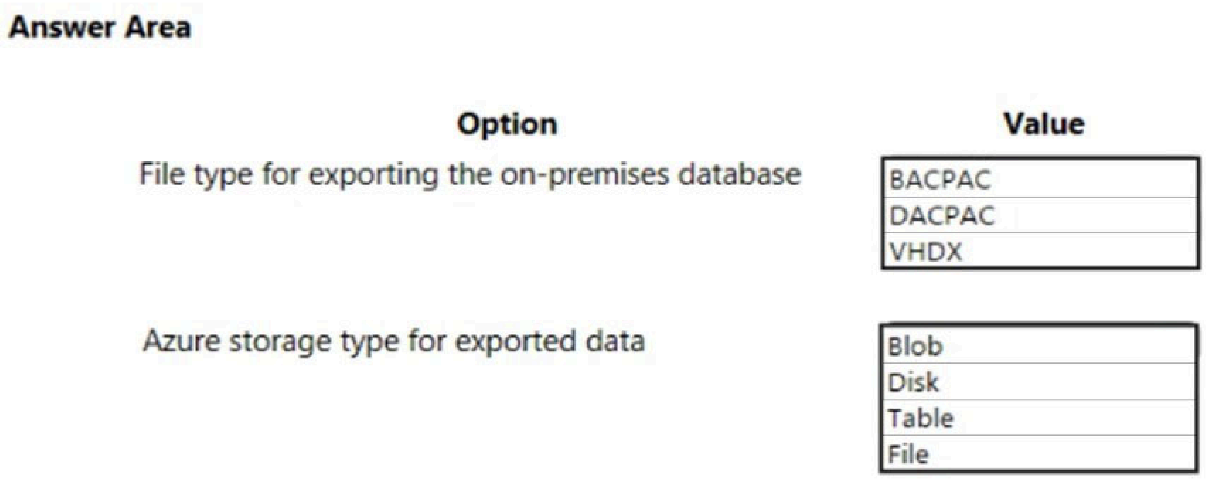
Azure SQL Database.

You recommend a solution to migrate Database1 to an Azure SQL Database instance.

What should you recommend? To answer, select the appropriate configuration in the answer area.

NOTE: Each correct selection is worth one point.

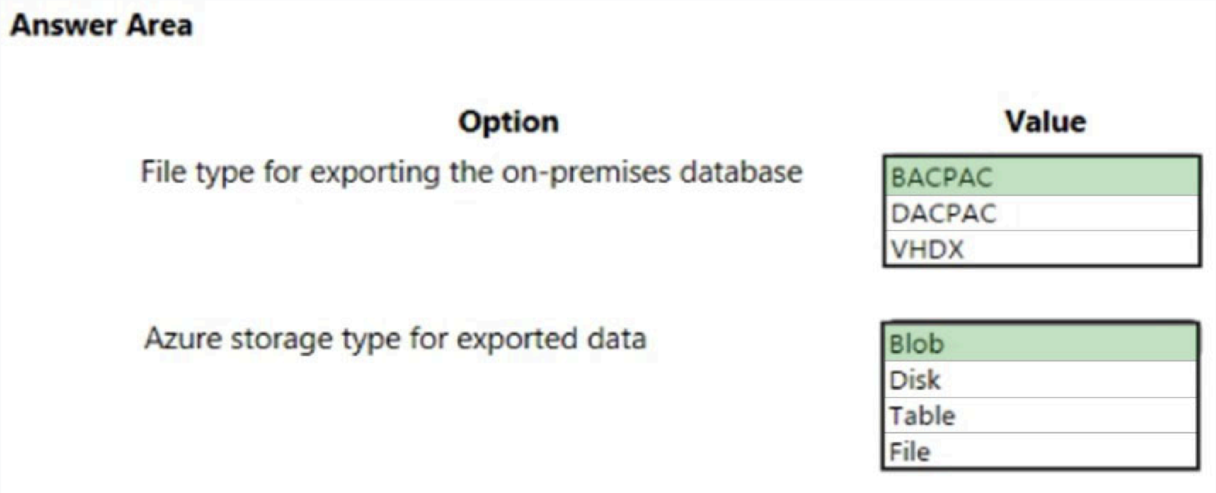
Hot Area:



**Answer Area**

| Option | Value |
|---|---|
| File type for exporting the on-premises database | BACPAC<br>DACPAC<br>VHDX |
| Azure storage type for exported data | Blob<br>Disk<br>Table<br>File |

**Suggested Answer:**



**Answer Area**

| Option | Value |
|---|---|
| File type for exporting the on-premises database | [✓ BACPAC]<br>DACPAC<br>VHDX |
| Azure storage type for exported data | [✓ Blob]<br>Disk<br>Table<br>File |

References:

https://docs.microsoft.com/en-us/azure/sql-database/sql-database-import

**Yaswant** Highly Voted 4 years, 10 months ago

The explaination is

If we have an on-premises database and we have a new app(ex: app1). Now app1 can only access data from azure databases. So our plan is to

1. Import data from onpremises to azure sql db.

Backup the on premises db into a bacpac file -> Upload the bacpac file to a blob storage container -> Go to azure sql db you created and in overview click on backup and give the information about sa( blob container) and click backup.

upvoted 47 times

**chaoxes** Highly Voted 4 years, 6 months ago

Box 1: BACKPACK

Box2: BLOB

upvoted 8 times

**cadio30** Most Recent 4 years ago

BACPAC contains the schema and data of the specific database while Azure Blob Storage is the storage that can handle such file and size.

upvoted 1 times

⊟ **sjain91** 4 years, 2 months ago

1: Bacpac

2: Blob

upvoted 1 times

Question #16 *Topic 1*

You are designing an application. You plan to use Azure SQL Database to support the application.
The application will extract data from the Azure SQL Database and create text documents. The text documents will be placed into a cloud-based storage solution.
The text storage solution must be accessible from an SMB network share.
You need to recommend a data storage solution for the text documents.
Which Azure data storage type should you recommend?

A. Queue

B. Files

C. Blob

D. Table

**Suggested Answer:** *B*
Azure Files enables you to set up highly available network file shares that can be accessed by using the standard Server Message Block (SMB) protocol.
Incorrect Answers:
A: The Azure Queue service is used to store and retrieve messages. It is generally used to store lists of messages to be processed asynchronously.
C: Blob storage is optimized for storing massive amounts of unstructured data, such as text or binary data. Blob storage can be accessed via HTTP or HTTPS but not via SMB.
D: Azure Table storage is used to store large amounts of structured data. Azure tables are ideal for storing structured, non-relational data.
Reference:
https://docs.microsoft.com/en-us/azure/storage/common/storage-introduction https://docs.microsoft.com/en-us/azure/storage/tables/table-storage-overview

**lovely** Highly Voted 5 years, 5 months ago
https://docs.microsoft.com/en-us/azure/storage/files/storage-how-to-use-files-windows
upvoted 13 times

**serger** Highly Voted 5 years ago
Yes SMB only available with FILES.
upvoted 13 times

**sjain91** Most Recent 4 years, 2 months ago
SMB is only available with Files as we are using test documents storage
upvoted 1 times

**Deepu1987** 4 years, 4 months ago
the keyword is SMB (Server Mesage Block) protocol based on it we can choose the ans choice FILES
https://docs.microsoft.com/en-us/azure/storage/files/storage-files-introduction pls refer this link
upvoted 2 times

**karishura** 4 years, 4 months ago
Correct answer
upvoted 2 times

**chaoxes** 4 years, 6 months ago
SMB protocol is for Azure Files. Files is the obvious answer.
upvoted 3 times

**syu31svc** 4 years, 6 months ago
Can only be files since SMB is mentioned
upvoted 4 times

**serger** 5 years, 1 month ago
SMB is available with FILES.
upvoted 3 times

Question #17 *Topic 1*

You are designing an application that will have an Azure virtual machine. The virtual machine will access an Azure SQL database. The database will not be accessible from the Internet.
You need to recommend a solution to provide the required level of access to the database.
What should you include in the recommendation?

A. Deploy an On-premises data gateway.

B. Add a virtual network to the Azure SQL server that hosts the database.

C. Add an application gateway to the virtual network that contains the Azure virtual machine.

D. Add a virtual network gateway to the virtual network that contains the Azure virtual machine.

**Suggested Answer:** *B*
When you create an Azure virtual machine (VM), you must create a virtual network (VNet) or use an existing VNet. You also need to decide how your VMs are intended to be accessed on the VNet.
Incorrect Answers:
C: Azure Application Gateway is a web traffic load balancer that enables you to manage traffic to your web applications.
D: A VPN gateway is a specific type of virtual network gateway that is used to send encrypted traffic between an Azure virtual network and an on-premises location over the public Internet.
Reference:
https://docs.microsoft.com/en-us/azure/virtual-machines/network-overview

**Yaswant** Highly Voted 4 years, 10 months ago
There are many ways to achieve this
1. One way is to use a service endpoint along with service end point policies.
2. Second way is by using an azure private link.
3. Third way is to go to azure sql server where the db is hosted and add a virtual network using firewall/virtual_network blade.
upvoted 23 times

**Treadmill** 4 years, 10 months ago
B correct: agree with Yaswant, but haven't found a source that presents the three options. I'd say A, C and D are not correct and therefore B is correct.

Here a graph:

https://azure.microsoft.com/de-de/blog/vnet-service-endpoints-for-azure-sql-database-now-generally-available/
upvoted 19 times

**syu31svc** Highly Voted 4 years, 6 months ago
B is the best answer
The on-premises data gateway acts as a bridge. It provides quick and secure data transfer between on-premises data, which is data that isn't in the cloud, and several Microsoft cloud services
Azure Application Gateway is a web traffic load balancer that enables you to manage traffic to your web applications
A VPN gateway is a specific type of virtual network gateway that is used to send encrypted traffic between an Azure virtual network and an on-premises location over the public Internet
upvoted 8 times

**watata** Most Recent 4 years, 4 months ago
you need an endpoint in a VNet
upvoted 1 times

**Deepu1987** 4 years, 4 months ago
the mentioned ans is correct pls check the link https://azure.microsoft.com/de-de/blog/vnet-service-endpoints-for-azure-sql-database-now-generally-available/ and line
It's shown in the picture clearly where C& D options dont relate at all
upvoted 1 times

**chaoxes** 4 years, 6 months ago

B. Add a virtual network to the Azure SQL server that hosts the database.

Is the correct answer.

upvoted 4 times

**Rohit77** 4 years, 11 months ago

The correct answer is "C"

upvoted 2 times

**hariharan1983_n** 4 years, 2 months ago

Please Note the option C and D is for virtual machine not for database.

upvoted 2 times

Question #18 *Topic 1*

HOTSPOT -

You are designing an application that will store petabytes of medical imaging data

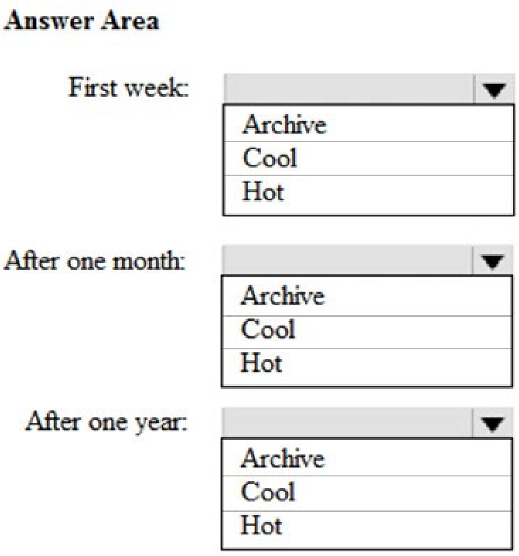
When the data is first created, the data will be accessed frequently during the first week. After one month, the data must be accessible within 30 seconds, but files will be accessed infrequently. After one year, the data will be accessed infrequently but must be accessible within five minutes.

You need to select a storage strategy for the data. The solution must minimize costs.

Which storage tier should you use for each time frame? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:



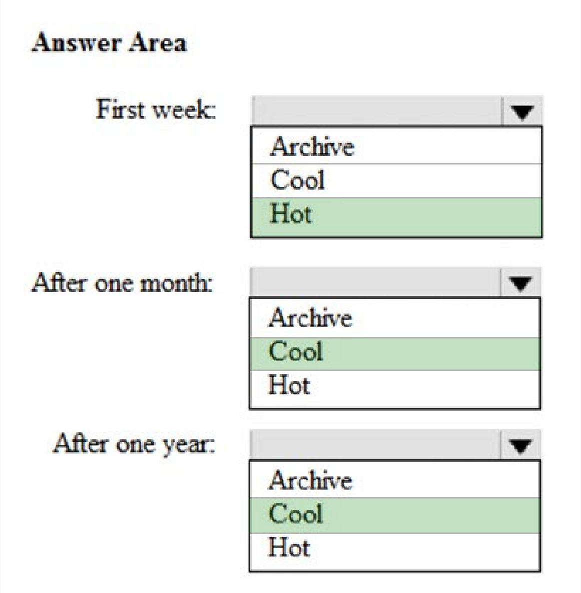
**Answer Area**

First week: Archive / Cool / Hot

After one month: Archive / Cool / Hot

After one year: Archive / Cool / Hot

**Suggested Answer:**



**Answer Area**

First week: Archive / Cool / [✓ Hot]

After one month: Archive / [✓ Cool] / Hot

After one year: Archive / [✓ Cool] / Hot

First week: Hot -

Hot - Optimized for storing data that is accessed frequently.

After one month: Cool -

Cool - Optimized for storing data that is infrequently accessed and stored for at least 30 days.

After one year: Cool -

Incorrect Answers:

Archive: Optimized for storing data that is rarely accessed and stored for at least 180 days with flexible latency requirements (on the order of hours).

References:

https://docs.microsoft.com/en-us/azure/storage/blobs/storage-blob-storage-tiers

**samok** Highly Voted 👍 5 years, 2 months ago

This is correct. Look at this from the docs:

"The archive access tier has the lowest storage cost. But it has higher data retrieval costs compared to the hot and cool tiers. Data in the archive tier can take several hours to retrieve."

upvoted 61 times

**Yaswant** Highly Voted 👍 4 years, 10 months ago

Hot : Frequent

Cool : Infrequent (30 days)

Archive : Rare (180 days) -> Data from archive tier can only be accessed by rehydrating the blob which may take up to several hours.

upvoted 15 times

**sjain91** Most Recent 🕒 4 years, 2 months ago

We can't use archive because data must be accessible. So:

Box 1: Hot

Box 2: Cold

Box 3: Cold

upvoted 5 times

**zic04** 4 years, 4 months ago

Hot cold cold : Correct

upvoted 1 times

**chaoxes** 4 years, 6 months ago

We can't use archive because data must be accessible. So:

Box 1: Hot

Box 2: Cold

Box 3: Cold

upvoted 3 times

**Andrexx** 4 years, 8 months ago

I agree with the answer by the reasons explained in the comments.

upvoted 1 times

**Rohit77** 4 years, 11 months ago

After One year The correct Answer is Archieve

upvoted 3 times

> **RajdeepRoy** 4 years, 11 months ago
>
> Data in the archive tier can take several hours to retrieve. But here the question asks to retrieve data in 5 minutes, hence it should be Cool Tire. Given answer is correct.
>
> upvoted 22 times
>
> **opawale** 4 years, 10 months ago
>
> Apart from the reason given by RajdeepRoy, it costs more to retrieve data from archived data and that would not be in line with the requirement of minimizing the cost.
>
> upvoted 5 times

**serger** 5 years, 1 month ago

Yes correct answer

upvoted 5 times

## Question #19 *Topic 1*

You are designing a data store that will store organizational information for a company. The data will be used to identify the relationships between users. The data will be stored in an Azure Cosmos DB database and will contain several million objects.

You need to recommend which API to use for the database. The API must minimize the complexity to query the user relationships. The solution must support fast traversals.

Which API should you recommend?

A. MongoDB

B. Table

C. Gremlin

D. Cassandra

**Suggested Answer:** *C*

Gremlin features fast queries and traversals with the most widely adopted graph query standard.

Reference:

https://docs.microsoft.com/th-th/azure/cosmos-db/graph-introduction?view=azurermps-5.7.0

**chaoxes** Highly Voted 4 years, 6 months ago

C. Gremlin API

When we talk about relationships and/or edges/nodes, Gremin API is the answer

upvoted 15 times

**syu31svc** Highly Voted 4 years, 6 months ago

identify the relationships between users

Answer is C

upvoted 7 times

## Question #20 *Topic 1*

HOTSPOT -

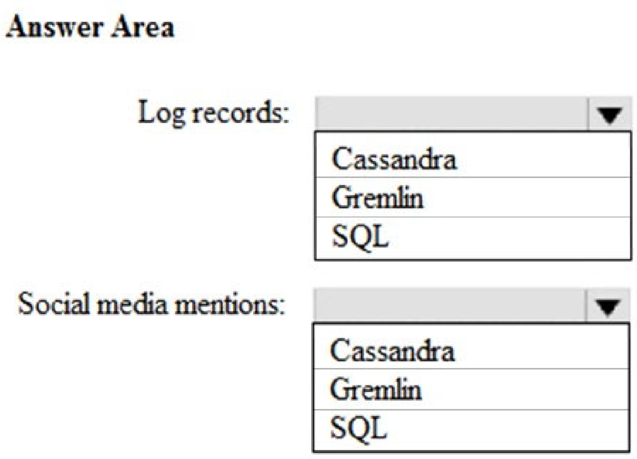
You are designing a new application that uses Azure Cosmos DB. The application will support a variety of data patterns including log records and social media relationships.

You need to recommend which Cosmos DB API to use for each data pattern. The solution must minimize resource utilization.

Which API should you recommend for each data pattern? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

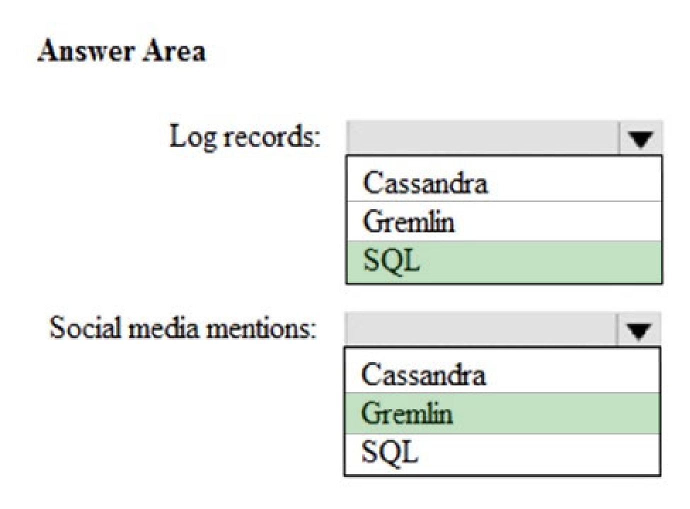


**Answer Area**

Log records: ▼
- Cassandra
- Gremlin
- SQL

Social media mentions: ▼
- Cassandra
- Gremlin
- SQL

**Suggested Answer:**



**Answer Area**

Log records: ▼
- Cassandra
- Gremlin
- [✓ SQL]

Social media mentions: ▼
- Cassandra
- [✓ Gremlin]
- SQL

Log records: SQL -

Social media mentions: Gremlin -

You can store the actual graph of followers using Azure Cosmos DB Gremlin API to create vertexes for each user and edges that maintain the "A-follows-B" relationships. With the Gremlin API, you can get the followers of a certain user and create more complex queries to suggest people in common. If you add to the graph the Content Categories that people like or enjoy, you can start weaving experiences that include smart content discovery, suggesting content that those people you follow like, or finding people that you might have much in common with.

Reference:

https://docs.microsoft.com/en-us/azure/cosmos-db/social-media-apps

⊟ **chaoxes** Highly Voted 4 years, 6 months ago

Log records: SQL (because logs are row oriented)

Social Media: Gremlin (because it uses relationships)

upvoted 27 times

⊟ **cadio30** Most Recent 4 years, 1 month ago

propose solution is correct.

Cassandra API is mostly used to handle high volume and real time data while the requirement is related to log records, SQL API is sufficient in this terms. Gremlin API is appropriate when we are talking about connection between entities.

Reference: https://acloudguru.com/blog/engineering/azure-cosmos-db-apis-use-cases-and-trade-offs

upvoted 4 times

⊟ **klasius** 5 years, 3 months ago

why SQL over Cassandra for log records?

upvoted 3 times

⊟ **z8zhong** 5 years, 3 months ago

log data are row-oriented so SQL handle them better, Cassandra are mainly for column-oriented data

upvoted 46 times

⊟ **kempstonjoystick** 5 years, 3 months ago

It also states in the documentation that Cassandra is only recommend to migrate existing Cassandra databases to CosmosDB. In all other cases, the SQL Api is recommended.

https://docs.microsoft.com/en-us/learn/modules/choose-api-for-cosmos-db/3-analyze-the-decision-criteria

upvoted 40 times

Question #21 *Topic 1*

You need to recommend a storage solution to store flat files and columnar optimized files. The solution must meet the following requirements:
✑ Store standardized data that data scientists will explore in a curated folder.
✑ Ensure that applications cannot access the curated folder.
✑ Store staged data for import to applications in a raw folder.
✑ Provide data scientists with access to specific folders in the raw folder and all the content the curated folder.
Which storage solution should you recommend?

A. Azure Synapse Analytics

B. Azure Blob storage

C. Azure Data Lake Storage Gen2

D. Azure SQL Database

**Suggested Answer:** *B*
Azure Blob Storage containers is a general purpose object store for a wide variety of storage scenarios. Blobs are stored in containers, which are similar to folders.
Incorrect Answers:
C: Azure Data Lake Storage is an optimized storage for big data analytics workloads.
Reference:
https://docs.microsoft.com/en-us/azure/architecture/data-guide/technology-choices/data-storage

**Sam9999** Highly Voted 5 years, 3 months ago
Shouldn't answer be C, there is no concept of folders and folder permissions in Azure storage.
upvoted 112 times

**Marcus1612** 3 years, 9 months ago
RBAC Security on Azure Blob can be scoped at the container level or above. With two containers (one for raw data and one for curated data) without folders, it would be possible to manage the security. BUT the current use case states that: " datascientists need access to specifics folders in the "raw" folder. You cannot manage security at this level with Azure Blob. You have to use Azure Data Lake with RBAC/ACLs. The right answer is C
upvoted 1 times

**kempstonjoystick** 5 years, 3 months ago
I agree, Azure Data Lake Stroage includes ACLs which can be applied to folder structures, which Blob Storage does not. Therefore the security requirements mean the answer should be ADLS
upvoted 13 times

**MLCL** 5 years, 2 months ago
There is the notion of public anonymous access in blob storage as well as shared access signatures, and of course RBAC can be implemented through Azure AD for Blobs and Queues, so the security requirements can be met.
Check this doc : https://docs.microsoft.com/en-us/azure/storage/common/storage-auth-aad-rbac-portal
upvoted 10 times

**Yuri1101** 5 years, 2 months ago
Agree, especially it is only required to handle standardized data. There is no need to use ADLS.
upvoted 3 times

**Leonido** 5 years, 2 months ago
However, strictly speaking, in BLOB storage, data not stored in folders, just the name of the blob will include the folder name. So if the requirement is to store in folder, it have to be ADLS
upvoted 16 times

**Leonido** 5 years, 2 months ago
Also, in blob, without RBAC you can only grant permission to the level of container.
upvoted 11 times

**lingjun** 4 years, 7 months ago

When an Azure role is assigned to an Azure AD security principal, Azure grants access to those resources for that security principal. Access can be scoped to the level of the subscription, the resource group, the storage account, or an individual container or queue. An Azure AD security principal may be a user, a group, an application service principal, or a managed identity for Azure resources.
https://docs.microsoft.com/de-de/azure/storage/common/storage-auth-aad
upvoted 2 times

**tes** 4 years ago
there is, it is called container.
upvoted 2 times

**HeB** Highly Voted 5 years, 2 months ago
Answer should definitely be C, Azure Data Lake Storage Gen2.
upvoted 45 times

**tes** Most Recent 4 years ago
The given answer is wrong and it should be C. The answer given states container is same as folder but it is not. A folder can have sub folders and access can be given only to sub folder. Where as in containers there are no sub containers hence the answer is wrong. Folder however can be given access in ADLS Gen2 using ACL so when we have a straight forward answer, why go with assumtion that 'container is same as folder'
upvoted 3 times

**azurenav** 4 years ago
Azure Data Lake Store Gen2 is a superset of Azure Blob storage capabilities. In the list below, some of the key differences between ADLS Gen2 and Blob storage are summarized.

ADLS Gen2 supports ACL and POSIX permissions allowing for more granular access control compared to Blob storage.
ADLS Gen2 introduces a hierarchical namespace. This is a true file system, unlike Blob Storage which has a flat namespace. This capability has a significant impact on performance, especially in big data analytics scenarios.
ADLS Gen2 is an HDFS-compatible store. This means that Apache Hadoop services can use data stored in ADLS Gen2. Azure Blob storage is not Hadoop-compatible.
upvoted 1 times

**cadio30** 4 years ago
ADLS is the appropriate solution here as it has ACL function.
upvoted 1 times

**Arjun16** 4 years, 1 month ago
In Question they mentioned about flat files and columnar optimized files(Binary Files) and Containers are similar to folders, so Azure storage is Correct
upvoted 1 times

**cadio30** 4 years, 1 month ago
The requirements leads to using ADLS gen 2 as it can manage the folder level using ACL
upvoted 1 times

**davita8** 4 years, 2 months ago
C. Azure Data Lake Storage Gen2
upvoted 2 times

**rmk4ever** 4 years, 2 months ago
Columnar optimized file for Raw, enriched and curated structure with Folder level access
Ans is ADLS
ref:
https://www.dremio.com/data-lake/adls/
https://medium.com/microsoftazure/building-your-data-lake-on-adls-gen2-3f196fc6b430
upvoted 2 times

**Deepu1987** 4 years, 4 months ago
The given answer is correct as when you check the below link
https://docs.microsoft.com/en-us/azure/architecture/data-guide/technology-choices/data-storage it's clearly mentioned that ADLS can be used with certain restrictions it can be accessed via az synapse using poly base feature. There are certain performance tuning guidelines but in qn it's asked it need to be easily accessed by data scientistists as per the conditions we can go with blob storage
upvoted 1 times

**AyeshJr** 4 years, 4 months ago

I will choose Azure Datalake on the only fact that the question did ask for columnar optimized files and this is available in Datalake and not Azure storage account

upvoted 2 times

**mohowzeh** 4 years, 6 months ago

A folder can be created in a blob (e.g. via button "Create folder" in the portal) but such a folder is virtual. Using Azure Storage Explorer (presently v1.17.0), one can verify that an SAS can be created on a blob container, but not on a folder within a blob.

Still, multiple containers could be created where each container maps to one group of users in the security requirements. This is not forbidden in the question. If each container has one or more folders, all requirements would still be met, making answer B a "minimum viable answer".

However, I agree that answer C is the best and most flexible. Using Azure Storage Explorer, one can easily verify that the option "Manage Access Control Lists" is available on an individual folder.

upvoted 2 times

**M0e** 4 years, 8 months ago

The given answer is clearly incorrect. All the points that are mentioned in the questions are hints to use ADLS Gen 2.

upvoted 3 times

**monumentalcrankiness** 4 years, 8 months ago

The answer also mentions that the files are supposed to be explored by Data Scientists in curated folder. ADLS Gen 2 hooked up with Databricks or Azure Synapse Analytics is a ready-made solution for this kind of exploration.

upvoted 3 times

**monumentalcrankiness** 4 years, 8 months ago

I think correct answer should be ADLS Gen 2.

upvoted 2 times

**yilpiz** 4 years, 10 months ago

raw, curated folder, folder level access all characteristics of ADLS

upvoted 6 times

**Bob123456** 4 years, 10 months ago

I believe there is actually only a single layer of containers. You can virtually create a "file-system" like layered storage, but in reality everything will be in 1 layer, the container in which it is.

So Answer should be DATA LAKE

upvoted 1 times

Question #22 *Topic 1*

Your company is an online retailer that can have more than 100 million orders during a 24-hour period, 95 percent of which are placed between 16:30 and 17:00.

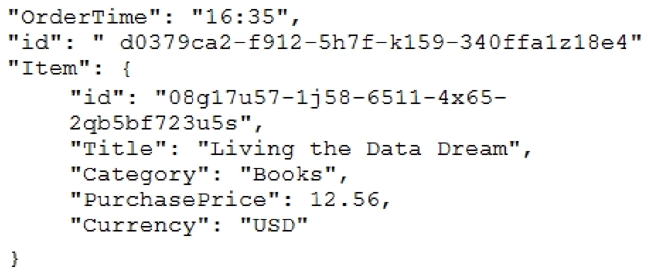
All the orders are in US dollars. The current product line contains the following three item categories:

✑ Games with 15,123 items

✑ Books with 35,312 items

✑ Pens with 6,234 items

You are designing an Azure Cosmos DB data solution for a collection named Orders Collection. The following documents is a typical order in Orders Collection.



```
"OrderTime": "16:35",
"id": " d0379ca2-f912-5h7f-k159-340ffa1z18e4"
"Item": {
     "id": "08g17u57-1j58-6511-4x65-
     2qb5bf723u5s",
     "Title": "Living the Data Dream",
     "Category": "Books",
     "PurchasePrice": 12.56,
     "Currency": "USD"
}
```

Orders Collection is expected to have a balanced read/write-intensive workload.

Which partition key provides the most efficient throughput?

A. Item/Category

B. OrderTime

C. Item/Currency

D. Item/id

**Suggested Answer:** *A*

Choose a partition key that has a wide range of values and access patterns that are evenly spread across logical partitions. This helps spread the data and the activity in your container across the set of logical partitions, so that resources for data storage and throughput can be distributed across the logical partitions.

Choose a partition key that spreads the workload evenly across all partitions and evenly over time. Your choice of partition key should balance the need for efficient partition queries and transactions against the goal of distributing items across multiple partitions to achieve scalability.

Candidates for partition keys might include properties that appear frequently as a filter in your queries. Queries can be efficiently routed by including the partition key in the filter predicate.

Reference:

https://docs.microsoft.com/en-us/azure/cosmos-db/partitioning-overview#choose-partitionkey

**kempstonjoystick** Highly Voted 👍 5 years, 3 months ago

Given there are 100 million orders in a 24 hour period, and there are only three catgegories, is Item/Id not a better solution, otherwise the category will cause significant hotspots?

upvoted 67 times

**Taddi10** 4 years, 11 months ago

I think if the id was an integrer (inremental foe exemple ) it can be a good partition key but with this format i think category is the best choice

upvoted 7 times

**MamadouNiang** Highly Voted 👍 5 years, 1 month ago

2 paragraphs below the link given in microsoft docs, there is an interesting answer :

Using item ID as the partition key

If your container has a property that has a wide range of possible values, it is likely a great partition key choice. One possible example of such a property is the item ID. For small read-heavy containers or write-heavy containers of any size, the item ID is naturally a great choice for the partition key.

The item ID is a great partition key choice for the following reasons:

There are a wide range of possible values (one unique item ID per item).
Because there is a unique item ID per item, the item ID does a great job at evenly balancing RU consumption and data storage.
You can easily do efficient point reads since you'll always know an item's partition key if you know its item ID.
upvoted 35 times

**Treadmill** 4 years, 10 months ago
D correct: Source as above quoted https://docs.microsoft.com/en-us/azure/cosmos-db/partitioning-overview
upvoted 2 times

**IAMKPR** Most Recent 4 years, 1 month ago
Answer should be "item/id". You can find almost similar example in below link.
https://docs.microsoft.com/en-us/learn/modules/monitor-and-scale-cosmos-db/5-partition-lesson
upvoted 5 times

**MMM777** 4 years ago
This example is definitely VERY similar to the question and explains why several of the proposed values are not good choices, and also shows "Item/Id" to be a decent choice.
upvoted 1 times

**davita8** 4 years, 2 months ago
D. Item/id is the answer
upvoted 6 times

**Deepu1987** 4 years, 4 months ago
Given solution is right where we choose the item/category. It's explained in detail in the below link https://medium.com/walmartglobaltech/deep-dive-azure-cosmos-partitions-and-partitionkey-14e898f371cd this concept is of major focus as question may not be exactly asked in exam we need to need to know the concept of physical & logical partitions pre-requisites & Partition key as well.
upvoted 1 times

**BobFar** 4 years, 1 month ago
the item/id is the correct solution, regarding to the explanation in the link that you posted, all the documents related to the item/id will store in same partition.
upvoted 1 times

**TaherAli2020** 4 years, 4 months ago
If you use the Item/Category property as a partition key, then it has a small cardinality. Even if the documents are evenly distributed across the collection, for large collections, any category might outgrow a single partition.

If the categories aren't evenly distributed across the documents in the collection, then the problem is even worse. The dominant category restricts the ability of Azure Cosmos DB to scale.

Item/Category is not a good choice for the partition key.
https://docs.microsoft.com/en-us/learn/modules/monitor-and-scale-cosmos-db/5-partition-lesson
upvoted 14 times

**cadio30** 4 years ago
perfect! the link provided clear states the strategy of optimizing partition.
upvoted 2 times

**tejasjoshi** 3 years, 11 months ago
Superb ! Its crystal clear now. Partition should be on Item/id. Requesting all to go through above link.
upvoted 2 times

**sturcu** 4 years, 4 months ago
Nice link. It is exactly the case from the ex.
upvoted 1 times

**TkSQL** 4 years, 1 month ago
this link is the answer to all the confusion here
upvoted 3 times

**syu31svc** 4 years, 6 months ago
From https://docs.microsoft.com/en-us/azure/cosmos-db/partitioning-overview#choose-partitionkey:
"Have a high cardinality. In other words, the property should have a wide range of possible values."

D is the answer

upvoted 4 times

**brcdbrcd** 4 years, 7 months ago

item/id for sure.
see the section "Propose partition key values for the collection" at:
https://docs.microsoft.com/en-us/learn/modules/monitor-and-scale-cosmos-db/5-partition-lesson

upvoted 5 times

**lingjun** 4 years, 7 months ago

Candidates for partition keys might include properties that appear frequently as a filter in your queries. Queries can be efficiently routed by including the partition key in the filter predicate.
Item ID will not appear as a filter most likely

upvoted 1 times

**lingjun** 4 years, 7 months ago

For small read-heavy containers or write-heavy containers of any size, Item-ID is naturally good choice. In this case, we have balanced read/write workload

upvoted 1 times

**M0e** 4 years, 8 months ago

Given the discussion here: https://docs.microsoft.com/en-us/learn/modules/monitor-and-scale-cosmos-db/5-partition-lesson, "Item/id" is the correct answer

upvoted 12 times

**monumentalcrankiness** 4 years, 8 months ago

I shall go with D. Item/Id
Item/Category is out. It will only create 3 logical partitions, that also unevenly distributed. A logical distribution has a size cap of 20 GB. With 100 million orders per day, it won't be very hard to reach that limit quickly.

OrderTime is out. 16:30 to 17:00 spike shall create a hotspot problem.

Item/Currency is out. Only 1 value "USD" will result in everything cramming up one logical partition.

Only Item/id is left. So this is the answer.

upvoted 8 times

**Shivam131** 4 years, 9 months ago

your partition key should:

Be a property that has a value which does not change. If a property is your partition key, you can't update that property's value.
Have a high cardinality. In other words, the property should have a wide range of possible values.
Spread request unit (RU) consumption and data storage evenly across all logical partitions. This ensures even RU consumption and storage distribution across your physical partitions.

upvoted 1 times

**Ash666** 4 years, 10 months ago

https://docs.microsoft.com/en-us/azure/cosmos-db/partition-data#logical-partitions
https://docs.microsoft.com/en-us/azure/cosmos-db/partitioning-overview
https://www.examtopics.com/exams/microsoft/dp-201/view/6/

D Item/ID

Category doesn't distribute RU evenly across partitions. Low cardinality.

upvoted 3 times

**Yaswant** 4 years, 10 months ago

Consider we have provisioned a throughput of 1200 request units and we know that throughput can be provisioned in cosmos db only at a container level or at a database level.
In our case we consider our online retailer to be Walkart. Now walkart has an account in cosmosdb and they have a document db with coresql api.
Now walkart has created a container named orders in their cosmos account and provisioned 1200ru's.
Now consider the case of choosing a partition key. Considering they have 1200 customer id-s and if they use id as partition key they will have their

throughput spread across partitions which makes their unused throughput in vain as customers come buy and go and it makes a hotspot. Now if we choose product category as partition we'll be having a balanced throughput and read-write.

upvoted 1 times

⊟ **krisspark** 4 years, 11 months ago

these comments causing further confusing for new bees as it's not able to draw whats final correct answer.. I would go by Item/Category only... this combo may not give repeated values as item would be different in same category.. item/id might create super heavy number of partitions

upvoted 1 times

⊟ **LeonLeon** 5 years ago

In this case A is correct indeed. See the reference and be aware of the read/write balancing. The read is as important as the throuput.

Partition keys for read-heavy containers
For most containers, the above criteria is all you need to consider when picking a partition key. For large read-heavy containers, however, you might want to choose a partition key that appears frequently as a filter in your queries. Queries can be efficiently routed to only the relevant physical partitions by including the partition key in the filter predicate.
If most of your workload's requests are queries and most of your queries have an equality filter on the same property, this property can be a good partition key choice. For example, if you frequently run a query that filters on UserID, then selecting UserID as the partition key would reduce the number of cross-partition queries

upvoted 7 times

⊟ **Sudipta3009** 4 years, 11 months ago

Ur explanation is correct

upvoted 1 times

⊟ **BHAWS** 5 years ago

Choose a partition key that has a wide range of values,so the data is evenly spread across logical partitioning. Hence I suggest the answer is item/category

upvoted 3 times

Question #23 *Topic 1*

You have a MongoDB database that you plan to migrate to an Azure Cosmos DB account that uses the MongoDB API.
During testing, you discover that the migration takes longer than expected.
You need to recommend a solution that will reduce the amount of time it takes to migrate the data.
What are two possible recommendations to achieve this goal? Each correct answer presents a complete solution.
NOTE: Each correct selection is worth one point.

A. Increase the Request Units (RUs).

B. Turn off indexing.

C. Add a write region.

D. Create unique indexes.

E. Create compound indexes.

**Suggested Answer:** *AB*
A: Increase the throughput during the migration by increasing the Request Units (RUs).
For customers that are migrating many collections within a database, it is strongly recommend to configure database-level throughput. You must make this choice when you create the database. The minimum database-level throughput capacity is 400 RU/sec. Each collection sharing database-level throughput requires at least 100 RU/sec.
B: By default, Azure Cosmos DB indexes all your data fields upon ingestion. You can modify the indexing policy in Azure Cosmos DB at any time. In fact, it is often recommended to turn off indexing when migrating data, and then turn it back on when the data is already in Cosmos DB.
Reference:
https://docs.microsoft.com/bs-latn-ba/Azure/cosmos-db/mongodb-pre-migration

**alexvno** Highly Voted 👍 4 years, 11 months ago
Correct
upvoted 20 times

**chaoxes** Highly Voted 👍 4 years, 6 months ago
A. Increase the request units (RUs)
B. Turn off indexing
upvoted 12 times

**cadio30** Most Recent ⏱ 4 years, 1 month ago
Propose solution is correct, by default azure cosmos db create an index though the feature could be toggle to prevent it from happening.

Reference: https://docs.microsoft.com/bs-latn-ba/Azure/cosmos-db/mongodb-post-migration
upvoted 1 times

**sjain91** 4 years, 2 months ago
turn off indexing and increase the request units - Answer A is correct
upvoted 1 times

**MSFTLearn** 4 years, 7 months ago
Indexing makes write operation slower.

"By default, indexing policy is set to automatic. It's achieved by setting the automatic property in the indexing policy to true. Setting this property to true allows Azure CosmosDB to automatically index documents as they are written."

https://docs.microsoft.com/en-us/azure/cosmos-db/index-policy
upvoted 4 times

**sunil_kalra** 4 years, 5 months ago
yes, we can turn that indexing off and turn it back on after data load is complete
upvoted 2 times

**M0e** 4 years, 8 months ago

The explanation for B. is incorrect. MongoDB API only creates an index for _id field. From the documentation: "The Azure Cosmos DB's API for MongoDB server version 3.6 automatically indexes the _id field only. This field can't be dropped. It automatically enforces the uniqueness of the _id field per shard key. To index additional fields, you apply the MongoDB index-management commands. This default indexing policy differs from the Azure Cosmos DB SQL API, which indexes all fields by default."

So, I think B can not the correct answer.

upvoted 5 times

**djangodev** 4 years, 8 months ago

I think its A & D. The link that is provided in explanation, does not mention about the turning off indexes, however it mentions Creating Unix Indexes. Any suggestion?

upvoted 3 times

**syu31svc** 4 years, 6 months ago

I agree on this one

upvoted 1 times

**Leonido** 5 years, 2 months ago

Just a thought - If I migrate from Amazon and I have many locations there, it will make sense to have multiple write sites and run migration in parallel from several different locations. That will server as a migration accelerator.

upvoted 2 times

Question #24 *Topic 1*

You need to recommend a storage solution for a sales system that will receive thousands of small files per minute. The files will be in JSON, text, and CSV formats. The files will be processed and transformed before they are loaded into a data warehouse in Azure Synapse Analytics. The files must be stored and secured in folders.
Which storage solution should you recommend?

A. Azure Data Lake Storage Gen2

B. Azure Cosmos DB

C. Azure SQL Database

D. Azure Blob storage

**Suggested Answer:** *A*
Azure provides several solutions for working with CSV and JSON files, depending on your needs. The primary landing place for these files is either Azure Storage or Azure Data Lake Store.1
Azure Data Lake Storage is an optimized storage for big data analytics workloads.
Incorrect Answers:
D: Azure Blob Storage containers is a general purpose object store for a wide variety of storage scenarios. Blobs are stored in containers, which are similar to folders.
Reference:
https://docs.microsoft.com/en-us/azure/architecture/data-guide/scenarios/csv-and-json

**Yaswant** Highly Voted 👍 4 years, 10 months ago
Blob Storage -> Object Storage (Binary / Flatfiles)
DataLake -> Various formats (CSV, Json, Avro, Parquet.../ Folders)
upvoted 38 times

**cadio30** Most Recent ⏲ 4 years ago
ADLS is the appropriate solution for this requirement as it was indicated the use of "folder" and is a good holder of big data files.
upvoted 2 times

**Saravjeet** 4 years, 1 month ago
Even I am thinking to opt ADLS2 but the only thing to mention is several files per minute which might impact adls2 which is good for bigdata work loads as compared to blob? Which one is correct any source?
upvoted 1 times

**sjain91** 4 years, 2 months ago
Azure datalake storage gen 2
upvoted 4 times

**Deepu1987** 4 years, 4 months ago
The k/word to lookout for is "folders" which means ADLS Gen 2 which is built on top of Az Blob Strg it's a container - vir dir.. where as ADLS Gen 2 is like accumulation of files & it's like a folder.
upvoted 4 times

**Hardik17** 4 years, 6 months ago
ADLS is good for analytics solutions. This requirement has that, that is why ADLS
upvoted 1 times

**syu31svc** 4 years, 6 months ago
The files must be stored and secured in folders.
A is the answer for sure
upvoted 4 times

**timebeing** 4 years, 9 months ago
You can also query JSON files directly from Azure Blob Storage without importing them into Azure SQL. For a complete example of this approach, see Work with JSON files with Azure SQL. Currently this option isn't available for CSV files.
upvoted 1 times

**Jatinmaya** 4 years, 10 months ago

I think it will be be ADLS as it supports all file format and will handle any flow of small files as long as we are not retaining them for a longer period there should not be any problem.

upvoted 1 times

⊟ **Bob123456** 4 years, 10 months ago

Blob is correct

upvoted 1 times

⊟ **chaoxes** 4 years, 6 months ago

It is not. Correct answer is Azure Data Lake Gen 2. It is build on top of Blob Storage + hierarchical namespace (folders). The questions includes file in folders as requirement.

upvoted 5 times

⊟ **Bob123456** 4 years, 10 months ago

I too agree . because of small files per minute , Which is not ideal for Datalake . Correct answer is BLOB STORAGE.

upvoted 2 times

⊟ **Jzerpa_ccs** 4 years, 8 months ago

Question say folder, not containers. I think correct answer is ADLS

upvoted 5 times

⊟ **M0e** 4 years, 8 months ago

Small files issue was with DLS Gen 1. I think because Gen 2 is using Blob Storage in the background, it is not the case with Gen 2 any more. So, the given solution is correct.

upvoted 2 times

⊟ **arkadipb** 4 years, 7 months ago

ADLS Gen 2 is everything Blob is, plus hierarchical capabilities

upvoted 7 times

⊟ **mirr84** 4 years, 12 months ago

There is many small files, and file types are text types, what means the DLS solution, which is BigData like storage is a bad idea. DLS distributed file storage like big files and types of Parquet (columnar optimized). So the correct answer should be Blob Storage in my opinion

upvoted 3 times

⊟ **peppele** 4 years, 11 months ago

Folders = ADLS

upvoted 33 times

Question #25 *Topic 1*

You are designing an Azure Cosmos DB database that will support vertices and edges.
Which Cosmos DB API should you include in the design?

A. SQL

B. Cassandra

C. Gremlin

D. Table

**Suggested Answer:** *C*
The Azure Cosmos DB Gremlin API can be used to store massive graphs with billions of vertices and edges.
Reference:
https://docs.microsoft.com/en-us/azure/cosmos-db/graph-introduction

**chaoxes** Highly Voted 4 years, 6 months ago
1000000% is C Gremlin API.
When we talk about edges/nodes and relationships, its Gremlin API
upvoted 15 times

**Deepu1987** Most Recent 4 years, 4 months ago
True. It's also used in Socai n/ws , Recommendation engines , Geospatial, IoT
upvoted 2 times

**syu31svc** 4 years, 6 months ago
100% is C
upvoted 4 times

Question #26 *Topic 1*

You are designing a big data storage solution. The solution must meet the following requirements:

✑ Provide unlimited account sizes.

✑ Support a hierarchical file system.

✑ Be optimized for parallel analytics workloads.

Which storage solution should you use?

A. Azure Data Lake Storage Gen2

B. Azure Blob storage

C. Apache HBase in Azure HDInsight

D. Azure Cosmos DB

**Suggested Answer:** *A*

Azure Data Lake Storage is optimized performance for parallel analytics workloads

A key mechanism that allows Azure Data Lake Storage Gen2 to provide file system performance at object storage scale and prices is the addition of a hierarchical namespace. This allows the collection of objects/files within an account to be organized into a hierarchy of directories and nested subdirectories in the same way that the file system on your computer is organized.

Reference:

https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-namespace

**syu31svc** Highly Voted 4 years, 6 months ago

100% is A

upvoted 16 times

**achamizo** Most Recent 3 years, 11 months ago

ADLS Gen 2 is the correct

upvoted 1 times

**IAMKPR** 4 years, 1 month ago

Keyword is "Hierarchical", so it should be ADLS Gen 2

upvoted 2 times

**sjain91** 4 years, 2 months ago

Azure data lake storage gen2

upvoted 2 times

## Question #27 *Topic 1*

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.

After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You plan to store delimited text files in an Azure Data Lake Storage account that will be organized into department folders.

You need to configure data access so that users see only the files in their respective department folder.

Solution: From the storage account, you enable a hierarchical namespace, and you use RBAC.

Does this meet the goal?

A. Yes

B. No

**Suggested Answer:** *B*

Disable the hierarchical namespace. And instead of RBAC use access control lists (ACLs).

Note: Azure Data Lake Storage implements an access control model that derives from HDFS, which in turn derives from the POSIX access control model.

Blob container ACLs does not support the hierarchical namespace, so it must be disabled.

Reference:

https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-known-issues https://docs.microsoft.com/en-us/azure/data-lake-store/data-lake-store-access-control

**Yaswant** Highly Voted 👍 4 years, 10 months ago

RBAC -> Container level.

ACL -> Each file and directory in your account.

*NO*

upvoted 30 times

**Abhilvs** Highly Voted 👍 5 years ago

'No' is correct. When you set ACL, if RBAC is enabled on that container, it takes precedence over ACL. So, RBAC should be disabled when using ACL

upvoted 9 times

**sjain91** Most Recent ⏲ 4 years, 2 months ago

Answer: No

Azure RBAC : Storage accounts, containers. Cross resource Azure role assignments at subscription or resource group level.

upvoted 1 times

**vaseva1** 4 years, 2 months ago

Answer: No

Azure RBAC : Storage accounts, containers. Cross resource Azure role assignments at subscription or resource group level.

ACL : Directory, file

upvoted 1 times

**Pavanm34** 4 years, 5 months ago

Data lake gen2 with hierarchical namespace support ACLS . Currently we can not set it up from storage explorer and portal.

Support for setting access control lists (ACLs) recursively

The ability to apply ACL changes recursively from parent directory to child items is generally available. In the current release of this capability, you can apply ACL changes by using PowerShell, Azure CLI, and the .NET, Java, and Python SDK. Support is not yet available for the Azure portal, or Azure Storage Explorer.

upvoted 1 times

**syu31svc** 4 years, 6 months ago

Answer given is correct

https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-namespace

https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control

upvoted 1 times

**brcdbrcd** 4 years, 6 months ago

Answer: No

The only correct case: ACL & HNS enabled.

Azure RBAC and ACL both require the user (or application) to have an identity in Azure AD. Azure RBAC lets you grant "coarse-grain" access to storage account data, such as read or write access to all of the data in a storage account, while ACLs let you grant "fine-grained" access, such as write access to a specific directory or file.

https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control-model

upvoted 5 times

**rmk4ever** 4 years, 9 months ago

Ans: NO.

General-purpose V2 -->Blob container ACL- Not yet supported

You can set ACLs on the root folder of the container but not the container itself.

Can't use ACL in data lake. (can't use in HNS enabled storage account)

Ref: https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-supported-blob-storage-features

upvoted 1 times

> **rmk4ever** 4 years, 9 months ago
>
> Sorry, please ignore the first one.
>
> For Gen2 - can use ACL with HNS
>
> ref: https://docs.microsoft.com/en-us/azure/storage/blobs/recursive-access-control-lists?tabs=azure-powershell
>
> upvoted 1 times

**Ash666** 4 years, 10 months ago

No. We need ACL.

upvoted 1 times

**freia** 4 years, 11 months ago

HNS should not be disabled. "Access control via ACLs is enabled for a storage account as long as the Hierarchical Namespace (HNS) feature is turned ON." (https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control)

upvoted 4 times

**Abhilvs** 5 years ago

ACLs are granular and only evaluated when RBAC if permissions aren't granted with RBAC.

upvoted 1 times

**azurearch** 5 years, 1 month ago

answer should be yes. In RBAC, minimum level of scope to implement security is at container level. Folder level auth is not possible. It needs ACL for that. No reason to disable HNS (data lake ) for that, we can use POSIX permissions provided by data lake to implement folder level permissions.

upvoted 4 times

> **runningman** 5 years, 1 month ago
>
> Isn't the solution saying that RBAC is wrong? If at folder level Auth, RBAC is not possible, then No is correct. Thoughts?
>
> upvoted 3 times
>
> > **drdean** 5 years ago
> >
> > Yes that makes sense to me
> >
> > upvoted 1 times

## Question #28 *Topic 1*

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You plan to store delimited text files in an Azure Data Lake Storage account that will be organized into department folders.
You need to configure data access so that users see only the files in their respective department folder.
Solution: From the storage account, you disable a hierarchical namespace, and you use RBAC (role-based access control).
Does this meet the goal?

A. Yes

B. No

**Suggested Answer:** *B*
Instead of RBAC use access control lists (ACLs).
Note: Azure Data Lake Storage implements an access control model that derives from HDFS, which in turn derives from the POSIX access control model.
Blob container ACLs does not support the hierarchical namespace, so it must be disabled.
Reference:
https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-known-issues https://docs.microsoft.com/en-us/azure/data-lake-store/data-lake-store-access-control

**Ous01** 4 years, 1 month ago

In ADLS, once the hierarchal name space is enabled. It can't be disabled.

upvoted 3 times

**Marcus1612** 3 years, 9 months ago

The answer is NO but the explanation should be updated. ADLS support ACLs and hierarchy cannot be disabled on Azure Blob Storage once it was enabled

upvoted 1 times

**sjain91** 4 years, 2 months ago

Answer: No
Azure RBAC : Storage accounts, containers. Cross resource Azure role assignments at subscription or resource group level.

upvoted 3 times

**riteshsinha18** 4 years, 3 months ago

correct answer

upvoted 2 times

Question #29 *Topic 1*

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You plan to store delimited text files in an Azure Data Lake Storage account that will be organized into department folders.
You need to configure data access so that users see only the files in their respective department folder.
Solution: From the storage account, you disable a hierarchical namespace, and you use access control lists (ACLs).
Does this meet the goal?

A. Yes

B. No

**Suggested Answer:** *A*
Azure Data Lake Storage implements an access control model that derives from HDFS, which in turn derives from the POSIX access control model.
Blob container ACLs does not support the hierarchical namespace, so it must be disabled.
Reference:
https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-known-issues https://docs.microsoft.com/en-us/azure/data-lake-store/data-lake-store-access-control

**passnow** Highly Voted 4 years, 11 months ago
sometimes u guys commenting confuse people
upvoted 37 times

**kempstonjoystick** Highly Voted 5 years, 3 months ago
The question is unclear in this instance, as it doesn't specify whether the ADLS is v1 or v2. For v1, Hierarchical namespaces must be off, for v2 they need to be on:

"Do I have to enable support for ACLs?
No. Access control via ACLs is enabled for a storage account as long as the Hierarchical Namespace (HNS) feature is turned ON."

https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control
upvoted 35 times

**samok** 5 years, 2 months ago
You are correct. I believe this is an old question, before Gen2 was available. In current exams, they ought to specify which Gen they are referring to.
upvoted 6 times

**M0e** 4 years, 8 months ago
I think Gen 1 is not covered in the exams any more. So having the assumption that the question talks about Gen 2, the answer here is No.
upvoted 7 times

**satyamkishoresingh** Most Recent 3 years, 10 months ago
Solution: From the storage account, you disable a hierarchical namespace, and you use access control lists (ACLs).

if disable hierarchical namespace , then the case has to be NO
upvoted 1 times

**eurekamike** 4 years ago
enable hierarchical namespace, then access control lists
upvoted 1 times

**azurenav** 4 years ago
Enable HNS and ACL -- This is 100% correct
upvoted 1 times

**Ous01** 4 years, 1 month ago

One the storage account is created. We can't enable or disable Namespace. The storage account must be re-created. I don't understand why the answer is Yes. It should be no in my opinion.

upvoted 1 times

**cadio30** 4 years, 1 month ago

From the question standpoint, it is pertaining to ADLS Gen 2 in which is it requires to enable the "hierarchical namespace" to utilize the functionality of Data Lake then we could configure the ACL in the folder level. Therefore, the answer is NO.

upvoted 1 times

**Apox** 4 years, 2 months ago

I believe the answer should be "YES":
The requirement is that data is organized into folders (hence, you have to enable hierarchical namespace") and the users should only see their respective folders. The only way to give users fine-grained access to folders in ADLS Gen2 is to use Access Control Lists. If this is not used you will have to use RBAC and this can only give access to ALL of the data in a storage account or ALL of the data in the container, which will not fulfill the requirement.

It is also unlikely that Shared Access Signatures (SAS) should be used. The reason is that this is internal and you want to have a concept of who actually access what (and they likely have users set up in AAD). SAS is more often used in the context of applications than users, and therefore this is not the right answer either. Hence, hierarchical namespace and ACL should be used and the answer to this question is "YES".

upvoted 1 times

> **BobFar** 4 years, 1 month ago
>
> what about this ?
> Do I have to enable support for ACLs?
> No. Access control via ACLs is enabled for a storage account as long as the Hierarchical Namespace (HNS) feature is turned ON.
>
> If HNS is turned OFF, the Azure Azure RBAC authorization rules still apply.
> in the below link?
> https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control
>
> upvoted 1 times

**sdas1** 4 years, 5 months ago

Refer: https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control
Access control via ACLs is enabled for a storage account as long as the Hierarchical Namespace (HNS) feature is turned ON.

If HNS is turned OFF, the Azure Azure RBAC authorization rules still apply.

upvoted 4 times

> **BobFar** 4 years, 1 month ago
>
> that is exactly what I found
> https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control
>
> upvoted 1 times

**sd_dp200** 4 years, 5 months ago

isn't hierarchical namespace a fundamental property of data lake storage that separates it from blob storage type? why are they saying disable HNS then?

upvoted 1 times

> **cadio30** 4 years ago
>
> in ADLS Gen 1, there is no such feature that could disable the HNS while in Gen 2 this is possible.
>
> upvoted 1 times

**sturcu** 4 years, 5 months ago

The Question n is out dated, it refresh to gen1. In gen2 there is no need to Disable Hierarchical Namespace

upvoted 1 times

**mohowzeh** 4 years, 5 months ago

In storage V2, you can only create ACL's on a container with hierarchical namespace enabled. You cannot disable hierarchical namespace and have an ACL at the same time. Hence, the goal is not met.
Test this yourself in Azure. Create two storage accounts: one with hierarchical namespace disabled (the "blob account"), and one with it enabled (the "data lake account"). Create a container in each. Install Azure Data Explorer on your local machine, then follow the instructions on this page:
https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-explorer#managing-access

You will see that ACL's are an option on the data lake container, but not on the blob container. Hence, disabling the hierarchical namespace makes it impossible to have an ACL on the containers in that account. The configuration as given in the question is therefore not meeting the goal.

upvoted 3 times

**BobFar** 4 years, 1 month ago

Do I have to enable support for ACLs?

No. Access control via ACLs is enabled for a storage account as long as the Hierarchical Namespace (HNS) feature is turned ON.

If HNS is turned OFF, the Azure Azure RBAC authorization rules still apply.

https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control

upvoted 1 times

**BungyTex** 4 years, 6 months ago

It clearly says the data is arranged into folders by department. If you don't have HNS you don't have the folders.

upvoted 1 times

**syu31svc** 4 years, 6 months ago

Answer is No; enable not disable the namespace

upvoted 1 times

**rmk4ever** 4 years, 9 months ago

New update: https://docs.microsoft.com/en-us/azure/storage/blobs/recursive-access-control-lists?tabs=azure-powershell

upvoted 1 times

**yilpiz** 4 years, 10 months ago

Question clearly states Azure Data Lake Storage. Why he is talking about blob?

upvoted 1 times

**Yaswant** 4 years, 10 months ago

Enable heirarchial namespace and use ACL's

This is the one of the option i got in recent exam.

upvoted 14 times

**Porus** 4 years, 10 months ago

whats the answer

upvoted 1 times

**treebeard** 4 years, 8 months ago

This is what I found @ MS Docs:

'Access control via ACLs is enabled for a storage account as long as the Hierarchical Namespace (HNS) feature is turned ON.

If HNS is turned OFF, the Azure RBAC authorization rules still apply.'

Ref: https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control

upvoted 4 times

Question #30 *Topic 1*

You plan to store 100 GB of data used by a line-of-business (LOB) app.
You need to recommend a data storage solution for the data. The solution must meet the following requirements:
✑ Minimize storage costs.
✑ Natively support relational queries.
✑ Provide a recovery time objective (RTO) of less than one minute.
What should you include in the recommendation?

A. Azure Cosmos DB

B. Azure SQL Database

C. Azure Synapse Analytics

D. Azure Blob storage

**Suggested Answer:** *D*
Incorrect Answers:
A: Azure Cosmos DB would require an SQL API.

**Nehuuu** Highly Voted 👍 5 years, 3 months ago
Should it not be SQL Database?
relational queries are supported in SQL DB and SQL DWH, however if cost becomes a factor, it should be SQL DB.
upvoted 90 times

**chaoxes** 4 years, 6 months ago
Yes, answer is B. Azure SQL Database.

Cosmos DB - expensive
SQL DW - expensive
Blob Storage - cheap, but doesn't support SQL relational queries
upvoted 23 times

**EricN** Highly Voted 👍 5 years, 3 months ago
The answer should be B. Azure SQL Database manual database failover can be achieved in 30s. https://docs.microsoft.com/en-us/azure/sql-database/sql-database-business-continuity
upvoted 36 times

**Niteen** 5 years, 1 month ago
Yes, and it can be stored Point in time recovery with fastest. So Ans should be SQL DB
upvoted 2 times

**Bro_its_guru** Most Recent 🕘 3 years ago
Option D. Azure Blob Storage is correct answer 1000 percent sure. Not B.
upvoted 1 times

**Sasha_in_San_Francisco** 3 years, 9 months ago
I also believe it is B - Azure SQL Database. FYI: SkillCertPro has the wrong answer! Very confusing. :-(
upvoted 1 times

**savin** 4 years ago
should be Azure SQL DB
upvoted 2 times

**ismaelrihawi** 4 years, 1 month ago
Relational queries support is only provided by a Azure SQL Database
upvoted 1 times

**cadio30** 4 years, 1 month ago
Definitely B is the answer
upvoted 1 times

**davita8** 4 years, 2 months ago

B. Azure SQL Database

upvoted 2 times

**chirag1234** 4 years, 2 months ago

Blob storage is correct.

-- minimum cost

-- support SQL Relational queries by creating an external table

-- hot tier to recover fast

upvoted 2 times

**chakanirban** 4 years ago

Natively support relational queries. hence Azure SQL DB

upvoted 3 times

**Nik71** 4 years, 3 months ago

The Query Blob Contents API applies a simple Structured Query Language (SQL) statement on a blob's contents and returns only the queried subset of the data.

so BLOB is correct

upvoted 1 times

**syu31svc** 4 years, 6 months ago

https://azure.microsoft.com/en-us/updates/azure-sql-db-published-first-in-industry-business-continuity-sla-for-a-relational-database-service/:

"100% SLA for a 30 second recovery time objective (RTO)"

upvoted 2 times

**sandGrain** 4 years, 7 months ago

Azure SQL DB .... 100%

upvoted 5 times

**BCYT** 4 years, 7 months ago

Mark it for myself, SQL DB

upvoted 4 times

**Bob123456** 4 years, 10 months ago

It should be sql database instead of blob

upvoted 3 times

**Yaswant** 4 years, 10 months ago

Azure SQL Database

DataStorage : Around 1TB (Max).

Relational queries : Native support.

RTO (Manual failover) : 30seconds.

upvoted 10 times

**passnow** 4 years, 11 months ago

B does not require sql api support

upvoted 1 times

**proca** 4 years, 11 months ago

The answer should be B.

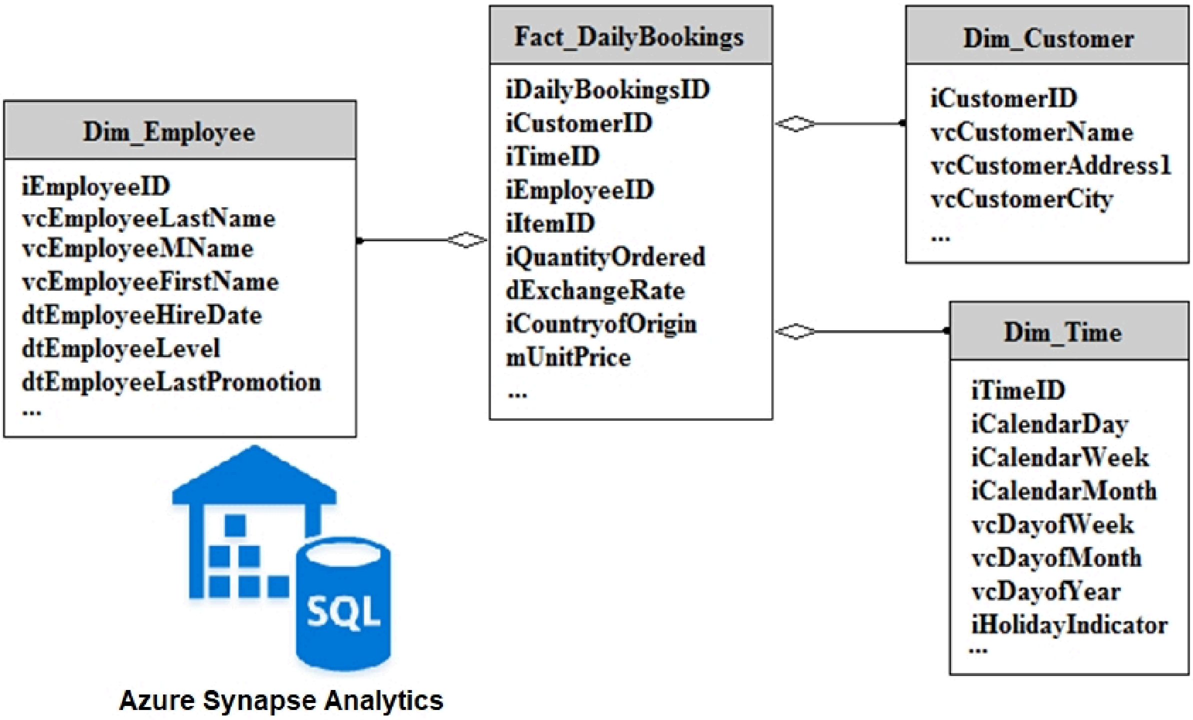
Azure SQL Database Business Critical tier configured with geo-replication has a guarantee of Recovery time objective (RTO) of 30 sec for 100% of deployed h

https://azure.microsoft.com/en-us/support/legal/sla/sql-database/v1_4/#:~:text=of%20deployed%20hours.-,Azure%20SQL%20Database%20Business%20Critical%20tier%20configured%20with%20geo%2Dreplicatior

upvoted 2 times

HOTSPOT -

You have a data model that you plan to implement in a data warehouse in Azure Synapse Analytics as shown in the following exhibit.



Fact_DailyBookings
iDailyBookingsID
iCustomerID
iTimeID
iEmployeeID
iItemID
iQuantityOrdered
dExchangeRate
iCountryofOrigin
mUnitPrice
...

Dim_Customer
iCustomerID
vcCustomerName
vcCustomerAddress1
vcCustomerCity
...

Dim_Employee
iEmployeeID
vcEmployeeLastName
vcEmployeeMName
vcEmployeeFirstName
dtEmployeeHireDate
dtEmployeeLevel
dtEmployeeLastPromotion
...

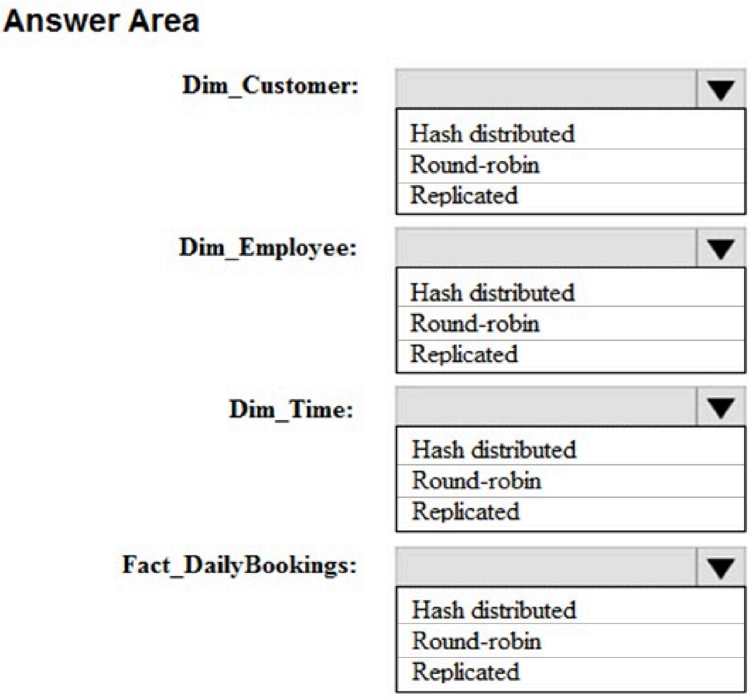
Dim_Time
iTimeID
iCalendarDay
iCalendarWeek
iCalendarMonth
vcDayofWeek
vcDayofMonth
vcDayofYear
iHolidayIndicator
...

All the dimension tables will be less than 2 GB after compression, and the fact table will be approximately 6 TB.

Which type of table should you use for each table? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:



Answer Area

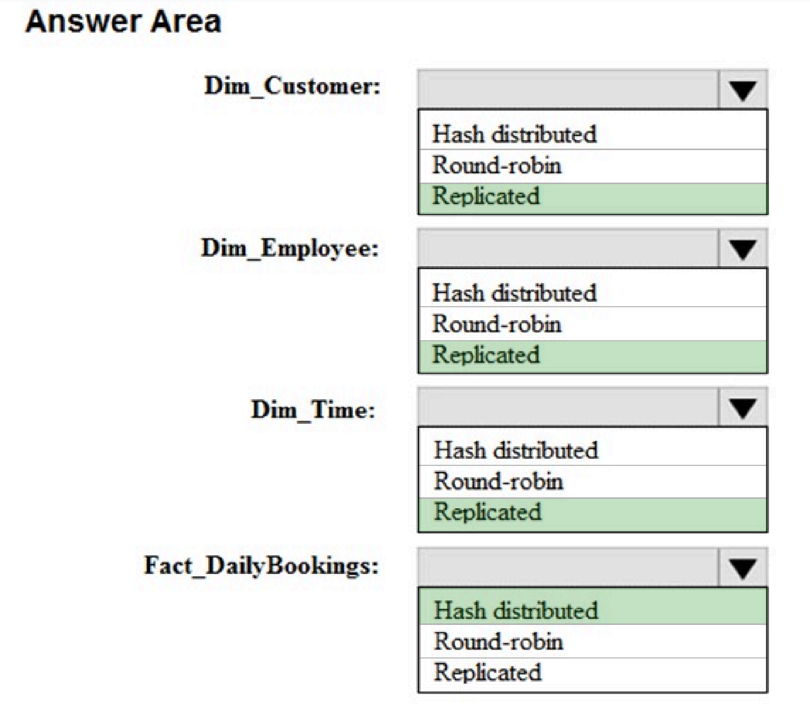
Dim_Customer:
Hash distributed
Round-robin
Replicated

Dim_Employee:
Hash distributed
Round-robin
Replicated

Dim_Time:
Hash distributed
Round-robin
Replicated

Fact_DailyBookings:
Hash distributed
Round-robin
Replicated

**Suggested Answer:**



Answer Area

Dim_Customer:
- Hash distributed
- Round-robin
- [✓ Replicated]

Dim_Employee:
- Hash distributed
- Round-robin
- [✓ Replicated]

Dim_Time:
- Hash distributed
- Round-robin
- [✓ Replicated]

Fact_DailyBookings:
- [✓ Hash distributed]
- Round-robin
- Replicated

Box 1: Replicated -
Replicated tables are ideal for small star-schema dimension tables, because the fact table is often distributed on a column that is not compatible with the connected dimension tables. If this case applies to your schema, consider changing small dimension tables currently implemented as round-robin to replicated.

Box 2: Replicated -

Box 3: Replicated -

Box 4: Hash-distributed -
For Fact tables use hash-distribution with clustered columnstore index. Performance improves when two hash tables are joined on the same distribution column.
Reference:
https://azure.microsoft.com/en-us/updates/reduce-data-movement-and-make-your-queries-more-efficient-with-the-general-availability-of-replicated-tables/ https://azure.microsoft.com/en-us/blog/replicated-tables-now-generally-available-in-azure-sql-data-warehouse/

**anamaster** Highly Voted 4 years, 2 months ago
correct
upvoted 13 times

**IAMKPR** Highly Voted 4 years, 1 month ago
Small Dimension Tables --> Replicated
Large Fact Tables --> Hash Distributed
upvoted 12 times

**erssiws** Most Recent 4 years ago
all the dimension tables are replicated since they are below 2G.
The fact table should be round-robin in my view
upvoted 1 times

**rk_datageek** 4 years, 1 month ago
In case if the dimension tables are larger than 2GB, should it be "Round-Robin" or still "Replicated" is a good option
upvoted 1 times

Question #32 *Topic 1*

You are designing a data storage solution for a database that is expected to grow to 50 TB. The usage pattern is singleton inserts, singleton updates, and reporting.
Which storage solution should you use?

A. Azure SQL Database elastic pools

B. Azure Synapse Analytics

C. Azure Cosmos DB that uses the Gremlin API

D. Azure SQL Database Hyperscale

**Suggested Answer:** *D*
A Hyperscale database is an Azure SQL database in the Hyperscale service tier that is backed by the Hyperscale scale-out storage technology. A Hyperscale database supports up to 100 TB of data and provides high throughput and performance, as well as rapid scaling to adapt to the workload requirements. Scaling is transparent to the application ג€" connectivity, query processing, etc. work like any other Azure SQL database.
Incorrect Answers:
A: SQL Database elastic pools are a simple, cost-effective solution for managing and scaling multiple databases that have varying and unpredictable usage demands. The databases in an elastic pool are on a single Azure SQL Database server and share a set number of resources at a set price. Elastic pools in Azure
SQL Database enable SaaS developers to optimize the price performance for a group of databases within a prescribed budget while delivering performance elasticity for each database.
B: Rather than SQL Data Warehouse, consider other options for operational (OLTP) workloads that have large numbers of singleton selects.
Reference:
https://docs.microsoft.com/en-us/azure/sql-database/sql-database-service-tier-hyperscale-faq

**riteshsinha18** Highly Voted 4 years, 3 months ago
correct answer
upvoted 9 times

**kimalto452** Most Recent 3 years, 8 months ago
synapse support singleton operations...
upvoted 1 times

**achamizo** 3 years, 11 months ago
Singlenton operations = transactional operations, so only relational databases supports it. Hyperscale supports up to 100TB.
upvoted 1 times

**toandm** 4 years, 1 month ago
can anyone tell what 'singleton inserts, singleton updates' mean? I've been looking on the internet but no solid definition
upvoted 1 times

> **[Removed]** 4 years, 1 month ago
> Go through the ACID concept
> upvoted 1 times

> **Steve92873197** 4 years, 1 month ago
> Would take that to mean single row inserts, or updates
> upvoted 1 times

**cadio30** 4 years, 1 month ago
azure synapse is better for OLAP purposes while for OLTP, use of Azure SQL is the best choice. as the requirement states it requires to perform insert and update.
upvoted 1 times

**aksoumi** 4 years, 2 months ago
But isn't Datawarehouse (Synapse Analytics) a right choice considering that we have table geometries (hash distribution) to efficiently perform inserts updates and deletes?
upvoted 1 times

⊟ **savin** 4 years ago

No DWH is not a place for singleton updates

upvoted 1 times

⊟ **DongDuong** 4 years, 2 months ago

Keyword here is 100TB based on the support link.

upvoted 10 times

Question #33 *Topic 1*

HOTSPOT -

You are designing a solution that will use Azure Table storage. The solution will log records in the following entity.

```
DepartmentName+EmployeeID (string)
Year+Month+Day+Hour+EventID (string)
FirstName (string)
LastName (string)
EventType (string)
EventTimestamp (datetime)
EventText (string)
```

You are evaluating which partition key to use based on the following two scenarios:

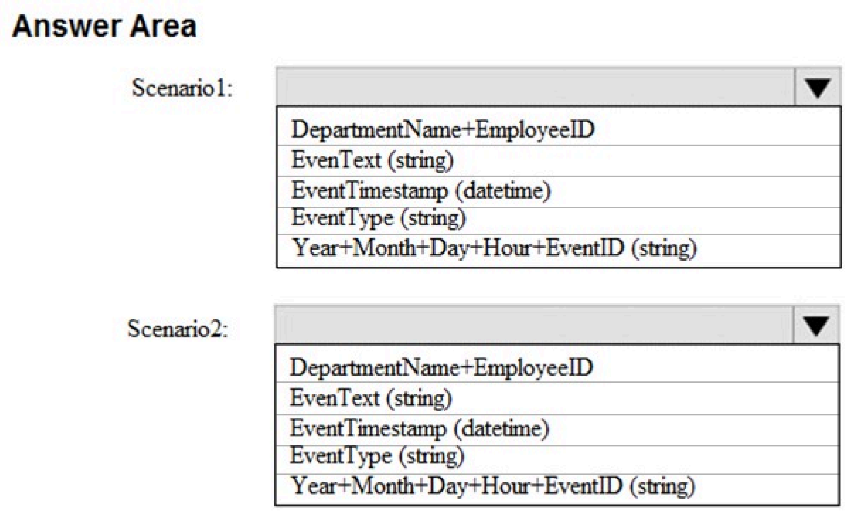
✑ Scenario1: Minimize hotspots under heavy write workloads.

✑ Scenario2: Ensure that date lookups are as efficient as possible for read workloads.

Which partition key should you use for each scenario? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:



**Answer Area**

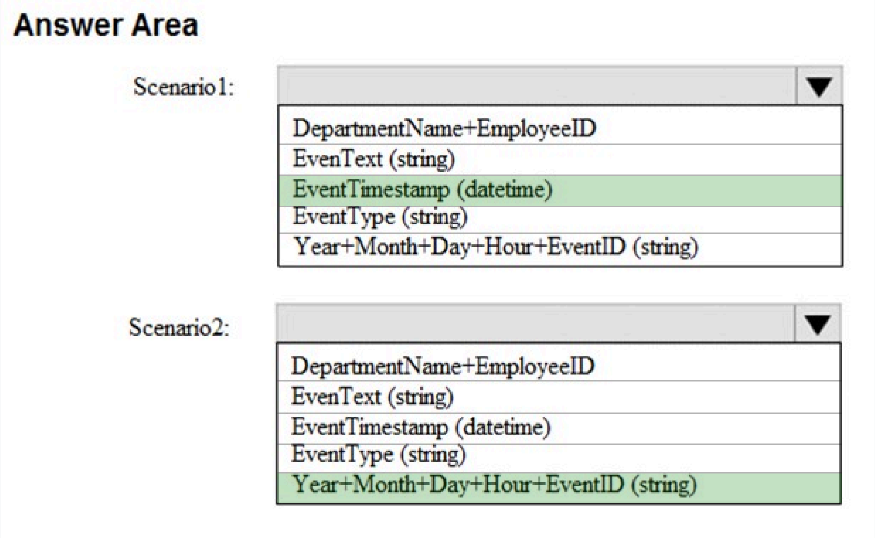
Scenario1: ▼

- DepartmentName+EmployeeID
- EventText (string)
- EventTimestamp (datetime)
- EventType (string)
- Year+Month+Day+Hour+EventID (string)

Scenario2: ▼

- DepartmentName+EmployeeID
- EventText (string)
- EventTimestamp (datetime)
- EventType (string)
- Year+Month+Day+Hour+EventID (string)

**Suggested Answer:**



**Answer Area**

Scenario1: ▼

- DepartmentName+EmployeeID
- EventText (string)
- [✓ EventTimestamp (datetime)]
- EventType (string)
- Year+Month+Day+Hour+EventID (string)

Scenario2: ▼

- DepartmentName+EmployeeID
- EventText (string)
- EventTimestamp (datetime)
- EventType (string)
- [✓ Year+Month+Day+Hour+EventID (string)]

References:

https://docs.microsoft.com/en-us/rest/api/storageservices/designing-a-scalable-partitioning-strategy-for-azure-table-storage

**Yuri1101** Highly Voted 5 years, 2 months ago

Scenerio1: DepartmentName+EmployeeID

upvoted 78 times

**JamesCho** 5 years, 1 month ago

Even if 1-2 departments have more employees than other departments, practically not all employees will not sign-up for events all at a time.

upvoted 1 times

**vistran** 5 years, 1 month ago

thats the answer as per

https://docs.microsoft.com/en-us/azure/cosmos-db/table-storage-design-guide#solution-11

upvoted 20 times

**Treadmill** 4 years, 10 months ago

Scenario 1: Department+EmployeeID = avoids hotspots on inserts which happen at the same time
Scenario 2: Year+month+day+hour+EventID = date is included as a string for date lookups

Wrong
Datetime: The partition key value (For example: "Andrew"). The partition key value can be of string or numeric types.

https://docs.microsoft.com/en-us/azure/cosmos-db/partitioning-overview

https://docs.microsoft.com/en-us/azure/cosmos-db/table-storage-design-guide#solution-11

upvoted 23 times

**groy** Highly Voted 4 years, 9 months ago

Correct final answers.....!!

1: Department+EmployeeID = avoids hotspots on inserts which happen at the same time
2: Year+month+day+hour+EventID = date is included as a string for date lookups

upvoted 43 times

**Pairon** Most Recent 4 years, 3 months ago

I agree with the given answer.
"You could also partition your data by a Date or DateTime attribute (or some part of)." (https://trycatch.me/data-partitioning-strategy-in-cosmosdb/), so datetime field can be used as partition key.

upvoted 2 times

**Pairon** 4 years, 3 months ago

Not sure about DepartmentName+EmployeeID: the department name could change.

upvoted 2 times

**AhmedReda** 5 years ago

For Scenario 2 : is it Year-Month-Day-Hour-EventID ??

upvoted 2 times

**Abhilvs** 5 years ago

The timestamp is of milliseconds precision, in that case, it won't lead to hot partitions.

upvoted 2 times

**Ash666** 4 years, 10 months ago

Can't set date time type as partition key.
Either string or numeric type.

upvoted 4 times

**Mathster** 5 years, 1 month ago

Surname cannot be a partition key because ~750 records have a null value.

upvoted 1 times

**azurearch** 5 years, 1 month ago

choosing timestamp would create multiple partitions and affects insert operations. composite key could be the right choice here

upvoted 1 times

**Tombarc** 5 years, 2 months ago

DepartmentName+EmployeeID could still result in hot partitions as there might be departments with many more employees than others. I'd say both scenarios would have a combination of "Year+month+day+hour_EventID" to suffice the requirements, and then a "rowkey" would be used to distinguish between the two.

https://docs.microsoft.com/en-us/rest/api/storageservices/designing-a-scalable-partitioning-strategy-for-azure-table-storage#r

upvoted 7 times

**azurearch** 5 years, 1 month ago

the scenario is for write heavy workload, on a certain hour there could be many events causing hot spots.

upvoted 1 times

⊟ **azurearch** 5 years, 1 month ago

for each event it would be return in a separate partition since we are adding event id to the partition key. that would introduce write latency. dept + empid makes logical. can be the answer.

upvoted 2 times

⊟ **zb99** 5 years, 2 months ago

Timestamp would actually be the worst possible partition option for hotspots. Will result in automatic range partitioning, causing all writes to go to a single partition: https://docs.microsoft.com/en-us/rest/api/storageservices/designing-a-scalable-partitioning-strategy-for-azure-table-storage

upvoted 6 times

⊟ **apz333** 5 years, 2 months ago

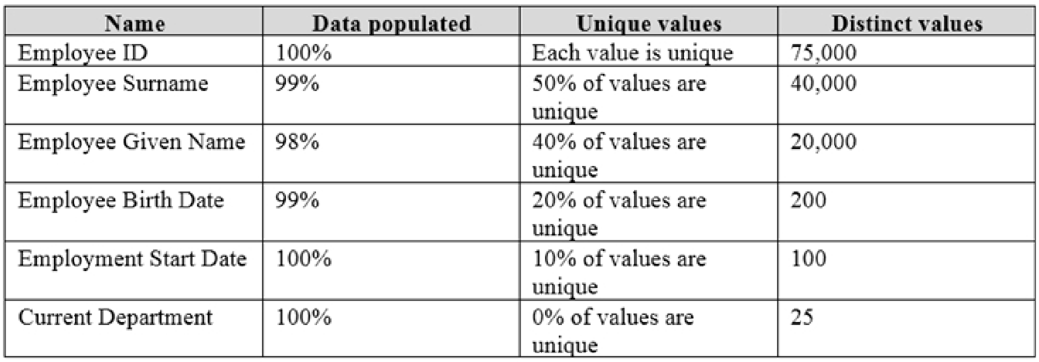
Besides, you have to use string values for PartitionKey, and "EventTimestamp" is a datetime type. I don't think you could use it at all unless you convert it to string.

upvoted 2 times

Question #34 *Topic 1*

DRAG DROP -

You have data on the 75,000 employees of your company. The data contains the properties shown in the following table.



| Name | Data populated | Unique values | Distinct values |
|---|---|---|---|
| Employee ID | 100% | Each value is unique | 75,000 |
| Employee Surname | 99% | 50% of values are unique | 40,000 |
| Employee Given Name | 98% | 40% of values are unique | 20,000 |
| Employee Birth Date | 99% | 20% of values are unique | 200 |
| Employment Start Date | 100% | 10% of values are unique | 100 |
| Current Department | 100% | 0% of values are unique | 25 |

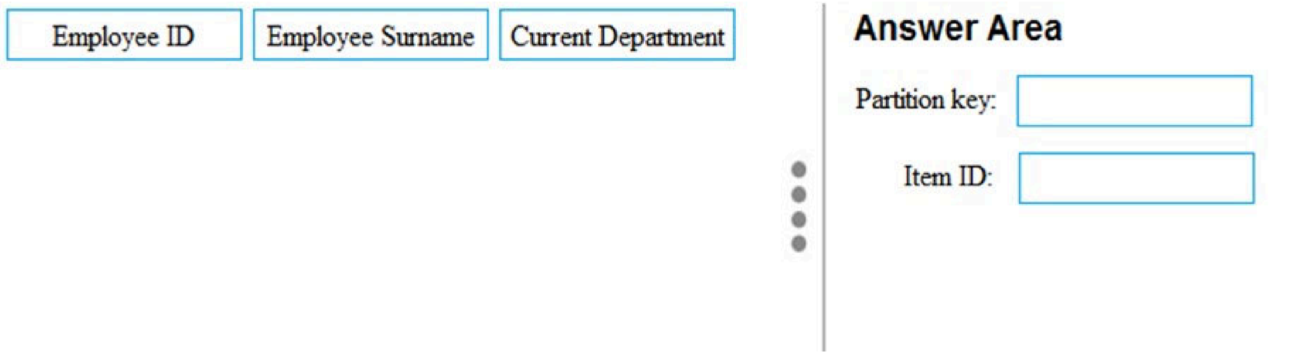
You need to store the employee data in an Azure Cosmos DB container. Most queries on the data will filter by the Current Department and the Employee

Surname properties.

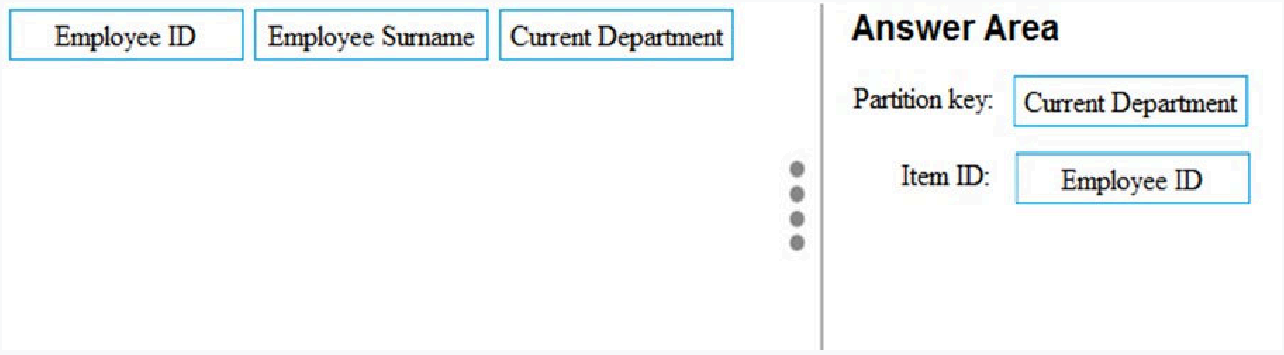
Which partition key and item ID should you use for the container? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Select and Place:



**Suggested Answer:**



Partition key: Current Department

Item ID: Employee ID -

Reference:

https://docs.microsoft.com/en-us/rest/api/storageservices/designing-a-scalable-partitioning-strategy-for-azure-table-storage

**Luke97** Highly Voted 5 years, 2 months ago

I think the partition key should be Department rather than Surname. The reason for this is the read latency. As the question stated, "most of query would filter by Current Department or Employer Surname". Having Surname as partition key, you would have 40,000 partition (40,000 unique value), and when you filter your query by Department, you query will need to go through 40,000 partition which would be real bad on performance. Other another hand, having Department as partition key, you would have 25 partition, and to filter surname query, it would be much faster compare to query on 40,000 partition.

upvoted 80 times

**Gashurb** 5 years, 2 months ago

Yepp, i agree too. 40k partitions can't be good.
upvoted 1 times

**Isio05** 4 years, 6 months ago

Current department suggests it's something can change, therefore it can't be parition key. Reasoning that we shouldn't use Surname because it will result in many logical partitions is completely wrong. It's even clearly stated in docs that even item id (with only unique) values is a valid option (however here Surname is more appropriate as we use it a predicate in queries).
upvoted 2 times

**francisco94** 4 years, 6 months ago

Wrong, Department can change thus it cant be a partitioning key, moreover your argument is that it would be better if partitioning key would be Department and filtering on Surname because you would need to access only 25 partitions. Yes 25 partition of thousand of values! It is bad either way... but the better one is surname.
upvoted 4 times

**Manue** 5 years, 1 month ago

From https://docs.microsoft.com/en-gb/azure/cosmos-db/partitioning-overview:
"For all containers, your partition key should:
Be a property that has a value which does not change. If a property is your partition key, you can't update that property's value.
Have a high cardinality. In other words, the property should have a wide range of possible values.
Spread request unit (RU) consumption and data storage evenly across all logical partitions. This ensures even RU consumption and storage distribution across your physical partitions."
Firstly, "Current Department" is something that could change. Secondly, "25" is not high cardinality, and does not guarantee even distribution of data. E.g. if that was a huge IT company, 50k could be in the Engineering department, 50 in HHRR, 50 in MKT, etc.
So I think it should be "Surname" and EmployeeID.
upvoted 57 times

**LiamRT** 3 years, 7 months ago

The 'Sales' department will not change it's name. An employee may transfer from 'Sales' to 'Engineering' but that causes no issue to the partitioning.
upvoted 1 times

**aksoumi** 4 years, 2 months ago

agree 100%
upvoted 1 times

**Dhaval_Azure** 4 years ago

No. It can't be as "Surname" as its values are populated only 99%. Will empty value in Partition key works? I think we need a column that is 100% populated. it can be EmpoyeID or the current department.
So, I am thinking to go with Key: Employer ID & Item Id: current department
upvoted 1 times

**Ard** Highly Voted 5 years, 2 months ago

i think the answer should be partition by surname ( as it has more unique values than department) and employeeId as itemid since it's unique.
upvoted 31 times

**manasa203** Most Recent 2 years ago

I think the answer is correct. if you see the data populated column, for the surname it's 99%. A partition key column should not have null values. for department it's 100%, hence department is the best choice here
upvoted 1 times

**satyamkishoresingh** 3 years, 8 months ago

Isn't employee ID a good candidate for partitioning ?
upvoted 1 times

**Marcus1612** 3 years, 9 months ago

The anwser is wrong ! look at this because the Department could change.
https://docs.microsoft.com/en-gb/azure/cosmos-db/partitioning-overview#choose-partitionkey
The answer would be good for a large containers. But in this use case we have a small one. Partition strategy depends on the container size. Since we do not have an Read-Heavy container. We shoud use a property that does not change. Partion key = EmployeeID . (both Department and Surnames can change).
" For large read-heavy containers, however, you might want to choose a partition key that appears frequently as a filter in your queries. Queries can be efficiently routed to only the relevant physical partitions by including the partition key in the filter predicate.

If most of your workload's requests are queries and most of your queries have an equality filter on the same property, this property can be a good partition key choice."

upvoted 1 times

⊟ **hello_there_** 3 years, 10 months ago

The partition key should be employee_id. From the documentation (https://docs.microsoft.com/en-us/azure/cosmos-db/partitioning-overview):

For all containers, your partition key should:

Be a property that has a value which does not change. If a property is your partition key, you can't update that property's value.

Have a high cardinality. In other words, the property should have a wide range of possible values.

Spread request unit (RU) consumption and data storage evenly across all logical partitions. This ensures even RU consumption and storage distribution across your physical partitions.
currentDepartment has a low cardinality and can change. Lastname can change (people get married) and 1% has null lastname, which creates one large partition and thus uneven distribution. The same documentation states: "For small read-heavy containers or write-heavy containers of any size, the item ID is naturally a great choice for the partition key.". This container certainly qualifies as small, it's just 75.000 employee records.

upvoted 1 times

⊟ **Durga123** 4 years, 1 month ago

lot of confusion here. what is the correct answer?

upvoted 3 times

⊟ **alain2** 4 years, 1 month ago

pk: Employee Surname
id: Employee Id

upvoted 6 times

⊟ **Deepu1987** 4 years, 4 months ago

I agree with the given solution partion key - current dept
item id - emp id

upvoted 5 times

⊟ **syu31svc** 4 years, 6 months ago

The answer given is correct.
Put aside all the theory and concepts about partitioning and just think about it:
A company has different departments and each department has its own employees. Between name/surname and ID, ID is definitely the better identifier.

upvoted 2 times

⊟ **ttAsh** 4 years, 6 months ago

partition key should be current department(populated 100%) as surname is only 99% populated. we cannot have a partition key as NULL/ not populated.

upvoted 9 times

⊟ **BitchNigga** 4 years ago

Finally someone said it

upvoted 3 times

⊟ **captainbee** 3 years, 11 months ago

But also department is a field that can chagne quite easily, which is something that partitions cannot do. So ultimately this question sucks, but if at gunpoint I had to take one of them, I'd go with Surname.

upvoted 2 times

⊟ **essdeecee** 4 years, 8 months ago

I suspect it's surname rather than department. Firstly there are simply too few variants, its also "current department" so might is likely to change. Surname is similarly bad on the changeable nature (assume a woman getting married e.g.) but assuming all else it's better than department.

upvoted 1 times

⊟ **tdaou** 4 years, 9 months ago

I agree with the suggested answer, I would also argue that the distribution of names will be highly uneven and would result in partitions of very different sizes, including 40,000 with one unique entry. So Current Department by elimination really.

upvoted 3 times

**zglat** 4 years, 10 months ago

'Current' department suggests that it changes. Surnames change all the time. As a result neither of them are good choice for partition keys. I believe that leaves Employee ID

upvoted 2 times

**Ash666** 4 years, 10 months ago

From the docs:

For all containers, your partition key should:
Be a property that has a value which does not change. If a property is your partition key, you can't update that property's value.

So current dept can't be partition key. So obviously it's surname.

Employee ID should be item ID

upvoted 8 times

**Ash666** 4 years, 10 months ago

https://docs.microsoft.com/en-us/azure/cosmos-db/partitioning-overview

upvoted 1 times

**envy** 4 years, 11 months ago

I think the answer is correct as
https://docs.microsoft.com/en-us/azure/cosmos-db/partitioning-overview#choose-partitionkey
If your container could grow to more than a few physical partitions, then you should make sure you pick a partition key that minimizes cross-partition queries. Your container will require more than a few physical partitions when either of the following are true:
• Your container will have over 30,000 RU's provisioned
Your container will store over 100 GB of data

surname has 40,000 values, which " more than a few physical partitions", we should pick a partition key that minimizes cross-partition queries and used in filter. which is "Current Department"

upvoted 4 times

**Needium** 4 years, 3 months ago

Your Partition key should be a value that does not change cos you would not be able to change it. More so nothing in this question suggests the size of the database to be so large or would have over 30000 RUs provisioned. Yes, the nulls in the surname and the fact that surname could even change is a concern, but Surname is very unlikely to change compared to Current Department. Current tells us it is even very volatile.

I would rather have the Surname as Partitioning key.
Thanks for raising this point though, it is worth considering too

upvoted 2 times

**Mathster** 5 years, 1 month ago

Surname cannot be a partition key because ~750 records have a null value.

upvoted 5 times

**drdean** 5 years ago

That's not the worst thing in the world https://sqlstudies.com/2017/05/03/partitioning-on-a-nullable-column/

upvoted 1 times

**HeB** 5 years, 1 month ago

I think the answer for Partition Key should be Employee Surname. It has a wider range and more unique values, see:
https://docs.microsoft.com/nl-nl/azure/cosmos-db/partitioning-overview#choose-partitionkey

upvoted 9 times

Question #35 *Topic 1*

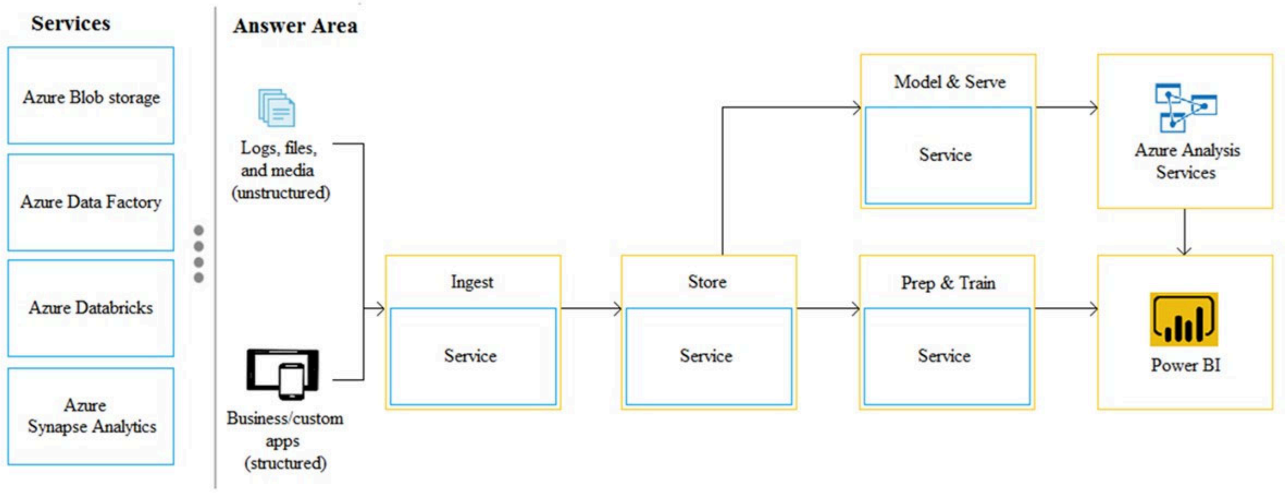
DRAG DROP -

You need to design a data architecture to bring together all your data at any scale and provide insights into all your users through the use of analytical dashboards, operational reports, and advanced analytics.

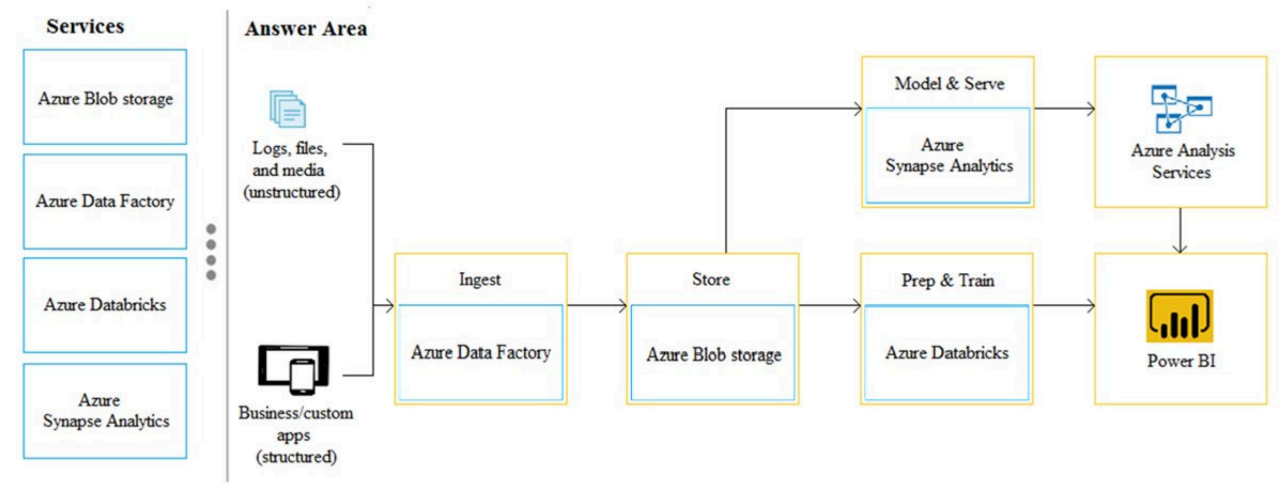
How should you complete the architecture? To answer, drag the appropriate Azure services to the correct locations in the architecture. Each service may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point.

Select and Place:



**Suggested Answer:**



Ingest: Azure Data Factory -

Store: Azure Blob storage -

Model & Serve: Azure Synapse Analytics

Load data into Azure Synapse Analytics.

Prep & Train: Azure Databricks.

Extract data from Azure Blob storage.

Reference:

https://docs.microsoft.com/en-us/azure/azure-databricks/databricks-extract-load-sql-data-warehouse

**gabry75** Highly Voted 4 years, 3 months ago

Correct.

https://docs.microsoft.com/en-us/azure/architecture/solution-ideas/articles/advanced-analytics-on-big-data

upvoted 14 times

**cadio30** 4 years, 1 month ago

The solution would be different if the link provided is the basis as Azure Synapse nowadays could perform the ADF (ingest) , Databricks (Prep & Train) and Model & Serve. Its like an all-in-one package.

upvoted 2 times

## Question #36 *Topic 1*

HOTSPOT -

You are designing an enterprise data warehouse in Azure Synapse Analytics that will store website traffic analytic in a star schema.

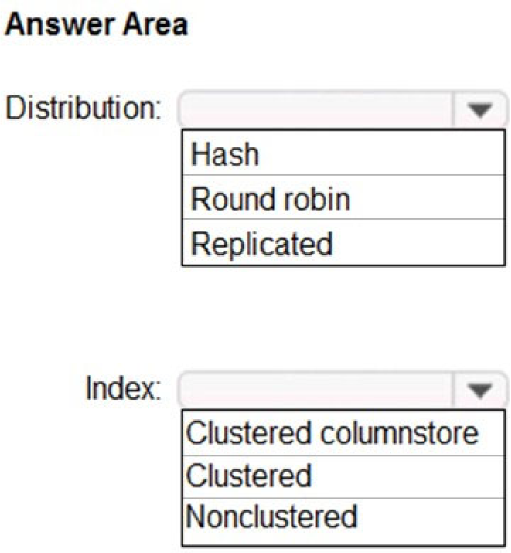
You plan to have a fact table for website visits. The table will be approximately 5 GB.

You need to recommend which distribution type and index type to use for the table. The solution must provide the fastest query performance.

What should you recommend? To answer, select the appropriate options in the answer area

NOTE: Each correct selection is worth one point.

Hot Area:

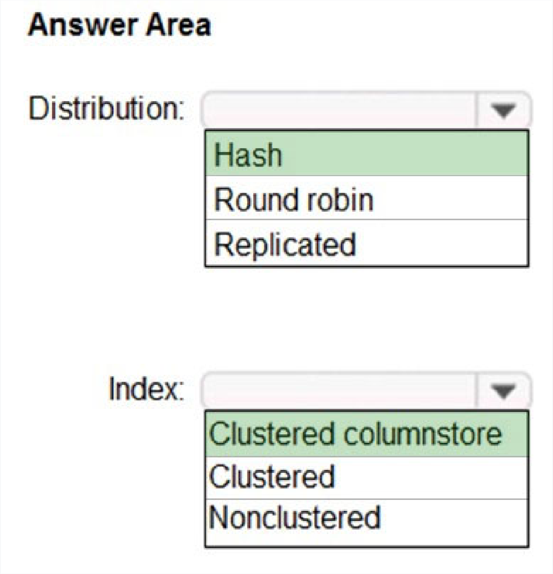


**Answer Area**

Distribution:
- Hash
- Round robin
- Replicated

Index:
- Clustered columnstore
- Clustered
- Nonclustered

**Suggested Answer:**



**Answer Area**

Distribution:
- [✓ Hash]
- Round robin
- Replicated

Index:
- [✓ Clustered columnstore]
- Clustered
- Nonclustered

Reference:
https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-tables-distribute https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-tables-index

**dbdev** Highly Voted 4 years, 1 month ago

The answer is straightforward and correct. Even, there is no need to put a comment here.

upvoted 8 times

**aksoumi** Most Recent 4 years, 2 months ago

How should we decide if it is clustered or a non clustered index? Usually if there are over a million rows (or more than 60 million rows) we use clustered index, but these aren't mentioned in question.

upvoted 1 times

**Apox** 4 years, 2 months ago

Clustered indexes and non-clustered indexes only outperform clustered columnstore indexes when a single row needs to be quickly retrieved with extreme speed. So, for highly selective filters this is the right choice. Since this is a star schema where filters may vary, clustered columnstore indexes are the better choice as this generally provides the best overall query performance (and is best for large tables).

upvoted 8 times

⊟ **BigMF** 4 years ago

The question also mentions analytics implying that it is not intended for “single row” queries. So, I agree with this reasoning.

upvoted 1 times

⊟ **Pairon** 4 years, 3 months ago

Shouldn't be "Round Robin" in the first box? Isn't it more efficient compared to "Hash" since avoids computing the partitions?

upvoted 1 times

⊟ **anamaster** 4 years, 2 months ago

no, since "The solution must provide the fastest query performance."

upvoted 5 times

⊟ **bdloko** 4 years, 2 months ago

Hash for performance requirement

upvoted 5 times

⊟ **eurekamike** 4 years ago

loading time: round robin

query time: hash

upvoted 1 times

⊟ **cadio30** 4 years, 1 month ago

round-robin is use in staging tables

upvoted 1 times

## Question #37 *Topic 1*

You plan to deploy a reporting database to Azure. The database will contain 30 GB of data. The amount of data will increase by 300 MB each year. Rarely will the database be accessed during the second and third weeks of each month. During the first and fourth week of each month, new data will be loaded each night.

You need to recommend a solution for the planned database. The solution must meet the following requirements:

✑ Minimize costs.

✑ Minimize administrative effort.

What should you recommend?

A. an Azure HDInsight cluster

B. Azure SQL Database Hyperscale

C. Azure SQL Database Business Critical

D. Azure SQL Database serverless

**Suggested Answer:** *D*

Serverless is a compute tier for single Azure SQL Databases that automatically scales compute based on workload demand and bills for the amount of compute used per second. The serverless compute tier also automatically pauses databases during inactive periods when only storage is billed and automatically resumes databases when activity returns.

Incorrect Answers:

A: Azure HDInsight is a managed Apache Hadoop service that lets you run Apache Spark, Apache Hive, Apache Kafka, Apache HBase, and more in the cloud.

B, C: Azure SQL Database Hyperscale and Azure SQL Database Business Critical are based on SQL Server database engine architecture that is adjusted for the cloud environment in order to ensure 99.99% availability even in the cases of infrastructure failures.

Reference:

https://docs.microsoft.com/en-us/azure/azure-sql/database/serverless-tier-overview https://docs.microsoft.com/en-us/azure/hdinsight/
https://docs.microsoft.com/en-us/azure/azure-sql/database/service-tier-hyperscale

**syu31svc** Highly Voted 4 years, 6 months ago

A, B and C are wrong for sure

upvoted 5 times

**Kampai787** 4 years, 6 months ago

SI, es SI

upvoted 1 times

**IAMKPR** Most Recent 4 years, 1 month ago

Keyword "Minimize administrative effort" ---> Serverless

upvoted 3 times

## Question #38 *Topic 1*

You are designing a solution for the ad hoc analysis of data in Azure Databricks notebooks. The data will be stored in Azure Blob storage.
You need to ensure that Blob storage will support the recovery of the data if the data is overwritten accidentally.
What should you recommend?

A. Enable soft delete.

B. Add a resource lock.

C. Enable diagnostics logging.

D. Use read-access geo-redundant storage (RA-GRS).

**Suggested Answer:** *A*
Soft delete protects blob data from being accidentally or erroneously modified or deleted. When soft delete is enabled for a storage account, blobs, blob versions
(preview), and snapshots in that storage account may be recovered after they are deleted, within a retention period that you specify.
Reference:
https://docs.microsoft.com/en-us/azure/storage/blobs/soft-delete-overview

**jsonify** Highly Voted 👍 4 years, 2 months ago

A is the correct answer because when you enable blob soft delete for a storage account, you specify a retention period for deleted objects of between 1 and 365 days. The retention period indicates how long the data remains available after it is deleted or overwritten.

upvoted 9 times

**Sudhansu21** Most Recent 🕒 4 years, 1 month ago

A. soft delete is the correct answer. It will allow you to recover quickly.

upvoted 1 times

**sjain91** 4 years, 2 months ago

Enable soft delete

upvoted 2 times

Question #39 *Topic 1*

You are planning a solution that combines log data from multiple systems. The log data will be downloaded from an API and stored in a data store.

You plan to keep a copy of the raw data as well as some transformed versions of the data. You expect that there will be at least 2 TB of log files.

The data will be used by data scientists and applications.

You need to recommend a solution to store the data in Azure. The solution must minimize costs.

What storage solution should you recommend?

A. Azure Data Lake Storage Gen2

B. Azure Synapse Analytics

C. Azure SQL Database

D. Azure Cosmos DB

**Suggested Answer:** *A*

To land the data in Azure storage, you can move it to Azure Blob storage or Azure Data Lake Store Gen2. In either location, the data should be stored in text files.

PolyBase and the COPY statement can load from either location.

Incorrect Answers:

B: Azure Synapse Analytics, uses distributed query processing architecture that takes advantage of the scalability and flexibility of compute and storage resources. Use Azure Synapse Analytics transform and move the data.

Reference:

https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/design-elt-data-loading

**ArunMonika** 3 years, 6 months ago

Given answer is correct

upvoted 1 times

**Ysandee** 4 years ago

Agree with the answer provided.

upvoted 3 times

Question #40 *Topic 1*

You are designing a serving layer for data. The design must meet the following requirements:

✑ Authenticate users by using Azure Active Directory (Azure AD).

✑ Serve as a hot path for data.

✑ Support query scale out.

✑ Support SQL queries.

What should you include in the design?

A. Azure Data Lake Storage

B. Azure Cosmos DB

C. Azure Blob storage

D. Azure Synapse Analytics

**Suggested Answer:** *B*

Do you need serving storage that can serve as a hot path for your data? If yes, narrow your options to those that are optimized for a speed serving layer. This would be Cosmos DB among the options given in this question.

Note: Analytical data stores that support querying of both hot-path and cold-path data are collectively referred to as the serving layer, or data serving storage.

There are several options for data serving storage in Azure, depending on your needs:

✑ Azure Synapse Analytics

✑ Azure Cosmos DB

✑ Azure Data Explorer

Azure SQL Database -

▪

✑ SQL Server in Azure VM

✑ HBase/Phoenix on HDInsight

✑ Hive LLAP on HDInsight

✑ Azure Analysis Services

Incorrect Answers:

A, C: Azure Data Lake Storage & Azure Blob storage are not data serving storage in Azure.

Reference:

https://docs.microsoft.com/en-us/azure/architecture/data-guide/technology-choices/analytical-data-stores

**suman13** Highly Voted 4 years, 2 months ago

scaleout+hotspot = cosmosdb

hotspot = azuresqldb

scaleout = azure synapse

upvoted 19 times

**tamil1006** Most Recent 4 years, 1 month ago

Cosmos db is correct one because ( speed servicing not possible in azure Synapse)https://docs.microsoft.com/en-us/azure/architecture/data-guide/technology-choices/analytical-data-stores

upvoted 1 times

**tamil1006** 4 years, 1 month ago

Do you need serving storage that can serve as a hot path for your data? If yes, narrow your options to those that are optimized for a speed serving layer --> in this way speed serving layer supports by cosmos db and not by synapse

upvoted 1 times

**Geo_Barros** 4 years, 3 months ago

I think that Azure Synapse Analytics could also be an answer for this question.

upvoted 2 times

**szpinat** 4 years, 3 months ago

Synapse is note optimized for speed serving layer.

upvoted 11 times

**cadio30** 4 years, 1 month ago

Agree on this

upvoted 1 times

Question #41 *Topic 1*

You are designing a storage solution for streaming data that is processed by Azure Databricks. The solution must meet the following requirements:
✑ The data schema must be fluid.
✑ The source data must have a high throughput.
✑ The data must be available in multiple Azure regions as quickly as possible.
What should you include in the solution to meet the requirements?

A. Azure Cosmos DB

B. Azure Synapse Analytics

C. Azure SQL Database

D. Azure Data Lake Storage

**Suggested Answer:** *A*
Azure Cosmos DB is Microsoft's globally distributed, multi-model database. Azure Cosmos DB enables you to elastically and independently scale throughput and storage across any number of Azure's geographic regions. It offers throughput, latency, availability, and consistency guarantees with comprehensive service level agreements (SLAs).
You can read data from and write data to Azure Cosmos DB using Databricks.
Note on fluid schema:
If you are managing data whose structures are constantly changing at a high rate, particularly if transactions can come from external sources where it is difficult to enforce conformity across the database, you may want to consider a more schema-agnostic approach using a managed NoSQL database service like Azure
Cosmos DB.
Reference:
https://docs.databricks.com/data/data-sources/azure/cosmosdb-connector.html https://docs.microsoft.com/en-us/azure/cosmos-db/relational-nosql

**davem0193** 4 years ago

I guess anywhere a question says 'multi region', the answer has to be cosmosdb :)

upvoted 2 times

**sjain91** 4 years, 2 months ago

Cosmos DB 100%

upvoted 4 times

Question #42 *Topic 1*

You are designing a log storage solution that will use Azure Blob storage containers.

CSV log files will be generated by a multi-tenant application. The log files will be generated for each customer at five-minute intervals. There will be more than

5,000 customers. Typically, the customers will query data generated on the day the data was created.

You need to recommend a naming convention for the virtual directories and files. The solution must minimize the time it takes for the customers to query the log files.

What naming convention should you recommend?

A. {year}/{month}/{day}/{hour}/{minute}/{CustomerID}.csv

B. {year}/{month}/{day}/{CustomerID}/{hour}/{minute}.csv

C. {minute}/{hour}/{day}/{month}/{year}/{CustomeriD}.csv

D. {CustomerID}/{year}/{month}/{day}/{hour}/{minute}.csv

**Suggested Answer:** *B*

Reference:

https://docs.microsoft.com/en-us/azure/cdn/cdn-azure-diagnostic-logs

**Geo_Barros** Highly Voted 4 years, 3 months ago

In my opinion, option "D" would be the right one.

upvoted 42 times

**cadio30** 4 years, 1 month ago

Referencing the link that was provided in the solution, it was stated in the blob path that it started using the "profile name" then proceed with the datetime stamp. It make sense that 'D' is the appropriate answer in this question.

upvoted 4 times

**rahul_t** Highly Voted 4 years, 2 months ago

I think B is correct. We want to minimize the time it takes for customers to query log files. 'Typically, the customers will query data generated on the day the data was created'. So it makes sense to include the path for a particular day i.e {Year}/{Month}/{Day} close to the start. Once we have reached a particular day then we will want to filter for a particular Customer so {Year}/{Month}/{Day}/{CustomerID}. Then we will want to aggregate down to hour and minute. The only other viable option will be D. The reason I think {CustomerID} should NOT be at the beginning of the path is in the case a Customer wants to query data related to multiple CustomerIDs on the same day.

upvoted 11 times

**Marcus1612** Most Recent 3 years, 9 months ago

I think the key word is "Multi-tenant". It appears to me that the logs for a single customer need to be under its own branch. D is the right answer

upvoted 3 times

**J4C7** 3 years, 10 months ago

what is correct answer i'm confused between B and D?

upvoted 1 times

**msn1712** 4 years ago

Why now A be the correct answer? On the link - https://docs.microsoft.com/en-us/azure/cdn/cdn-azure-diagnostic-logs, it's mentioned:

The name of the blob follows the following naming convention:

resourceId=/SUBSCRIPTIONS/{Subscription Id}/RESOURCEGROUPS/{Resource Group Name}/PROVIDERS/MICROSOFT.CDN/PROFILES/{Profile Name}/ENDPOINTS/{Endpoint Name}/ y={Year}/m={Month}/d={Day}/h={Hour}/m={Minutes}/PT1H.json

y={Year}/m={Month}/d={Day}/h={Hour}/m={Minutes}/PT1H.json

upvoted 1 times

**Alekx42** 4 years ago

Since it is stated that this is a multi-tenant application, customers would not (and probably should not be able to) query data of other customers. This makes D the right answer.

Moreover, while it said that typically the queries are done on the same day the data is created, this does not exclude the possibility of making queries

that range across multiple days or months. With solution B this becomes unpleasant, since you cannot just query year/month since that will return data of all customers for that month. With solution D all queries are easier, since customerID/year/month returns immediately all the data for that customer of that month.
Basically, while it is true that both B and D allow for rapid quering of data for a single customer for a single day, B is worse for all queries that want data of more than 1 day.

upvoted 4 times

**tes** 4 years ago

"this does not exclude the possibility of making queries " that is additional assumption made the person who is supposed to answer it.

upvoted 1 times

**BigMF** 4 years ago

All of these options are poor in my opinion and therefore hard to choose a “best” option. If it were me, I’d go with this: {CustomerID}/{year}/{month}/{day}/{CustomerID}_{year}{month}{day}{hour}{minute}.csv. This allows a customer to go directly to their folder and drill down quickly to the day they need. It also has the added benefit of the files being named intelligently and not just a “single bit of info”.csv. It also allows for easier maintenance down the road when customers leave by allowing you to easily archive or delete their data simply by archiving or deleting their folder. All that being said, I would go with D because I don’t think it is any slower for a customer to search for their data following that path than any of the others and in fact probably quicker. Also, it would provide easier maintenance down the road.

upvoted 1 times

**Mandar77** 4 years ago

I think, Answer B is correct. This is how you would like to restrict the access. question says, customer will access log information on the same day. So if you organize containers on year - month -day -customer - hour - time way, every customer has to come to day folder of that year and month and go to his container to get logs for the day.
If you organize container based on customer - year - month -day - hour - time, every customer has to traverse the long search path to get to day to get the logs. With option B, searching path would be optimum considering requirement

upvoted 3 times

**BigMF** 4 years ago

This logic is flawed because the customer still has to traverse a long search path when they drill down into the folder structure. You either traverse it to begin with or later in the drill down.

upvoted 1 times

**tanza** 4 years, 1 month ago

I think answer is A

upvoted 4 times

**Apox** 4 years, 2 months ago

I am certain that B is wrong. Why should Customer ID be put randomly in between the data formats?

I think D is the right answer and the reason is that each "/" takes you to a new directory (folder). As a hierarchy it would make the most sense to have a folder per customer, and then sort by date/time. Source: "Blob Path Format" Section here: https://docs.microsoft.com/en-us/azure/cdn/cdn-azure-diagnostic-logs#blob-path-format

upvoted 4 times

**KRV** 4 years, 1 month ago

By the looks of the question overall your argument holds good however if you read the question carefully it says ...
1. customers will query data generated on the day the data was created --> means it should start with a year to day granularity then
2. log files will be generated for each customer at five-minute intervals --> Now you are left with 2 options either organize by customer ID / hr/min or hr/min customer ID , given the case and nothing is explicility mentioned it is safe to assume that queries will be more customer centric and then within customer at a point in time and hence answer A happens to be logically more correct in the context of question !
{year}/{month}/{day}/{CustomerID}/{hour}/{minute}.csv

upvoted 4 times

**maynard13x8** 4 years, 2 months ago

Answer is correct. D is wrong because you duplicate year and month folders. It is also worse option because consumers query data of the day so, when you set the name, you already have all the data you are interested in.

upvoted 2 times

**Kevin89** 4 years, 2 months ago

The name of the blob follows the following naming convention:

resourceId=/SUBSCRIPTIONS/{Subscription Id}/RESOURCEGROUPS/{Resource Group Name}/PROVIDERS/MICROSOFT.CDN/PROFILES/{Profile

Name}/ENDPOINTS/{Endpoint Name}/ y={Year}/m={Month}/d={Day}/h={Hour}/m={Minutes}/PT1H.json

so it should actually be answer a

upvoted 3 times

⊟ **Nik71** 4 years, 3 months ago

confuse between A and B after reviewing https://docs.microsoft.com/en-us/azure/cdn/cdn-azure-diagnostic-logs feels like why we avoid A here.

upvoted 1 times

⊟ **Neha14n** 4 years, 3 months ago

Typically, the customers will query data generated on the day the data was created.

This line clears query will be specific to date not customer. Or else D would be correct answer

upvoted 3 times

⊟ **DongDuong** 4 years, 2 months ago

agree, in this case B is more suitable

upvoted 1 times

⊟ **AlexD332** 4 years, 3 months ago

still not clear as query should be optimized for customers - they won't request not their data.

upvoted 4 times

## Question #43 *Topic 1*

You are designing an Azure Cosmos DB database that will contain news articles.
The articles will have the following properties: Category, Created Datetime, Publish Datetime, Author, Headline, Body Text, and Publish Status. Multiple articles will be published in each category daily, but no two stories in a category will be published simultaneously.
Headlines may be updated over time. Publish Status will have the following values: draft, published, updated, and removed. Most articles will remain in the published or updated status. Publish Datetime will be populated only when Publish Status is set to published.
You will serve the latest articles to websites for users to consume.
You need to recommend a partition key for the database container. The solution must ensure that the articles are served to the websites as quickly as possible.
Which partition key should you recommend?

A. Publish Status

B. Category + Created Datetime

C. Headline

D. Publish Date + random suffix

**Suggested Answer:** *B*
You can form a partition key by concatenating multiple property values into a single artificial partitionKey property. These keys are referred to as synthetic keys.
Incorrect Answers:
D: Publish Datetime will be populated only when Publish Status is set to published.
Reference:
https://docs.microsoft.com/en-us/azure/cosmos-db/synthetic-partition-keys

**anamaster** Highly Voted 4 years, 2 months ago
and the publish status and headline will change
upvoted 6 times

**hello_there_** 3 years, 10 months ago
And publish date as well, from null to some value. Changing partition keys is not allowed, so only possible answer is B
upvoted 2 times

**cadio30** Most Recent 4 years, 1 month ago
the propose solution is correct. Publish datetime is not an option here as the partition key should be in string or integer
upvoted 3 times

Question #44 *Topic 1*

You are designing a product catalog for a customer. The product data will be stored in Azure Cosmos DB. The product properties will be different for each product and additional properties will be added to products as needed.
Which Cosmos DB API should you use to provision the database?

A. Cassandra API

B. Core (SQL) API

C. Gremlin API

**Suggested Answer:** *A*
Cassandrsa is a type of NoSQL database.
NoSQL database (sometimes called as Not Only SQL) is a database that provides a mechanism to store and retrieve data other than the tabular relations used in relational databases.
Incorrect Answers:
B: Core (SQL) API is a relational database which does not fit this scenario.
C: Gremlin is the graph traversal language of Apache TinkerPop. Gremlin is a functional, data-flow language that enables users to succinctly express complex traversals on (or queries of) their application's property graph.
Reference:
https://www.tutorialspoint.com/cassandra/cassandra_introduction.htm

**rmk4ever** Highly Voted 4 years, 2 months ago
Ans: Core (SQL) API
ref:
https://docs.microsoft.com/en-us/learn/modules/choose-api-for-cosmos-db/4-use-the-core-sql-api-to-store-a-product-catalog
upvoted 33 times

**maynard13x8** Highly Voted 4 years, 2 months ago
As Microsoft recommend, when you create a cosmos dB from scratch and there isn't any previous work that you could reuse, you should use sql api , unless you need relationships between data, in which case you should use gremnlin.
upvoted 9 times

**rmk4ever** 4 years, 2 months ago
you are right
upvoted 1 times

**Mily94** 4 years, 1 month ago
are you sure? "he product properties will be different for each product and additional properties will be added to products as needed" indicates NoSQL (Cassandra)
upvoted 5 times

**corebit** Most Recent 3 years, 6 months ago
"You've decided to look at how the new project is going to store the catalog for your customer facing e-commerce site. The sales team is likely to need support for adding new product categories quickly. The team had issues in the past as the old system that was using a relational database was too structured. Any necessary changes to add properties to products required downtime to update the table schemas, queries, and databases."

"Supporting new product categories is an important requirement for your project, and the Core (SQL) schema is flexible and requires a schemaless data store."

https://docs.microsoft.com/en-us/learn/modules/choose-api-for-cosmos-db/4-use-the-core-sql-api-to-store-a-product-catalog
upvoted 1 times

**tes** 4 years ago
"Cassandra This API isn't a good choice in this particular scenario, because the schema is unknown and will change over time."
https://docs.microsoft.com/en-us/learn/modules/choose-api-for-cosmos-db/4-use-the-core-sql-api-to-store-a-product-catalog
upvoted 2 times

**cadio30** 4 years, 1 month ago

By all means it is "SQL API"

Reference: https://docs.microsoft.com/en-us/learn/modules/choose-api-for-cosmos-db/

upvoted 1 times

**toandm** 4 years, 1 month ago

Answer is Core (SQL) API. Core (SQL) API is a document database, which is also NoSQL database. It has the name SQL because you can use SQL language to query it, not because it is relational DB

upvoted 3 times

**Hrabia** 4 years, 1 month ago

A. Cassandra API is correct. the product catalog was an example for cassandra api in the MS Learning Path for this exam.

upvoted 2 times

**HeywwooodJab** 4 years ago

are you sure about that? https://docs.microsoft.com/en-us/learn/modules/choose-api-for-cosmos-db/4-use-the-core-sql-api-to-store-a-product-catalog

"Cassandra - This API isn't a good choice in this particular scenario, because the schema is unknown and will change over time."

upvoted 2 times

Question #45 *Topic 1*

You work for a finance company.
You need to design a business network analysis solution that meets the following requirements:
✑ Analyzes the flow of transactions between the Azure environments of the company's various partner organizations
✑ Supports Gremlin (graph) queries
What should you include in the solution?

A. Azure Cosmos DB

B. Azure Synapse

C. Azure Analysis Services

D. Azure Data Lake Storage Gen2

**Suggested Answer:** *A*
Gremlin is one of the most popular query languages for exploring and analyzing data modeled as property graphs. There are many graph-database vendors out there that support Gremlin as their query language, in particular Azure Cosmos DB which is one of the world's first self-managed, geo-distributed, multi-master capable graph databases.
Azure Synapse Link for Azure Cosmos DB is a cloud native hybrid transactional and analytical processing (HTAP) capability that enables you to run near real-time analytics over operational data. Synapse Link creates a tight seamless integration between Azure Cosmos DB and Azure Synapse Analytics.
Reference:
https://jayanta-mondal.medium.com/analyzing-and-improving-the-performance-azure-cosmos-db-gremlin-queries-7f68bbbac2c
https://docs.microsoft.com/en-us/azure/cosmos-db/synapse-link-use-cases

⊟ **SandeshVartak** Highly Voted 👍 4 years ago

Gremlin API is supported by Cosmos DB only.

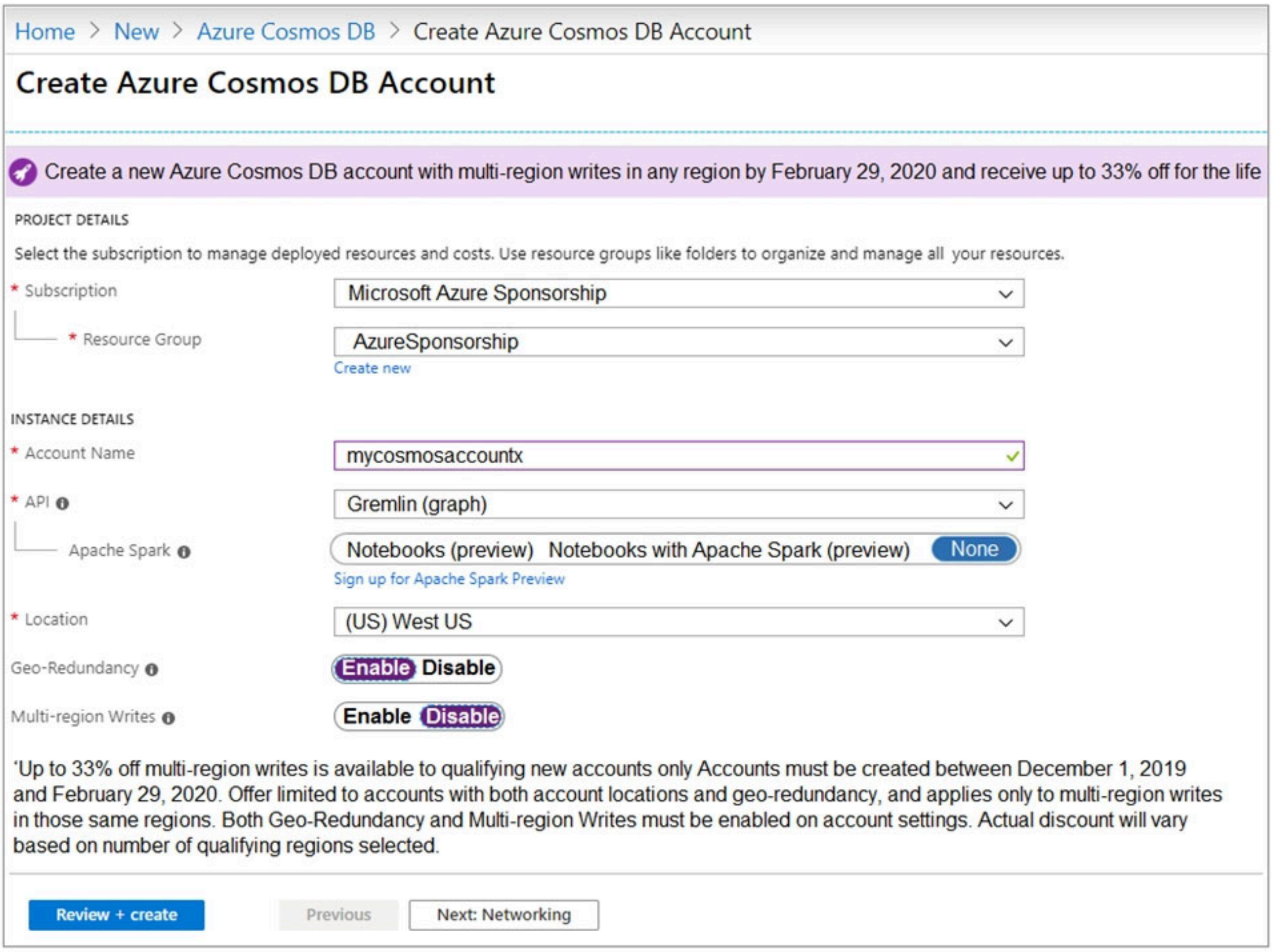
upvoted 5 times

Question #46 *Topic 1*

HOTSPOT -

You are evaluating the use of an Azure Cosmos DB account for a new database.

The proposed account will be configured as shown in the following exhibit.



Home > New > Azure Cosmos DB > Create Azure Cosmos DB Account

## Create Azure Cosmos DB Account

Create a new Azure Cosmos DB account with multi-region writes in any region by February 29, 2020 and receive up to 33% off for the life

PROJECT DETAILS

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

* Subscription: Microsoft Azure Sponsorship

* Resource Group: AzureSponsorship

Create new

INSTANCE DETAILS

* Account Name: mycosmosaccountx ✓

* API: Gremlin (graph)

Apache Spark: Notebooks (preview) Notebooks with Apache Spark (preview) [✓ None]

Sign up for Apache Spark Preview

* Location: (US) West US

Geo-Redundancy: [✓ Enable] Disable

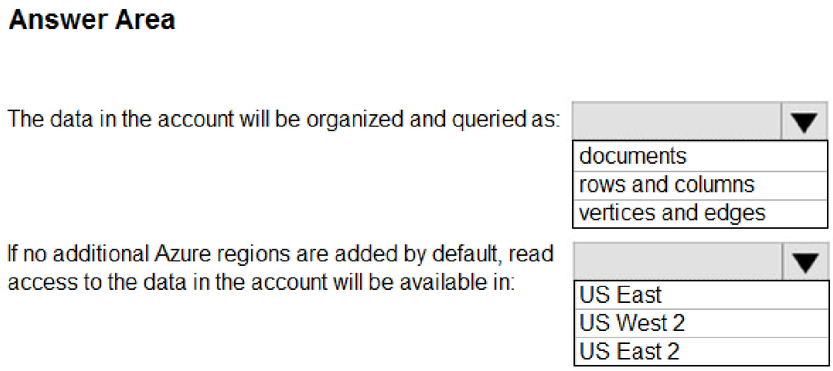
Multi-region Writes: Enable [✓ Disable]

'Up to 33% off multi-region writes is available to qualifying new accounts only Accounts must be created between December 1, 2019 and February 29, 2020. Offer limited to accounts with both account locations and geo-redundancy, and applies only to multi-region writes in those same regions. Both Geo-Redundancy and Multi-region Writes must be enabled on account settings. Actual discount will vary based on number of qualifying regions selected.

Use the drop-down menus to select the answer choice that completes each statement based on the information presented in the graphic.

NOTE: Each correct selection is worth one point.

Hot Area:



### Answer Area

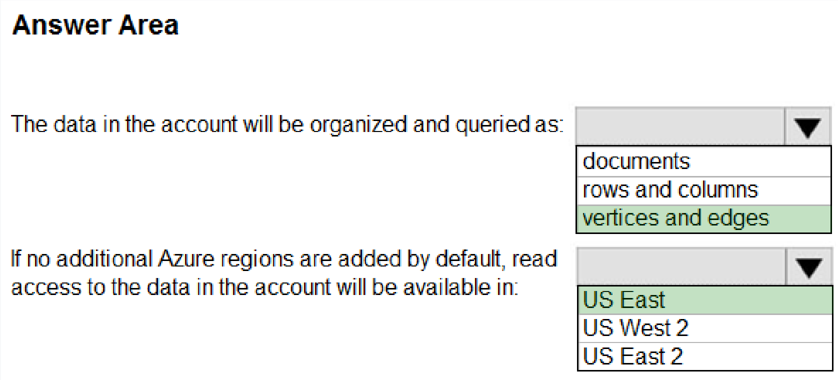
The data in the account will be organized and queried as: ▼

- documents
- rows and columns
- vertices and edges

If no additional Azure regions are added by default, read access to the data in the account will be available in: ▼

- US East
- US West 2
- US East 2

**Suggested Answer:**



**Answer Area**

The data in the account will be organized and queried as: ▼

- documents
- rows and columns
- 🟩 vertices and edges

If no additional Azure regions are added by default, read access to the data in the account will be available in: ▼

- 🟩 US East
- US West 2
- US East 2

Box 1: vertices and edges -

Gremlin API is selected.

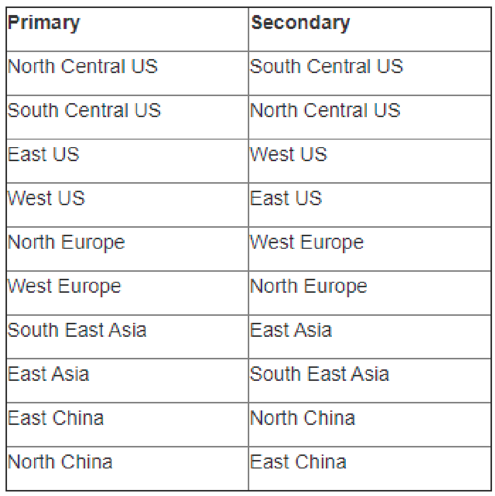
You can use the Gremlin language to create graph entities (vertices and edges), modify properties within those entities, perform queries and traversals, and delete entities.

Box 2: US East -

The (US) West US is selected as the primary location and geo- redundancy is enabled.

The secondary location for West US is East US.

Note: When a storage account is created, the customer chooses the primary location for their storage account. However, the secondary location for the storage account is fixed and customers do not have the ability to change this. The following table shows the current primary and secondary location pairings:



| Primary | Secondary |
|---|---|
| North Central US | South Central US |
| South Central US | North Central US |
| East US | West US |
| West US | East US |
| North Europe | West Europe |
| West Europe | North Europe |
| South East Asia | East Asia |
| East Asia | South East Asia |
| East China | North China |
| North China | East China |

Reference:

https://docs.microsoft.com/en-us/azure/cosmos-db/gremlin-support https://technet2.github.io/Wiki/blogs/windowsazurestorage/windows-azure-storage-redundancy-options-and-read-access-geo-redundant-storage.html

⊟ **Mittun** 4 years, 2 months ago

US West is selected in the picture - hence 2nd answer is US West.

upvoted 2 times

⊟ **DongDuong** 4 years, 2 months ago

Wrong, as explained The (US) West US is selected as the primary location and geo-redundancy is enabled. The secondary location for West US is East US. So East US is the correct answer.

upvoted 27 times

⊟ **suvenk** 4 years ago

Thats right they fall under the paired regions, hence it is US East.

upvoted 1 times

⊟ **cadio30** 4 years ago

i agree with the explanation, some of the regions do have corresponding default secondary location whenever the geo-redundancy is enable and if there are no specified location in it.

upvoted 1 times

Question #47 *Topic 1*

You are designing a streaming solution that must meet the following requirements:

✑ Accept input data from an Azure IoT hub.

✑ Write aggregated data to Azure Cosmos DB.

✑ Calculate minimum, maximum, and average sensor readings every five minutes.

✑ Define calculations by using a SQL query.

✑ Deploy to multiple environments by using Azure Resource Manager templates.

What should you include in the solution?

A. Azure Functions

B. Azure HDInsight with Spark Streaming

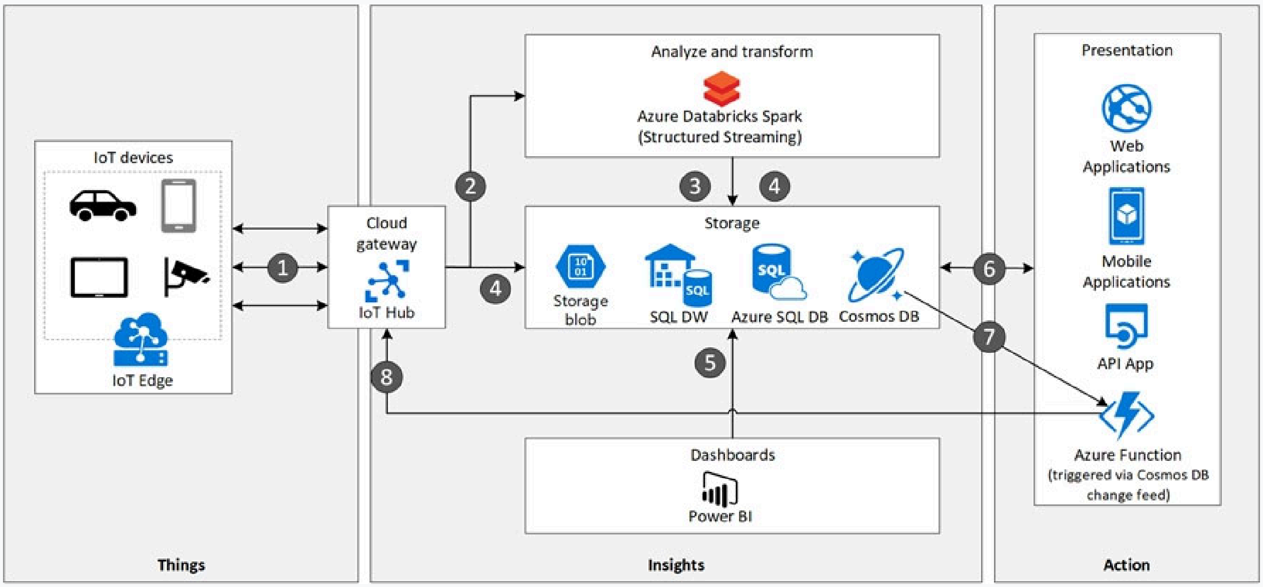
C. Azure Databricks

D. Azure Stream Analytics

**Suggested Answer:** *C*

Cosmos DB is ideally suited for IoT solutions. Cosmos DB can ingest device telemetry data at high rates.

Architecture -



Data flow -

1. Events generated from IoT devices are sent to the analyze and transform layer through Azure IoT Hub as a stream of messages. Azure IoT Hub stores streams of data in partitions for a configurable amount of time.

2. Azure Databricks, running Apache Spark Streaming, picks up the messages in real time from IoT Hub, processes the data based on the business logic and sends the data to Serving layer for storage. Spark Streaming can provide real time analytics such as calculating moving averages, min and max values over time periods.

3. Device messages are stored in Cosmos DB as JSON documents.

Reference:

https://docs.microsoft.com/en-us/azure/architecture/solution-ideas/articles/iot-using-cosmos-db

⊟ **suman13** Highly Voted 👍 4 years, 2 months ago

ARM template can only be used to create databricks workspace but not for the application(notebooks),cluster etc. whereas it can be used for stream analytics. Hence Stream Analytics should be the answer here

upvoted 32 times

⊟ **KRV** Highly Voted 👍 4 years, 2 months ago

Since the question clearly specifies to define calculations by using a SQL query as well as the points that minimum, maximum, and average sensor readings are to be calculated every five minutes

using Windows functions within the Azure stream analytics would be a straight and preffered option.

One can always use Azure Databricks , however for minimal code using SQL and windows functions , the best possible solution ideally should be Azure Stream Analytics.

upvoted 13 times

⊟ **Ankush1994** Most Recent 3 years, 10 months ago

Azure Databricks, running Apache Spark Streaming, picks up the messages in real time from IoT Hub, processes the data based on the business logic and sends the data to Serving layer for storage.

upvoted 1 times

⊟ **tes** 4 years ago

In the link given there is an alternatives section which states for streaming Stream Analytics could be used as an alternative. So that is another plus to the argument that Stream Analytics should be the answer.

upvoted 1 times

⊟ **davem0193** 4 years ago

The architectural diagram provided as part of the solution clearly shows that it needs to be databricks although Stream analytics makes more sense. Solution provided is correct - it is databricks

upvoted 1 times

⊟ **tes** 4 years ago

in the same page alternatives are provided and one is stream analytics. ARM template deploy of jobs are possible there. Where as DBR notebooks cannot be deployed through arm templates

upvoted 2 times

⊟ **cadio30** 4 years, 1 month ago

D. Azure Stream Analytics is the appropriate solution for the requirements

upvoted 4 times

⊟ **maciejt** 4 years, 2 months ago

Azure functions is also present in the architecture. Why incorrect answer then?

upvoted 1 times

⊟ **SK1984** 4 years, 2 months ago

Why not D. Azure Stream analytics ?

upvoted 3 times

⊟ **maynard13x8** 4 years, 2 months ago

I think because asa jobs queries are not exactly sql. If that si not te case

upvoted 2 times

⊟ **maynard13x8** 4 years, 2 months ago

I would also choose asa (Azure stream analytics).

upvoted 3 times

Question #48 *Topic 1*

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.

After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You plan to store delimited text files in an Azure Data Lake Storage account that will be organized into department folders.

You need to configure data access so that users see only the files in their respective department folder.

Solution: From the storage account, you enable a hierarchical namespace, and you use access control lists (ACLs).

Does this meet the goal?

A. Yes

B. No

**Suggested Answer:** *B*

Azure Data Lake Storage implements an access control model that derives from HDFS, which in turn derives from the POSIX access control model.

Blob container ACLs does not support the hierarchical namespace, so it must be disabled.

Reference:

https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-known-issues

**niwe** Highly Voted 4 years, 2 months ago

Hierarchical namespace, must be enabled, so No.

https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-acl-dotnet

upvoted 12 times

**azurenav** 4 years ago

Question has "Enable",please check

upvoted 8 times

**Sennkumar** 4 years ago

It says enable, please check the question. Answer is "Yes"

upvoted 7 times

**cadio30** 4 years, 1 month ago

The link provided already stated that it requires to "ENABLE" the Hierarchical namespace which is included in the question. The appropriate answer is "Yes". Also this configuration is only available in Azure Data Lake Storage Gen 2 (azure storage blob)

upvoted 6 times

**Deasto** 4 years, 1 month ago

The question saying "you enable a hierarchical namespace" though. Did the question change?

Solution: From the storage account, you enable a hierarchical namespace, and you use access control lists (ACLs).

upvoted 7 times

**niwe** 4 years, 1 month ago

Yes you are right question has change

upvoted 5 times

**Psycho** 4 years, 1 month ago

So, the answer is Yes, right?

upvoted 7 times

**BobFar** 4 years ago

right, the answer is YES

upvoted 5 times

**crissw22** 4 years, 1 month ago

so, YES is the answer

upvoted 9 times

**satyamkishoresingh** Most Recent 3 years, 10 months ago

This should be Yes as solution Fit the problem statement ,

Enable Hierarchical namespace + ACL

upvoted 1 times

⊟ **tes** 4 years ago

you enable a hierarchical namespace = then the storage account becomes Gen2

Enable ACL: Gen2 automatically has ACL

so the answer is Yes

upvoted 1 times

⊟ **tes** 4 years ago

sorry ignore this, wrong answer. I cannot delete it. ACL is there in Gen1.

upvoted 1 times

⊟ **davita8** 4 years, 2 months ago

B. No The answer

upvoted 1 times

⊟ **Apox** 4 years, 2 months ago

The answer is "NO"

Hierarchical namespace must be enabled to have a folder structure and actually be an ADLS account (or else it is regular blob)

However, it is correct to use ACLs, as this is the only mechanism to give "finer grain" level of access to directories and files. (except Shared Access Signature, but this would make more sense to use for external users for e.g. a limited amount of time)

Source: https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control-model

upvoted 2 times

⊟ **azurenav** 4 years ago

Question has "Enable",please check

upvoted 1 times

⊟ **aksoumi** 4 years, 2 months ago

Also, note below from the Azure Documentation. In order to create ADLS account you have to enable Hierarchical option, else it is not ADLS. Hence correct ANSWER is "NO"

"You'll create a Data Lake Storage Gen2 account the same way you create an Azure Blob store, but with one setting difference. In Advanced, in the Data Lake Storage Gen2 (preview) section, next to Hierarchical namespace, select Enabled."

upvoted 3 times

⊟ **azurenav** 4 years ago

Question has "Enable",please check

upvoted 1 times

Question #49 *Topic 1*

You need to design a solution to support the storage of datasets. The solution must meet the following requirements:

✑ Send email alerts when new datasets are added.

✑ Control access to collections of datasets by using Azure Active Directory groups.

Support the storage of Microsoft Excel, Comma Separated Values (CSV), and zip files.

▪

What should you include in the solution?

A. Azure SQL Database

B. Azure Storage

C. Azure Cosmos DB

D. Azure HDInsight

**Suggested Answer:** *B*

Reference:

https://docs.microsoft.com/en-us/azure/architecture/data-guide/technology-choices/data-storage

Design data processing solutions

**KpKo** Highly Voted 👍 4 years, 1 month ago

Agreed, the answer is correct.

upvoted 5 times

**DingDongSingSong** Most Recent ⏱ 3 years, 3 months ago

The answer here should be CosmosDB. Access to datasets needs to be controlled. Storage account only provides container level access security at the most granular level which means all datasets will be available for anyone with access to storage account container and the respective blobs. That does not meet the access security requirement.

upvoted 1 times

**satyamkishoresingh** 3 years, 10 months ago

ACD are definitely not so B should be the answer

upvoted 1 times

**memo43** 4 years, 1 month ago

answer is CORRECT

upvoted 4 times

Question #1 *Topic 2*

You are designing an Azure Databricks interactive cluster. The cluster will be used infrequently and will be configured for auto-termination.
You need to ensure that the cluster configuration is retained indefinitely after the cluster is terminated. The solution must minimize costs.
What should you do?

A. Clone the cluster after it is terminated.

B. Terminate the cluster manually when processing completes.

C. Create an Azure runbook that starts the cluster every 90 days.

D. Pin the cluster.

**Suggested Answer:** *D*
To keep an interactive cluster configuration even after it has been terminated for more than 30 days, an administrator can pin a cluster to the cluster list.
Reference:
https://docs.azuredatabricks.net/clusters/clusters-manage.html#automatic-termination

**syu31svc** Highly Voted 👍 4 years, 6 months ago

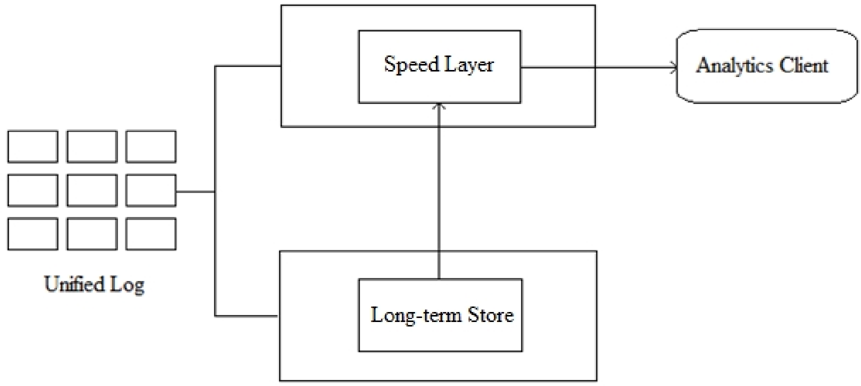
This is same as Topic 1 Qn 12
Answer is D

upvoted 12 times

Question #2 *Topic 2*

HOTSPOT -

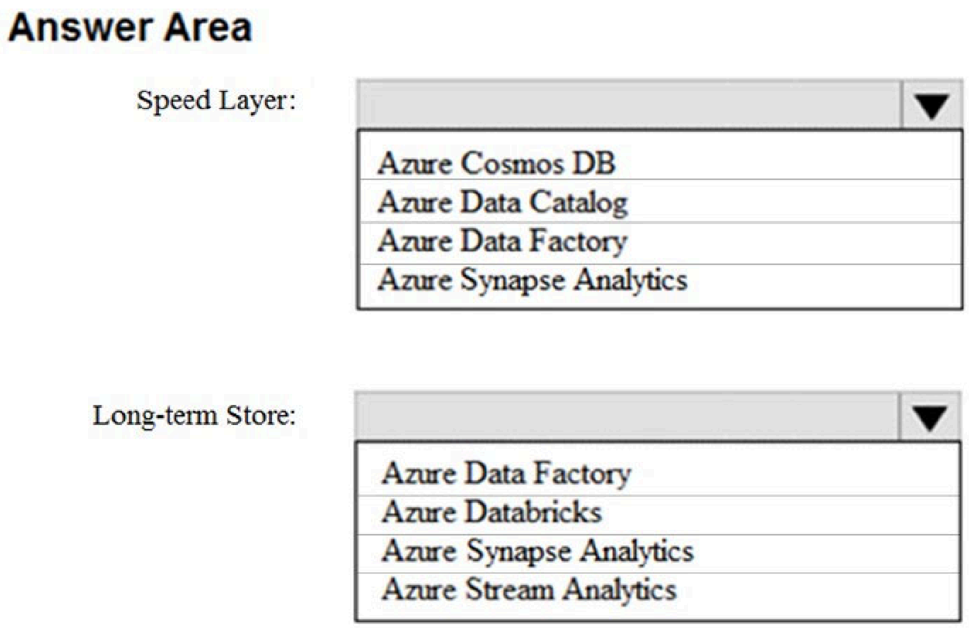
You are planning a design pattern based on the Kappa architecture as shown in the exhibit.



Which Azure service should you use for each layer? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:



## Answer Area

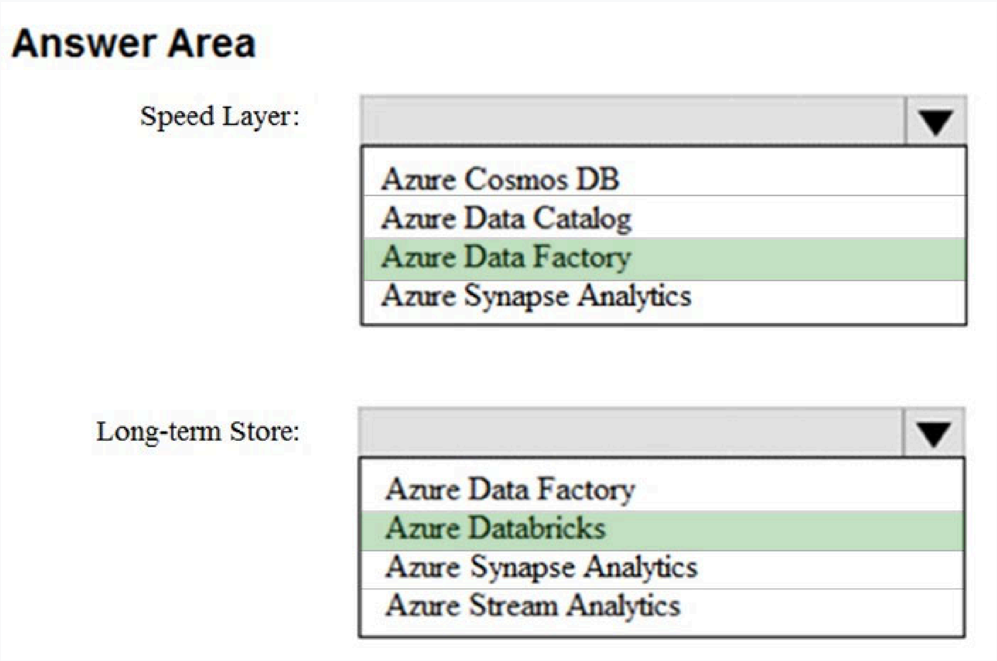
Speed Layer:

- Azure Cosmos DB
- Azure Data Catalog
- Azure Data Factory
- Azure Synapse Analytics

Long-term Store:

- Azure Data Factory
- Azure Databricks
- Azure Synapse Analytics
- Azure Stream Analytics

**Suggested Answer:**



## Answer Area

Speed Layer:

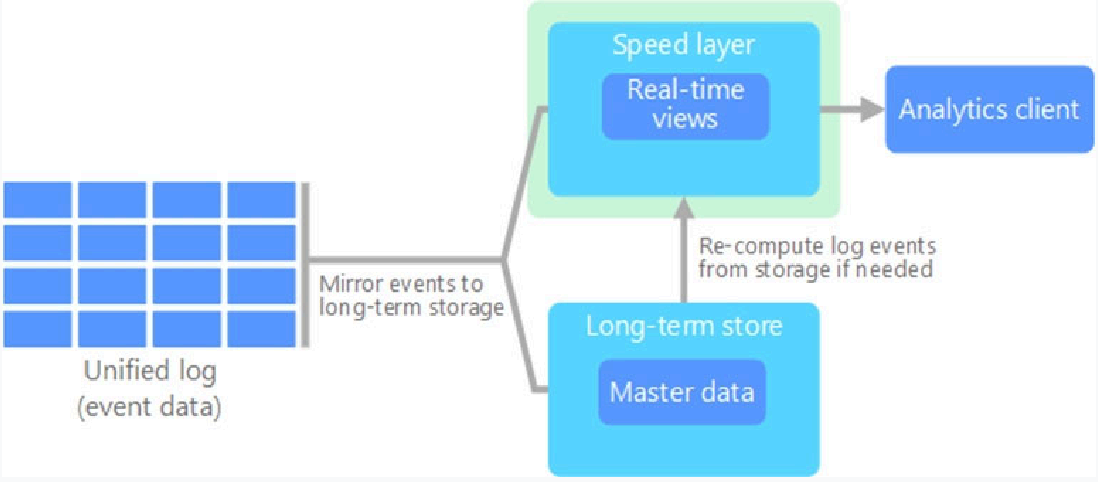
- Azure Cosmos DB
- Azure Data Catalog
- [✓ Azure Data Factory]
- Azure Synapse Analytics

Long-term Store:

- Azure Data Factory
- [✓ Azure Databricks]
- Azure Synapse Analytics
- Azure Stream Analytics

Layer 1: Azure Data Factory -

Layer 2: Azure Databricks -
Azure Databricks is fully integrated with Azure Data Factory .



Reference:
https://docs.microsoft.com/en-us/azure/architecture/data-guide/big-data/

⊟ **Anagarika** Highly Voted 4 years, 3 months ago
Speed layer: Cosmos DB, Long-term store: Synapse
upvoted 71 times

⊟ **cadio30** 4 years, 1 month ago
Agree with this solution
upvoted 5 times

⊟ **Mandar77** 4 years ago
Agree. I saw same question in another sample test .. Answer mention above is right.
upvoted 4 times

⊟ **TIVIND** Highly Voted 4 years, 1 month ago
Speed layer: Cosmos , Long-term : Synapse
upvoted 6 times

⊟ **satyamkishoresingh** Most Recent 3 years, 10 months ago
Cosmos DB + ASA fit appropriately for this question I think.
upvoted 1 times

⊟ **mogashe12** 4 years, 2 months ago
The kappa architecture was proposed by Jay Kreps as an alternative to the lambda architecture. It has the same basic goals as the lambda architecture, but with an important distinction: All data flows through a single path, using a stream processing system ,Analytical data store
Azure Synapse Analytics, Azure Data Explorer, HBase, Spark, or Hive. Processed real-time data can be stored in a relational database such Synapse Analytics, Azure Data Explorer, a NoSQL store such as HBase, or as files in distributed storage over which Spark or Hive tables can be defined and queried.
https://docs.microsoft.com/en-us/azure/architecture/data-guide/big-data/real-time-processing
upvoted 1 times

⊟ **RedSquirrel** 4 years, 3 months ago
speed layer should be Cosmos DB, based on following article. Long term is correct. Data stored in Parquets.
https://azure.microsoft.com/en-us/blog/lambda-architecture-using-azure-cosmosdb-faster-performance-low-tco-low-devops/
upvoted 1 times

⊟ **malay1232489** 4 years, 3 months ago
There must be a reason why you said that
upvoted 1 times

⊟ **H_S** 4 years, 3 months ago
i don't think this's the right answer
upvoted 3 times

Question #3 *Topic 2*

You need to design a telemetry data solution that supports the analysis of log files in real time.
Which two Azure services should you include in the solution? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.

A. Azure Databricks

B. Azure Data Factory

C. Azure Event Hubs

D. Azure Data Lake Storage Gen2

**Suggested Answer:** *AC*
You connect a data ingestion system with Azure Databricks to stream data into an Apache Spark cluster in near real-time. You set up data ingestion system using
Azure Event Hubs and then connect it to Azure Databricks to process the messages coming through.
Note: Azure Event Hubs is a highly scalable data streaming platform and event ingestion service, capable of receiving and processing millions of events per second. Event Hubs can process and store events, data, or telemetry produced by distributed software and devices. Data sent to an event hub can be transformed and stored using any real-time analytics provider or batching/storage adapters.
Reference:
https://docs.microsoft.com/en-us/azure/azure-databricks/databricks-stream-from-eventhubs

**maynard13x8** Highly Voted 4 years, 2 months ago

Correct answer.

upvoted 11 times

**cadio30** Most Recent 4 years, 1 month ago

Propose solution is correct

upvoted 2 times

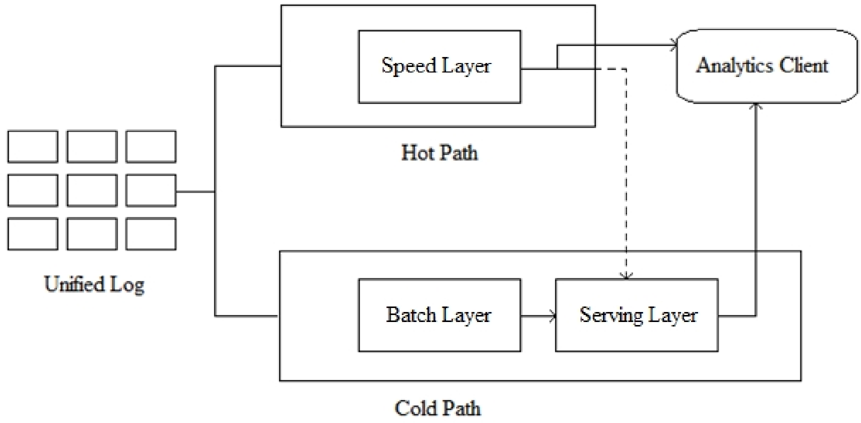
**bdloko** 4 years, 2 months ago

A: for realtime analysis.
B: for stream injection.

upvoted 1 times

Question #4 *Topic 2*

You are planning a design pattern based on the Lambda architecture as shown in the exhibit.



Which Azure service should you use for the hot path?

A. Azure Databricks

B. Azure Data Lake Storage Gen2

C. Azure Data Factory

D. Azure Synapse Analytics

**Suggested Answer:** *A*

In Azure, all of the following data stores will meet the core requirements supporting real-time processing:

✑ Apache Spark in Azure Databricks

✑ Azure Stream Analytics

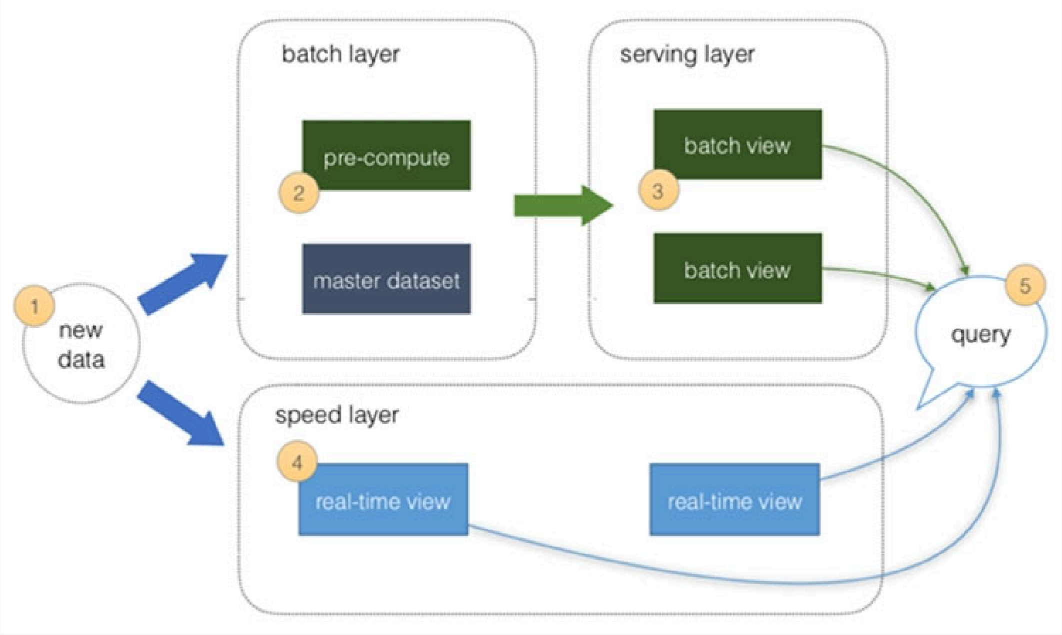
✑ HDInsight with Spark Streaming

✑ HDInsight with Storm

✑ Azure Functions

✑ Azure App Service WebJobs

Note: Lambda architectures use batch-processing, stream-processing, and a serving layer to minimize the latency involved in querying big data.



Reference:

https://azure.microsoft.com/en-us/blog/lambda-architecture-using-azure-cosmosdb-faster-performance-low-tco-low-devops/

https://docs.microsoft.com/en-us/azure/architecture/data-guide/technology-choices/stream-processing

**111222333** 4 years, 1 month ago

Why not Azure Stream Analytics? A speed layer (hot path) analyzes data in real time. Databricks is for batch processing.

https://docs.microsoft.com/en-us/azure/architecture/data-guide/big-data/#lambda-architecture

upvoted 3 times

⊟ **111222333** 4 years, 1 month ago

Sorry, I see now that Azure Stream Analytics is not a proposed solution, so Databricks is the best option among others.

upvoted 4 times

⊟ **bdloko** 4 years, 2 months ago

Streaming vs Batching...

upvoted 1 times

Question #5 *Topic 2*

You are designing an audit strategy for an Azure SQL Database environment.
You need to recommend a solution to provide real-time notifications for potential security breaches. The solution must minimize development effort.
Which destination should you include in the recommendation?

A. Azure Blob storage

B. Azure Synapse Analytics

C. Azure Event Hubs

D. Azure Log Analytics

**Suggested Answer:** *D*
Auditing for Azure SQL Database and SQL Data Warehouse tracks database events and writes them to an audit log in your Azure storage account, Log Analytics workspace or Event Hubs.
Alerts in Azure Monitor can identify important information in your Log Analytics repository. They are created by alert rules that automatically run log searches at regular intervals, and if results of the log search match particular criteria, then an alert record is created and it can be configured to perform an automated response.
Reference:
https://docs.microsoft.com/en-us/azure/sql-database/sql-database-auditing https://docs.microsoft.com/en-us/azure/azure-monitor/learn/tutorial-response

**bdloko** Highly Voted 4 years, 2 months ago
Not C: Solution must minimize development effort.
D: Alerts in Azure Monitor can identify important information in your Log Analytics repository.
upvoted 14 times

**cadio30** Most Recent 4 years, 1 month ago
Appropriate answer is azure log analytics as the it can be configured easily by sending the logs into the component
upvoted 2 times

**maynard13x8** 4 years, 2 months ago
Destination should be Azure storage because it's the place where the logs are saved when you enable auditing.
upvoted 3 times

**maynard13x8** 4 years, 2 months ago
Sorry. D is correct. Bdloko is right.
upvoted 3 times

Question #6 *Topic 2*

You need to design a real-time stream solution that uses Azure Functions to process data uploaded to Azure Blob Storage.
The solution must meet the following requirements:
Support up to 1 million blobs.

▪

✑ Scaling must occur automatically.
✑ Costs must be minimized.
What should you recommend?

A. Deploy the Azure Function in an App Service plan and use a Blob trigger.

B. Deploy the Azure Function in a Consumption plan and use an Event Grid trigger.

C. Deploy the Azure Function in a Consumption plan and use a Blob trigger.

D. Deploy the Azure Function in an App Service plan and use an Event Grid trigger.

**Suggested Answer:** *C*
Create a function, with the help of a blob trigger template, which is triggered when files are uploaded to or updated in Azure Blob storage.
You use a consumption plan, which is a hosting plan that defines how resources are allocated to your function app. In the default Consumption Plan, resources are added dynamically as required by your functions. In this serverless hosting, you only pay for the time your functions run.
When you run in an App Service plan, you must manage the scaling of your function app.
Reference:
https://docs.microsoft.com/en-us/azure/azure-functions/functions-create-storage-blob-triggered-function

**sdas1** Highly Voted 👍 4 years, 3 months ago
The solution is B - Deploy the Azure Function in a Consumption plan and use an Event Grid trigger. 1M blobs will cripple the ability of blob trigger to provide the events.
The EventGrid trigger is instantaneous, so it depends on your needs.
upvoted 33 times

**vrmei** 4 years ago
High-scale: High scale can be loosely defined as containers that have more than 100,000 blobs in them or storage accounts that have more than 100 blob updates per second.
upvoted 1 times

**H_S** 4 years, 3 months ago
you're right
In addition, storage logs are created on a "best effort" basis. There's no guarantee that all events are captured. Under some conditions, logs may be missed.

If you require faster or more reliable blob processing, consider creating a queue message when you create the blob. Then use a queue trigger instead of a blob trigger to process the blob. Another option is to use Event Grid; see the tutorial Automate resizing uploaded images using Event Grid.
https://docs.microsoft.com/en-us/azure/azure-functions/functions-bindings-storage-blob-trigger?tabs=csharp
upvoted 1 times

**H_S** 4 years, 3 months ago
Because it's a real time processing, it has to be appservice plan THE CORRECT ANSWER IS D
upvoted 2 times

**jms309** 4 years, 3 months ago
I would say it is B as because of thee High number of blobs that can land it has to be definitely an Event Grid Trigger (See https://docs.microsoft.com/es-es/azure/azure-functions/functions-bindings-storage-blob-trigger?tabs=csharp#alternatives). Also, due to cost minimized, it should be a consumption plan. Even being a consumption plan the trigger delay will be minimum and fit to the conditions. However if you have a service plan you have to had the resources needed allocated even if you are using them or not.
upvoted 3 times

**miod** 4 years, 2 months ago
https://docs.microsoft.com/en-us/azure/azure-functions/functions-bindings-storage-blob-trigger?tabs=csharp#alternatives

upvoted 2 times

**cadio30** Most Recent 4 years, 1 month ago

The requirement is leading to the answer below as it was focusing to minimize cost.
B. Deploy the Azure Function in a Consumption plan and use an Event Grid trigger.

Reference: https://docs.microsoft.com/en-us/azure/azure-functions/functions-scale

upvoted 2 times

**davita8** 4 years, 2 months ago

B. Deploy the Azure Function in a Consumption plan and use an Event Grid trigger.

upvoted 3 times

**david112sdsf** 4 years, 3 months ago

It doesn't say you have to minimize latency.

upvoted 1 times

**eurekamike** 4 years ago

real-time = minimize latency

upvoted 3 times

Question #7 *Topic 2*

A company purchases IoT devices to monitor manufacturing machinery. The company uses an IoT appliance to communicate with the IoT devices.

The company must be able to monitor the devices in real-time.

You need to design the solution.

What should you recommend?

A. Azure Data Factory instance using Azure PowerShell

B. Azure Analysis Services using Microsoft Visual Studio

C. Azure Stream Analytics cloud job using Azure PowerShell

D. Azure Data Factory instance using Microsoft Visual Studio

**Suggested Answer:** *C*

Stream Analytics is a cost-effective event processing engine that helps uncover real-time insights from devices, sensors, infrastructure, applications and data quickly and easily.

Monitor and manage Stream Analytics resources with Azure PowerShell cmdlets and powershell scripting that execute basic Stream Analytics tasks.

Note: Visual Studio 2019 and Visual Studio 2017 also support Stream Analytics Tools.

Reference:

https://cloudblogs.microsoft.com/sqlserver/2014/10/29/microsoft-adds-iot-streaming-analytics-data-production-and-workflow-services-to-azure/

**SebK** Highly Voted 👍 5 years ago

I agree with answer C. But shoudn't be Azure Stream Analytics edge job ?

upvoted 21 times

> **watata** 4 years, 4 months ago
>
> yes, it should be this option
>
> upvoted 4 times

**tyler4kn** Highly Voted 👍 4 years, 8 months ago

In the exam, the answer is "Azure Stream Analytics with Azure Portal", not PowerShell. That is the correct answer.

upvoted 16 times

**realtp9** Most Recent 🕓 3 years, 10 months ago

on 31st August, it was Azure Stream Analytics Edge Application using Microsoft Visual Studio

upvoted 1 times

**syu31svc** 4 years, 6 months ago

It can only be stream analytics

upvoted 4 times

**Arsa** 4 years, 10 months ago

correct

upvoted 6 times

Question #8 *Topic 2*

HOTSPOT -

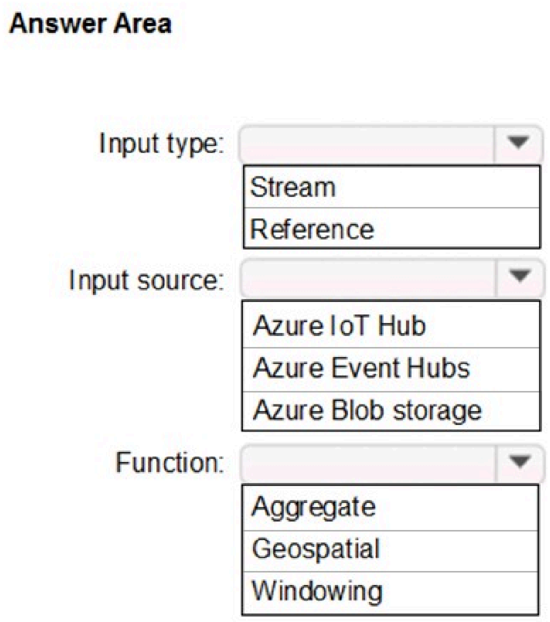
You plan to create a real-time monitoring app that alerts users when a device travels more than 200 meters away from a designated location.

You need to design an Azure Stream Analytics job to process the data for the planned app. The solution must minimize the amount of code developed and the number of technologies used.

What should you include in the Stream Analytics job? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:



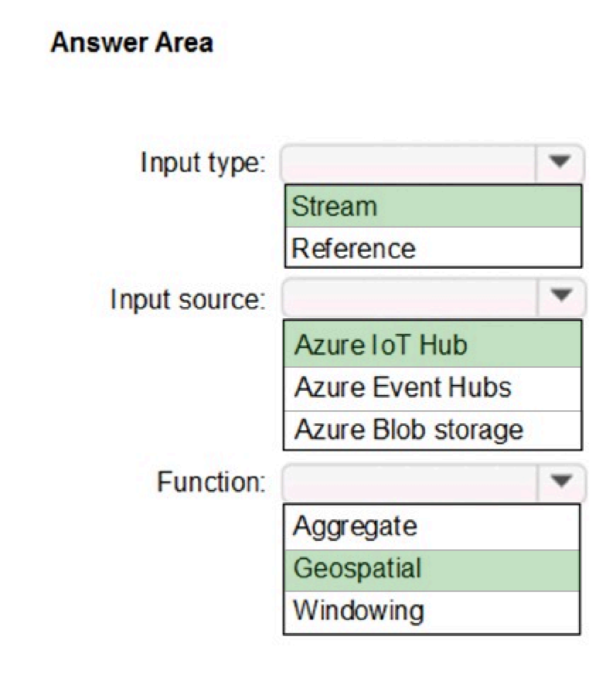
**Answer Area**

Input type:
- Stream
- Reference

Input source:
- Azure IoT Hub
- Azure Event Hubs
- Azure Blob storage

Function:
- Aggregate
- Geospatial
- Windowing

**Suggested Answer:**



**Answer Area**

Input type:
- [✓ Stream]
- Reference

Input source:
- [✓ Azure IoT Hub]
- Azure Event Hubs
- Azure Blob storage

Function:
- Aggregate
- [✓ Geospatial]
- Windowing

Input type: Stream -

You can process real-time IoT data streams with Azure Stream Analytics.

Input source: Azure IoT Hub -

In a real-world scenario, you could have hundreds of these sensors generating events as a stream. Ideally, a gateway device would run code to push these events to Azure Event Hubs or Azure IoT Hubs.

Function: Geospatial -

With built-in geospatial functions, you can use Azure Stream Analytics to build applications for scenarios such as fleet management, ride

sharing, connected cars, and asset tracking.
Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-get-started-with-azure-stream-analytics-to-process-data-from-iot-devices https://docs.microsoft.com/en-us/azure/stream-analytics/geospatial-scenarios

**AaronZ** Highly Voted 5 years, 2 months ago
Just finished my test, there is no second drop box for input source, only Input Type and Function. Maybe Event Hub and IoT hub both are correct now, so they removed the question.
upvoted 60 times

> **ChinChin** 5 years, 1 month ago
> Correct
> upvoted 2 times

**jsad** Highly Voted 5 years, 2 months ago
It says that the app needs to alert users
upvoted 14 times

**aksoumi** Most Recent 4 years, 2 months ago
How about box 1? Is the input a stream or a reference?
upvoted 2 times

> **anamaster** 4 years, 2 months ago
> stream
> upvoted 3 times

**nehab0101** 4 years, 10 months ago
yes both are correct event hub and Iot Hub
upvoted 3 times

**Arsa** 4 years, 10 months ago
correct answers
upvoted 3 times

**AhmedReda** 5 years ago
I thinks Answer = IoT, as per that link of Geospatial function
https://docs.microsoft.com/en-us/azure/stream-analytics/geospatial-scenarios
upvoted 7 times

> **cadio30** 4 years, 1 month ago
> good reference
> upvoted 1 times

**Abhilvs** 5 years ago
I guess it is IoT hub because IoT has the ability to integrate with Edge devices
upvoted 2 times

**MLCL** 5 years, 2 months ago
I think bothe Event Hub and IoT hub work in this scenario.
upvoted 2 times

**mclawson1966** 5 years, 3 months ago
Why IoT Hub? I don't see a requirement for bi-directional...
upvoted 4 times

## Question #9 *Topic 2*

A company purchases IoT devices to monitor manufacturing machinery. The company uses an IoT appliance to communicate with the IoT devices.

The company must be able to monitor the devices in real-time.

You need to design the solution.

What should you recommend?

A. Azure Data Factory instance using the Azure portal

B. Azure Analysis Services using Microsoft Visual Studio

C. Azure Stream Analytics Edge application using Microsoft Visual Studio

D. Azure Analysis Services using the Azure portal

**Suggested Answer:** *C*

Azure Stream Analytics (ASA) on IoT Edge empowers developers to deploy near-real-time analytical intelligence closer to IoT devices so that they can unlock the full value of device-generated data.

Reference:

https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-edge

**Nik71** Highly Voted 👍 4 years, 3 months ago

https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-tools-for-visual-studio-edge-jobs

upvoted 8 times

## Question #10 *Topic 2*

You plan to ingest streaming social media data by using Azure Stream Analytics. The data will be stored in files in Azure Data Lake Storage, and then consumed by using Azure Databricks and PolyBase in Azure Synapse Analytics.

You need to recommend a Stream Analytics data output format to ensure that the queries from Databricks and PolyBase against the files encounter the fewest possible errors. The solution must ensure that the files can be queried quickly and that the data type information is retained.

What should you recommend?

A. Avro

B. CSV

C. Parquet

D. JSON

**Suggested Answer:** *A*

The Avro format is great for data and message preservation.

Avro schema with its support for evolution is essential for making the data robust for streaming architectures like Kafka, and with the metadata that schema provides, you can reason on the data. Having a schema provides robustness in providing meta-data about the data stored in Avro records which are self- documenting the data.

References:

http://cloudurable.com/blog/avro/index.html

**felmasri** Highly Voted 4 years, 3 months ago

I think this Answer is wrong since polybase does not support Avro.

I will pick Parquet

upvoted 52 times

**jms309** Highly Voted 4 years, 3 months ago

I understand that Databricks and Polybase will consume the data independently ... So, based on that premise the selected output format from Synapse Stream Analytics should be a format compatible with both. Since, we need the file format to be a distributed file format for speed up the queries, the only possible solutions are AVRO and Parquet. As, AVRO is no a valid solution as Polybase doesn't support this format, the only possible answer is PARQUET

upvoted 15 times

**massnonn** Most Recent 3 years, 7 months ago

for me the correct answer is parquet

upvoted 1 times

**dumpi** 4 years ago

Parquet is correct answer I verify

upvoted 3 times

**KpKo** 4 years, 1 month ago

Agreed with Parquet

upvoted 2 times

**cadio30** 4 years, 1 month ago

Both services uses CSV and parquet as input files though parquet is the candidate for this requirement as it is the recommended file format for azure databricks and is also supported by polybase

upvoted 2 times

**davita8** 4 years, 2 months ago

C. Parquet

upvoted 3 times

**maciejt** 4 years, 2 months ago

JSON and CSV don't define the types strongly and we need to preserve the data types, so those 2 are exuded.

Parquet is better optimized for read, avro is for write and requirement is to make queries fast, so parquet.

https://www.datanami.com/2018/05/16/big-data-file-formats-demystified/

upvoted 7 times

⊟ **Nik71** 4 years, 3 months ago

its Parquet file format

upvoted 2 times

⊟ **al9887655** 4 years, 3 months ago

Polybase support requirement eliminates Avro. Not sure what the right answer is.

upvoted 1 times

⊟ **H_S** 4 years, 3 months ago

avro is not supported by polybase, but why not CSV

upvoted 1 times

⊟ **H_S** 4 years, 3 months ago

https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-define-outputs

it's PARKET

upvoted 2 times

⊟ **kz_data** 4 years, 3 months ago

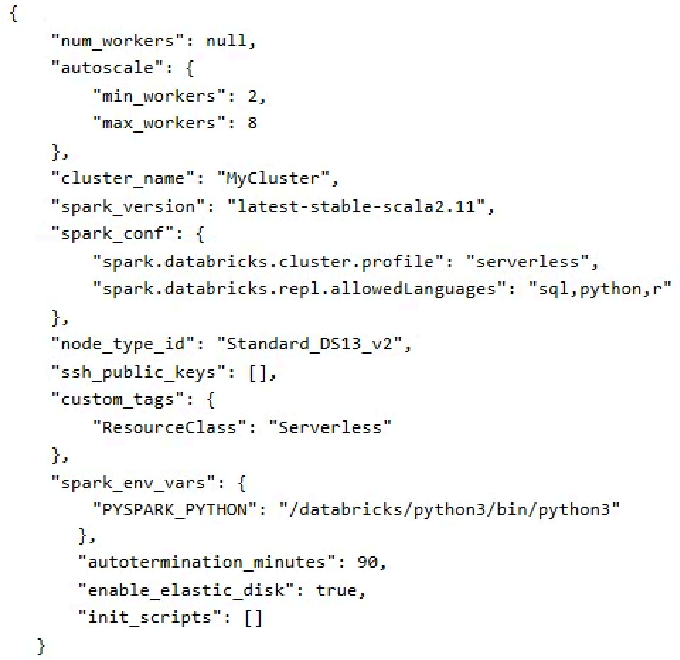
I think Parquet is the right Answer

upvoted 1 times

Question #11 Topic 2

HOTSPOT -

The following code segment is used to create an Azure Databricks cluster.

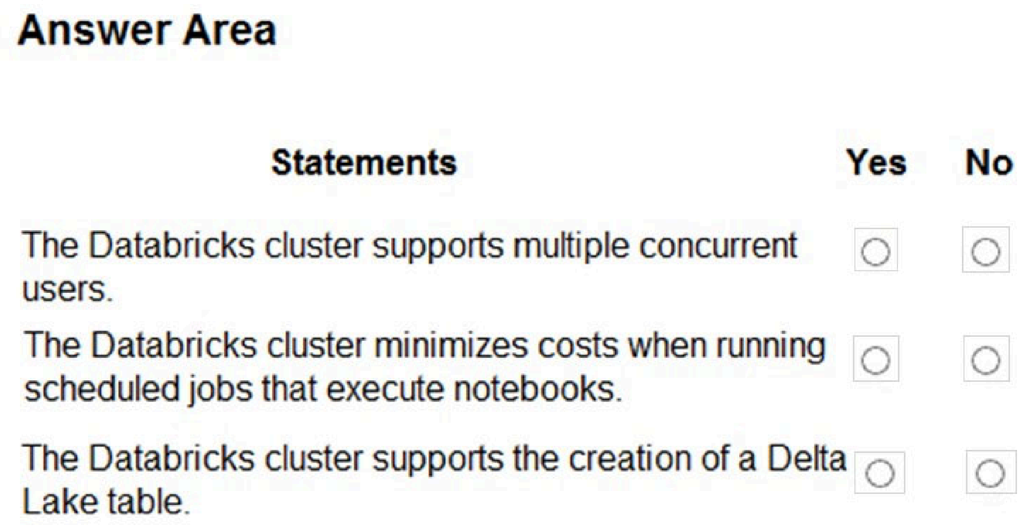


```
{
    "num_workers": null,
    "autoscale": {
        "min_workers": 2,
        "max_workers": 8
    },
    "cluster_name": "MyCluster",
    "spark_version": "latest-stable-scala2.11",
    "spark_conf": {
        "spark.databricks.cluster.profile": "serverless",
        "spark.databricks.repl.allowedLanguages": "sql,python,r"
    },
    "node_type_id": "Standard_DS13_v2",
    "ssh_public_keys": [],
    "custom_tags": {
        "ResourceClass": "Serverless"
    },
    "spark_env_vars": {
        "PYSPARK_PYTHON": "/databricks/python3/bin/python3"
      },
      "autotermination_minutes": 90,
      "enable_elastic_disk": true,
      "init_scripts": []
  }
```

For each of the following statements, select Yes if the statement is true. Otherwise, select No.

NOTE: Each correct selection is worth one point.

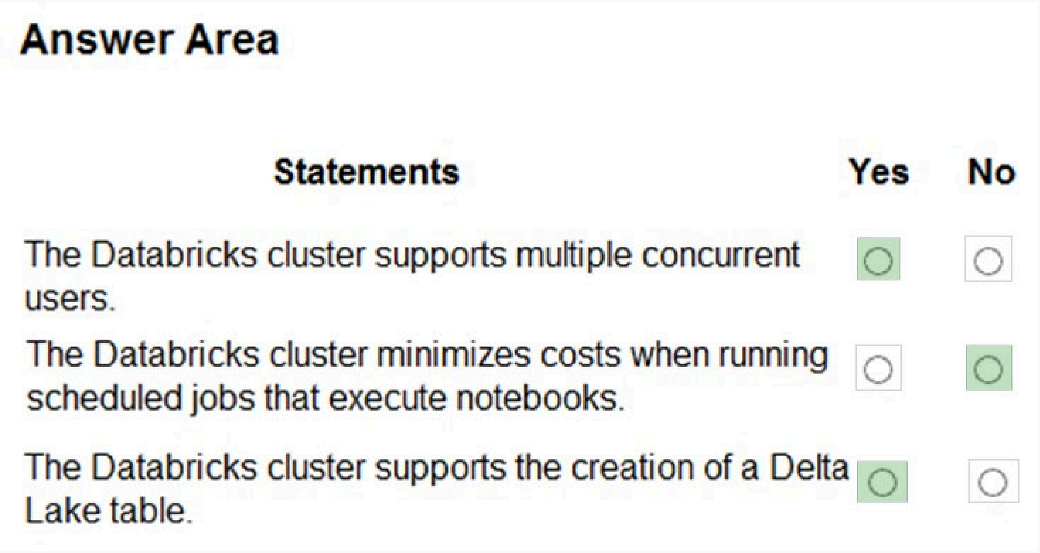
Hot Area:



**Answer Area**

| Statements | Yes | No |
|---|---|---|
| The Databricks cluster supports multiple concurrent users. | ○ | ○ |
| The Databricks cluster minimizes costs when running scheduled jobs that execute notebooks. | ○ | ○ |
| The Databricks cluster supports the creation of a Delta Lake table. | ○ | ○ |

Suggested Answer:



**Answer Area**

| Statements | Yes | No |
|---|---|---|
| The Databricks cluster supports multiple concurrent users. | [✓ ○] | ○ |
| The Databricks cluster minimizes costs when running scheduled jobs that execute notebooks. | ○ | [✓ ○] |
| The Databricks cluster supports the creation of a Delta Lake table. | [✓ ○] | ○ |

Box 1: Yes -

Box 2: No -
autotermination_minutes: Automatically terminates the cluster after it is inactive for this time in minutes. If not set, this cluster will not be automatically terminated.
If specified, the threshold must be between 10 and 10000 minutes. You can also set this value to 0 to explicitly disable automatic termination.

Box 3: Yes -
References:
https://docs.databricks.com/dev-tools/api/latest/clusters.html

**rmk4ever** Highly Voted 4 years, 9 months ago

1. Yes
A cluster mode of 'High Concurrency' is selected, unlike all the others which are 'Standard'. This results in a worker type of Standard_DS13_v2.
ref: https://adatis.co.uk/databricks-cluster-sizing/

2. NO
recommended: New Job Cluster.
When you run a job on a new cluster, the job is treated as a data engineering (job) workload subject to the job workload pricing. When you run a job on an existing cluster, the job is treated as a data analytics (all-purpose) workload subject to all-purpose workload pricing.
ref: https://docs.microsoft.com/en-us/azure/databricks/jobs
Scheduled batch workload- Launch new cluster via job
ref: https://docs.databricks.com/administration-guide/capacity-planning/cmbp.html#plan-capacity-and-control-cost

3.YES
Delta Lake on Databricks allows you to configure Delta Lake based on your workload patterns.
ref: https://docs.databricks.com/delta/index.html
upvoted 30 times

**cadio30** 4 years, 1 month ago

This explanation is entirely correct.

the first item is referencing 'high concurrency' and one could check this while creating an interactive cluster.

second item, a new job cluster should be created for job purposes as the existing all purpose cluster has different pricing. refer to the url provided at the bottom

lastly, delta lake is configurable in the mentioned cluster version

Reference: https://docs.microsoft.com/en-us/azure/databricks/jobs#cluster-config-tips
upvoted 4 times

**cadio30** 4 years, 1 month ago

btw, first item hint is when you see 'serverless' it automatically indicates the 'high concurrency' cluster mode
upvoted 6 times

**Leonido** Highly Voted 5 years, 2 months ago

My take on it:
Yes to multiple users - fits to support high concurrency since no scala support
Yes to efficiency - autostop and autoscale
Yes to the delta store - elastic disk (not 100% sure about that)
upvoted 17 times

**knightkkd** 4 years, 8 months ago

Auto termination is not configured for high concurrency clusters. so this cluster does not support high concurrency. So the answer should be
No
Yes
No
refer
https://docs.databricks.com/clusters/clusters-manage.html#automatic-termination

upvoted 2 times

**D_Duke** 4 years, 8 months ago

Auto termination is not configured for high concurrency clusters BY DEFAULT, yet you can still enable and configure it.

upvoted 5 times

**awitick** 4 years, 5 months ago

exactly

upvoted 1 times

**karma_wins** Most Recent 4 years, 2 months ago

it seems serverless corresponds to "high concurrency" as per this blogpost - https://databricks.com/blog/2017/06/07/databricks-serverless-next-generation-resource-management-for-apache-spark.html

upvoted 3 times

**sdas1** 4 years, 5 months ago

The answer is correct. I am able to create a High Concurrency cluster as per given json config.

upvoted 2 times

**sdas1** 4 years, 5 months ago

Cluster Mode - High Concurrency
Databricks Runtime Version
7.4 (includes Apache Spark 3.0.1, Scala 2.12)
NewThis Runtime version supports only Python 3.
Autopilot Options

Enable autoscaling

Terminate after
120
minutes of inactivity
Worker Type
Standard_DS13_v2
56.0 GB Memory, 8 Cores, 2 DBU
Min Workers
2
Max Workers
8
Driver Type
Standard_DS13_v2
56.0 GB Memory, 8 Cores, 2 DBU

upvoted 2 times

**sdas1** 4 years, 5 months ago

```
{
"autoscale": {
"min_workers": 2,
"max_workers": 8
},
"cluster_name": "cluster2",
"spark_version": "7.4.x-scala2.12",
"spark_conf": {
"spark.databricks.repl.allowedLanguages": "sql,python,r",
"spark.databricks.cluster.profile": "serverless"
},
"node_type_id": "Standard_DS13_v2",
"driver_node_type_id": "Standard_DS13_v2",
"ssh_public_keys": [],
"custom_tags": {
"ResourceClass": "Serverless"
},
"spark_env_vars": {
"PYSPARK_PYTHON": "/databricks/python3/bin/python3"
},
"autotermination_minutes": 120,
"enable_elastic_disk": true,
"cluster_source": "UI",
"init_scripts": [],
"cluster_id": "0116-203628-tins636"
}
```

upvoted 1 times

**sdas1** 4 years, 4 months ago

As per below link, High Concurrency clusters are configured to not terminate automatically. But while configuring High Concurrency, I am able to set the autotermination_minutes=120

https://docs.microsoft.com/en-us/azure/databricks/clusters/configure

upvoted 2 times

**zarga** 4 years, 5 months ago

1. YES
2. NO (use job custer to reduce cost rather than high concurency)
3. NO (we can use Delta lake starting from spark 2.4.2 based on scala 2.12.x. In this example the cluster definition is based on scala 2.11)

upvoted 4 times

**syu31svc** 4 years, 6 months ago

allowed languages are R SQL and Python -> High concurrency cluster
autoscaling is enabled as seen by min and max nodes -> minimise cost definitely
no CREATE TABLE syntax -> no Delta Lake table
Yes Yes No

upvoted 5 times

**lingjun** 4 years, 7 months ago

1. High Concurrency "Yes" because of following config:

```
"spark_conf": {
"spark.databricks.cluster.profile": "serverless",
"spark.databricks.repl.allowedLanguages": "sql,python,r"
},
```

2. minimise cost "No", because there is no auto scale config as below:

```
"autoscale": {
"min_workers": 2,
"max_workers": 8
},
```

upvoted 1 times

**lingjun** 4 years, 7 months ago

sorry, ignore the second point.

upvoted 1 times

**Yaswant** 4 years, 10 months ago

I think for part 2 of question "NO" is the right answer. Let's say we have three scheduled jobs with a difference of 180 minutes each that had to be run throughout the day. Since we have set the auto-termination to 90 minutes the cluster after executing the first schedule job remains active for 90 minutes so we'll have to pay for it. Which in turn doesn't minimize cost.

upvoted 1 times

**passnow** 4 years, 11 months ago

Data Lakes Support All Data Types
A data lake holds big data from many sources in a raw, granular format. It can store structured, semi-structured, or unstructured data, which means data can be kept in a more flexible format so we can transform it when we're ready to use . I stick with the default answer

upvoted 2 times

**shaktiprasad88** 4 years, 11 months ago

I think Answer is
Yes

No

No

The given Configuration is for Interactive Cluster -(My Sample Interactive Cluster with Delta Enabled)

```
{
"autoscale": {
"min_workers": 2,
"max_workers": 8
},
"cluster_name": "dev_work",
"spark_version": "6.6.x-scala2.11",
"spark_conf": {
"spark.databricks.delta.preview.enabled": "true"
},
"node_type_id": "Standard_DS3_v2",
"driver_node_type_id": "Standard_DS3_v2",
"ssh_public_keys": [],
"custom_tags": {},
"spark_env_vars": {},
"autotermination_minutes": 120,
"enable_elastic_disk": true,
"cluster_source": "UI",
"init_scripts": [],
"cluster_id": "0529-111838-patch496"
}
```

upvoted 3 times

**brcdbrcd** 4 years, 6 months ago

But it says: The Databricks cluster supports the creation of a Delta Lake table.

It is a spark cluster and it "supports" if it is needed. So I would say Yes.

upvoted 1 times

**dip17** 4 years, 11 months ago

High Concurrency does not support Auto termination; Auto-scaling minimizes the cost. So, No, Yes, Yes

upvoted 4 times

**alexvno** 4 years, 11 months ago

First - True

Optimized to run concurrent SQL, Phyton and R workloads" Doesn't support Scala. Previously known as SERVERLESS

upvoted 1 times

**AhmedReda** 5 years ago

This link shows that standard for single user, so i think High concurrency clusters for concurrency : https://docs.microsoft.com/en-us/azure/databricks/clusters/configure

Standard clusters

---------------------------

Standard clusters are recommended for a single user. Standard can run workloads developed in any language: Python, R, Scala, and SQL.

1) No

2) Yes :autoscale enabled and auto-termination was decreased from 120 default to 90

3) Yes

upvoted 6 times

**essdeecee** 4 years, 8 months ago

Standard_DS13_v2 is a High Concurrency Cluster Mode, if I select High Concurrency the Worker Type defaults to Standard_DS13_v2

upvoted 2 times

**Abhilvs** 5 years ago

Yes - Standard_DS13_V2 is cluster mode for High concurrency

No- It's an interactive cluster

Yes - I'm not sure, it seems like it is default setting when SQL API is chosen.

upvoted 2 times

**Nehuuu** 5 years, 3 months ago

In part 2 of the question, I have a confusion, in the datbricks config, the auto termination is set to 90 mins, and hence there is a provision of automatically getting the cluster down and minimizing cost. Had it been 0, it would to be auto termination disabled.

Any thoughtS?

upvoted 2 times

**avestabrzn** 5 years, 3 months ago

I think it talks about running a job on a job cluster instead of an interactive cluster. Not sure..

upvoted 3 times

**Yuri1101** 5 years, 2 months ago

I think part 2 should be yes

upvoted 4 times

**Mathster** 5 years, 1 month ago

To minimize the cost, it shoud be set to the lower value = 10. Since it is set to 90, it means the cluster can run for nothing during the next 90 minutes after the last schedule job which is not cost-efficient so the answer "NO" is correct for this one.
YES/NO/YES seams to be the correct answer.

upvoted 21 times

**andreeavi** 4 years, 6 months ago

High Concurrency clusters are configured to not terminate automatically. https://docs.microsoft.com/en-us/azure/databricks/clusters/configure

upvoted 1 times

**andreeavi** 4 years, 6 months ago

ignore it. it's not set by default

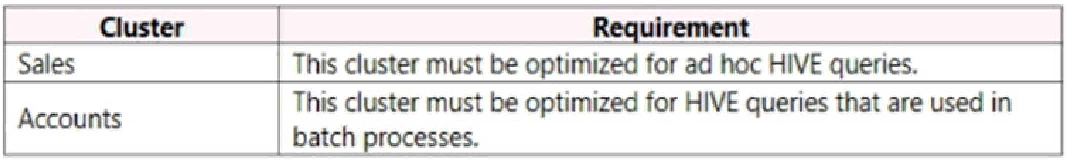
upvoted 1 times

Question #12 Topic 2

HOTSPOT -

A company stores large datasets in Azure, including sales transactions and customer account information.

You must design a solution to analyze the data. You plan to create the following HDInsight clusters:



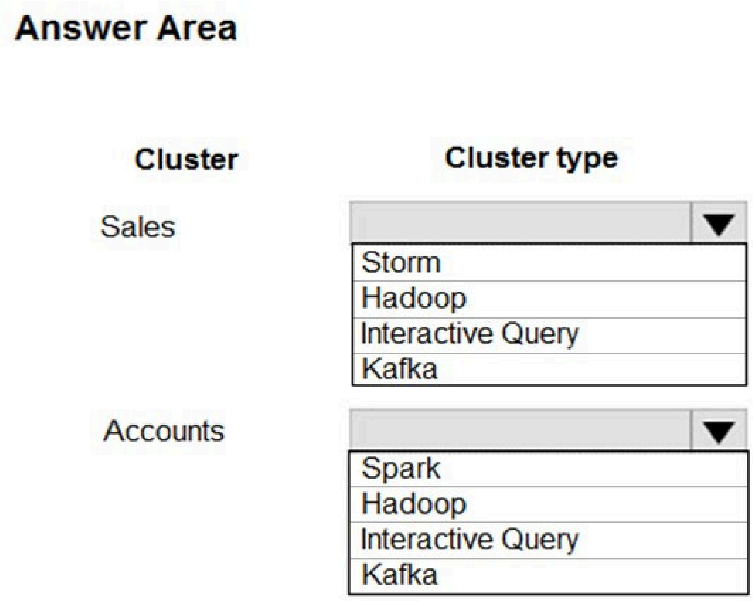
| Cluster | Requirement |
|---|---|
| Sales | This cluster must be optimized for ad hoc HIVE queries. |
| Accounts | This cluster must be optimized for HIVE queries that are used in batch processes. |

You need to ensure that the clusters support the query requirements.

Which cluster types should you recommend? To answer, select the appropriate configuration in the answer area.

NOTE: Each correct selection is worth one point.

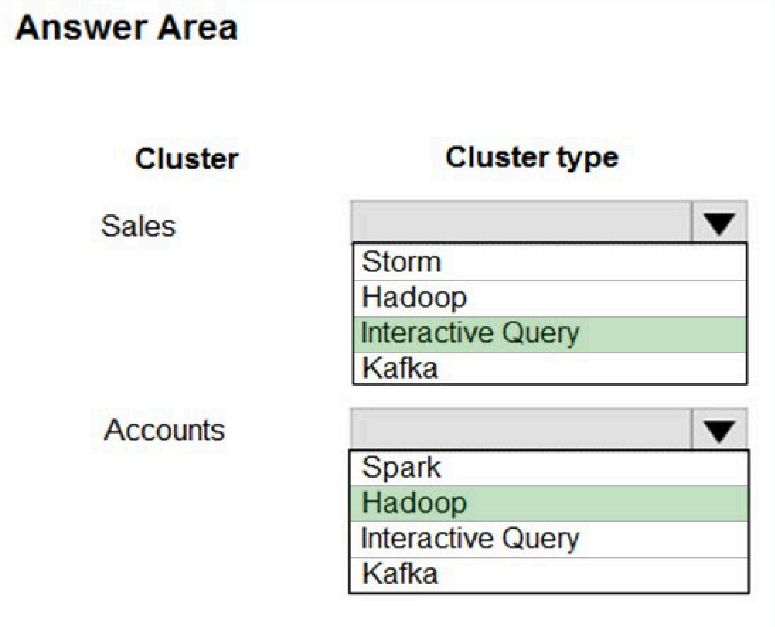
Hot Area:



**Answer Area**

| Cluster | Cluster type |
|---|---|
| Sales | Storm<br>Hadoop<br>Interactive Query<br>Kafka |
| Accounts | Spark<br>Hadoop<br>Interactive Query<br>Kafka |

**Suggested Answer:**



**Answer Area**

| Cluster | Cluster type |
|---|---|
| Sales | Storm<br>Hadoop<br>[✓ Interactive Query]<br>Kafka |
| Accounts | Spark<br>[✓ Hadoop]<br>Interactive Query<br>Kafka |

Box 1: Interactive Query -

Choose Interactive Query cluster type to optimize for ad hoc, interactive queries.

Box 2: Hadoop -

Choose Apache Hadoop cluster type to optimize for Hive queries used as a batch process.

Note: In Azure HDInsight, there are several cluster types and technologies that can run Apache Hive queries. When you create your HDInsight cluster, choose the appropriate cluster type to help optimize performance for your workload needs.

For example, choose Interactive Query cluster type to optimize for ad hoc, interactive queries. Choose Apache Hadoop cluster type to optimize

for Hive queries used as a batch process. Spark and HBase cluster types can also run Hive queries.
Reference:
https://docs.microsoft.com/bs-latn-ba/azure/hdinsight/hdinsight-hadoop-optimize-hive-query?toc=%2Fko-kr%2Fazure%2Fhdinsight%2Finteractive-query%2FTOC.json&bc=%2Fbs-latn-ba%2Fazure%2Fbread%2Ftoc.json

**M0e** Highly Voted 👍 4 years, 8 months ago

HDInsight is not covered in the exam any more.

upvoted 10 times

**REZ82** Highly Voted 👍 5 years ago

Cluster types in HDInsight
https://docs.microsoft.com/bs-latn-ba/azure/hdinsight/hdinsight-overview#cluster-types-in-hdinsight

upvoted 6 times

**H_S** Most Recent ⏲ 4 years, 3 months ago

NOT ANY MORE IN THE DP-201

upvoted 1 times

**Deepu1987** 4 years, 4 months ago

The given soln is correct
Interactive qury - In-memory caching for interactive and faster Hive queries
hadoop - A framework that uses HDFS, YARN resource management, and a simple MapReduce programming model to process and analyze batch data in parallel.
We're receiving qns on this topic until dp-201 is removed

upvoted 2 times

**chaoxes** 4 years, 6 months ago

Answer is correct.

Sales: Interactive Queries
In-memory caching for interactive and faster Hive queries.

Accounts: Hadoop
A framework that uses HDFS, YARN resource management, and a simple MapReduce programming model to process and analyze batch data in parallel.

Source: https://docs.microsoft.com/bs-latn-ba/azure/hdinsight/hdinsight-overview#cluster-types-in-hdinsight

upvoted 6 times

**syu31svc** 4 years, 6 months ago

Answer is correct

upvoted 3 times

Question #13 *Topic 2*

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.

After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You have streaming data that is received by Azure Event Hubs and stored in Azure Blob storage. The data contains social media posts that relate to a keyword of

Contoso.

You need to count how many times the Contoso keyword and a keyword of Litware appear in the same post every 30 seconds. The data must be available to

Microsoft Power BI in near real-time.

Solution: You use Azure Data Factory and an event trigger to detect when new blobs are created. You use mapping data flows in Azure Data Factory to aggregate and filter the data, and then send the data to an Azure SQL database. You consume the data in Power BI by using DirectQuery mode.

Does the solution meet the goal?

A. Yes

B. No

**Suggested Answer:** *B*

**cadio30** Highly Voted 4 years, 1 month ago

Answer is NO. The propose solution to utilize ADF doesn't have real-time capability. Instead use of Azure Stream Analytics as it can count the data with the use of window function and output the data directly to the PowerBI dataset.

upvoted 11 times

**AlexD332** Highly Voted 4 years, 3 months ago

scenario: You need to count how many times .. appear in the same post every 30 seconds.

Solution: You use Azure Data Factory and an event trigger

Answer: No - you don't need an event trigger - you need a schedule trigger each 30 seconds

upvoted 8 times

**sjain91** Most Recent 4 years, 2 months ago

Answer: No - you don't need an event trigger - you need a schedule trigger each 30 seconds

upvoted 4 times

**maciejt** 4 years, 2 months ago

Mapping data flow needs few minutes to start up spark cluster, it's good for batch ETL, but not suitable for real time stream processing.

upvoted 4 times

> **karma_wins** 4 years, 2 months ago
>
> reply from maciejt seems only logical to me for "No" to this question i.e. since mapping dataflow needs few mins to spark up the cluster.
>
> upvoted 1 times

**mohowzeh** 4 years, 5 months ago

correction: in my previous post, forget the 30-word distance. But the "break a post into words" remains the reason why the proposed solution does not fully meet the requirements IMHO

upvoted 1 times

**mohowzeh** 4 years, 5 months ago

Answer is correct in my opinion. The proposed approach is missing a step where you break out the social media post into words and investigate the 30-word "distance".

upvoted 1 times

**cm19** 4 years, 5 months ago

ADF is not a suitable solution for realtime feeds.When Streaming analytics can do the job directly no need of triggers to identify new blobs.So the answer looks correct to me.

upvoted 5 times

**HPotter** 4 years, 6 months ago

Why ? There should be an explanation. Solution provided seems reasonable

upvoted 2 times

Question #14 *Topic 2*

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You have streaming data that is received by Azure Event Hubs and stored in Azure Blob storage. The data contains social media posts that relate to a keyword of
Contoso.
You need to count how many times the Contoso keyword and a keyword of Litware appear in the same post every 30 seconds. The data must be available to
Microsoft Power BI in near real-time.
Solution: You create an Azure Stream Analytics job that uses an input from Event Hubs to count the posts that have the specified keywords, and then send the data to an Azure SQL database. You consume the data in Power BI by using DirectQuery mode.
Does the solution meet the goal?

A. Yes

B. No

**Suggested Answer:** *A*
Reference:
https://docs.microsoft.com/en-us/power-bi/service-real-time-streaming https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-twitter-sentiment-analysis-trends

**vivekazure** 3 years, 5 months ago
Unless Power BI connects to SQL DB thru Import mode and schedule a automatic refresh, Data can never be available in near real-time.
upvoted 1 times

**Saravjeet** 4 years, 1 month ago
What I think is the proposed solution is correct, as we can generate the report directly through power BI by connecting it to ASA but not through direct query mode, I am not able to find the source as ASA while referring to DirectQuery. Refer the link: https://docs.microsoft.com/en-us/power-bi/connect-data/power-bi-data-sources

So we have to use sql db if we use directquery mode. Thanks.
upvoted 1 times

**cadio30** 4 years, 1 month ago
The propose solution is feasible as the data can be stored in Azure SQL DB then use of direct mode from Power BI retrieves the latest data while 'import' connectivity mode requires schedule to refresh the dataset. And as mentioned in the requirement, it states 'near real time' unless it is explicitly label as 'real time' then Azure Stream Analytics is the most suited solution.
upvoted 1 times

**Nik71** 4 years, 3 months ago
You can add Power BI as an output within Azure Stream Analytics (ASA), and then visualize those data streams in the Power BI service in real time

https://docs.microsoft.com/en-us/power-bi/connect-data/service-real-time-streaming#pushing-data-to-datasets
upvoted 3 times

**Nik71** 4 years, 3 months ago
Answer should be No why we output data from Stream to SQL db we can direct output to power BI
upvoted 3 times

**chirag1234** 4 years, 2 months ago
but they have not mentioned real-time, it is near real-time so it should be year
upvoted 3 times

## Question #15 *Topic 2*

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You have streaming data that is received by Azure Event Hubs and stored in Azure Blob storage. The data contains social media posts that relate to a keyword of
Contoso.
You need to count how many times the Contoso keyword and a keyword of Litware appear in the same post every 30 seconds. The data must be available to
Microsoft Power BI in near real-time.
Solution: You use Azure Databricks to create a Scala notebook. You use a Structured Streaming job to connect to the event hub that counts the posts that have the specified keywords, and then writes the data to a Delta table. You consume the data in Power BI by using DirectQuery mode.
Does the solution meet the goal?

A. Yes

B. No

**Suggested Answer:** *B*

**Shrikant_Kulkarni** Highly Voted 4 years, 7 months ago
this question is outdated. this should be perfectly possible.
upvoted 14 times

**cadio30** Most Recent 4 years, 1 month ago
The propose solution is feasible as the PowerBI can integrate to Azure Databricks. Unless the requirement changes to 'real time' then Azure Stream Analytics is suited service.

Reference: https://azure.microsoft.com/nl-nl/blog/structured-streaming-with-databricks-into-power-bi-cosmos-db/
upvoted 3 times

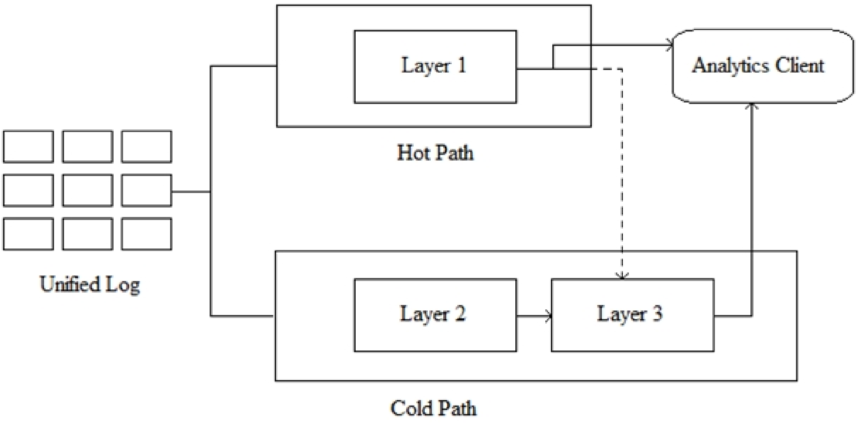
**mohowzeh** 4 years, 5 months ago
https://databricks.com/blog/2020/10/30/announcing-azure-databricks-power-bi-connector-public-preview.html
upvoted 2 times

**ACSC** 4 years, 5 months ago
So, the answer is Yes, the solution meets the goal.
upvoted 2 times

**ACSC** 4 years, 5 months ago
I mean, it is possible, but not near real-time. The answer is No.
upvoted 5 times

Question #16 *Topic 2*

You are planning a design pattern based on the Lambda architecture as shown in the exhibit.



Which Azure service should you use for the hot path?

A. Azure Synapse Analytics

B. Azure SQL Database

C. Azure Cosmos DB

D. Azure Data Catalog

**Suggested Answer:** *C*

In Azure, all of the following data stores will meet the core requirements supporting real-time processing:

✑ Apache Spark in Azure Databricks

✑ Azure Stream Analytics

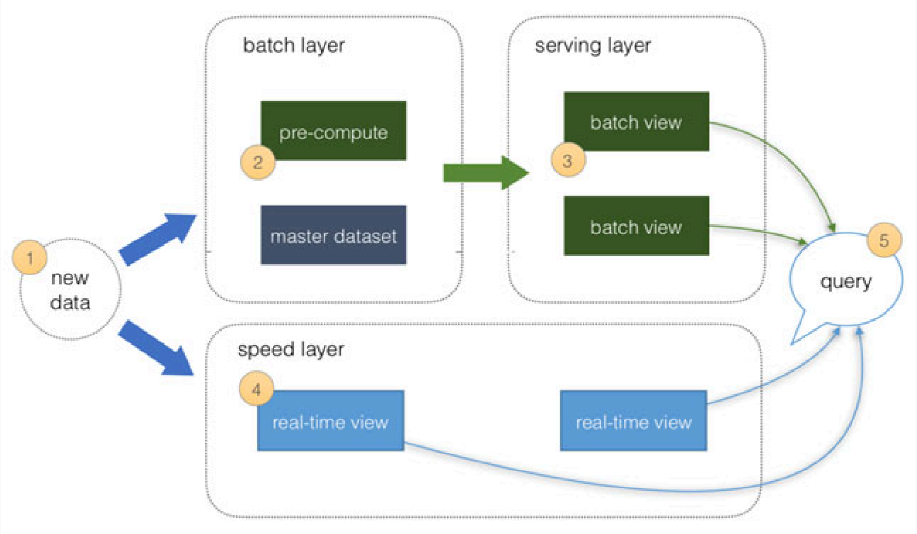
✑ HDInsight with Spark Streaming

✑ HDInsight with Storm

✑ Azure Functions

✑ Azure App Service WebJobs

Note: Lambda architectures use batch-processing, stream-processing, and a serving layer to minimize the latency involved in querying big data.



Reference:

https://azure.microsoft.com/en-us/blog/lambda-architecture-using-azure-cosmosdb-faster-performance-low-tco-low-devops/

https://docs.microsoft.com/en-us/azure/architecture/data-guide/technology-choices/stream-processing https://docs.microsoft.com/en-us/azure/cosmos-db/lambda-architecture

**H_S** Highly Voted 4 years, 3 months ago

https://azure.microsoft.com/en-us/blog/lambda-architecture-using-azure-cosmosdb-faster-performance-low-tco-low-devops/

upvoted 8 times

**maciejt** Most Recent 4 years, 2 months ago

Why not synapse? It also can use spark databricks notebooks...

upvoted 1 times

**Hrabia** 4 years, 1 month ago

synapse is cold. it's basicaly a warehouse

upvoted 8 times

## Question #17 *Topic 2*

You are designing an enterprise data warehouse in Azure Synapse Analytics. You plan to load millions of rows of data into the data warehouse each day.

You must ensure that staging tables are optimized for data loading.

You need to design the staging tables.

What type of tables should you recommend?

A. Round-robin distributed table

B. Hash-distributed table

C. Replicated table

D. External table

**Suggested Answer:** *A*

To achieve the fastest loading speed for moving data into a data warehouse table, load data into a staging table. Define the staging table as a heap and use round-robin for the distribution option.

Incorrect:

Not B: Consider that loading is usually a two-step process in which you first load to a staging table and then insert the data into a production data warehouse table. If the production table uses a hash distribution, the total time to load and insert might be faster if you define the staging table with the hash distribution.

Loading to the staging table takes longer, but the second step of inserting the rows to the production table does not incur data movement across the distributions.

Reference:

https://docs.microsoft.com/en-us/azure/sql-data-warehouse/guidance-for-loading-data

**riteshsinha18** Highly Voted 4 years, 3 months ago

correct

upvoted 8 times

**memo43** 4 years, 1 month ago

keyword: staging table

upvoted 3 times

Question #18 *Topic 2*

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You have streaming data that is received by Azure Event Hubs and stored in Azure Blob storage. The data contains social media posts that relate to a keyword of
Contoso.
You need to count how many times the Contoso keyword and a keyword of Litware appear in the same post every 30 seconds. The data must be available to
Microsoft Power BI in near real-time.
Solution: You create an Azure Stream Analytics job that uses an input from Event Hubs to count the posts that have the specified keywords, and then send the data directly to Power BI.
Does the solution meet the goal?

A. Yes

B. No

**Suggested Answer:** *B*
DirectQuery mode is required for automatic page refresh.
Reference:
https://docs.microsoft.com/en-us/power-bi/service-real-time-streaming https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-twitter-sentiment-analysis-trends

*Community vote distribution*
A (100%)

**stormraider** Highly Voted 4 years, 3 months ago
Should be yes, Stream analytics can directly send the output to Power BI
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-power-bi-dashboard
upvoted 32 times

**niwe** 4 years, 2 months ago
Link is for real-time, questions says "The data must be available to Microsoft Power BI in near real-time." So is, B-No
upvoted 2 times

**alf99** 4 years, 1 month ago
MS documentations confirms this point:

The analysis of the streaming data is performed with Stream Analytics itself. A Stream Analytics job is created that contains the Event Hubs as input streams, with a query that performs stream processing of the two event hubs to correlate the records in the two data streams. An output is defined to Cosmos DB that stores the correlated results written as JSON documents to a Cosmos DB document database.

Power BI then uses Cosmos DB as a source for a dashboard of the correlated records. You could also have Power BI point
directly from Stream Analytics; however, this data would not be persisted in a data store.
upvoted 1 times

**cadio30** 4 years, 1 month ago
The link already stated that the setup is feasible in 'REAL TIME'
upvoted 2 times

**dakku987** Most Recent 1 year, 6 months ago
**Selected Answer: A**
chat gpt says
Based on the information provided, it seems like the solution has the potential to meet the goal, but there are some details that need clarification and adjustments. Therefore, I would say it's a "Yes, with caveats." Ensure that you address the specific details mentioned in the evaluation to make the solution more robust and aligned with the goal of counting keyword occurrences and sending the data to Power BI in near real-time.
upvoted 1 times

**Ssv2030** 3 years, 9 months ago

the question says "The data must be available to Microsoft Power BI in near real-time", but ASA +Power BI is real-time combination as per:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-power-bi-dashboard

And, there is difference between real time and near real time. So, I feel the answer is 'NO'
can someone pls confirm?

upvoted 1 times

**BigMF** 4 years ago

I think the key phrase is "...then send the data directly to Power BI". Power BI is NOT a data store and does not "accept" data, it goes and gets it. Therefore, I believe the answer to be NO.

upvoted 2 times

**Mandar77** 4 years ago

You have to use power Bi in direct query mode

upvoted 2 times

**MMM777** 4 years ago

Should be YES. Another doc showing it is possible - BI can only query every 1 second - but that is near real time:
https://docs.microsoft.com/en-us/azure/stream-analytics/power-bi-output

upvoted 2 times

**Nik71** 4 years, 3 months ago

https://docs.microsoft.com/en-us/power-bi/connect-data/service-real-time-streaming#pushing-data-to-datasets
Answer is Yes

upvoted 2 times

**I** 4 years, 3 months ago

You create an Azure Stream Analytics job that uses an input from Event Hubs to count the posts that have the specified keywords, and then send the data to an Azure SQL database. You consume the data in Power BI by using DirectQuery mode.

upvoted 2 times

**I** 4 years, 3 months ago

Answer is No.

upvoted 1 times

**obz** 4 years, 3 months ago

You can aggregate and send directly to Power BI with Azure stream analytics if you want to achieve near-real time reporting. Using a intermediary SQL Database will be a waste of cost & management and won't allow you to achieve near real-time reporting. The Answer is YES.

upvoted 7 times

## Question #19 *Topic 2*

A company has an application that uses Azure SQL Database as the data store.
The application experiences a large increase in activity during the last month of each year.
You need to manually scale the Azure SQL Database instance to account for the increase in data write operations.
Which scaling method should you recommend?

A. Scale up by using elastic pools to distribute resources.

B. Scale out by sharding the data across databases.

C. Scale up by increasing the database throughput units.

**Suggested Answer:** *C*
As of now, the cost of running an Azure SQL database instance is based on the number of Database Throughput Units (DTUs) allocated for the database. When determining the number of units to allocate for the solution, a major contributing factor is to identify what processing power is needed to handle the volume of expected requests.
Running the statement to upgrade/downgrade your database takes a matter of seconds.
Incorrect Answers:
A: Elastic pools is used if there are two or more databases.
Reference:
https://www.skylinetechnologies.com/Blog/Skyline-Blog/August_2017/dynamically-scale-azure-sql-database

**Arsa** Highly Voted 4 years, 10 months ago
correct answer
upvoted 16 times

**Deepu1987** Most Recent 4 years, 4 months ago
When we say increase it' s Scale up by increasing the DTUs of Azure SQL DB
upvoted 1 times

**chaoxes** 4 years, 6 months ago
C. Scale up by increasing the database throughput units.
upvoted 2 times

**syu31svc** 4 years, 6 months ago
Single database so C is the answer
upvoted 2 times

## Question #20 *Topic 2*

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You have an Azure Data Lake Storage account that contains a staging zone.
You need to design a daily process to ingest incremental data from the staging zone, transform the data by executing an R script, and then insert the transformed data into a data warehouse in Azure Synapse Analytics.
Solution: You use an Azure Data Factory schedule trigger to execute a pipeline that copies the data to a staging table in the data warehouse, and then uses a stored procedure to execute the R script.
Does this meet the goal?

A. Yes

B. No

**Suggested Answer:** *A*
If you need to transform data in a way that is not supported by Data Factory, you can create a custom activity with your own data processing logic and use the activity in the pipeline. You can create a custom activity to run R scripts on your HDInsight cluster with R installed.
Reference:
https://docs.microsoft.com/en-US/azure/data-factory/transform-data

**Nieswurz** Highly Voted 4 years, 10 months ago
The proposed solution seems to let the R function do the loading into Synapse. The answer then should be 'no', but more likely the description seems again to be incomplete.
upvoted 16 times

**Pairon** 4 years, 3 months ago
I agree with you. The insert operation into the DWH should be come after the R script process.
upvoted 2 times

**cadio30** 4 years, 1 month ago
Agree with the statement and as of this year, the Azure Synapse doesn't support R language unless it is executed against Azure SQL Manage Instance
upvoted 5 times

**azurrematt123** Most Recent 4 years ago
Looks like executing R is possible(sp_execute_external_script), please review the link.
https://docs.microsoft.com/en-us/sql/relational-databases/system-stored-procedures/sp-execute-external-script-transact-sql?view=sql-server-ver15
upvoted 2 times

**azurrematt123** 4 years ago
My bad it only applies to SQL Server 2016 & Azure SQL Managed Instance, Moderator please dont post this.
upvoted 1 times

**tes** 4 years ago
there is no moderator human
upvoted 1 times

**dbdev** 4 years, 1 month ago
https://www.mssqltips.com/sqlservertip/6622/stored-procedure-in-sql-server-with-r-code/
The R function can be used inside stored procedure activity, so answer makes sense to me.
upvoted 1 times

**dbdev** 4 years, 1 month ago
https://docs.microsoft.com/en-US/azure/data-factory/transform-data#custom-activity
upvoted 1 times

**KRV** 4 years ago
yes R function can be used within the Stored procedure for SQL Server or Azure SQL Managed instance , however the statement states that the data is loaded using SQL Synapse which does not support R at this time

upvoted 1 times

**I** 4 years, 3 months ago

The answer should be No. Because MS always supply options for users so it won't engage in one specific programming language such as R.

upvoted 1 times

**Madhumita88** 4 years, 3 months ago

I am also not cleared with this answer

upvoted 1 times

**S3** 4 years, 5 months ago

I am still not clear as to what is the correct answer

upvoted 1 times

**syu31svc** 4 years, 6 months ago

Solution proposed is on data pipeline and orchestration so I would say yes

upvoted 1 times

**AJMorgan591** 4 years, 9 months ago

Should use a tumbling window trigger in ADF for incremental loading.

https://docs.microsoft.com/en-us/azure/data-factory/solution-template-copy-new-files-lastmodifieddate

upvoted 2 times

> **sandGrain** 4 years, 7 months ago
>
> You can do incremental load using schedule trigger. Does not have to be Tumbling window
>
> upvoted 1 times

**Bob123456** 4 years, 10 months ago

can we run a stored procedure to execute the R script ?? I don't think so.

upvoted 2 times

> **VMLearn** 4 years, 6 months ago
>
> possible
>
> upvoted 1 times

> **Bob123456** 4 years, 10 months ago
>
> https://docs.microsoft.com/en-us/sql/machine-learning/tutorials/quickstart-r-create-script?view=sql-server-ver15
>
> i believe this answers the question . Answer should be 'yes'
>
> upvoted 6 times
>
> > **pablocg** 4 years, 7 months ago
> >
> > That is for sql server and managed instances and the question is about Azure Synapse Analytics
> >
> > upvoted 4 times

**Nieswurz** 4 years, 10 months ago

The explanation of the answer contains R on an HDInsight-Cluster. This kind of solution is stated to be incorrect in another questions explanation - in favor of an Azure function.

upvoted 2 times

## Question #21 *Topic 2*

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You have an Azure Data Lake Storage account that contains a staging zone.
You need to design a daily process to ingest incremental data from the staging zone, transform the data by executing an R script, and then insert the transformed data into a data warehouse in Azure Synapse Analytics.
Solution: You schedule an Azure Databricks job that executes an R notebook, and then inserts the data into the data warehouse.
Does this meet the goal?

A. Yes

B. No

**Suggested Answer:** *B*
You should use an Azure Data Factory, not an Azure Databricks job.
Reference:
https://docs.microsoft.com/en-US/azure/data-factory/transform-data

**avix** Highly Voted 4 years, 9 months ago
But I can do with this too!
upvoted 20 times

**andreeavi** 4 years, 5 months ago
it is possible, but first you need to ingest data from staging source
upvoted 3 times

**Kalo** 4 years, 4 months ago
with a mount in DBFS, we can ingest data from ADLS
upvoted 4 times

**Shrikant_Kulkarni** Highly Voted 4 years, 7 months ago
answer should be yes.
upvoted 17 times

**cadio30** Most Recent 4 years, 1 month ago
This requirement is possible with the use R script in Azure Databricks job. Therefore, answer should be 'Yes'
upvoted 2 times

**mohowzeh** 4 years, 5 months ago
A scheduled daily Databricks job does the trick. Data Factory isn't the only tool that can bring data from one place to another... Answer should be yes.
upvoted 4 times

**Psycho360** 4 years, 6 months ago
Who is gonna stop me from using Databricks. There seems to be no technical limitation in this approach
upvoted 3 times

**Akva** 4 years, 7 months ago
I think it should be YES.

https://docs.microsoft.com/en-us/azure/databricks/scenarios/databricks-extract-load-sql-data-warehouse
upvoted 9 times

Question #22 *Topic 2*

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You have an Azure Data Lake Storage account that contains a staging zone.
You need to design a daily process to ingest incremental data from the staging zone, transform the data by executing an R script, and then insert the transformed data into a data warehouse in Azure Synapse Analytics.
Solution: You use an Azure Data Factory schedule trigger to execute a pipeline that executes an Azure Databricks notebook, and then inserts the data into the data warehouse.
Does this meet the goal?

A. Yes

B. No

**Suggested Answer:** *B*
Use a stored procedure, not an Azure Databricks notebook to invoke the R script.
Reference:
https://docs.microsoft.com/en-US/azure/data-factory/transform-data

**Nieswurz** Highly Voted 4 years, 10 months ago
This should be the correct answer.
upvoted 27 times

**andreeavi** 4 years, 5 months ago
first step is to ingest data..
upvoted 1 times

**maynard13x8** 4 years, 2 months ago
I think notebooks are only interactive. It should be a job cluster. Any opinions?
upvoted 2 times

**Bhagya123456** 3 years, 10 months ago
Now your comment is ambiguous. Do you mean correct answer provided in that case 'NO' is answer or the Solution provided is correct and it will do the Job, in this case 'Yes' will be the answer...
upvoted 4 times

**bakamon** Most Recent 2 years, 1 month ago
Yes, this solution meets the goal. You can use an Azure Data Factory schedule trigger to execute a pipeline that copies the data to a staging table in the data warehouse, and then uses a stored procedure to execute the R script. This will allow you to ingest incremental data from the staging zone, transform the data by executing an R script, and then insert the transformed data into a data warehouse in Azure Synapse Analytics on a daily basis.
upvoted 1 times

**Ssv2030** 3 years, 9 months ago
The answer should be NO because:
1. we can't assume that the Azure Databricks notebook will execute/run the transform R script, it is not mentioned that Azure Databricks notebook will run the R script
2. for incremental loads in ADF, I think a tumbling trigger should be used.
can someone pls confirm?
upvoted 1 times

**MMM777** 4 years ago
Answer should be YES: ADF can trigger a Databricks notebook (not required to be user-driven):

https://docs.microsoft.com/en-us/azure/data-factory/transform-data-using-databricks-notebook
upvoted 4 times

**cadio30** 4 years, 1 month ago
The answer is Yes. R script is executed in the azure databricks notebook and once the transformation is completed then the mount the Azure Synapse to load the data.

Reference: https://docs.microsoft.com/en-us/azure/databricks/scenarios/databricks-extract-load-sql-data-warehouse

upvoted 1 times

**AJMorgan591** 4 years, 9 months ago

Should use a tumbling window trigger in ADF for incremental loading.

https://docs.microsoft.com/en-us/azure/data-factory/solution-template-copy-new-files-lastmodifieddate

upvoted 2 times

**BungyTex** 4 years, 6 months ago

Don't have to, can just use a regular schedule no problem.

upvoted 1 times

**avix** 4 years, 9 months ago

I'm surprised as I ran R in Azure Databrick

upvoted 2 times

**Nieswurz** 4 years, 10 months ago

The solution template mentioned by Bob123456 does not fit, as --- per description --- the R script is to be run when the data is still located in the data lake. After the R based transformation, the result is to be loaded to the DWH. This type of processing would need polybase for accessing the data lake, which is not mentioned here.

upvoted 3 times

**apandey** 4 years, 9 months ago

Databricks notebook can use mount to access data lake. Notebook is correct answer

upvoted 2 times

**Bob123456** 4 years, 10 months ago

this is incorrect

https://docs.microsoft.com/en-us/sql/machine-learning/tutorials/quickstart-r-create-script?view=sql-server-ver15

upvoted 1 times

## Question #23 *Topic 2*

A company purchases IoT devices to monitor manufacturing machinery. The company uses an Azure IoT Hub to communicate with the IoT devices.

The company must be able to monitor the devices in real-time.

You need to design the solution.

What should you recommend?

A. Azure Data Factory instance using Azure Portal

B. Azure Analysis Services using Microsoft Visual Studio

C. Azure Stream Analytics Edge application using Microsoft Visual Studio

D. Azure Data Factory instance using Microsoft Visual Studio

**Suggested Answer:** *C*

Azure Stream Analytics (ASA) on IoT Edge empowers developers to deploy near-real-time analytical intelligence closer to IoT devices so that they can unlock the full value of device-generated data.

You can use Visual Studio plugin to create an ASA Edge job.

Reference:

https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-edge

**satyamkishoresingh** 3 years, 10 months ago

It's a duplicate Question

upvoted 1 times

**mosheshito** 4 years ago

Correct

upvoted 1 times

Question #24 *Topic 2*

A company has a real-time data analysis solution that is hosted on Microsoft Azure. The solution uses Azure Event Hub to ingest data and an Azure Stream
Analytics cloud job to analyze the data. The cloud job is configured to use 120 Streaming Units (SU).
You need to optimize performance for the Azure Stream Analytics job.
Which two actions should you perform? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.

A. Implement event ordering

B. Scale the SU count for the job up

C. Implement Azure Stream Analytics user-defined functions (UDF)

D. Scale the SU count for the job down

E. Implement query parallelization by partitioning the data output

F. Implement query parallelization by partitioning the data output

**Suggested Answer:** *BF*
Scale out the query by allowing the system to process each input partition separately.
F: A Stream Analytics job definition includes inputs, a query, and output. Inputs are where the job reads the data stream from.
Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-parallelization

**NabilR** Highly Voted 4 years, 7 months ago
Answer is correct. F should specify "Input"
upvoted 25 times

**Nik71** 4 years, 3 months ago
yep Input need to be partitioned not output
upvoted 2 times

**pablocg** Highly Voted 4 years, 7 months ago
I have seen this question and answer before but I don't think it is correct as it specifically mentions optimize performance.

In Microsoft's documentation, it specifies partitioning input and output to leverage parallelization, so I think E and F should be the answer.

https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-parallelization
upvoted 6 times

**tes** 4 years ago
"E and F should be the answer"? What is the different between E and F when it is copy pasted?
upvoted 1 times

**anamaster** 4 years, 2 months ago
Isn't the input already partitioned?
upvoted 1 times

**Arusham** Most Recent 4 years ago
This question came in my DP 200 exam
I gave on 24th June 2021
upvoted 2 times

**Qrm_1972** 4 years ago
The correct answer: B & F

In F: you can change the latest word ( output >>>> input )!
A. Implement event ordering
B. Scale the SU count for the job up
C. Implement Azure Stream Analytics user-defined functions (UDF)

D. Scale the SU count for the job down

E. Implement query parallelization by partitioning the data output

F. Implement query parallelization by partitioning the data input

upvoted 5 times

**cadio30** 4 years, 1 month ago

Answer are Scale up the Streaming Units then include partition on 'Input' (done in the query)

upvoted 1 times

**I** 4 years, 3 months ago

E and F are exactly same.

upvoted 2 times

**rajat009** 4 years, 6 months ago

i saw this question in dp-200, but option E,F wasnt there either

upvoted 2 times

**anamaster** 4 years, 2 months ago

I had exactly the same question on dp-200

upvoted 3 times

**rajat009** 4 years, 6 months ago

DP-200 question not 201

upvoted 4 times

**syu31svc** 4 years, 6 months ago

Scaling the SU count is correct

partition the output not input

so B is correct

Either E or F is right since there is a typo of output twice

upvoted 2 times

Question #25 *Topic 2*

You manage a process that performs analysis of daily web traffic logs on an HDInsight cluster. Each of the 250 web servers generates approximately
10megabytes (MB) of log data each day. All log data is stored in a single folder in Microsoft Azure Data Lake Storage Gen 2.
You need to improve the performance of the process.
Which two changes should you make? Each correct answer presents a complete solution.
NOTE: Each correct selection is worth one point.

A. Combine the daily log files for all servers into one file

B. Increase the value of the mapreduce.map.memory parameter

C. Move the log files into folders so that each day's logs are in their own folder

D. Increase the number of worker nodes

E. Increase the value of the hive.tez.container.size parameter

**Suggested Answer:** *AC*
A: Typically, analytics engines such as HDInsight and Azure Data Lake Analytics has a per-five overhead. If you store your data as many small files, this can negatively affect performance. In general, organize your data into larger sized files for better performance (256MB to 100GB in size). Some engines and applications might have trouble efficiently processing files that are greater than 100GB in size.
C: For Hive workloads, partition pruning of time-series data can help some queries read only a subset of the data which improves performance.
Those pipelines that ingest time-series data, often place their files with a very structured naming for files and folders. Below is a very common example we see for data is structured by date:
\DataSet\YYYY\MM\DD\datafile_YYYY_MM_DD.tsv
Notice that the datetime information appears both as folders and in the filename.
Reference:
https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-performance-tuning-guidance

**DannyDaj** Highly Voted 👍 4 years, 5 months ago
This question is also in the DP-200 exam. Same with the previous question.
upvoted 13 times

**azurrematt123** Most Recent 🕓 4 years ago
Agreed this is a question from DP-200 but wondering if this is part of DP-201 as well?
upvoted 1 times

## Question #26 *Topic 2*

A company purchases IoT devices to monitor manufacturing machinery. The company uses an IoT appliance to communicate with the IoT devices.

The company must be able to monitor the devices in real-time.

You need to design the solution.

What should you recommend?

A. Azure Analysis Services using Azure Portal

B. Azure Analysis Services using Microsoft Visual Studio

C. Azure Stream Analytics Edge application using Microsoft Visual Studio

D. Azure Data Factory instance using Microsoft Visual Studio

**Suggested Answer:** *C*

Stream Analytics is a cost-effective event processing engine that helps uncover real-time insights from devices, sensors, infrastructure, applications and data quickly and easily.

Visual Studio 2019 and Visual Studio 2017 support Stream Analytics Tools.

Note: You can also monitor and manage Stream Analytics resources with Azure PowerShell cmdlets and powershell scripting that execute basic Stream Analytics tasks.

Reference:

https://cloudblogs.microsoft.com/sqlserver/2014/10/29/microsoft-adds-iot-streaming-analytics-data-production-and-workflow-services-to-azure/ https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-tools-for-visual-studio-install

⊟ **ZodiaC** Highly Voted 4 years ago

3 times DAMN

upvoted 6 times

⊟ **Harshit1905** Most Recent 2 years, 5 months ago

Now I cannot miss this question in exam . Lol

upvoted 2 times

⊟ **IAMKPR** 4 years, 1 month ago

It's a repeated question. Already appeared once before.

upvoted 1 times

⊟ **dbdev** 4 years, 1 month ago

even twice

upvoted 7 times

Question #27 *Topic 2*

You are designing an anomaly detection solution for streaming data from an Azure IoT hub. The solution must meet the following requirements:

✑ Send the output to Azure Synapse.

✑ Identify spikes and dips in time series data.

✑ Minimize development and configuration effort

Which should you include in the solution?

A. Azure Databricks

B. Azure Stream Analytics

C. Azure SQL Database

**Suggested Answer:** *B*

You can identify anomalies by routing data via IoT Hub to a built-in ML model in Azure Stream Analytics.

Reference:

https://docs.microsoft.com/en-us/learn/modules/data-anomaly-detection-using-azure-iot-hub/

**cadio30** 4 years, 1 month ago

Propose solution is correct and better to perform hands-on to identify the other possible destination of data using Azure Stream Analytics

upvoted 4 times

**aksoumi** 4 years, 2 months ago

I am not sure if this is the correct answer. Never heard of Azure streaming analytics output being sent to Azure synapse/Data warehouse

upvoted 2 times

**Sherinm** 4 years, 2 months ago

It is possible.

https://docs.microsoft.com/en-us/azure/stream-analytics/azure-synapse-analytics-output#:~:text=Azure%20Stream%20Analytics%20jobs%20can,rates%20up%20to%20200MB%2Fsec.&text=To%20use%20Azure%20Synapse%20as,have%2

upvoted 4 times

**sdas1** 4 years, 3 months ago

answer is correct.

https://azure.github.io/iot-workshop-asset-tracking/step-003-anomaly-detection/

upvoted 4 times

Question #28 *Topic 2*

You are designing an Azure Data Factory pipeline for processing data. The pipeline will process data that is stored in general-purpose standard Azure storage.
You need to ensure that the compute environment is created on-demand and removed when the process is completed.
Which type of activity should you recommend?

A. Databricks Python activity

B. Data Lake Analytics U-SQL activity

C. HDInsight Pig activity

D. Databricks Jar activity

**Suggested Answer:** *C*
The HDInsight Pig activity in a Data Factory pipeline executes Pig queries on your own or on-demand HDInsight cluster.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/transform-data-using-hadoop-pig

**GabiN** Highly Voted 5 years, 4 months ago
According to Microsoft documentation: https://docs.microsoft.com/en-us/azure/data-factory/transform-data only 4 external transformations can be executed on-demand: HDInsight MapReduce Activity, HDInsight Hive Activity, HDInsight Pig Activity and HDInsight Streaming Activity. On-demand means that the computing environment is automatically created by the Data Factory service before a job is submitted to process data and removed when the job is completed. Therefore, the correct answer is C.
upvoted 53 times

**methodidacte** Highly Voted 5 years, 5 months ago
I agree with the solution C : "With on-demand HDInsight linked service, a HDInsight cluster is created every time a slice needs to be processed unless there is an existing live cluster (timeToLive) and is deleted when the processing is done." But why are the others false ?
upvoted 7 times

**H_S** Most Recent 4 years, 3 months ago
NOT IN THE DP-201 ANY MORE
upvoted 6 times

**Deepu1987** 4 years, 4 months ago
I would go with HDInsight Pig activity - rather than option A as per the given condition in the question where we're using ADLS n data bricks is ideally used during ADLS Gen2
upvoted 1 times

**syu31svc** 4 years, 6 months ago
I would agree with the answer
From https://docs.microsoft.com/en-us/azure/data-factory/v1/data-factory-compute-linked-services#:~:text=When%20the%20job%20is%20finished,cluster%20management%2C%20and%20bootstrapping%20actions.:
"Data Factory automatically creates the compute environment before a job is submitted for processing data. When the job is finished, Data Factory removes the compute environment."
"The Azure Storage linked service to be used by the on-demand cluster for storing and processing data. The HDInsight cluster is created in the same region as this storage account.
Currently, you can't create an on-demand HDInsight cluster that uses Azure Data Lake Store as the storage. If you want to store the result data from HDInsight processing in Data Lake Store, use Copy Activity to copy the data from Blob storage to Data Lake Store."
upvoted 1 times

**GraceCyborg** 4 years, 7 months ago
HDinsight is not in dp201 anymore
upvoted 2 times

**Abhilvs** 5 years ago
Azure Databricks also supports on-demand. when running from Az Datafactory, Databricks cluster gets created as an Automated cluster and destroyed after completion. The question is ambiguous.
upvoted 2 times

⊟ **Runi** 5 years ago

The HDInsight Pig activity in a Data Factory pipeline executes Pig queries on your own or on-demand Windows/Linux-based HDInsight cluster. See Pig activity article for details about this activity.

Same as Mapreduce , streaming and hive activity - mentioned explicitly "on your own or on-demand" and based on on demand "On-Demand: In this case, the computing environment is fully managed by Data Factory. It is automatically created by the Data Factory service before a job is submitted to process data and removed when the job is completed. You can configure and control granular settings of the on-demand compute environment for job execution, cluster management, and bootstrapping actions." However, python or jar activities doesn't do any on-demand process. So answer is C.

upvoted 1 times

⊟ **Leonido** 5 years, 2 months ago

It's the strange question. Every one of them could answer the demand.

upvoted 3 times

⊟ **azurearch** 5 years, 1 month ago

The Azure Databricks Python Activity in a Data Factory pipeline runs a Python file in your Azure Databricks cluster. This article builds on the data transformation activities article, which presents a general overview of data transformation and the supported transformation activities. Azure Databricks is a managed platform for running Apache Spark.

upvoted 1 times

⊟ **Narender_Bhadrecha** 5 years, 4 months ago

A is also correct answer.

upvoted 2 times

⊟ **mustaphaa** 5 years, 5 months ago

A and D are correct too, u can use automatic created cluster option in linked services

upvoted 1 times

Question #29 *Topic 2*

A company installs IoT devices to monitor its fleet of delivery vehicles. Data from devices is collected from Azure Event Hub.
The data must be transmitted to Power BI for real-time data visualizations.
You need to recommend a solution.
What should you recommend?

A. Azure HDInsight with Spark Streaming

B. Apache Spark in Azure Databricks

C. Azure Stream Analytics

D. Azure HDInsight with Storm

**Suggested Answer:** *C*
Step 1: Get your IoT hub ready for data access by adding a consumer group.
Step 2: Create, configure, and run a Stream Analytics job for data transfer from your IoT hub to your Power BI account.
Step 3: Create and publish a Power BI report to visualize the data.
Reference:
https://docs.microsoft.com/en-us/azure/iot-hub/iot-hub-live-data-visualization-in-power-bi

**azurearch** Highly Voted 👍 5 years, 1 month ago
data is already collected from event hub as per question, hence stream analytics is correct
upvoted 33 times

**Deepu1987** Most Recent ⏲ 4 years, 4 months ago
It's ASA - Azure Stream Analytics where it's fastest way to view real time data visualizations
upvoted 2 times

**chaoxes** 4 years, 6 months ago
C. Azure Stream Analytics

It is most efficient with Event Hubs
upvoted 2 times

**syu31svc** 4 years, 6 months ago
It can only be stream analytics
upvoted 1 times

**runningman** 5 years, 1 month ago
Is 'B' wrong/no good because it says Apache Spark with Databricks? If 'Databricks' was by itself, wouldn't that be an acceptable answer? Databricks can model and serve to BI.
upvoted 2 times

**Abhilvs** 5 years ago
Azure Databricks is suitable for complex analysis. With Strem analytics, one can query event data and perform the required analysis on it. upon the data can directly send to Power BI without any other interface between, SA has Power BI as an output stream.
upvoted 5 times

**Yuri1101** 5 years, 2 months ago
The given reference is incorrect. Should use Event Hub.
https://docs.microsoft.com/en-us/azure/event-hubs/event-hubs-tutorial-visualize-anomalies
upvoted 2 times

Question #30 *Topic 2*

You have a Windows-based solution that analyzes scientific data. You are designing a cloud-based solution that performs real-time analysis of the data.
You need to design the logical flow for the solution.
Which two actions should you recommend? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.

A. Send data from the application to an Azure Stream Analytics job.

B. Use an Azure Stream Analytics job on an edge device. Ingress data from an Azure Data Factory instance and build queries that output to Power BI.

C. Use an Azure Stream Analytics job in the cloud. Ingress data from the Azure Event Hub instance and build queries that output to Power BI.

D. Use an Azure Stream Analytics job in the cloud. Ingress data from an Azure Event Hub instance and build queries that output to Azure Data Lake Storage.

E. Send data from the application to Azure Data Lake Storage.

F. Send data from the application to an Azure Event Hub instance.

**Suggested Answer:** *CF*
Stream Analytics has first-class integration with Azure data streams as inputs from three kinds of resources:
✑ Azure Event Hubs
✑ Azure IoT Hub
✑ Azure Blob storage
Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-define-inputs

⊟ **anupit** Highly Voted 👍 4 years, 11 months ago
CF is the correct one, no confusion
upvoted 36 times

⊟ **Arsa** Highly Voted 👍 4 years, 10 months ago
CF correct
upvoted 15 times

⊟ **cadio30** Most Recent 🕒 4 years, 1 month ago
As stated C and F are the correct solutions though if it is needed to configure it sequentially then F and C are the appropriate steps to accomplish it.

Event Hub > Azure Stream Analytics (Cloud) > Power BI (output for real time analysis)
upvoted 1 times

⊟ **Deepu1987** 4 years, 4 months ago
we can't o/p azure data lake storage as we have a better option Power BI as we can run Power BI directly on ASA
upvoted 1 times

⊟ **Deepu1987** 4 years, 4 months ago
The 1st step would be read option F & then option C in order to understand the logical flow
D cannot be included as per the given scenario.
upvoted 1 times

⊟ **spiitr** 4 years, 4 months ago
Why not CD? and how do you meet "performs real-time analysis of the data." through event hub?
upvoted 1 times

⊟ **spiitr** 4 years, 4 months ago
I guess these are two actions or steps part of same solution. If so CF is correct. First action is F and then perform C.
upvoted 1 times

⊟ **Sriniv** 4 years, 4 months ago
Yes for real time analysis input through event hub and output to BI
upvoted 1 times

⊟ **JMCun** 4 years, 5 months ago

Why is the output to DLake v2 option wrong? Thanks

upvoted 5 times

⊟ **maciejt** 4 years, 2 months ago

I think that data lake cannot be tread as last step to consume data, there would be needed a third step to consume data from DL. PowerBI visualization can be threated as final data consumption

upvoted 1 times

⊟ **ExamW** 4 years, 11 months ago

The answer should be DF

upvoted 1 times

⊟ **ExamW** 4 years, 11 months ago

Ignore this. CF is correct

upvoted 40 times

⊟ **Israel2** 4 years, 11 months ago

Could it be D? Can you run Power BI directly on Streaming Analytics? Or do you need direct query for real time?

upvoted 2 times

⊟ **ExamW** 4 years, 11 months ago

https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-power-bi-dashboard It seems you can...

upvoted 2 times

⊟ **knightkkd** 4 years, 8 months ago

Direct Query works in real time when you have Power BI Premium subcription

upvoted 2 times

DRAG DROP -

You are designing a Spark job that performs batch processing of daily web log traffic.

When you deploy the job in the production environment, it must meet the following requirements:

✑ Run once a day.

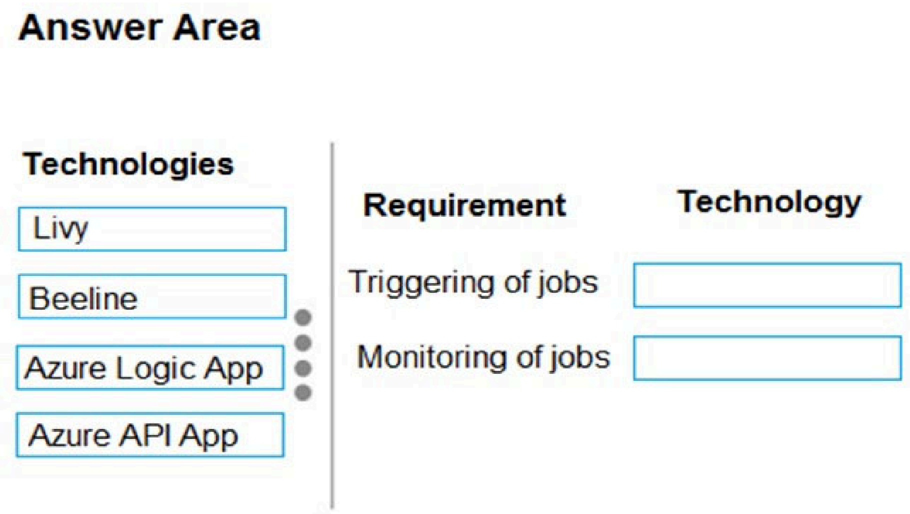
✑ Display status information on the company intranet as the job runs.

You need to recommend technologies for triggering and monitoring jobs.

Which technologies should you recommend? To answer, drag the appropriate technologies to the correct locations. Each technology may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point.

Select and Place:



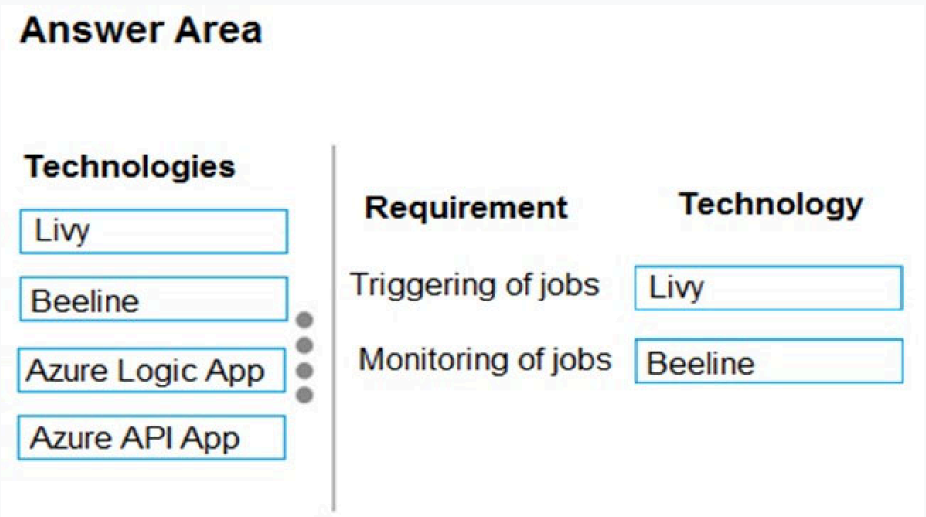
## Answer Area

**Technologies**

- Livy
- Beeline
- Azure Logic App
- Azure API App

| Requirement | Technology |
|---|---|
| Triggering of jobs | |
| Monitoring of jobs | |

**Suggested Answer:**



## Answer Area

**Technologies**

- Livy
- Beeline
- Azure Logic App
- Azure API App

| Requirement | Technology |
|---|---|
| Triggering of jobs | Livy |
| Monitoring of jobs | Beeline |

Box 1: Livy -

You can use Livy to run interactive Spark shells or submit batch jobs to be run on Spark.

Box 2: Beeline -

Apache Beeline can be used to run Apache Hive queries on HDInsight. You can use Beeline with Apache Spark.

Note: Beeline is a Hive client that is included on the head nodes of your HDInsight cluster. Beeline uses JDBC to connect to HiveServer2, a service hosted on your

HDInsight cluster. You can also use Beeline to access Hive on HDInsight remotely over the internet.

Reference:

https://docs.microsoft.com/en-us/azure/hdinsight/spark/apache-spark-livy-rest-interface https://docs.microsoft.com/en-us/azure/hdinsight/hadoop/apache-hadoop-use-hive-beeline

**H_S** Highly Voted 4 years, 3 months ago

NOT ANY MORE IN DP-201

upvoted 6 times

⊟ **jacint** 4 years ago

are you sure? how do you know?

upvoted 1 times

⊟ **akram786** Highly Voted 👍 4 years, 3 months ago

Beeline used for running hive queries not for monitoring

upvoted 5 times

⊟ **dakku987** Most Recent 🕓 1 year, 6 months ago

Beeline or Livy for triggering the Spark job.

Azure Logic Apps for monitoring and updating status on the company intranet.

upvoted 1 times

⊟ **seby** 4 years ago

Guys, then which is the answer ?

upvoted 1 times

⊟ **Makar** 4 years, 3 months ago

answer is right https://livy.incubator.apache.org/

upvoted 2 times

## Question #32 *Topic 2*

You have an Azure Databricks workspace named workspace1 in the Standard pricing tier. Workspace1 contains an all-purpose cluster named cluster1.

You need to reduce the time it takes for cluster1 to start and scale up. The solution must minimize costs.

What should you do first?

A. Upgrade workspace1 to the Premium pricing tier.

B. Create a pool in workspace1.

C. Configure a global init script for workspace1.

D. Create a cluster policy in workspace1.

**Suggested Answer:** *B*

Databricks Pools increase the productivity of both Data Engineers and Data Analysts. With Pools, Databricks customers eliminate slow cluster start and auto- scaling times. Data Engineers can reduce the time it takes to run short jobs in their data pipeline, thereby providing better SLAs to their downstream teams.

Reference:

https://databricks.com/blog/2019/11/11/databricks-pools-speed-up-data-pipelines.html

⊟ **VG2007** Highly Voted 👍 4 years, 1 month ago

Correct.

upvoted 7 times

⊟ **PHULU** Most Recent 🕒 3 years, 8 months ago

B is the answer

upvoted 1 times

Question #33 *Topic 2*

HOTSPOT -

You are designing a solution to process data from multiple Azure event hubs in near real-time.

Once processed, the data will be written to an Azure SQL database.

The solution must meet the following requirements:

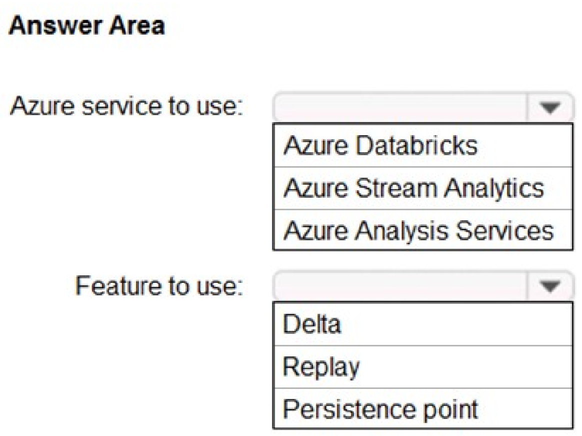
✑ Support the auditing of resource and data changes.

✑ Support data versioning and rollback.

What should you recommend? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

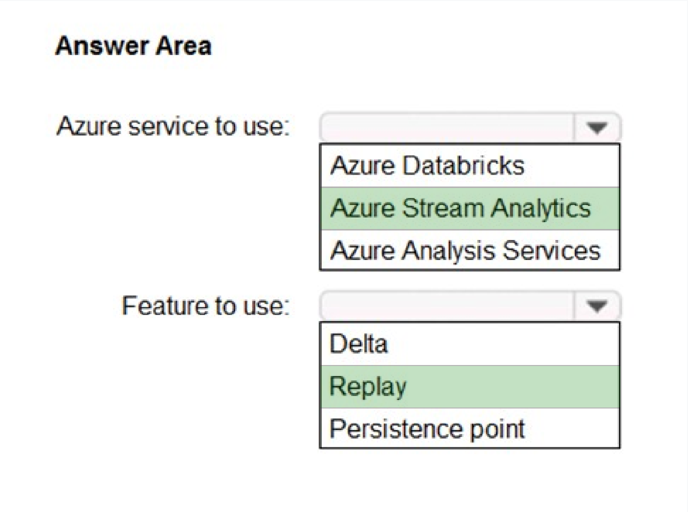


**Answer Area**

Azure service to use:
- Azure Databricks
- Azure Stream Analytics
- Azure Analysis Services

Feature to use:
- Delta
- Replay
- Persistence point

**Suggested Answer:**



**Answer Area**

Azure service to use:
- Azure Databricks
- [✓ Azure Stream Analytics]
- Azure Analysis Services

Feature to use:
- Delta
- [✓ Replay]
- Persistence point

Box 1: Azure Stream Analytics -

Users can now ingest, process, view, and analyze real-time streaming data into a table directly from a database in Azure SQL Database. They do so in the Azure portal using Azure Stream Analytics.

In the Azure portal, you can select an events source (Event Hub/IoT Hub), view incoming real-time events, and select a table to store events.

Stream Analytics leverages versioning of reference data to augment streaming data with the reference data that was valid at the time the event was generated.

This ensures repeatability of results.

Box 2: Replay -

Reference data is versioned, enabling to always get the same results, even when we ג€replayג€ the stream.

Reference:

https://docs.microsoft.com/en-us/azure/azure-sql/database/stream-data-stream-analytics-integration https://azure.microsoft.com/en-us/updates/additional-support-for-managed-identity-and-new-features-in-azure-stream-analytics/

obz Highly Voted 4 years, 3 months ago

Azure streaming analytics and replay are about job recovery.
I would definitely go for Azure Databricks and Delta.

https://databricks.com/blog/2019/02/04/introducing-delta-time-travel-for-large-scale-data-lakes.html
upvoted 38 times

**hichemck** 4 years ago
Also azure stream analytics does not support multiple inputs
upvoted 1 times

**cadio30** 4 years, 1 month ago
Agree with this solution. Better to compare the delta table of Azure Databricks against the Azure Stream Analytics
upvoted 2 times

**aditya_064** Highly Voted 4 years, 2 months ago
Data is to be written to Azure SQL Database, Azure Databricks doesn't support it as a sink - https://docs.microsoft.com/en-us/azure/architecture/data-guide/technology-choices/stream-processing. Therefore ASA is the only otpion left along with Replay feature. Replay may not be the best method but since ASA has to be included this will do. Also Azure SQL Database can handle rollbacks and auditing by itself. So given solution is correct.
upvoted 8 times

**dakku987** Most Recent 1 year, 6 months ago
Azure synpse analytics and delta

Azure Service to Use: Azure Synapse Analytics (formerly SQL Data Warehouse):

Azure Synapse Analytics is a comprehensive analytics service that brings together big data and data warehousing. It supports near real-time analytics and integrates with Azure SQL Database.
Feature to Use: Delta:

Delta is a feature associated with Azure Synapse Analytics that provides capabilities like versioning, change tracking, and rollback options. It is designed to handle data versioning and changes efficiently.
upvoted 1 times

**nefarious_smalls** 3 years, 1 month ago
Job Replay is simply about recovering your stream to its prior state in the case of a service upgrade or any manual intervention in which you stop the stream. It has nothing to do with versioning or rolling back to a specific version. That is Delta.
upvoted 1 times

**muni53** 3 years, 9 months ago
azure databricks and delta tables. Databricks can use delta feature to load required snapshot. ADB can load data to sql db.
upvoted 1 times

**zarga** 3 years, 11 months ago
Azure databricks and Delta
We can push data from databricks to azure sql db using jdbc driver.
upvoted 2 times

**Tracy_Anderson** 3 years, 11 months ago
https://www.sqlshack.com/load-data-into-azure-sql-database-from-azure-databricks/
upvoted 1 times

**VG2007** 4 years, 1 month ago
Not sure what is the confusion.. refrence links are already given in and looks correct ..
eference:
https://docs.microsoft.com/en-us/azure/azure-sql/database/stream-data-stream-analytics-integration https://azure.microsoft.com/en-us/updates/additional-support-for-managed-identity-and-new-features-in-azure-stream-analytics/
upvoted 3 times

**niwe** 4 years, 2 months ago
The given answer, I think i correct
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-concepts-checkpoint-replay
upvoted 2 times

**KRV** 4 years, 2 months ago

I think the questions talks about data versioning and auditing data and date related changes and not about the job recovery , and hence I think Azure Databricks and Delta happens to be logically correct with corelation to the question asked for Data changes and not job related changes or recovery

upvoted 2 times

Question #34 Topic 2

DRAG DROP -

You are designing a real-time processing solution for maintenance work requests that are received via email. The solution will perform the following actions:

✑ Store all email messages in an archive.

✑ Access weather forecast data by using the Python SDK for Azure Open Datasets.

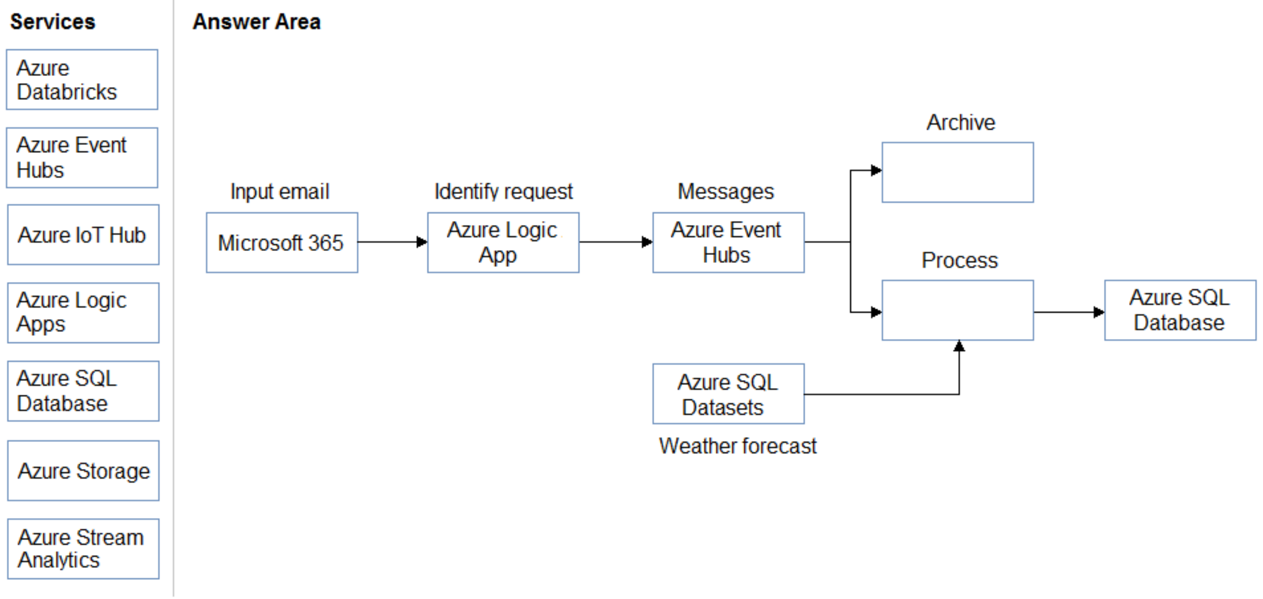
✑ Identify high priority requests that will be affected by poor weather conditions and store the requests in an Azure SQL database.

The solution must minimize costs.

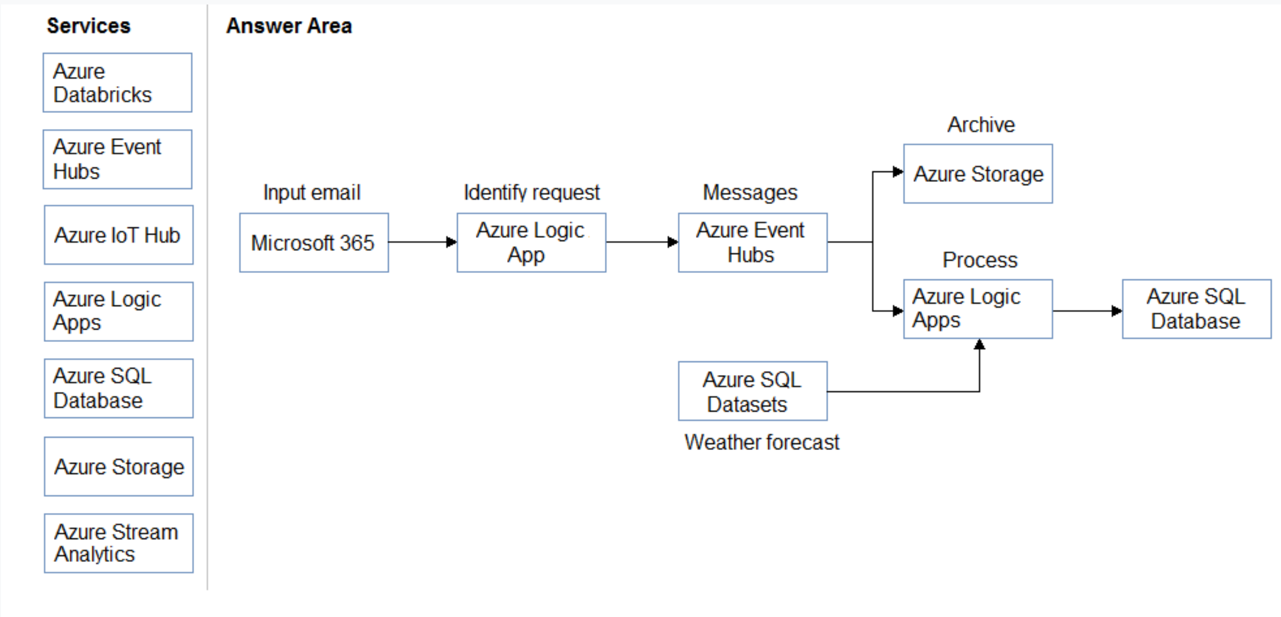
How should you complete the solution? To answer, drag the appropriate services to the correct locations. Each service may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point.

Select and Place:



**Suggested Answer:**



Box 1: Azure Storage -

Azure Event Hubs enables you to automatically capture the streaming data in Event Hubs in an Azure Blob storage or Azure Data Lake Storage Gen 1 or Gen 2 account of your choice, with the added flexibility of specifying a time or size interval. Setting up Capture is fast, there are no administrative costs to run it, and it scales automatically with Event Hubs throughput units. Event Hubs Capture is the easiest way to load streaming data into Azure, and enables you to focus on data processing rather than on data capture.

Box 2: Azure Logic Apps -

You can monitor and manage events sent to Azure Event Hubs from inside a logic app with the Azure Event Hubs connector. That way, you can

create logic apps that automate tasks and workflows for checking, sending, and receiving events from your Event Hub.
Reference:
https://docs.microsoft.com/en-us/azure/event-hubs/event-hubs-capture-overview https://docs.microsoft.com/en-us/azure/connectors/connectors-create-api-azure-event-hubs

**Jony** Highly Voted 4 years, 3 months ago
I believe it is Databricks in the second box due to the Python - Azure Open Datasets.
upvoted 36 times

**vrmei** 4 years ago
Databrick is the correct answer
upvoted 1 times

**H_S** 4 years, 3 months ago
it's for sure azure databricks
upvoted 5 times

**Debjit** 4 years, 3 months ago
How? The processing unit has to pick up incoming events and then process. ASA is best fit for it
upvoted 3 times

**cadio30** 4 years, 1 month ago
ASA doesn't have Python SDK on it
upvoted 1 times

**devpool** 4 years, 3 months ago
Jony is right.
https://docs.microsoft.com/en-us/azure/open-datasets/overview-what-are-open-datasets
upvoted 6 times

**vrmei** 4 years ago
Valid link
upvoted 1 times

**AlexD332** Highly Voted 4 years, 3 months ago
Can we use Stream Analytics instead of the Logic app?
upvoted 8 times

**MMM777** Most Recent 4 years ago
The NOAA data options only show Databricks or Synapse, which was not an option in the question, and Azure Notebooks which was in preview and no longer available (and also not an option in the question):

https://azure.microsoft.com/en-us/services/open-datasets/catalog/noaa-integrated-surface-data/#AzureDatabricks
upvoted 1 times

**Alka3** 4 years, 2 months ago
The key here is 'Identify high priority requests that will be affected by poor weather conditions..' - This can be done efficiently by Logic app.
upvoted 4 times

**BobFar** 4 years, 1 month ago
check this link
https://docs.microsoft.com/en-us/azure/open-datasets/overview-what-are-open-datasets
upvoted 1 times

**sdas1** 4 years, 3 months ago
Azure open dataset can be accessed from databricks or any Python environment with or without Spark. Hence, the second box should be Databricks.
https://docs.microsoft.com/en-us/azure/open-datasets/overview-what-are-open-datasets
upvoted 3 times

**ekko1224** 4 years, 3 months ago
Function Apps would be the way to go for any custom code needed in logic apps.
Python in Azure Functions is still in Preview, so it's not recommended for production use.
You can refer to this for more information on creating a new python function app.
https://social.msdn.microsoft.com/Forums/en-US/1b46dcb2-2832-4861-a214-e49e85247d53/how-can-i-run-a-python-script-in-logic-apps

upvoted 1 times

⊟ **jms309** 4 years, 3 months ago

I think they are not in preview right now, this thread is from 2018, so it is not up to date

upvoted 1 times

⊟ **H_S** 4 years, 3 months ago

yet databricks is good condidate

upvoted 2 times

⊟ **maynard13x8** 4 years, 2 months ago

I think Logic App (in which you can use Python by mean of function App) o stream analytics are better.

upvoted 1 times

Question #35 *Topic 2*

You have a large amount of sensor data stored in an Azure Data Lake Storage Gen2 account. The files are in the Parquet file format.
New sensor data will be published to Azure Event Hubs.
You need to recommend a solution to add the new sensor data to the existing sensor data in real-time. The solution must support the interactive querying of the entire dataset.
Which type of server should you include in the recommendation?

A. Azure SQL Database

B. Azure Cosmos DB

C. Azure Stream Analytics

D. Azure Databricks

**Suggested Answer:** *C*
Azure Stream Analytics is a fully managed PaaS offering that enables real-time analytics and complex event processing on fast moving data streams.
By outputting data in parquet format into a blob store or a data lake, you can take advantage of Azure Stream Analytics to power large scale streaming extract, transfer, and load (ETL), to run batch processing, to train machine learning algorithms, or to run interactive queries on your historical data.
Reference:
https://azure.microsoft.com/en-us/blog/new-capabilities-in-stream-analytics-reduce-development-time-for-big-data-apps/

**sdas1** Highly Voted 4 years, 3 months ago
As per below link the answer is correct.
https://azure.microsoft.com/en-in/blog/new-capabilities-in-stream-analytics-reduce-development-time-for-big-data-apps/
upvoted 5 times

**cadio30** Highly Voted 4 years, 1 month ago
Both Azure Databricks and Azure Stream analytics can output data to parquet format and have interactive queries as well. For simplicity, I'll choose Azure Stream Analytics
upvoted 5 times

> **cadio30** 4 years ago
> Opt to choose Azure Databricks instead of Azure Streaming Analytics due to the keywork 'large dataset'
>
> Reference: https://techcommunity.microsoft.com/t5/analytics-on-azure/azure-stream-analytics-real-time-analytics-for-big-data-made/ba-p/549621
> upvoted 2 times

> **BobFar** 4 years, 1 month ago
> ASA doesn't support Parquet format.!
> upvoted 2 times
>
> > **BobFar** 4 years, 1 month ago
> > I was wrong, it supports now
> >
> > https://azure.microsoft.com/en-us/updates/stream-analytics-offers-native-support-for-parquet-format/#:~:text=Azure%20Stream%20Analytics%20now%20offers,in%20the%20Big%20Data%20ecosystems.
> > upvoted 1 times

> **mbravo** 4 years, 1 month ago
> One of the requirements is to be able to interactively query the whole (possibly very large) dataset according to the scenario. This requirement alone is a perfect fit for Spark. I highly doubt there is a sensible way to achieve this with ASA. Therefore I vote for Databricks.
> upvoted 1 times

**daradev** Most Recent 3 years, 11 months ago
By outputting data in parquet format into a blob store or a data lake, you can take advantage of Azure Stream Analytics to power large scale streaming extract, transfer, and load (ETL), to run batch processing, to train machine learning algorithms, or to run interactive queries on your historical data.

Soure: https://azure.microsoft.com/en-in/blog/new-capabilities-in-stream-analytics-reduce-development-time-for-big-data-apps/

upvoted 1 times

**hello_there_** 3 years, 10 months ago

What this quote says is that ASA can output parquet format to blob storage, so that another tool can then run interactive queries on the data. ASA itself can't do interactive queries on parquet in blob storage, which is what is required here. I'd go with databricks.

upvoted 1 times

**VG2007** 4 years, 1 month ago

Native support for egress in Apache parquet format into Azure Blob Storage is now generally available. Parquet is a columnar format enabling efficient big data processing. By outputting data in parquet format into a blob store or a data lake, you can take advantage of Azure Stream Analytics to power large scale streaming extract, transfer, and load (ETL), to run batch processing, to train machine learning algorithms, or to run interactive queries on your historical data. We are now announcing general availability of this feature for egress to Azure Blob Storage.

upvoted 3 times

**jms309** 4 years, 3 months ago

I think that Databrick is a good answer. I'm not sure if Azure Stream Analytics is another right answer but maybe there are two possibilities

upvoted 2 times

**anamaster** 4 years, 2 months ago

interactive querying eliminates ASA

upvoted 3 times

**niwe** 4 years, 2 months ago

Azure Stream Analytics does not support Parquet data format.

https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-stream-analytics-query-patterns

upvoted 2 times

**saifone** 4 years, 1 month ago

It does as of July 2019 https://azure.microsoft.com/en-us/updates/stream-analytics-offers-native-support-for-parquet-format/

upvoted 2 times

**sdas1** 4 years, 3 months ago

As per below link the answer is correct.

new-capabilities-in-stream-analytics-reduce-development-time-for-big-data-app

upvoted 1 times

**YOMYOM** 4 years, 3 months ago

is C really the correct answer pls?

upvoted 2 times

**H_S** 4 years, 3 months ago

i think it's D because the interactive querying of the entire dataset.

entire dataset/interative isn't possible with A.stream

upvoted 7 times

Question #36 *Topic 2*

HOTSPOT -

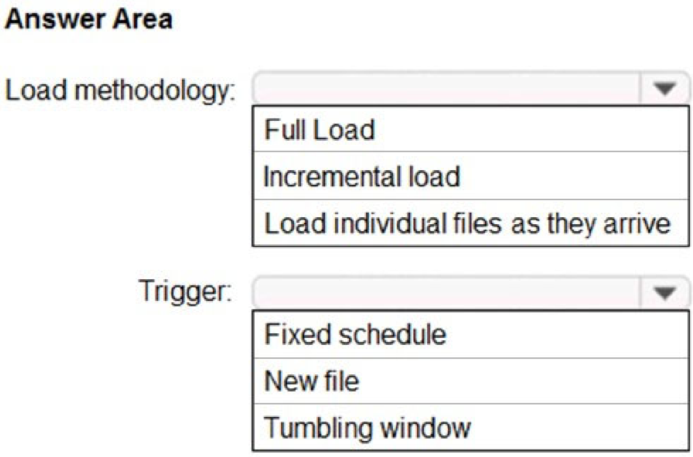
You have an Azure Storage account that generates 200,000 new files daily. The file names have a format of {YYYY}/{MM}/{DD}/{HH}/{CustomerID}.csv.

You need to design an Azure Data Factory solution that will load new data from the storage account to an Azure Data Lake once hourly. The solution must minimize load times and costs.

How should you configure the solution? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point

Hot Area:

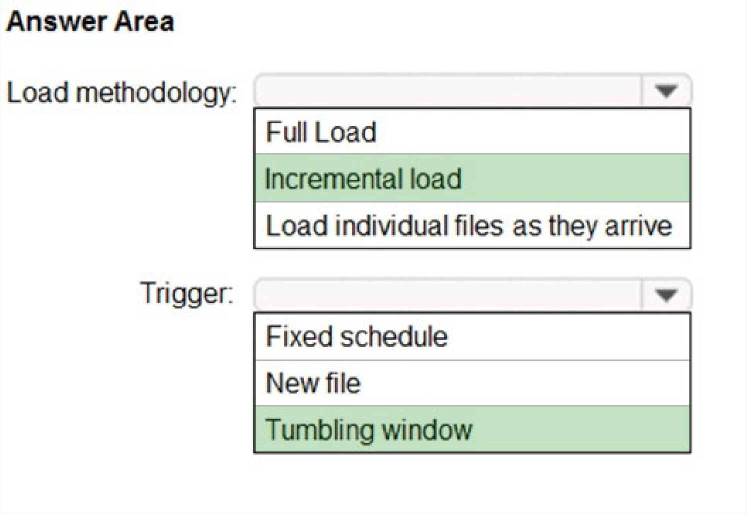


**Answer Area**

Load methodology:
- Full Load
- Incremental load
- Load individual files as they arrive

Trigger:
- Fixed schedule
- New file
- Tumbling window

**Suggested Answer:**



**Answer Area**

Load methodology:
- Full Load
- [✓ Incremental load]
- Load individual files as they arrive

Trigger:
- Fixed schedule
- New file
- [✓ Tumbling window]

Box 1: Incremental load -

When you start to build the end to end data integration flow the first challenge is to extract data from different data stores, where incrementally (or delta) loading data after an initial full load is widely used at this stage. Now, ADF provides a new capability for you to incrementally copy new or changed files only by

LastModifiedDate from a file-based store. By using this new feature, you do not need to partition the data by time-based folder or file name. The new or changed file will be automatically selected by its metadata LastModifiedDate and copied to the destination store.

Box 2: Tumbling window -

Tumbling window triggers are a type of trigger that fires at a periodic time interval from a specified start time, while retaining state. Tumbling windows are a series of fixed-sized, non-overlapping, and contiguous time intervals. A tumbling window trigger has a one-to-one relationship with a pipeline and can only reference a singular pipeline.

Reference:

https://azure.microsoft.com/en-us/blog/incrementally-copy-new-files-by-lastmodifieddate-with-azure-data-factory/

https://docs.microsoft.com/en-us/azure/data-factory/how-to-create-tumbling-window-trigger

**satyamkishoresingh** 3 years, 10 months ago

I believe it can be tumbling as well as fixed schedule as they both can do fixed hour

upvoted 1 times

**mbravo** 4 years ago

According the MS documentation, incremental loads are used together with tumbling window. Tumbling window is used in both of these examples where we are performing an incremental load from Blob Storage.

https://docs.microsoft.com/en-us/azure/data-factory/tutorial-incremental-copy-lastmodified-copy-data-tool

and

https://docs.microsoft.com/en-us/azure/data-factory/tutorial-incremental-copy-partitioned-file-name-copy-data-tool

upvoted 3 times

**MMM777** 4 years ago

Tumbling Window trigger is a "smarter" run - what if the pipeline takes longer than an hour to run?

https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipeline-execution-triggers#trigger-type-comparison

upvoted 1 times

**BobFar** 4 years ago

The appropriate solutions are 'incremental load' and schedule is mentioned as once hourly which is 'Tumbling window' the answer is correct!

upvoted 2 times

**cadio30** 4 years, 1 month ago

The appropriate solutions are 'incremental load' and 'fixed schedule' as the basis is 1 hour trigger and the use of tumbling window requires further configuration than the mentioned schedule earlier. It would be better if there is an option to use 'storage events' as the ADF will trigger if a blob is created or deleted.

Reference: https://www.mssqltips.com/sqlservertip/6061/create-tumbling-window-trigger-in-azure-data-factory-adf/

upvoted 3 times

**cadio30** 4 years ago

In the event the requirement requires to take in consideration the load processing time then tumbling window is the appropriate configuration as the trigger won't overlap.

upvoted 1 times

**tamil1006** 4 years, 1 month ago

tumbling window will be used for stream analytics...

upvoted 1 times

**Amy007** 4 years, 2 months ago

But schedule is mentioned as once hourly , why would it be Tumbling window ?

upvoted 4 times

**Debjit** 4 years, 3 months ago

If its tumbling window then why not new individual file as they arrive? Tumbling window works only when a new event occurs

upvoted 1 times

**Debjit** 4 years, 3 months ago

ignore. The answer is correct

upvoted 4 times

Question #37 *Topic 2*

You are designing a solution that will copy Parquet files stored in an Azure Blob storage account to an Azure Data Lake Storage Gen2 account.
The data will be loaded daily to the data lake and will use a folder structure of {Year}/{Month}/{Day}/.
You need to design a daily Azure Data Factory data load to minimize the data transfer between the two accounts.
Which two configurations should you include in the design? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.

A. Delete the files in the destination before loading new data.

B. Filter by the last modified date of the source files.

C. Delete the source files after they are copied.

D. Specify a file naming pattern for the destination.

**Suggested Answer:** *BC*
B: To copy a subset of files under a folder, specify folderPath with a folder part and fileName with a wildcard filter.
C: After completion: Choose to do nothing with the source file after the data flow runs, delete the source file, or move the source file. The paths for the move are relative.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/connector-azure-data-lake-storage

**phi618t** Highly Voted 👍 4 years ago
If you choose C. Delete the source files after they are copied, why do you choose B. Filter by the last modified date of the source files? I prefer BD.
upvoted 12 times

**Marcus1612** 3 years, 9 months ago
This is a basic question. Copy data from one place to another. The requirements are : 1- need to minimize transfert and 2- need to adapte data to the destination folder structure. Filter on LastModifiedDate will copy everything that have changed since the latest load while minimizing the data transfert. Specifying the file naming pattern allows to copy data at the right place to the destination Data Lake. The answer is BD
upvoted 2 times

**Bhagya123456** 3 years, 10 months ago
How naming pattern gonna minimize the Data Transfer? BC should be correct answer.
upvoted 1 times

**Wendy_DK** Highly Voted 👍 4 years, 1 month ago
Correct answer is BC.
In the source option of copy activities. There are three choices: 1. No Action 2. Delete Source files 3. Move
upvoted 10 times

**BigMF** Most Recent ⏲ 4 years ago
A is obviously out and you're are not going to do both B and C so D is in by default. Your only choice at that point is B or C to go along with D. In my experience, you cannot rely 100% on any job to run every single day (assuming this process is daily). Therefore, if the job does not run for one or more days, if you were to choose B you would only copy over the most recent files and there would be files left in the storage account. Therefore, my choice would be to not filter and load everything that is in the storage account and then delete the files once they have been copied. So, C and D are my choices.
upvoted 4 times

**YLiu** 3 years, 9 months ago
B ensures minimized data transfer. If it copies everything every time, then data transfer is not minimized.
upvoted 1 times

**mter2007** 4 years, 2 months ago
I would like to choose CD.
upvoted 3 times

**maciejt** 4 years, 2 months ago
The was no requirement what to do with original files, so why i the world anwer C - delete them???
upvoted 3 times

**BobFar** 4 years ago

I guess to make sure you dont read the file again!

upvoted 1 times

⊟ **Nik71** 4 years, 3 months ago

C seems not correcct as to deletion you can do life cycle mgmt in storage, so D should be second answer.

upvoted 2 times

⊟ **AlexD332** 4 years, 3 months ago

thought it's the only logical choice but they said copy activity not moving files

upvoted 1 times

⊟ **H_S** 4 years, 3 months ago

I think it"s BD

upvoted 22 times

⊟ **etl** 4 years, 3 months ago

Wildcard path: Using a wildcard pattern will instruct ADF to loop through each matching folder and file in a single Source transformation. This is an effective way to process multiple files within a single flow. Add multiple wildcard matching patterns with the + sign that appears when hovering over your existing wildcard pattern.

From your source container, choose a series of files that match a pattern. Only container can be specified in the dataset. Your wildcard path must therefore also include your folder path from the root folder.

upvoted 1 times

⊟ **etl** 4 years, 3 months ago

yes BD.. i think you are right

upvoted 4 times

⊟ **maciejt** 4 years, 2 months ago

but this applies to finding a source files and D was about destintion file naming pattern... which there were no requirement to change the file name

upvoted 2 times

⊟ **cadio30** 4 years, 1 month ago

Agree with the answer B and D as this kind of setup doesn't perform any deletion from both storages which lessen the processing.

upvoted 3 times

Question #38 *Topic 2*

You have a C# application that process data from an Azure IoT hub and performs complex transformations.
You need to replace the application with a real-time solution. The solution must reuse as much code as possible from the existing application.

A. Azure Databricks

B. Azure Event Grid

C. Azure Stream Analytics

D. Azure Data Factory

**Suggested Answer:** *C*
Azure Stream Analytics on IoT Edge empowers developers to deploy near-real-time analytical intelligence closer to IoT devices so that they can unlock the full value of device-generated data. UDF are available in C# for IoT Edge jobs
Azure Stream Analytics on IoT Edge runs within the Azure IoT Edge framework. Once the job is created in Stream Analytics, you can deploy and manage it using
IoT Hub.
References:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-edge

*Community vote distribution*

C (100%)

**Wendy_DK** Highly Voted 4 years, 1 month ago

Correct answer is C

upvoted 12 times

**Bhagya123456** 3 years, 10 months ago

Nope It has to be Databricks.

upvoted 2 times

**rmk4ever** Highly Voted 4 years, 1 month ago

Ans : A
Apache Spark in Azure Databricks supports C#.
Ref: https://docs.microsoft.com/en-us/azure/architecture/data-guide/technology-choices/stream-processing

upvoted 10 times

**Saravjeet** 4 years, 1 month ago

This seems to be correct answer as Databricks, as from the available options only Databricks have the support for C#. so the question says maximum reusability of the code, hence Databricks must be the correct answer. Thanks.

upvoted 2 times

**BigMF** 4 years ago

ASA supports C#: https://azure.microsoft.com/en-us/blog/supercharge-your-azure-stream-analytics-query-with-c-code/

upvoted 1 times

**dinesh_tng** 4 years ago

That is limited only to IoT Edge devices....Az Databricks looks more relevant

upvoted 2 times

**dakku987** Most Recent 1 year, 6 months ago

Selected Answer: C

Azure synapse analytics

Azure Databricks is a big data analytics platform. While it's powerful, it may not be the most straightforward choice if the goal is to reuse existing C# code for real-time processing.

Given the requirements, Azure Stream Analytics is likely the most appropriate option for a real-time solution with the potential for code reuse.

chat gpt

upvoted 1 times

⊟ **satyamkishoresingh** 3 years, 10 months ago

question says complex transformation, I think ASA is not the right fit for that kind calculation. I would go with databricks

upvoted 1 times

⊟ **maciejt** 4 years, 2 months ago

why not databricks? can also utilize c# code and process data in real time

upvoted 5 times

⊟ **coldog86** 4 years, 2 months ago

SPARK does real time processing, not Databricks

https://docs.microsoft.com/en-us/azure/architecture/data-guide/big-data/real-time-processing

upvoted 5 times

⊟ **dinesh_tng** 4 years ago

yeah, but spark is core of Databricks solution

upvoted 1 times

Question #39 *Topic 2*

HOTSPOT -

You are designing an Azure Data Factory solution that will download up to 5 TB of data from several REST APIs.

The solution must meet the following staging requirements:

✑ Ensure that the data can be landed quickly and in parallel to a staging area.

✑ Minimize the need to return to the API sources to retrieve the data again should a later activity in the pipeline fail.

The solution must meet the following analysis requirements:

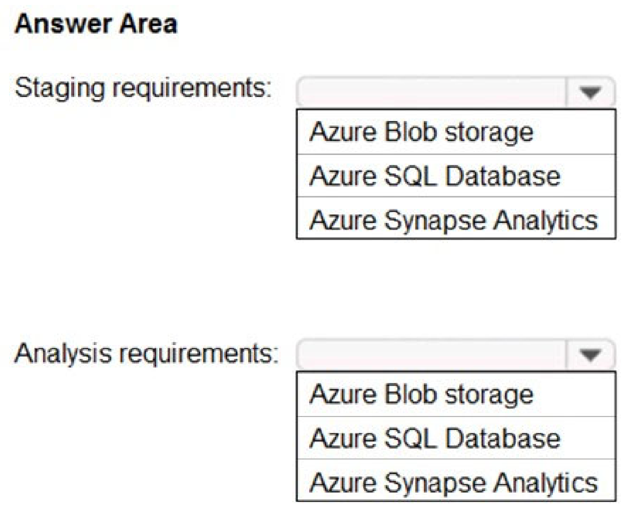
✑ Ensure that the data can be loaded in parallel.

✑ Ensure that users and applications can query the data without requiring an additional compute engine.

What should you include in the solution to meet the requirements? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:



**Answer Area**

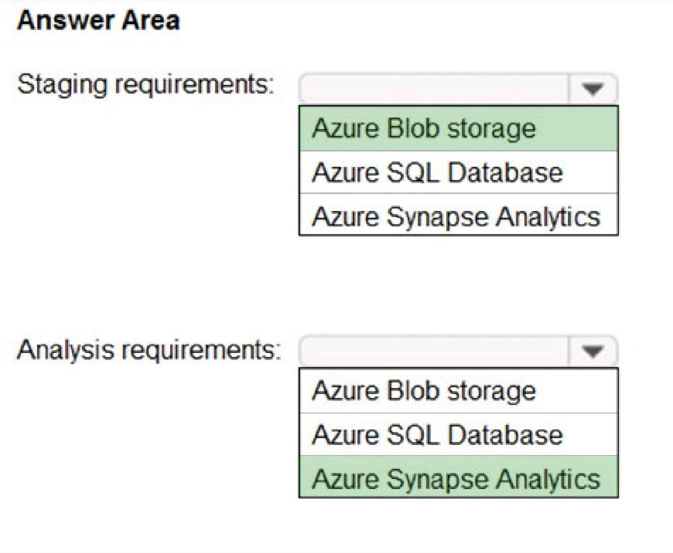
Staging requirements:

- Azure Blob storage
- Azure SQL Database
- Azure Synapse Analytics

Analysis requirements:

- Azure Blob storage
- Azure SQL Database
- Azure Synapse Analytics

**Suggested Answer:**



**Answer Area**

Staging requirements:

- [✓ Azure Blob storage]
- Azure SQL Database
- Azure Synapse Analytics

Analysis requirements:

- Azure Blob storage
- Azure SQL Database
- [✓ Azure Synapse Analytics]

Box 1: Azure Blob storage -

When you activate the staging feature, first the data is copied from the source data store to the staging storage (bring your own Azure Blob or Azure Data Lake

Storage Gen2).

Box 2: Azure Synapse Analytics -

The Azure Synapse Analytics connector in copy activity provides built-in data partitioning to copy data in parallel.

Reference:

https://docs.microsoft.com/en-us/azure/data-factory/copy-activity-performance-features https://docs.microsoft.com/en-us/azure/data-factory/connector-azure-sql-data-warehouse

**Marcus1612** 3 years, 9 months ago

Look at this: https://docs.microsoft.com/en-us/azure/data-factory/copy-activity-performance-features

When you activate the staging feature, first the data is copied from the source data store to the staging storage (bring your own Azure Blob or Azure Data Lake Storage Gen2). Next, the data is copied from the staging to the sink data store. The copy activity automatically manages the two-stage flow for you, and also cleans up temporary data from the staging storage after the data movement is complete.

upvoted 2 times

**anamaster** 4 years, 2 months ago

correct, but the explanation for synapse is that ASA allows querying

upvoted 2 times

## Question #40 *Topic 2*

A company purchases IoT devices to monitor manufacturing machinery. The company uses an IoT appliance to communicate with the IoT devices.
The company must be able to monitor the devices in real-time.
You need to design the solution.
What should you recommend?

A. Azure Data Factory instance using Azure Portal

B. Azure Analysis Services using Microsoft Visual Studio

C. Azure Stream Analytics cloud job using Azure Portal

D. Azure Data Factory instance using Azure Portal

**Suggested Answer:** *C*
The Stream Analytics query language allows to perform CEP (Complex Event Processing) by offering a wide array of functions for analyzing streaming data. This query language supports simple data manipulation, aggregation and analytics functions, geospatial functions, pattern matching and anomaly detection. You can edit queries in the portal or using our development tools, and test them using sample data that is extracted from a live stream.
Note: Stream Analytics is a cost-effective event processing engine that helps uncover real-time insights from devices, sensors, infrastructure, applications and data quickly and easily.
Monitor and manage Stream Analytics resources with Azure PowerShell cmdlets and powershell scripting that execute basic Stream Analytics tasks.
Reference:
https://cloudblogs.microsoft.com/sqlserver/2014/10/29/microsoft-adds-iot-streaming-analytics-data-production-and-workflow-services-to-azure/ https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-introduction

**AliceIS** Highly Voted 👍 4 years ago
appears fourth time
upvoted 6 times

**ZodiaC** Most Recent 🕒 4 years ago
to much duplicated
upvoted 2 times

**NamishBansal** 4 years, 1 month ago
a and d are same answer
upvoted 1 times

**anamaster** 4 years, 2 months ago
c but edge job
upvoted 3 times

Question #41 *Topic 2*

You are designing a real-time stream solution based on Azure Functions. The solution will process data uploaded to Azure Blob Storage.
The solution requirements are as follows:
✑ Support up to 1 million blobs.
✑ Scaling must occur automatically.
✑ Costs must be minimized.
What should you recommend?

A. Deploy the Azure Function in an App Service plan and use a Blob trigger.

B. Deploy the Azure Function in a Consumption plan and use an Event Grid trigger.

C. Deploy the Azure Function in a Consumption plan and use a Blob trigger.

D. Deploy the Azure Function in an App Service plan and use an Event Grid trigger.

**Suggested Answer:** *C*
Create a function, with the help of a blob trigger template, which is triggered when files are uploaded to or updated in Azure Blob storage.
You use a consumption plan, which is a hosting plan that defines how resources are allocated to your function app. In the default Consumption Plan, resources are added dynamically as required by your functions. In this serverless hosting, you only pay for the time your functions run.
When you run in an App Service plan, you must manage the scaling of your function app.
Reference:
https://docs.microsoft.com/en-us/azure/azure-functions/functions-create-storage-blob-triggered-function

*Community vote distribution*
B (100%)

**sdas1** Highly Voted 4 years, 3 months ago
The solution is B - Deploy the Azure Function in a Consumption plan and use an Event Grid trigger. 1M blobs will cripple the ability of blob trigger to provide the events.
The EventGrid trigger is instantaneous, so it depends on your needs.
upvoted 39 times

**Wendy_DK** Highly Voted 4 years, 1 month ago
It repeated question
upvoted 8 times

**dakku987** Most Recent 1 year, 6 months ago
Selected Answer: B
B. Deploy the Azure Function in a Consumption plan and use an Event Grid trigger.

Consumption Plan: Azure Functions in a Consumption Plan automatically scales based on the number of incoming events. It is a serverless option where you only pay for the actual execution of functions.

Event Grid Trigger: Using an Event Grid trigger allows you to respond to events in Azure Blob Storage, such as new blobs being created. This is a more event-driven and scalable approach compared to polling for changes.

chat gpt
upvoted 1 times

**satyamkishoresingh** 3 years, 8 months ago
repeated . Answer B
upvoted 2 times

**davita8** 4 years, 2 months ago
C. Deploy the Azure Function in a Consumption plan and use a Blob trigger.
upvoted 4 times

**davita8** 4 years, 2 months ago
B. Deploy the Azure Function in a Consumption plan and use an Event Grid trigger.
upvoted 2 times

⊟ **karma_wins** 4 years, 2 months ago

"B" is correct because https://docs.microsoft.com/en-us/azure/azure-functions/functions-bindings-storage-blob-trigger?tabs=csharp#event-grid-trigger

upvoted 4 times

⊟ **cadio30** 4 years, 1 month ago

Agree with the propose solution as the url states the scenario to utilize event trigger

upvoted 1 times

⊟ **v_gul** 4 years, 2 months ago

The solution is D - Deploy the Azure Function in an App Service plan and use an Event Grid trigger.

App Service plan - ... If you need low latency in your blob triggered functions, consider running your function app in an App Service plan. SOURCE -> https://docs.microsoft.com/en-us/azure/azure-functions/functions-create-storage-blob-triggered-function

Event Grid trigger - ... The Event Grid trigger also has built-in support for blob events. Use Event Grid instead of the Blob storage trigger for the following scenarios:

...

High-scale: High scale can be loosely defined as containers that have more than 100,000 blobs in them or storage accounts that have more than 100 blob updates per second.

SOURCE - > https://docs.microsoft.com/en-gb/azure/azure-functions/functions-bindings-storage-blob-trigger?tabs=csharp

upvoted 2 times

⊟ **Apox** 4 years, 2 months ago

Since minimizing costs is a requirement, and low latency is not, the correct answer should be a "consumption plan". Otherwise I agree with Event Grid Trigger, as this should be used if there are more than 100 000 blobs.

upvoted 2 times

⊟ **H_S** 4 years, 3 months ago

When your function app runs in the default Consumption plan, there may be a delay of up to several minutes between the blob being added or updated and the function being triggered. If you need low latency in your blob triggered functions, consider running your function app in an App Service plan.

upvoted 2 times

⊟ **szpinat** 4 years, 3 months ago

"You can also use an Event Grid trigger with your Blob storage account."

https://docs.microsoft.com/en-gb/azure/azure-functions/functions-bindings-storage-blob-trigger?tabs=csharp

upvoted 1 times

Question #42 *Topic 2*

You plan to migrate data to Azure SQL Database.
The database must remain synchronized with updates to Microsoft Azure and SQL Server.
You need to set up the database as a subscriber.
What should you recommend?

A. Azure Data Factory

B. SQL Server Data Tools

C. Data Migration Assistant

D. SQL Server Agent for SQL Server 2017 or later

E. SQL Server Management Studio 17.9.1 or later

**Suggested Answer:** *E*
To set up the database as a subscriber we need to configure database replication. You can use SQL Server Management Studio to configure replication. Use the latest versions of SQL Server Management Studio in order to be able to use all the features of Azure SQL Database.
Reference:
https://www.sqlshack.com/sql-server-database-migration-to-azure-sql-database-using-sql-server-transactional-replication/

*Community vote distribution*
D (100%)

⊟ **extraego** Highly Voted 4 years, 10 months ago
The answer E is correct. It's not asking how to migrate but what tool to use for setting up a transactional replication. "Replication can be configured by using SQL Server Management Studio". https://docs.microsoft.com/en-us/azure/azure-sql/database/replication-to-sql-database
upvoted 29 times

⊟ **cadio30** 4 years, 1 month ago
This is the correct answer as it needs configuration on the SSMS side to perform the replication

Reference: https://www.sqlshack.com/sql-server-database-migration-to-azure-sql-database-using-sql-server-transactional-replication/
upvoted 1 times

⊟ **AhmedReda** Highly Voted 5 years ago
I think he means the updated versions of the services and tools not the data
upvoted 5 times

⊟ **MLCL** 4 years, 12 months ago
Exactly
upvoted 2 times

⊟ **dakku987** Most Recent 1 year, 6 months ago
Selected Answer: D
D. SQL Server Agent for SQL Server 2017 or later

Explanation:

SQL Server Agent: SQL Server Agent is a component of SQL Server that enables the scheduling and automation of administrative tasks, including replication. You can use SQL Server Agent to set up and manage replication subscriptions.
Azure SQL Database supports various types of replication, such as transactional replication or snapshot replication, depending on your synchronization requirements. SQL Server Agent is commonly used for managing replication in SQL Server environments.
upvoted 1 times

⊟ **syu31svc** 4 years, 6 months ago
E is the best answer though I would say SQL Data Sync is the right solution for this question
upvoted 3 times

⊟ **Ikrom** 4 years, 10 months ago

The given answer "E: SQL Server Management Studio 17.9.1 or later" is correct based on the remarks from: https://docs.microsoft.com/en-us/azure/azure-sql/database/replication-to-sql-database.

1. Replication can be configured by using SQL Server Management Studio or by executing Transact-SQL statements on the publisher. You cannot configure replication by using the Azure portal.
2. Replication can only use SQL Server authentication logins to connect to Azure SQL Database.

upvoted 3 times

⊟ **pravinDataSpecialist** 5 years ago

I Believe the answer should be SQL Data Sync which is not there in the options

upvoted 3 times

⊟ **Abhitm** 5 years, 1 month ago

I believe the answer should be B (SQL Data sync tool)
https://docs.microsoft.com/en-us/azure/azure-sql/database/sql-data-sync-data-sql-server-sql-database

upvoted 3 times

⊟ **drdean** 5 years ago

That's now what B says but were it an option I agree SQL Data Sync would be correct

upvoted 3 times

⊟ **extraego** 4 years, 10 months ago

SQL Server Data Tools (SSDT) allows Visual Studio to create a SQL Server database project. It is nothing to do with migration/replication. It's a development tool.

upvoted 3 times

HOTSPOT -

You are designing a solution for a company. You plan to use Azure Databricks.

You need to recommend workloads and tiers to meet the following requirements:

✑ Provide managed clusters for running production jobs.

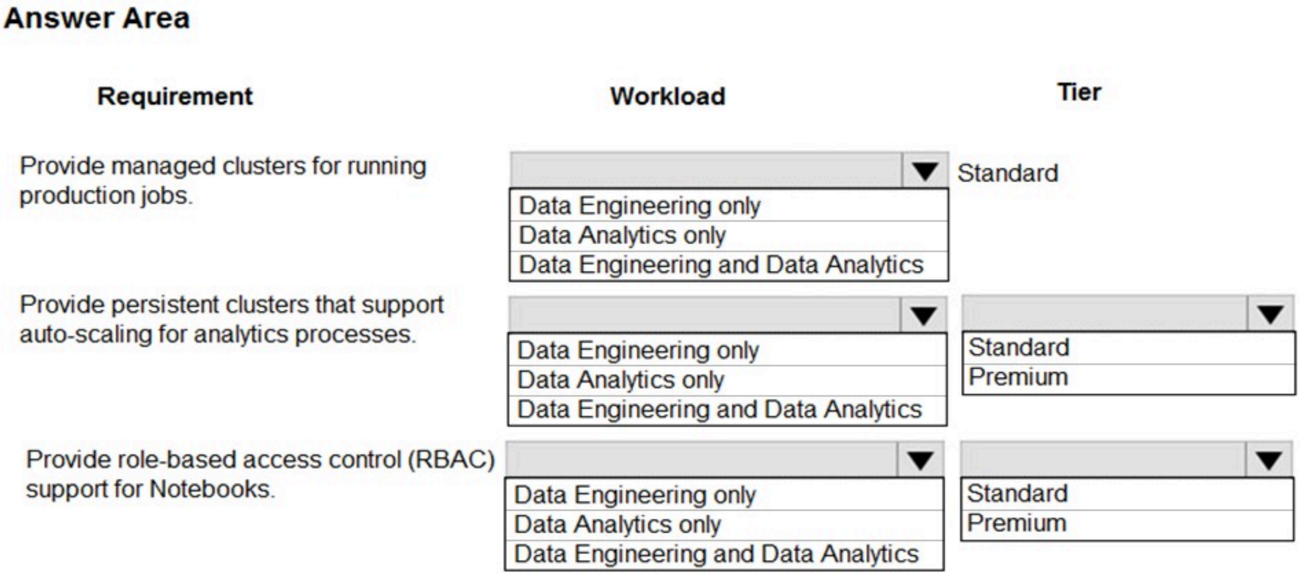
✑ Provide persistent clusters that support auto-scaling for analytics processes.

✑ Provide role-based access control (RBAC) support for Notebooks.

What should you recommend? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

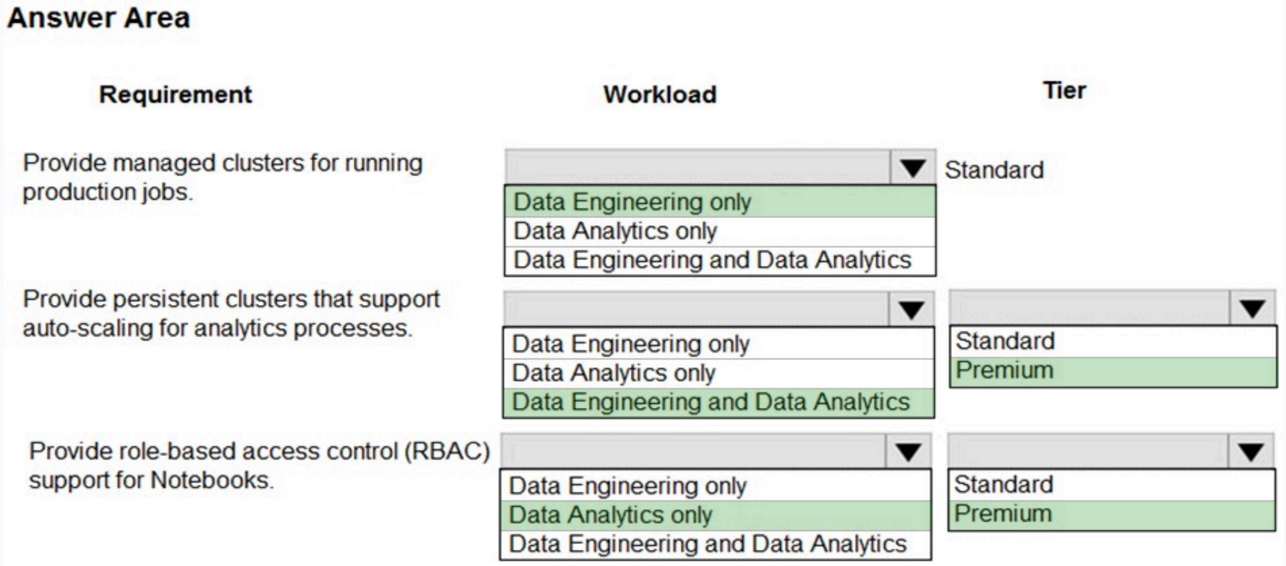
Hot Area:



**Answer Area**

| Requirement | Workload | Tier |
|---|---|---|
| Provide managed clusters for running production jobs. | Data Engineering only<br>Data Analytics only<br>Data Engineering and Data Analytics | Standard |
| Provide persistent clusters that support auto-scaling for analytics processes. | Data Engineering only<br>Data Analytics only<br>Data Engineering and Data Analytics | Standard<br>Premium |
| Provide role-based access control (RBAC) support for Notebooks. | Data Engineering only<br>Data Analytics only<br>Data Engineering and Data Analytics | Standard<br>Premium |

**Suggested Answer:**



**Answer Area**

| Requirement | Workload | Tier |
|---|---|---|
| Provide managed clusters for running production jobs. | [✓ Data Engineering only]<br>Data Analytics only<br>Data Engineering and Data Analytics | Standard |
| Provide persistent clusters that support auto-scaling for analytics processes. | Data Engineering only<br>Data Analytics only<br>[✓ Data Engineering and Data Analytics] | Standard<br>[✓ Premium] |
| Provide role-based access control (RBAC) support for Notebooks. | Data Engineering only<br>[✓ Data Analytics only]<br>Data Engineering and Data Analytics | Standard<br>[✓ Premium] |

Box 1: Data Engineering Only -

Box 2: Data Engineering and Data Analytics

Box 3: Standard -

Box 4: Data Analytics only -

Box 5: Premium -

Premium required for RBAC. Data Analytics Premium Tier provide interactive workloads to analyze data collaboratively with notebooks

Reference:

https://azure.microsoft.com/en-us/pricing/details/databricks/

**Filippo** Highly Voted 5 years, 2 months ago

Based on MS documentation (https://azure.microsoft.com/en-us/pricing/details/databricks/) the answers should be: Box 1 = Data Engineering and Data Analytics, Box 2 = Data Analytics only, Box 3 = Standard, Box 4 = Data Engineering and Data Analytics, Box 5 = Premium

upvoted 125 times

**Niteen** 5 years, 1 month ago

It's perfect.

upvoted 4 times

**Niteen** 5 years, 1 month ago

Filippo - It's perfect.

upvoted 2 times

**rohitbinnani** 4 years, 11 months ago

I did validate too. it is perfect.

upvoted 2 times

**D_Duke** 4 years, 8 months ago

Yes, perfect!

upvoted 2 times

**cadio30** 4 years, 1 month ago

Perfect answer with citation of reference. Cheers!

upvoted 1 times

**cadio30** 4 years ago

To justify this solution feel free to check out the url below

Reference: https://www.azure.cn/en-us/pricing/details/databricks/

upvoted 2 times

**ToNiOZ45** Highly Voted 4 years, 12 months ago

After spending too much time on examtopics, I'm a robot!

upvoted 20 times

**hsetin** Most Recent 3 years, 9 months ago

i thought Autoscaling requires premium??

upvoted 1 times

**Bhagya123456** 3 years, 10 months ago

The pricing plans have different names now and many of the features has been combined. So this question is not valid today.

upvoted 2 times

**RThakor** 3 years, 11 months ago

so many updates for same question so which answer is correct one ?

upvoted 1 times

**vrmei** 4 years ago

https://sql-stack.com/2018/11/29/azure-databricks-workloads-and-job-scheduling/

1. Production Job (scheduled operations) - Data Engineering Only, Standard Tier
2. Persistent for analytics, auto-scaling - Data Analytics Only, Standard Tier
3. Role-based access - For both Data Engineering and Data Analytics, Premium Tier

upvoted 1 times

**sturcu** 4 years, 3 months ago

https://azure.microsoft.com/en-us/pricing/details/databricks/
Persistent clusters for analytics is only available for All Purpose(data Analyst) Tier

upvoted 1 times

**ThijsN** 4 years, 5 months ago

Best article I have found that explains this: https://sql-stack.com/2018/11/29/azure-databricks-workloads-and-job-scheduling/#:~:text=The%20Data%20Engineering%20workload%20is,the%20duration%20of%20the%20job.&text=This%20workload%20is%20also%20designe

1. Data engineering and standard (data engineering is meant for jobs, will tear itself down)

2. Data analytics and standard (is meant for ad-hoc analysis, will stay up)

3. Data engineer and analytics and premium (RBAC is a premium feature)

upvoted 7 times

**suman13** 4 years, 2 months ago

perfect

upvoted 1 times

**Ab5381** 4 years, 6 months ago

Box 1: Data Engineering (Jobs Compute) and Data Analytics (All-Purpose Compute)

Box 2: Data Engineering (Jobs Compute) and Data Analytics (All-Purpose Compute)

Box 3: Standard

Box 4: Data Engineering (Jobs Compute) and Data Analytics (All-Purpose Compute)

Box 5: Premium

upvoted 5 times

**littlebear1** 4 years, 6 months ago

the original answer is correct

upvoted 2 times

**syu31svc** 4 years, 6 months ago

https://azure.microsoft.com/en-us/pricing/details/databricks/

Credit to Filippo for the link and mapping of answers to the details in the link

upvoted 1 times

**ttAsh** 4 years, 6 months ago

from an efficiency point of view why cant we need Box 1 be Data Engineering only (Jobs Light Compute), as its less price and does the production job.

upvoted 1 times

**Arsa** 4 years, 10 months ago

Box1 = Data Analytics & Data Engineering

Box2 = Data Analytics & Data Engineering

Box4 = Standard

Box3 = Data Analytics & Data Engineering

Box5 = Premium

upvoted 1 times

**Arsa** 4 years, 10 months ago

ignore this.

Box1 = Data Engineering

Box2 = Data Analytics only

Box4 = Standard

Box3 = Data Analytics only

Box5 = Premium

upvoted 1 times

**M0e** 4 years, 8 months ago

ignore this!

Box 1: Data Engineering and Data Analytics

Box 2: Data Analytics only

Box 3: Standard

Box 4: Data Engineering and Data Analytics

Box 5: Premium

source: https://azure.microsoft.com/en-us/pricing/details/databricks/

upvoted 6 times

**dsyouness** 4 years, 8 months ago

True source : https://databricks.com/fr/product/azure-pricing

upvoted 1 times

**Mittleme** 4 years, 11 months ago

Phew . these original answers are messed up and confusing . even though we have correct answers in the link. the answers provided are totally different

upvoted 3 times

**User27069** 4 years, 7 months ago

https://azure.microsoft.com/en-us/pricing/details/databricks/

Reading this link, it looks like Data Analytics has been rebranded to "All purpose Compute", and Data Engineering to "Jobs Compute". In which case, the answers given by wak are correct.

upvoted 5 times

**wak** 4 years, 11 months ago

Box1 = Data Analytics & Data Engineering

Box2 = Data Analytics

Box3 = Data Analytics & Data Engineering

Box4 = Standard

Box5 = Premium

As per https://azure.microsoft.com/en-us/pricing/details/databricks/

upvoted 11 times

**vvt** 5 years ago

As per https://azure.microsoft.com/en-us/pricing/details/databricks/ box 1,2,4 should be Data Engineering and Data Analytics, 5,4=Premium

upvoted 1 times

**AhmedReda** 5 years ago

I think Box 4 = Standard as mentioned in the table and per @Filippo

Table header is Standard tier features in the below link

==============================================

https://azure.microsoft.com/en-us/pricing/details/databricks/

upvoted 1 times

**Treadmill** 4 years, 10 months ago

Persistent clusters for analytics is a standard tier (box 4) feature for data analytics (box 3) workload. Filippo has the provided the correct answer.

upvoted 1 times

**drdean** 5 years ago

https://databricks.com/product/azure-pricing

upvoted 1 times

**drdean** 5 years ago

Box 2 should be premium for - Optimized autoscaling of compute

upvoted 2 times

**mohowzeh** 4 years, 5 months ago

Box 3 can be Standard. See the "Autopilot" option in the Standard plan on this link: https://databricks.com/product/azure-pricing

upvoted 1 times

## Question #44 *Topic 2*

You design data engineering solutions for a company.

A project requires analytics and visualization of large set of data. The project has the following requirements:

✑ Notebook scheduling

✑ Cluster automation

✑ Power BI Visualization

You need to recommend the appropriate Azure service. Your solution must minimize the number of services required.

Which Azure service should you recommend?

A. Azure Batch

B. Azure Stream Analytics

C. Azure Databricks

D. Azure HDInsight

**Suggested Answer:** *C*

A databrick job is a way of running a notebook or JAR either immediately or on a scheduled basis.

Azure Databricks has two types of clusters: interactive and job. Interactive clusters are used to analyze data collaboratively with interactive notebooks. Job clusters are used to run fast and robust automated workloads using the UI or API.

You can visualize Data with Azure Databricks and Power BI Desktop.

Reference:

https://docs.azuredatabricks.net/user-guide/clusters/index.html https://docs.azuredatabricks.net/user-guide/jobs.html

**VG2007** Highly Voted 4 years, 1 month ago

No doubts.. Databricks.. Correct answer!!!

upvoted 9 times

Question #45 *Topic 2*

HOTSPOT -

You design data engineering solutions for a company.

You must integrate on-premises SQL Server data into an Azure solution that performs Extract-Transform-Load (ETL) operations have the following requirements:

✑ Develop a pipeline that can integrate data and run notebooks.

✑ Develop notebooks to transform the data.

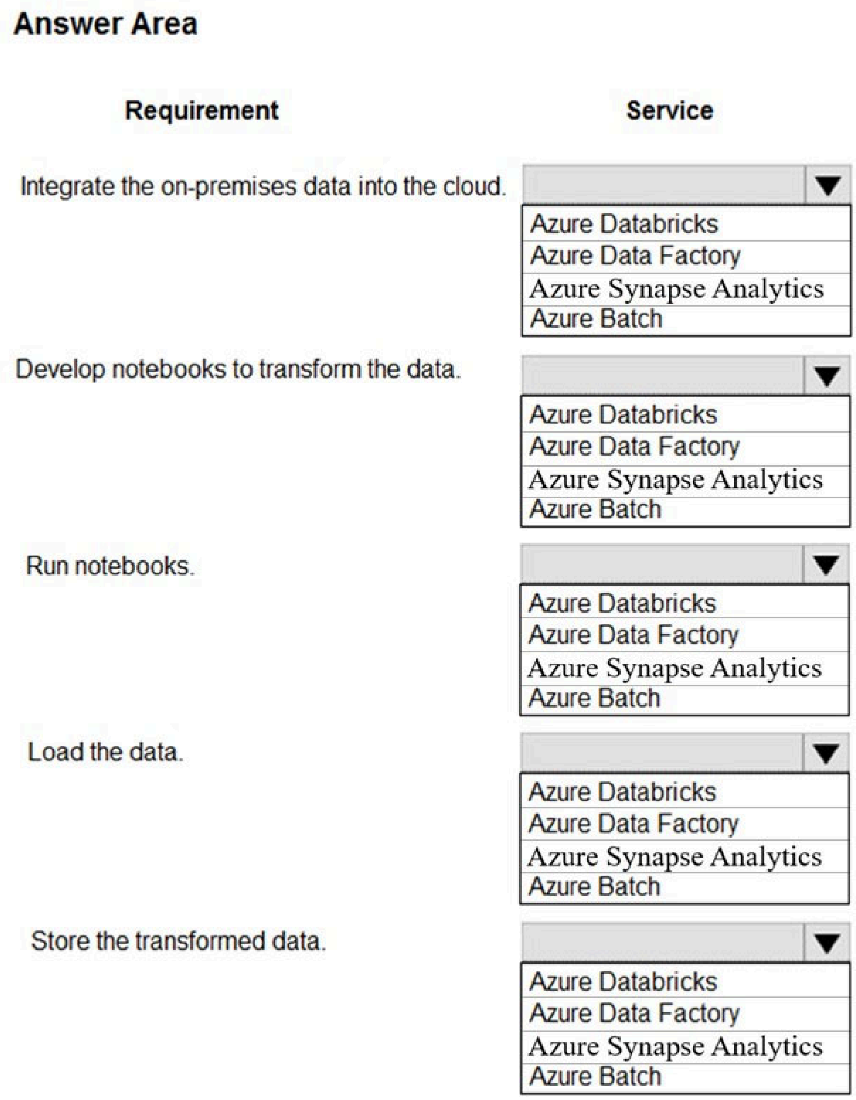
✑ Load the data into a massively parallel processing database for later analysis.

You need to recommend a solution.

What should you recommend? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

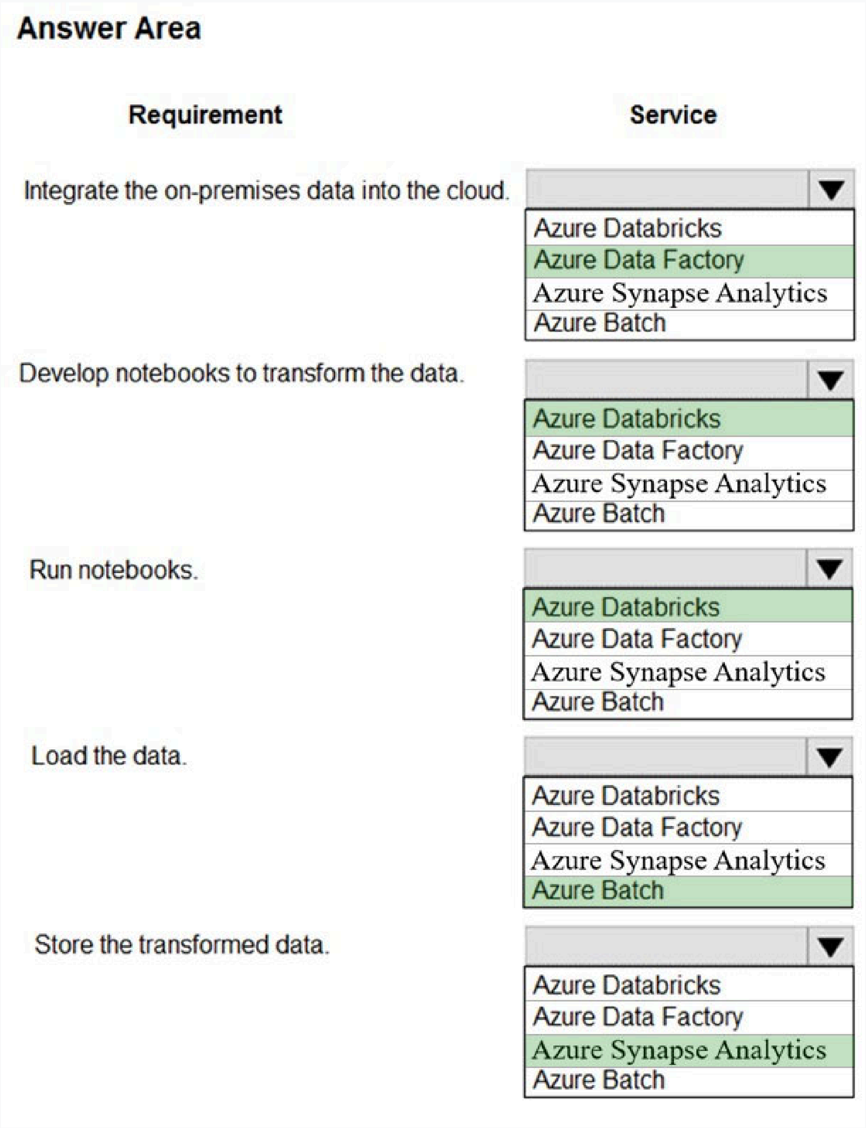
Hot Area:



**Answer Area**

| Requirement | Service |
|---|---|
| Integrate the on-premises data into the cloud. | Azure Databricks<br>Azure Data Factory<br>Azure Synapse Analytics<br>Azure Batch |
| Develop notebooks to transform the data. | Azure Databricks<br>Azure Data Factory<br>Azure Synapse Analytics<br>Azure Batch |
| Run notebooks. | Azure Databricks<br>Azure Data Factory<br>Azure Synapse Analytics<br>Azure Batch |
| Load the data. | Azure Databricks<br>Azure Data Factory<br>Azure Synapse Analytics<br>Azure Batch |
| Store the transformed data. | Azure Databricks<br>Azure Data Factory<br>Azure Synapse Analytics<br>Azure Batch |

Suggested Answer:



## Answer Area

| Requirement | Service |
|---|---|
| Integrate the on-premises data into the cloud. | Azure Databricks<br>[✓ Azure Data Factory]<br>Azure Synapse Analytics<br>Azure Batch |
| Develop notebooks to transform the data. | [✓ Azure Databricks]<br>Azure Data Factory<br>Azure Synapse Analytics<br>Azure Batch |
| Run notebooks. | [✓ Azure Databricks]<br>Azure Data Factory<br>Azure Synapse Analytics<br>Azure Batch |
| Load the data. | Azure Databricks<br>Azure Data Factory<br>Azure Synapse Analytics<br>[✓ Azure Batch] |
| Store the transformed data. | Azure Databricks<br>Azure Data Factory<br>[✓ Azure Synapse Analytics]<br>Azure Batch |

**Needium** Highly Voted 4 years, 3 months ago

I would rather have

Integrate on premises data to Cloud : ADF

Develop notebooks to Transform Data : DataBricks

Run Notebooks : ADF (Azure Databricks notebooks can be run within an ADF pipeline)

Load the Data : Use ADF to load the Data

Store the Transformed Data: Azure Synapse Analyses

upvoted 29 times

**maciejt** 4 years, 2 months ago

Exactly that was my take before seeing the solution.

upvoted 1 times

**cadio30** 4 years, 1 month ago

Azure databricks can handle the loading of data from the notebook to the external tables of Azure Synapse unless the requirement is explicitly to export the file to another storage then use of ADF is the appropriate

upvoted 2 times

**Wendy_DK** Highly Voted 4 years, 1 month ago

Given answer is right.

Remember requirement: Load the data into a massively parallel processing database for later analysis.

ADF and Batch can work together.

ref: https://docs.microsoft.com/en-us/azure/data-factory/v1/data-factory-data-processing-using-batch

upvoted 8 times

**BobFar** 4 years ago

I am agree with you.

upvoted 1 times

**Bhagya123456** Most Recent 3 years, 10 months ago

Given Solution is 100% Correct.

Do not confuse people with absurd arguments. I can do all the activities through Synapse Analysis also. That doesn't mean I will choose 5 times Synapse Analyses.

upvoted 2 times

**tes** 4 years ago

Just one change Run notebook is better done from ADF as we can orchestrate the sequence better. When run from databricks, it may not know the time of data retrieveal and also the next step, Azure Batch cannot be called from ADB

upvoted 1 times

**Ous01** 4 years, 1 month ago

Why note using Databricks to load the data? When the notebook finishes the process, it also can load the data into Synapse. Databricks can easily uploads results to Synapse, Azure SQL, and Azure Cosmos DB.

upvoted 3 times

**VG2007** 4 years, 1 month ago

Given Solution is correct.. no confusions..

why anyone will use ADB to develop notebook and then use ADF to run them unless it is specifically specified ?

upvoted 4 times

> **Larrave** 3 years, 7 months ago
>
> Because they were asking for a Data Engineering solution and having everything handled within one orchestration/etl tool makes definitely sense.
>
> upvoted 1 times

**davita8** 4 years, 2 months ago

Load the data - Azure data factory

transformed data-azure sql data warehouse

upvoted 3 times

**aditya_064** 4 years, 2 months ago

Shouldn't Load the data (Box 4) be Azure Synapse Analytics ? It's the only one with a MPP engine, which is exactly what is mentioned in the question

upvoted 2 times

**maciejt** 4 years, 2 months ago

Why Azure Batch is better than ADF to load data?

ADF could be used to: Integrate from on-prem to azure, invoke notebook (developed in data bricks), then load data into warehouse, all within one pipeline.

upvoted 1 times

> **BobFar** 4 years ago
>
> I guess for loading the data into a massively parallel processing database , azure data batch is the better solution.
>
> https://docs.microsoft.com/en-us/azure/data-factory/v1/data-factory-data-processing-using-batch
>
> upvoted 1 times

**Geo_Barros** 4 years, 3 months ago

Regarding loading the data, I think Azure Data Factory could also be an appropriate answer.

upvoted 3 times

**H_S** 4 years, 3 months ago

azure data factory could be used to load the data too

upvoted 3 times

Question #46 *Topic 2*

A company plans to use Apache Spark Analytics to analyze intrusion detection data.
You need to recommend a solution to analyze network and system activities for malicious activities and policy violations. The solution must minimize administrative efforts.
What should you recommend?

A. Azure Data Factory

B. Azure Data Lake Storage

C. Azure Databricks

D. Azure HDInsight

**Suggested Answer:** *D*
With Azure HDInsight you can set up Azure Monitor alerts that will trigger when the value of a metric or the results of a query meet certain conditions. You can condition on a query returning a record with a value that is greater than or less than a certain threshold, or even on the number of results returned by a query. For example, you could create an alert to send an email if a Spark job fails or if a Kafka disk usage becomes over 90 percent full.
Reference:
https://azure.microsoft.com/en-us/blog/monitoring-on-azure-hdinsight-part-4-workload-metrics-and-logs/

*Community vote distribution*
C (100%)

**suman13** Highly Voted 4 years, 2 months ago
this is typical use case for Azure databricks that can use spark analytics paltform. so the answer should be databricks
upvoted 28 times

**rahul_t** 4 years, 2 months ago
I agree: https://azure.microsoft.com/es-es/blog/three-critical-analytics-use-cases-with-microsoft-azure-databricks/
upvoted 4 times

**cadio30** 4 years, 1 month ago
Second agree with the propose solution
upvoted 1 times

**dakku987** Most Recent 1 year, 6 months ago
**Selected Answer: C**
Azure Databricks: Azure Databricks is a fast, easy, and collaborative Apache Spark-based analytics platform. It simplifies the deployment and management of Apache Spark clusters, making it an excellent choice for large-scale data processing, including analyzing network and system activities for security purposes. It integrates with various Azure services and provides collaborative notebooks for data scientists and analysts.
upvoted 1 times

**Bhagya123456** 3 years, 10 months ago
HDInsight is not in syllabus. So anytime you have a confusion with Databricks and HDInsight better go with Databricks.
upvoted 2 times

**satyamkishoresingh** 3 years, 10 months ago
I will go with the databricks for the analytics of intrusion data!
upvoted 2 times

**BobFar** 4 years ago
Answer should be Azure databricks.
Intrusion detection is one of use case for Azure databricks.
https://azure.microsoft.com/es-es/blog/three-critical-analytics-use-cases-with-microsoft-azure-databricks/
upvoted 2 times

**dbdev** 4 years, 1 month ago
The answer provided is correct.
upvoted 1 times

**Dymize** 4 years, 1 month ago

are you sure?

upvoted 1 times

⊟ dbdev 4 years ago

nope.. it's Databricks for sure

upvoted 2 times

Question #47 *Topic 2*

You are planning a streaming data solution that will use Azure Databricks. The solution will stream sales transaction data from an online store.
The solution has the following specifications:
✑ The output data will contain items purchased, quantity, line total sales amount, and line total tax amount.
✑ Line total sales amount and line total tax amount will be aggregated in Databricks.
✑ Sales transactions will never be updated. Instead, new rows will be added to adjust a sale.
You need to recommend an output mode for the dataset that will be processed by using Structured Streaming. The solution must minimize duplicate data.
What should you recommend?

A. Append

B. Complete

C. Update

**Suggested Answer:** *A*
Append Mode: Only new rows appended in the result table since the last trigger are written to external storage. This is applicable only for the queries where existing rows in the Result Table are not expected to change.
Incorrect Answers:
B: Complete Mode: The entire updated result table is written to external storage. It is up to the storage connector to decide how to handle the writing of the entire table.
C: Update Mode: Only the rows that were updated in the result table since the last trigger are written to external storage. This is different from Complete Mode in that Update Mode outputs only the rows that have changed since the last trigger. If the query doesn't contain aggregations, it is equivalent to Append mode.
Reference:
https://docs.microsoft.com/en-us/azure/databricks/getting-started/spark/streaming

**toandm** Highly Voted 4 years, 1 month ago
Same question in DP 200
upvoted 5 times

**lorenzoV** Most Recent 2 years, 10 months ago
there will be a new line for each updated transaction (record). So 'append' is correct
upvoted 1 times

**nefarious_smalls** 3 years, 1 month ago
I think the answer is correct because it says data will be aggregated using Databricks. As far as the streaming mode The data is only being appended. Aggregations will be calculated separately.
upvoted 1 times

**MayankSh** 4 years ago
Sales transactions will never be updated --> No updates meaning, no need to perform merge operation or updates, Hence append is the correct answer
upvoted 1 times

**erssiws** 4 years ago
The required conditions are confusing:
condition2-> update
condition3-> append
upvoted 2 times

**maynard13x8** 4 years, 2 months ago
I think it should be update because of the possible new additions of new data to already copied rows. Any opinions?
upvoted 1 times

**maynard13x8** 4 years, 2 months ago
sorry, I haven't read third condition. I think answer is correct.
upvoted 5 times

Question #48 Topic 2

HOTSPOT -

You have an Azure event hub named retailhub that has 16 partitions. Transactions are posted to retailhub. Each transaction includes the transaction ID, the individual line items, and the payment details. The transaction ID is used as the partition key.

You are designing an Azure Stream Analytics job to identify potentially fraudulent transactions at a retail store. The job will use retailhub as the input. The job will output the transaction ID, the individual line items, the payment details, a fraud score, and a fraud indicator.

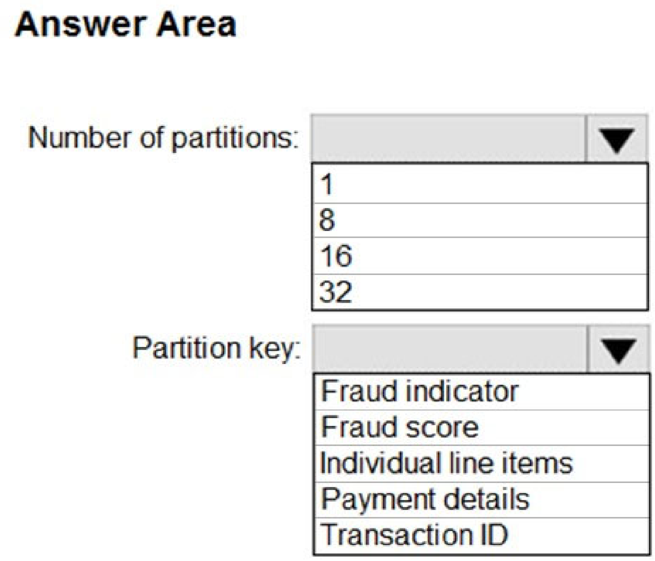
You plan to send the output to an Azure event hub named fraudhub.

You need to ensure that the fraud detection solution is highly scalable and processes transactions as quickly as possible.

How should you structure the output of the Stream Analytics job? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:



Answer Area

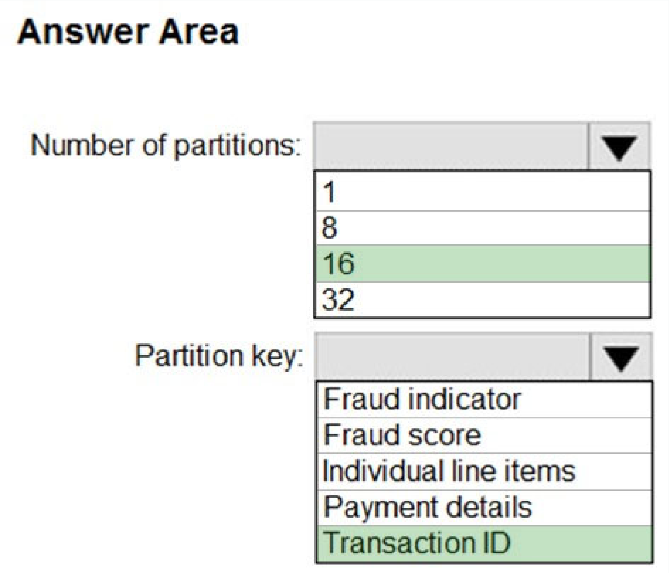
Number of partitions: ▼

1
8
16
32

Partition key: ▼

Fraud indicator
Fraud score
Individual line items
Payment details
Transaction ID

**Suggested Answer:**



Answer Area

Number of partitions: ▼

1
8
[✓ 16]
32

Partition key: ▼

Fraud indicator
Fraud score
Individual line items
Payment details
[✓ Transaction ID]

Box 1: 16 -

For Event Hubs you need to set the partition key explicitly.

An embarrassingly parallel job is the most scalable scenario in Azure Stream Analytics. It connects one partition of the input to one instance of the query to one partition of the output.

Box 2: Transaction ID -

Reference:

https://docs.microsoft.com/en-us/azure/event-hubs/event-hubs-features#partitions

**IAMKPR** Highly Voted 4 years, 1 month ago

Given answer and explanation are correct.

upvoted 9 times

Question #49 *Topic 2*

DRAG DROP -

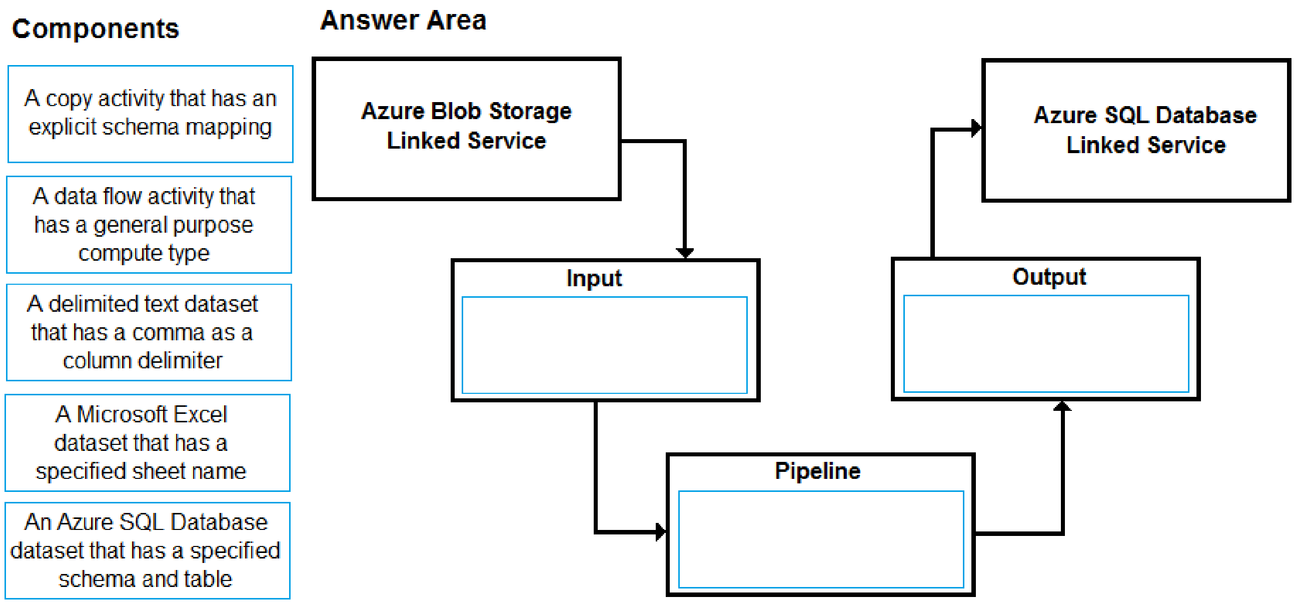
You have a CSV file in Azure Blob storage. The file does NOT have a header row.

You need to use Azure Data Factory to copy the file to an Azure SQL database. The solution must minimize how long it takes to copy the file.

How should you configure the copy process? To answer, drag the appropriate components to the correct locations. Each component may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point.

Select and Place:



Components

- A copy activity that has an explicit schema mapping
- A data flow activity that has a general purpose compute type
- A delimited text dataset that has a comma as a column delimiter
- A Microsoft Excel dataset that has a specified sheet name
- An Azure SQL Database dataset that has a specified schema and table

Answer Area

Azure Blob Storage Linked Service

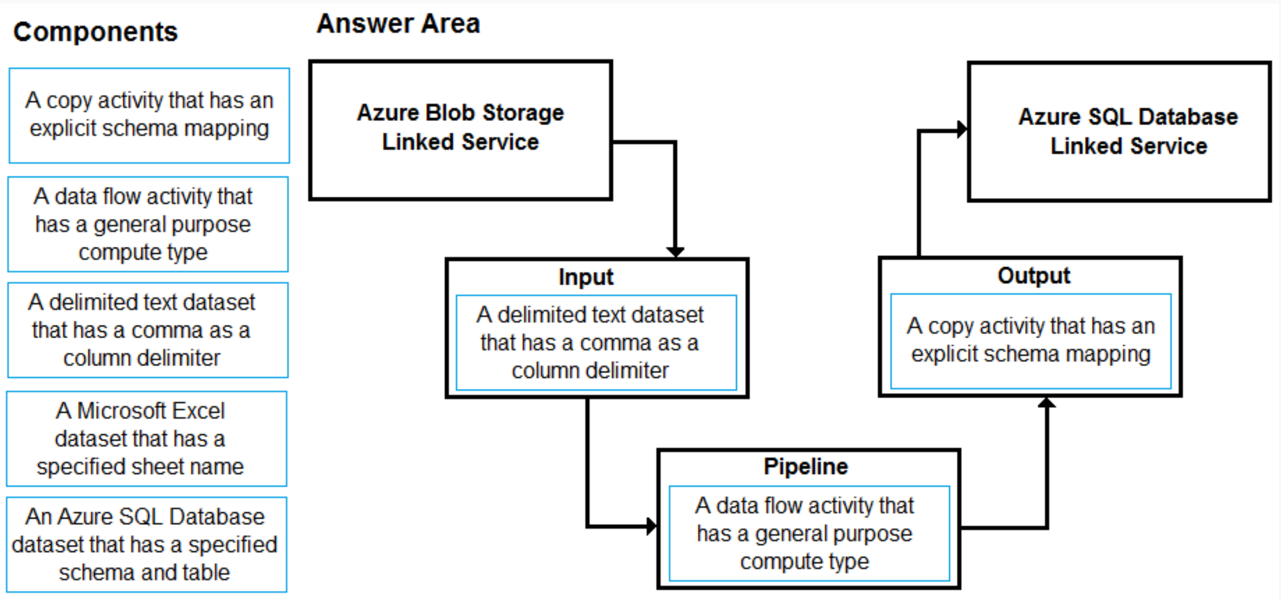
Input

Pipeline

Output

Azure SQL Database Linked Service

**Suggested Answer:**



Components

- A copy activity that has an explicit schema mapping
- A data flow activity that has a general purpose compute type
- A delimited text dataset that has a comma as a column delimiter
- A Microsoft Excel dataset that has a specified sheet name
- An Azure SQL Database dataset that has a specified schema and table

Answer Area

Azure Blob Storage Linked Service

Input: A delimited text dataset that has a comma as a column delimiter

Pipeline: A data flow activity that has a general purpose compute type

Output: A copy activity that has an explicit schema mapping

Azure SQL Database Linked Service

Input: A delimited text dataset that has a comma a column delimiter columnDelimiter: The character(s) used to separate columns in a file. The default value is comma ,. When the column delimiter is defined as empty string, which means no delimiter, the whole line is taken as a single column.

Pipeline: A data flow activity that has a general purpose compute type

When you're transforming data in mapping data flows, you can read and write files from Azure Blob storage.

Output: A copy activity that has an explicit schema mapping

Use Copy Activity in Azure Data Factory to copy data from and to Azure SQL Database, and use Data Flow to transform data in Azure SQL Database.

Reference:

https://docs.microsoft.com/en-us/azure/data-factory/format-delimited-text https://docs.microsoft.com/en-us/azure/data-factory/connector-azure-sql-database

**suman13** Highly Voted 4 years, 2 months ago

2n box: copy activity

3rd box: azure sqldb dataset

upvoted 62 times

**AngelRio** Highly Voted 4 years ago

Input: A delimited text dataset...

Output: Azure SQL DB dataset...

Pipeline: Copy Activity ....

upvoted 20 times

**BitchNigga** Most Recent 4 years ago

I have performed this activity during my course. I am 200% sure that second one is copy activity an third one is sql db with fixed schema

upvoted 9 times

**cadio30** 4 years, 1 month ago

First layer corresponds to linked services while the second layer is for the dataset and lastly the pipeline level. Therefore second layer is CSV and Azure SQL Database then copy activity.

upvoted 1 times

**maciejt** 4 years, 2 months ago

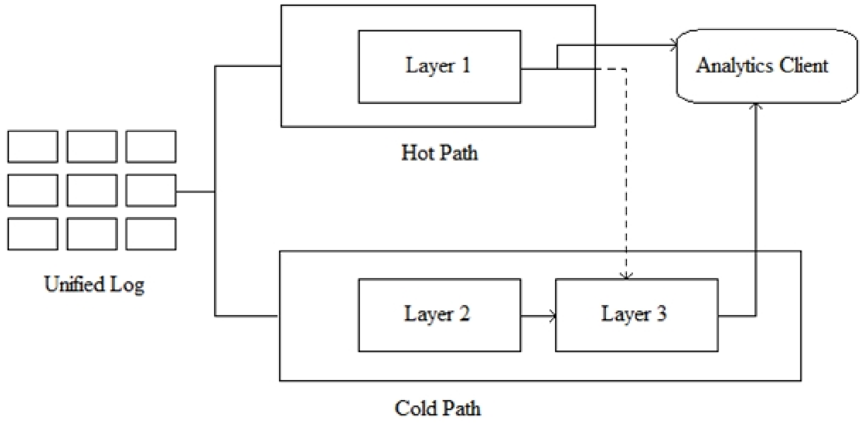
This is completely wrong. Both middle boxes are an abstraction layer, so if left box is a dataset of input, then right box is a dataset for output. There is no requirement to transform the data, only copy, so pipeline consists only of copy activity. If we were using data flow to copy, then we would not need copy activity, because data flow could copy directly to sql database, but requirement is performance and data flow need to start up the cluster that it is run at, while copy activity works instantly.

upvoted 7 times

## Question #50 *Topic 2*

DRAG DROP -

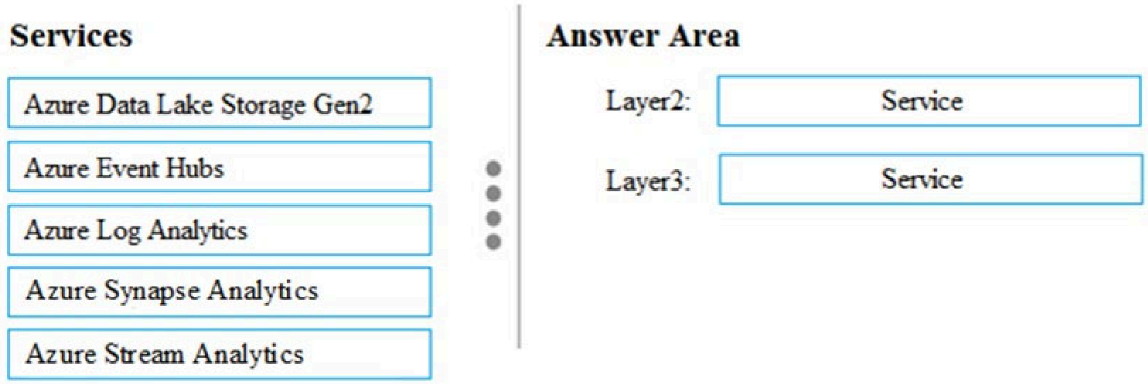
You are planning a design pattern based on the Lambda architecture as shown in the exhibit.



Which Azure services should you use for the cold path? To answer, drag the appropriate services to the correct layers. Each service may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point.

Select and Place:



**Services**

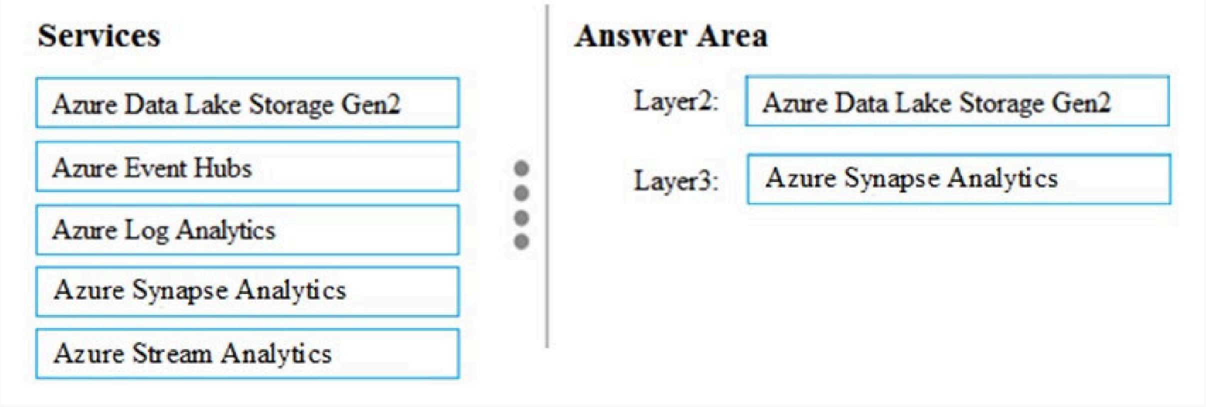
- Azure Data Lake Storage Gen2
- Azure Event Hubs
- Azure Log Analytics
- Azure Synapse Analytics
- Azure Stream Analytics

**Answer Area**

Layer2: Service

Layer3: Service

**Suggested Answer:**



**Services**

- Azure Data Lake Storage Gen2
- Azure Event Hubs
- Azure Log Analytics
- Azure Synapse Analytics
- Azure Stream Analytics

**Answer Area**

Layer2: Azure Data Lake Storage Gen2

Layer3: Azure Synapse Analytics

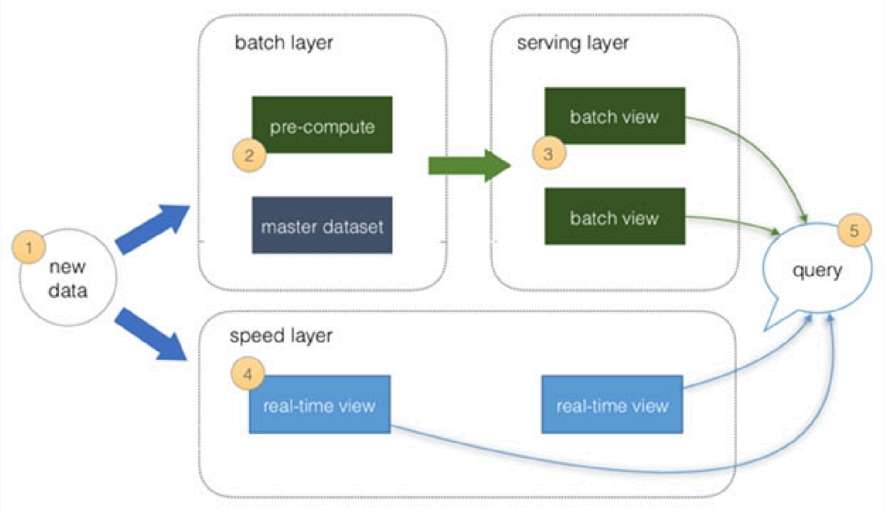
Layer 2: Azure Data Lake Storage Gen2

Layer 3: Azure Synapse Analytics

Azure Synapse Analytics can be used for batch processing.

Note: Layer 1 = speed layer, layer 2 = batch layer, layer 3 = serving layer

Note 2: Lambda architectures use batch-processing, stream-processing, and a serving layer to minimize the latency involved in querying big data.



Reference:

https://azure.microsoft.com/en-us/blog/lambda-architecture-using-azure-cosmosdb-faster-performance-low-tco-low-devops/

https://docs.microsoft.com/en-us/azure/architecture/data-guide/technology-choices/batch-processing

⊟ **Anagarika** Highly Voted 4 years, 3 months ago

Correct answer

upvoted 8 times

⊟ **cadio30** 4 years, 1 month ago

Entirely correct

upvoted 2 times

⊟ **Srinivasnaveen** Most Recent 4 years ago

Layer 2 : Azure data lake GEN2

Layer 3: ASA

upvoted 2 times

⊟ **Srinivasnaveen** 4 years ago

Sorry ASA : Azure Synapse Analytics

upvoted 1 times

Question #51 *Topic 2*

You need to recommend an Azure Cosmos DB solution that meets the following requirements:

✑ All data that was NOT modified during the last 30 days must be purged automatically.

✑ The solution must NOT affect ongoing user requests.

What should you recommend using to purge the data?

A. an Azure Cosmos DB stored procedure executed by an Azure logic app

B. an Azure Cosmos DB REST API Delete Document operation called by an Azure function

C. Time To Live (TTL) setting in Azure Cosmos DB

D. an Azure Cosmos DB change feed queried by an Azure function

**Suggested Answer:** *C*

Reference:

https://docs.microsoft.com/en-us/azure/cosmos-db/time-to-live

**memo43** Highly Voted 👍 4 years, 1 month ago

answer is CORRECT

upvoted 7 times

Question #52 *Topic 2*

You are planning a solution to aggregate streaming data that originates in Apache Kafka and is output to Azure Data Lake Storage Gen2. The developers who will implement the stream processing solution use Java.
Which service should you recommend using to process the streaming data?

A. Azure Data Factory

B. Azure Databricks

C. Azure Event Hubs

D. Azure Stream Analytics

**Suggested Answer:** *B*
Reference:
https://docs.microsoft.com/en-us/azure/architecture/data-guide/technology-choices/stream-processing

**memo43** Highly Voted 👍 4 years, 1 month ago
answer is CORRECT
-->> Apache Spark in Azure Databricks
upvoted 5 times

**DingDongSingSong** Most Recent 🕘 3 years, 3 months ago
Stream analytics does not integrate with Kafka. Look at the documentation provided. Therefore, the answer is Databricks
upvoted 2 times

**muni53** 3 years, 9 months ago
is it due to fact that jar can be uploaded and used in adb. Question mentions about java developers. think that is reason to pick adb instead of stream analytics
upvoted 1 times

**YLiu** 3 years, 9 months ago
Why not Azure Stream Analytics?
upvoted 1 times

Question #53 Topic 2

DRAG DROP -

Your company has a custom human resources (HR) app named App1 that is deployed to mobile devices.

You are designing a solution that will use real-time metrics to indicate how employees use App1. The solution must meet the following requirements:

✑ Use hierarchy data exported monthly from the company's HR enterprise application.

✑ Process approximately 1 GB of telemetry data per day.

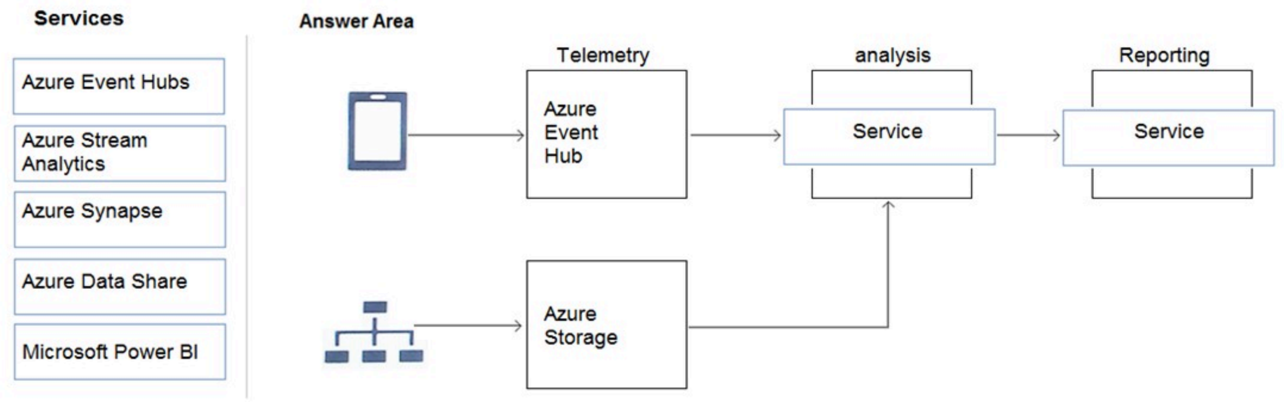
✑ Minimize costs.

You need to recommend which Azure services to use to meet the requirements.

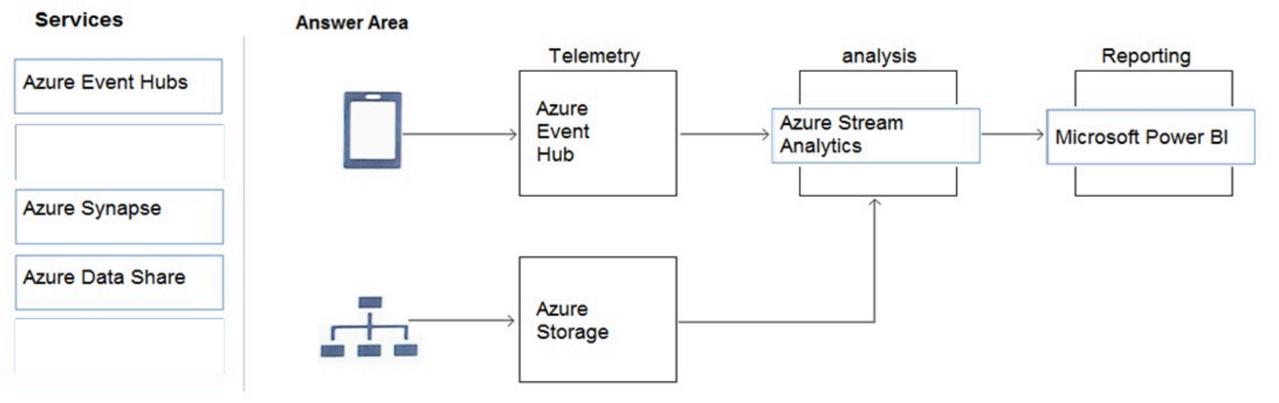
Which two services should you recommend? To answer, drag the appropriate services to the correct targets. Each service may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point.

Select and Place:



**Suggested Answer:**



Reference:

https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-introduction

**erssiws** Highly Voted 4 years ago

the proposed answer makes sense

upvoted 6 times

**DingDongSingSong** Most Recent 3 years, 3 months ago

Why PowerBI and not Synapse Analytics. The question makes no mention of dashboards. Dashboards is the last step. You have to analyze the data before you can produce dashboards. So it should be Stream + Synapse

upvoted 1 times

## Question #54 *Topic 2*

You are planning an Azure solution that will aggregate streaming data.

The input data will be retrieved from tab-separated values (TSV) files in Azure Blob storage.

You need to output the maximum value from a specific column for every two-minute period in near real-time. The output must be written to Blob storage as a

Parquet file.

What should you use?

A. Azure Data Factory and mapping data flows

B. Azure Data Factory and wrangling data flows

C. Azure Stream Analytics window functions

D. Azure Databricks and Apache Spark SQL window functions

**Suggested Answer:** *C*

Reference:

https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-define-outputs#parquet-output-batching-window-properties

**cadio30** Highly Voted 4 years, 1 month ago

Appropriate answer is C as the Azure Stream Analytics can perform aggregation syntax using MAX then output it in blob storage/ADLS Gen2

upvoted 13 times

**erssiws** Most Recent 4 years ago

The Databricks spark SQL by default works with batch mode. It won't fulfill the near real-time requirement

upvoted 1 times

**Rob77** 4 years, 1 month ago

Both ASA and Databricks should be able to do it now

upvoted 1 times

**niwe** 4 years, 1 month ago

I think correct answer is D

upvoted 3 times

**ZodiaC** 4 years ago

Make no SENSE

upvoted 1 times

Question #55 *Topic 2*

You are designing a statistical analysis solution that will use custom proprietary Python functions on near real-time data from Azure Event Hubs.
You need to recommend which Azure service to use to perform the statistical analysis. The solution must minimize latency.
What should you recommend?

A. Azure Synapse Analytics

B. Azure Stream Analytics

C. Azure Databricks

D. Azure SQL Database

**Suggested Answer:** *B*
Reference:
https://docs.microsoft.com/en-us/azure/event-hubs/process-data-azure-stream-analytics
Design for data security and compliance

**Mily94** Highly Voted 👍 4 years, 1 month ago
Shouldn't be C (Databricks)? I think that the keyword is Python.
upvoted 32 times

**cadio30** 4 years, 1 month ago
Provided with the limited option, Azure Databricks is the appropriate solution as it can accommodate Python script
upvoted 3 times

**alf99** Highly Voted 👍 4 years, 1 month ago
https://docs.microsoft.com/en-us/azure/stream-analytics/functions-overview

Azure Stream Analytics supports the following four function types:

JavaScript user-defined functions
JavaScript user-defined aggregates
C# user-defined functions (using Visual Studio)
Azure Machine Learning

Python is not on list so Databricks must be the right choice

Answer: C
upvoted 14 times

**massnonn** Most Recent 🕒 3 years, 7 months ago
why not A)Azure Synapse Analytics? support the python and nearl real time instead databricks is batch, or not?
upvoted 2 times

**deeps1390** 3 years, 8 months ago
Correct ans is B
https://docs.microsoft.com/en-us/azure/event-hubs/event-hubs-python-get-started-send
upvoted 1 times

**Deedubya** 3 years, 10 months ago
I agree it should be Databricks. Nowhere in Microsoft documentation does it say Azre Stream Analytics supports python.
https://docs.microsoft.com/en-us/azure/architecture/data-guide/technology-choices/stream-processing

Python Reference document to show all integrated Azure services also does not list Azure Stream Analytics.
https://docs.microsoft.com/en-us/python/api/overview/azure/?view=azure-python

If someone finds something updated, please post!
upvoted 1 times

**erssiws** 4 years ago

Should be the Databricks which has better support for UDF in python

upvoted 1 times

**Mandar77** 4 years ago

You have to perform statistical analysis that what question says. Means what ever date you get on that you need to do it. I have found the link https://stackoverflow.com/questions/58097539/execute-azure-steaming-analytics-queries-from-a-python-script which talks about executing queries using python. The answer seems correct.

upvoted 1 times

**BitchNigga** 4 years ago

Custom proprietary functions are written manually which need to be packaged and imported so yes databricks

upvoted 1 times

**DragonBlake** 4 years ago

Answer is correct. ASA supports python https://docs.microsoft.com/en-us/python/api/overview/azure/mgmt-streamanalytics-readme?view=azure-python-preview

upvoted 1 times

**Ninja1** 4 years, 1 month ago

Agree, Databricks should be the answer

upvoted 2 times