

Free
PDF

On a Mission to make 100+ Certified Cloud Engineer in next 30 Days

CertyIQ

DP-203

Real Exam Questions & Answers



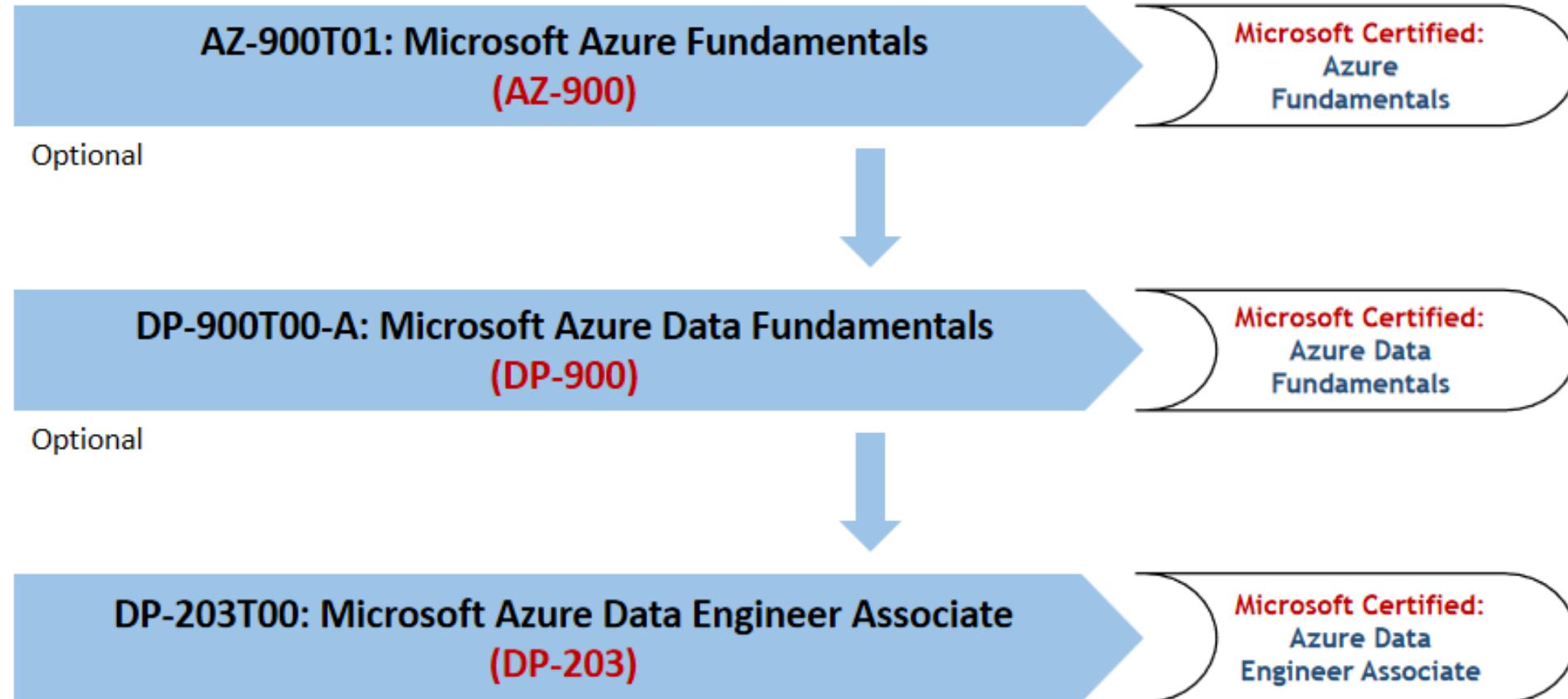
Latest Exam Questions
All Topics Covered-2023
Added New Que'ss

IMPORTANT QUESTIONS



Subscribe

Microsoft Azure Data Engineer Certification Path



Data Engineering on Microsoft Azure (DP-203) Exam Format

Exam Name	Data Engineering on Microsoft Azure
Exam Code	DP-203
Exam Cost	USD 165
Exam Format	Multiple Choice and Multiple Response Questions
Total Questions	40-60 Questions
Passing Score	700 out of 1000
Exam Duration	130 Minutes
Languages	English, Chinese (Simplified), German, French, Spanish, Chinese (Traditional), Italian, Indonesian (Indonesia), Arabic (Saudi Arabia), Russian, Portuguese (Brazil), Japanese, Korean

You have an Azure Synapse workspace named MyWorkspace that contains an Apache Spark database named mytestdb.

You run the following command in an Azure Synapse Analytics Spark pool in MyWorkspace.

```
CREATE TABLE mytestdb.myParquetTable(
```

```
EmployeeID int,
```

```
EmployeeName string,
```

```
EmployeeStartDate date)
```

```
USING Parquet -
```

You then use Spark to insert a row into mytestdb.myParquetTable. The row contains the following data.

EmployeeName	EmployeeID	EmployeeStartDate
Alice	24	2020-01-25

One minute later, you execute the following query from a serverless SQL pool in MyWorkspace.

```
SELECT EmployeeID -
```

```
FROM mytestdb.dbo.myParquetTable
```

```
WHERE EmployeeName = 'Alice';
```

What will be returned by the query?

A. 24

B. an error

C. a null value

CertyIQ@gmail.com



You have an Azure Synapse workspace named MyWorkspace that contains an Apache Spark database named mytestdb.

You run the following command in an Azure Synapse Analytics Spark pool in MyWorkspace.

```
CREATE TABLE mytestdb.myParquetTable(
```

```
EmployeeID int,
```

```
EmployeeName string,
```

```
EmployeeStartDate date)
```

```
USING Parquet -
```

You then use Spark to insert a row into mytestdb.myParquetTable. The row contains the following data.

EmployeeName	EmployeeID	EmployeeStartDate
Alice	24	2020-01-25

One minute later, you execute the following query from a serverless SQL pool in MyWorkspace.

```
SELECT EmployeeID -
```

```
FROM mytestdb.dbo.myParquetTable
```

```
WHERE EmployeeName = 'Alice';
```

What will be returned by the query?

A. 24

B. an error

C. a null value

CertyIQ@gmail.com



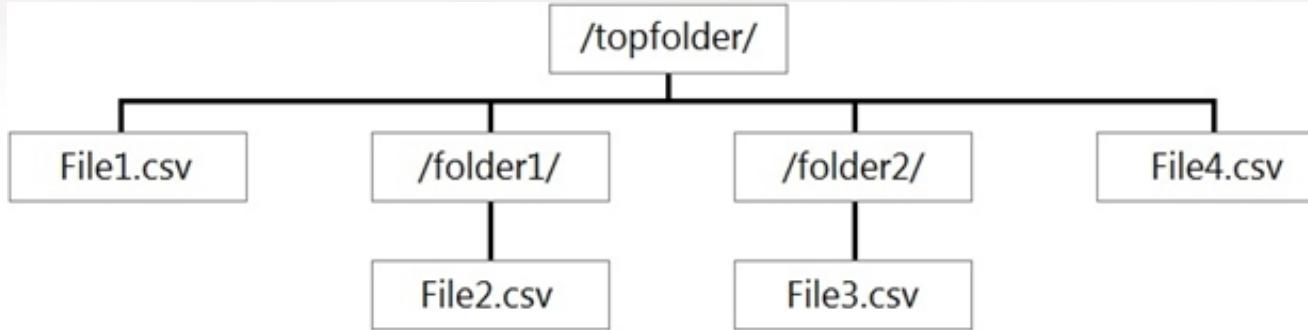
Explanation

Correct Answer: B

It will return an error. table name will be converted to lowercase

You have files and folders in Azure Data Lake Storage Gen2 for an Azure Synapse workspace as shown in the following exhibit.

CertyIQ@gmail.com



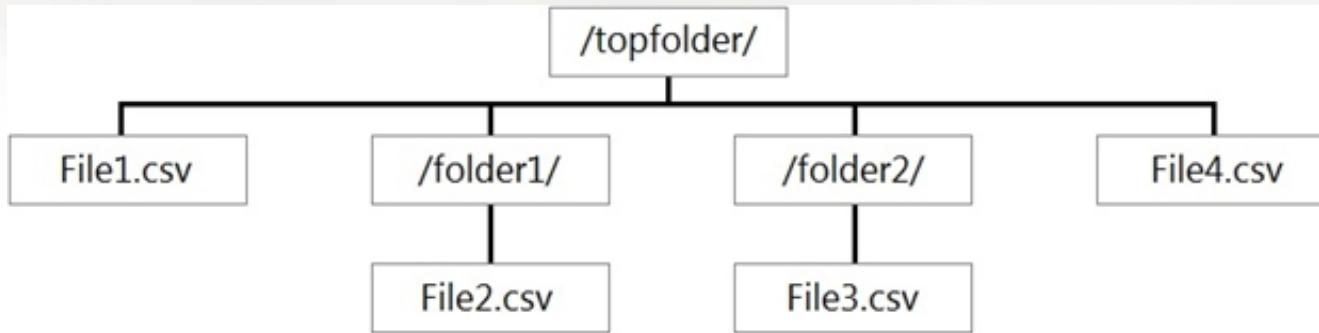
You create an external table named ExtTable that has LOCATION='/topfolder/'.

When you query ExtTable by using an Azure Synapse Analytics serverless SQL pool, which files are returned?

- A. File2.csv and File3.csv only
- B. File1.csv and File4.csv only
- C. File1.csv, File2.csv, File3.csv, and File4.csv
- D. File1.csv only

You have files and folders in Azure Data Lake Storage Gen2 for an Azure Synapse workspace as shown in the following exhibit.

CertyIQ@gmail.com



You create an external table named ExtTable that has LOCATION='/topfolder/'.

When you query ExtTable by using an Azure Synapse Analytics serverless SQL pool, which files are returned?

- A. File2.csv and File3.csv only
- B. File1.csv and File4.csv only**
- C. File1.csv, File2.csv, File3.csv, and File4.csv
- D. File1.csv only

Explanation

Correct Answer: B

In case of a serverless pool a wildcard should be added to the location.

HOTSPOT -

CertyIQ@gmail.com

You are planning the deployment of Azure Data Lake Storage Gen2.

You have the following two reports that will access the data lake:

- ☞ Report1: Reads three columns from a file that contains 50 columns.
- ☞ Report2: Queries a single record based on a timestamp.

You need to recommend in which format to store the data in the data lake to support the reports. The solution must minimize read times.

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

NOTE: Each correct selection is worth one point.

Answer Area

Hot Area:

Report1:

Avro	▼
CSV	
Parquet	
TSV	

Report2:

Avro	▼
CSV	
Parquet	
TSV	

HOTSPOT -

CertyIQ@gmail.com

You are planning the deployment of Azure Data Lake Storage Gen2.

You have the following two reports that will access the data lake:

☞ Report1: Reads three columns from a file that contains 50 columns.

☞ Report2: Queries a single record based on a timestamp.

You need to recommend in which format to store the data in the data lake to support the reports. The solution must minimize read times.

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

NOTE: Each correct selection is worth one point.

Answer Area

Hot Area:

Report1:

Avro
CSV
Parquet
TSV

Report2:

Avro
CSV
Parquet
TSV

Explanation

1: Parquet - column-oriented binary file format

2: AVRO - Row based format, and has logical type timestamp

HOTSPOT -

You need to output files from Azure Data Factory.

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

NOTE: Each correct selection is worth one point.

Hot Area:

CertyIQ@gmail.com

Columnar format:

Avro
GZip
Parquet
TXT

JSON with a timestamp:

Avro
GZip
Parquet
TXT

HOTSPOT -

You need to output files from Azure Data Factory.

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Columnar format:

Avro
GZip
Parquet
TXT

Correct Answer:

JSON with a timestamp:

Avro
GZip
Parquet
TXT

CertyIQ@gmail.com



Explanation

Box 1: Parquet -

Parquet stores data in columns, while Avro stores data in a row-based format. By their very nature, column-oriented data stores are optimized for read-heavy analytical workloads, while row-based databases are best for write-heavy transactional workloads.

Box 2: Avro -

An Avro schema is created using JSON format.

AVRO supports timestamps.

Note: Azure Data Factory supports the following file formats (not GZip or TXT).

HOTSPOT -

CertyIQ@gmail.com

You use Azure Data Factory to prepare data to be queried by Azure Synapse Analytics serverless SQL pools.



Files are initially ingested into an Azure Data Lake Storage Gen2 account as 10 small JSON files. Each file contains the same data attributes and data from a subsidiary of your company. You need to move the files to a different folder and transform the data to meet the following requirements:

- Provide the fastest possible query times.
- Automatically infer the schema from the underlying files.

How should you configure the Data Factory copy activity? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area: **Answer Area**

Copy behavior:

Flatten hierarchy
Merge files
Preserve hierarchy

Sink file type:

CSV
JSON
Parquet
TXT

HOTSPOT -

CertyIQ@gmail.com

You use Azure Data Factory to prepare data to be queried by Azure Synapse Analytics serverless SQL pools.

Files are initially ingested into an Azure Data Lake Storage Gen2 account as 10 small JSON files. Each file contains the same data attributes and data from a subsidiary of your company. You need to move the files to a different folder and transform the data to meet the following requirements:

- Provide the fastest possible query times.
- Automatically infer the schema from the underlying files.

How should you configure the Data Factory copy activity? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Copy behavior:

Flatten hierarchy
Merge files
Preserve hierarchy

Sink file type:

CSV
JSON
Parquet
TXT

Explanation

Correct Answer:

1. Merge Files

2 Parquet -

Azure Data Factory parquet format is supported for Azure Data Lake Storage Gen2.

Parquet supports the schema property.

Question 6

DRAG DROP -

You need to create a partitioned table in an Azure Synapse Analytics dedicated SQL pool.

How should you complete the Transact-SQL statement? To answer, drag the appropriate values to the correct targets. Each value 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:

CertyIQ@gmail.com

Values

CLUSTERED INDEX
COLLATE
DISTRIBUTION
PARTITION
PARTITION FUNCTION
PARTITION SCHEME

Answer Area

```
CREATE TABLE table1
(
    ID INTEGER,
    col1 VARCHAR(10),
    col2 VARCHAR(10)
) WITH
(
    [ ] = HASH(ID),
    [ ] (ID RANGE LEFT FOR VALUES (1, 1000000, 2000000))
);
```

DRAG DROP -

You need to create a partitioned table in an Azure Synapse Analytics dedicated SQL pool.

How should you complete the Transact-SQL statement? To answer, drag the appropriate values to the correct targets. Each value 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:

Values

CLUSTERED INDEX
COLLATE
DISTRIBUTION
PARTITION
PARTITION FUNCTION
PARTITION SCHEME

Answer Area

```
CREATE TABLE table1
(
    ID INTEGER,
    col1 VARCHAR(10),
    col2 VARCHAR(10)
) WITH
(
    DISTRIBUTION = HASH(ID),
    PARTITION (ID RANGE LEFT FOR VALUES (1, 1000000, 2000000))
);
```

CertyIQ@gmail.com



Explanation

Box 1: DISTRIBUTION -

Table distribution options include **DISTRIBUTION = HASH (distribution_column_name)**, assigns each row to one distribution by hashing the value stored in **distribution_column_name**.

Explanation

Box 2: PARTITION - Table partition options. Syntax:

PARTITION (partition_column_name RANGE [LEFT | RIGHT] FOR VALUES ([boundary_value [...]n]))

You need to design an Azure Synapse Analytics dedicated SQL pool that meets the following requirements:

- Can return an employee record from a given point in time.
- Maintains the latest employee information.
- Minimizes query complexity.

How should you model the employee data?

- A. as a temporal table
- B. as a SQL graph table
- C. as a degenerate dimension table
- D. as a Type 2 slowly changing dimension (SCD) table

CertyIQ@gmail.com



You need to design an Azure Synapse Analytics dedicated SQL pool that meets the following requirements:

- Can return an employee record from a given point in time.
- Maintains the latest employee information.
- Minimizes query complexity.

How should you model the employee data?

- A. as a temporal table
- B. as a SQL graph table
- C. as a degenerate dimension table
- D. as a Type 2 slowly changing dimension (SCD) table**

CertyIQ@gmail.com

Explanation

Correct Answer: D

A Type 2 SCD supports versioning of dimension members. Often the source system doesn't store versions, so the data warehouse load process detects and manages changes in a dimension table. In this case, the dimension table must use a surrogate key to provide a unique reference to a version of the dimension member. It also includes columns that define the date range validity of the version (for example, StartDate and EndDate) and possibly a flag column (for example, IsCurrent) to easily filter by current dimension members.

You have an enterprise-wide Azure Data Lake Storage Gen2 account. The data lake is accessible only through an Azure virtual network named VNET1.

CertyIQ@gmail.com

You are building a SQL pool in Azure Synapse that will use data from the data lake.

Your company has a sales team. All the members of the sales team are in an Azure Active Directory group named Sales. POSIX controls are used to assign the Sales group access to the files in the data lake.

You plan to load data to the SQL pool every hour.

You need to ensure that the SQL pool can load the sales data from the data lake.

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

NOTE: Each area selection is worth one point.

- A. Add the managed identity to the Sales group.
- B. Use the managed identity as the credentials for the data load process.
- C. Create a shared access signature (SAS).
- D. Add your Azure Active Directory (Azure AD) account to the Sales group.
- E. Use the shared access signature (SAS) as the credentials for the data load process.
- F. Create a managed identity.



You have an enterprise-wide Azure Data Lake Storage Gen2 account. The data lake is accessible only through an Azure virtual network named VNET1.

CertyIQ@gmail.com

You are building a SQL pool in Azure Synapse that will use data from the data lake.

Your company has a sales team. All the members of the sales team are in an Azure Active Directory group named Sales. POSIX controls are used to assign the Sales group access to the files in the data lake.

You plan to load data to the SQL pool every hour.

You need to ensure that the SQL pool can load the sales data from the data lake.

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

NOTE: Each area selection is worth one point.

- A. Add the managed identity to the Sales group.
- B. Use the managed identity as the credentials for the data load process.
- C. Create a shared access signature (SAS).
- D. Add your Azure Active Directory (Azure AD) account to the Sales group.
- E. Use the shared access signature (SAS) as the credentials for the data load process.
- F. Create a managed identity.

Explanation

Correct Answer: ABF

The managed identity grants permissions to the dedicated SQL pools in the workspace.

You have an Azure Data Lake Storage Gen2 container that contains 100 TB of data. You need to ensure that the data in the container is available for read workloads in a secondary region if an outage occurs in the primary region. The solution must minimize costs. Which type of data redundancy should you use?

CertyIQ@gmail.com

- A. geo-redundant storage (GRS)
- B. read-access geo-redundant storage (RA-GRS)
- C. zone-redundant storage (ZRS)
- D. locally-redundant storage (LRS)

You have an Azure Data Lake Storage Gen2 container that contains 100 TB of data. You need to ensure that the data in the container is available for read workloads in a secondary region if an outage occurs in the primary region. The solution must minimize costs. Which type of data redundancy should you use?

- A. geo-redundant storage (GRS)
- B. read-access geo-redundant storage (RA-GRS)**
- C. zone-redundant storage (ZRS)
- D. locally-redundant storage (LRS)

CertyIQ@gmail.com



Explanation

Correct Answer: B

Geo-redundant storage (with GRS or GZRS) replicates your data to another physical location in the secondary region to protect against regional outages.

However, that data is available to be read only if the customer or Microsoft initiates a failover from the primary to secondary region. When you enable read access to the secondary region, your data is available to be read at all times, including in a situation where the primary region becomes unavailable.

You need to ensure that the data lake will remain available if a data center fails in the primary Azure region. The solution must minimize costs.

CertyIQ@gmail.com

Which type of replication should you use for the storage account?

- A. geo-redundant storage (GRS)
- B. geo-zone-redundant storage (GZRS)
- C. locally-redundant storage (LRS)
- D. zone-redundant storage (ZRS)

You need to ensure that the data lake will remain available if a data center fails in the primary Azure region. The solution must minimize costs.

CertyIQ@gmail.com

Which type of replication should you use for the storage account?

- A. geo-redundant storage (GRS)
- B. geo-zone-redundant storage (GZRS)
- C. locally-redundant storage (LRS)
- D. zone-redundant storage (ZRS)**

Explanation

Correct Answer: D

Zone-redundant storage (ZRS) copies your data synchronously across three Azure availability zones in the primary region.

You have a SQL pool in Azure Synapse.

CertyIQ@gmail.com

You plan to load data from Azure Blob storage to a staging table. Approximately 1 million rows of data will be loaded daily. The table will be truncated before each daily load.

You need to create the staging table. The solution must minimize how long it takes to load the data to the staging table.

How should you configure the table? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Distribution:

- Hash
- Replicated
- Round-robin

Indexing:

- Clustered
- Clustered columnstore
- Heap

Partitioning:

- Date
- None

You have a SQL pool in Azure Synapse.

CertyIQ@gmail.com

You plan to load data from Azure Blob storage to a staging table. Approximately 1 million rows of data will be loaded daily. The table will be truncated before each daily load.

You need to create the staging table. The solution must minimize how long it takes to load the data to the staging table.

How should you configure the table? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Distribution:

Hash
Replicated
Round-robin

Indexing:

Clustered
Clustered columnstore
Heap

Partitioning:

Date
None

Explanation

1. Search for “Use round-robin for the staging table.”
2. Search for: “A heap table can be especially useful for loading data, such as a staging table”
3. None

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 Storage account that contains 100 GB of files. The files contain rows of text and numerical values. 75% of the rows contain description data that has an average length of 1.1 MB.

You plan to copy the data from the storage account to an enterprise data warehouse in Azure Synapse Analytics.

You need to prepare the files to ensure that the data copies quickly.

Solution: You convert the files to compressed delimited text files.

Does this meet the goal?

- A. Yes
- B. No

CertyIQ@gmail.com



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.

CertyIQ@gmail.com



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 Storage account that contains 100 GB of files. The files contain rows of text and numerical values. 75% of the rows contain description data that has an average length of 1.1 MB.

You plan to copy the data from the storage account to an enterprise data warehouse in Azure Synapse Analytics.

You need to prepare the files to ensure that the data copies quickly.

Solution: You convert the files to compressed delimited text files.

Does this meet the goal?

- A. Yes
- B. No

Explanation

Correct Answer: Yes

convert the files to compressed delimited text files.

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 Storage account that contains 100 GB of files. The files contain rows of text and numerical values. 75% of the rows contain description data that has an average length of 1.1 MB.

You plan to copy the data from the storage account to an enterprise data warehouse in Azure Synapse Analytics.

You need to prepare the files to ensure that the data copies quickly.

Solution: You copy the files to a table that has a columnstore index.

Does this meet the goal?

- A. Yes
- B. No

CertyIQ@gmail.com



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.

CertyIQ@gmail.com

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 Storage account that contains 100 GB of files. The files contain rows of text and numerical values. 75% of the rows contain description data that has an average length of 1.1 MB.

You plan to copy the data from the storage account to an enterprise data warehouse in Azure Synapse Analytics.

You need to prepare the files to ensure that the data copies quickly.

Solution: You copy the files to a table that has a columnstore index.

Does this meet the goal?

- A. Yes
- B. No



Explanation

Correct Answer: No

Instead convert the files to compressed delimited text files.

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 Storage account that contains 100 GB of files. The files contain rows of text and numerical values. 75% of the rows contain description data that has an average length of 1.1 MB.

You plan to copy the data from the storage account to an enterprise data warehouse in Azure Synapse Analytics.

You need to prepare the files to ensure that the data copies quickly.

Solution: You modify the files to ensure that each row is more than 1 MB.

Does this meet the goal?

- A. Yes
- B. No

CertyIQ@gmail.com



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.

CertyIQ@gmail.com



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 Storage account that contains 100 GB of files. The files contain rows of text and numerical values. 75% of the rows contain description data that has an average length of 1.1 MB.

You plan to copy the data from the storage account to an enterprise data warehouse in Azure Synapse Analytics.

You need to prepare the files to ensure that the data copies quickly.

Solution: You modify the files to ensure that each row is more than 1 MB.

Does this meet the goal?

- A. Yes
- B. No

Explanation

Correct Answer: No

Instead convert the files to compressed delimited text files.

Rows need to have less than 1 MB. A batch size between 100 K to 1M rows is the recommended baseline for determining optimal batch size capacity.

You build a data warehouse in an Azure Synapse Analytics dedicated SQL pool.

CertyIQ@gmail.com

Analysts write a complex SELECT query that contains multiple JOIN and CASE statements to transform data for use in inventory reports. The inventory reports will use the data and additional WHERE parameters depending on the report. The reports will be produced once daily.

You need to implement a solution to make the dataset available for the reports. The solution must minimize query times.

What should you implement?

- A. an ordered clustered columnstore index
- B. a materialized view
- C. result set caching
- D. a replicated table

You build a data warehouse in an Azure Synapse Analytics dedicated SQL pool. Analysts write a complex SELECT query that contains multiple JOIN and CASE statements to transform data for use in inventory reports. The inventory reports will use the data and additional WHERE parameters depending on the report. The reports will be produced once daily. You need to implement a solution to make the dataset available for the reports. The solution must minimize query times.
What should you implement?

- A. an ordered clustered columnstore index
- B. a materialized view**
- C. result set caching
- D. a replicated table

CertyIQ@gmail.com



Explanation

Correct Answer: B

Materialized views for dedicated SQL pools in Azure Synapse provide a low maintenance method for complex analytical queries to get fast performance without any query change.

You have an Azure Synapse Analytics workspace named WS1 that contains an Apache Spark pool named Pool1.

CertyIQ@gmail.com

You plan to create a database named DB1 in Pool1.

You need to ensure that when tables are created in DB1, the tables are available automatically as external tables to the built-in serverless SQL pool.

Which format should you use for the tables in DB1?

- A. CSV
- B. ORC
- C. JSON
- D. Parquet

You have an Azure Synapse Analytics workspace named WS1 that contains an Apache Spark pool named Pool1.

CertyIQ@gmail.com

You plan to create a database named DB1 in Pool1.

You need to ensure that when tables are created in DB1, the tables are available automatically as external tables to the built-in serverless SQL pool.

Which format should you use for the tables in DB1?

- A. CSV
- B. ORC
- C. JSON
- D. Parquet

Explanation

Correct Answer: AD

For each Spark external table based on Parquet or CSV and located in Azure Storage, an external table is created in a serverless SQL pool database. As such, you can shut down your Spark pools and still query Spark external tables from serverless SQL pool

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.

CertyIQ@gmail.com

Which service should you recommend using to process the streaming data?

- A. Azure Event Hubs
- B. Azure Data Factory
- C. Azure Stream Analytics
- D. Azure Databricks

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.

CertyIQ@gmail.com

Which service should you recommend using to process the streaming data?

- A. Azure Event Hubs
- B. Azure Data Factory
- C. Azure Stream Analytics
- D. Azure Databricks**

Explanation

Correct Answer: D

Azure Databricks

You plan to implement an Azure Data Lake Storage Gen2 container that will contain CSV files.

The size of the files will vary based on the number of events that occur per hour.

File sizes range from 4 KB to 5 GB.

You need to ensure that the files stored in the container are optimized for batch processing.

What should you do?

- A. Convert the files to JSON
- B. Convert the files to Avro
- C. Compress the files
- D. Merge the files

CertyIQ@gmail.com



You plan to implement an Azure Data Lake Storage Gen2 container that will contain CSV files.

The size of the files will vary based on the number of events that occur per hour.

File sizes range from 4 KB to 5 GB.

You need to ensure that the files stored in the container are optimized for batch processing.

What should you do?

- A. Convert the files to JSON
- B. Convert the files to Avro**
- C. Compress the files
- D. Merge the files

CertyIQ@gmail.com



Explanation

Correct Answer: B

Avro supports batch and is very relevant for streaming.

Note: Avro is framework developed within Apache's Hadoop project.

It is a row-based storage format which is widely used as a serialization process. AVRO stores its schema in JSON format making it easy to read and interpret by any program. The data itself is stored in binary format by doing it compact and efficient.

You are designing a financial transactions table in an Azure Synapse Analytics dedicated SQL pool.

CertyIQ@gmail.com

The table will have a clustered columnstore index and will include the following columns:

- TransactionType: 40 million rows per transaction type
- CustomerSegment: 4 million per customer segment
- TransactionMonth: 65 million rows per month

AccountType: 500 million per account type

You have the following query requirements:

- Analysts will most commonly analyze transactions for a given month.
- Transactions analysis will typically summarize transactions by transaction type, customer segment, and/or account type

You need to recommend a partition strategy for the table to minimize query times.

On which column should you recommend partitioning the table?

- A. CustomerSegment
- B. AccountType
- C. TransactionType
- D. TransactionMonth

You are designing a financial transactions table in an Azure Synapse Analytics dedicated SQL pool.

CertyIQ@gmail.com

The table will have a clustered columnstore index and will include the following columns:

- TransactionType: 40 million rows per transaction type
- CustomerSegment: 4 million per customer segment
- TransactionMonth: 65 million rows per month

AccountType: 500 million per account type

You have the following query requirements:

- Analysts will most commonly analyze transactions for a given month.
- Transactions analysis will typically summarize transactions by transaction type, customer segment, and/or account type

You need to recommend a partition strategy for the table to minimize query times.

On which column should you recommend partitioning the table?

- A. CustomerSegment
- B. AccountType
- C. TransactionType
- D. TransactionMonth**

Explanation

Correct Answer: D

For optimal compression and performance of clustered columnstore tables, a minimum of 1 million rows per distribution and partition is needed. Before partitions are created, dedicated SQL pool already divides each table into 60 distributed databases.

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. JSON
- B. Parquet
- C. CSV
- D. Avro

CertyIQ@gmail.com



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.

CertyIQ@gmail.com

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. JSON
- B. Parquet**
- C. CSV
- D. Avro

Explanation

Correct Answer: B

Need Parquet to support both Databricks and PolyBase.

You are designing a partition strategy for a fact table in an Azure Synapse Analytics dedicated SQL pool. The table has the following specifications:

- Contain sales data for 20,000 products.
- Use hash distribution on a column named ProductID.
- Contain 2.4 billion records for the years 2019 and 2020.

Which number of partition ranges provides optimal compression and performance for the clustered columnstore index?

- A. 40
- B. 240
- C. 400
- D. 2,400

CertyIQ@gmail.com



You are designing a partition strategy for a fact table in an Azure Synapse Analytics dedicated SQL pool. The table has the following specifications:

- Contain sales data for 20,000 products.
- Use hash distribution on a column named ProductID.
- Contain 2.4 billion records for the years 2019 and 2020.

Which number of partition ranges provides optimal compression and performance for the clustered columnstore index?

A. 40

B. 240

C. 400

D. 2,400

CertyIQ@gmail.com



Explanation

Correct Answer: A

Each partition should have around 1 millions records. Dedicated SQL pools already have 60 partitions.

We have the formula: Records/(Partitions*60)= 1 million

Partitions= Records/(1 million * 60)

Partitions= $2.4 \times 1,000,000,000 / (1,000,000 \times 60) = 40$

You are implementing a batch dataset in the Parquet format.

Data files will be produced be using Azure Data Factory and stored in Azure Data Lake Storage Gen2. The files will be consumed by an Azure Synapse Analytics serverless SQL pool.

You need to minimize storage costs for the solution.

What should you do?

- A. Use Snappy compression for the files.
- B. Use OPENROWSET to query the Parquet files.
- C. Create an external table that contains a subset of columns from the Parquet files.
- D. Store all data as string in the Parquet files.

CertyIQ@gmail.com



You are implementing a batch dataset in the Parquet format.

Data files will be produced be using Azure Data Factory and stored in Azure Data Lake Storage Gen2. The files will be consumed by an Azure Synapse Analytics serverless SQL pool.

You need to minimize storage costs for the solution.

What should you do?

- A. Use Snappy compression for the files.
- B. Use OPENROWSET to query the Parquet files.
- C. Create an external table that contains a subset of columns from the Parquet files.**
- D. Store all data as string in the Parquet files.



CertyIQ@gmail.com
CertyIQ@gmail.com

Explanation

Correct Answer: C

An external table points to data located in Hadoop, Azure Storage blob, or Azure Data Lake Storage. External tables are used to read data from files or write data to files in Azure Storage. With Synapse SQL, you can use external tables to read external data using dedicated SQL pool or serverless SQL pool.

You are designing a data mart for the human resources (HR) department at your company.

The data mart will contain employee information and employee transactions.

From a source system, you have a flat extract that has the following fields:

- EmployeeID
- FirstName -
- LastName
- Recipient
- GrossAmount
- TransactionID
- GovernmentID
- NetAmountPaid
- TransactionDate

You need to design a star schema data model in an Azure Synapse Analytics dedicated SQL pool for the data mart.

Which two tables should you create? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. a dimension table for Transaction
- B. a dimension table for EmployeeTransaction
- C. a dimension table for Employee
- D. a fact table for Employee
- E. a fact table for Transaction



CertyIQ@gmail.com

You are designing a data mart for the human resources (HR) department at your company.

The data mart will contain employee information and employee transactions.

From a source system, you have a flat extract that has the following fields:

- EmployeeID
- FirstName -
- LastName
- Recipient
- GrossAmount
- TransactionID
- GovernmentID
- NetAmountPaid
- TransactionDate

CertyIQ@gmail.com

Explanation

Correct Answer: CE

C: Dimension tables contain attribute data that might change but usually changes infrequently.

E: Fact tables contain quantitative data that are commonly generated in a transactional system, and then loaded into the dedicated SQL pool.

You need to design a star schema data model in an Azure Synapse Analytics dedicated SQL pool for the data mart.

Which two tables should you create? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. a dimension table for Transaction
- B. a dimension table for EmployeeTransaction
- C. a dimension table for Employee**
- D. a fact table for Employee
- E. a fact table for Transaction**

You have an Azure Synapse Analytics Apache Spark pool named Pool1.

You plan to load JSON files from an Azure Data Lake Storage Gen2 container into the tables in Pool1. The structure and data types vary by file.

You need to load the files into the tables. The solution must maintain the source data types. What should you do?



CertyIQ@gmail.com

- A. Use a Conditional Split transformation in an Azure Synapse data flow.
- B. Use a Get Metadata activity in Azure Data Factory.
- C. Load the data by using the OPENROWSET Transact-SQL command in an Azure Synapse Analytics serverless SQL pool.
- D. Load the data by using PySpark.

You have an Azure Synapse Analytics Apache Spark pool named Pool1. You plan to load JSON files from an Azure Data Lake Storage Gen2 container into the tables in Pool1. The structure and data types vary by file. You need to load the files into the tables. The solution must maintain the source data types. What should you do?

- A. Use a Conditional Split transformation in an Azure Synapse data flow.
- B. Use a Get Metadata activity in Azure Data Factory.
- C. Load the data by using the OPENROWSET Transact-SQL command in an Azure Synapse Analytics serverless SQL pool.
- D. Load the data by using PySpark.**

**CertyIQ@gmail.com**

Explanation

Correct Answer: D

it's about Apache Spark pool, not serverless SQL pool.

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. Configure a global init script for workspace1.
- B. Create a cluster policy in workspace1.
- C. Upgrade workspace1 to the Premium pricing tier.
- D. Create a pool in workspace1.

CertyIQ@gmail.com



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. Configure a global init script for workspace1.
- B. Create a cluster policy in workspace1.
- C. Upgrade workspace1 to the Premium pricing tier.
- D. Create a pool in workspace1.**



CertyIQ@gmail.com

Explanation

Correct Answer: D

You can use Databricks Pools to Speed up your Data Pipelines and Scale Clusters Quickly.

Databricks Pools, a managed cache of virtual machine instances that enables clusters to start and scale 4 times faster.

HOTSPOT -

You have an Azure Data Lake Storage Gen2 service.

You need to design a data archiving solution that meets the following requirements:

- ☞ Data that is older than five years is accessed infrequently but must be available within one second when requested.
- ☞ Data that is older than seven years is NOT accessed.
- ☞ Costs must be minimized while maintaining the required availability.

How should you manage the data? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area: **Answer Area**

Data over five years old:

- Delete the blob.
- Move to archive storage.
- Move to cool storage.
- Move to hot storage.

Data over seven years old:

- Delete the blob.
- Move to archive storage.
- Move to cool storage.
- Move to hot storage.



CertyIQ@gmail.com

HOTSPOT -

You have an Azure Data Lake Storage Gen2 service.

You need to design a data archiving solution that meets the following requirements:

- ☞ Data that is older than five years is accessed infrequently but must be available within one second when requested.
- ☞ Data that is older than seven years is NOT accessed.
- ☞ Costs must be minimized while maintaining the required availability.

How should you manage the data? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area: **Answer Area**

Data over five years old:

- Delete the blob.
- Move to archive storage.
- Move to cool storage.
- Move to hot storage.

Data over seven years old:

- Delete the blob.
- Move to archive storage.
- Move to cool storage.
- Move to hot storage.

CertyIQ@gmail.com



Explanation

Correct Answer:

Box1: - Delete the blob

Box2: - Move to archive storage

You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Table1.

Table1 contains the following:

- One billion rows
- A clustered columnstore index
- A hash-distributed column named Product Key
- A column named Sales Date that is of the date data type and cannot be null

Thirty million rows will be added to Table1 each month.

You need to partition Table1 based on the Sales Date column. The solution must optimize query performance and data loading.

How often should you create a partition?

- A. once per month
- B. once per year
- C. once per day
- D. once per week



CertyIQ@gmail.com

You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Table1.

CertyIQ@gmail.com

Table1 contains the following:

- One billion rows
- A clustered columnstore index
- A hash-distributed column named Product Key
- A column named Sales Date that is of the date data type and cannot be null

Thirty million rows will be added to Table1 each month.

You need to partition Table1 based on the Sales Date column. The solution must optimize query performance and data loading.

How often should you create a partition?

- A. once per month
- B. once per year**
- C. once per day
- D. once per week

Explanation

Correct Answer: B

Need a minimum 1 million rows per distribution. Each table is 60 distributions. 30 millions rows is added each month.

Need 2 months to get a minimum of 1 million rows per distribution in a new partition.

You have an Azure Databricks workspace that contains a Delta Lake dimension table named Table1.

Table1 is a Type 2 slowly changing dimension (SCD) table.

You need to apply updates from a source table to Table1.

Which Apache Spark SQL operation should you use?

- A. CREATE
- B. UPDATE
- C. ALTER
- D. MERGE

CertyIQ@gmail.com



You have an Azure Databricks workspace that contains a Delta Lake dimension table named Table1.

Table1 is a Type 2 slowly changing dimension (SCD) table.

You need to apply updates from a source table to Table1.

Which Apache Spark SQL operation should you use?

- A. CREATE
- B. UPDATE
- C. ALTER
- D. MERGE**

CertyIQ@gmail.com



Explanation

Correct Answer: D

The Delta provides the ability to infer the schema for data input which further reduces the effort required in managing the schema changes. The Slowly Changing Dimension (SCD) Type 2 records all the changes made to each key in the dimensional table. These operations require updating the existing rows to mark the previous values of the keys as old and then inserting new rows as the latest values. Also, Given a source table with the updates and the target table with dimensional data, SCD Type 2 can be expressed with the merge.

You are designing an Azure Data Lake Storage solution that will transform raw JSON files for use in an analytical workload.

CertyIQ@gmail.com

You need to recommend a format for the transformed files. The solution must meet the following requirements:

- Contain information about the data types of each column in the files.
- Support querying a subset of columns in the files.
- Support read-heavy analytical workloads.
- Minimize the file size.

What should you recommend?

- A. JSON
- B. CSV
- C. Apache Avro
- D. Apache Parquet

You are designing an Azure Data Lake Storage solution that will transform raw JSON files for use in an analytical workload.

CertyIQ@gmail.com

You need to recommend a format for the transformed files. The solution must meet the following requirements:

- Contain information about the data types of each column in the files.
- Support querying a subset of columns in the files.
- Support read-heavy analytical workloads.
- Minimize the file size.

What should you recommend?

- A. JSON
- B. CSV
- C. Apache Avro
- D. Apache Parquet**

Explanation

Correct Answer: D

Parquet, an open-source file format for Hadoop, stores nested data structures in a flat columnar format.

Compared to a traditional approach where data is stored in a row-oriented approach, Parquet file format is more efficient in terms of storage and performance.

You plan to create a dimension table in Azure Synapse Analytics that will be less than 1 GB.

You need to create the table to meet the following requirements:

- Provide the fastest query time.
- Minimize data movement during queries.

Which type of table should you use?

- A. replicated
- B. hash distributed
- C. heap
- D. round-robin

CertyIQ@gmail.com



You plan to create a dimension table in Azure Synapse Analytics that will be less than 1 GB.

You need to create the table to meet the following requirements:

- Provide the fastest query time.
- Minimize data movement during queries.

Which type of table should you use?

A. replicated

B. hash distributed

C. heap

D. round-robin

CertyIQ@gmail.com



Explanation

Correct Answer: A

A replicated table has a full copy of the table accessible on each Compute node. Replicating a table removes the need to transfer data among Compute nodes before a join or aggregation. Since the table has multiple copies, replicated tables work best when the table size is less than 2 GB compressed. 2 GB is not a hard limit.

You are designing a dimension table in an Azure Synapse Analytics dedicated SQL pool. You need to create a surrogate key for the table. The solution must provide the fastest query performance.
What should you use for the surrogate key?

CertyIQ@gmail.com

- A. a GUID column
- B. a sequence object
- C. an IDENTITY column

You are designing a dimension table in an Azure Synapse Analytics dedicated SQL pool. You need to create a surrogate key for the table. The solution must provide the fastest query performance.
What should you use for the surrogate key?

- A. a GUID column
- B. a sequence object
- C. an IDENTITY column**

CertyIQ@gmail.com



Explanation

Correct Answer: C

Use IDENTITY to create surrogate keys using dedicated SQL pool in AzureSynapse Analytics.

Note: A surrogate key on a table is a column with a unique identifier for each row. The key is not generated from the table data. Data modelers like to create surrogate keys on their tables when they design data warehouse models. You can use the IDENTITY property to achieve this goal simply and effectively without affecting load performance.

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

Function:

Aggregate
Geospatial
Windowing



CertyIQ@gmail.com

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

Correct Answer:

Input type: Stream Reference

Function: Geospatial Aggregate Windowing



CertyIQ@gmail.com

Explanation

Correct Answer:

Input type: Stream -

You can process real-time IoT data streams with Azure Stream Analytics.

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.

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. Implement Azure Stream Analytics user-defined functions (UDF).
- C. Implement query parallelization by partitioning the data output.
- D. Scale the SU count for the job up.
- E. Scale the SU count for the job down.
- F. Implement query parallelization by partitioning the data input.

CertyIQ@gmail.com



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. Implement Azure Stream Analytics user-defined functions (UDF).
- C. Implement query parallelization by partitioning the data output.**
- D. Scale the SU count for the job up.
- E. Scale the SU count for the job down.
- F. Implement query parallelization by partitioning the data input.**

CertyIQ@gmail.com



Explanation

Correct Answer:

Partition input and output.

You need to trigger an Azure Data Factory pipeline when a file arrives in an Azure Data Lake Storage Gen2 container.

CertyIQ@gmail.com

Which resource provider should you enable?

- A. Microsoft.Sql
- B. Microsoft.Automation
- C. Microsoft.EventGrid
- D. Microsoft.EventHub

You need to trigger an Azure Data Factory pipeline when a file arrives in an Azure Data Lake Storage Gen2 container.

Which resource provider should you enable?



CertyIQ@gmail.com

- A. Microsoft.Sql
- B. Microsoft.Automation
- C. Microsoft.EventGrid**
- D. Microsoft.EventHub

Explanation

Correct Answer: C

Event-driven architecture (EDA) is a common data integration pattern that involves production, detection, consumption, and reaction to events. Data integration scenarios often require Data Factory customers to trigger pipelines based on events happening in storage account, such as the arrival or deletion of a file in Azure Blob Storage account. Data Factory natively integrates with Azure Event Grid, which lets you trigger pipelines on such events.

You plan to perform batch processing in Azure Databricks once daily.

CertyIQ@gmail.com

Which type of Databricks cluster should you use?

- A. High Concurrency
- B. automated
- C. interactive

You plan to perform batch processing in Azure Databricks once daily.

CertyIQ@gmail.com

Which type of Databricks cluster should you use?

- A. High Concurrency
- B. automated**
- C. interactive

Explanation

Correct Answer: B

Automated Databricks clusters are the best for jobs and automated batch processing.

Note: Azure Databricks has two types of clusters: interactive and automated. You use interactive clusters to analyze data collaboratively with interactive notebooks. You use automated clusters to run fast and robust automated jobs.

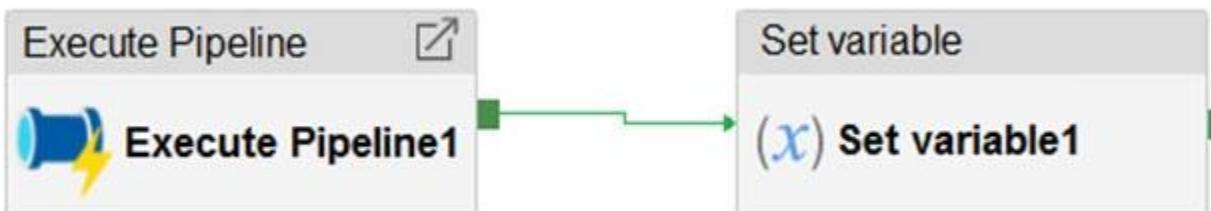
You have an Azure Data Factory instance that contains two pipelines named Pipeline1 and Pipeline2.

CertyIQ@gmail.com

Pipeline1 has the activities shown in the following exhibit.



Pipeline2 has the activities shown in the following exhibit.



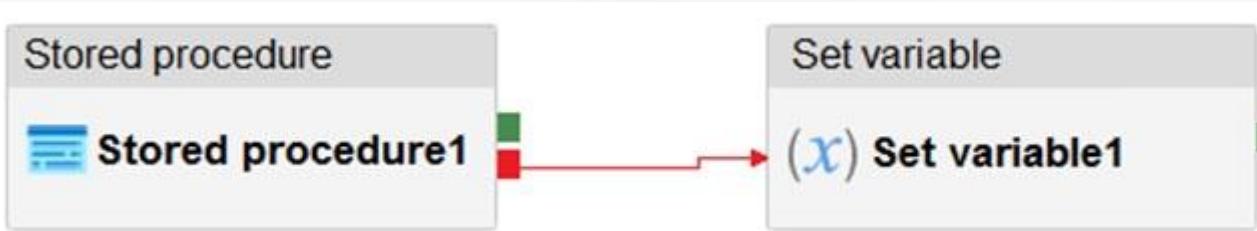
You execute Pipeline2, and Stored procedure1 in Pipeline1 fails.
What is the status of the pipeline runs?

- A. Pipeline1 and Pipeline2 succeeded.
- B. Pipeline1 and Pipeline2 failed.
- C. Pipeline1 succeeded and Pipeline2 failed.
- D. Pipeline1 failed and Pipeline2 succeeded.

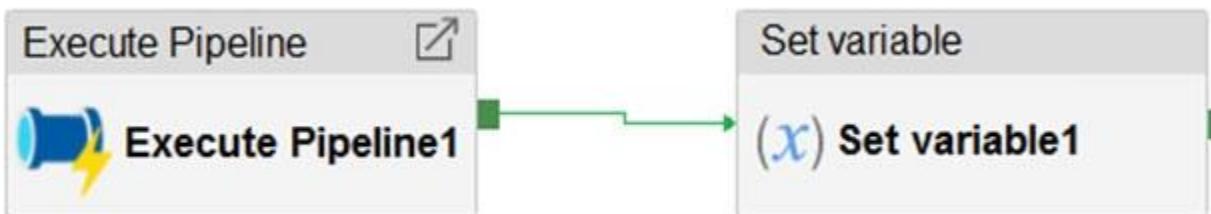
You have an Azure Data Factory instance that contains two pipelines named Pipeline1 and Pipeline2.

CertyIQ@gmail.com

Pipeline1 has the activities shown in the following exhibit.



Pipeline2 has the activities shown in the following exhibit.



You execute Pipeline2, and Stored procedure1 in Pipeline1 fails.
What is the status of the pipeline runs?

- A. Pipeline1 and Pipeline2 succeeded.
- B. Pipeline1 and Pipeline2 failed.
- C. Pipeline1 succeeded and Pipeline2 failed.
- D. Pipeline1 failed and Pipeline2 succeeded.

CertyIQ@gmail.com

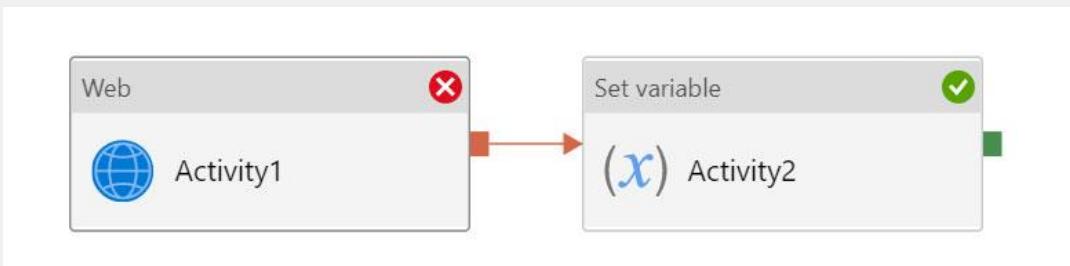
Explanation

Correct Answer: A

Activities are linked together via dependencies. A dependency has a condition of one of the following: Succeeded, Failed, Skipped, or Completed.

Consider Pipeline1:

If we have a pipeline with two activities where Activity2 has a failure dependency on Activity1, the pipeline will not fail just because Activity1 failed. If Activity1 fails and Activity2 succeeds, the pipeline will succeed. This scenario is treated as a try-catch block by Data Factory.



The failure dependency means this pipeline reports success.

Note:

If we have a pipeline containing Activity1 and Activity2, and Activity2 has a success dependency on Activity1, it will only execute if Activity1 is successful. In this scenario, if Activity1 fails, the pipeline will fail.

You have an Azure Data Factory that contains 10 pipelines.

You need to label each pipeline with its main purpose of either ingest, transform, or load.

The labels must be available for grouping and filtering when using the monitoring experience in Data Factory.

What should you add to each pipeline?

- A. a resource tag
- B. a correlation ID
- C. a run group ID
- D. an annotation

CertyIQ@gmail.com



You have an Azure Data Factory that contains 10 pipelines.

You need to label each pipeline with its main purpose of either ingest, transform, or load.

The labels must be available for grouping and filtering when using the monitoring experience in Data Factory.

What should you add to each pipeline?

- A. a resource tag
- B. a correlation ID
- C. a run group ID
- D. an annotation**

CertyIQ@gmail.com



Explanation

Correct Answer: D

Annotations are additional, informative tags that you can add to specific factory resources: pipelines, datasets, linked services, and triggers. By adding annotations, you can easily filter and search for specific factory resources.

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 Databricks
- C. Azure Stream Analytics
- D. Azure SQL Database

CertyIQ@gmail.com



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 Databricks**
- C. Azure Stream Analytics
- D. Azure SQL Database

CertyIQ@gmail.com



Explanation

Correct Answer: B

Azure Databricks supports near real-time data from Azure Event Hubs. And includes support for R, SQL, Python, Scala, and Java.

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 Stream Analytics solution that will analyze Twitter data.

You need to count the tweets in each 10-second window. The solution must ensure that each tweet is counted only once.

Solution: You use a hopping window that uses a hop size of 5 seconds and a window size 10 seconds.

Does this meet the goal?

A. Yes

B. No

CertyIQ@gmail.com



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 Stream Analytics solution that will analyze Twitter data.

You need to count the tweets in each 10-second window. The solution must ensure that each tweet is counted only once.

Solution: You use a hopping window that uses a hop size of 5 seconds and a window size 10 seconds.

Does this meet the goal?

A. Yes

B. No

CertyIQ@gmail.com



Explanation

Correct Answer: B

Instead use a tumbling window. Tumbling windows are a series of fixed-sized, non-overlapping and contiguous time intervals.

You need to schedule an Azure Data Factory pipeline to execute when a new file arrives in an Azure Data Lake Storage Gen2 container.

Which type of trigger should you use?

- A. on-demand
- B. tumbling window
- C. schedule
- D. event

CertyIQ@gmail.com



You need to schedule an Azure Data Factory pipeline to execute when a new file arrives in an Azure Data Lake Storage Gen2 container.

Which type of trigger should you use?

- A. on-demand
- B. tumbling window
- C. schedule**
- D. event

CertyIQ@gmail.com



Explanation

Correct Answer: C

In Azure Data Factory, continuous integration and delivery (CI/CD) means moving Data Factory pipelines from one environment (development, test, production) to another.

End

On a Mission to make 100+ Certified Cloud Engineer in next 30 Days

CertyIQ

CertyIQ@gmail.com

Please **Subscribe** to CertyIQ Channel
To get notification for upcoming dumps

SUBSCRIBE



DP-203 Important Que's



Free
PDF

On a Mission to make 100+ Certified Cloud Engineer in next 30 Days

CertyIQ

DP-203

Real Exam Questions & Answers



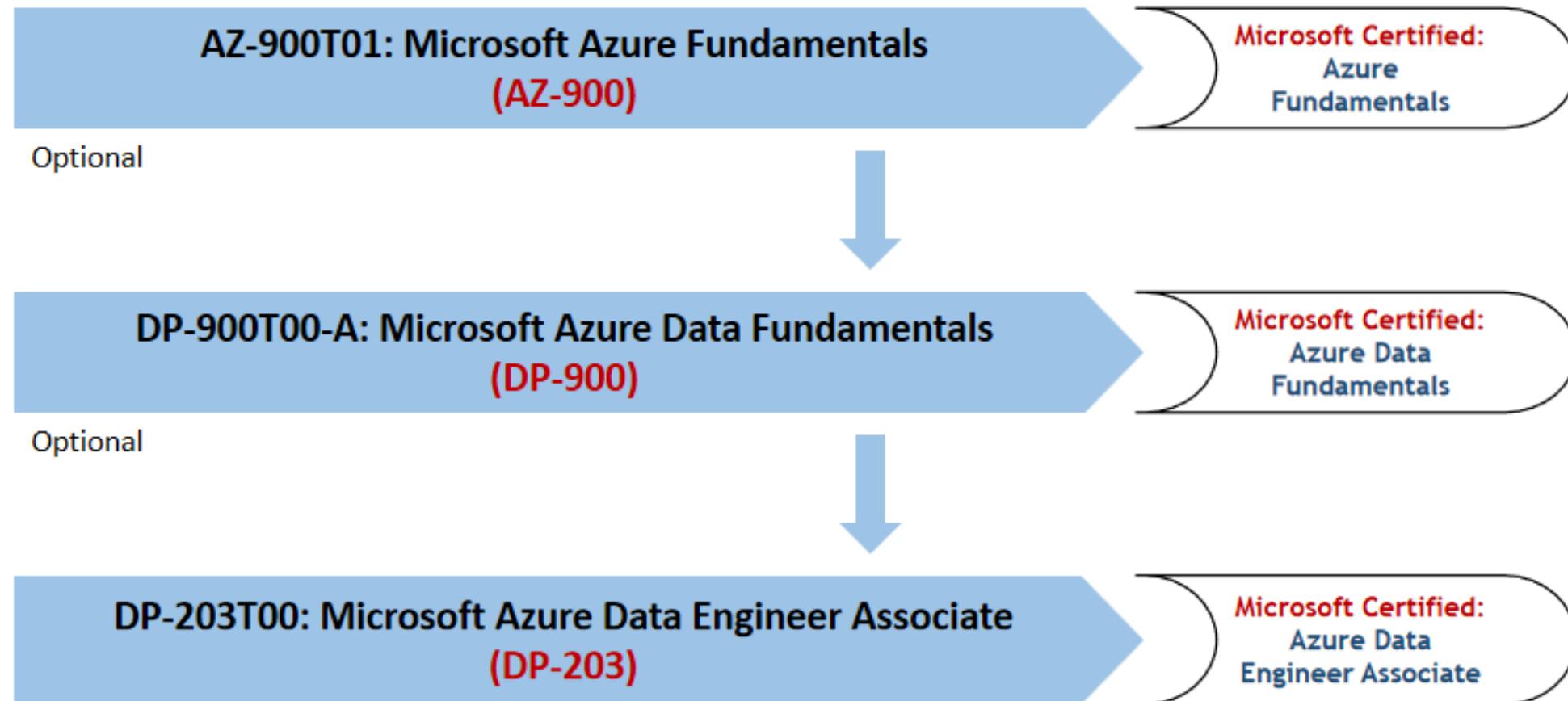
Latest Exam Questions
All Topics Covered-2023
Added New Que'ss

IMPORTANT QUESTIONS



Subscribe

Microsoft Azure Data Engineer Certification Path



Data Engineering on Microsoft Azure (DP-203) Exam Format

Exam Name	Data Engineering on Microsoft Azure
Exam Code	DP-203
Exam Cost	USD 165
Exam Format	Multiple Choice and Multiple Response Questions
Total Questions	40-60 Questions
Passing Score	700 out of 1000
Exam Duration	130 Minutes
Languages	English, Chinese (Simplified), German, French, Spanish, Chinese (Traditional), Italian, Indonesian (Indonesia), Arabic (Saudi Arabia), Russian, Portuguese (Brazil), Japanese, Korean

You have an Azure Synapse workspace named MyWorkspace that contains an Apache Spark database named mytestdb.

You run the following command in an Azure Synapse Analytics Spark pool in MyWorkspace.

```
CREATE TABLE mytestdb.myParquetTable(
```

```
EmployeeID int,
```

```
EmployeeName string,
```

```
EmployeeStartDate date)
```

```
USING Parquet -
```

You then use Spark to insert a row into mytestdb.myParquetTable. The row contains the following data.

EmployeeName	EmployeeID	EmployeeStartDate
Alice	24	2020-01-25

One minute later, you execute the following query from a serverless SQL pool in MyWorkspace.

```
SELECT EmployeeID -
```

```
FROM mytestdb.dbo.myParquetTable
```

```
WHERE EmployeeName = 'Alice';
```

What will be returned by the query?

A. 24

B. an error

C. a null value

CertyIQ@gmail.com



You have an Azure Synapse workspace named MyWorkspace that contains an Apache Spark database named mytestdb.

You run the following command in an Azure Synapse Analytics Spark pool in MyWorkspace.

```
CREATE TABLE mytestdb.myParquetTable(
```

```
EmployeeID int,
```

```
EmployeeName string,
```

```
EmployeeStartDate date)
```

```
USING Parquet -
```

You then use Spark to insert a row into mytestdb.myParquetTable. The row contains the following data.

EmployeeName	EmployeeID	EmployeeStartDate
Alice	24	2020-01-25

One minute later, you execute the following query from a serverless SQL pool in MyWorkspace.

```
SELECT EmployeeID -
```

```
FROM mytestdb.dbo.myParquetTable
```

```
WHERE EmployeeName = 'Alice';
```

What will be returned by the query?

A. 24

B. an error

C. a null value

CertyIQ@gmail.com



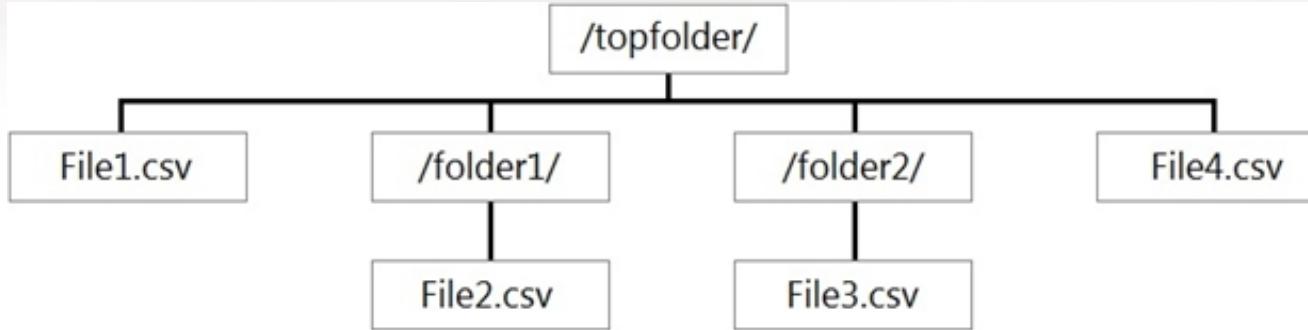
Explanation

Correct Answer: B

It will return an error. table name will be converted to lowercase

You have files and folders in Azure Data Lake Storage Gen2 for an Azure Synapse workspace as shown in the following exhibit.

CertyIQ@gmail.com



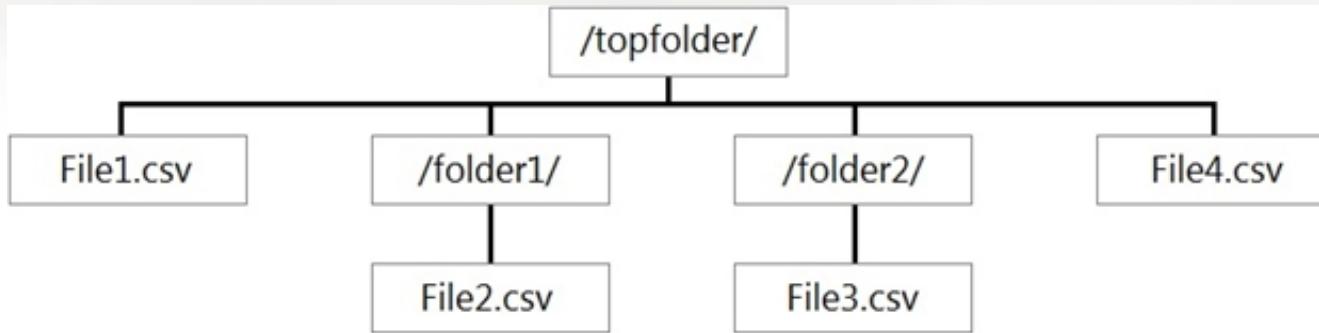
You create an external table named ExtTable that has LOCATION='/topfolder/'.

When you query ExtTable by using an Azure Synapse Analytics serverless SQL pool, which files are returned?

- A. File2.csv and File3.csv only
- B. File1.csv and File4.csv only
- C. File1.csv, File2.csv, File3.csv, and File4.csv
- D. File1.csv only

You have files and folders in Azure Data Lake Storage Gen2 for an Azure Synapse workspace as shown in the following exhibit.

CertyIQ@gmail.com



You create an external table named ExtTable that has LOCATION='/topfolder/'.

When you query ExtTable by using an Azure Synapse Analytics serverless SQL pool, which files are returned?

- A. File2.csv and File3.csv only
- B. File1.csv and File4.csv only**
- C. File1.csv, File2.csv, File3.csv, and File4.csv
- D. File1.csv only

Explanation

Correct Answer: B

In case of a serverless pool a wildcard should be added to the location.

HOTSPOT -

CertyIQ@gmail.com

You are planning the deployment of Azure Data Lake Storage Gen2.

You have the following two reports that will access the data lake:

- ☞ Report1: Reads three columns from a file that contains 50 columns.
- ☞ Report2: Queries a single record based on a timestamp.

You need to recommend in which format to store the data in the data lake to support the reports. The solution must minimize read times.

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

NOTE: Each correct selection is worth one point.

Answer Area

Hot Area:

Report1:

Avro	▼
CSV	
Parquet	
TSV	

Report2:

Avro	▼
CSV	
Parquet	
TSV	

HOTSPOT -

CertyIQ@gmail.com

You are planning the deployment of Azure Data Lake Storage Gen2.

You have the following two reports that will access the data lake:

☞ Report1: Reads three columns from a file that contains 50 columns.

☞ Report2: Queries a single record based on a timestamp.

You need to recommend in which format to store the data in the data lake to support the reports. The solution must minimize read times.

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

NOTE: Each correct selection is worth one point.

Answer Area

Hot Area:

Report1:

Avro
CSV
Parquet
TSV

Report2:

Avro
CSV
Parquet
TSV

Explanation

1: Parquet - column-oriented binary file format

2: AVRO - Row based format, and has logical type timestamp

HOTSPOT -

You need to output files from Azure Data Factory.

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

NOTE: Each correct selection is worth one point.

Hot Area:

CertyIQ@gmail.com

Columnar format:

Avro
GZip
Parquet
TXT

JSON with a timestamp:

Avro
GZip
Parquet
TXT

HOTSPOT -

You need to output files from Azure Data Factory.

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Columnar format:

Avro
GZip
Parquet
TXT

Correct Answer:

JSON with a timestamp:

Avro
GZip
Parquet
TXT

CertyIQ@gmail.com



Explanation

Box 1: Parquet -

Parquet stores data in columns, while Avro stores data in a row-based format. By their very nature, column-oriented data stores are optimized for read-heavy analytical workloads, while row-based databases are best for write-heavy transactional workloads.

Box 2: Avro -

An Avro schema is created using JSON format.

AVRO supports timestamps.

Note: Azure Data Factory supports the following file formats (not GZip or TXT).

HOTSPOT -

CertyIQ@gmail.com

You use Azure Data Factory to prepare data to be queried by Azure Synapse Analytics serverless SQL pools.



Files are initially ingested into an Azure Data Lake Storage Gen2 account as 10 small JSON files. Each file contains the same data attributes and data from a subsidiary of your company. You need to move the files to a different folder and transform the data to meet the following requirements:

- Provide the fastest possible query times.
- Automatically infer the schema from the underlying files.

How should you configure the Data Factory copy activity? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area: **Answer Area**

Copy behavior:

Flatten hierarchy
Merge files
Preserve hierarchy

Sink file type:

CSV
JSON
Parquet
TXT

HOTSPOT -

CertyIQ@gmail.com

You use Azure Data Factory to prepare data to be queried by Azure Synapse Analytics serverless SQL pools.

Files are initially ingested into an Azure Data Lake Storage Gen2 account as 10 small JSON files. Each file contains the same data attributes and data from a subsidiary of your company. You need to move the files to a different folder and transform the data to meet the following requirements:

- Provide the fastest possible query times.
- Automatically infer the schema from the underlying files.

How should you configure the Data Factory copy activity? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Copy behavior:

Flatten hierarchy
Merge files
Preserve hierarchy

Sink file type:

CSV
JSON
Parquet
TXT

Explanation

Correct Answer:

1. Merge Files

2 Parquet -

Azure Data Factory parquet format is supported for Azure Data Lake Storage Gen2.

Parquet supports the schema property.

Question 6

DRAG DROP -

You need to create a partitioned table in an Azure Synapse Analytics dedicated SQL pool.

How should you complete the Transact-SQL statement? To answer, drag the appropriate values to the correct targets. Each value 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:

CertyIQ@gmail.com

Values

CLUSTERED INDEX
COLLATE
DISTRIBUTION
PARTITION
PARTITION FUNCTION
PARTITION SCHEME

Answer Area

```
CREATE TABLE table1
(
    ID INTEGER,
    col1 VARCHAR(10),
    col2 VARCHAR(10)
) WITH
(
    [ ] = HASH(ID),
    [ ] (ID RANGE LEFT FOR VALUES (1, 1000000, 2000000))
);
```

DRAG DROP -

You need to create a partitioned table in an Azure Synapse Analytics dedicated SQL pool.

How should you complete the Transact-SQL statement? To answer, drag the appropriate values to the correct targets. Each value 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:

Values

CLUSTERED INDEX
COLLATE
DISTRIBUTION
PARTITION
PARTITION FUNCTION
PARTITION SCHEME

Answer Area

```
CREATE TABLE table1
(
    ID INTEGER,
    col1 VARCHAR(10),
    col2 VARCHAR(10)
) WITH
(
    DISTRIBUTION = HASH(ID),
    PARTITION (ID RANGE LEFT FOR VALUES (1, 1000000, 2000000))
);
```

CertyIQ@gmail.com



Explanation

Box 1: DISTRIBUTION -

Table distribution options include **DISTRIBUTION = HASH (distribution_column_name)**, assigns each row to one distribution by hashing the value stored in **distribution_column_name**.

Explanation

Box 2: PARTITION - Table partition options. Syntax:

PARTITION (partition_column_name RANGE [LEFT | RIGHT] FOR VALUES ([boundary_value [...]n]))

You need to design an Azure Synapse Analytics dedicated SQL pool that meets the following requirements:

- Can return an employee record from a given point in time.
- Maintains the latest employee information.
- Minimizes query complexity.

How should you model the employee data?

- A. as a temporal table
- B. as a SQL graph table
- C. as a degenerate dimension table
- D. as a Type 2 slowly changing dimension (SCD) table

CertyIQ@gmail.com



You need to design an Azure Synapse Analytics dedicated SQL pool that meets the following requirements:

- Can return an employee record from a given point in time.
- Maintains the latest employee information.
- Minimizes query complexity.

How should you model the employee data?

- A. as a temporal table
- B. as a SQL graph table
- C. as a degenerate dimension table
- D. as a Type 2 slowly changing dimension (SCD) table**

CertyIQ@gmail.com

Explanation

Correct Answer: D

A Type 2 SCD supports versioning of dimension members. Often the source system doesn't store versions, so the data warehouse load process detects and manages changes in a dimension table. In this case, the dimension table must use a surrogate key to provide a unique reference to a version of the dimension member. It also includes columns that define the date range validity of the version (for example, StartDate and EndDate) and possibly a flag column (for example, IsCurrent) to easily filter by current dimension members.

You have an enterprise-wide Azure Data Lake Storage Gen2 account. The data lake is accessible only through an Azure virtual network named VNET1.

CertyIQ@gmail.com

You are building a SQL pool in Azure Synapse that will use data from the data lake.

Your company has a sales team. All the members of the sales team are in an Azure Active Directory group named Sales. POSIX controls are used to assign the Sales group access to the files in the data lake.

You plan to load data to the SQL pool every hour.

You need to ensure that the SQL pool can load the sales data from the data lake.

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

NOTE: Each area selection is worth one point.

- A. Add the managed identity to the Sales group.
- B. Use the managed identity as the credentials for the data load process.
- C. Create a shared access signature (SAS).
- D. Add your Azure Active Directory (Azure AD) account to the Sales group.
- E. Use the shared access signature (SAS) as the credentials for the data load process.
- F. Create a managed identity.



You have an enterprise-wide Azure Data Lake Storage Gen2 account. The data lake is accessible only through an Azure virtual network named VNET1.

CertyIQ@gmail.com

You are building a SQL pool in Azure Synapse that will use data from the data lake.

Your company has a sales team. All the members of the sales team are in an Azure Active Directory group named Sales. POSIX controls are used to assign the Sales group access to the files in the data lake.

You plan to load data to the SQL pool every hour.

You need to ensure that the SQL pool can load the sales data from the data lake.

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

NOTE: Each area selection is worth one point.

- A. Add the managed identity to the Sales group.
- B. Use the managed identity as the credentials for the data load process.
- C. Create a shared access signature (SAS).
- D. Add your Azure Active Directory (Azure AD) account to the Sales group.
- E. Use the shared access signature (SAS) as the credentials for the data load process.
- F. Create a managed identity.

Explanation

Correct Answer: ABF

The managed identity grants permissions to the dedicated SQL pools in the workspace.

You have an Azure Data Lake Storage Gen2 container that contains 100 TB of data. You need to ensure that the data in the container is available for read workloads in a secondary region if an outage occurs in the primary region. The solution must minimize costs. Which type of data redundancy should you use?

- A. geo-redundant storage (GRS)
- B. read-access geo-redundant storage (RA-GRS)
- C. zone-redundant storage (ZRS)
- D. locally-redundant storage (LRS)

CertyIQ@gmail.com



You have an Azure Data Lake Storage Gen2 container that contains 100 TB of data. You need to ensure that the data in the container is available for read workloads in a secondary region if an outage occurs in the primary region. The solution must minimize costs. Which type of data redundancy should you use?

- A. geo-redundant storage (GRS)
- B. read-access geo-redundant storage (RA-GRS)**
- C. zone-redundant storage (ZRS)
- D. locally-redundant storage (LRS)

CertyIQ@gmail.com



Explanation

Correct Answer: B

Geo-redundant storage (with GRS or GZRS) replicates your data to another physical location in the secondary region to protect against regional outages.

However, that data is available to be read only if the customer or Microsoft initiates a failover from the primary to secondary region. When you enable read access to the secondary region, your data is available to be read at all times, including in a situation where the primary region becomes unavailable.

You need to ensure that the data lake will remain available if a data center fails in the primary Azure region. The solution must minimize costs.

CertyIQ@gmail.com

Which type of replication should you use for the storage account?

- A. geo-redundant storage (GRS)
- B. geo-zone-redundant storage (GZRS)
- C. locally-redundant storage (LRS)
- D. zone-redundant storage (ZRS)

You need to ensure that the data lake will remain available if a data center fails in the primary Azure region. The solution must minimize costs.

CertyIQ@gmail.com

Which type of replication should you use for the storage account?

- A. geo-redundant storage (GRS)
- B. geo-zone-redundant storage (GZRS)
- C. locally-redundant storage (LRS)
- D. zone-redundant storage (ZRS)**

Explanation

Correct Answer: D

Zone-redundant storage (ZRS) copies your data synchronously across three Azure availability zones in the primary region.

You have a SQL pool in Azure Synapse.

CertyIQ@gmail.com

You plan to load data from Azure Blob storage to a staging table. Approximately 1 million rows of data will be loaded daily. The table will be truncated before each daily load.

You need to create the staging table. The solution must minimize how long it takes to load the data to the staging table.

How should you configure the table? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Distribution:

- Hash
- Replicated
- Round-robin

Indexing:

- Clustered
- Clustered columnstore
- Heap

Partitioning:

- Date
- None

You have a SQL pool in Azure Synapse.

CertyIQ@gmail.com

You plan to load data from Azure Blob storage to a staging table. Approximately 1 million rows of data will be loaded daily. The table will be truncated before each daily load.

You need to create the staging table. The solution must minimize how long it takes to load the data to the staging table.

How should you configure the table? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Distribution:

Hash
Replicated
Round-robin

Indexing:

Clustered
Clustered columnstore
Heap

Partitioning:

Date
None

Explanation

1. Search for “Use round-robin for the staging table.”
2. Search for: “A heap table can be especially useful for loading data, such as a staging table”
3. None

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 Storage account that contains 100 GB of files. The files contain rows of text and numerical values. 75% of the rows contain description data that has an average length of 1.1 MB.

You plan to copy the data from the storage account to an enterprise data warehouse in Azure Synapse Analytics.

You need to prepare the files to ensure that the data copies quickly.

Solution: You convert the files to compressed delimited text files.

Does this meet the goal?

- A. Yes
- B. No

CertyIQ@gmail.com



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.

CertyIQ@gmail.com



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 Storage account that contains 100 GB of files. The files contain rows of text and numerical values. 75% of the rows contain description data that has an average length of 1.1 MB.

You plan to copy the data from the storage account to an enterprise data warehouse in Azure Synapse Analytics.

You need to prepare the files to ensure that the data copies quickly.

Solution: You convert the files to compressed delimited text files.

Does this meet the goal?

- A. Yes
- B. No

Explanation

Correct Answer: Yes

convert the files to compressed delimited text files.

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 Storage account that contains 100 GB of files. The files contain rows of text and numerical values. 75% of the rows contain description data that has an average length of 1.1 MB.

You plan to copy the data from the storage account to an enterprise data warehouse in Azure Synapse Analytics.

You need to prepare the files to ensure that the data copies quickly.

Solution: You copy the files to a table that has a columnstore index.

Does this meet the goal?

- A. Yes
- B. No

CertyIQ@gmail.com



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.

CertyIQ@gmail.com

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 Storage account that contains 100 GB of files. The files contain rows of text and numerical values. 75% of the rows contain description data that has an average length of 1.1 MB.

You plan to copy the data from the storage account to an enterprise data warehouse in Azure Synapse Analytics.

You need to prepare the files to ensure that the data copies quickly.

Solution: You copy the files to a table that has a columnstore index.

Does this meet the goal?

- A. Yes
- B. No



Explanation

Correct Answer: No

Instead convert the files to compressed delimited text files.

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 Storage account that contains 100 GB of files. The files contain rows of text and numerical values. 75% of the rows contain description data that has an average length of 1.1 MB.

You plan to copy the data from the storage account to an enterprise data warehouse in Azure Synapse Analytics.

You need to prepare the files to ensure that the data copies quickly.

Solution: You modify the files to ensure that each row is more than 1 MB.

Does this meet the goal?

- A. Yes
- B. No

CertyIQ@gmail.com



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.

CertyIQ@gmail.com



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 Storage account that contains 100 GB of files. The files contain rows of text and numerical values. 75% of the rows contain description data that has an average length of 1.1 MB.

You plan to copy the data from the storage account to an enterprise data warehouse in Azure Synapse Analytics.

You need to prepare the files to ensure that the data copies quickly.

Solution: You modify the files to ensure that each row is more than 1 MB.

Does this meet the goal?

- A. Yes
- B. No

Explanation

Correct Answer: No

Instead convert the files to compressed delimited text files.

Rows need to have less than 1 MB. A batch size between 100 K to 1M rows is the recommended baseline for determining optimal batch size capacity.

You build a data warehouse in an Azure Synapse Analytics dedicated SQL pool.

CertyIQ@gmail.com

Analysts write a complex SELECT query that contains multiple JOIN and CASE statements to transform data for use in inventory reports. The inventory reports will use the data and additional WHERE parameters depending on the report. The reports will be produced once daily.

You need to implement a solution to make the dataset available for the reports. The solution must minimize query times.

What should you implement?

- A. an ordered clustered columnstore index
- B. a materialized view
- C. result set caching
- D. a replicated table

You build a data warehouse in an Azure Synapse Analytics dedicated SQL pool. Analysts write a complex SELECT query that contains multiple JOIN and CASE statements to transform data for use in inventory reports. The inventory reports will use the data and additional WHERE parameters depending on the report. The reports will be produced once daily. You need to implement a solution to make the dataset available for the reports. The solution must minimize query times.
What should you implement?

- A. an ordered clustered columnstore index
- B. a materialized view**
- C. result set caching
- D. a replicated table

CertyIQ@gmail.com



Explanation

Correct Answer: B

Materialized views for dedicated SQL pools in Azure Synapse provide a low maintenance method for complex analytical queries to get fast performance without any query change.

You have an Azure Synapse Analytics workspace named WS1 that contains an Apache Spark pool named Pool1.

CertyIQ@gmail.com

You plan to create a database named DB1 in Pool1.

You need to ensure that when tables are created in DB1, the tables are available automatically as external tables to the built-in serverless SQL pool.

Which format should you use for the tables in DB1?

- A. CSV
- B. ORC
- C. JSON
- D. Parquet

You have an Azure Synapse Analytics workspace named WS1 that contains an Apache Spark pool named Pool1.

CertyIQ@gmail.com

You plan to create a database named DB1 in Pool1.

You need to ensure that when tables are created in DB1, the tables are available automatically as external tables to the built-in serverless SQL pool.

Which format should you use for the tables in DB1?

- A. CSV
- B. ORC
- C. JSON
- D. Parquet

Explanation

Correct Answer: AD

For each Spark external table based on Parquet or CSV and located in Azure Storage, an external table is created in a serverless SQL pool database. As such, you can shut down your Spark pools and still query Spark external tables from serverless SQL pool

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.

CertyIQ@gmail.com

Which service should you recommend using to process the streaming data?

- A. Azure Event Hubs
- B. Azure Data Factory
- C. Azure Stream Analytics
- D. Azure Databricks

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.

CertyIQ@gmail.com

Which service should you recommend using to process the streaming data?

- A. Azure Event Hubs
- B. Azure Data Factory
- C. Azure Stream Analytics
- D. Azure Databricks**

Explanation

Correct Answer: D

Azure Databricks

You plan to implement an Azure Data Lake Storage Gen2 container that will contain CSV files.

The size of the files will vary based on the number of events that occur per hour.

File sizes range from 4 KB to 5 GB.

You need to ensure that the files stored in the container are optimized for batch processing.

What should you do?

- A. Convert the files to JSON
- B. Convert the files to Avro
- C. Compress the files
- D. Merge the files

CertyIQ@gmail.com



You plan to implement an Azure Data Lake Storage Gen2 container that will contain CSV files.

The size of the files will vary based on the number of events that occur per hour.

File sizes range from 4 KB to 5 GB.

You need to ensure that the files stored in the container are optimized for batch processing.

What should you do?

- A. Convert the files to JSON
- B. Convert the files to Avro**
- C. Compress the files
- D. Merge the files

CertyIQ@gmail.com



Explanation

Correct Answer: B

Avro supports batch and is very relevant for streaming.

Note: Avro is framework developed within Apache's Hadoop project.

It is a row-based storage format which is widely used as a serialization process. AVRO stores its schema in JSON format making it easy to read and interpret by any program. The data itself is stored in binary format by doing it compact and efficient.

You are designing a financial transactions table in an Azure Synapse Analytics dedicated SQL pool.

CertyIQ@gmail.com

The table will have a clustered columnstore index and will include the following columns:

- TransactionType: 40 million rows per transaction type
- CustomerSegment: 4 million per customer segment
- TransactionMonth: 65 million rows per month

AccountType: 500 million per account type

You have the following query requirements:

- Analysts will most commonly analyze transactions for a given month.
- Transactions analysis will typically summarize transactions by transaction type, customer segment, and/or account type

You need to recommend a partition strategy for the table to minimize query times.

On which column should you recommend partitioning the table?

- A. CustomerSegment
- B. AccountType
- C. TransactionType
- D. TransactionMonth

You are designing a financial transactions table in an Azure Synapse Analytics dedicated SQL pool.

CertyIQ@gmail.com

The table will have a clustered columnstore index and will include the following columns:

- TransactionType: 40 million rows per transaction type
- CustomerSegment: 4 million per customer segment
- TransactionMonth: 65 million rows per month

AccountType: 500 million per account type

You have the following query requirements:

- Analysts will most commonly analyze transactions for a given month.
- Transactions analysis will typically summarize transactions by transaction type, customer segment, and/or account type

You need to recommend a partition strategy for the table to minimize query times.

On which column should you recommend partitioning the table?

- A. CustomerSegment
- B. AccountType
- C. TransactionType
- D. TransactionMonth**

Explanation

Correct Answer: D

For optimal compression and performance of clustered columnstore tables, a minimum of 1 million rows per distribution and partition is needed. Before partitions are created, dedicated SQL pool already divides each table into 60 distributed databases.

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. JSON
- B. Parquet
- C. CSV
- D. Avro

CertyIQ@gmail.com



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.

CertyIQ@gmail.com

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. JSON
- B. Parquet**
- C. CSV
- D. Avro

Explanation

Correct Answer: B

Need Parquet to support both Databricks and PolyBase.

You are designing a partition strategy for a fact table in an Azure Synapse Analytics dedicated SQL pool. The table has the following specifications:

- Contain sales data for 20,000 products.
- Use hash distribution on a column named ProductID.
- Contain 2.4 billion records for the years 2019 and 2020.

Which number of partition ranges provides optimal compression and performance for the clustered columnstore index?

- A. 40
- B. 240
- C. 400
- D. 2,400

CertyIQ@gmail.com



You are designing a partition strategy for a fact table in an Azure Synapse Analytics dedicated SQL pool. The table has the following specifications:

- Contain sales data for 20,000 products.
- Use hash distribution on a column named ProductID.
- Contain 2.4 billion records for the years 2019 and 2020.

Which number of partition ranges provides optimal compression and performance for the clustered columnstore index?

A. 40

B. 240

C. 400

D. 2,400

CertyIQ@gmail.com



Explanation

Correct Answer: A

Each partition should have around 1 millions records. Dedicated SQL pools already have 60 partitions.

We have the formula: Records/(Partitions*60)= 1 million

Partitions= Records/(1 million * 60)

Partitions= $2.4 \times 1,000,000,000 / (1,000,000 \times 60) = 40$

You are implementing a batch dataset in the Parquet format.

Data files will be produced be using Azure Data Factory and stored in Azure Data Lake Storage Gen2. The files will be consumed by an Azure Synapse Analytics serverless SQL pool.

You need to minimize storage costs for the solution.

What should you do?

- A. Use Snappy compression for the files.
- B. Use OPENROWSET to query the Parquet files.
- C. Create an external table that contains a subset of columns from the Parquet files.
- D. Store all data as string in the Parquet files.

CertyIQ@gmail.com



You are implementing a batch dataset in the Parquet format.

Data files will be produced be using Azure Data Factory and stored in Azure Data Lake Storage Gen2. The files will be consumed by an Azure Synapse Analytics serverless SQL pool.

You need to minimize storage costs for the solution.

What should you do?

- A. Use Snappy compression for the files.
- B. Use OPENROWSET to query the Parquet files.
- C. Create an external table that contains a subset of columns from the Parquet files.**
- D. Store all data as string in the Parquet files.



CertyIQ@gmail.com
CertyIQ@gmail.com

Explanation

Correct Answer: C

An external table points to data located in Hadoop, Azure Storage blob, or Azure Data Lake Storage. External tables are used to read data from files or write data to files in Azure Storage. With Synapse SQL, you can use external tables to read external data using dedicated SQL pool or serverless SQL pool.

You are designing a data mart for the human resources (HR) department at your company.

The data mart will contain employee information and employee transactions.

From a source system, you have a flat extract that has the following fields:

- EmployeeID
- FirstName -
- LastName
- Recipient
- GrossAmount
- TransactionID
- GovernmentID
- NetAmountPaid
- TransactionDate

You need to design a star schema data model in an Azure Synapse Analytics dedicated SQL pool for the data mart.

Which two tables should you create? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. a dimension table for Transaction
- B. a dimension table for EmployeeTransaction
- C. a dimension table for Employee
- D. a fact table for Employee
- E. a fact table for Transaction



CertyIQ@gmail.com

You are designing a data mart for the human resources (HR) department at your company.

The data mart will contain employee information and employee transactions.

From a source system, you have a flat extract that has the following fields:

- EmployeeID
- FirstName -
- LastName
- Recipient
- GrossAmount
- TransactionID
- GovernmentID
- NetAmountPaid
- TransactionDate



CertyIQ@gmail.com

Explanation

Correct Answer: CE

C: Dimension tables contain attribute data that might change but usually changes infrequently.

E: Fact tables contain quantitative data that are commonly generated in a transactional system, and then loaded into the dedicated SQL pool.

You need to design a star schema data model in an Azure Synapse Analytics dedicated SQL pool for the data mart.

Which two tables should you create? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. a dimension table for Transaction
- B. a dimension table for EmployeeTransaction
- C. a dimension table for Employee**
- D. a fact table for Employee
- E. a fact table for Transaction**

You have an Azure Synapse Analytics Apache Spark pool named Pool1.

You plan to load JSON files from an Azure Data Lake Storage Gen2 container into the tables in Pool1. The structure and data types vary by file.

You need to load the files into the tables. The solution must maintain the source data types. What should you do?

- A. Use a Conditional Split transformation in an Azure Synapse data flow.
- B. Use a Get Metadata activity in Azure Data Factory.
- C. Load the data by using the OPENROWSET Transact-SQL command in an Azure Synapse Analytics serverless SQL pool.
- D. Load the data by using PySpark.

CertyIQ@gmail.com



You have an Azure Synapse Analytics Apache Spark pool named Pool1. You plan to load JSON files from an Azure Data Lake Storage Gen2 container into the tables in Pool1. The structure and data types vary by file. You need to load the files into the tables. The solution must maintain the source data types. What should you do?

- A. Use a Conditional Split transformation in an Azure Synapse data flow.
- B. Use a Get Metadata activity in Azure Data Factory.
- C. Load the data by using the OPENROWSET Transact-SQL command in an Azure Synapse Analytics serverless SQL pool.
- D. Load the data by using PySpark.**



CertyIQ@gmail.com

Explanation

Correct Answer: D

it's about Apache Spark pool, not serverless SQL pool.

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. Configure a global init script for workspace1.
- B. Create a cluster policy in workspace1.
- C. Upgrade workspace1 to the Premium pricing tier.
- D. Create a pool in workspace1.

CertyIQ@gmail.com



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. Configure a global init script for workspace1.
- B. Create a cluster policy in workspace1.
- C. Upgrade workspace1 to the Premium pricing tier.
- D. Create a pool in workspace1.**



CertyIQ@gmail.com

Explanation

Correct Answer: D

You can use Databricks Pools to Speed up your Data Pipelines and Scale Clusters Quickly.

Databricks Pools, a managed cache of virtual machine instances that enables clusters to start and scale 4 times faster.

HOTSPOT -

You have an Azure Data Lake Storage Gen2 service.

You need to design a data archiving solution that meets the following requirements:

- ☞ Data that is older than five years is accessed infrequently but must be available within one second when requested.
- ☞ Data that is older than seven years is NOT accessed.
- ☞ Costs must be minimized while maintaining the required availability.

How should you manage the data? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area: **Answer Area**

Data over five years old:

- Delete the blob.
- Move to archive storage.
- Move to cool storage.
- Move to hot storage.

Data over seven years old:

- Delete the blob.
- Move to archive storage.
- Move to cool storage.
- Move to hot storage.



CertyIQ@gmail.com

HOTSPOT -

You have an Azure Data Lake Storage Gen2 service.

You need to design a data archiving solution that meets the following requirements:

- Data that is older than five years is accessed infrequently but must be available within one second when requested.
- Data that is older than seven years is NOT accessed.
- Costs must be minimized while maintaining the required availability.

How should you manage the data? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area: **Answer Area**

Data over five years old:

- Delete the blob.
- Move to archive storage.
- Move to cool storage.
- Move to hot storage.

Data over seven years old:

- Delete the blob.
- Move to archive storage.
- Move to cool storage.
- Move to hot storage.

CertyIQ@gmail.com



Explanation

Correct Answer:

Box1: - Delete the blob

Box2: - Move to archive storage

You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Table1.

Table1 contains the following:

- One billion rows
- A clustered columnstore index
- A hash-distributed column named Product Key
- A column named Sales Date that is of the date data type and cannot be null

Thirty million rows will be added to Table1 each month.

You need to partition Table1 based on the Sales Date column. The solution must optimize query performance and data loading.

How often should you create a partition?

- A. once per month
- B. once per year
- C. once per day
- D. once per week



CertyIQ@gmail.com

You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Table1.

CertyIQ@gmail.com

Table1 contains the following:

- One billion rows
- A clustered columnstore index
- A hash-distributed column named Product Key
- A column named Sales Date that is of the date data type and cannot be null

Thirty million rows will be added to Table1 each month.

You need to partition Table1 based on the Sales Date column. The solution must optimize query performance and data loading.

How often should you create a partition?

- A. once per month
- B. once per year**
- C. once per day
- D. once per week

Explanation

Correct Answer: B

Need a minimum 1 million rows per distribution. Each table is 60 distributions. 30 millions rows is added each month.

Need 2 months to get a minimum of 1 million rows per distribution in a new partition.

You have an Azure Databricks workspace that contains a Delta Lake dimension table named Table1.

Table1 is a Type 2 slowly changing dimension (SCD) table.

You need to apply updates from a source table to Table1.

Which Apache Spark SQL operation should you use?

- A. CREATE
- B. UPDATE
- C. ALTER
- D. MERGE

CertyIQ@gmail.com



You have an Azure Databricks workspace that contains a Delta Lake dimension table named Table1.

Table1 is a Type 2 slowly changing dimension (SCD) table.

You need to apply updates from a source table to Table1.

Which Apache Spark SQL operation should you use?

- A. CREATE
- B. UPDATE
- C. ALTER
- D. MERGE**

CertyIQ@gmail.com



Explanation

Correct Answer: D

The Delta provides the ability to infer the schema for data input which further reduces the effort required in managing the schema changes. The Slowly Changing Dimension (SCD) Type 2 records all the changes made to each key in the dimensional table. These operations require updating the existing rows to mark the previous values of the keys as old and then inserting new rows as the latest values. Also, Given a source table with the updates and the target table with dimensional data, SCD Type 2 can be expressed with the merge.

You are designing an Azure Data Lake Storage solution that will transform raw JSON files for use in an analytical workload.

CertyIQ@gmail.com

You need to recommend a format for the transformed files. The solution must meet the following requirements:

- Contain information about the data types of each column in the files.
- Support querying a subset of columns in the files.
- Support read-heavy analytical workloads.
- Minimize the file size.

What should you recommend?

- A. JSON
- B. CSV
- C. Apache Avro
- D. Apache Parquet

You are designing an Azure Data Lake Storage solution that will transform raw JSON files for use in an analytical workload.

CertyIQ@gmail.com

You need to recommend a format for the transformed files. The solution must meet the following requirements:

- Contain information about the data types of each column in the files.
- Support querying a subset of columns in the files.
- Support read-heavy analytical workloads.
- Minimize the file size.

What should you recommend?

- A. JSON
- B. CSV
- C. Apache Avro
- D. Apache Parquet**

Explanation

Correct Answer: D

Parquet, an open-source file format for Hadoop, stores nested data structures in a flat columnar format.

Compared to a traditional approach where data is stored in a row-oriented approach, Parquet file format is more efficient in terms of storage and performance.

You plan to create a dimension table in Azure Synapse Analytics that will be less than 1 GB.

You need to create the table to meet the following requirements:

- Provide the fastest query time.
- Minimize data movement during queries.

Which type of table should you use?

- A. replicated
- B. hash distributed
- C. heap
- D. round-robin

CertyIQ@gmail.com



You plan to create a dimension table in Azure Synapse Analytics that will be less than 1 GB.

You need to create the table to meet the following requirements:

- Provide the fastest query time.
- Minimize data movement during queries.

Which type of table should you use?

A. replicated

B. hash distributed

C. heap

D. round-robin

CertyIQ@gmail.com



Explanation

Correct Answer: A

A replicated table has a full copy of the table accessible on each Compute node. Replicating a table removes the need to transfer data among Compute nodes before a join or aggregation. Since the table has multiple copies, replicated tables work best when the table size is less than 2 GB compressed. 2 GB is not a hard limit.

You are designing a dimension table in an Azure Synapse Analytics dedicated SQL pool. You need to create a surrogate key for the table. The solution must provide the fastest query performance.
What should you use for the surrogate key?

CertyIQ@gmail.com

- A. a GUID column
- B. a sequence object
- C. an IDENTITY column

You are designing a dimension table in an Azure Synapse Analytics dedicated SQL pool. You need to create a surrogate key for the table. The solution must provide the fastest query performance.
What should you use for the surrogate key?

- A. a GUID column
- B. a sequence object
- C. an IDENTITY column**

CertyIQ@gmail.com



Explanation

Correct Answer: C

Use IDENTITY to create surrogate keys using dedicated SQL pool in AzureSynapse Analytics.

Note: A surrogate key on a table is a column with a unique identifier for each row. The key is not generated from the table data. Data modelers like to create surrogate keys on their tables when they design data warehouse models. You can use the IDENTITY property to achieve this goal simply and effectively without affecting load performance.

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

Function:

Aggregate
Geospatial
Windowing



CertyIQ@gmail.com

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

Correct Answer:

Input type: Stream Reference

Function: Geospatial Aggregate Windowing



CertyIQ@gmail.com

Explanation

Correct Answer:

Input type: Stream -

You can process real-time IoT data streams with Azure Stream Analytics.

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.

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. Implement Azure Stream Analytics user-defined functions (UDF).
- C. Implement query parallelization by partitioning the data output.
- D. Scale the SU count for the job up.
- E. Scale the SU count for the job down.
- F. Implement query parallelization by partitioning the data input.

CertyIQ@gmail.com



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. Implement Azure Stream Analytics user-defined functions (UDF).
- C. Implement query parallelization by partitioning the data output.**
- D. Scale the SU count for the job up.
- E. Scale the SU count for the job down.
- F. Implement query parallelization by partitioning the data input.**

CertyIQ@gmail.com



Explanation

Correct Answer:

Partition input and output.

You need to trigger an Azure Data Factory pipeline when a file arrives in an Azure Data Lake Storage Gen2 container.

CertyIQ@gmail.com

Which resource provider should you enable?

- A. Microsoft.Sql
- B. Microsoft.Automation
- C. Microsoft.EventGrid
- D. Microsoft.EventHub

You need to trigger an Azure Data Factory pipeline when a file arrives in an Azure Data Lake Storage Gen2 container.

Which resource provider should you enable?



CertyIQ@gmail.com

- A. Microsoft.Sql
- B. Microsoft.Automation
- C. Microsoft.EventGrid**
- D. Microsoft.EventHub

Explanation

Correct Answer: C

Event-driven architecture (EDA) is a common data integration pattern that involves production, detection, consumption, and reaction to events. Data integration scenarios often require Data Factory customers to trigger pipelines based on events happening in storage account, such as the arrival or deletion of a file in Azure Blob Storage account. Data Factory natively integrates with Azure Event Grid, which lets you trigger pipelines on such events.

You plan to perform batch processing in Azure Databricks once daily.

CertyIQ@gmail.com

Which type of Databricks cluster should you use?

- A. High Concurrency
- B. automated
- C. interactive

You plan to perform batch processing in Azure Databricks once daily.

CertyIQ@gmail.com

Which type of Databricks cluster should you use?

- A. High Concurrency
- B. automated**
- C. interactive

Explanation

Correct Answer: B

Automated Databricks clusters are the best for jobs and automated batch processing.

Note: Azure Databricks has two types of clusters: interactive and automated. You use interactive clusters to analyze data collaboratively with interactive notebooks. You use automated clusters to run fast and robust automated jobs.

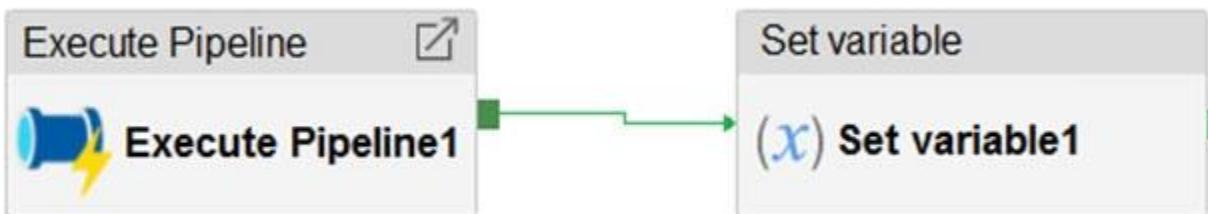
You have an Azure Data Factory instance that contains two pipelines named Pipeline1 and Pipeline2.

CertyIQ@gmail.com

Pipeline1 has the activities shown in the following exhibit.



Pipeline2 has the activities shown in the following exhibit.



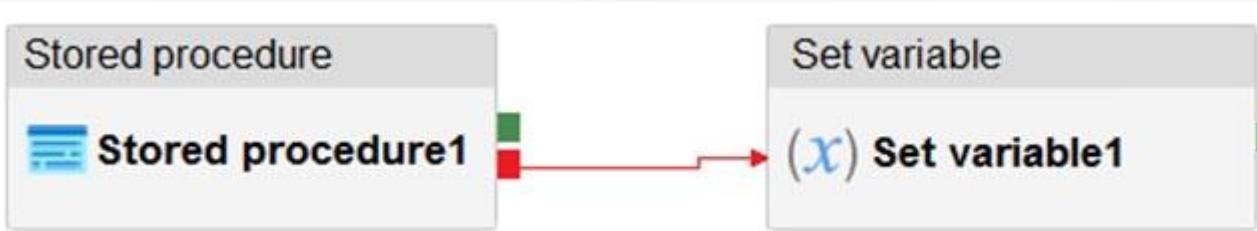
You execute Pipeline2, and Stored procedure1 in Pipeline1 fails.
What is the status of the pipeline runs?

- A. Pipeline1 and Pipeline2 succeeded.
- B. Pipeline1 and Pipeline2 failed.
- C. Pipeline1 succeeded and Pipeline2 failed.
- D. Pipeline1 failed and Pipeline2 succeeded.

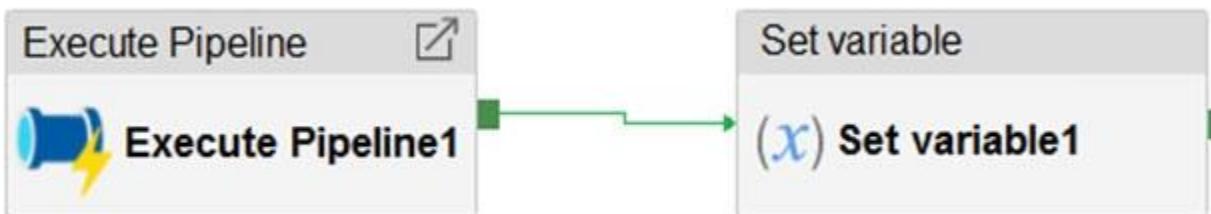
You have an Azure Data Factory instance that contains two pipelines named Pipeline1 and Pipeline2.

CertyIQ@gmail.com

Pipeline1 has the activities shown in the following exhibit.



Pipeline2 has the activities shown in the following exhibit.



You execute Pipeline2, and Stored procedure1 in Pipeline1 fails.
What is the status of the pipeline runs?

- A. Pipeline1 and Pipeline2 succeeded.
- B. Pipeline1 and Pipeline2 failed.
- C. Pipeline1 succeeded and Pipeline2 failed.
- D. Pipeline1 failed and Pipeline2 succeeded.

CertyIQ@gmail.com

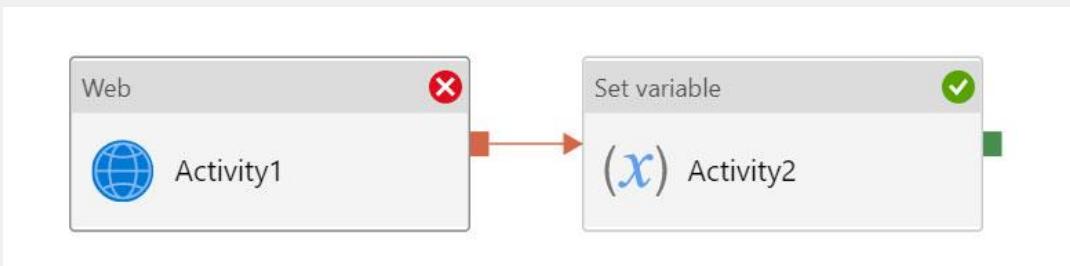
Explanation

Correct Answer: A

Activities are linked together via dependencies. A dependency has a condition of one of the following: Succeeded, Failed, Skipped, or Completed.

Consider Pipeline1:

If we have a pipeline with two activities where Activity2 has a failure dependency on Activity1, the pipeline will not fail just because Activity1 failed. If Activity1 fails and Activity2 succeeds, the pipeline will succeed. This scenario is treated as a try-catch block by Data Factory.



The failure dependency means this pipeline reports success.

Note:

If we have a pipeline containing Activity1 and Activity2, and Activity2 has a success dependency on Activity1, it will only execute if Activity1 is successful. In this scenario, if Activity1 fails, the pipeline will fail.

You have an Azure Data Factory that contains 10 pipelines.

You need to label each pipeline with its main purpose of either ingest, transform, or load.

The labels must be available for grouping and filtering when using the monitoring experience in Data Factory.

What should you add to each pipeline?

- A. a resource tag
- B. a correlation ID
- C. a run group ID
- D. an annotation

CertyIQ@gmail.com



You have an Azure Data Factory that contains 10 pipelines.

You need to label each pipeline with its main purpose of either ingest, transform, or load.

The labels must be available for grouping and filtering when using the monitoring experience in Data Factory.

What should you add to each pipeline?

- A. a resource tag
- B. a correlation ID
- C. a run group ID
- D. an annotation**

CertyIQ@gmail.com



Explanation

Correct Answer: D

Annotations are additional, informative tags that you can add to specific factory resources: pipelines, datasets, linked services, and triggers. By adding annotations, you can easily filter and search for specific factory resources.

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 Databricks
- C. Azure Stream Analytics
- D. Azure SQL Database

CertyIQ@gmail.com



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 Databricks**
- C. Azure Stream Analytics
- D. Azure SQL Database

CertyIQ@gmail.com



Explanation

Correct Answer: B

Azure Databricks supports near real-time data from Azure Event Hubs. And includes support for R, SQL, Python, Scala, and Java.

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 Stream Analytics solution that will analyze Twitter data.

You need to count the tweets in each 10-second window. The solution must ensure that each tweet is counted only once.

Solution: You use a hopping window that uses a hop size of 5 seconds and a window size 10 seconds.

Does this meet the goal?

A. Yes

B. No

CertyIQ@gmail.com



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 Stream Analytics solution that will analyze Twitter data.

You need to count the tweets in each 10-second window. The solution must ensure that each tweet is counted only once.

Solution: You use a hopping window that uses a hop size of 5 seconds and a window size 10 seconds.

Does this meet the goal?

A. Yes

B. No

CertyIQ@gmail.com



Explanation

Correct Answer: B

Instead use a tumbling window. Tumbling windows are a series of fixed-sized, non-overlapping and contiguous time intervals.

You need to schedule an Azure Data Factory pipeline to execute when a new file arrives in an Azure Data Lake Storage Gen2 container.

Which type of trigger should you use?

- A. on-demand
- B. tumbling window
- C. schedule
- D. event

CertyIQ@gmail.com



You need to schedule an Azure Data Factory pipeline to execute when a new file arrives in an Azure Data Lake Storage Gen2 container.

Which type of trigger should you use?

- A. on-demand
- B. tumbling window
- C. schedule**
- D. event

CertyIQ@gmail.com



Explanation

Correct Answer: C

In Azure Data Factory, continuous integration and delivery (CI/CD) means moving Data Factory pipelines from one environment (development, test, production) to another.

End

On a Mission to make 100+ Certified Cloud Engineer in next 30 Days

CertyIQ

CertyIQ@gmail.com

Please **Subscribe** to CertyIQ Channel
To get notification for upcoming dumps

SUBSCRIBE



DP-203 Important Que's



DP-203

Real Exam Questions & Answers



Latest Exam Questions
All Topics Covered-2023
Added New Que's



Get Certified Today!

New Course Covered

Subscribe

Question 41

CertyIQ

DRAG DROP -

You need to build a solution to ensure that users can query specific files in an Azure Data Lake Storage Gen2 account from an Azure Synapse Analytics serverless SQL pool.

Which three actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

NOTE: More than one order of answer choices is correct. You will receive credit for any of the correct orders you select.
Select and Place:

Actions

Answer Area

Create an external file format object

Create an external data source

Create a query that uses Create Table as Select

Create a table

Create an external table



Correct Answer:

Actions	Answer Area
	Create an external data source
	Create an external file format object
<p>Create a query that uses Create Table as Select</p> <p>Create a table</p>	<p>>Create an external table</p> 

Explanation:

Correct Answer:

Step 1: Create an external data source

You can create external tables in Synapse SQL pools via the following steps:

1. CREATE EXTERNAL DATA SOURCE to reference an external Azure storage and specify the credential that should be used to access the storage.

2. CREATE EXTERNAL FILE FORMAT to describe format of CSV or Parquet files.

3. CREATE EXTERNAL TABLE on top of the files placed on the data source with the same file format.

Step 2: Create an external file format object

Creating an external file format is a prerequisite for creating an external table.

Step 3: Create an external table

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables>

Question 42

CertyIQ

You are designing a data mart for the human resources (HR) department at your company. The data mart will contain employee information and employee transactions.

From a source system, you have a flat extract that has the following fields:

- EmployeeID
- FirstName -
- LastName
- Recipient
- GrossAmount
- TransactionID
- GovernmentID
- NetAmountPaid
- TransactionDate

You need to design a star schema data model in an Azure Synapse Analytics dedicated SQL pool for the data mart.

Which two tables should you create? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. a dimension table for Transaction
- B. a dimension table for EmployeeTransaction
- C. a dimension table for Employee
- D. a fact table for Employee
- E. a fact table for Transaction

Explanation:

Correct Answer: CE

C: Dimension tables contain attribute data that might change but usually changes infrequently. For example, a customer's name and address are stored in a dimension table and updated only when the customer's profile changes. To minimize the size of a large fact table, the customer's name and address don't need to be in every row of a fact table. Instead, the fact table and the dimension table can share a customer ID. A query can join the two tables to associate a customer's profile and transactions.

E: Fact tables contain quantitative data that are commonly generated in a transactional system, and then loaded into the dedicated SQL pool. For example, a retail business generates sales transactions every day, and then loads the data into a dedicated SQL pool fact table for analysis.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-overview>

Question 43

CertyIQ

You are designing a dimension table for a data warehouse. The table will track the value of the dimension attributes over time and preserve the history of the data by adding new rows as the data changes.

Which type of slowly changing dimension (SCD) should you use?

- A. Type 0
- B. Type 1
- C. Type 2
- D. Type 3

Explanation:

Correct Answer: C

A Type 2 SCD supports versioning of dimension members. Often the source system doesn't store versions, so the data warehouse load process detects and manages changes in a dimension table. In this case, the dimension table must use a surrogate key to provide a unique reference to a version of the dimension member. It also includes columns that define the date range validity of the version (for example, StartDate and EndDate) and possibly a flag column (for example, IsCurrent) to easily filter by current dimension members.

Incorrect Answers:

B: A Type 1 SCD always reflects the latest values, and when changes in source data are detected, the dimension table data is overwritten.

D: A Type 3 SCD supports storing two versions of a dimension member as separate columns. The table includes a column for the current value of a member plus either the original or previous value of the member. So Type 3 uses additional columns to track one key instance of history, rather than storing additional rows to track each change like in a Type 2 SCD.

Reference:

<https://docs.microsoft.com/en-us/learn/modules/populate-slowly-changing-dimensions-azure-synapse-analytics-pipelines/3-choose-between-dimension-types>

Question 44

CertyIQ

DRAG DROP -

You have data stored in thousands of CSV files in Azure Data Lake Storage Gen2. Each file has a header row followed by a properly formatted carriage return (/r) and line feed (/n).

You are implementing a pattern that batch loads the files daily into a dedicated SQL pool in Azure Synapse Analytics by using PolyBase.

You need to skip the header row when you import the files into the data warehouse. Before building the loading pattern, you need to prepare the required database objects in Azure Synapse Analytics.

Which three actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

NOTE: Each correct selection is worth one point

Select and Place:

Actions

Create a database scoped credential that uses Azure Active Directory Application and a Service Principal Key

Create an external data source that uses the abfs location

Use CREATE EXTERNAL TABLE AS SELECT (CETAS) and configure the reject options to specify reject values or percentages

Create an external file format and set the First_Row option

Answer Area



Actions

Create a database scoped credential that uses Azure Active Directory Application and a Service Principal Key

Create an external data source that uses the abfs location

Use CREATE EXTERNAL TABLE AS SELECT (CETAS) and configure the reject options to specify reject values or percentages

Create an external file format and set the First_Row option

Answer Area

Create a database scoped credential that uses Azure Active Directory Application and a Service Principal Key

Create an external data source that uses the abfs location

Create an external file format and set the First_Row option

Use CREATE EXTERNAL TABLE AS SELECT (CETAS) and configure the reject options to specify reject values or percentages

Explanation:

- Correct Answer: 1) create database scoped credentials
- 2) create external source
- 3) create file format
- 4) create external table (it not supports CTAS)

Question 45

CertyIQ

HOTSPOT -

You are building an Azure Synapse Analytics dedicated SQL pool that will contain a fact table for transactions from the first half of the year 2020.

You need to ensure that the table meets the following requirements:

⇒ Minimizes the processing time to delete data that is older than 10 years

⇒ Minimizes the I/O for queries that use year-to-date values

How should you complete the Transact-SQL statement? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

```
CREATE TABLE [dbo].[FactTransaction]
```

```
(
```

```
    [TransactionTypeID] int NOT NULL
```

```
,
```

```
    [TransactionDateID] int NOT NULL
```

```
,
```

```
    [CustomerID] int NOT NULL
```

```
,
```

```
    [RecipientID] int NOT NULL
```

```
,
```

```
    [Amount] money NOT NU::
```

```
)
```

```
WITH
```

```
(
```

CLUSTERED COLUMNSTORE INDEX	▼
DISTRIBUTION	▼
PARTITION	▼
TRUNCATE_TARGET	▼

```
[TransactionDateID]
```

```
[TransactionDateID], [TransactionTypeID]
```

```
HASH([TransactionTypeID])
```

```
ROUND_ROBIN
```

RANGE RIGHT FOR VALUES

```
(20200101, 20200201, 20200301, 20200401, 20200501, 20200601)
```

Answer Area

```
CREATE TABLE [dbo].[FactTransaction]
(
    [TransactionTypeID] int NOT NULL
    , [TransactionDateID] int NOT NULL
    , [CustomerID] int NOT NULL
    , [RecipientID] int NOT NULL
    , [Amount] money NOT NU:::
)
```

WITH

(▼	
CLUSTERED COLUMNSTORE INDEX		
DISTRIBUTION		
PARTITION		
TRUNCATE_TARGET		
(▼	RANGE RIGHT FOR VALUES
[TransactionDateID]		
[TransactionDateID], [TransactionTypeID]		
HASH([TransactionTypeID])		
ROUND_ROBIN		

(20200101,20200201,20200301,20200401,20200501,20200601)

Explanation:

Correct Answer:

Box 1: PARTITION -

RANGE RIGHT FOR VALUES is used with PARTITION.

Part 2: [TransactionDateID]

Partition on the date column.

Example: Creating a RANGE RIGHT partition function on a datetime column

The following partition function partitions a table or index into 12 partitions, one for each month of a year's worth of values in a datetime column.

```
CREATE PARTITION FUNCTION [myDateRangePF1] (datetime)
AS RANGE RIGHT FOR VALUES ('20030201', '20030301', '20030401',
'20030501', '20030601', '20030701', '20030801',
'20030901', '20031001', '20031101', '20031201');
```

Reference:

<https://docs.microsoft.com/en-us/sql/t-sql/statements/create-partition-function-transact-sql>

Question 46

CertyIQ

You are performing exploratory analysis of the bus fare data in an Azure Data Lake Storage Gen2 account by using an Azure Synapse Analytics serverless SQL pool.

You execute the Transact-SQL query shown in the following exhibit.

```

SELECT
    payment_type,
    SUM(fare_amount) AS fare_total
FROM OPENROWSET (
    BULK 'csv/busfare/tripdata_2020*.csv',
    DATA_SOURCE = 'BusData',
    FORMAT = 'CSV', PARSER_VERSION = '2.0',
    FIRSTROW = 2
)
WITH (
    payment_type INT 10,
    fare_amount FLOAT 11
) AS nyc
GROUP BY payment_type
ORDER BY payment_type;

```

What do the query results include?

- A. Only CSV files in the tripdata_2020 subfolder.
- B. All files that have file names that beginning with "tripdata_2020".
- C. All CSV files that have file names that contain "tripdata_2020".
- D. Only CSV that have file names that beginning with "tripdata_2020"

Explanation:

Correct Answer :D

Question 47

CertyIQ

DRAG DROP -

You use PySpark in Azure Databricks to parse the following JSON input.

```
{
  "persons": [
    {
      "name": "Keith",
      "age": 30,
      "dogs": ["Fido", "Fluffy"]
    },
    {
      "name": "Donna",
      "age": 46,
      "dogs": ["Spot"]
    }
  ]
}
```

You need to output the data in the following tabular format.

owner	age	dog
Keith	30	Fido
Keith	30	Fluffy
Donna	46	Spot

How should you complete the PySpark code? To answer, drag the appropriate values to the correct targets. Each value 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:

Values	Answer Area
alias	dbutils.fs.put("/tmp/source.json", source_json, True) source_df = spark.read.option("multiline", "true").json("/tmp/source.json")
array_union	persons = source_df. <input type="text"/> Value <input type="text"/> Value ("persons").alias("persons")
createDataFrame	persons_dogs = persons.select(col("persons.name").alias("owner"), col("persons.age").alias("age"), explode <input type="text"/> Value ("dog"))
explode	("persons-dogs"). display(persons_dogs)
select	
translate	

Values	Answer Area
array_union	dbutils.fs.put("/tmp/source.json", source_json, True) source_df = spark.read.option("multiline", "true").json("/tmp/source.json")
createDataFrame	persons = source_df. select <input type="text"/> explode <input type="text"/> ("persons").alias("persons") persons_dogs = persons.select(col("persons.name").alias("owner"), col("persons.age").alias("age"), explode <input type="text"/> alias <input type="text"/> ("dog")) ("persons-dogs"). display(persons_dogs)
translate	

Explanation:

Correct Answer:

Box 1: select -

Box 2: explode -

Box 3: alias -

pyspark.sql.Column.alias returns this column aliased with a new name or names (in the case of expressions that return more than one column, such as explode).

Reference:

<https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.sql.Column.alias.html>

<https://docs.microsoft.com/en-us/azure/databricks/sql/language-manual/functions/explode>

Question 48

CertyIQ

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

Answer Area

First week:

Archive
Cool
Hot

After one month:

Archive
Cool
Hot

After one year:

Archive
Cool
Hot

Correct Answer:

Explanation:

Correct Answer:

Box 1: Hot -

Hot tier - An online tier optimized for storing data that is accessed or modified frequently. The Hot tier has the highest storage costs, but the lowest access costs.

Box 2: Cool -

Cool tier - An online tier optimized for storing data that is infrequently accessed or modified. Data in the Cool tier should be stored for a minimum of 30 days. The Cool tier has lower storage costs and higher access costs compared to the Hot tier.

Box 3: Cool -

Not Archive tier - An offline tier optimized for storing data that is rarely accessed, and that has flexible latency

requirements, on the order of hours. Data in the Archive tier should be stored for a minimum of 180 days.

	Premium performance	Hot tier	Cool tier	Archive tier
Availability	99.9%	99.9%	99%	Offline
Availability (RA-GRS reads)	N/A	99.99%	99.9%	Offline
Usage charges	Higher storage costs, lower access, and transaction cost	Higher storage costs, lower access, and transaction costs	Lower storage costs, higher access, and transaction costs	Lowest storage costs, highest access, and transaction costs
Minimum object size	N/A	N/A	N/A	N/A
Minimum storage duration	N/A	N/A	30 days ¹	180 days
Latency (Time to first byte)	Single-digit milliseconds	milliseconds	milliseconds	hours ²

Reference:

<https://docs.microsoft.com/en-us/azure/storage/blobs/access-tiers-overview> <https://www.altaro.com/hyper-v/azure-archive-storage/>

Question 49

CertyIQ

You have an Azure Synapse Analytics Apache Spark pool named Pool1.

You plan to load JSON files from an Azure Data Lake Storage Gen2 container into the tables in Pool1. The structure and data types vary by file.

You need to load the files into the tables. The solution must maintain the source data types.

What should you do?

- A. Use a Conditional Split transformation in an Azure Synapse data flow.
- B. Use a Get Metadata activity in Azure Data Factory.
- C. Load the data by using the OPENROWSET Transact-SQL command in an Azure Synapse Analytics serverless SQL pool.
- D. Load the data by using PySpark.

Explanation:

Correct Answer:D it's about Apache Spark pool, not serverless SQL pool.

Question 50

CertyIQ

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. Configure a global init script for workspace1.
- B. Create a cluster policy in workspace1.
- C. Upgrade workspace1 to the Premium pricing tier.
- D. Create a pool in workspace1.

Explanation:

Correct Answer: **D**

You can use Databricks Pools to Speed up your Data Pipelines and Scale Clusters Quickly.

Databricks Pools, a managed cache of virtual machine instances that enables clusters to start and scale 4 times faster.

Reference:

<https://databricks.com/blog/2019/11/11/databricks-pools-speed-up-data-pipelines.html>

Question 51

CertyIQ

HOTSPOT -

You are building an Azure Stream Analytics job that queries reference data from a product catalog file. The file is updated daily.

The reference data input details for the file are shown in the Input exhibit. (Click the Input tab.)

Input Details ×

products

Test Delete

Container

Create new Use existing

Path pattern (i)

Date format

Time format

Event serialization format * (i)

Delimiter (i)

Encoding (i)

Save

- If the chosen resource and the stream analytics job are located in different regions, you will be billed to move data between regions.

The storage account container view is shown in the Refdata exhibit. (Click the Refdata tab.)

refdata
Container

Search (Ctrl + /) «

Upload Add Directory Refresh Rename Delete

Overview Access Control (IAM)

Settings

- Access policy
- Properties
- Metadata

Authentication method: Access key ([Switch to Azure AD User Account](#))
Location: refdata / 2020-03-20

Search blobs by prefix (case-sensitive)

Name
<input type="checkbox"/> [..]
<input type="checkbox"/> product.csv

You need to configure the Stream Analytics job to pick up the new reference data.

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Path pattern:

- {date}/product.csv
- {date}/{time}/product.csv
- product.csv
- */product.csv

Date format:

- MM/DD/YYYY
- YYYY/MM/DD
- YYYY-DD-MM
- YYYY-MM-DD

Answer Area

Path pattern:

{date}/product.csv
{date}/{time}/product.csv
product.csv
*/product.csv

Date format:

MM/DD/YYYY
YYYY/MM/DD
YYYY-DD-MM
YYYY-MM-DD

Explanation:

Correct Answer: 1. {date}/{time}/product.csv More detailed things should be put at the last. 2. YYYY-MM-DD if you choose YYYY/MM/DD, the system will think this is a file path.

Question 52

CertyIQ

HOTSPOT -

You have the following Azure Stream Analytics query.

WITH

```
step1 AS (SELECT *
    FROM input1
    PARTITION BY StateID
    INTO 10),
step2 AS (SELECT *
    FROM input2
    PARTITION BY StateID
    INTO 10)

SELECT *
INTO output
FROM step1
PARTITION BY StateID
UNION
SELECT * INTO output
    FROM step2
    PARTITION BY StateID
```

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 query combines two streams of partitioned data.	<input type="radio"/>	<input type="radio"/>
The stream scheme key and count must match the output scheme.	<input type="radio"/>	<input type="radio"/>
Providing 60 streaming units will optimize the performance of the query.	<input checked="" type="radio"/>	<input type="radio"/>

Answer Area

Statements	Yes	No
Correct Answer: The query combines two streams of partitioned data.	<input checked="" type="radio"/>	<input type="radio"/>
The stream scheme key and count must match the output scheme.	<input checked="" type="radio"/>	<input type="radio"/>
Providing 60 streaming units will optimize the performance of the query.	<input checked="" type="radio"/>	<input type="radio"/>

Explanation:

Correct Answer: YES.

Since it does use a UNION and UNION combines. No matter it repartitions the result is the combination of two sources, a UNION of two sources. Am I missing something here?

Box 2: Yes -

When joining two streams of data explicitly repartitioned, these streams must have the same partition key and partition count.

Box 3: Yes -

Streaming Units (SUs) represents the computing resources that are allocated to execute a Stream Analytics job. The higher the number of SUs, the more CPU and memory resources are allocated for your job.

In general, the best practice is to start with 6 SUs for queries that don't use PARTITION BY.

Here there are 10 partitions, so $6 \times 10 = 60$ SUs is good.

Note: Remember, Streaming Unit (SU) count, which is the unit of scale for Azure Stream Analytics, must be adjusted so the number of physical resources available to the job can fit the partitioned flow. In general, six SUs is a good number to assign to each partition. In case there are insufficient resources assigned to the job, the system will only apply the repartition if it benefits the job.

Reference:

<https://azure.microsoft.com/en-in/blog/maximize-throughput-with-repartitioning-in-azure-stream-analytics/>

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-streaming-unit-consumption>

HOTSPOT -

You are building a database in an Azure Synapse Analytics serverless SQL pool.
You have data stored in Parquet files in an Azure Data Lake Storege Gen2 container.
Records are structured as shown in the following sample.

```
{  
  "id": 123,  
  "address_housenumber": "19c",  
  "address_line": "Memory Lane",  
  "applicant1_name": "Jane",  
  "applicant2_name": "Dev"  
}
```

The records contain two applicants at most.

You need to build a table that includes only the address fields.

How should you complete the Transact-SQL statement? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

```
applications  
CREATE EXTERNAL TABLE  
CREATE TABLE  
CREATE VIEW  
  
WITH (  
    LOCATION = 'applications/',  
    DATA_SOURCE = applications_ds,  
    FILE_FORMAT = applications_file_format  
)  
AS  
SELECT id, [address_housenumber] as addresshousenumber, [address_line1] as addressline1  
FROM  
    (BULK 'https://contosol1.dfs.core.windows.net/applications/year=/*/*.parquet',  
CROSS APPLY  
OPENJSON  
OPENROWSET  
  
FORMAT='PARQUET') AS [r]  
GO
```

Answer Area

```
applications  
CREATE EXTERNAL TABLE  
CREATE TABLE  
CREATE VIEW  
  
WITH (  
    LOCATION = 'applications/',  
    DATA_SOURCE = applications_ds,  
    FILE_FORMAT = applications_file_format  
)  
AS  
SELECT id, [address_housenumber] as addresshousenumber, [address_line1] as addressline1  
FROM  
    (BULK 'https://contosol1.dfs.core.windows.net/applications/year=/*/*.parquet',  
CROSS APPLY  
OPENJSON  
OPENROWSET  
  
FORMAT='PARQUET') AS [r]  
GO
```

Explanation:

Correct Answer:

Box 1: CREATE EXTERNAL TABLE -

An external table points to data located in Hadoop, Azure Storage blob, or Azure Data Lake Storage. External tables are used to read data from files or write data to files in Azure Storage. With Synapse SQL, you can use external tables to read external data using dedicated SQL pool or serverless SQL pool.

Syntax:

```
CREATE EXTERNAL TABLE { database_name.schema_name.table_name | schema_name.table_name | table_name }
( <column_definition> [ ,...n ] )
WITH (
    LOCATION = 'folder_or_filepath',
    DATA_SOURCE = external_data_source_name,
    FILE_FORMAT = external_file_format_name)
```

Box 2. OPENROWSET -

When using serverless SQL pool, CETAS is used to create an external table and export query results to Azure Storage Blob or Azure Data Lake Storage Gen2.

Example:

AS -

```
SELECT decennialTime, stateName, SUM(population) AS population
```

FROM -

```
OPENROWSET(BULK
'https://azureopendatastorage.blob.core.windows.net/censusdatacontainer/release/us_population_county/year=/*/*.parquet',
FORMAT='PARQUET') AS [r]
GROUP BY decennialTime, stateName
```

GO -

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables>

Question 54

CertyIQ

HOTSPOT -

You have an Azure Synapse Analytics dedicated SQL pool named Pool1 and an Azure Data Lake Storage Gen2 account named Account1.

You plan to access the files in Account1 by using an external table.

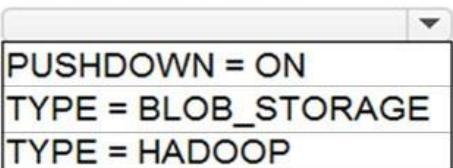
You need to create a data source in Pool1 that you can reference when you create the external table.

How should you complete the Transact-SQL statement? To answer, select the appropriate options in the answer area.

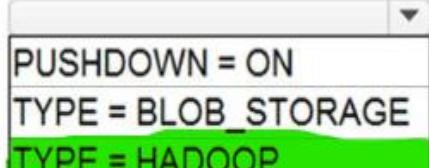
NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

```
CREATE EXTERNAL DATA SOURCE source1
WITH
( LOCATION = 'https://account1.
    
    .core.windons.net',
    
)
)
```

Answer Area

```
CREATE EXTERNAL DATA SOURCE source1
WITH
( LOCATION = 'https://account1.
    
    .core.windons.net',
    
)
)
```

Explanation:

Correct Answer: 1. dfs when the Storage Account is an ADLS
2. Hadoop, it is stated in the documentation below

https://docs.microsoft.com/en-us/sql/t-sql/statements/create-external-data-source-transact-sql?view=sql-server-ver16&tabs=dedicated#type--hadoop--blob_storage--2

Question 55

CertyIQ

You have an Azure subscription that contains an Azure Blob Storage account named storage1 and an Azure Synapse Analytics dedicated SQL pool named Pool1.

You need to store data in storage1. The data will be read by Pool1. The solution must meet the following requirements:

Enable Pool1 to skip columns and rows that are unnecessary in a query.

- Automatically create column statistics.

⇒ Minimize the size of files.

Which type of file should you use?

- A. JSON
- **B. Parquet**
- C. Avro
- D. CSV

Explanation:

Correct Answer: **B**

Automatic creation of statistics is turned on for Parquet files. For CSV files, you need to create statistics manually until automatic creation of CSV files statistics is supported.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-statistics>

Question 56

CertyIQ

DRAG DROP -

You plan to create a table in an Azure Synapse Analytics dedicated SQL pool.

Data in the table will be retained for five years. Once a year, data that is older than five years will be deleted.

You need to ensure that the data is distributed evenly across partitions. The solution must minimize the amount of time required to delete old data.

How should you complete the Transact-SQL statement? To answer, drag the appropriate values to the correct targets. Each value 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:

Values	Answer Area
CustomerKey	<pre>CREATE TABLE [dbo].[FactSales] ([ProductKey] int NOT NULL , [OrderDateKey] int NOT NULL , [CustomerKey] int NOT NULL , [SalesOrderNumber] nvarchar (20) NOT NULL , [OrderQuantity] smallint NOT NULL , [UnitPrice] money NOT NULL) WITH (CLUSTERED COLUMNSTORE INDEX , DISTRIBUTION = <input type="text" value="Value"/> ([ProductKey]) , PARTITION ([<input type="text" value="Value"/>] RANGE RIGHT FOR VALUES (20170101,20180101,20190101,20200101,20210101))</pre>
HASH	
ROUND_ROBIN	
REPLICATE	
OrderDateKey	
SalesOrderNumber	

Values	Answer Area
CustomerKey	<code>CREATE TABLE [dbo].[FactSales]</code>
	<code>(</code>
	<code>[ProductKey] int NOT NULL</code>
	<code>[OrderDateKey] int NOT NULL</code>
	<code>[CustomerKey] int NOT NULL</code>
	<code>[SalesOrderNumber] nvarchar (20) NOT NULL</code>
	<code>[OrderQuantity] smallint NOT NULL</code>
	<code>[UnitPrice] money NOT NULL</code>
	<code>)</code>
	<code>WITH</code>
	<code>(CLUSTERED COLUMNSTORE INDEX</code>
	<code>, DISTRIBUTION = ROUND_ROBIN ([ProductKey])</code>
	<code>, PARTITION ([OrderDateKey] RANGE RIGHT FOR VALUES</code>
SalesOrderNumber	<code>(20170101.20180101.20190101.20200101.20210101)</code>

Explanation:

Correct Answer:

While you are right, that Round-Robin guarantees an even distribution, it is only recommended to use on small tables < 2 GB (see your link). Using the Hash of the ProductKey will also allow for an even distribution but in a more efficient manner. Also, the Syntax here would be wrong if you would insert Round-Robin. As in that case it would only say: "DISTRIBUTION = ROUND-ROBIN" (no ProductKey)

Box 2: OrderDateKey -

In most cases, table partitions are created on a date column.

A way to eliminate rollbacks is to use Metadata Only operations like partition switching for data management. For example, rather than execute a DELETE statement to delete all rows in a table where the order_date was in October of 2001, you could partition your data early. Then you can switch out the partition with data for an empty partition from another table.

Reference:

<https://docs.microsoft.com/en-us/sql/t-sql/statements/create-table-azure-sql-data-warehouse>

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/best-practices-dedicated-sql-pool>

Question 57

CertyIQ

HOTSPOT -

You have an Azure Data Lake Storage Gen2 service.

You need to design a data archiving solution that meets the following requirements:

- ☞ Data that is older than five years is accessed infrequently but must be available within one second when requested.
- ☞ Data that is older than seven years is NOT accessed.
- ☞ Costs must be minimized while maintaining the required availability.

How should you manage the data? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Data over five years old:

- Delete the blob.
- Move to archive storage.
- Move to cool storage.
- Move to hot storage.

Data over seven years old:

- Delete the blob.
- Move to archive storage.
- Move to cool storage.
- Move to hot storage.

Answer Area

Data over five years old:

- Delete the blob.**
- Move to archive storage.
- Move to cool storage.
- Move to hot storage.

Data over seven years old:

- Delete the blob.
- Move to archive storage.**
- Move to cool storage.
- Move to hot storage.

Explanation:

Correct Answer: Deleting data older than 7 years is not an option available in the answer list. Be careful of the gotcha: 'Delete the blob' is an option but it would delete all the data, included the ones that are e.g. 5 years old. So you can't choose that answer. So the next best thing to do is to put it into archive.

Question 58

CertyIQ

HOTSPOT -

You plan to create an Azure Data Lake Storage Gen2 account.

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

⇒ Provides the highest degree of data resiliency

⇒ Ensures that content remains available for writes if a primary data center fails

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Replication mechanism:

- | |
|--|
| Change feed |
| Zone-redundant storage (ZRS) |
| Read-access geo-redundant storage (RA-GRS) |
| Read-access geo-zone-redundant storage (RA-GZRS) |

Failover process:

- | |
|---|
| Failover initiated by Microsoft |
| Failover manually initiated by the customer |
| Failover automatically initiated by an Azure Automation job |

Answer Area

Replication mechanism:

- | |
|--|
| Change feed |
| Zone-redundant storage (ZRS) |
| Read-access geo-redundant storage (RA-GRS) |
| Read-access geo-zone-redundant storage (RA-GZRS) |

Failover process:

- | |
|---|
| Failover initiated by Microsoft |
| Failover manually initiated by the customer |
| Failover automatically initiated by an Azure Automation job |

Explanation:

Correct Answer: Microsoft recommends RA-GZRS for maximum availability and durability for your applications." Failover initiated by Microsoft. Customer-managed account failover is not yet supported in accounts that have a hierarchical namespace (Azure Data Lake Storage Gen2). To learn more, see Blob storage features available in Azure Data Lake Storage Gen2.

You need to implement a Type 3 slowly changing dimension (SCD) for product category data in an Azure Synapse Analytics dedicated SQL pool.

You have a table that was created by using the following Transact-SQL statement.

```
CREATE TABLE [DB0].[DimProduct] (
    [ProductKey] [int] IDENTITY(1,1) NOT NULL,
    [ProductSourceID] [int] NOT NULL,
    [ProductName] [nvarchar](100) NOT NULL,
    [Color] [nvarchar](15) NULL,
    [SellStartDate] [date] NOT NULL,
    [SellEndDate] [date] NULL,
    [RowInsertedDateTime] [datetime] NOT NULL,
    [RowUpdatedDateTime] [datetime] NOT NULL,
    [ETLAuditID] [int] NOT NULL
)
```

Which two columns should you add to the table? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

A.

[EffectiveEndDate] [datetime] NULL,

B.

[CurrentProductCategory] [nvarchar](100) NOT NULL,

C.

[ProductCategory] [nvarchar](100) NOT NULL,

D.

[EffectiveStartDate] [datetime] NOT NULL,

E.

[OriginalProductCategory] [nvarchar](100) NOT NULL,

Explanation:

Correct Answer: BE

A Type 3 SCD supports storing two versions of a dimension member as separate columns. The table includes a column for the current value of a member plus either the original or previous value of the member. So Type 3 uses additional columns to track one key instance of history, rather than storing additional rows to track each change like in a Type 2 SCD.

This type of tracking may be used for one or two columns in a dimension table. It is not common to use it for many members of the same table. It is often used in combination with Type 1 or Type 2 members.

CustomerID	FirstName	LastName	CurrentEmail	OriginalEmail	CompanyName	InsertedDate	ModifiedDate
2	Keith	Harris	keith0@aw.com	keith0@aw.com	Progressive Sports	2021-03-20	2021-03-20
3	Donna	Carreras	donna0@aw.com	donna0@aw.com	A Bike Store	2021-03-20	2021-03-20

CustomerID	FirstName	LastName	CurrentEmail	OriginalEmail	CompanyName	InsertedDate	ModifiedDate
2	Keith	Harris	keith0@aw.com	keith0@aw.com	Progressive Sports	2021-03-20	2021-03-20
3	Donna	Carreras	dc3@aw.com	donna0@aw.com	A Bike Store	2021-03-20	2021-03-22

Reference:

<https://k21academy.com/microsoft-azure/azure-data-engineer-dp203-q-a-day-2-live-session-review/>

prw709528

Question 60

CertyIQ

DRAG DROP -

You have an Azure subscription.

You plan to build a data warehouse in an Azure Synapse Analytics dedicated SQL pool named pool1 that will contain staging tables and a dimensional model.

Pool1 will contain the following tables.

Name	Number of rows	Update frequency	Description
Common.Date	7,300	New rows inserted yearly	<ul style="list-style-type: none"> Contains one row per date for the last 20 years Contains columns named Year, Month, Quarter, and IsWeekend
Marketing.WebSessions	1,500,500,000	Hourly inserts and updates	Fact table that contains counts of and updates sessions and page views, including foreign key values for date, channel, device, and medium
Staging.WebSessions	300,000	Hourly truncation and inserts	Staging table for web session data, truncation and including descriptive fields for inserts channel, device, and medium

You need to design the table storage for pool1. The solution must meet the following requirements:

☞ Maximize the performance of data loading operations to Staging.WebSessions.

☞ Minimize query times for reporting queries against the dimensional model.

Which type of table distribution should you use for each table? To answer, drag the appropriate table distribution types to the correct tables. Each table distribution type 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:

Table distribution types

- Hash
- Replicated
- Round-robin

Answer Area

Common.Data:

Marketing.Web.Sessions:

Staging. Web.Sessions:

Correct Answer:

Table distribution types

- Hash
- Replicated
- Round-robin

Answer Area

Common.Data:

Replicated

Marketing.Web.Sessions:

Hash

Staging. Web.Sessions:

Round-robin

Explanation:

Correct Answer:

Box 1: Replicated -

The best table storage option for a small table is to replicate it across all the Compute nodes.

Box 2: Hash -

Hash-distribution improves query performance on large fact tables.

Box 3: Round-robin -

Round-robin distribution is useful for improving loading speed.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute>

Question 61

CertyIQ

HOTSPOT -

You have an Azure Synapse Analytics dedicated SQL pool.

You need to create a table named FactInternetSales that will be a large fact table in a dimensional model.

FactInternetSales will contain 100 million rows and two columns named SalesAmount and OrderQuantity. Queries executed on FactInternetSales will aggregate the values in SalesAmount and OrderQuantity from the last year for a specific product. The solution must minimize the data size and query execution time.

How should you complete the code? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

```
CREATE TABLE [dbo].[FactInternetSales]
(
    [ProductKey] int NOT NULL
    , [OrderDateKey] int NOT NULL
    , [CustomerKey] int NOT NULL
    , [PromotionKey] int NOT NULL
    , [SalesOrderNumber] nvarchar(20) NOT NULL
    , [OrderQuantity] smallint NOT NULL
    , [UnitPrice] money NOT NULL
    , [SalesAmount] money NOT NULL
)
WITH
(
    CLUSTERED COLUMNSTORE INDEX
    CLUSTERED INDEX ([OrderDateKey])
    HEAP
    INDEX on [ProductKey]
    , DISTRIBUTION =
);

```

Hash([OrderDateKey])
Hash([ProductKey])
REPLICATE
ROUND_ROBIN

Answer Area

```
CREATE TABLE [dbo].[FactInternetSales]
(
    [ProductKey] int NOT NULL
    , [OrderDateKey] int NOT NULL
    , [CustomerKey] int NOT NULL
    , [PromotionKey] int NOT NULL
    , [SalesOrderNumber] nvarchar(20) NOT NULL
    , [OrderQuantity] smallint NOT NULL
    , [UnitPrice] money NOT NULL
    , [SalesAmount] money NOT NULL
)
WITH
(
    CLUSTERED COLUMNSTORE INDEX
    CLUSTERED INDEX ([OrderDateKey])
    HEAP
    INDEX on [ProductKey]
    , DISTRIBUTION =
);

```

The code shows the creation of a table named FactInternetSales with various columns and constraints. It includes a clustered columnstore index, a clustered index on OrderDateKey, and a heap. The distribution of data is specified using Hash functions for OrderDateKey and ProductKey, and options for Replicate or RoundRobin.

Explanation:

Correct Answer:

Box 1: (CLUSTERED COLUMNSTORE INDEX)

CLUSTERED COLUMNSTORE INDEX -

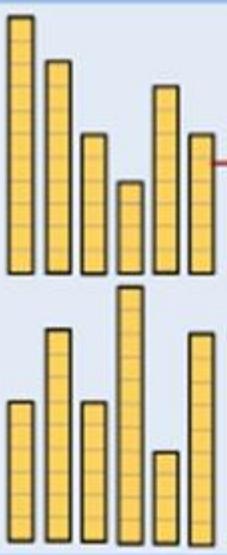
Columnstore indexes are the standard for storing and querying large data warehousing fact tables. This index uses column-based data storage and query processing to achieve gains up to 10 times the query performance in your data warehouse over traditional row-oriented storage. You can also achieve gains up to 10 times the data compression over the uncompressed data size. Beginning with SQL Server 2016 (13.x) SP1, columnstore indexes enable operational analytics: the ability to run performant real-time analytics on a transactional workload.

Note: Clustered columnstore index

A clustered columnstore index is the physical storage for the entire table.

Clustered Columnstore Index – Physical Storage

Columnstore



Compressed column segments

+



Rows in the index, but not in the columnstore

Deltastore

To reduce fragmentation of the column segments and improve performance, the columnstore index might store some data temporarily into a clustered index called a deltastore and a B-tree list of IDs for deleted rows. The deltastore operations are handled behind the scenes. To return the correct query results, the clustered columnstore index combines query results from both the columnstore and the deltastore.

Question 62

CertyIQ

You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Table1. Table1 contains the following:

- ☞ One billion rows
- ☞ A clustered columnstore index
- ☞ A hash-distributed column named Product Key
- ☞ A column named Sales Date that is of the date data type and cannot be null

Thirty million rows will be added to Table1 each month.

You need to partition Table1 based on the Sales Date column. The solution must optimize query performance and data loading.

How often should you create a partition?

- A. once per month
- B. once per year
- C. once per day
- D. once per week

Explanation:

Correct Answer: : B

Need a minimum 1 million rows per distribution. Each table is 60 distributions. 30 millions rows is added each month. Need 2 months to get a minimum of 1 million rows per distribution in a new partition.

Note: When creating partitions on clustered columnstore tables, it is important to consider how many rows belong to each partition. For optimal compression and performance of clustered columnstore tables, a minimum of 1 million rows per distribution and partition is needed. Before partitions are created, dedicated SQL pool already divides each table into 60 distributions.

Any partitioning added to a table is in addition to the distributions created behind the scenes. Using this example, if the sales fact table contained 36 monthly partitions, and given that a dedicated SQL pool has 60 distributions, then the sales fact table should contain 60 million rows per month, or 2.1 billion rows when all months are populated. If a table contains fewer than the recommended minimum number of rows per partition, consider using fewer partitions in order to increase the number of rows per partition.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-partition>

Question 63

CertyIQ

You have an Azure Databricks workspace that contains a Delta Lake dimension table named Table1.

Table1 is a Type 2 slowly changing dimension (SCD) table.

You need to apply updates from a source table to Table1.

Which Apache Spark SQL operation should you use?

- A. CREATE
- B. UPDATE
- C. ALTER
- D. MERGE

Explanation:

Correct Answer: D

The Delta provides the ability to infer the schema for data input which further reduces the effort required in managing the schema changes. The Slowly Changing Dimension (SCD) Type 2 records all the changes made to each key in the dimensional table. These operations require updating the existing rows to mark the previous values of the keys as old and then inserting new rows as the latest values. Also, Given a source table with the updates and the target table with dimensional data,

SCD Type 2 can be expressed with the merge.

Example:

```
// Implementing SCD Type 2 operation using merge function
customersTable
.as("customers")
.merge(
stagedUpdates.as("staged_updates"),
"customers.customerId = mergeKey")
.whenMatched("customers.current = true AND customers.address <> staged_updates.address")
.updateExpr(Map(
"current" -> "false",
"endDate" -> "staged_updates.effectiveDate"))
.whenNotMatched()
.insertExpr(Map(
"customerid" -> "staged_updates.customerId",
"address" -> "staged_updates.address",
"current" -> "true",
"effectiveDate" -> "staged_updates.effectiveDate",
"endDate" -> "null"))
.execute()
```

}

Reference:

<https://www.projectpro.io/recipes/what-is-slowly-changing-data-scd-type-2-operation-delta-table-databricks>

Question 64

CertyIQ

You are designing an Azure Data Lake Storage solution that will transform raw JSON files for use in an analytical workload.

You need to recommend a format for the transformed files. The solution must meet the following requirements:

- Contain information about the data types of each column in the files.
- Support querying a subset of columns in the files.
- Support read-heavy analytical workloads.
- Minimize the file size.

What should you recommend?

- A. JSON
- B. CSV
- C. Apache Avro
- D. Apache Parquet

Explanation:

Correct Answer: *D*

Parquet, an open-source file format for Hadoop, stores nested data structures in a flat columnar format.

Compared to a traditional approach where data is stored in a row-oriented approach, Parquet file format is more efficient in terms of storage and performance.

It is especially good for queries that read particular columns from a *wide* (with many columns) table since only needed columns are read, and IO is minimized.

Incorrect:

Not C:

The Avro format is the ideal candidate for storing data in a data lake landing zone because:

1. Data from the landing zone is usually read as a whole for further processing by downstream systems (the row-based format is more efficient in this case).
2. Downstream systems can easily retrieve table schemas from Avro files (there is no need to store the schemas separately in an external meta store).
3. Any source schema change is easily handled (schema evolution).

Reference:

<https://www.clairvoyant.ai/blog/big-data-file-formats>

Question 65

CertyIQ

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 Storage account that contains 100 GB of files. The files contain rows of text and numerical values. 75% of the rows contain description data that has an average length of 1.1 MB.

You plan to copy the data from the storage account to an enterprise data warehouse in Azure Synapse Analytics.

You need to prepare the files to ensure that the data copies quickly.

Solution: You modify the files to ensure that each row is less than 1 MB.
Does this meet the goal?

- A. Yes
- B. No

Explanation:

Correct Answer: A

Polybase loads rows that are smaller than 1 MB.

Note on Polybase Load: PolyBase is a technology that accesses external data stored in Azure Blob storage or Azure Data Lake Store via the T-SQL language.

Extract, Load, and Transform (ELT)

Extract, Load, and Transform (ELT) is a process by which data is extracted from a source system, loaded into a data warehouse, and then transformed.

The basic steps for implementing a PolyBase ELT for dedicated SQL pool are:

Extract the source data into text files.

Land the data into Azure Blob storage or Azure Data Lake Store.

Prepare the data for loading.

Load the data into dedicated SQL pool staging tables using PolyBase.

Transform the data.

Insert the data into production tables.

Reference:

[https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/service-capacity-limits](https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-service-capacity-limits) <https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/load-data-overview>

Question 66

CertyIQ

You plan to create a dimension table in Azure Synapse Analytics that will be less than 1 GB.

You need to create the table to meet the following requirements:

- Provide the fastest query time.
- Minimize data movement during queries.

Which type of table should you use?

- A. replicated
- B. hash distributed
- C. heap
- D. round-robin

Explanation:

Correct Answer: A

A replicated table has a full copy of the table accessible on each Compute node. Replicating a table removes the need to transfer data among Compute nodes before a join or aggregation. Since the table has multiple copies, replicated tables work best when the table size is less than 2 GB compressed. 2 GB is not a hard limit.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/design-guidance-for-replicated-tables>

You are designing a dimension table in an Azure Synapse Analytics dedicated SQL pool. You need to create a surrogate key for the table. The solution must provide the fastest query performance. What should you use for the surrogate key?

- A. a GUID column
- B. a sequence object
- C. an IDENTITY column

Explanation:

Correct Answer: C

Use IDENTITY to create surrogate keys using dedicated SQL pool in AzureSynapse Analytics.

Note: A surrogate key on a table is a column with a unique identifier for each row. The key is not generated from the table data. Data modelers like to create surrogate keys on their tables when they design data warehouse models. You can use the IDENTITY property to achieve this goal simply and effectively without affecting load performance.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-identi>

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

Function:

Aggregate
Geospatial
Windowing

Answer Area

Input type:

Stream
Reference

Correct Answer:

Function:

Aggregate
Geospatial
Windowing

Explanation:

Correct Answer:

Input type: Stream -

You can process real-time IoT data streams with Azure Stream Analytics.

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.

Note: 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.

Reference:

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-get-started-with-azure-stream-analytics-to-process-data-from-iot-device>

<https://docs.microsoft.com/en-us/azure/stream-analytics/geospatial-scenarios>

Question 69

CertyIQ

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. Implement Azure Stream Analytics user-defined functions (UDF).
- C. Implement query parallelization by partitioning the data output.
- D. Scale the SU count for the job up.

- E. Scale the SU count for the job down.
- F. Implement query parallelization by partitioning the data input.

Explanation:

Correct Answer:

Partition input and output. REF: <https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-parallelization>

Question 70

CertyIQ

You need to trigger an Azure Data Factory pipeline when a file arrives in an Azure Data Lake Storage Gen2 container.

Which resource provider should you enable?

- A. Microsoft.Sql
- B. Microsoft.Automation
- C. Microsoft.EventGrid
- D. Microsoft.EventHub

Explanation:

Correct Answer: C

Event-driven architecture (EDA) is a common data integration pattern that involves production, detection, consumption, and reaction to events. Data integration scenarios often require Data Factory customers to trigger pipelines based on events happening in storage account, such as the arrival or deletion of a file in Azure Blob Storage account. Data Factory natively integrates with Azure Event Grid, which lets you trigger pipelines on such events.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/how-to-create-event-trigger>

<https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipeline-execution-triggers>

Azure Event Grids – Event-driven publish-subscribe model (think reactive programming) Azure Event Hubs – Multiple source big data streaming pipeline (think telemetry data) In this case its more suitable vs Event Hubs.

Question 71

CertyIQ

You plan to perform batch processing in Azure Databricks once daily.

Which type of Databricks cluster should you use?

- A. High Concurrency
- B. automated
- C. interactive

Explanation:

Correct Answer:B

Automated Databricks clusters are the best for jobs and automated batch processing.

Note: Azure Databricks has two types of clusters: interactive and automated. You use interactive clusters to analyze data collaboratively with interactive notebooks. You use automated clusters to run fast and robust automated jobs.

Example: Scheduled batch workloads (data engineers running ETL jobs)

This scenario involves running batch job JARs and notebooks on a regular cadence through the Databricks platform.

The suggested best practice is to launch a new cluster for each run of critical jobs. This helps avoid any issues (failures, missing SLA, and so on) due to an existing workload (noisy neighbor) on a shared cluster.

Reference:

<https://docs.microsoft.com/en-us/azure/databricks/clusters/create>

<https://docs.databricks.com/administration-guide/cloud-configurations/aws/cmbp.html#scenario-3-scheduled-batch-workloads-data-engineers-running-etl-jobs>

CertyIQ

Question 72

HOTSPOT -

You are processing streaming data from vehicles that pass through a toll booth.

You need to use Azure Stream Analytics to return the license plate, vehicle make, and hour the last vehicle passed during each 10-minute window.

How should you complete the query? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

```
WITH LastInWindow AS
(
    SELECT
        COUNT(*) AS COUNT,
        MAX(Time) AS LastEventTime
    FROM
        Input TIMESTAMP BY Time
    GROUP BY
        HoppingWindow(minute, 10)
)
SELECT
    Input.License_plate,
    Input.Make,
    Input.Time
FROM
    Input TIMESTAMP BY Time
INNER JOIN LastInWindow
ON
    DATEADD(minute, Input.Time, LastInWindow.LastEventTime) BETWEEN 0 AND 10
    AND Input.Time = LastInWindow.LastEventTime
```

Answer Area

```
WITH LastInWindow AS
(
    SELECT
        [▼] (Time) AS LastEventTime
        COUNT
        MAX
        MIN
        TOPONE
    FROM
        Input TIMESTAMP BY Time
    GROUP BY
        [▼] (minute, 10)
        HoppingWindow
        SessionWindow
        SlidingWindow
        TumblingWindow
)
SELECT
    Input.License_plate,
    Input.Make,
    Input.Time
FROM
    Input TIMESTAMP BY Time
    INNER JOIN LastInWindow
    ON [▼] (minute, Input, LastInWindow) BETWEEN 0 AND 10
    DATEADD
    DATEDIFF
    DATENAME
    DATEPART
AND Input.Time = LastInWindow.LastEventTime
```

Explanation:

Correct Answer:

Box 1: MAX -

The first step on the query finds the maximum time stamp in 10-minute windows, that is the time stamp of the last event for that window. The second step joins the results of the first query with the original stream to find the event that match the last time stamps in each window.

Query:

```
WITH LastInWindow AS -
(
    SELECT -
        MAX(Time) AS LastEventTime -
    FROM -
```

Input TIMESTAMP BY Time -

GROUP BY -

TumblingWindow(minute, 10)

)

SELECT -

Input.License_plate,

Input.Make,

Input.Time -

FROM -

Input TIMESTAMP BY Time -

INNER JOIN LastInWindow -

ON DATEDIFF(minute, Input, LastInWindow) BETWEEN 0 AND 10

AND Input.Time = LastInWindow.LastEventTime

Box 2: TumblingWindow -

Tumbling windows are a series of fixed-sized, non-overlapping and contiguous time intervals.

Box 3: DATEDIFF -

DATEDIFF is a date-specific function that compares and returns the time difference between two DateTime fields, for more information, refer to date functions.

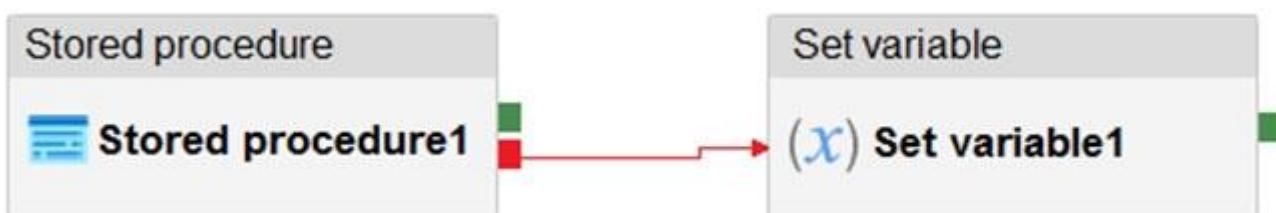
Reference:

<https://docs.microsoft.com/en-us/stream-analytics-query/tumbling-window-azure-stream-analytics>

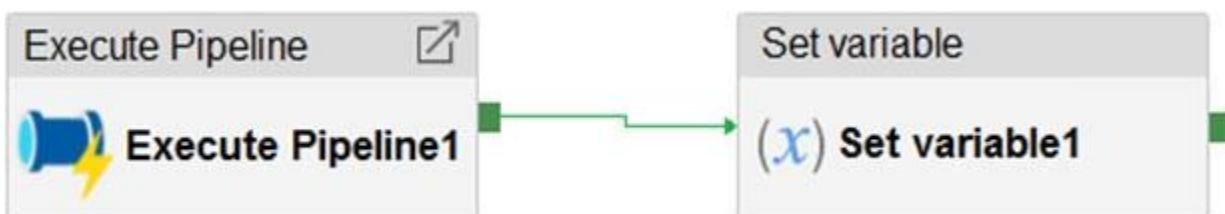
Question 73

CertyIQ

You have an Azure Data Factory instance that contains two pipelines named Pipeline1 and Pipeline2. Pipeline1 has the activities shown in the following exhibit.



Pipeline2 has the activities shown in the following exhibit.



You execute Pipeline2, and Stored procedure1 in Pipeline1 fails.

What is the status of the pipeline runs?

- A. Pipeline1 and Pipeline2 succeeded

- B. Pipeline1 and Pipeline2 failed.
- C. Pipeline1 succeeded and Pipeline2 failed.
- D. Pipeline1 failed and Pipeline2 succeeded.

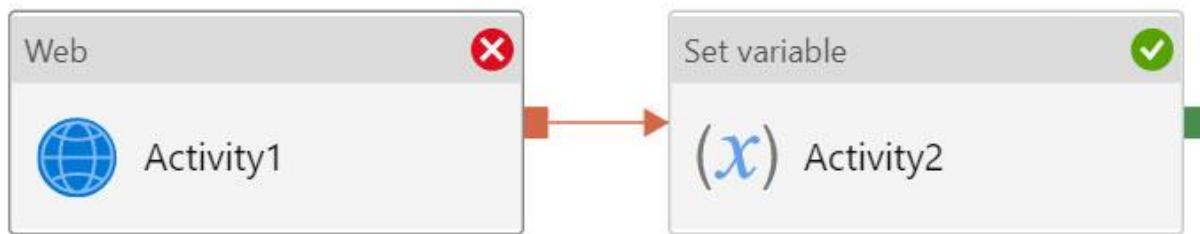
Explanation:

Correct Answer A

Activities are linked together via dependencies. A dependency has a condition of one of the following: Succeeded, Failed, Skipped, or Completed.

Consider Pipeline1:

If we have a pipeline with two activities where Activity2 has a failure dependency on Activity1, the pipeline will not fail just because Activity1 failed. If Activity1 fails and Activity2 succeeds, the pipeline will succeed. This scenario is treated as a try-catch block by Data Factory.



The failure dependency means this pipeline reports success.

Note:

If we have a pipeline containing Activity1 and Activity2, and Activity2 has a success dependency on Activity1, it will only execute if Activity1 is successful. In this scenario, if Activity1 fails, the pipeline will fail.

Reference:

<https://datasavvy.me/category/azure-data-factory/>

Answer is A. The trick is the fact that pipeline 1 only has a Failure dependency between the activity's. In this situation this results in a Succeeded pipeline if the stored procedure failed. If also the success connection was linked to a follow up activity, and the SP would fail, the pipeline would be indeed marked as failed.

Question 74

CertyIQ

HOTSPOT -

A company plans to use Platform-as-a-Service (PaaS) to create the new data pipeline process. The process must meet the following requirements:

Ingest:

- ☞ Access multiple data sources.
- ☞ Provide the ability to orchestrate workflow.
- ☞ Provide the capability to run SQL Server Integration Services packages.

Store:

- ☞ Optimize storage for big data workloads.
- ☞ Provide encryption of data at rest.
- ☞ Operate with no size limits.

Prepare and Train:

- ☞ Provide a fully-managed and interactive workspace for exploration and visualization.
- ☞ Provide the ability to program in R, SQL, Python, Scala, and Java.

Provide seamless user authentication with Azure Active Directory.

Model & Serve:

⇒ Implement native columnar storage.

⇒ Support for the SQL language

⇒ Provide support for structured streaming.

You need to build the data integration pipeline.

Which technologies should you use? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Architecture requirement	Technology
Ingest	<div style="border: 1px solid black; padding: 5px;"><p>Logic Apps</p><p>Azure Data Factory</p><p>Azure Automation</p></div>
Store	<div style="border: 1px solid black; padding: 5px;"><p>Azure Data Lake Storage</p><p>Azure Blob storage</p><p>Azure files</p></div>
Prepare and Train	<div style="border: 1px solid black; padding: 5px;"><p>HDInsight Apache Spark cluster</p><p>Azure Databricks</p><p>HDInsight Apache Storm cluster</p></div>
Model and Serve	<div style="border: 1px solid black; padding: 5px;"><p>HDInsight Apache Kafka cluster</p><p>Azure Synapse Analytics</p><p>Azure Data Lake Storage</p></div>

Answer Area

Architecture requirement

Technology

Ingest

Logic Apps
Azure Data Factory
Azure Automation

Store

Correct Answer:

Azure Data Lake Storage
Azure Blob storage
Azure files

Prepare and Train

HDInsight Apache Spark cluster
Azure Databricks
HDInsight Apache Storm cluster

Model and Serve

HDInsight Apache Kafka cluster
Azure Synapse Analytics
Azure Data Lake Storage

Explanation:

Correct Answer

Ingest: Azure Data Factory -

Azure Data Factory pipelines can execute SSIS packages.

In Azure, the following services and tools will meet the core requirements for pipeline orchestration, control flow, and data movement: Azure Data Factory, Oozie on HDInsight, and SQL Server Integration Services (SSIS).

Store: Data Lake Storage -

Data Lake Storage Gen1 provides unlimited storage.

Note: Data at rest includes information that resides in persistent storage on physical media, in any digital format.

Microsoft Azure offers a variety of data storage solutions to meet different needs, including file, disk, blob, and table storage. Microsoft also provides encryption to protect Azure SQL Database, Azure Cosmos DB, and Azure Data Lake.

Prepare and Train: Azure Databricks

Azure Databricks provides enterprise-grade Azure security, including Azure Active Directory integration.

With Azure Databricks, you can set up your Apache Spark environment in minutes, autoscale and collaborate on shared projects in an interactive workspace.

Azure Databricks supports Python, Scala, R, Java and SQL, as well as data science frameworks and libraries including TensorFlow, PyTorch and scikit-learn.

Model and Serve: Azure Synapse Analytics

Azure Synapse Analytics/ SQL Data Warehouse stores data into relational tables with columnar storage.

Azure SQL Data Warehouse connector now offers efficient and scalable structured streaming write support for SQL Data Warehouse. Access SQL Data

Warehouse from Azure Databricks using the SQL Data Warehouse connector.

Note: As of November 2019, Azure SQL Data Warehouse is now Azure Synapse Analytics.

Reference:

<https://docs.microsoft.com/bs-latn-ba/azure/architecture/data-guide/technology-choices/pipeline-orchestration-data-movement>

<https://docs.microsoft.com/en-us/azure/azure-databricks/what-is-azure-databricks>

Question 75

CertyIQ

DRAG DROP -

You have the following table named Employees.

first_name	last_name	hire_date	employee_type
Jane	Doe	2019-08-23	new
Ben	Smith	2017-12-15	Standard

You need to calculate the employee_type value based on the hire_date value.

How should you complete the Transact-SQL statement? To answer, drag the appropriate values to the correct targets. Each value 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:

Values

Answer Area

```
SELECT
    *,
    CASE
        WHEN hire_date >= '2019-01-01' THEN 'New'
        ELSE 'Standard'
    END AS employee_type
FROM
    employees
```

CASE

ELSE

OVER

PARTITION BY

ROW_NUMBER

FROM

employees

	Values	Answer Area
Correct Answer:	<div style="display: flex; justify-content: space-around;"> <div>CASE</div> <div>ELSE</div> <div>OVER</div> <div>PARTITION BY</div> <div>ROW_NUMBER</div> </div>	<pre> SELECT *, CASE WHEN hire_date >= '2019-01-01' THEN 'New' ELSE 'Standard' END AS employee_type FROM employees </pre>

Explanation:

Correct Answer

Box 1: CASE -

CASE evaluates a list of conditions and returns one of multiple possible result expressions.

CASE can be used in any statement or clause that allows a valid expression. For example, you can use CASE in statements such as SELECT, UPDATE,

DELETE and SET, and in clauses such as select_list, IN, WHERE, ORDER BY, and HAVING.

Syntax: Simple CASE expression:

CASE input_expression -

WHEN when_expression THEN result_expression [...n]

[ELSE else_result_expression]

END -

Box 2: ELSE -

Reference:

<https://docs.microsoft.com/en-us/sql/t-sql/language-elements/case-transact-sql>

Question 76

CertyIQ

DRAG DROP -

You have an Azure Synapse Analytics workspace named WS1.

You have an Azure Data Lake Storage Gen2 container that contains JSON-formatted files in the following format.

```
{
  "id": "66532691-ab20-11ea-8b1d-936b3ec64e54",
  "context": {
    "data": {
      "eventTime": "2020-06-10T13:43:34.553Z",
      "samplingRate": "100.0",
      "isSynthetic": "false"
    },
    "session": {
      "isFirst": "false",
      "id": "38619c14-7a23-4687-8268-95862c5326b1"
    },
    "custom": {
      "dimensions": [
        {
          "customerInfo": {
            "ProfileType": "ExpertUser",
            "RoomName": "",
            "CustomerName": "diamond",
            "UserName": "XXXX@yahoo.com"
          }
        },
        {
          "customerInfo": {
            "ProfileType": "Novice",
            "RoomName": "",
            "CustomerName": "topaz",
            "UserName": "XXXX@outlook.com"
          }
        }
      ]
    }
  }
}
```

You need to use the serverless SQL pool in WS1 to read the files.

How should you complete the Transact-SQL statement? To answer, drag the appropriate values to the correct targets. Each value 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:

Values

Answer Area

```
select*  
  
FROM  
[ ] (   
  
    BULK 'https://contoso.blob.core.windows.net/contosodw',  
    FORMAT= 'CSV',  
    fieldterminator = '0x0b',  
    fieldquote = '0x0b',  
    rowterminator = '0x0b'  
)  
with (id varchar(50),  
      contextdateeventTime varchar(50) '$.context.data.eventTime',  
      contextdatasamplingRate varchar(50) '$.context.data.samplingRate',  
      contextdataisSynthetic varchar(50) '$.context.data.isSynthetic',  
      contextsessionisFirst varchar(50) '$.context.session.isFirst',  
      contextsession varchar(50) '$.context.session.id',  
      contextcustomdimensions varchar(max) '$.context.custom.dimensions'  
) as q  
cross apply [ ] (contextcustomdimensions)  
  
with ( ProfileType varchar(50) '$.customerInfo.ProfileType',  
       RoomName varchar(50) '$.customerInfo.RoomName',  
       CustomerName varchar(50) '$.customerInfo.CustomerName',  
       UserName varchar(50) '$.customerInfo.UserName'  
)
```

Values

Answer Area

```
select*  
  
FROM  
[ ] (   
  
    BULK 'https://contoso.blob.core.windows.net/contosodw',  
    FORMAT= 'CSV',  
    fieldterminator = '0x0b',  
    fieldquote = '0x0b',  
    rowterminator = '0x0b'  
)  
with (id varchar(50),  
      contextdateeventTime varchar(50) '$.context.data.eventTime',  
      contextdatasamplingRate varchar(50) '$.context.data.samplingRate',  
      contextdataisSynthetic varchar(50) '$.context.data.isSynthetic',  
      contextsessionisFirst varchar(50) '$.context.session.isFirst',  
      contextsession varchar(50) '$.context.session.id',  
      contextcustomdimensions varchar(max) '$.context.custom.dimensions'  
) as q  
cross apply [ ] openjson (contextcustomdimensions)  
  
with ( ProfileType varchar(50) '$.customerInfo.ProfileType',  
       RoomName varchar(50) '$.customerInfo.RoomName',  
       CustomerName varchar(50) '$.customerInfo.CustomerName',  
       UserName varchar(50) '$.customerInfo.UserName'  
)
```

Explanation:

Correct Answer

Box 1: openrowset -

The easiest way to see to the content of your CSV file is to provide file URL to OPENROWSET function, specify csv FORMAT.

Example:

```
SELECT *
FROM OPENROWSET(
    BULK 'csv/population/population.csv',
    DATA_SOURCE = 'SqlOnDemandDemo',
    FORMAT = 'CSV', PARSER_VERSION = '2.0',
    FIELDTERMINATOR = ',',
    ROWTERMINATOR = '\n'
```

Box 2: openjson -

You can access your JSON files from the Azure File Storage share by using the mapped drive, as shown in the following example:

```
SELECT book.* FROM -
OPENROWSET(BULK N't:\books\books.json', SINGLE_CLOB) AS json
CROSS APPLY OPENJSON(BulkColumn)
WITH( id nvarchar(100), name nvarchar(100), price float,
pages_i int, author nvarchar(100)) AS book
```

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/query-single-csv-file>

e <https://docs.microsoft.com/en-us/sql/relational-databases/json/import-json-documents-into-sql-server>

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/query-json-files>

Question 77

CertyIQ

DRAG DROP -

You have an Apache Spark DataFrame named temperatures. A sample of the data is shown in the following table.

Date	Temp
...	...
18-01-2021	3
19-01-2021	4
20-01-2021	2
21-01-2021	2
...	...

You need to produce the following table by using a Spark SQL query.

Year	JAN	FEB	MAR	APR	MAY
2019	2.3	4.1	5.2	7.6	9.2
2020	2.4	4.2	4.9	7.8	9.1
2021	2.6	5.3	3.4	7.9	9.5

How should you complete the query? To answer, drag the appropriate values to the correct targets. Each value 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:

Values Answer Area

```
SELECT * FROM (
    SELECT YEAR(Date) Year, MONTH(Date) Month, Temp
    FROM temperatures
    WHERE date BETWEEN DATE '2019-01-01' AND DATE '2021-08-31'
)
(
    AVG ( [ ] (Temp AS DECIMAL(4, 1)))
FOR Month in (
    1 JAN, 2 FEB, 3 MAR, 4 APR, 5 MAY, 6 JUN,
    7 JUL, 8 AUG, 9 SEP, 10 OCT, 11 NOV, 12 DEC
)
)
ORDER BY Year ASC
```

Values Answer Area

```
SELECT * FROM (
    SELECT YEAR(Date) Year, MONTH(Date) Month, Temp
    FROM temperatures
    WHERE date BETWEEN DATE '2019-01-01' AND DATE '2021-08-31'
)
(
    PIVOT (
        AVG ( CAST [ ] (Temp AS DECIMAL(4, 1)))
FOR Month in (
    1 JAN, 2 FEB, 3 MAR, 4 APR, 5 MAY, 6 JUN,
    7 JUL, 8 AUG, 9 SEP, 10 OCT, 11 NOV, 12 DEC
)
)
ORDER BY Year ASC
```

Correct Answer:

Explanation:

Correct Answer

Box 1: PIVOT -

PIVOT rotates a table-valued expression by turning the unique values from one column in the expression into multiple columns in the output. And PIVOT runs aggregations where they're required on any remaining column values that are wanted in the final output.

Incorrect Answers:

UNPIVOT carries out the opposite operation to PIVOT by rotating columns of a table-valued expression into column values.

Box 2: CAST -

If you want to convert an integer value to a DECIMAL data type in SQL Server use the `CAST()` function.

Example:

`SELECT -`

`CAST(12 AS DECIMAL(7,2)) AS decimal_value;`

Here is the result:

`decimal_value`

`12.00`

Reference:

<https://learnsql.com/cookbook/how-to-convert-an-integer-to-a-decimal-in-sql-server>

/ <https://docs.microsoft.com/en-us/sql/t-sql/queries/from-using-pivot-and-unpivot>

Question 78

CertyIQ

You have an Azure Data Factory that contains 10 pipelines.

You need to label each pipeline with its main purpose of either ingest, transform, or load. The labels must be available for grouping and filtering when using the monitoring experience in Data Factory.

What should you add to each pipeline?

- A. a resource tag
- B. a correlation ID
- C. a run group ID
- D. an annotation

Explanation:

Correct Answer : D

Annotations are additional, informative tags that you can add to specific factory resources: pipelines, datasets, linked services, and triggers. By adding annotations, you can easily filter and search for specific factory resources.

Reference:

<https://www.cathrinewilhelmsen.net/annotations-user-properties-azure-data-factory/>

Question 79

CertyIQ

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.	<input type="radio"/>	<input type="radio"/>
The Databricks cluster minimizes costs when running scheduled jobs that execute notebooks.	<input type="radio"/>	<input type="radio"/>
The Databricks cluster supports the creation of a Delta Lake table.	<input type="radio"/>	<input type="radio"/>

Answer Area

	Statements	Yes	No
Correct Answer:	The Databricks cluster supports multiple concurrent users.	<input checked="" type="radio"/>	<input type="radio"/>
	The Databricks cluster minimizes costs when running scheduled jobs that execute notebooks.	<input type="radio"/>	<input checked="" type="radio"/>
	The Databricks cluster supports the creation of a Delta Lake table.	<input checked="" type="radio"/>	<input type="radio"/>

Explanation:

Correct Answer

Box 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.

Box 2: No -

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.

Box 3: Yes -

Delta Lake on Databricks allows you to configure Delta Lake based on your workload patterns.

Reference:

<https://adatis.co.uk/databricks-cluster-sizing/>

<https://docs.microsoft.com/en-us/azure/databricks/jobs>

<https://docs.databricks.com/administration-guide/capacity-planning/cmbp.html>

<https://docs.databricks.com/delta/index.html>

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>

Question 80

CertyIQ

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 Databricks
- C. Azure Stream Analytics
- D. Azure SQL Database

Explanation:

Correct Answer B

Stream Analytics supports "extending SQL language with JavaScript and C# user-defined functions (UDFs)". There is no mention of Python support; hence Stream Analytics is not correct.

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-introduction>

Azure Databricks supports near real-time data from Azure Event Hubs. And includes support for R, SQL, Python, Scala, and Java. So I will go for option B.

End of Part 2



Looking for all PDFs of DP-203?

Please drop us mail at: - certyiq@gmail.com

Please find the videos of this AZ-900/AI-900/AZ-305/
AZ-104 /DP-900/ SC-900 and other Microsoft exam
series on

CertyIQ Official YouTube channel (FREE PDFs): -

Please [Subscribe](#) to CertyIQ YouTube Channel to get notified for latest exam dumps by clicking on the below image, it will redirect to the [CertyIQ](#) YouTube page.



Connect with us @ [LinkedIn](#) [Telegram](#)

Contact us for other dumps: - certyiqpdf@gmail.com

For any other enquiry, please drop us a mail at
certyiqofficial@gmail.com

DP-203

Real Exam Questions & Answers



Latest Exam Questions
All Topics Covered-2023
Added New Que's



Get Certified Today!

New Course Covered

Subscribe

Question 1

CertyIQ

You have a table in an Azure Synapse Analytics dedicated SQL pool. The table was created by using the following Transact-SQL statement

```
CREATE TABLE [dbo].[DimEmployee] (
    [EmployeeKey] [int] IDENTITY(1,1) NOT NULL,
    [EmployeeID] [int] NOT NULL,
    [FirstName] [varchar](100) NOT NULL,
    [LastName] [varchar](100) NOT NULL,
    [JobTitle] [varchar](100) NULL,
    [LastHireDate] [date] NULL,
    [StreetAddress] [varchar](500) NOT NULL,
    [City] [varchar](200) NOT NULL,
    [StateProvince] [varchar](50) NOT NULL,
    [Portalcode] [varchar](10) NOT NULL
)
```

You need to alter the table to meet the following requirements:

- ⇒ Ensure that users can identify the current manager of employees.

- Support creating an employee reporting hierarchy for your entire company.
 - Provide fast lookup of the managers' attributes such as name and job title.
- Which column should you add to the table?

- A. [ManagerEmployeeID] [smallint] NULL
- B. [ManagerEmployeeKey] [smallint] NULL
- C. [ManagerEmployeeKey] [int] NULL**
- D. [ManagerName] [varchar](200) NULL

Explanation:

Correct Answer: C

We need an extra column to identify the Manager. Use the data type as the EmployeeKey column, an int column.

Reference:

<https://docs.microsoft.com/en-us/analysis-services/tabular-models/hierarchies-ssas-tabular>

Question 2

CertyIQ

You have an Azure Synapse workspace named MyWorkspace that contains an Apache Spark database named mytestdb.

You run the following command in an Azure Synapse Analytics Spark pool in MyWorkspace.

```
CREATE TABLE mytestdb.myParquetTable(  
EmployeeID int,  
EmployeeName string,  
EmployeeStartDate date)
```

USING Parquet -

You then use Spark to insert a row into mytestdb.myParquetTable. The row contains the following data.

EmployeeName	EmployeeID	EmployeeStartDate
Alice	24	2020-01-25

One minute later, you execute the following query from a serverless SQL pool in MyWorkspace.

```
SELECT EmployeeID -  
FROM mytestdb.dbo.myParquetTable  
WHERE EmployeeName = 'Alice';
```

What will be returned by the query?

- A. 24
- B. an error**
- C. a null value

Explanation:

Correct Answer: B. It will return an error. table name will be converted to lowercase <https://learn.microsoft.com/en-us/azure/synapse-analytics/metadata/table>

Question 3

CertyIQ

DRAG DROP -

You have a table named SalesFact in an enterprise data warehouse in Azure Synapse Analytics. SalesFact contains sales data from the past 36 months and has the following characteristics:

- ⇒ Is partitioned by month
- ⇒ Contains one billion rows
- ⇒ Has clustered columnstore index

At the beginning of each month, you need to remove data from SalesFact that is older than 36 months as quickly as possible.

Which three actions should you perform in sequence in a stored procedure? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

Select and Place:

Actions

Answer Area

Switch the partition containing the stale data from SalesFact to SalesFact_Work.
Truncate the partition containing the stale data.
Drop the SalesFact_Work table.
Create an empty table named SalesFact_Work that has the same schema as SalesFact.
Execute a DELETE statement where the value in the Date column is more than 36 months ago.
Copy the data to a new table by using CREATE TABLE AS SELECT (CTAS).

Correct

Answer:

Actions

Answer Area

Switch the partition containing the stale data from SalesFact to SalesFact_Work.
Truncate the partition containing the stale data.
Drop the SalesFact_Work table.
Create an empty table named SalesFact_Work that has the same schema as SalesFact.
Execute a DELETE statement where the value in the Date column is more than 36 months ago.
Copy the data to a new table by using CREATE TABLE AS SELECT (CTAS).

Create an empty table named SalesFact_Work that has the same schema as SalesFact.
Switch the partition containing the stale data from SalesFact to SalesFact_Work.
Drop the SalesFact_Work table.

Explanation:

Correct Answer:

Step 1: Create an empty table named SalesFact_work that has the same schema as SalesFact.

Step 2: Switch the partition containing the stale data from SalesFact to SalesFact_Work.

SQL Data Warehouse supports partition splitting, merging, and switching. To switch partitions between two tables, you must ensure that the partitions align on their respective boundaries and that the table definitions match.

Loading data into partitions with partition switching is a convenient way stage new data in a table that is not visible to users the switch in the new data.

Step 3: Drop the SalesFact_Work table.

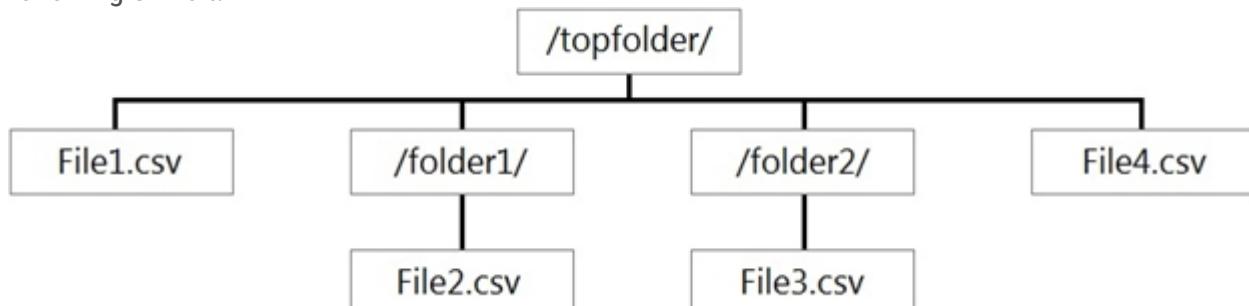
Reference:

<https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-tables-partitions>

Question 4

CertyIQ

You have files and folders in Azure Data Lake Storage Gen2 for an Azure Synapse workspace as shown in the following exhibit.



You create an external table named ExtTable that has LOCATION='/topfolder/'.

When you query ExtTable by using an Azure Synapse Analytics serverless SQL pool, which files are returned?

- A. File2.csv and File3.csv only
- B. File1.csv and File4.csv only
- C. File1.csv, File2.csv, File3.csv, and File4.csv
- D. File1.csv only

Explanation:

Correct Answer: B

In case of a serverless pool a wildcard should be added to the location.

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables?tabs=hadoop#arguments-create-external-table>

Question 5

CertyIQ

HOTSPOT -

You are planning the deployment of Azure Data Lake Storage Gen2.

You have the following two reports that will access the data lake:

- ☞ Report1: Reads three columns from a file that contains 50 columns.
- ☞ Report2: Queries a single record based on a timestamp.

You need to recommend in which format to store the data in the data lake to support the reports. The solution must minimize read times.

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Report1: 

Avro
CSV
Parquet
TSV

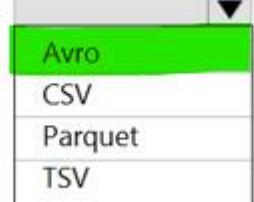
Report2: 

Avro
CSV
Parquet
TSV

Answer Area

Report1: 

Avro
CSV
Parquet
TSV

Report2: 

Avro
CSV
Parquet
TSV

Explanation:

Correct Answer:

1. Parquet
2. Avro

<https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-best-practices>

"Consider using the Avro file format in cases where your I/O patterns are more write heavy, or the query patterns favor retrieving multiple rows of records in their entirety.

Consider Parquet and ORC file formats when the I/O patterns are more read heavy or when the query patterns are focused on a subset of columns in the records."

Question 6

CertyIQ

You are designing the folder structure for an Azure Data Lake Storage Gen2 container.

Users will query data by using a variety of services including Azure Databricks and Azure Synapse Analytics

serverless SQL pools. The data will be secured by subject area. Most queries will include data from the current year or current month.

Which folder structure should you recommend to support fast queries and simplified folder security?

- A. /{SubjectArea}/{DataSource}/{DD}/{MM}/{YYYY}/{FileData}_{YYYY}_{MM}_{DD}.csv
- B. /{DD}/{MM}/{YYYY}/{SubjectArea}/{DataSource}/{FileData}_{YYYY}_{MM}_{DD}.csv
- C. /{YYYY}/{MM}/{DD}/{SubjectArea}/{DataSource}/{FileData}_{YYYY}_{MM}_{DD}.csv
- D. /{SubjectArea}/{DataSource}/{YYYY}/{MM}/{DD}/{FileData}_{YYYY}_{MM}_{DD}.csv

Explanation:

Correct Answer: : D

There's an important reason to put the date at the end of the directory structure. If you want to lock down certain regions or subject matters to users/groups, then you can easily do so with the POSIX permissions. Otherwise, if there was a need to restrict a certain security group to viewing just the UK data or certain planes, with the date structure in front a separate permission would be required for numerous directories under every hour directory. Additionally, having the date structure in front would exponentially increase the number of directories as time went on.

Note: In IoT workloads, there can be a great deal of data being landed in the data store that spans across numerous products, devices, organizations, and customers. It's important to pre-plan the directory layout for organization, security, and efficient processing of the data for down-stream consumers. A general template to consider might be the following layout:

Question 7

CertyIQ

HOTSPOT -

You need to output files from Azure Data Factory.

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Columnar format:

Avro
GZip
Parquet
TXT

JSON with a timestamp:

Avro
GZip
Parquet
TXT

Answer Area

Columnar format:

Avro
GZip
Parquet
TXT

Correct Answer:

JSON with a timestamp:

Avro
GZip
Parquet
TXT

Explanation:

Correct Answer:

Box 1: Parquet -

Parquet stores data in columns, while Avro stores data in a row-based format. By their very nature, column-oriented data stores are optimized for read-heavy analytical workloads, while row-based databases are best for write-heavy transactional workloads.

Box 2: Avro -

An Avro schema is created using JSON format.

AVRO supports timestamps.

Note: Azure Data Factory supports the following file formats (not GZip or TXT).

Avro format -

- ☞ Binary format
- ☞ Delimited text format
- ☞ Excel format
- ☞ JSON format
- ☞ ORC format
- ☞ Parquet format
- ☞ XML format

Reference:

<https://www.datanami.com/2018/05/16/big-data-file-formats-demystified>

Question 8

CertyIQ

HOTSPOT -

You use Azure Data Factory to prepare data to be queried by Azure Synapse Analytics serverless SQL pools. Files are initially ingested into an Azure Data Lake Storage Gen2 account as 10 small JSON files. Each file contains the same data attributes and data from a subsidiary of your company.

You need to move the files to a different folder and transform the data to meet the following requirements:

- ☞ Provide the fastest possible query times.
- ☞ Automatically infer the schema from the underlying files.

How should you configure the Data Factory copy activity? To answer, select the appropriate options in the

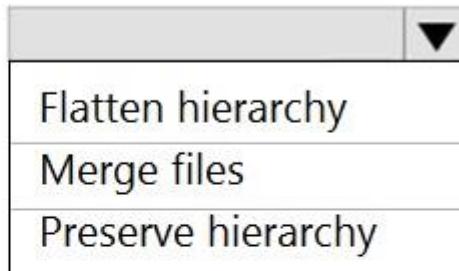
answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Copy behavior:

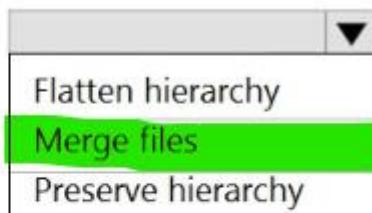


Sink file type:



Answer Area

Copy behavior:



Sink file type:



Explanation:

Correct Answer:

1. Merge Files
- 2 Parquet -

Azure Data Factory parquet format is supported for Azure Data Lake Storage Gen2.

Parquet supports the schema property.

Reference:

<https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-introduction> <https://docs.microsoft.com/en-us/azure/data-factory/format-parquet> <https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-performance-tuning-guidance>

HOTSPOT -

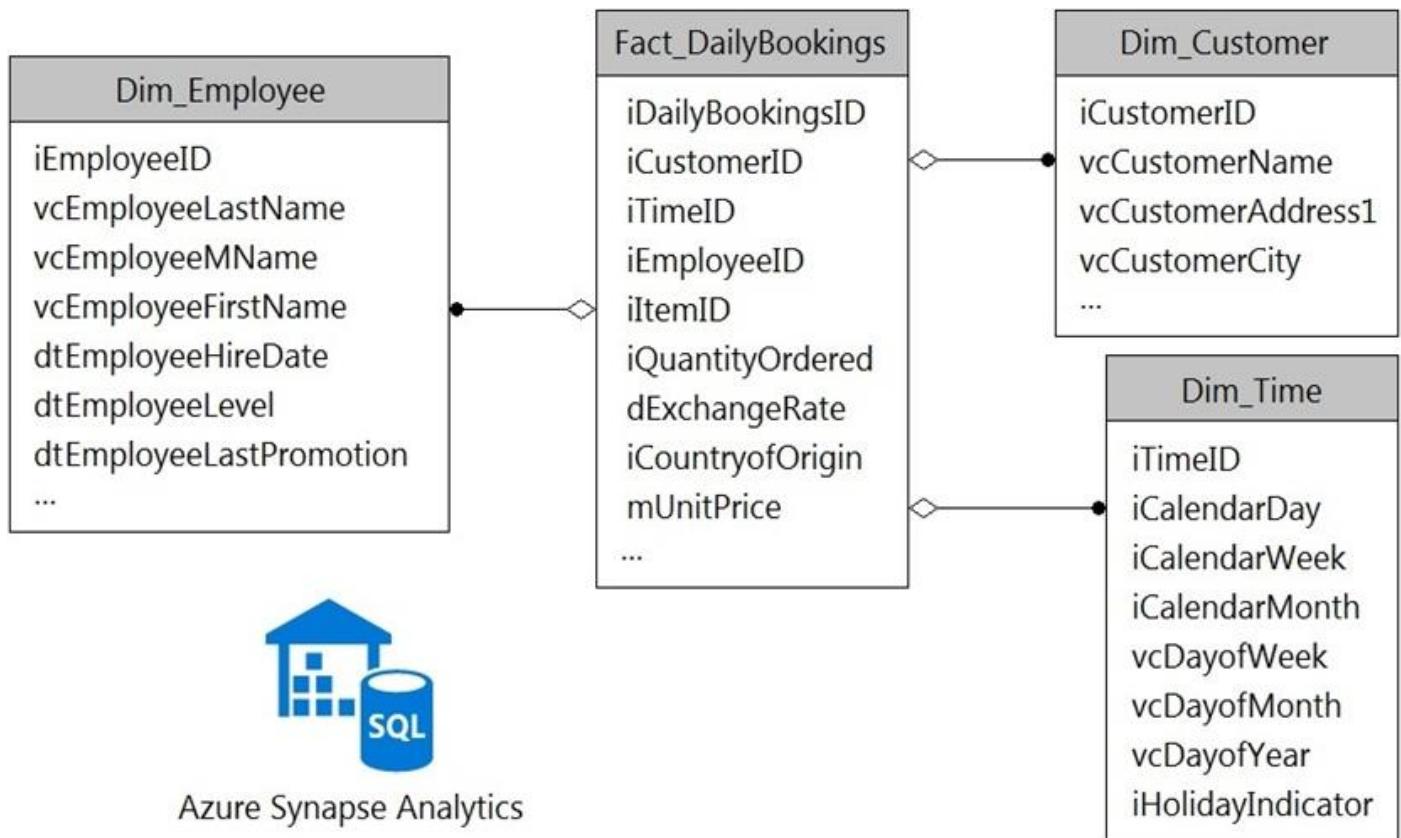
You have a data model that you plan to implement in a data warehouse in Azure Synapse Analytics as shown in the following exhibit.

All the dimension tables will be less than 2 GB after compression, and the fact table will be approximately 6 TB. The dimension tables will be relatively static with very few data inserts and updates.

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:



All the dimension tables will be less than 2 GB after compression, and the fact table will be approximately 6 TB. The dimension tables will be relatively static with very few data inserts and updates.

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

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

Explanation:

Correct Answer:

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/>

HOTSPOT -

You have an Azure Data Lake Storage Gen2 container.

Data is ingested into the container, and then transformed by a data integration application. The data is NOT modified after that. Users can read files in the container but cannot modify the files.

You need to design a data archiving solution that meets the following requirements:

- ☞ New data is accessed frequently and must be available as quickly as possible.
- ☞ Data that is older than five years is accessed infrequently but must be available within one second when requested.
- ☞ Data that is older than seven years is NOT accessed. After seven years, the data must be persisted at the lowest cost possible.
- ☞ Costs must be minimized while maintaining the required availability.

How should you manage the data? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point

Hot Area:

Answer Area

Five-year-old data:

Delete the blob.
Move to archive storage.
Move to cool storage.
Move to hot storage.

Seven-year-old data:

Delete the blob.
Move to archive storage.
Move to cool storage.
Move to hot storage.

Answer Area

Five-year-old data:

Delete the blob.
Move to archive storage.
Move to cool storage.
Move to hot storage.

Correct Answer:

Seven-year-old data:

Delete the blob.
Move to archive storage.
Move to cool storage.
Move to hot storage.

Explanation:

Correct Answer:

Box 1: Move to cool storage -

Box 2: Move to archive storage -

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.

The following table shows a comparison of premium performance block blob storage, and the hot, cool, and archive access tiers.

Question11

CertyIQ

DRAG DROP -

You need to create a partitioned table in an Azure Synapse Analytics dedicated SQL pool.

How should you complete the Transact-SQL statement? To answer, drag the appropriate values to the correct targets. Each value 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:

Values

CLUSTERED INDEX
COLLATE
DISTRIBUTION
PARTITION
PARTITION FUNCTION
PARTITION SCHEME

Answer Area

```
CREATE TABLE table1
(
    ID INTEGER,
    col1 VARCHAR(10),
    col2 VARCHAR(10)
) WITH
(
    = HASH(ID),
    (ID RANGE LEFT FOR VALUES (1, 1000000, 2000000))
);
```

Values
CLUSTERED INDEX
COLLATE
DISTRIBUTION
PARTITION
PARTITION FUNCTION
PARTITION SCHEME

Answer Area

```
CREATE TABLE table1
(
    ID INTEGER,
    col1 VARCHAR(10),
    col2 VARCHAR(10)
) WITH
(
    DISTRIBUTION = HASH(ID),
    PARTITION (ID RANGE LEFT FOR VALUES (1, 1000000, 2000000))
);
```

Explanation:

Correct Answer

Box 1: DISTRIBUTION -

Table distribution options include DISTRIBUTION = HASH (distribution_column_name), assigns each row to one distribution by hashing the value stored in distribution_column_name.

Box 2: PARTITION -

Table partition options. Syntax:

PARTITION (partition_column_name RANGE [LEFT | RIGHT] FOR VALUES ([boundary_value [...]]))

Reference:

<https://docs.microsoft.com/en-us/sql/t-sql/statements/create-table-azure-sql-data-warehouse>

Question 12

CertyIQ

You need to design an Azure Synapse Analytics dedicated SQL pool that meets the following requirements:

- ⇒ Can return an employee record from a given point in time.
- ⇒ Maintains the latest employee information.
- ⇒ Minimizes query complexity.

How should you model the employee data?

- A. as a temporal table
- B. as a SQL graph table
- C. as a degenerate dimension table
- D. as a Type 2 slowly changing dimension (SCD) table

Explanation:

Correct Answer: : D

A Type 2 SCD supports versioning of dimension members. Often the source system doesn't store versions, so the data warehouse load process detects and manages changes in a dimension table. In this case, the dimension table must use a surrogate key to provide a unique reference to a version of the dimension member. It also includes columns that define the date range validity of the version (for example, StartDate and EndDate) and possibly a flag column (for example, IsCurrent) to easily filter by current dimension members.

Reference:

<https://docs.microsoft.com/en-us/learn/modules/populate-slowly-changing-dimensions-azure-synapse-analytics-pipelines/3-choose-between-dimension-types>

Question 13

CertyIQ

You have an enterprise-wide Azure Data Lake Storage Gen2 account. The data lake is accessible only through an Azure virtual network named VNET1.

You are building a SQL pool in Azure Synapse that will use data from the data lake.

Your company has a sales team. All the members of the sales team are in an Azure Active Directory group named Sales. POSIX controls are used to assign the Sales group access to the files in the data lake.

You plan to load data to the SQL pool every hour.

You need to ensure that the SQL pool can load the sales data from the data lake.

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

NOTE: Each area selection is worth one point.

- A. Add the managed identity to the Sales group.
- B. Use the managed identity as the credentials for the data load process.
- C. Create a shared access signature (SAS).
- D. Add your Azure Active Directory (Azure AD) account to the Sales group.
- E. Use the shared access signature (SAS) as the credentials for the data load process.
- F. Create a managed identity

Explanation:

Correct Answer: ABF

The managed identity grants permissions to the dedicated SQL pools in the workspace.

Azure AD -

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/security/synapse-workspace-managed-identity>

Question 14

CertyIQ

HOTSPOT -

You have an Azure Synapse Analytics dedicated SQL pool that contains the users shown in the following table.

Name	Role
User1	Server admin
User2	db_datereader

User1 executes a query on the database, and the query returns the results shown in the following exhibit.

```

1  SELECT c.name,
2      tbl.name AS table_name,
3      typ.name AS datatype,
4      c.is_masked,
5      c.masking_function
6  FROM sys.masked_columns AS c
7  INNER JOIN sys.tables AS tbl ON c.[object_id] = tbl.[object_id]
8  INNER JOIN sys.types typ ON c.user_type_id = typ.user_type_id
9  WHERE is_masked = 1;
10

```

Results Messages

	name	table_name	datatype	is_masked	masking_function
1	BirthDate	DimCustomer	date	1	default()
2	Gender	DimCustomer	nvarchar	1	default()
3	EmailAddress	DimCustomer	nvarchar	1	email()
4	YearlyIncome	DimCustomer	money	1	default()

User1 is the only user who has access to the unmasked data.

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

When User2 queries the YearlyIncome column,
the values returned will be [answer choice].

a random number
the values stored in the database
XXXX
0

When User1 queries the BirthDate column, the
values returned will be [answer choice].

a random date
the values stored in the database
XXXX
1900-01-01

Correct
Answer:

Answer Area

When User2 queries the YearlyIncome column,
the values returned will be [answer choice].

a random number
the values stored in the database
XXXX
0

When User1 queries the BirthDate column, the
values returned will be [answer choice].

a random date
the values stored in the database
XXXX
1900-01-01

Explanation:

Correct Answer:

Box 1: 0 -

The YearlyIncome column is of the money data type.

The Default masking function: Full masking according to the data types of the designated fields

☞ Use a zero value for numeric data types (bigint, bit, decimal, int, money, numeric, smallint, smallmoney, tinyint, float, real).

Box 2: the values stored in the database

Users with administrator privileges are always excluded from masking, and see the original data without any mask.

Reference:

<https://docs.microsoft.com/en-us/azure/azure-sql/database/dynamic-data-masking-overview>

Question 15

CertyIQ

You have an enterprise data warehouse in Azure Synapse Analytics. Using PolyBase, you create an external table named [Ext].[Items] to query Parquet files stored in Azure Data Lake Storage Gen2 without importing the data to the data warehouse. The external table has three columns.

You discover that the Parquet files have a fourth column named ItemID. Which command should you run to add the ItemID column to the external table?

A.

```
ALTER EXTERNAL TABLE [Ext].[Items]
    ADD [ItemID] int;
```

B.

```
DROP EXTERNAL FILE FORMAT parquetfile1;
CREATE EXTERNAL FILE FORMAT parquetfile1
WITH (
    FORMAT_TYPE = PARQUET,
    DATA_COMPRESSION = 'org.apache.hadoop.io.compress.SnappyCodec'
);
```

C.

```
DROP EXTERNAL TABLE [Ext].[Items]
CREATE EXTERNAL TABLE [Ext].[Items]
([ItemID] [int] NULL,
 [ItemName] nvarchar(50) NULL,
 [ItemType] nvarchar(20) NULL,
 [ItemDescription] nvarchar(250))
WITH
(
    LOCATION= '/Items/',
    DATA_SOURCE = AzureDataLakeStore,
    FILE_FORMAT = PARQUET,
    REJECT_TYPE = VALUE,
    REJECT_VALUE = 0
);
```

D.

```
ALTER TABLE [Ext].[Items]
ADD [ItemID] int;
```

Explanation:

Correct Answer:C

Incorrect Answers:

A, D: Only these Data Definition Language (DDL) statements are allowed on external tables:

☛ CREATE TABLE and DROP TABLE

☛ CREATE STATISTICS and DROP STATISTIC

☛ CREATE VIEW and DROP VIEW

[Reference: https://docs.microsoft.com/en-us/sql/t-sql/statements/create-external-table-transact-sql](https://docs.microsoft.com/en-us/sql/t-sql/statements/create-external-table-transact-sql)

Question 16

CertyIQ

You have two Azure Storage accounts named Storage1 and Storage2. Each account holds one container and has the hierarchical namespace enabled. The system has files that contain data stored in the Apache Parquet format. You need to copy folders and files from Storage1 to Storage2 by using a Data Factory copy activity. The solution must meet the following requirements:

- ☞ No transformations must be performed.
- ☞ The original folder structure must be retained.
- ☞ Minimize time required to perform the copy activity.

How should you configure the copy activity? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Source dataset type:

Binary
Parquet
Delimited text

Copy activity copy behavior:

FlattenHierarchy
MergeFiles
PreserveHierarchy

Answer Area

Source dataset type:

Binary
Parquet
Delimited text

Copy activity copy behavior:

FlattenHierarchy
MergeFiles
PreserveHierarchy

Explanation:

Correct Answer: I tend to believe that would be binary the correct answer.

Box 2: PreserveHierarchy -

PreserveHierarchy (default): Preserves the file hierarchy in the target folder. The relative path of the source file to the source folder is identical to the relative path of the target file to the target folder.

Incorrect Answers:

☞ FlattenHierarchy: All files from the source folder are in the first level of the target folder. The target files have autogenerated names.

☞ MergeFiles: Merges all files from the source folder to one file. If the file name is specified, the merged file name is

the specified name. Otherwise, it's an autogenerated file name.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/format-parquet>

[t https://docs.microsoft.com/en-us/azure/data-factory/connector-azure-data-lake-storage](https://docs.microsoft.com/en-us/azure/data-factory/connector-azure-data-lake-storage)

Question 17

CertyIQ

You have an Azure Data Lake Storage Gen2 container that contains 100 TB of data.

You need to ensure that the data in the container is available for read workloads in a secondary region if an outage occurs in the primary region. The solution must minimize costs.

Which type of data redundancy should you use?

- A. geo-redundant storage (GRS)
- B. read-access geo-redundant storage (RA-GRS)
- C. zone-redundant storage (ZRS)
- D. locally-redundant storage (LRS)

Explanation:

Correct Answer: B

Geo-redundant storage (with GRS or GZRS) replicates your data to another physical location in the secondary region to protect against regional outages.

However, that data is available to be read only if the customer or Microsoft initiates a failover from the primary to secondary region. When you enable read access to the secondary region, your data is available to be read at all times, including in a situation where the primary region becomes unavailable.

Incorrect Answers:

A: While Geo-redundant storage (GRS) is cheaper than Read-Access Geo-Redundant Storage (RA-GRS), GRS does NOT initiate automatic failover.

C, D: Locally redundant storage (LRS) and Zone-redundant storage (ZRS) provides redundancy within a single region..

Reference:

<https://docs.microsoft.com/en-us/azure/storage/common/storage-redundancy>

Question 18

CertyIQ

You need to ensure that the data lake will remain available if a data center fails in the primary Azure region. The solution must minimize costs.

Which type of replication should you use for the storage account?

- A. geo-redundant storage (GRS)
- B. geo-zone-redundant storage (GZRS)
- C. locally-redundant storage (LRS)
- D. zone-redundant storage (ZRS)

Explanation:

Correct Answer: D

Zone-redundant storage (ZRS) copies your data synchronously across three Azure availability zones in the primary region.

Incorrect Answers:

C: Locally redundant storage (LRS) copies your data synchronously three times within a single physical location in the

primary region. LRS is the least expensive replication option, but is not recommended for applications requiring high availability or durability

Reference:

<https://docs.microsoft.com/en-us/azure/storage/common/storage-redundancy>

Question 19

CertyIQ

You have a SQL pool in Azure Synapse.

You plan to load data from Azure Blob storage to a staging table. Approximately 1 million rows of data will be loaded daily. The table will be truncated before each daily load.

You need to create the staging table. The solution must minimize how long it takes to load the data to the staging table.

How should you configure the table? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Distribution:

Hash
Replicated
Round-robin

Indexing:

Clustered
Clustered columnstore
Heap

Partitioning:

Date
None

Answer Area

Distribution:

Hash
Replicated
Round-robin

Indexing:

Clustered
Clustered columnstore
Heap

Partitioning:

Date
None

Explanation:

Correct Answer:

Round-Robin (1), Heap (2), None (3). Within this doc: [#1. Search for "Use round-robin for the staging table." #2. Search for: "A heap table can be especially useful for loading data, such as a staging table,..."](https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-overview) Within this doc:

[#](https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-partition?context=azure/synapse-analytics/context/context)

3. Partitioning by date is useful when stage destination has data because you can hide the inserting data's new partition (to keep users from hitting it), complete the load and then unhide the new partition. However, in this question it states, "the table will be truncated before each daily load", so, it appears it's a true Staging table and there are no users with access, no existing data, and I see no reason to have a Date partition. To me, such a partition would do nothing but slow the load.

Question 20

CertyIQ

You are designing a fact table named FactPurchase in an Azure Synapse Analytics dedicated SQL pool. The table contains purchases from suppliers for a retail store. FactPurchase will contain the following columns.

Name	Data type	Nullable
PurchaseKey	Bigint	No
DateKey	Int	No
SupplierKey	Int	No
StockItemKey	Int	No
PurchaseOrderID	Int	Yes
OrderedQuantity	Int	No
OrderedOuters	Int	No
ReceivedOuters	Int	No
Package	Nvarchar(50)	No
IsOrderFinalized	Bit	No
LineageKey	Int	No

FactPurchase will have 1 million rows of data added daily and will contain three years of data. Transact-SQL queries similar to the following query will be executed daily.

SELECT -
SupplierKey, StockItemKey, IsOrderFinalized, COUNT(*)

FROM FactPurchase -

WHERE DateKey >= 20210101 -

AND DateKey <= 20210131 -
GROUP By SupplierKey, StockItemKey, IsOrderFinalized
Which table distribution will minimize query times?

- A. replicated
- **B. hash-distributed on PurchaseKey**
- C. round-robin
- D. hash-distributed on IsOrderFinalized

Explanation:

Correct Answer: : B

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

To balance the parallel processing, select a distribution column that:

- ☞ Has many unique values. The column can have duplicate values. All rows with the same value are assigned to the same distribution. Since there are 60 distributions, some distributions can have > 1 unique values while others may end with zero values.
- ☞ Does not have NULLs, or has only a few NULLs.
- ☞ Is not a date column.

Incorrect Answers:

C: Round-robin tables are useful for improving loading speed.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute>

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables->

Question 21

CertyIQ

HOTSPOT -

From a website analytics system, you receive data extracts about user interactions such as downloads, link clicks, form submissions, and video plays.

The data contains the following columns.

Name	Sample value
Date	15 Jan 2021
EventCategory	Videos
EventAction	Play
EventLabel	Contoso Promotional
ChannelGrouping	Social
TotalEvents	150
UniqueEvents	120
SessionWithEvents	99

You need to design a star schema to support analytical queries of the data. The star schema will contain four tables including a date dimension.

To which table should you add each column? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

EventCategory:

DimChannel
DimDate
DimEvent
FactEvents

ChannelGrouping:

DimChannel
DimDate
DimEvent
FactEvents

TotalEvents:

DimChannel
DimDate
DimEvent
FactEvents

Answer Area

EventCategory:

DimChannel
DimDate
DimEvent
FactEvents

Correct Answer:

ChannelGrouping:

DimChannel
DimDate
DimEvent
FactEvents

TotalEvents:

DimChannel
DimDate
DimEvent
FactEvents

Explanation:

Correct Answer:

Box 1: DimEvent -

Box 2: DimChannel -

Box 3: FactEvents -

Fact tables store observations or events, and can be sales orders, stock balances, exchange rates, temperatures, etc

Reference:

<https://docs.microsoft.com/en-us/power-bi/guidance/star-schema>

Question 22

CertyIQ

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 Storage account that contains 100 GB of files. The files contain rows of text and numerical values. 75% of the rows contain description data that has an average length of 1.1 MB.

You plan to copy the data from the storage account to an enterprise data warehouse in Azure Synapse Analytics.

You need to prepare the files to ensure that the data copies quickly.

Solution: You convert the files to compressed delimited text files.

Does this meet the goal?

- A. Yes
- B. No

Explanation:

Correct Answer: A

convert the files to compressed delimited text files.

<https://azure.microsoft.com/en-gb/blog/increasing-polybase-row-width-limitation-in-azure-sql-data-warehouse/>

Question 23

CertyIQ

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 Storage account that contains 100 GB of files. The files contain rows of text and numerical values. 75% of the rows contain description data that has an average length of 1.1 MB.

You plan to copy the data from the storage account to an enterprise data warehouse in Azure Synapse Analytics.

You need to prepare the files to ensure that the data copies quickly.

Solution: You copy the files to a table that has a columnstore index.

Does this meet the goal?

- A. Yes
- B. No

Explanation:

Correct Answer: B

Instead convert the files to compressed delimited text files.

Reference:

<https://docs.microsoft.com/en-us/azure/sql-data-warehouse/guidance-for-loading-data>

From the documentation, loads to heap table are faster than indexed tables. So, better to use heap table than columnstore index table in this case

Question 24

CertyIQ

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 Storage account that contains 100 GB of files. The files contain rows of text and numerical values. 75% of the rows contain description data that has an average length of 1.1 MB.

You plan to copy the data from the storage account to an enterprise data warehouse in Azure Synapse Analytics.

You need to prepare the files to ensure that the data copies quickly.

Solution: You modify the files to ensure that each row is more than 1 MB.

Does this meet the goal?

- A. Yes
- B. No

Explanation:

Correct Answer: No, rows need to have less than 1 MB. A batch size between 100 K to 1M rows is the recommended baseline for determining optimal batch size capacity.

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-temporary>

Question 25

CertyIQ

You build a data warehouse in an Azure Synapse Analytics dedicated SQL pool. Analysts write a complex SELECT query that contains multiple JOIN and CASE statements to transform data for use in inventory reports. The inventory reports will use the data and additional WHERE parameters depending on the report. The reports will be produced once daily. You need to implement a solution to make the dataset available for the reports. The solution must minimize query times. What should you implement?

- A. an ordered clustered columnstore index
- B. a materialized view
- C. result set caching
- D. a replicated table

Explanation:

Correct Answer: : B

Materialized views for dedicated SQL pools in Azure Synapse provide a low maintenance method for complex analytical queries to get fast performance without any query change.

Incorrect Answers:

C: One daily execution does not make use of result cache caching.

Note: When result set caching is enabled, dedicated SQL pool automatically caches query results in the user database for repetitive use. This allows subsequent query executions to get results directly from the persisted cache so recomputation is not needed. Result set caching improves query performance and reduces compute resource usage. In addition, queries using cached results set do not use any concurrency slots and thus do not count against existing concurrency limits.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/performance-tuning-materialized-views>
<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/performance-tuning-result-set-caching>

Question 26

CertyIQ

You have an Azure Synapse Analytics workspace named WS1 that contains an Apache Spark pool named Pool1.

You plan to create a database named DB1 in Pool1.

You need to ensure that when tables are created in DB1, the tables are available automatically as external tables to the built-in serverless SQL pool.

Which format should you use for the tables in DB1?

- A. CSV
- B. ORC
- C. JSON
- D. Parquet

Explanation:

Correct Answer: AD

"For each Spark external table based on Parquet or CSV and located in Azure Storage, an external table is created in a serverless SQL pool database.

As such, you can shut down your Spark pools and still query Spark external tables from serverless SQL pool."

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-temporary>

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 Event Hubs
- B. Azure Data Factory
- C. Azure Stream Analytics
- D. Azure Databricks

Explanation:

Correct Answer: D

The following tables summarize the key differences in capabilities for stream processing technologies in Azure.

General capabilities -

Capability	Azure Stream Analytics	HDInsight with Spark Streaming	Apache Spark in Azure Databricks	HDInsight with Storm
Programmability	Stream analytics query language, JavaScript	C#/F# ↗, Java, Python, Scala	C#/F# ↗, Java, Python, R, Scala	C#, Java

Integration capabilities -

Capability	Azure Stream Analytics	HDInsight with Spark Streaming	Apache Spark in Azure Databricks	HDInsight with Storm
Inputs	Azure Event Hubs, IoT Hubs, Azure IoT Hub, Azure Blob storage	Event Hubs, IoT Hub, Kafka, HDFS, Storage Blobs, Azure Data Lake Store	Event Hubs, IoT Hub, Kafka, HDFS, Storage Blobs, Azure Data Lake Store	Event Hubs, IoT Hub, Storage Blobs, Azure Data Lake Store
Sinks	Azure Data Lake Store, Azure SQL Database, Storage Blobs, Event	HDFS, Kafka, Storage Blobs, Azure Data Lake Store, Cosmos DB	HDFS, Kafka, Storage Blobs, Azure Data Lake Store	Event Hubs, Service Bus, Kafka

Reference:

<https://docs.microsoft.com/en-us/azure/architecture/data-guide/technology-choices/stream-processing>

Question 28

CertyIQ

You plan to implement an Azure Data Lake Storage Gen2 container that will contain CSV files. The size of the files will vary based on the number of events that occur per hour.

File sizes range from 4 KB to 5 GB.

You need to ensure that the files stored in the container are optimized for batch processing.

What should you do?

- A. Convert the files to JSON
- B. Convert the files to Avro
- C. Compress the files
- D. Merge the files

Explanation:

Correct Answer: B

Avro supports batch and is very relevant for streaming.

Note: Avro is framework developed within Apache's Hadoop project. It is a row-based storage format which is widely used as a serialization process. AVRO stores its schema in JSON format making it easy to read and interpret by any program. The data itself is stored in binary format by doing it compact and efficient.

Reference:

<https://www.adaltas.com/en/2020/07/23/benchmark-study-of-different-file-format/>

HOTSPOT -

You store files in an Azure Data Lake Storage Gen2 container. The container has the storage policy shown in the following exhibit.

```
{  
    "rules": [  
        {  
            "enabled": true,  
            "name": "contosorule",  
            "type": "Lifecycle",  
            "definition": {  
                "actions": {  
                    "version": {  
                        "delete": {  
                            "daysAfterCreationGreaterThanOrEqual": 60  
                        }  
                    },  
                    "baseBlob": {  
                        "tierToCool": {  
                            "daysAfterModificationGreaterThanOrEqual":  
                                30  
                        },  
                        "move": {  
                            "daysAfterModificationGreaterThanOrEqual": 30  
                        }  
                    }  
                },  
                "filters": {  
                    "blobTypes": [  
                        "blockBlob"  
                    ],  
                    "prefixMatch": [  
                        "container1/contoso"  
                    ]  
                }  
            }  
        }  
    ]  
}
```

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 files are [answer choice] after 30 days:

▼
deleted from the container
moved to archive storage
moved to cool storage
moved to hot storage

The storage policy applies to [answer choice]:

▼
container1/contoso.csv
container1/docs/contoso.json
container1/mycontoso/contoso.csv

Correct Answer:

Answer Area

The files are [answer choice] after 30 days:

▼
deleted from the container
moved to archive storage
moved to cool storage
moved to hot storage

The storage policy applies to [answer choice]:

▼
container1/contoso.csv
container1/docs/contoso.json
container1/mycontoso/contoso.csv

Explanation:

Correct Answer:

Box 1: moved to cool storage -

The ManagementPolicyBaseBlob.TierToCool property gets or sets the function to tier blobs to cool storage. Support blobs currently at Hot tier.

Box 2: container1/contoso.csv -

As defined by prefixMatch.

prefixMatch: An array of strings for prefixes to be matched. Each rule can define up to 10 case-sensitive prefixes. A prefix string must start with a container name.

Reference:

<https://docs.microsoft.com/en-us/dotnet/api/microsoft.azure.management.storage.fluent.models.managementpolicybaseblob.tiertocool>

Question 30

CertyIQ

You are designing a financial transactions table in an Azure Synapse Analytics dedicated SQL pool. The table will have a clustered columnstore index and will include the following columns:

⌚ TransactionType: 40 million rows per transaction type

⌚ CustomerSegment: 4 million per customer segment

⌚ TransactionMonth: 65 million rows per month

⌚ AccountType: 500 million per account type

You have the following query requirements:

⌚ Analysts will most commonly analyze transactions for a given month.

⌚ Transactions analysis will typically summarize transactions by transaction type, customer segment, and/or account type

You need to recommend a partition strategy for the table to minimize query times.

On which column should you recommend partitioning the table?

- A. CustomerSegment
- B. AccountType
- C. TransactionType
- D. TransactionMonth

Explanation:

Correct Answer: D

For optimal compression and performance of clustered columnstore tables, a minimum of 1 million rows per distribution and partition is needed. Before partitions are created, dedicated SQL pool already divides each table into 60 distributed databases.

Example: Any partitioning added to a table is in addition to the distributions created behind the scenes. Using this example, if the sales fact table contained 36 monthly partitions, and given that a dedicated SQL pool has 60 distributions, then the sales fact table should contain 60 million rows per month, or 2.1 billion rows when all months are populated. If a table contains fewer than the recommended minimum number of rows per partition, consider using fewer partitions in order to increase the number of rows per partition.

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-temporary>

Question 31

CertyIQ

HOTSPOT -

You have an Azure Data Lake Storage Gen2 account named account1 that stores logs as shown in the following table.

Type	Designated retention period
Application	360 days
Infrastructure	60 days

You do not expect that the logs will be accessed during the retention periods.

You need to recommend a solution for account1 that meets the following requirements:

☞ Automatically deletes the logs at the end of each retention period

☞ Minimizes storage costs

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

To minimize storage costs:

Store the infrastructure logs and the application logs in the Archive access tier	▼
Store the infrastructure logs and the application logs in the Cool access tier	
Store the infrastructure logs in the Cool access tier and the application logs in the Archive access tier	

To delete logs automatically:

Azure Data Factory pipelines	▼
Azure Blob storage lifecycle management rules	
Immutable Azure Blob storage time-based retention policies	

Correct Answer:

Answer Area

To minimize storage costs:

- Store the infrastructure logs and the application logs in the Archive access tier
- Store the infrastructure logs and the application logs in the Cool access tier
- Store the infrastructure logs in the Cool access tier and the application logs in the Archive access tier**

To delete logs automatically:

- Azure Data Factory pipelines
- Azure Blob storage lifecycle management rules**
- Immutable Azure Blob storage time-based retention policies

Explanation:

Correct Answer

Box 1: Store the infrastructure logs in the Cool access tier and the application logs in the Archive access tier

For infrastructure logs: Cool tier - An online tier optimized for storing data that is infrequently accessed or modified.

Data in the cool tier should be stored for a minimum of 30 days. The cool tier has lower storage costs and higher access costs compared to the hot tier.

For application logs: Archive tier - An offline tier optimized for storing data that is rarely accessed, and that has flexible latency requirements, on the order of hours.

Data in the archive tier should be stored for a minimum of 180 days.

Box 2: Azure Blob storage lifecycle management rules

Blob storage lifecycle management offers a rule-based policy that you can use to transition your data to the desired access tier when your specified conditions are met. You can also use lifecycle management to expire data at the end of its life.

Reference:

<https://docs.microsoft.com/en-us/azure/storage/blobs/access-tiers-overview>

Question 32

CertyIQ

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. JSON
- B. Parquet**
- C. CSV
- D. Avro

Explanation:

Correct Answer: **B**

Need Parquet to support both Databricks and PolyBase.

Reference:

<https://docs.microsoft.com/en-us/sql/t-sql/statements/create-external-file-format-transact-sql>

Question 33

CertyIQ

You have an Azure Synapse Analytics dedicated SQL pool named Pool1. Pool1 contains a partitioned fact table named dbo.Sales and a staging table named stg.Sales that has the matching table and partition definitions.

You need to overwrite the content of the first partition in dbo.Sales with the content of the same partition in stg.Sales. The solution must minimize load times.

What should you do?

- A. Insert the data from stg.Sales into dbo.Sales.
- B. Switch the first partition from dbo.Sales to stg.Sales.
- C. Switch the first partition from stg.Sales to dbo.Sales.
- D. Update dbo.Sales from stg.Sales.

Explanation:

Correct Answer: The below link describes a similar scenario. Answer should be C <https://medium.com/@cocci.g/switch-partitions-in-azure-synapse-sql-dw-1e0e32309872>

Question 34

CertyIQ

You are designing a slowly changing dimension (SCD) for supplier data in an Azure Synapse Analytics dedicated SQL pool.

You plan to keep a record of changes to the available fields.

The supplier data contains the following columns.

Name	Description
SupplierSystemID	Unique supplier ID in an enterprise resource planning (ERP) system
SupplierName	Name of the supplier company
SupplierAddress1	Address of the supplier company
SupplierAddress2	Second address of the supplier company
SupplierCity	City of the supplier company
SupplierStateProvince	State or province of the supplier company
SupplierCountry	Country of the supplier company
SupplierPostalCode	Postal code of the supplier company
SupplierDescription	Free-text description of the supplier company
SupplierCategory	Category of goods provided by the supplier company

Which three additional columns should you add to the data to create a Type 2 SCD? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. surrogate primary key
- B. effective start date
- C. business key
- D. last modified date
- E. effective end date
- F. foreign key

Explanation:

Correct Answer:

The answer is ABE. A type 2 SCD requires a surrogate key to uniquely identify each record when versioning. See <https://docs.microsoft.com/en-us/learn/modules/populate-slowly-changing-dimensions-azure-synapse-analytics-pipelines/3-choose-between-dimension-types> under SCD Type 2 "the dimension table must use a surrogate key to provide a unique reference to a version of the dimension member." A business key is already part of this table - SupplierSystemID. The column is derived from the source data.

Question 35

CertyIQ

HOTSPOT -

You have a Microsoft SQL Server database that uses a third normal form schema.

You plan to migrate the data in the database to a star schema in an Azure Synapse Analytics dedicated SQL pool.

You need to design the dimension tables. The solution must optimize read operations.

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Transform data for the dimension tables by:

	▼
Maintaining to a third normal form	▼
Normalizing to a fourth normal form	▼
Denormalizing to a second normal form	▼

For the primary key columns in the dimension tables, use:

	▼
New IDENTITY columns	▼
A new computed column	▼
The business key column from the source sys	▼

Correct Answer:

Answer Area

Transform data for the dimension tables by:

	▼
Maintaining to a third normal form	▼
Normalizing to a fourth normal form	▼
Denormalizing to a second normal form	▼

For the primary key columns in the dimension tables, use:

	▼
New IDENTITY columns	▼
A new computed column	▼
The business key column from the source sys	▼

Explanation:

Correct Answer:

Box 1: Denormalize to a second normal form

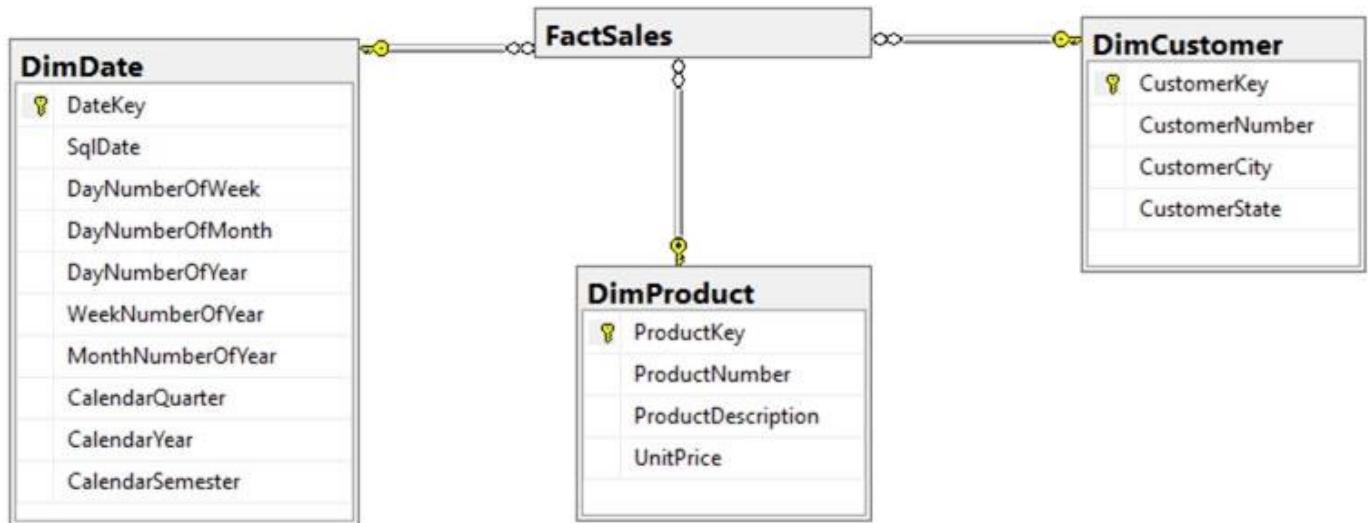
Denormalization is the process of transforming higher normal forms to lower normal forms via storing the join of higher normal form relations as a base relation.

Denormalization increases the performance in data retrieval at cost of bringing update anomalies to a database.

Box 2: New identity columns -

The collapsing relations strategy can be used in this step to collapse classification entities into component entities to obtain flat dimension tables with single-part keys that connect directly to the fact table. The single-part key is a surrogate key generated to ensure it remains unique over time.

Example:



Note: A surrogate key on a table is a column with a unique identifier for each row. The key is not generated from the table data. Data modelers like to create surrogate keys on their tables when they design data warehouse models. You can use the IDENTITY property to achieve this goal simply and effectively without affecting load performance.

Reference:

<https://www.mssqltips.com/sqlservertip/5614/explore-the-role-of-normal-forms-in-dimensional-modeling/>

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-identity>

Question 36

CertyIQ

HOTSPOT -

You plan to develop a dataset named Purchases by using Azure Databricks. Purchases will contain the following columns:

- ⌚ ProductID
- ⌚ ItemPrice
- ⌚ LineTotal
- ⌚ Quantity
- ⌚ StoreID
- ⌚ Minute
- ⌚ Month
- ⌚ Hour

Year -

-
- ⌚ Day

You need to store the data to support hourly incremental load pipelines that will vary for each Store ID. The solution must minimize storage costs.

How should you complete the code? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

df.write

.bucketBy
.partitionBy
.range
.sortBy

.mode("append")

.csv("/Purchases")
.json("/Purchases")
.parquet("/Purchases")
.saveAsTable("/Purchases")

(*)
("StoreID", "Hour")
("StoreID", "Year", "Month", "Day", "Hour")

Answer Area

df.write

.bucketBy
.partitionBy
.range
.sortBy

.mode("append")

.csv("/Purchases")
.json("/Purchases")
.parquet("/Purchases")
.saveAsTable("/Purchases")

(*)
("StoreID", "Hour")
("StoreID", "Year", "Month", "Day", "Hour")

Explanation:

Correct Answer: Box 1: partitionBy -

We should overwrite at the partition level.

Example:

```
df.write.partitionBy("y","m","d")
.mode(SaveMode.Append)
.parquet("/data/hive/warehouse/db_name.db/" + tableName)
```

Box 2: ("StoreID", "Year", "Month", "Day", "Hour", "StoreID")

Box 3: parquet("/Purchases")

Reference:

<https://intellipaat.com/community/11744/how-to-partition-and-write-dataframe-in-spark-without-deleting-partitions-with-no-new-data>

Question 37

CertyIQ

You are designing a partition strategy for a fact table in an Azure Synapse Analytics dedicated SQL pool. The table has the following specifications:

- ☛ Contain sales data for 20,000 products.

Use hash distribution on a column named ProductID.

Contain 2.4 billion records for the years 2019 and 2020.

Which number of partition ranges provides optimal compression and performance for the clustered columnstore index?

- A. 40
- B. 240
- C. 400
- D. 2,400

Explanation:

Correct Answer:A

Each partition should have around 1 millions records. Dedicated SQL pools already have 60 partitions.

We have the formula: Records/(Partitions*60)= 1 million

Partitions= Records/(1 million * 60)

Partitions= $2.4 \times 1,000,000,000 / (1,000,000 * 60) = 40$

Note: Having too many partitions can reduce the effectiveness of clustered columnstore indexes if each partition has fewer than 1 million rows. Dedicated SQL pools automatically partition your data into 60 databases. So, if you create a table with 100 partitions, the result will be 6000 partitions.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/best-practices-dedicated-sql-pool>

Question 38

CertyIQ

HOTSPOT -

You are creating dimensions for a data warehouse in an Azure Synapse Analytics dedicated SQL pool.

You create a table by using the Transact-SQL statement shown in the following exhibit.

```
CREATE TABLE [dbo].[DimProduct] (
    [ProductKey] [int] IDENTITY(1,1) NOT NULL,
    [ProductSourceID] [int] NOT NULL,
    [ProductName] [nvarchar](100) NOT NULL,
    [ProductNumber] [nvarchar](25) NOT NULL,
    [Color] [nvarchar](15) NULL,
    [Size] [nvarchar](5) NULL,
    [Weight] [decimal](8, 2) NULL,
    [ProductCategory] [nvarchar](100) NULL,
    [SellStartDate] [date] NOT NULL,
    [SellEndDate] [date] NULL,
    [RowInsertedDateTime] [datetime] NOT NULL,
    [RowUpdatedDateTime] [datetime] NOT NULL,
    [ETLAuditID] [int] NOT NULL
)
```

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

DimProduct is a [answer choice] slowly changing dimension (SCD).

Type 0
Type 1
Type 2

The ProductKey column is [answer choice].

a surrogate key
a business key
an audit column

Answer Area

DimProduct is a [answer choice] slowly changing dimension (SCD).

Type 0
Type 1
Type 2

The ProductKey column is [answer choice].

a surrogate key
a business key
an audit column

Explanation:

Correct Answer:

Box 1: Type 2 -

A Type 2 SCD supports versioning of dimension members. Often the source system doesn't store versions, so the data warehouse load process detects and manages changes in a dimension table. In this case, the dimension table must use a surrogate key to provide a unique reference to a version of the dimension member. It also includes columns that define the date range validity of the version (for example, StartDate and EndDate) and possibly a flag column (for example, IsCurrent) to easily filter by current dimension members.

Incorrect Answers:

A Type 1 SCD always reflects the latest values, and when changes in source data are detected, the dimension table data is overwritten. Agree on the surrogate key, exactly. "In data warehousing, IDENTITY functionality is particularly important as it makes easier the creation of surrogate keys." Why ProductKey is certainly not a business key: "The IDENTITY value in Synapse is not guaranteed to be unique if the user explicitly inserts a duplicate value with 'SET IDENTITY_INSERT ON' or reseeds IDENTITY". Business key is an index which identifies uniqueness of a row and here Microsoft says that identity doesn't guarantee uniqueness.

References

<https://azure.microsoft.com/en-us/blog/identity-now-available/>

with-azure-sql-data-warehouse/ <https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-identity>

You are designing a fact table named FactPurchase in an Azure Synapse Analytics dedicated SQL pool. The table contains purchases from suppliers for a retail store. FactPurchase will contain the following columns.

Name	Data type	Nullable
PurchaseKey	Bigint	No
DateKey	Int	No
SupplierKey	Int	No
StockItemKey	Int	No
PurchaseOrderID	Int	Yes
OrderedQuantity	Int	No
OrderedOuters	Int	No
ReceivedOuters	Int	No
Package	Nvarchar(50)	No
IsOrderFinalized	Bit	No
LineageKey	Int	No

FactPurchase will have 1 million rows of data added daily and will contain three years of data. Transact-SQL queries similar to the following query will be executed daily.

```
SELECT -
SupplierKey, StockItemKey, COUNT(*)
FROM FactPurchase -
WHERE DateKey >= 20210101 -
AND DateKey <= 20210131 -
GROUP By SupplierKey, StockItemKey
```

Which table distribution will minimize query times?

- A. replicated
- B. hash-distributed on PurchaseKey
- C. round-robin
- D. hash-distributed on DateKey

Explanation:

Correct Answer:B

Hash-distributed tables improve query performance on large fact tables, and are the focus of this article. Round-robin tables are useful for improving loading speed.

Incorrect:

Not D: Do not use a date column. . All data for the same date lands in the same distribution. If several users are all filtering on the same date, then only 1 of the 60 distributions do all the processing work.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute>

You are implementing a batch dataset in the Parquet format.

Data files will be produced by using Azure Data Factory and stored in Azure Data Lake Storage Gen2. The files will be consumed by an Azure Synapse Analytics serverless SQL pool.

You need to minimize storage costs for the solution.

What should you do?

- A. Use Snappy compression for the files.
- B. Use OPENROWSET to query the Parquet files.
- C. Create an external table that contains a subset of columns from the Parquet files.
- D. Store all data as string in the Parquet files.

Explanation:

Correct Answer: C

An external table points to data located in Hadoop, Azure Storage blob, or Azure Data Lake Storage. External tables are used to read data from files or write data to files in Azure Storage. With Synapse SQL, you can use external tables to read external data using dedicated SQL pool or serverless SQL pool.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables>

End of Part 1



Want all PDFs of DP-203?

Please drop us mail at: - certyiq@gmail.com

Please find the videos of this AZ-900/AI-900/AZ-305/
AZ-104 /DP-900/ SC-900 and other Microsoft exam
series on

CertyIQ Official YouTube channel (FREE PDFs): -

Please [Subscribe](#) to CertyIQ YouTube Channel to get notified for latest exam dumps by clicking on the below image, it will redirect to the [CertyIQ](#) YouTube page.



Connect with us @ [LinkedIn](#) [Telegram](#)

Contact us for other dumps: certyiqpdf@gmail.com

For any other enquiry, please drop us a mail at
certyiqofficial@gmail.com

DP-203

Real Exam Questions & Answers



Latest Exam Questions
All Topics Covered-2023
Added New Que's



Get Certified Today!

New Course Covered

Subscribe

Question 81

CertyIQ

HOTSPOT -

You have an enterprise data warehouse in Azure Synapse Analytics that contains a table named FactOnlineSales. The table contains data from the start of 2009 to the end of 2012.

You need to improve the performance of queries against FactOnlineSales by using table partitions. The solution must meet the following requirements:

- ⇒ Create four partitions based on the order date.
- ⇒ Ensure that each partition contains all the orders placed during a given calendar year.

How should you complete the T-SQL command? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

```
CREATE TABLE [dbo].FactOnlineSales
([OnlineSalesKey] [int] NOT NULL,
[OrderDateKey] [datetime] NOT NULL,
[StoreKey] [int] NOT NULL,
[ProductKey] [int] NOT NULL,
[CustomerKey] [int] NOT NULL,
[SalesOrderNumber] [nvarchar] (20) NOT NULL,
[SalesQuantity] [int] NOT NULL,
[SalesAmount] [money] NOT NULL,
[UnitPrice] [money] NULL)
WITH (CLUSTERED COLUMNSTORE INDEX)
PARTITION ([OrderDateKey] RANGE
```

	▼
RIGHT	
LEFT	

FOR VALUES

(▼)
20090101,20121231		
20100101,20110101,20120101		
20090101,20100101,20110101,20120101		

Answer Area

```
CREATE TABLE [dbo].FactOnlineSales
([OnlineSalesKey] [int] NOT NULL,
[OrderDateKey] [datetime] NOT NULL,
[StoreKey] [int] NOT NULL,
[ProductKey] [int] NOT NULL,
[CustomerKey] [int] NOT NULL,
[SalesOrderNumber] [nvarchar] (20) NOT NULL,
[SalesQuantity] [int] NOT NULL,
[Correct Answer: SalesAmount] [money] NOT NULL,
[UnitPrice] [money] NULL)
WITH (CLUSTERED COLUMNSTORE INDEX)
PARTITION ([OrderDateKey] RANGE
```

	▼
RIGHT	
LEFT	

FOR VALUES

(▼)
20090101,20121231		
20100101,20110101,20120101		
20090101,20100101,20110101,20120101		

Explanation:

Correct Answer Range Left or Right, both are creating similar partition but there is difference in comparison

For example: in this scenario, when you use LEFT and 20100101,20110101,20120101

Partition will be, datecol<=20100101, datecol>20100101 and datecol<=20110101, datecol>20110101 and

datecol<=20120101, datecol>20120101

But if you use range RIGHT and 20100101,20110101,20120101

Partition will be, datecol<20100101, datecol>=20100101 and datecol<20110101, datecol>=20110101 and datecol<20120101, datecol>=20120101

In this example, Range RIGHT will be suitable for calendar comparison Jan 1st to Dec 31st

Reference:

<https://docs.microsoft.com/en-us/sql/t-sql/statements/create-partition-function-transact-sql?view=sql-server-ver15>

Question 82

CertyIQ

You need to implement a Type 3 slowly changing dimension (SCD) for product category data in an Azure Synapse Analytics dedicated SQL pool.

You have a table that was created by using the following Transact-SQL statement.

```
CREATE TABLE [DBO].[DimProduct] (
[ProductKey] [int] IDENTITY(1,1) NOT NULL,
[ProductSourceID] [int] NOT NULL,
[ProductName] [nvarchar] (100) NULL,
[Color] [nvarchar] (15) NULL,
[SellStartDate] [date] NOT NULL,
[SellEndDate] [date] NULL,
[RowInsertedDateTime] [datetime] NOT NULL,
[RowUpdatedDateTime] [datetime] NOT NULL,
[ETLAuditID] [int] NOT NULL
)
```

Which two columns should you add to the table? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. [EffectiveStartDate] [datetime] NOT NULL,
- B. [CurrentProductCategory] [nvarchar] (100) NOT NULL,
- C. [EffectiveEndDate] [datetime] NULL,
- D. [ProductCategory] [nvarchar] (100) NOT NULL,
- E. [OriginalProductCategory] [nvarchar] (100) NOT NULL,

Explanation:

Correct Answer BE

A Type 3 SCD supports storing two versions of a dimension member as separate columns. The table includes a column for the current value of a member plus either the original or previous value of the member. So Type 3 uses additional columns to track one key instance of history, rather than storing additional rows to track each change like in a Type 2 SCD.

This type of tracking may be used for one or two columns in a dimension table. It is not common to use it for many members of the same table. It is often used in combination with Type 1 or Type 2 members.

CustomerID	FirstName	LastName	CurrentEmail	OriginalEmail	CompanyName	InsertedDate	ModifiedDate
2	Keith	Harris	keith0@aw.com	keith0@aw.com	Progressive Sports	2021-03-20	2021-03-20
3	Donna	Carreras	donna0@aw.com	donna0@aw.com	A Bike Store	2021-03-20	2021-03-20

CustomerID	FirstName	LastName	CurrentEmail	OriginalEmail	CompanyName	InsertedDate	ModifiedDate
2	Keith	Harris	keith0@aw.com	keith0@aw.com	Progressive Sports	2021-03-20	2021-03-20
3	Donna	Carreras	dc3@aw.com	donna0@aw.com	A Bike Store	2021-03-20	2021-03-22

Reference:

<https://k21academy.com/microsoft-azure/azure-data-engineer-dp203-q-a-day-2-live-session-review/>

Question 83

CertyIQ

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 Stream Analytics solution that will analyze Twitter data.

You need to count the tweets in each 10-second window. The solution must ensure that each tweet is counted only once.

Solution: You use a hopping window that uses a hop size of 10 seconds and a window size of 10 seconds.

Does this meet the goal?

- A. Yes
- B. No

Explanation:

Correct Answer The answer "Yes". Hopping window with hop size equals window size should be the same as Tumbling window.

Note that a tumbling window is simply a hopping window whose 'hop' is equal to its 'size'.

<https://learn.microsoft.com/en-us/stream-analytics-query/hopping-window-azure-stream-analytics>

Question 84

CertyIQ

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 Stream Analytics solution that will analyze Twitter data.

You need to count the tweets in each 10-second window. The solution must ensure that each tweet is counted only once.

Solution: You use a hopping window that uses a hop size of 5 seconds and a window size 10 seconds.

Does this meet the goal?

- A. Yes
- B. No

Explanation:

Correct Answer **B**

Instead use a tumbling window. Tumbling windows are a series of fixed-sized, non-overlapping and contiguous time intervals.

Reference:

<https://docs.microsoft.com/en-us/stream-analytics-query/tumbling-window-azure-stream-analytics>

Question 85

CertyIQ

HOTSPOT -

You are building an Azure Stream Analytics job to identify how much time a user spends interacting with a feature on a webpage.

The job receives events based on user actions on the webpage. Each row of data represents an event. Each event has a type of either 'start' or 'end'.

You need to calculate the duration between start and end events.

How should you complete the query? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

```

SELECT
    [user],
    feature,
    DATEADD(
        DATEDIFF(
            DATEPART(
                second,
                ISFIRST
                LAST
                TOPONE
            )
        )
    ) OVER (PARTITION BY [user], feature LIMIT DURATION(hour, 1) WHEN Event = 'start'),
    (Time) OVER (PARTITION BY [user], feature LIMIT DURATION(hour, 1) WHEN Event = 'start') -
    (Time) OVER (PARTITION BY [user], feature LIMIT DURATION(hour, 1) WHEN Event = 'end') -
    (Time) as duration
FROM input TIMESTAMP BY Time
WHERE
    Event = 'end'
  
```

Correct Answer:

Answer Area

```

SELECT
    [user],
    feature,
    DATEADD(
        DATEDIFF(
            DATEPART(
                second,
                ISFIRST
                LAST
                TOPONE
            )
        )
    ) OVER (PARTITION BY [user], feature LIMIT DURATION(hour, 1) WHEN Event = 'start'),
    (Time) OVER (PARTITION BY [user], feature LIMIT DURATION(hour, 1) WHEN Event = 'start') -
    (Time) OVER (PARTITION BY [user], feature LIMIT DURATION(hour, 1) WHEN Event = 'end') -
    (Time) as duration
FROM input TIMESTAMP BY Time
WHERE
    Event = 'end'
  
```

Explanation:

Correct Answer

Box 1: DATEDIFF -

DATEDIFF function returns the count (as a signed integer value) of the specified datepart boundaries crossed between the specified startdate and enddate.

Syntax: DATEDIFF (datepart , startdate, enddate)

Box 2: LAST -

The LAST function can be used to retrieve the last event within a specific condition. In this example, the condition is an event of type Start, partitioning the search by PARTITION BY user and feature. This way, every user and feature is treated independently when searching for the Start event. LIMIT DURATION limits the search back in time to 1 hour between the End and Start events.

Example:

```
SELECT -  
[user],  
feature,  
DATEDIFF(  
second,  
LAST(Time) OVER (PARTITION BY [user], feature LIMIT DURATION(hour,  
1) WHEN Event = 'start'),
```

Time) as duration -

FROM input TIMESTAMP BY Time -

WHERE -

Event = 'end'

Reference:

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-stream-analytics-query-patterns>

Question 86

CertyIQ

You are creating an Azure Data Factory data flow that will ingest data from a CSV file, cast columns to specified types of data, and insert the data into a table in an Azure Synapse Analytic dedicated SQL pool. The CSV file contains three columns named username, comment, and date.

The data flow already contains the following:

- Ⓐ A source transformation.
- Ⓐ A Derived Column transformation to set the appropriate types of data.
- Ⓐ A sink transformation to land the data in the pool.

You need to ensure that the data flow meets the following requirements:

- Ⓐ All valid rows must be written to the destination table.
- Ⓐ Truncation errors in the comment column must be avoided proactively.
- Ⓐ Any rows containing comment values that will cause truncation errors upon insert must be written to a file

in blob storage.

Which two actions should you perform? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. To the data flow, add a sink transformation to write the rows to a file in blob storage.
- B. To the data flow, add a Conditional Split transformation to separate the rows that will cause truncation errors.
- C. To the data flow, add a filter transformation to filter out rows that will cause truncation errors.
- D. Add a select transformation to select only the rows that will cause truncation errors.

Explanation:

Correct Answer : AB

B: Example:

1. This conditional split transformation defines the maximum length of "title" to be five. Any row that is less than or equal to five will go into the GoodRows stream.

Any row that is larger than five will go into the BadRows stream.

Conditional Split Settings

Output stream name * ErrorRows

Incoming stream * TypeCast

Split on First matching condition

Split condition

STREAM NAMES	COLUMN
GoodRows	length(title) <= 5
BadRows	Rows that do not meet any condition will use this output stream

2. This conditional split transformation defines the maximum length of "title" to be five. Any row that is less than or equal to five will go into the GoodRows stream.

Any row that is larger than five will go into the BadRows stream.

A:

3. Now we need to log the rows that failed. Add a sink transformation to the BadRows stream for logging. Here, we'll "auto-map" all of the fields so that we have logging of the complete transaction record. This is a text-delimited CSV file output to a single file in Blob Storage. We'll call the log file "badrows.csv".

Sink

Settings

Mapping

Optimize

Inspect

Data Preview

Clear the folder

Add dynamic content [Alt+P]

File name option *

Default

Pattern

Per partition

As data in column

Output to single file

badrows.csv

Quote All

Conditionally distributing the data in title groups, based on columns [1]

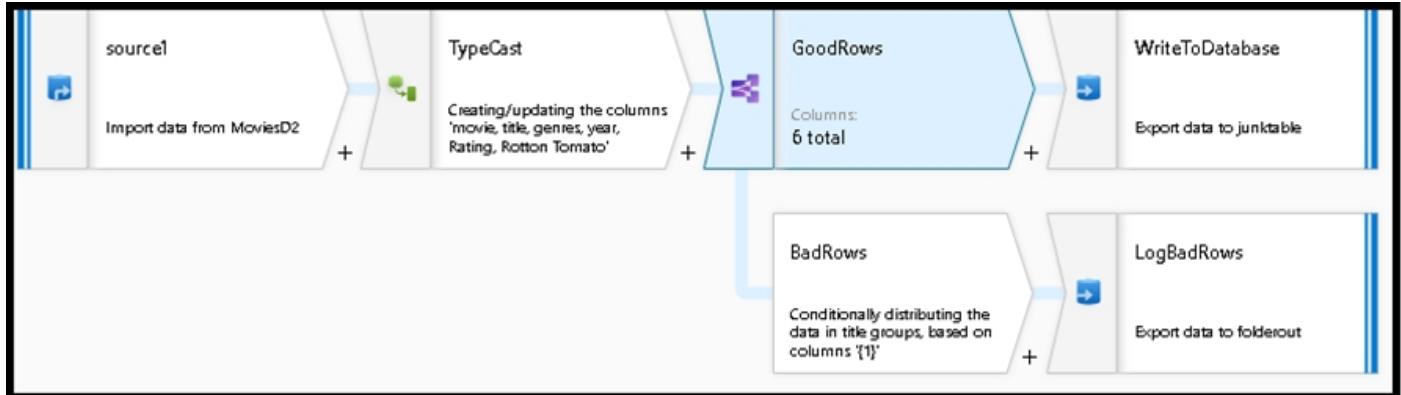
LogBadRows

Columns: 6 total

4. The completed data flow is shown below. We are now able to split off error rows to avoid the SQL truncation errors

and put those entries into a log file.

Meanwhile, successful rows can continue to write to our target database.



Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/how-to-data-flow-error-rows>

Question 87

CertyIQ

DRAG DROP -

You need to create an Azure Data Factory pipeline to process data for the following three departments at your company: Ecommerce, retail, and wholesale. The solution must ensure that data can also be processed for the entire company. How should you complete the Data Factory data flow script? To answer, drag the appropriate values to the correct targets. Each value 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:

Values

```
all, ecommerce, retail, wholesale
dept=='ecommerce', dept=='retail',
dept=='wholesale'
dept=='ecommerce', dept==
'wholesale', dept=='retail'
disjoint: false
disjoint: true
ecommerce, retail, wholesale, all
```

Answer Area

```
CleanData
split(
    [REDACTED]
    [REDACTED]
) ~> SplitByDept@([REDACTED])
```

Values

```
all, ecommerce, retail, wholesale  
dept=='ecommerce', dept=='retail'  
dept=='wholesale'  
dept=='ecommerce', dept=='  
'wholesale', dept=='retail'  
disjoint: false  
disjoint: true  
ecommerce, retail, wholesale, all
```

Answer Area

```
CleanData  
    split(  
        dept=='ecommerce', dept=='retail'  
        dept=='wholesale'  
            disjoint: true  
    ) ~> SplitByDept@(  
        ecommerce, retail, wholesale, all  
)
```

Explanation:

Correct Answer The conditional split transformation routes data rows to different streams based on matching conditions.

The conditional split transformation is similar to a CASE decision structure in a programming language. The transformation evaluates expressions, and based on the results, directs the data row to the specified stream.

Box 1: dept=='ecommerce', dept=='retail', dept=='wholesale'

First we put the condition. The order must match the stream labeling we define in Box 3.

Syntax:

```
<incomingStream>  
split(  
<conditionalExpression1>  
<conditionalExpression2>  
...  
disjoint: {true | false}  
) ~> <splitTx>@(stream1, stream2, ..., <defaultStream>)
```

"disjoint" should be True, so that data can be sent to all matching conditions. In this way the "all" output can get the data from every department, which ensures that "data can also be processed by the entire company".

Box 3: ecommerce, retail, wholesale, all

Label the streams -

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/data-flow-conditional-split>

Question 88

CertyIQ

DRAG DROP -

You have an Azure Data Lake Storage Gen2 account that contains a JSON file for customers. The file contains two attributes named FirstName and LastName.

You need to copy the data from the JSON file to an Azure Synapse Analytics table by using Azure Databricks. A new column must be created that concatenates the FirstName and LastName values.

You create the following components:

- ☞ A destination table in Azure Synapse
- ☞ An Azure Blob storage container

» A service principal

Which five actions should you perform in sequence next in a Databricks notebook? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

Select and Place:

Actions

- Mount the Data Lake Storage onto DBFS.
- Write the results to a table in Azure Synapse.
- Perform transformations on the file.
- Specify a temporary folder to stage the data.
- Write the results to Data Lake Storage.
- Read the file into a data frame.
- Drop the data frame.
- Perform transformations on the data frame.

Answer Area

Actions

- Mount the Data Lake Storage onto DBFS.
- Read the file into a data frame.
- Perform transformations on the data frame.
- Specify a temporary folder to stage the data.
- Write the results to a table in Azure Synapse.

Answer Area

Correct Answer:

Explanation:

Correct Answer

Step 1: Mount the Data Lake Storage onto DBFS

Begin with creating a file system in the Azure Data Lake Storage Gen2 account.

Step 2: Read the file into a data frame.

You can load the json files as a data frame in Azure Databricks.

Step 3: Perform transformations on the data frame.

Step 4: Specify a temporary folder to stage the data

Specify a temporary folder to use while moving data between Azure Databricks and Azure Synapse.

Step 5: Write the results to a table in Azure Synapse.

You upload the transformed data frame into Azure Synapse. You use the Azure Synapse connector for Azure Databricks to directly upload a dataframe as a table in a Azure Synapse.

Reference:

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

HOTSPOT -

You build an Azure Data Factory pipeline to move data from an Azure Data Lake Storage Gen2 container to a database in an Azure Synapse Analytics dedicated SQL pool.

Data in the container is stored in the following folder structure.

/in/{YYYY}/{MM}/{DD}/{HH}/{mm}

The earliest folder is /in/2021/01/01/00/00. The latest folder is /in/2021/01/15/01/45.

You need to configure a pipeline trigger to meet the following requirements:

- ⇒ Existing data must be loaded.
- ⇒ Data must be loaded every 30 minutes.
- ⇒ Late-arriving data of up to two minutes must be included in the load for the time at which the data should have arrived.

How should you configure the pipeline trigger? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Type:

Event
On-demand
Schedule
Tumbling window

Additional properties:

Prefix: /in/, Event: Blob created
Recurrence: 30 minutes, Start time: 2021-01-01T00:00
Recurrence: 30 minutes, Start time: 2021-01-01T00:00, Delay: 2 minutes
Recurrence: 32 minutes, Start time: 2021-01-15T01:45

Correct Answer:

Answer Area

Type:

Event
On-demand
Schedule
Tumbling window

Additional properties:

Prefix: /in/, Event: Blob created
Recurrence: 30 minutes, Start time: 2021-01-01T00:00
Recurrence: 30 minutes, Start time: 2021-01-01T00:00, Delay: 2 minutes
Recurrence: 32 minutes, Start time: 2021-01-15T01:45

Explanation:

Correct Answer

Box 1: Tumbling window -

To be able to use the Delay parameter we select Tumbling window.

Box 2:

Recurrence: 30 minutes, not 32 minutes

Delay: 2 minutes.

The amount of time to delay the start of data processing for the window. The pipeline run is started after the expected execution time plus the amount of delay.

The delay defines how long the trigger waits past the due time before triggering a new run. The delay doesn't alter the window startTime.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/how-to-create-tumbling-window-trigger>

Question 90

CertyIQ

HOTSPOT -

You are designing a near real-time dashboard solution that will visualize streaming data from remote sensors that connect to the internet. The streaming data must be aggregated to show the average value of each 10-second interval. The data will be discarded after being displayed in the dashboard.

The solution will use Azure Stream Analytics and must meet the following requirements:

- ⇒ Minimize latency from an Azure Event hub to the dashboard.
- ⇒ Minimize the required storage.
- ⇒ Minimize development effort.

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

NOTE: Each correct selection is worth one point

Hot Area:

Answer Area

Azure Stream Analytics input type:

Azure Event Hub
Azure SQL Database
Azure Stream Analytics
Microsoft Power BI

Azure Stream Analytics output type:

Azure Event Hub
Azure SQL Database
Azure Stream Analytics
Microsoft Power BI

Aggregation query location:

Azure Event Hub
Azure SQL Database
Azure Stream Analytics
Microsoft Power BI

Answer Area

Azure Stream Analytics input type:

Azure Event Hub
Azure SQL Database
Azure Stream Analytics
Microsoft Power BI

Azure Stream Analytics output type:

Correct Answer:

Azure Event Hub
Azure SQL Database
Azure Stream Analytics
Microsoft Power BI

Aggregation query location:

Azure Event Hub
Azure SQL Database
Azure Stream Analytics
Microsoft Power BI

Explanation:

Correct Answer

Reference:

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

Question 91

CertyIQ

DRAG DROP -

You have an Azure Stream Analytics job that is a Stream Analytics project solution in Microsoft Visual Studio. The job accepts data generated by IoT devices in the JSON format.

You need to modify the job to accept data generated by the IoT devices in the Protobuf format.

Which three actions should you perform from Visual Studio on sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

Select and Place:

Actions

- Change the Event Serialization Format to Protobuf in the input.json file of the job and reference the DLL.
- Add an Azure Stream Analytics Custom Deserializer Project (.NET) project to the solution.
- Add .NET deserializer code for Protobuf to the custom deserializer project.
- Add .NET deserializer code for Protobuf to the Stream Analytics project.
- Add an Azure Stream Analytics Application project to the solution.

Answer Area

Correct Answer:

Actions

- Change the Event Serialization Format to Protobuf in the input.json file of the job and reference the DLL.
- Add an Azure Stream Analytics Custom Deserializer Project (.NET) project to the solution.
- Add .NET deserializer code for Protobuf to the custom deserializer project.
- Add .NET deserializer code for Protobuf to the Stream Analytics project.
- Add an Azure Stream Analytics Application project to the solution.

Answer Area

- Add an Azure Stream Analytics Custom Deserializer Project (.NET) project to the solution.
- Add .NET deserializer code for Protobuf to the custom deserializer project.
- Change the Event Serialization Format to Protobuf in the input.json file of the job and reference the DLL.

Explanation:

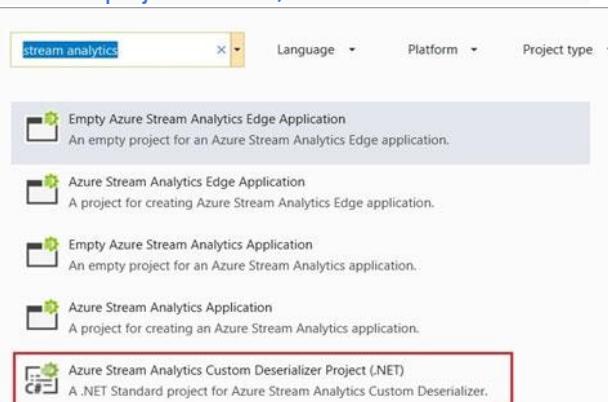
Correct Answer

Step 1: Add an Azure Stream Analytics Custom Deserializer Project (.NET) project to the solution.

Create a custom deserializer -

1. Open Visual Studio and select File > New > Project. Search for Stream Analytics and select Azure Stream Analytics Custom Deserializer Project (.NET). Give the project a name, like Protobuf Deserializer.

Create a new project



2. In Solution Explorer, right-click your Protobuf Deserializer project and select Manage NuGet Packages from the menu. Then install the Microsoft.Azure.StreamAnalytics and Google.Protobuf NuGet packages.
3. Add the MessageBodyProto class and the MessageBodyDeserializer class to your project.
4. Build the Protobuf Deserializer project.

Step 2: Add .NET deserializer code for Protobuf to the custom deserializer project

Azure Stream Analytics has built-in support for three data formats: JSON, CSV, and Avro. With custom .NET deserializers, you can read data from other formats such as Proto

I Buffer, Bond and other user defined formats for both cloud and edge jobs.

Absolutely: <https://docs.microsoft.com/en-us/azure/stream-analytics/custom-deserializer>

Question 92

CertyIQ

You have an Azure Storage account and a data warehouse in Azure Synapse Analytics in the UK South region. You need to copy blob data from the storage account to the data warehouse by using Azure Data Factory. The solution must meet the following requirements:

- ⇒ Ensure that the data remains in the UK South region at all times.
- ⇒ Minimize administrative effort.

Which type of integration runtime should you use?

- A. Azure integration runtime
- B. Azure-SSIS integration runtime
- C. Self-hosted integration runtime

Explanation:

Correct Answer A

IR type	Public network	Private network
Azure	Data Flow Data movement Activity dispatch	
Self-hosted	Data movement Activity dispatch	Data movement Activity dispatch
Azure-SSIS	SSIS package execution	SSIS package execution

Incorrect Answers:

C: Self-hosted integration runtime is to be used On-premises.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/concepts-integration-runtime>

Question 93

CertyIQ

HOTSPOT -

You have an Azure SQL database named Database1 and two Azure event hubs named HubA and HubB. The data consumed from each source is shown in the following table.

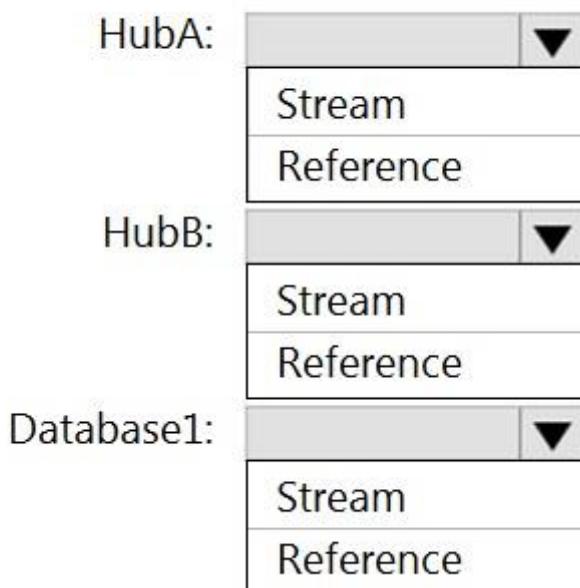
Source	Data
Database1	Driver's name Driver's license number
HubA	Ride route Ride distance Ride duration
HubB	Ride fare Ride payment

You need to implement Azure Stream Analytics to calculate the average fare per mile by driver.

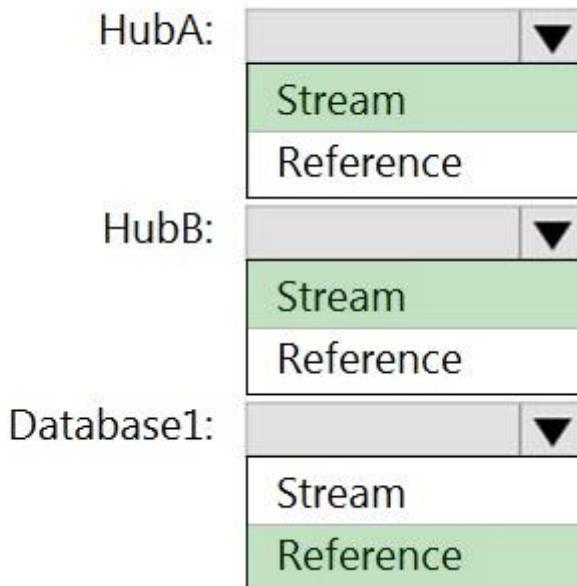
How should you configure the Stream Analytics input for each source? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Answer Area



Explanation:

Correct Answer

HubA: Stream -

HubB: Stream -

Database1: Reference -

Reference data (also known as a lookup table) is a finite data set that is static or slowly changing in nature, used to perform a lookup or to augment your data streams. For example, in an IoT scenario, you could store metadata about sensors (which don't change often) in reference data and join it with real time IoT data streams. Azure Stream Analytics loads reference data in memory to achieve low latency stream processing

Reference:

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-use-reference-data>

Question 94

CertyIQ

You have an Azure Stream Analytics job that receives clickstream data from an Azure event hub. You need to define a query in the Stream Analytics job. The query must meet the following requirements:

- ☞ Count the number of clicks within each 10-second window based on the country of a visitor.
- ☞ Ensure that each click is NOT counted more than once.

How should you define the Query?

- A. SELECT Country, Avg(*) AS Average FROM ClickStream TIMESTAMP BY CreatedAt GROUP BY Country, SlidingWindow(second, 10)
- B. SELECT Country, Count(*) AS Count FROM ClickStream TIMESTAMP BY CreatedAt GROUP BY Country, TumblingWindow(second, 10)
- C. SELECT Country, Avg(*) AS Average FROM ClickStream TIMESTAMP BY CreatedAt GROUP BY Country, HoppingWindow(second, 10, 2)

- D. SELECT Country, Count(*) AS Count FROM ClickStream TIMESTAMP BY CreatedAt GROUP BY Country, SessionWindow(second, 5, 10)

Explanation:

Correct Answer B

tumbling window functions are used to segment a data stream into distinct time segments and perform a function against them, such as the example below. The key differentiators of a Tumbling window are that they repeat, do not overlap, and an event cannot belong to more than one tumbling window.

Example:

Incorrect Answers:

A: Sliding windows, unlike Tumbling or Hopping windows, output events only for points in time when the content of the window actually changes. In other words, when an event enters or exits the window. Every window has at least one event, like in the case of Hopping windows, events can belong to more than one sliding window.

C: Hopping window functions hop forward in time by a fixed period. It may be easy to think of them as Tumbling windows that can overlap, so events can belong to more than one Hopping window result set. To make a Hopping window the same as a Tumbling window, specify the hop size to be the same as the window size.

D: Session windows group events that arrive at similar times, filtering out periods of time where there is no data.

Reference:

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-window-functions>

Question 95

CertyIQ

HOTSPOT -

You are building an Azure Analytics query that will receive input data from Azure IoT Hub and write the results to Azure Blob storage.

You need to calculate the difference in the number of readings per sensor per hour.

How should you complete the query? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

```
SELECT sensorId,
       growth = reading -
               ▼ (reading) OVER (PARTITION BY sensorId ▼ (hour,1))
               □ LAG
               □ LAST
               □ LEAD
               □ LIMIT DURATION
               □ OFFSET
               □ WHEN
FROM input
```

Answer Area

```
SELECT sensorId,
       growth = reading -
               ▼ (reading) OVER (PARTITION BY sensorId ▼ (hour,1))
               □ LAG
               □ LAST
               □ LEAD
               □ LIMIT DURATION
               □ OFFSET
               □ WHEN
FROM input
```

Explanation:

Correct Answer

Box 1: LAG -

The LAG analytic operator allows one to look up a previous event in an event stream, within certain constraints. It is very useful for computing the rate of growth of a variable, detecting when a variable crosses a threshold, or when a condition starts or stops being true.

Box 2: LIMIT DURATION -

Example: Compute the rate of growth, per sensor:

```
SELECT sensorId,  
growth = reading -  
LAG(reading) OVER (PARTITION BY sensorId LIMIT DURATION(hour, 1))
```

FROM input -

Reference:

<https://docs.microsoft.com/en-us/stream-analytics-query/lag-azure-stream-analytics>

Question 96

CertyIQ

You need to schedule an Azure Data Factory pipeline to execute when a new file arrives in an Azure Data Lake Storage Gen2 container.

Which type of trigger should you use?

- A. on-demand
- B. tumbling window
- C. schedule
- D. event

Explanation:

Correct Answer D

Event-driven architecture (EDA) is a common data integration pattern that involves production, detection, consumption, and reaction to events. Data integration scenarios often require Data Factory customers to trigger pipelines based on events happening in storage account, such as the arrival or deletion of a file in Azure

Blob Storage account.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/how-to-create-event-trigger>

Question 97

CertyIQ

You are developing a solution that will stream to Azure Stream Analytics. The solution will have both streaming data and reference data.

Which input type should you use for the reference data?

- A. Azure Cosmos DB
- **B. Azure Blob storage**
- C. Azure IoT Hub
- D. Azure Event Hubs

Explanation:

Correct Answer: *B*

Stream Analytics supports Azure Blob storage and Azure SQL Database as the storage layer for Reference Data.

Reference:

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-use-reference-data>

Question 98

CertyIQ

You are designing an Azure Stream Analytics job to process incoming events from sensors in retail environments.

You need to process the events to produce a running average of shopper counts during the previous 15 minutes, calculated at five-minute intervals.

Which type of window should you use?

- A. snapshot
- B. tumbling
- **C. hopping**
- D. sliding

Explanation:

Correct Answer: *C*

Unlike tumbling windows, hopping windows model scheduled overlapping windows. A hopping window specification consist of three parameters: the timeunit, the windowsize (how long each window lasts) and the hopsize (by how much each window moves forward relative to the previous one).

Reference:

<https://docs.microsoft.com/en-us/stream-analytics-query/hopping-window-azure-stream-analytics>

Question 99

CertyIQ

HOTSPOT -

You are designing a monitoring solution for a fleet of 500 vehicles. Each vehicle has a GPS tracking device that sends data to an Azure event hub once per minute.

You have a CSV file in an Azure Data Lake Storage Gen2 container. The file maintains the expected geographical area in which each vehicle should be.

You need to ensure that when a GPS position is outside the expected area, a message is added to another event hub for processing within 30 seconds. The solution must minimize cost.

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Service:

- An Azure Synapse Analytics Apache Spark pool
- An Azure Synapse Analytics serverless SQL pool
- Azure Data Factory
- Azure Stream Analytics

Window:

- Hopping
- No window
- Session
- Tumbling

Analysis type:

- Event pattern matching
- Lagged record comparison
- Point within polygon
- Polygon overlap

Service:

- An Azure Synapse Analytics Apache Spark pool
- An Azure Synapse Analytics serverless SQL pool
- Azure Data Factory
- Azure Stream Analytics

Window:

- Hopping
- No window
- Session
- Tumbling

Analysis type:

- Event pattern matching
- Lagged record comparison
- Point within polygon
- Polygon overlap

Explanation:

Correct Answer:

Box 1: Azure Stream Analytics -

Box 2: NO Window

Box 3: Point within polygon -

Reference:

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-window-functions>

Question 100

CertyIQ

You are designing an Azure Databricks table. The table will ingest an average of 20 million streaming events per day.

You need to persist the events in the table for use in incremental load pipeline jobs in Azure Databricks. The solution must minimize storage costs and incremental load times.

What should you include in the solution?

- A. Partition by DateTime fields.
- B. Sink to Azure Queue storage.
- C. Include a watermark column.
- D. Use a JSON format for physical data storage.

Explanation:

Correct Answer: *B*

The Databricks ABS-AQS connector uses Azure Queue Storage (AQS) to provide an optimized file source that lets you find new files written to an Azure Blob storage (ABS) container without repeatedly listing all of the files. This provides two major advantages:

- ⇒ Lower latency: no need to list nested directory structures on ABS, which is slow and resource intensive.
- ⇒ Lower costs: no more costly LIST API requests made to ABS.

Reference:

<https://docs.microsoft.com/en-us/azure/databricks/spark/latest/structured-streaming/aqs>

Question 101

CertyIQ

HOTSPOT -

You have a self-hosted integration runtime in Azure Data Factory.

The current status of the integration runtime has the following configurations:

- ⇒ Status: Running
- ⇒ Type: Self-Hosted
- ⇒ Version: 4.4.7292.1
- ⇒ Running / Registered Node(s): 1/1
- ⇒ High Availability Enabled: False
- ⇒ Linked Count: 0
- ⇒ Queue Length: 0
- ⇒ Average Queue Duration: 0.00s

The integration runtime has the following node details:

- ⇒ Name: X-M
- ⇒ Status: Running
- ⇒ Version: 4.4.7292.1

Available Memory: 7697MB
CPU Utilization: 6%
Network (In/Out): 1.21KBps/0.83KBps
Concurrent Jobs (Running/Limit): 2/14
Role: Dispatcher/Worker
Credential Status: In Sync

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

If the X-M node becomes unavailable, all executed pipelines will:

fail until the node comes back online
switch to another integration runtime
exceed the CPU limit

The number of concurrent jobs and the CPU usage indicate that the Concurrent Jobs (Running/Limit) value should be:

raised
lowered
left as is

Answer Area

If the X-M node becomes unavailable, all executed pipelines will:

fail until the node comes back online
switch to another integration runtime
exceed the CPU limit

Correct Answer:

The number of concurrent jobs and the CPU usage indicate that the Concurrent Jobs (Running/Limit) value should be:

raised
lowered
left as is

Explanation:

Correct Answer:

Box 1: fail until the node comes back online

We see: High Availability Enabled: False

Note: Higher availability of the self-hosted integration runtime so that it's no longer the single point of failure in your big data solution or cloud data integration with Data Factory.

Box 2: lowered -

We see:

Concurrent Jobs (Running/Limit): 2/14

CPU Utilization: 6%

Note: When the processor and available RAM aren't well utilized, but the execution of concurrent jobs reaches a node's limits, scale up by increasing the number of concurrent jobs that a node can run

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/create-self-hosted-integration-runtime>

Question 102

CertyIQ

You have an Azure Databricks workspace named workspace1 in the Standard pricing tier.

You need to configure workspace1 to support autoscaling all-purpose clusters. The solution must meet the following requirements:

- ☞ Automatically scale down workers when the cluster is underutilized for three minutes.
- ☞ Minimize the time it takes to scale to the maximum number of workers.
- ☞ Minimize costs.

What should you do first?

- A. Enable container services for workspace1.
- B. Upgrade workspace1 to the Premium pricing tier.
- C. Set Cluster Mode to High Concurrency.
- D. Create a cluster policy in workspace1.

Explanation:

Correct Answer: B

For clusters running Databricks Runtime 6.4 and above, optimized autoscaling is used by all-purpose clusters in the Premium plan

Optimized autoscaling:

Scales up from min to max in 2 steps.

Can scale down even if the cluster is not idle by looking at shuffle file state.

Scales down based on a percentage of current nodes.

On job clusters, scales down if the cluster is underutilized over the last 40 seconds.

On all-purpose clusters, scales down if the cluster is underutilized over the last 150 seconds.

The spark.databricks.aggressiveWindowDownS Spark configuration property specifies in seconds how often a cluster makes down-scaling decisions. Increasing the value causes a cluster to scale down more slowly. The maximum value is 600.

Note: Standard autoscaling -

Starts with adding 8 nodes. Thereafter, scales up exponentially, but can take many steps to reach the max. You can customize the first step by setting the spark.databricks.autoscaling.standardFirstStepUp Spark configuration property.

Scales down only when the cluster is completely idle and it has been underutilized for the last 10 minutes.

Scales down exponentially, starting with 1 node.

Reference:

<https://docs.databricks.com/clusters/configure.html>

Question 103

CertyIQ

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 Stream Analytics solution that will analyze Twitter data.

You need to count the tweets in each 10-second window. The solution must ensure that each tweet is counted only once.

Solution: You use a tumbling window, and you set the window size to 10 seconds.

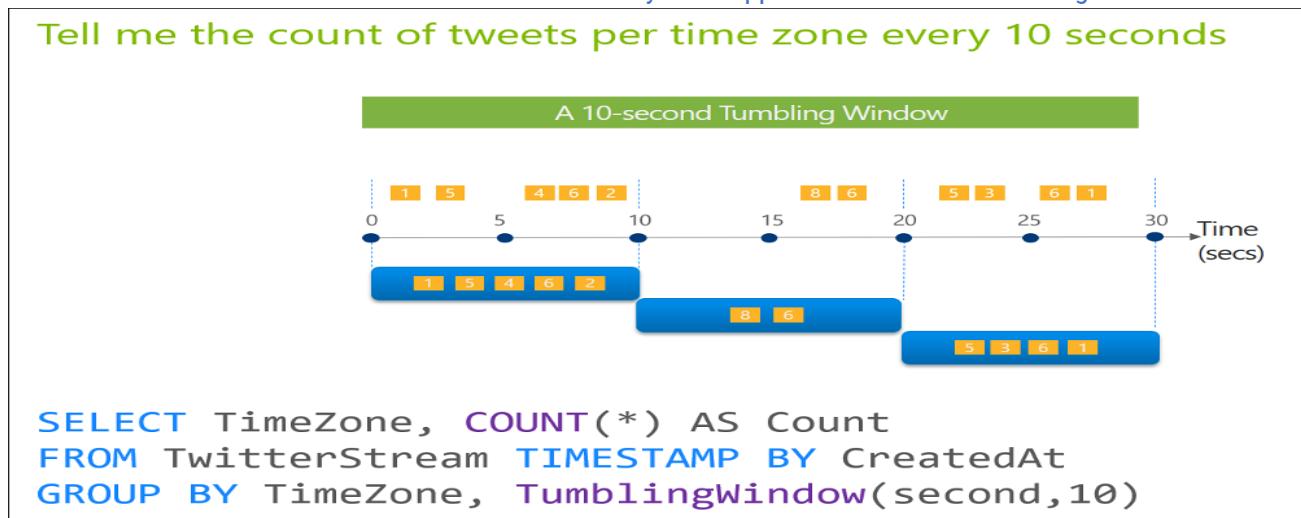
Does this meet the goal?

- A. Yes
- B. No

Explanation:

Correct Answer: :A

Tumbling windows are a series of fixed-sized, non-overlapping and contiguous time intervals. The following diagram illustrates a stream with a series of events and how they are mapped into 10-second tumbling windows.



Reference:

<https://docs.microsoft.com/en-us/stream-analytics-query/tumbling-window-azure-stream-analytics>

Question 104

CertyIQ

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 Stream Analytics solution that will analyze Twitter data.

You need to count the tweets in each 10-second window. The solution must ensure that each tweet is counted only once.

Solution: You use a session window that uses a timeout size of 10 seconds.
Does this meet the goal?

- A. Yes
- B. No

Explanation:

Correct Answer: : *B*

Instead use a tumbling window. Tumbling windows are a series of fixed-sized, non-overlapping and contiguous time intervals.

Reference:

<https://docs.microsoft.com/en-us/stream-analytics-query/tumbling-window-azure-stream-analytics>

Question 105

CertyIQ

You use Azure Stream Analytics to receive data from Azure Event Hubs and to output the data to an Azure Blob Storage account.

You need to output the count of records received from the last five minutes every minute.

Which windowing function should you use?

- A. Session
- B. Tumbling
- C. Sliding
- D. Hopping

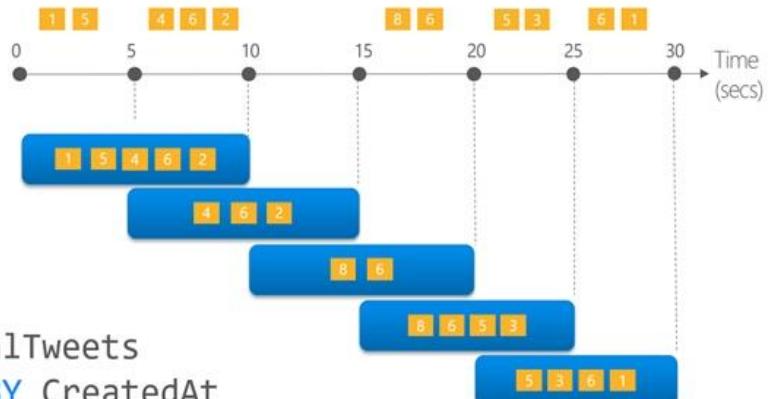
Explanation:

Correct Answer: *D*

Hopping window functions hop forward in time by a fixed period. It may be easy to think of them as Tumbling windows that can overlap and be emitted more often than the window size. Events can belong to more than one Hopping window result set. To make a Hopping window the same as a Tumbling window, specify the hop size to be the same as the window size.

Every 5 seconds give me the count of Tweets over the last 10 seconds

A 10-second Hopping Window with a 5-second "Hop"



```
SELECT Topic, COUNT(*) AS TotalTweets  
FROM TwitterStream TIMESTAMP BY CreatedAt  
GROUP BY Topic, HoppingWindow(second, 10 , 5)
```

Reference:

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-window-functions>

Question 106

CertyIQ

HOTSPOT -

You configure version control for an Azure Data Factory instance as shown in the following exhibit.

The screenshot shows the Azure Data Factory version control configuration page. On the left, there's a sidebar with icons for Home, Pipelines, Triggers, Global parameters, Security, Customer managed key, and Managed private endpoints. The main area is titled 'Git repository' and contains the following settings:

Setting	Value
Repository type	Azure DevOps Git
Azure DevOps Account	CONTOSO
Project name	Data
Repository name	dwh_batchetl
Collaboration branch	main
Publish branch	adf_publish
Root folder	/

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

Azure Resource Manager (ARM) templates for the pipeline assets are stored in [answer choice]

/
adf_publish
main
Parameterization template

A Data Factory Azure Resource Manager (ARM) template named contososales can be found in [answer choice]

/
/contososales
/dwh_batchetl/adf_publish/contososales
/main

Correct Answer:

Answer Area

Azure Resource Manager (ARM) templates for the pipeline assets are stored in [answer choice]

/
adf_publish
main
Parameterization template

A Data Factory Azure Resource Manager (ARM) template named contososales can be found in [answer choice]

/
/contososales
/dwh_batchetl/adf_publish/contososales
/main

Explanation:

Correct Answer:

Box 1: adf_publish -

The Publish branch is the branch in your repository where publishing related ARM templates are stored and updated. By default, it's adf_publish.

Box 2: / dwh_batchetl/adf_publish/contososales

Note: RepositoryName (here dwh_batchetl): Your Azure Repos code repository name. Azure Repos projects contain Git repositories to manage your source code as your project grows. You can create a new repository or use an existing repository that's already in your project.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/source-control>

Question 107

CertyIQ

HOTSPOT -

You are designing an Azure Stream Analytics solution that receives instant messaging data from an Azure Event Hub.

You need to ensure that the output from the Stream Analytics job counts the number of messages per time

zone every 15 seconds.

How should you complete the Stream Analytics query? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Select TimeZone, count (*) AS MessageCount

FROM MessageStream

	▼
LAST	
OVER	
SYSTEM.TIMESTAMP()	
TIMESTAMP BY	

CreatedAt

GROUP BY TimeZone,

	▼
HOPPINGWINDOW	
SESSIONWINDOW	
SLIDINGWINDOW	
TUMBLINGWINDOW	

(second, 15)

Answer Area

Select TimeZone, count (*) AS MessageCount

FROM MessageStream

	▼
LAST	
OVER	
SYSTEM.TIMESTAMP()	
TIMESTAMP BY	

CreatedAt

GROUP BY TimeZone,

	▼
HOPPINGWINDOW	
SESSIONWINDOW	
SLIDINGWINDOW	
TUMBLINGWINDOW	

(second, 15)

Explanation:

Correct Answer:

Box 1: timestamp by -

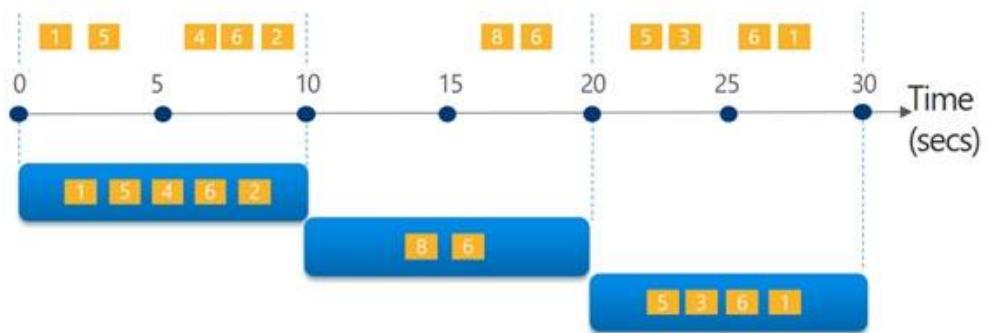
Box 2: TUMBLINGWINDOW -

Tumbling window functions are used to segment a data stream into distinct time segments and perform a function

against them, such as the example below. The key differentiators of a Tumbling window are that they repeat, do not overlap, and an event cannot belong to more than one tumbling window.

Tell me the count of Tweets per time zone every 10 seconds

A 10-second Tumbling Window



```
SELECT TimeZone, COUNT(*) AS Count
FROM TwitterStream TIMESTAMP BY CreatedAt
GROUP BY TimeZone, TumblingWindow(second,10)
```

Reference:

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-window-functions>

Question 108

CertyIQ

HOTSPOT -

You have an Azure Data Factory instance named ADF1 and two Azure Synapse Analytics workspaces named WS1 and WS2.

ADF1 contains the following pipelines:

- ☞ P1: Uses a copy activity to copy data from a nonpartitioned table in a dedicated SQL pool of WS1 to an Azure Data Lake Storage Gen2 account
- ☞ P2: Uses a copy activity to copy data from text-delimited files in an Azure Data Lake Storage Gen2 account to a nonpartitioned table in a dedicated SQL pool of WS2

You need to configure P1 and P2 to maximize parallelism and performance.

Which dataset settings should you configure for the copy activity if each pipeline? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

P1:

- Set the Copy method to Bulk insert
- Set the Copy method to PolyBase
- Set the Isolation level to Repeatable read
- Set the Partition option to Dynamic range

P2:

- Set the Copy method to Bulk insert
- Set the Copy method to PolyBase
- Set the Isolation level to Repeatable read
- Set the Partition option to Dynamic range

Answer Area

P1:

- Set the Copy method to Bulk insert
- Set the Copy method to PolyBase
- Set the Isolation level to Repeatable read
- Set the Partition option to Dynamic range

Suggested Answer:

P2:

- Set the Copy method to Bulk insert
- Set the Copy method to PolyBase
- Set the Isolation level to Repeatable read
- Set the Partition option to Dynamic range

Explanation:

Correct Answer: P1: Set the partition option to "Dynamic range "

P2: PolyBase

regarding to P1

<https://docs.microsoft.com/en-us/azure/data-factory/connector-azure-sql-data-warehouse?tabs=data-factory#parallel-copy-from-synapse-analytics>

Scenario: "Full load from large table, without physical partitions.." ->

Suggested settings: Partition options: Dynamic range partition.

Question 109

CertyIQ

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	

Answer Area

Load methodology:

Full Load
Incremental Load
Load individual files as they arrive

Correct Answer:

Trigger:

Fixed schedule
New file
Tumbling window

Explanation:

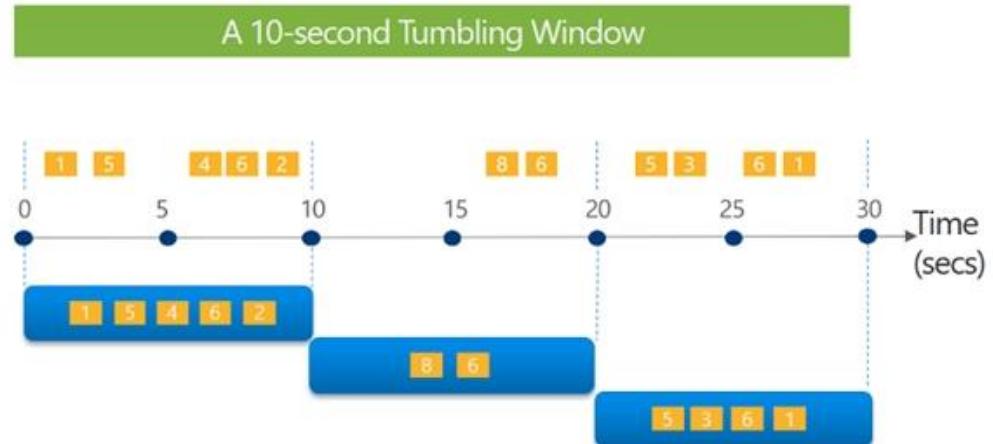
Correct Answer:

Box 1: Incremental load -

Box 2: Tumbling window -

Tumbling windows are a series of fixed-sized, non-overlapping and contiguous time intervals. The following diagram illustrates a stream with a series of events and how they are mapped into 10-second tumbling windows.

Tell me the count of tweets per time zone every 10 seconds



```
SELECT TimeZone, COUNT(*) AS Count
FROM TwitterStream TIMESTAMP BY CreatedAt
GROUP BY TimeZone, TumblingWindow(second,10)
```

Reference:

<https://docs.microsoft.com/en-us/stream-analytics-query/tumbling-window-azure-stream-analytics>

Question 110

CertyIQ

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 create an Azure Databricks workspace that has a tiered structure. The workspace will contain the following three workloads:

- Ⓐ A workload for data engineers who will use Python and SQL.
- Ⓑ A workload for jobs that will run notebooks that use Python, Scala, and SQL.
- Ⓒ A workload that data scientists will use to perform ad hoc analysis in Scala and R.

The enterprise architecture team at your company identifies the following standards for Databricks environments:

- Ⓓ The data engineers must share a cluster.
- Ⓔ The job cluster will be managed by using a request process whereby data scientists and data engineers provide packaged notebooks for deployment to the cluster.
- Ⓕ All the data scientists must be assigned their own cluster that terminates automatically after 120 minutes of inactivity. Currently, there are three data scientists.

You need to create the Databricks clusters for the workloads.

Solution: You create a Standard cluster for each data scientist, a Standard cluster for the data engineers, and a High Concurrency cluster for the jobs.

Does this meet the goal?

- A. Yes
- B. No

Explanation:

Correct Answer: *B*

We need a High Concurrency cluster for the data engineers and the jobs.

Note: Standard clusters are recommended for a single user. Standard can run workloads developed in any language: Python, R, Scala, and SQL.

A high concurrency cluster is a managed cloud resource. The key benefits of high concurrency clusters are that they provide Apache Spark-native fine-grained sharing for maximum resource utilization and minimum query latencies.

Reference:

<https://docs.azuredatabricks.net/clusters/configure.html>

Question 111

CertyIQ

You have the following Azure Data Factory pipelines:

- Ⓐ Ingest Data from System1
- Ⓑ Ingest Data from System2
- Ⓒ Populate Dimensions
- Ⓓ Populate Facts

Ingest Data from System1 and Ingest Data from System2 have no dependencies. Populate Dimensions must execute after Ingest Data from System1 and Ingest Data from System2. Populate Facts must execute after Populate Dimensions pipeline. All the pipelines must execute every eight hours.

What should you do to schedule the pipelines for execution?

- A. Add an event trigger to all four pipelines.
- B. Add a schedule trigger to all four pipelines.
- C. Create a patient pipeline that contains the four pipelines and use a schedule trigger.
- D. Create a patient pipeline that contains the four pipelines and use an event trigger.

Explanation:

Correct Answer: : C

Schedule trigger: A trigger that invokes a pipeline on a wall-clock schedule.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipeline-execution-triggers>

Question 112

CertyIQ

DRAG DROP -

You are responsible for providing access to an Azure Data Lake Storage Gen2 account.

Your user account has contributor access to the storage account, and you have the application ID and access key.

You plan to use PolyBase to load data into an enterprise data warehouse in Azure Synapse Analytics.

You need to configure PolyBase to connect the data warehouse to storage account.

Which three components should you create in sequence? To answer, move the appropriate components from the list of components to the answer area and arrange them in the correct order.

Select and Place:

Components

Answer Area

a database scoped credential
an asymmetric key
an external data source
a database encryption key
an external file format



Components

a database scoped credential

an asymmetric key

an external data source

a database encryption key

an external file format

Answer Area

a database scoped credential

an external data source

an external file format



Explanation:

Correct Answer: 1.- A database scoped credential

2.- an External data sorce

3.- a external file format

Question 113

CertyIQ

You are monitoring an Azure Stream Analytics job by using metrics in Azure.

You discover that during the last 12 hours, the average watermark delay is consistently greater than the configured late arrival tolerance.

What is a possible cause of this behavior?

- A. Events whose application timestamp is earlier than their arrival time by more than five minutes arrive as inputs.
- B. There are errors in the input data.
- C. The late arrival policy causes events to be dropped.
- D. The job lacks the resources to process the volume of incoming data

Explanation:

Correct Answer: : D

Watermark Delay indicates the delay of the streaming data processing job.

There are a number of resource constraints that can cause the streaming pipeline to slow down. The watermark delay metric can rise due to:

1. Not enough processing resources in Stream Analytics to handle the volume of input events. To scale up resources, see Understand and adjust Streaming Units.
2. Not enough throughput within the input event brokers, so they are throttled. For possible solutions, see Automatically scale up Azure Event Hubs throughput units.
3. Output sinks are not provisioned with enough capacity, so they are throttled. The possible solutions vary widely based

on the flavor of output service being used.

Reference:

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-time-handling>

Question 114

CertyIQ

HOTSPOT -

You are building an Azure Stream Analytics job to retrieve game data.

You need to ensure that the job returns the highest scoring record for each five-minute time interval of each game.

How should you complete the Stream Analytics query? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

SELECT

Collect(Score)	▼
CollectTop(1) OVER(ORDER BY Score Desc)	▼
Game, MAX(Score)	▼
TopOne() OVER(PARTITION BY Game ORDER BY Score Desc)	▼

as HighestScore

FROM input TIMESTAMP BY CreatedAt

GROUP BY

Game	▼
Hopping(minute,5)	▼
Tumbling(minute,5)	▼
Windows(TumblingWindow(minute,5),Hopping(minute,5))	▼

Correct Answer:

Answer Area



SELECT

Collect(Score)	▼
CollectTop(1) OVER(ORDER BY Score Desc)	▼
Game, MAX(Score)	▼
TopOne() OVER(PARTITION BY Game ORDER BY Score Desc)	▼

as HighestScore

FROM input TIMESTAMP BY CreatedAt

GROUP BY

Game	▼
Hopping(minute,5)	▼
Tumbling(minute,5)	▼
Windows(TumblingWindow(minute,5),Hopping(minute,5))	▼

Explanation:

Correct Answer: TopOne() / Tumbling

Box 1: TopOne OVER(PARTITION BY Game ORDER BY Score Desc)

TopOne returns the top-rank record, where rank defines the ranking position of the event in the window according to the specified ordering. Ordering/ranking is based on event columns and can be specified in ORDER BY clause.

BOX 2: TUMBLING

Reference:

<https://docs.microsoft.com/en-us/stream-analytics-query/topone-azure-stream-analytics>

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-window-functions>

Question 115

CertyIQ

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

Explanation:

Correct Answer: No

you cannot execute the R script using a stored procedure activity

Synapse doesn't support R at the moment

<https://docs.microsoft.com/en-us/answers/questions/222624/is-azure-synapse-analytics-supporting-r-language.html>

Question 116

CertyIQ

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 create an Azure Databricks workspace that has a tiered structure. The workspace will contain the following three workloads:

- ⌚ A workload for data engineers who will use Python and SQL.
- ⌚ A workload for jobs that will run notebooks that use Python, Scala, and SQL.
- ⌚ A workload that data scientists will use to perform ad hoc analysis in Scala and R.

The enterprise architecture team at your company identifies the following standards for Databricks environments:

- ⌚ The data engineers must share a cluster.
- ⌚ The job cluster will be managed by using a request process whereby data scientists and data engineers provide packaged notebooks for deployment to the cluster.
- ⌚ All the data scientists must be assigned their own cluster that terminates automatically after 120 minutes of inactivity. Currently, there are three data scientists.

You need to create the Databricks clusters for the workloads.

Solution: You create a High Concurrency cluster for each data scientist, a High Concurrency cluster for the data engineers, and a Standard cluster for the jobs.

Does this meet the goal?

A. Yes

B. No

Explanation:

Correct Answer: NO - as High concurrency not support Scala
data scientist need scala so standard, jobs need scala so standard, so B but for different reasons.
<https://docs.microsoft.com/en-us/azure/databricks/clusters/configure>

Question 117

CertyIQ

You are designing an Azure Databricks cluster that runs user-defined local processes.

You need to recommend a cluster configuration that meets the following requirements:

- ⇒ Minimize query latency.
- ⇒ Maximize the number of users that can run queries on the cluster at the same time.
- ⇒ Reduce overall costs without compromising other requirements.

Which cluster type should you recommend?

A. Standard with Auto Termination

B. High Concurrency with Autoscaling

C. High Concurrency with Auto Termination

D. Standard with Autoscaling

Explanation:

Correct Answer: B

High concurrency cluster cannot terminated, so C is wrong.

Standard cluster cannot shared by multiple tasks, so A and D are wrong.

A High Concurrency cluster is a managed cloud resource. The key benefits of High Concurrency clusters are that they provide fine-grained sharing for maximum resource utilization and minimum query latencies. Databricks chooses the appropriate number of workers required to run your job. This is referred to as autoscaling. Autoscaling makes it easier to achieve high cluster utilization, because you don't need to provision the cluster to match a workload.

Incorrect Answers:

C: The cluster configuration includes an auto terminate setting whose default value depends on cluster mode: Standard and Single Node clusters terminate automatically after 120 minutes by default.

High Concurrency clusters do not terminate automatically by default.

Reference:

<https://docs.microsoft.com/en-us/azure/databricks/clusters/configure>

Question 118

CertyIQ

HOTSPOT -

You are building an Azure Data Factory solution to process data received from Azure Event Hubs, and then ingested into an Azure Data Lake Storage Gen2 container.

The data will be ingested every five minutes from devices into JSON files. The files have the following naming pattern.

`/{deviceType}/in/{YYYY}/{MM}/{DD}/{HH}/{deviceId}_{YYYY}{MM}{DD}{HH}{mm}.json`

You need to prepare the data for batch data processing so that there is one dataset per hour per deviceType. The solution must minimize read times.

How should you configure the sink for the copy activity? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Parameter:

- `@pipeline(),TriggerTime`
- `@pipeline(),TriggerType`
- `@trigger().outputs.windowStartTime`
- `@trigger().startTime`

Naming pattern:

- `/{deviceId}/out/{YYYY}/{MM}/{DD}/{HH}.json`
- `/{YYYY}/{MM}/{DD}/{deviceType}.json`
- `/{YYYY}/{MM}/{DD}/{HH}.json`
- `/{YYYY}/{MM}/{DD}/{HH}_{deviceType}.json`

Copy behavior:

- `Add dynamic content`
- `Flatten hierarchy`
- `Merge files`

Answer Area

Suggested Answer:

Parameter:

@pipeline(),TriggerTime
@pipeline(),TriggerType
@trigger().outputs.windowStartTime
@trigger().startTime

Naming pattern:

/{deviceID}/out/{YYYY}/{MM}/{DD}/{HH}.json
/{YYYY}/{MM}/{DD}/{deviceType}.json
/{YYYY}/{MM}/{DD}/{HH}.json
/{YYYY}/{MM}/{DD}/{HH}_{deviceType}.json

Copy behavior:

Add dynamic content
Flatten hierarchy
Merge files

Explanation:

Correct Answer:

- 1) @trigger().outputs.windowStartTime - this output is from a tumbling window trigger, and is required to identify the correct directory at the /{HH}/ level. Using windowStartTime will give the hour with complete data. The @trigger().startTime is for a schedule trigger, which corresponds to the hour for which data has not arrived yet.
- 2) /{YYYY}/{MM}/{DD}/{HH}_{deviceType}.json is the naming pattern to achieve an hourly dataset for each device type.
- 3) Multiple files for each device type will exist on the source side, since the naming pattern starts with {deviceID}... so the files must be merged in the sink to create a single file per device type.

Question 119

CertyIQ

DRAG DROP -

You are designing an Azure Data Lake Storage Gen2 structure for telemetry data from 25 million devices distributed across seven key geographical regions. Each minute, the devices will send a JSON payload of metrics to Azure Event Hubs.

You need to recommend a folder structure for the data. The solution must meet the following requirements:

- ⇒ Data engineers from each region must be able to build their own pipelines for the data of their respective region only.

⦿ The data must be processed at least once every 15 minutes for inclusion in Azure Synapse Analytics serverless SQL pools.

How should you recommend completing the structure? To answer, drag the appropriate values to the correct targets. Each value 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:

Values	Answer Area
{deviceID}	/ <input type="text"/> Value / <input type="text"/> Value / <input type="text"/> Value .json
{mm}/{HH}/{DD}/{MM}/{YYYY}	
{regionID}/{deviceID}	
{regionID}/raw	
{YYYY}/{MM}/{DD}/{HH}	
{YYYY}/{MM}/{DD}/{HH}/{mm}	
raw/{deviceID}	
raw/{regionID}	

Correct Answer:

Values	Answer Area
{deviceID}	/ <input type="text"/> raw/{regionID} / <input type="text"/> {YYYY}/{MM}/{DD}/{HH}/{mm} / <input type="text"/> {deviceID} .json
{mm}/{HH}/{DD}/{MM}/{YYYY}	
{regionID}/{deviceID}	
{regionID}/raw	
{YYYY}/{MM}/{DD}/{HH}	
{YYYY}/{MM}/{DD}/{HH}/{mm}	
raw/{deviceID}	
raw/{regionID}	

Explanation:

Correct Answer:

{raw/regionID}/{YYYY}/{MM}/{DD}/{HH}/{mm}/{deviceID}.json

{raw/regionID} is the first level because raw is the container name for the raw data. RegionID follows it for ease of managing security.

{YYYY}/{MM}/{DD}/{HH}/{mm}/{deviceID}.json instead of {deviceID}/{YYYY}/{MM}/{DD}/{HH}/{mm}.json. The primary reason is that you want your namespace structure to have as few folders as high up and narrow those down as you get deeper into your structure.

For example, if you have 1 year worth of data and 25 million devices, using {YYYY}/{MM}/{DD}/{HH}/{mm}/{deviceID} results in 2.1 million folders (1 year * 12 months * 30 days [estimate] * 24 hours * 60 minutes). If you start your folder structure with {deviceID}, you end up with 25 million folders - one for each device - before you even get to including the date in the hierarchy.

Box 1: {raw/regionID}

Box 2: {YYYY}/{MM}/{DD}/{HH}/{mm}

Box 3: {deviceID}

Reference:

<https://github.com/paolosalvatori/StreamAnalyticsAzureDataLakeStore/blob/master/README.md>

Question 120

CertyIQ

HOTSPOT -

You are implementing an Azure Stream Analytics solution to process event data from devices. The devices output events when there is a fault and emit a repeat of the event every five seconds until the fault is resolved. The devices output a heartbeat event every five seconds after a previous event if there are no faults present.

A sample of the events is shown in the following table.

DeviceID	EventType	EventTime
78cc5ht9-w357-684r-w4fr-kr16h6p9874e	HeartBeat	2020-12-01T19:00.000Z
78cc5ht9-w357-684r-w4fr-kr16h6p9874e	HeartBeat	2020-12-01T19:05.000Z
78cc5ht9-w357-684r-w4fr-kr16h6p9874e	TemperatureSensorFault	2020-12-01T19:07.000Z

You need to calculate the uptime between the faults.

How should you complete the Stream Analytics SQL query? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

```
SELECT  
DeviceID,  
MIN(EventTime) as StartTime,  
MAX(EventTime) as EndTime,  
DATEDIFF(second, MIN(EventTime), MAX(EventTime)) AS duration_in_seconds  
FROM input TIMESTAMP BY EventTime
```

▼
WHERE EventType='HeartBeat' WHERE LAG(EventType, 1) OVER (LIMIT DURATION(second,5)) <> EventType WHERE IsFirst(second,5) = 1

GROUP BY

DeviceID

▼
,SessionWindow(second, 5, 50000) OVER (PARTITION BY DeviceID) ,TumblingWindow(second,5) HAVING DATEDIFF(second, MIN(EventTime), MAX(EventTime)) > 5

Answer Area

```
SELECT  
DeviceID,  
MIN(EventTime) as StartTime,  
MAX(EventTime) as EndTime,  
DATEDIFF(second, MIN(EventTime), MAX(EventTime)) AS duration_in_seconds  
FROM input TIMESTAMP BY EventTime
```

▼
WHERE EventType='HeartBeat' WHERE LAG(EventType, 1) OVER (LIMIT DURATION(second,5)) <> EventType WHERE IsFirst(second,5) = 1

GROUP BY

DeviceID

▼
,SessionWindow(second, 5, 50000) OVER (PARTITION BY DeviceID) ,TumblingWindow(second,5) HAVING DATEDIFF(second, MIN(EventTime), MAX(EventTime)) > 5

Explanation:

Correct Answer:

WHERE EventType='HeartBeat'

Session window.

If we want to calculate the uptime between the faults, we must use session window for each device, we know that will be receiving events for each 5 seconds if there is no error, so when an error occurs (or if we reach the maximum size of the window) then a new event will not be received within the next 5 seconds and the window will close, calculating the uptime. However if We use Tumbling window, it's not possible to calculate the uptime beyond 5 seconds

End of Part 3



Please find the videos of this AZ-900/AI-900/AZ-305/
AZ-104 /DP-900/ SC-900 and other
Microsoft/AWS/GOOGLE/COMPTIA exam series on

CertyIQ Official YouTube channel (FREE PDFs): -

Please [Subscribe](#) to CertyIQ YouTube Channel to get notified for latest exam dumps by clicking on the below image, it will redirect to the [CertyIQ](#) YouTube page.

Connect with us @ [LinkedIn](#) [Telegram](#)



Contact us for other dumps: - certyiq@gmail.com

For any other enquiry, please drop us a mail at
certyiqofficial@gmail.com

DP-203

Real Exam Questions & Answers



Latest Exam Questions
All Topics Covered-2023
Added New Que's



Get Certified Today!

New Course Covered



Subscribe

Question 121

CertyIQ

You are creating a new notebook in Azure Databricks that will support R as the primary language but will also support Scala and SQL.

Which switch should you use to switch between languages?

- A. %<language>
- B. @<Language >
- C. \\[<language >]
- D. \\(<language >)

Explanation:

Correct Answer: A

To change the language in Databricks' cells to either Scala, SQL, Python or R, prefix the cell with '%', followed by the language.

%python //or r, scala, sql

Reference:

<https://www.theta.co.nz/news-blogs/tech-blog/enhancing-digital-twins-part-3-predictive-maintenance-with-azure-databricks>

Question 122

CertyIQ

You have an Azure Data Factory pipeline that performs an incremental load of source data to an Azure Data Lake Storage Gen2 account.

Data to be loaded is identified by a column named LastUpdatedDate in the source table.

You plan to execute the pipeline every four hours.

You need to ensure that the pipeline execution meets the following requirements:

- ☞ Automatically retries the execution when the pipeline run fails due to concurrency or throttling limits.
- ☞ Supports backfilling existing data in the table.

Which type of trigger should you use?

- A. event
- B. on-demand
- C. schedule
- D. tumbling window**

Explanation:

Correct Answer: D

<https://www.sqlshack.com/how-to-schedule-azure-data-factory-pipeline-executions-using-triggers/>

Azure Data Factory pipeline executions using Triggers:

- Schedule Trigger: The schedule trigger is used to execute the Azure Data Factory pipelines on a wall-clock schedule.
- Tumbling Window Trigger: Can be used to process history data. Also can define Delay, Max concurrency, retry policy etc.
- Event-Based Triggers : The event-based trigger executes the pipelines in response to a blob-related event, such as creating or deleting a blob file, in an Azure Blob Storage

In case of pipeline failures, tumbling window trigger can retry the execution of the referenced pipeline automatically, using the same input parameters, without the user intervention. This can be specified using the property "retryPolicy" in the trigger definition.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/how-to-create-tumbling-window-trigger>

Question 123

CertyIQ

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. Specify a file naming pattern for the destination.
- B. Delete the files in the destination before loading the data.
- C. Filter by the last modified date of the source files.
- D. Delete the source files after they are copied.

Explanation:

Correct Answer: AC

there is no point about deletion in source and might be the case that the data should stay in source too.

Copy only the daily files by using filtering.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/connector-azure-data-lake-storage>

Question 124

CertyIQ

You plan to build a structured streaming solution in Azure Databricks. The solution will count new events in five-minute intervals and report only events that arrive during the interval. The output will be sent to a Delta Lake table.

Which output mode should you use?

- A. update
- B. complete
- C. append

Explanation:

Correct Answer: C

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.

A: 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.databricks.com/getting-started/spark/streaming.html>

Question 125

CertyIQ

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 Synapse Analytics dedicated SQL pool that contains a table named Table1.

You have files that are ingested and loaded into an Azure Data Lake Storage Gen2 container named container1.

You plan to insert data from the files in container1 into Table1 and transform the data. Each row of data in the files will produce one row in the serving layer of

Table1.

You need to ensure that when the source data files are loaded to container1, the DateTime is stored as an additional column in Table1.

Solution: In an Azure Synapse Analytics pipeline, you use a data flow that contains a Derived Column transformation.

Does this meet the goal?

A. Yes

B. No

Explanation:

Correct Answer: A

"Data flows are available both in Azure Data Factory and Azure Synapse Pipelines"

"Use the derived column transformation to generate new columns in your data flow or to modify existing fields."

<https://docs.microsoft.com/en-us/azure/data-factory/data-flow-derived-column>

Question 126

CertyIQ

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 Synapse Analytics dedicated SQL pool that contains a table named Table1.

You have files that are ingested and loaded into an Azure Data Lake Storage Gen2 container named container1.

You plan to insert data from the files in container1 into Table1 and transform the data. Each row of data in the files will produce one row in the serving layer of

Table1.

You need to ensure that when the source data files are loaded to container1, the DateTime is stored as an additional column in Table1.

Solution: You use a dedicated SQL pool to create an external table that has an additional DateTime column.

Does this meet the goal?

A. Yes

B. No

Explanation:

Correct Answer: NO

An external table is based on a source flat file structure. It seems to make no sense to add additional date time columns to such a table.

Instead use the derived column transformation to generate new columns in your data flow or to modify existing fields.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/data-flow-derived-column>

Question 127

CertyIQ

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 Synapse Analytics dedicated SQL pool that contains a table named Table1.

You have files that are ingested and loaded into an Azure Data Lake Storage Gen2 container named container1.

You plan to insert data from the files in container1 into Table1 and transform the data. Each row of data in the files will produce one row in the serving layer of

Table1.

You need to ensure that when the source data files are loaded to container1, the DateTime is stored as an additional column in Table1.

Solution: You use an Azure Synapse Analytics serverless SQL pool to create an external table that has an additional DateTime column.

Does this meet the goal?

A. Yes

B. No

Explanation:

Correct Answer: NO

Instead use the derived column transformation to generate new columns in your data flow or to modify existing fields.

The job is to insert data from the files in container1 into Table1 (in the dedicated sql pool) and transform the data after that and we need to add a new additional column.

External table are just references to the data, only metadata is really stored in the sql pool.

Hence anything including external table will be not a solution.

If you follow the different proposed solutions from previous questions, the most efficient solution is to use derived column transformation.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/data-flow-derived-column>

Question 128

CertyIQ

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 Synapse Analytics dedicated SQL pool that contains a table named Table1.

You have files that are ingested and loaded into an Azure Data Lake Storage Gen2 container named container1.

You plan to insert data from the files in container1 into Table1 and transform the data. Each row of data in the files will produce one row in the serving layer of

Table1.

You need to ensure that when the source data files are loaded to container1, the DateTime is stored as an additional column in Table1.

Solution: In an Azure Synapse Analytics pipeline, you use a Get Metadata activity that retrieves the DateTime of the files.

Does this meet the goal?

A. Yes

B. No

Explanation:

Correct Answer: NO

Instead use a serverless SQL pool to create an external table with the extra column.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/create-use-external-tables>

Question 129

CertyIQ

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

Explanation:

Correct Answer: YES

You can execute R code in a notebook, and then call it from Data Factory.

You can check it at "Databricks Notebook activity" header:

<https://docs.microsoft.com/en-US/azure/data-factory/transform-data>

And also:

<https://docs.microsoft.com/en-us/azure/databricks/spark/latest/sparkr/overview>

Question 130

CertyIQ

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 mapping data flow, and then inserts the data into the data warehouse.

Does this meet the goal?

A. Yes

B. No

Explanation:

Correct Answer: NO

There is no R in ADF dataflow

Mapping Dataflows can't execute R code that is a requirement, so not meet the goal.

Question 131

CertyIQ

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

Explanation:

Correct Answer: YES

You can execute R code in a notebook, and then call it from Data Factory.

You can check it at "Databricks Notebook activity" header:

<https://docs.microsoft.com/en-US/azure/data-factory/transform-data>

And also:

<https://docs.microsoft.com/en-us/azure/databricks/spark/latest/sparkr/overview>

Question 132

CertyIQ

You plan to create an Azure Data Factory pipeline that will include a mapping data flow.

You have JSON data containing objects that have nested arrays.

You need to transform the JSON-formatted data into a tabular dataset. The dataset must have one row for each item in the arrays.

Which transformation method should you use in the mapping data flow?

A. new branch

B. unpivot

C. alter row

D. flatten

Explanation:

Correct Answer: D

Use the flatten transformation to take array values inside hierarchical structures such as JSON and unroll them into individual rows. This process is known as denormalization.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/data-flow-flatten>

Question 133

CertyIQ

You use Azure Stream Analytics to receive Twitter data from Azure Event Hubs and to output the data to an Azure Blob storage account.

You need to output the count of tweets during the last five minutes every five minutes. Each tweet must only be counted once.

Which windowing function should you use?

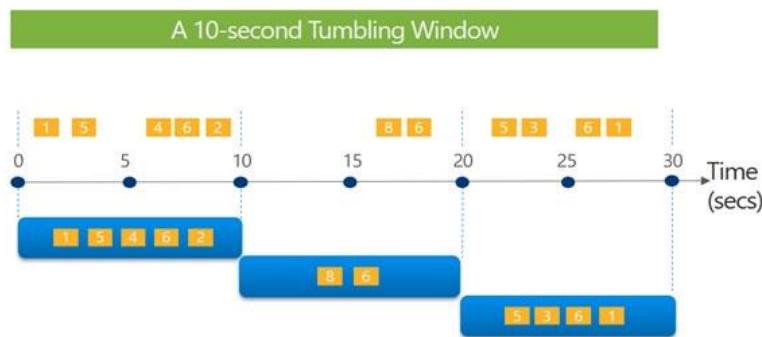
- A. a five-minute Sliding window
- B. a five-minute Session window
- C. a five-minute Hopping window that has a one-minute hop
- D. a five-minute Tumbling window**

Explanation:

Correct Answer: D

Tumbling window functions are used to segment a data stream into distinct time segments and perform a function against them, such as the example below. The key differentiators of a Tumbling window are that they repeat, do not overlap, and an event cannot belong to more than one tumbling window.

Tell me the count of Tweets per time zone every 10 seconds



```
SELECT TimeZone, COUNT(*) AS Count
FROM TwitterStream TIMESTAMP BY CreatedAt
GROUP BY TimeZone, TumblingWindow(second,10)
```

Reference:

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-window-functions>

Question 134

CertyIQ

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. Update
- B. Complete
- C. Append

Explanation:

Correct Answer: A

because " new rows will be added to adjust a sale" , that means that in the course of a day you must update de daily import with the new sales, the group by process generates new amounts, keep in mind that when it say "sales transactions will never be updated" its about the online store, not the aggregated rows.

By default, streams run in append mode, which adds new records to the table.

Incorrect Answers:

B: Complete mode: replace the entire table with every batch.

Reference:

<https://docs.databricks.com/delta/delta-streaming.html>

Question 135

CertyIQ

You have an enterprise data warehouse in Azure Synapse Analytics named DW1 on a server named Server1.

You need to determine the size of the transaction log file for each distribution of DW1.

What should you do?

- A. On DW1, execute a query against the sys.database_files dynamic management view.
- B. From Azure Monitor in the Azure portal, execute a query against the logs of DW1.
- C. Execute a query against the logs of DW1 by using the Get-AzOperationalInsightsSearchResult PowerShell cmdlet.
- D. On the master database, execute a query against the sys.dm_pdw_nodes_os_performance_counters dynamic management view.

Explanation:

Correct Answer: D

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-manage-monitor>

-- Transaction log size

SELECT

```
instance_name as distribution_db,  
cntr_value*1.0/1048576 as log_file_size_used_GB,  
pdw_node_id  
FROM sys.dm_pdw_nodes_os_performance_counters  
WHERE  
instance_name like 'Distribution_%'  
AND counter_name = 'Log File(s) Used Size (KB)'
```

Question 136

CertyIQ

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

Explanation:

Correct Answer: B

iOT is event hub of stream data so we need stream analytics for sure.

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/>

Question 137

CertyIQ

A company uses Azure Stream Analytics to monitor devices.

The company plans to double the number of devices that are monitored.

You need to monitor a Stream Analytics job to ensure that there are enough processing resources to handle the additional load.

Which metric should you monitor?

A. Early Input Events

B. Late Input Events

C. Watermark delay

D. Input Deserialization Errors

Explanation:

Correct Answer: C

There are a number of resource constraints that can cause the streaming pipeline to slow down. The watermark delay metric can rise due to:

Not enough processing resources in Stream Analytics to handle the volume of input events.

Not enough throughput within the input event brokers, so they are throttled.

Output sinks are not provisioned with enough capacity, so they are throttled. The possible solutions vary widely based on the flavor of output service being used.

Incorrect Answers:

A: Deserialization issues are caused when the input stream of your Stream Analytics job contains malformed messages.

Reference:

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-time-handling>

Question 138

CertyIQ

HOTSPOT -

You are designing an enterprise data warehouse in Azure Synapse Analytics that will store website traffic analytics 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

Answer Area

Distribution:

Hash
Round robin
Replicated

Index:

Clustered columnstore
Clustered
Nonclustered

Suggested Answer:

Explanation:

Correct Answer:

Box 1: Hash -

Consider using a hash-distributed table when:

The table size on disk is more than 2 GB.

The table has frequent insert, update, and delete operations.

Box 2: Clustered columnstore -

Clustered columnstore tables offer both the highest level of data compression and the best overall query performance.

Reference:

Question 139

CertyIQ

You have an Azure Stream Analytics job.

You need to ensure that the job has enough streaming units provisioned.

You configure monitoring of the SU % Utilization metric.

Which two additional metrics should you monitor? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

A. Backlogged Input Events

B. Watermark Delay

C. Function Events

D. Out of order Events

E. Late Input Events

Explanation:

Correct Answer: AB

To react to increased workloads and increase streaming units, consider setting an alert of 80% on the SU Utilization metric. Also, you can use watermark delay and backlogged events metrics to see if there is an impact.

Note: Backlogged Input Events: Number of input events that are backlogged. A non-zero value for this metric implies that your job isn't able to keep up with the number of incoming events. If this value is slowly increasing or consistently non-zero, you should scale out your job, by increasing the SUs.

Reference:

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-monitoring>

Question 140

CertyIQ

You have an activity in an Azure Data Factory pipeline. The activity calls a stored procedure in a data warehouse in Azure Synapse Analytics and runs daily.

You need to verify the duration of the activity when it ran last.

What should you use?

A. activity runs in Azure Monitor

B. Activity log in Azure Synapse Analytics

C. the sys.dm_pdw_wait_stats data management view in Azure Synapse Analytics

D. an Azure Resource Manager template

Explanation:

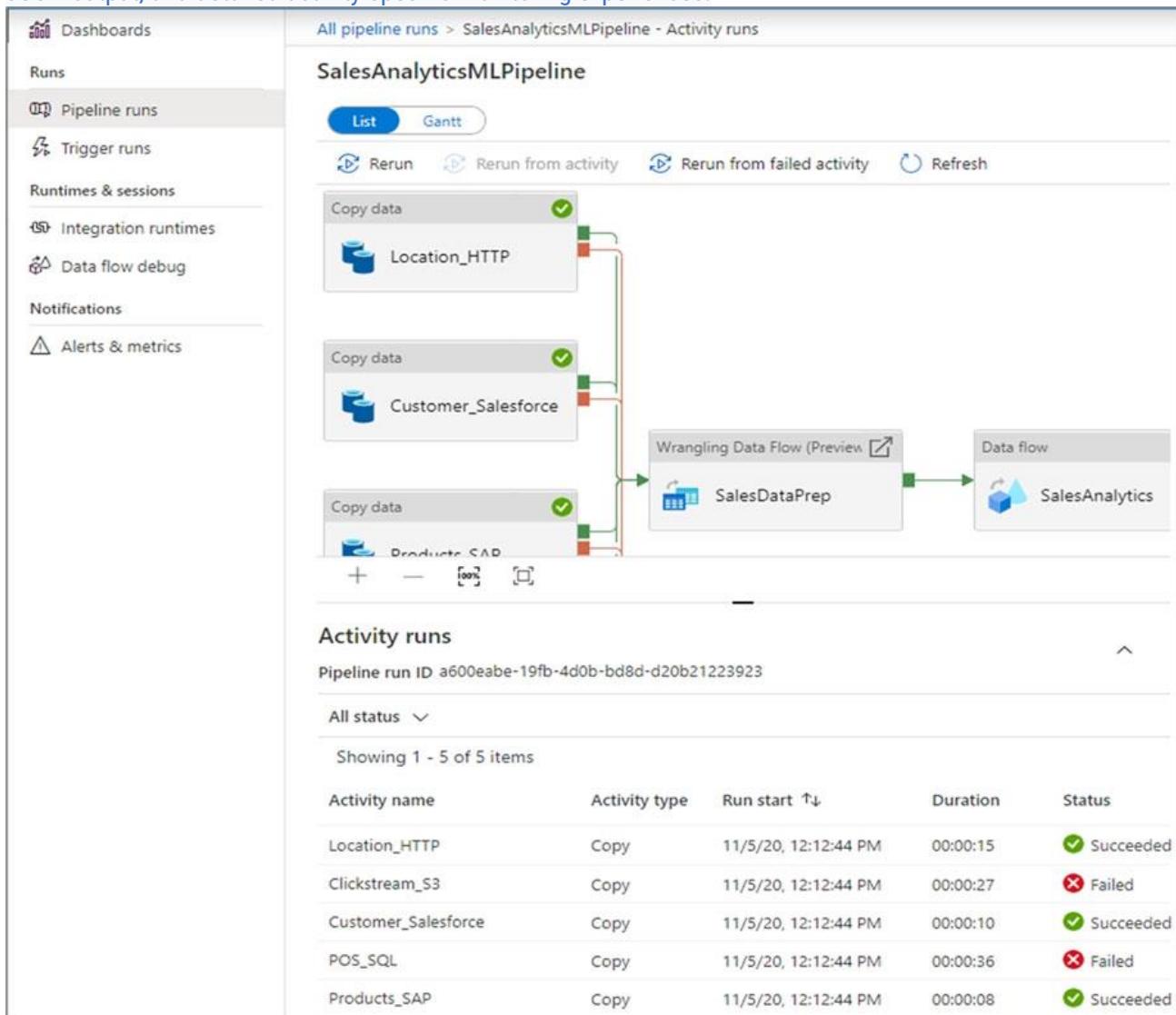
Correct Answer: A

Monitor activity runs. To get a detailed view of the individual activity runs of a specific pipeline run, click on the pipeline name.

Example:

Pipeline runs		
Triggered	Debug	Rerun Cancel Refresh Edit col
<input type="text"/> Search by run ID or name		Pacific Time (US & C... : Last 7 days)
Showing 1 - 21 items		S
<input type="checkbox"/> Pipeline name	Run start ↑	Run end
<input type="checkbox"/> S3ToDataLakeCopy	11/5/20, 6:00:18 AM	11/5/20, 6:03:18
<input type="checkbox"/> DatabricksJarPipeline	11/4/20, 6:04:11 PM	11/4/20, 6:10:18
<input type="checkbox"/> S3ToDataLakeCopy	11/4/20, 6:00:18 PM	11/4/20, 6:03:18
<input type="checkbox"/> S3ToDataLakeCopy	11/4/20, 6:00:19 AM	11/4/20, 6:04:18

The list view shows activity runs that correspond to each pipeline run. Hover over the specific activity run to get run-specific information such as the JSON input, JSON output, and detailed activity-specific monitoring experiences.



You can check the Duration.

Incorrect Answers:

C: sys.dm_pdw_wait_stats holds information related to the SQL Server OS state related to instances running on the different nodes.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/monitor-visually>

Question 141

CertyIQ

You have an Azure Data Factory pipeline that is triggered hourly.

The pipeline has had 100% success for the past seven days.

The pipeline execution fails, and two retries that occur 15 minutes apart also fail. The third failure returns the following error.

ErrorCode=UserErrorFileNotFoundException,'Type=Microsoft.DataTransfer.Common.Shared.HybridDeliveryException,Message=ADLS Gen2 operation failed for: Operation returned an invalid status code 'NotFound'. Account: 'contosoproduksouth'. Filesystem: wwi. Path: 'BIKES/CARBON/year=2021/month=01/day=10/hour=06'. ErrorCode: 'PathNotFoundException'. Message: 'The specified path does not exist.'. RequestId: '6d269b78-901f-001b-4924-e7a7bc000000'. TimeStamp: 'Sun, 10 Jan 2021 07:45:05'

What is a possible cause of the error?

- A. The parameter used to generate year=2021/month=01/day=10/hour=06 was incorrect.
- B. From 06:00 to 07:00 on January 10, 2021, there was no data in wwi/BIKES/CARBON.**
- C. From 06:00 to 07:00 on January 10, 2021, the file format of data in wwi/BIKES/CARBON was incorrect.
- D. The pipeline was triggered too early.

Explanation:

Correct Answer: B

The error message says a missing file, which matches with answer B: missing data from 06:00. The process had re-tried three times, 15 mins apart, which explains that the error was generated 07:45.

Question 142

CertyIQ

You have an Azure Synapse Analytics job that uses Scala.

You need to view the status of the job.

What should you do?

- A. From Synapse Studio, select the workspace. From Monitor, select SQL requests.
- B. From Azure Monitor, run a Kusto query against the AzureDiagnostics table.
- C. From Synapse Studio, select the workspace. From Monitor, select Apache Sparks applications.**
- D. From Azure Monitor, run a Kusto query against the SparkLoggingEvent_CL table.

Explanation:

Correct Answer: C

Use Synapse Studio to monitor your Apache Spark applications. To monitor running Apache Spark application Open Monitor, then select Apache Spark applications. To view the details about the Apache Spark applications that are running, select the submitting Apache Spark application and view the details. If the Apache Spark application is still running, you can monitor the progress.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/monitoring/apache-spark-applications>

Question 143

CertyIQ

DRAG DROP -

You have an Azure Data Lake Storage Gen2 account that contains a JSON file for customers. The file contains two attributes named FirstName and LastName.

You need to copy the data from the JSON file to an Azure Synapse Analytics table by using Azure Databricks. A new column must be created that concatenates the FirstName and LastName values.

You create the following components:

- Ⓐ A destination table in Azure Synapse
- Ⓑ An Azure Blob storage container
- Ⓒ A service principal

In which order should you perform the actions? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

Select and Place:

Actions

Answer Area

Mount the Data Lake Storage onto DBFS.

Write the results to a table in Azure Synapse.

Specify a temporary folder to stage the data.

Read the file into a data frame.

Perform transformations on the data frame.

Suggested Answer:

Actions

Answer Area

Mount the Data Lake Storage onto DBFS.

Mount the Data Lake Storage onto DBFS.

Write the results to a table in Azure Synapse.

Read the file into a data frame.

Specify a temporary folder to stage the data.

Perform transformations on the data frame.

Read the file into a data frame.

Specify a temporary folder to stage the data.

Perform transformations on the data frame.

Write the results to a table in Azure Synapse.

Explanation:

Correct Answer:

Step 1: Mount the Data Lake Storage onto DBFS

Begin with creating a file system in the Azure Data Lake Storage Gen2 account.

Step 2: Read the file into a data frame.

You can load the json files as a data frame in Azure Databricks.

Step 3: Perform transformations on the data frame.

Step 4: Specify a temporary folder to stage the data

Specify a temporary folder to use while moving data between Azure Databricks and Azure Synapse.

Step 5: Write the results to a table in Azure Synapse.

You upload the transformed data frame into Azure Synapse. You use the Azure Synapse connector for Azure Databricks to directly upload a dataframe as a table in a Azure Synapse.

Reference:

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

Question 144

CertyIQ

You have an Azure data factory named ADF1.

You currently publish all pipeline authoring changes directly to ADF1.

You need to implement version control for the changes made to pipeline artifacts. The solution must ensure that you can apply version control to the resources currently defined in the UX Authoring canvas for ADF1.

Which two actions should you perform? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. From the UX Authoring canvas, select Set up code repository..
- B. Create a Git repository.
- C. Create a GitHub action.
- D. Create an Azure Data Factory trigger.
- E. From the UX Authoring canvas, select Publish.
- F. From the UX Authoring canvas, run Publish All.

Explanation:

Correct Answer: AB

They are asking to "implement version control".

B Create Git repo

A From the UX Set up code repository

The documentation attached to this question states the first step is to set up a code repository from the UX and this question is around setting up version control, not saving your changes which is what F suggests

Question 145

CertyIQ

DRAG DROP -

You have an Azure subscription that contains an Azure Synapse Analytics workspace named workspace1. Workspace1 connects to an Azure DevOps repository named repo1. Repo1 contains a collaboration branch named main and a development branch named branch1. Branch1 contains an Azure Synapse pipeline named pipeline1.

In workspace1, you complete testing of pipeline1.

You need to schedule pipeline1 to run daily at 6 AM.

Which four actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

NOTE: More than one order of answer choices is correct. You will receive credit for any of the correct orders you select.

Select and Place:

Actions	Answer Area
Create a new branch in Repo1.	
Merge the changes from branch1 into main.	
Associate the schedule trigger with pipeline1.	 
Switch to Synapse live mode.	
Create a schedule trigger.	
Publish the contents of main.	

Suggested Answer:

Actions	Answer Area
Create a new branch in Repo1.	Create a schedule trigger.
Merge the changes from branch1 into main.	Associate the schedule trigger with pipeline1.
Associate the schedule trigger with pipeline1.	 
Switch to Synapse live mode.	Merge the changes from branch1 into main.
Create a schedule trigger.	Publish the contents of main.
Publish the contents of main.	

Explanation:

Correct Answer: 5-3-2-6

Question 146

CertyIQ

HOTSPOT -

You have an Azure subscription that contains an Azure Synapse Analytics dedicated SQL pool named Pool1 and an Azure Data Lake Storage account named storage1. Storage1 requires secure transfers.

You need to create an external data source in Pool1 that will be used to read .orc files in storage1.

How should you complete the code? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

```
CREATE EXTERNAL DATA SOURCE AzureDataLakeStore
```

WITH

```
( Location1 ` ://data@newyorktaxidataset.dfs.core.windows.net` ,  
    abfs  
    abfss  
    wasb  
    wasbs
```

```
credential = ADLS_credential ,
```

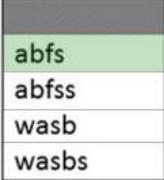
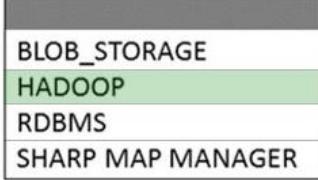
TYPE -

```
);
```

BLOB_STORAGE
HADOOP
RDBMS
SHARP MAP MANAGER

Suggested Answer:

Answer Area

```
CREATE EXTERNAL DATA SOURCE AzureDataLakeStore
WITH
( Location1 ` 
      ://data@newyorktaxidataset.dfs.core.windows.net' ,
  credential = ADLS_credential ,
  TYPE - 
) ;
```

Explanation:

Correct Answer:

abfss and Hadoop

Hint: Storage1 requires secure transfers --> The default option is to use enable secure SSL connections when provisioning Azure Data Lake Storage Gen2. When this is enabled, you must use abfss when a secure TLS/SSL connection is selected. abfss endpoint when your account has secure transfer enabled

Reference: <https://docs.microsoft.com/en-us/sql/t-sql/statements/create-external-data-source-transact-sql?view=azure-sqldw-latest&preserve-view=true&tabs=dedicated>

Question 147

CertyIQ

You have an Azure subscription that contains an Azure Synapse Analytics dedicated SQL pool named SQLPool1.

SQLPool1 is currently paused.

You need to restore the current state of SQLPool1 to a new SQL pool.

What should you do first?

- A. Create a workspace.
- B. Create a user-defined restore point.
- C. Resume SQLPool1.
- D. Create a new SQL pool.

Explanation:

Correct Answer: B

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-restore-active-paused-dw>

Question 148

CertyIQ

You are designing an Azure Synapse Analytics workspace.

You need to recommend a solution to provide double encryption of all the data at rest.

Which two components should you include in the recommendation? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

A. an X.509 certificate

B. an RSA key

C. an Azure virtual network that has a network security group (NSG)

D. an Azure Policy initiative

E. an Azure key vault that has purge protection enabled

Explanation:

Correct Answer: BE

Synapse workspaces encryption uses existing keys or new keys generated in Azure Key Vault. A single key is used to encrypt all the data in a workspace.

Synapse workspaces support RSA 2048 and 3072 byte-sized keys, and RSA-HSM keys.

The Key Vault itself needs to have purge protection enabled.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/security/workspaces-encryption>

Question 149

CertyIQ

You have an Azure Synapse Analytics serverless SQL pool named Pool1 and an Azure Data Lake Storage Gen2 account named storage1. The

AllowBlobPublicAccess property is disabled for storage1.

You need to create an external data source that can be used by Azure Active Directory (Azure AD) users to access storage from Pool1.

What should you create first?

- A. an external resource pool
- B. an external library
- C. database scoped credentials**
- D. a remote service binding

Explanation:

Correct Answer: C

Security -

User must have SELECT permission on an external table to read the data. External tables access underlying Azure storage using the database scoped credential defined in data source.

Note: A database scoped credential is a record that contains the authentication information that is required to connect to a resource outside SQL Server. Most credentials include a Windows user and password.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables>

<https://docs.microsoft.com/en-us/sql/t-sql/statements/create-database-scoped-credential-transact-sql>

Question 150

CertyIQ

You have an Azure Data Factory pipeline named Pipeline1. Pipeline1 contains a copy activity that sends data to an Azure Data Lake Storage Gen2 account.

Pipeline1 is executed by a schedule trigger.

You change the copy activity sink to a new storage account and merge the changes into the collaboration branch.

After Pipeline1 executes, you discover that data is NOT copied to the new storage account.

You need to ensure that the data is copied to the new storage account.

What should you do?

- A. Publish from the collaboration branch.**
- B. Create a pull request.
- C. Modify the schedule trigger.
- D. Configure the change feed of the new storage account.

Explanation:

Correct Answer:A

CI/CD lifecycle -

1. A development data factory is created and configured with Azure Repos Git. All developers should have permission to author Data Factory resources like pipelines and datasets.
2. A developer creates a feature branch to make a change. They debug their pipeline runs with their most recent changes

3. After a developer is satisfied with their changes, they create a pull request from their feature branch to the main or collaboration branch to get their changes reviewed by peers.

4. After a pull request is approved and changes are merged in the main branch, the changes get published to the development factory.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/continuous-integration-delivery>

Question 151

CertyIQ

You have an Azure Data Factory pipeline named pipeline1 that is invoked by a tumbling window trigger named Trigger1. Trigger1 has a recurrence of 60 minutes.

You need to ensure that pipeline1 will execute only if the previous execution completes successfully.

How should you configure the self-dependency for Trigger1?

A. offset: "-00:01:00" size: "00:01:00"

B. offset: "01:00:00" size: "-01:00:00"

C. offset: "01:00:00" size: "01:00:00"

D. offset: "-01:00:00" size: "01:00:00"

Explanation:

Correct Answer: D

Tumbling window self-dependency properties

In scenarios where the trigger shouldn't proceed to the next window until the preceding window is successfully completed, build a self-dependency. A self-dependency trigger that's dependent on the success of earlier runs of itself within the preceding hour will have the properties indicated in the following code.

Example code:

```
"name": "DemoSelfDependency",
"properties": {
  "runtimeState": "Started",
  "pipeline": {
    "pipelineReference": {
      "referenceName": "Demo",
      "type": "PipelineReference"
    }
  },
  "type": "TumblingWindowTrigger",
  "typeProperties": {
    "frequency": "Hour",
    "interval": 1,
```

```

"startTime": "2018-10-04T00:00:00Z",
"delay": "00:01:00",
"maxConcurrency": 50,
"retryPolicy": {
  "intervalInSeconds": 30
},
"dependsOn": [
{
  "type": "SelfDependencyTumblingWindowTriggerReference",
  "size": "01:00:00",
  "offset": "-01:00:00"
}
]
}
}
}
}

```

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/tumbling-window-trigger-dependency>

Question 152

CertyIQ

HOTSPOT -

You have an Azure Synapse Analytics pipeline named Pipeline1 that contains a data flow activity named Dataflow1.

Pipeline1 retrieves files from an Azure Data Lake Storage Gen 2 account named storage1.

Dataflow1 uses the AutoResolveIntegrationRuntime integration runtime configured with a core count of 128.

You need to optimize the number of cores used by Dataflow1 to accommodate the size of the files in storage1.

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

Hot Area:

Answer Area

To Pipeline1, add:

- A custom activity
- A Get Metadata activity
- An If Condition activity

For Dataflow1, set the core count by using:

- Dynamic content
- Parameters
- User properties

Suggested Answer:

Answer Area



To Pipeline1, add:

A custom activity
A Get Metadata activity
An If Condition activity

For Dataflow1, set the core count by using:

Dynamic content
Parameters
User properties

Explanation:

Correct Answer:

Box 1: A Get Metadata activity -

Dynamically size data flow compute at runtime

The Core Count and Compute Type properties can be set dynamically to adjust to the size of your incoming source data at runtime. Use pipeline activities like

Lookup or Get Metadata in order to find the size of the source dataset data. Then, use Add Dynamic Content in the Data Flow activity properties.

Box 2: Dynamic content -

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/control-flow-execute-data-flow-activity>

Question 153

CertyIQ

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 create an Azure Databricks workspace that has a tiered structure. The workspace will contain the following three workloads:

- ⦿ A workload for data engineers who will use Python and SQL.
- ⦿ A workload for jobs that will run notebooks that use Python, Scala, and SQL.
- ⦿ A workload that data scientists will use to perform ad hoc analysis in Scala and R.

The enterprise architecture team at your company identifies the following standards for Databricks environments:

- ⦿ The data engineers must share a cluster.
- ⦿ The job cluster will be managed by using a request process whereby data scientists and data engineers provide packaged notebooks for deployment to the cluster.

- ⦿ All the data scientists must be assigned their own cluster that terminates automatically after 120 minutes of inactivity. Currently, there are three data scientists.

You need to create the Databricks clusters for the workloads.

Solution: You create a Standard cluster for each data scientist, a High Concurrency cluster for the data engineers, and a Standard cluster for the jobs.

Does this meet the goal?

A. Yes

B. No

Explanation:

Correct Answer:

- data engineers: high concurrency cluster
- jobs: Standard cluster
- data scientists: Standard cluster

We would need a Standard cluster for the jobs to support Scala. High-concurrecy cluster does not support Scala.

Data Engineers - High Concurrency cluster as it provides for sharing . Also caters for SQL,Python and R.

Data Scientist - Standard Clusters which automatically terminates after 120 minutes and caters for Scala,SQL,Python and R.

JOBS- Standard Cluster

Question 154

CertyIQ

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 create an Azure Databricks workspace that has a tiered structure. The workspace will contain the following three workloads:

- ⦿ A workload for data engineers who will use Python and SQL.
- ⦿ A workload for jobs that will run notebooks that use Python, Scala, and SQL.
- ⦿ A workload that data scientists will use to perform ad hoc analysis in Scala and R.

The enterprise architecture team at your company identifies the following standards for Databricks environments:

- ⦿ The data engineers must share a cluster.
- ⦿ The job cluster will be managed by using a request process whereby data scientists and data engineers provide packaged notebooks for deployment to the cluster.

- ⇒ All the data scientists must be assigned their own cluster that terminates automatically after 120 minutes of inactivity. Currently, there are three data scientists.

You need to create the Databricks clusters for the workloads.

Solution: You create a Standard cluster for each data scientist, a High Concurrency cluster for the data engineers, and a High Concurrency cluster for the jobs.

Does this meet the goal?

A. Yes

B. No

Explanation:

Correct Answer: NO

High-concurrency clusters do not support Scala. So the answer is still 'No' but the reasoning is wrong.

Data scientist: STANDARD as need to run scala

Jobs: STANDARD as need to run scala

Data Engineers: High-concurrency clusters as better resource sharing

<https://docs.microsoft.com/en-us/azure/databricks/clusters/configure>

Question 155

CertyIQ

You are designing a folder structure for the files in an Azure Data Lake Storage Gen2 account. The account has one container that contains three years of data.

You need to recommend a folder structure that meets the following requirements:

- ⇒ Supports partition elimination for queries by Azure Synapse Analytics serverless SQL pools
- ⇒ Supports fast data retrieval for data from the current month
- ⇒ Simplifies data security management by department

Which folder structure should you recommend?

- A. \Department\DataSource\YYYY\MM\DataFile_YYYYMMDD.parquet**
- B. \DataSource\Department\YYYYMM\DataFile_YYYYMMDD.parquet
- C. \DD\MM\YYYY\Department\DataSource\DataFile_DDMMYY.parquet
- D. \YYYY\MM\DD\Department\DataSource\DataFile_YYYYMMDD.parquet

Explanation:

Correct Answer: A

Department top level in the hierarchy to simplify security management.

Month (MM) at the leaf/bottom level to support fast data retrieval for data from the current month.

You have an Azure subscription that contains an Azure Synapse Analytics dedicated SQL pool named Pool1. Pool1 receives new data once every 24 hours.

You have the following function.

```
create function dbo.udfFtoC(F decimal)
return decimal
as
begin
return (F - 32) * 5.0 / 9
end
```

You have the following query.

```
select avg_date, sensorid, avg_f, dbo.udfFtoC(avg_temperature) as avg_c from SensorTemps
where avg_date = @parameter
```

The query is executed once every 15 minutes and the @parameter value is set to the current date.

You need to minimize the time it takes for the query to return results.

Which two actions should you perform? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. Create an index on the avg_f column.
- B. Convert the avg_c column into a calculated column.**
- C. Create an index on the sensorid column.
- D. Enable result set caching.**
- E. Change the table distribution to replicate.

Explanation:

Correct Answer: AB

D: When result set caching is enabled, dedicated SQL pool automatically caches query results in the user database for repetitive use. This allows subsequent query executions to get results directly from the persisted cache so recomputation is not needed. Result set caching improves query performance and reduces compute resource usage. In addition, queries using cached results set do not use any concurrency slots and thus do not count against existing concurrency limits.

Incorrect:

Not A, not C: No joins so index not helpful.

Not E: What is a replicated table?

A replicated table has a full copy of the table accessible on each Compute node. Replicating a table removes the need to transfer data among Compute nodes before a join or aggregation. Since the table has multiple copies, replicated tables work best when the table size is less than 2 GB compressed. 2 GB is not a hard limit. If the data is static and does not change, you can replicate larger tables.

Reference:

Question 157

CertyIQ

DRAG DROP -

You have an Azure Active Directory (Azure AD) tenant that contains a security group named Group1. You have an Azure Synapse Analytics dedicated SQL pool named dw1 that contains a schema named schema1. You need to grant Group1 read-only permissions to all the tables and views in schema1. The solution must use the principle of least privilege.

Which three actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

NOTE: More than one order of answer choices is correct. You will receive credit for any of the correct orders you select.

Select and Place:

Actions

Answer Area

Create a database role named Role1 and grant Role1 SELECT permissions to schema1.

Create a database role named Role1 and grant Role1 SELECT permissions to dw1.

Assign the Azure role-based access control (Azure RBAC) Reader role for dw1 to Group1.

Create a database user in dw1 that represents Group1 and uses the FROM EXTERNAL PROVIDER clause.

Assign Role1 to the Group1 database user.

Suggested Answer:

Actions

Answer Area

Create a database role named Role1 and grant Role1 SELECT permissions to schema1.

Create a database user in dw1 that represents Group1 and uses the FROM EXTERNAL PROVIDER clause.

Create a database role named Role1 and grant Role1 SELECT permissions to dw1.

Create a database role named Role1 and grant Role1 SELECT permissions to schema1.

Assign the Azure role-based access control (Azure RBAC) Reader role for dw1 to Group1.

Assign Role1 to the Group1 database user.

Create a database user in dw1 that represents Group1 and uses the FROM EXTERNAL PROVIDER clause.

Assign Role1 to the Group1 database user.

Explanation:

Correct Answer:

Step 1: Create a database user named dw1 that represents Group1 and use the FROM EXTERNAL PROVIDER clause.

Step 2: Create a database role named Role1 and grant Role1 SELECT permissions to schema1.

Step 3: Assign Role1 to the Group1 database user.

Reference:

<https://docs.microsoft.com/en-us/azure/data-share/how-to-share-from-sql>

Question 158

CertyIQ

HOTSPOT -

You have an Azure subscription that contains a logical Microsoft SQL server named Server1. Server1 hosts an Azure Synapse Analytics SQL dedicated pool named Pool1.

You need to recommend a Transparent Data Encryption (TDE) solution for Server1. The solution must meet the following requirements:

☞ Track the usage of encryption keys.

Maintain the access of client apps to Pool1 in the event of an Azure datacenter outage that affects the availability of the encryption keys.

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

To track encryption key usage:

Always Encrypted
TDE with customer-managed keys
TDE with platform-managed keys

To maintain client app access in the event of a datacenter outage:

Create and configure Azure key vaults in two Azure regions.
Enable Advanced Data Security on Server1.
Implement the client apps by using a Microsoft .NET Framework data provider.

Suggested Answer:

Answer Area

To track encryption key usage:

Always Encrypted
TDE with customer-managed keys
TDE with platform-managed keys

To maintain client app access in the event of a datacenter outage:

Create and configure Azure key vaults in two Azure regions.
Enable Advanced Data Security on Server1.
Implement the client apps by using a Microsoft .NET Framework data provider.

Explanation:

Correct Answer:

Box 1: TDE with customer-managed keys

Customer-managed keys are stored in the Azure Key Vault. You can monitor how and when your key vaults are accessed, and by whom. You can do this by enabling logging for Azure Key Vault, which saves information in an Azure storage account that you provide.

Box 2: Create and configure Azure key vaults in two Azure regions

The contents of your key vault are replicated within the region and to a secondary region at least 150 miles away, but within the same geography to maintain high durability of your keys and secrets.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/security/workspaces-encryption>
<https://docs.microsoft.com/en-us/azure/key-vault/general/logging>

Question 159

CertyIQ

You plan to create an Azure Synapse Analytics dedicated SQL pool.

You need to minimize the time it takes to identify queries that return confidential information as defined by the company's data privacy regulations and the users who executed the queries.

Which two components should you include in the solution? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

A. sensitivity classification labels applied to columns that contain confidential information

B. resource tags for databases that contain confidential information

C. audit logs sent to a Log Analytics workspace

D. dynamic data masking for columns that contain confidential information

Explanation:

Correct Answer: C

A: You can classify columns manually, as an alternative or in addition to the recommendation-based classification:

Schema	Table	Column
SalesLT	Customer	FirstName
SalesLT	Customer	LastName
SalesLT	Customer	EmailAddress
SalesLT	Customer	Phone
SalesLT	Customer	PasswordHash
SalesLT	Customer	PasswordSalt
dbo	ErrorLog	UserName
SalesLT	Address	AddressLine1
SalesLT	Address	AddressLine2
SalesLT	Address	City
SalesLT	Address	PostalCode
SalesLT	CustomerAddress	AddressType
SalesLT	SalesOrderHeader	AccountNumber
SalesLT	SalesOrderHeader	CreditCardApprovalCode
SalesLT	SalesOrderHeader	TaxAmt

1. Select Add classification in the top menu of the pane.

2. In the context window that opens, select the schema, table, and column that you want to classify, and the information type and sensitivity label.

3. Select Add classification at the bottom of the context window.

C: An important aspect of the information-protection paradigm is the ability to monitor access to sensitive data. Azure SQL Auditing has been enhanced to include a new field in the audit log called `data_sensitivity_information`. This field logs the sensitivity classifications (labels) of the data that was returned by a query. Here's an example:

d	client_ip	application_name	duration_milliseconds	response_rows	affected_rows	connection_id	data_sensitivity_information
	7.125	Microsoft SQL Server Management Studio - Query	1	847	847	C244A066-2271...	Confidential - GDPR
	7.125	Microsoft SQL Server Management Studio - Query	2	32	32	C244A066-2271...	Confidential
	7.125	Microsoft SQL Server Management Studio - Query	41	32	32	A7088FD4-759E...	Confidential, Confidential - GDPR

Reference:

<https://docs.microsoft.com/en-us/azure/azure-sql/database/data-discovery-and-classification-overview>

Question 160

CertyIQ

You are designing an enterprise data warehouse in Azure Synapse Analytics that will contain a table named `Customers`. `Customers` will contain credit card information.

You need to recommend a solution to provide salespeople with the ability to view all the entries in Customers. The solution must prevent all the salespeople from viewing or inferring the credit card information.

What should you include in the recommendation?

- A. data masking
- B. Always Encrypted
- C. column-level security**
- D. row-level security

Explanation:

Correct Answer: C

Column-level security simplifies the design and coding of security in your application, allowing you to restrict column access to protect sensitive data.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/column-level-security>

End of Part 4



Please find the videos of this **AZ-900/AI-900/AZ-305/**

AZ-104 /DP-900/ SC-900 and other

Microsoft/AWS/GOOGLE/COMPTIA exam series on

CertyIQ Official YouTube channel (FREE PDFs): -

Please **Subscribe** to CertyIQ YouTube Channel to get notified for latest exam dumps by clicking on the below image, it will redirect to the **CertyIQ** YouTube page.



Connect with us @ [LinkedIn](#) [Telegram](#)

Contact us for other dumps: - certyiq@gmail.com

For any other enquiry, please drop us a mail at
certyiqofficial@gmail.com

DP-203

Real Exam Questions & Answers



Latest Exam Questions
All Topics Covered-2023
Added New Que's



Get Certified Today!

New Course Covered

Subscribe

Question 161

CertyIQ

You develop data engineering solutions for a company.

A project requires the deployment of data to Azure Data Lake Storage.

You need to implement role-based access control (RBAC) so that project members can manage the Azure Data Lake Storage resources.

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

NOTE: Each correct selection is worth one point.

- A. Create security groups in Azure Active Directory (Azure AD) and add project members.
- B. Configure end-user authentication for the Azure Data Lake Storage account.
- C. Assign Azure AD security groups to Azure Data Lake Storage.
- D. Configure Service-to-service authentication for the Azure Data Lake Storage account.
- E. Configure access control lists (ACL) for the Azure Data Lake Storage account.

Explanation:

Correct Answer: ACE

AC: Create security groups in Azure Active Directory. Assign users or security groups to Data Lake Storage Gen1 accounts.

E: Assign users or security groups as ACLs to the Data Lake Storage Gen1 file system

Reference:

<https://docs.microsoft.com/en-us/azure/data-lake-store/data-lake-store-secure-data>

Question 162

CertyIQ

You have an Azure Data Factory version 2 (V2) resource named Df1. Df1 contains a linked service.

You have an Azure Key vault named vault1 that contains an encryption key named key1.

You need to encrypt Df1 by using key1.

What should you do first?

- A. Add a private endpoint connection to vault1.
- B. Enable Azure role-based access control on vault1.
- C. Remove the linked service from Df1.**
- D. Create a self-hosted integration runtime.

Explanation:

Correct Answer: C

Linked services are much like connection strings, which define the connection information needed for Data Factory to connect to external resources.

You don't need to enable "RBAC", access policies is a default and more simple way to assign permissions, so B option is not necessary, but it is a requirement to delete the linked services to configure customer-managed key. So the correct answer is C - Delete linked services first.

<https://docs.microsoft.com/en-us/azure/key-vault/general/assign-access-policy?tabs=azure-portal>

<https://docs.microsoft.com/en-us/azure/data-factory/enable-customer-managed-key#enable-customer-managed-keys>

Incorrect Answers:

D: A self-hosted integration runtime copies data between an on-premises store and cloud storage.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/enable-customer-managed-key>

<https://docs.microsoft.com/en-us/azure/data-factory/concepts-linked-services>

<https://docs.microsoft.com/en-us/azure/data-factory/create-self-hosted-integration-runtime>

Question 163

CertyIQ

You are designing an Azure Synapse Analytics dedicated SQL pool.

You need to ensure that you can audit access to Personally Identifiable Information (PII).

What should you include in the solution?

- A. column-level security

- B. dynamic data masking
- C. row-level security (RLS)
- D. sensitivity classifications**

Explanation:

Correct Answer: D

Data Discovery & Classification is built into Azure SQL Database, Azure SQL Managed Instance, and Azure Synapse Analytics. It provides basic capabilities for discovering, classifying, labeling, and reporting the sensitive data in your databases.

Your most sensitive data might include business, financial, healthcare, or personal information. Discovering and classifying this data can play a pivotal role in your organization's information-protection approach. It can serve as infrastructure for:

- ⇒ Helping to meet standards for data privacy and requirements for regulatory compliance.
- ⇒ Various security scenarios, such as monitoring (auditing) access to sensitive data.
- ⇒ Controlling access to and hardening the security of databases that contain highly sensitive data.

Reference:

<https://docs.microsoft.com/en-us/azure/azure-sql/database/data-discovery-and-classification-overview>

Question 164

CertyIQ

HOTSPOT -

You have an Azure subscription that contains an Azure Data Lake Storage account. The storage account contains a data lake named DataLake1.

You plan to use an Azure data factory to ingest data from a folder in DataLake1, transform the data, and land the data in another folder.

You need to ensure that the data factory can read and write data from any folder in the DataLake1 file system. The solution must meet the following requirements:

- ⇒ Minimize the risk of unauthorized user access.
- ⇒ Use the principle of least privilege.
- ⇒ Minimize maintenance effort.

How should you configure access to the storage account for the data factory? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Use

Azure Active Directory (Azure AD)
a shared access signature (SAS)
a shared key

to authenticate by using

a managed identity
a stored access policy
an Authorization header

Answer Area

Suggested Answer:

Use

Azure Active Directory (Azure AD)
a shared access signature (SAS)
a shared key

to authenticate by using

a managed identity
a stored access policy
an Authorization header

Explanation:

Correct Answer:

Box 1: Azure Active Directory (Azure AD)

On Azure, managed identities eliminate the need for developers having to manage credentials by providing an identity for the Azure resource in Azure AD and using it to obtain Azure Active Directory (Azure AD) tokens.

Box 2: a managed identity -

A data factory can be associated with a managed identity for Azure resources, which represents this specific data factory. You can directly use this managed identity for Data Lake Storage Gen2 authentication, similar to using your own service principal. It allows this designated factory to access and copy data to or from your Data Lake Storage Gen2.

Note: The Azure Data Lake Storage Gen2 connector supports the following authentication types.

⇒ Account key authentication

⇒ Service principal authentication

⇒ Managed identities for Azure resources authentication

Reference:

<https://docs.microsoft.com/en-us/azure/active-directory/managed-identities-azure-resources/overview>

<https://docs.microsoft.com/en-us/azure/data-factory/connector-azure-data-lake-storage>

Question 165

CertyIQ

HOTSPOT -

You are designing an Azure Synapse Analytics dedicated SQL pool.

Groups will have access to sensitive data in the pool as shown in the following table.

Name	Enhanced access
Executives	No access to sensitive data
Analysts	Access to in-region sensitive data
Engineers	Access to all numeric sensitive data

You have policies for the sensitive data. The policies vary by region as shown in the following table.

Region	Data considered sensitive
RegionA	Financial, Personally Identifiable Information (PII)
RegionB	Financial, Personally Identifiable Information (PII), medical
RegionC	Financial, medical

You have a table of patients for each region. The tables contain the following potentially sensitive columns.

Name	Sensitive data	Description
CardOnFile	Financial	Debit/credit card number for charges
Height	Medical	Patient's height in cm
ContactEmail	PII	Email address for secure communications

You are designing dynamic data masking to maintain compliance.

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
Analysts in RegionA require dynamic data masking rules for [Patients_RegionA].	<input type="radio"/>	<input type="radio"/>
Engineers in RegionC require a dynamic data masking rule for [Patients_RegionA], [Height]	<input type="radio"/>	<input type="radio"/>
Engineers in RegionB require a dynamic data masking rule for [Patients_RegionB], [Height]	<input type="radio"/>	<input type="radio"/>

Answer Area

Statements	Yes	No
Suggested Answer: Analysts in RegionA require dynamic data masking rules for [Patients_RegionA].	<input type="radio"/>	<input checked="" type="radio"/>
Engineers in RegionC require a dynamic data masking rule for [Patients_RegionA], [Height]	<input type="radio"/>	<input checked="" type="radio"/>
Engineers in RegionB require a dynamic data masking rule for [Patients_RegionB], [Height]	<input type="radio"/>	<input checked="" type="radio"/>

Explanation:

Correct Answer: No, No, No.

Analysts have access to in-region sensitive data, so the first one should be No. Engineers have access to all numeric sensitive data, Height is patient's height in CM, so the second and third one should also No

Reference:

<https://docs.microsoft.com/en-us/azure/azure-sql/database/dynamic-data-masking-overview>

Question 166

CertyIQ

DRAG DROP -

You have an Azure Synapse Analytics SQL pool named Pool1 on a logical Microsoft SQL server named Server1. You need to implement Transparent Data Encryption (TDE) on Pool1 by using a custom key named key1. Which five actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

Select and Place:

Actions	Answer Area
Enable TDE on Pool1.	
Assign a managed identity to Server1.	
Configure key1 as the TDE protector for Server1.	◀ ▶
Add key1 to the Azure key vault.	
Create an Azure key vault and grant the managed identity permissions to the key vault.	

Actions	Answer Area
Enable TDE on Pool1.	Assign a managed identity to Server1.
Assign a managed identity to Server1.	Create an Azure key vault and grant the managed identity permissions to the key vault.
Configure key1 as the TDE protector for Server1.	◀ ▶
Add key1 to the Azure key vault.	Add key1 to the Azure key vault.
Create an Azure key vault and grant the managed identity permissions to the key vault.	Configure key1 as the TDE protector for Server1.
	Enable TDE on Pool1.

Explanation:

Correct Answer:

Step 1: Assign a managed identity to Server1

You will need an existing Managed Instance as a prerequisite.

Step 2: Create an Azure key vault and grant the managed identity permissions to the vault

Create Resource and setup Azure Key Vault.

Step 3: Add key1 to the Azure key vault

The recommended way is to import an existing key from a .pfx file or get an existing key from the vault. Alternatively, generate a new key directly in Azure Key

Vault.

Step 4: Configure key1 as the TDE protector for Server1

Provide TDE Protector key -

Step 5: Enable TDE on Pool1 -

Reference:

<https://docs.microsoft.com/en-us/azure/azure-sql/managed-instance/scripts/transparent-data-encryption-byok-powershell>

Question 167

CertyIQ

You have a data warehouse in Azure Synapse Analytics.

You need to ensure that the data in the data warehouse is encrypted at rest.

What should you enable?

A. Advanced Data Security for this database

B. Transparent Data Encryption (TDE)

C. Secure transfer required

D. Dynamic Data Masking

Explanation:

Correct Answer: B

Azure SQL Database currently supports encryption at rest for Microsoft-managed service side and client-side encryption scenarios.

⇒ Support for server encryption is currently provided through the SQL feature called Transparent Data Encryption.

⇒ Client-side encryption of Azure SQL Database data is supported through the Always Encrypted feature.

Reference:

<https://docs.microsoft.com/en-us/azure/security/fundamentals/encryption-atrest>

Question 168

CertyIQ

You are designing a streaming data solution that will ingest variable volumes of data.

You need to ensure that you can change the partition count after creation.

Which service should you use to ingest the data?

A. Azure Event Hubs Dedicated

B. Azure Stream Analytics

C. Azure Data Factory

D. Azure Synapse Analytics

Explanation:

Correct Answer: A

You can't change the partition count for an event hub after its creation except for the event hub in a dedicated cluster.

You can specify the number of partitions at the time of creating an event hub. In some scenarios, you may need to add partitions after the event hub has been created. This article describes how to dynamically add partitions to an existing event hub.

Dynamic additions of partitions is available only in premium and dedicated tiers of Event Hubs.

<https://docs.microsoft.com/en-us/azure/event-hubs/dynamically-add-partitions>

Reference:

<https://docs.microsoft.com/en-us/azure/event-hubs/event-hubs-features>

Question 169

CertyIQ

You are designing a date dimension table in an Azure Synapse Analytics dedicated SQL pool. The date dimension table will be used by all the fact tables.

Which distribution type should you recommend to minimize data movement during queries?

A. HASH

B. REPLICATE

C. ROUND_ROBIN

Explanation:

Correct Answer: B

A replicated table has a full copy of the table available on every Compute node. Queries run fast on replicated tables since joins on replicated tables don't require data movement. Replication requires extra storage, though, and isn't practical for large tables.

Incorrect Answers:

A: A hash distributed table is designed to achieve high performance for queries on large tables.

C: A round-robin table distributes table rows evenly across all distributions. The rows are distributed randomly. Loading data into a round-robin table is fast. Keep in mind that queries can require more data movement than the other distribution methods.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-overview>

Question 170

CertyIQ

HOTSPOT -

You develop a dataset named DBTBL1 by using Azure Databricks.

DBTBL1 contains the following columns:

- ⌚ SensorTypeID
- ⌚ GeographyRegionID
- ⌚ Year

- Month
- Day
- Hour
- Minute
- Temperature
- WindSpeed
- Other

You need to store the data to support daily incremental load pipelines that vary for each GeographyRegionID. The solution must minimize storage costs.

How should you complete the code? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

df.write

.bucketBy	(“*”)
.format	(“GeographyRegionID”)
.partitionBy	(“GeographyRegionID”, “Year”, “Month”, “Day”)
.sortBy	(“Year”, “Month”, “Day”, “GeographyRegionID”)
.mode (“append”)	
.csv(“/DBTBL1”)	
.json(“/DBTBL1”)	
.parquet(“/DBTBL1”)	
.saveAsTable(“/DBTBL1”)	

Answer Area

df.write

Suggested Answer:

.bucketBy	(“*”)
.format	(“GeographyRegionID”)
partitionBy	(“GeographyRegionID”, “Year”, “Month”, “Day”)
.sortBy	(“Year”, “Month”, “Day”, “GeographyRegionID”)
.mode (“append”)	
.csv(“/DBTBL1”)	
.json(“/DBTBL1”)	
parquet(“/DBTBL1”)	
.saveAsTable(“/DBTBL1”)	

Explanation:

Correct Answer:

1. Partition by
2. GeographyRegionID, Year, Month, Day
3. Parquet

Box 1: .partitionBy -

Incorrect Answers:

⇒ .format:

Method: format():

Arguments: "parquet", "csv", "txt", "json", "jdbc", "orc", "avro", etc.

⇒ .bucketBy:

Method: bucketBy()

Arguments: (numBuckets, col, col..., coln)

The number of buckets and names of columns to bucket by. Uses Hive's bucketing scheme on a filesystem.

Reference:

<https://www.oreilly.com/library/view/learning-spark-2nd/9781492050032/ch04.html> |

<https://docs.microsoft.com/en-us/azure/databricks/delta/delta-batch>

Question 171

CertyIQ

You are designing a security model for an Azure Synapse Analytics dedicated SQL pool that will support multiple companies.

You need to ensure that users from each company can view only the data of their respective company.

Which two objects should you include in the solution? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. a security policy
- B. a custom role-based access control (RBAC) role
- C. a predicate function
- D. a column encryption key
- E. asymmetric keys

Explanation:

Correct Answer: AB

A: Row-Level Security (RLS) enables you to use group membership or execution context to control access to rows in a database table. Implement RLS by using the CREATE SECURITY POLICYTransact-SQL statement.

B: Azure Synapse provides a comprehensive and fine-grained access control system, that integrates:

Azure roles for resource management and access to data in storage,

- ⇒ Synapse roles for managing live access to code and execution,
- ⇒ SQL roles for data plane access to data in SQL pools.

Reference:

<https://docs.microsoft.com/en-us/sql/relational-databases/security/row-level-security>

<https://docs.microsoft.com/en-us/azure/synapse-analytics/security/synapse-workspace-access-control-overview>

Question 172

CertyIQ

You have a SQL pool in Azure Synapse that contains a table named dbo.Customers. The table contains a column name Email.

You need to prevent nonadministrative users from seeing the full email addresses in the Email column. The users must see values in a format of aXXX@XXXX.com instead.

What should you do?

- A. From Microsoft SQL Server Management Studio, set an email mask on the Email column.
- B. From the Azure portal, set a mask on the Email column.
- C. From Microsoft SQL Server Management Studio, grant the SELECT permission to the users for all the columns in the dbo.Customers table except Email.
- D. From the Azure portal, set a sensitivity classification of Confidential for the Email column.

Explanation:

Correct Answer: A

The Email masking method, which exposes the first letter and replaces the domain with XXX.com using a constant string prefix in the form of an email address. aXX@XXXX.com

Reference:

<https://docs.microsoft.com/en-us/azure/azure-sql/database/dynamic-data-masking-overview>

Question 173

CertyIQ

You have an Azure Data Lake Storage Gen2 account named adls2 that is protected by a virtual network.

You are designing a SQL pool in Azure Synapse that will use adls2 as a source.

What should you use to authenticate to adls2?

- A. an Azure Active Directory (Azure AD) user
- B. a shared key
- C. a shared access signature (SAS)

D. a managed identity

Explanation:

Correct Answer: D

Managed Identity authentication is required when your storage account is attached to a VNet.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/quickstart-bulk-load-copy-tsql-examples>

Question 174

CertyIQ

HOTSPOT -

You have an Azure Synapse Analytics SQL pool named Pool1. In Azure Active Directory (Azure AD), you have a security group named Group1.

You need to control the access of Group1 to specific columns and rows in a table in Pool1.

Which Transact-SQL commands should you use? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

To control access to the columns:

	▼
CREATE CRYPTOGRAPHIC PROVIDER	
CREATE PARTITION FUNCTION	
CREATE SECURITY POLICY	
GRANT	

To control access to the rows:

	▼
CREATE CRYPTOGRAPHIC PROVIDER	
CREATE PARTITION FUNCTION	
CREATE SECURITY POLICY	
GRANT	

Suggested Answer:

Answer Area

To control access to the columns:

	▼
CREATE CRYPTOGRAPHIC PROVIDER	
CREATE PARTITION FUNCTION	
CREATE SECURITY POLICY	
GRANT	

To control access to the rows:

	▼
CREATE CRYPTOGRAPHIC PROVIDER	
CREATE PARTITION FUNCTION	
CREATE SECURITY POLICY	
GRANT	

Explanation:

Correct Answer:

Box 1: GRANT -

You can implement column-level security with the GRANT T-SQL statement. With this mechanism, both SQL and Azure Active Directory (Azure AD) authentication are supported.

Box 2: CREATE SECURITY POLICY -

Implement RLS by using the CREATE SECURITY POLICY Transact-SQL statement, and predicates created as inline table-valued functions.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/column-level-security>
<https://docs.microsoft.com/en-us/sql/relational-databases/security/row-level-security>

Question 175

CertyIQ

HOTSPOT -

You need to implement an Azure Databricks cluster that automatically connects to Azure Data Lake Storage Gen2 by using Azure Active Directory (Azure AD) integration.

How should you configure the new cluster? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Tier:

Premium	▼
Standard	

Advanced option to enable:

Azure Data Lake Storage Credential Passthrough	▼
Table Access Control	

Answer Area

Tier:

Premium	▼
Standard	

Suggested Answer:

Advanced option to enable:

Azure Data Lake Storage Credential Passthrough	▼
Table Access Control	

Explanation:

Correct Answer:

Box 1: Premium -

Credential passthrough requires an Azure Databricks Premium Plan

Box 2: Azure Data Lake Storage credential passthrough

You can access Azure Data Lake Storage using Azure Active Directory credential passthrough.

When you enable your cluster for Azure Data Lake Storage credential passthrough, commands that you run on that cluster can read and write data in Azure Data

Lake Storage without requiring you to configure service principal credentials for access to storage.

Reference:

<https://docs.microsoft.com/en-us/azure/databricks/security/credential-passthrough/adls-passthrough>

Question 176

CertyIQ

You are designing an Azure Synapse solution that will provide a query interface for the data stored in an Azure Storage account. The storage account is only accessible from a virtual network.

You need to recommend an authentication mechanism to ensure that the solution can access the source data.

What should you recommend?

- A. a managed identity
- B. anonymous public read access
- C. a shared key

Explanation:

Correct Answer: A

Managed Identity authentication is required when your storage account is attached to a VNet.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/quickstart-bulk-load-copy-tsql-examples>

Question 177

CertyIQ

You are developing an application that uses Azure Data Lake Storage Gen2.

You need to recommend a solution to grant permissions to a specific application for a limited time period.

What should you include in the recommendation?

- A. role assignments
- B. shared access signatures (SAS)
- C. Azure Active Directory (Azure AD) identities
- D. account keys

Explanation:

Correct Answer: B

A shared access signature (SAS) provides secure delegated access to resources in your storage account. With a SAS, you have granular control over how a client can access your data. For example:

What resources the client may access.

What permissions they have to those resources.

How long the SAS is valid.

Reference:

<https://docs.microsoft.com/en-us/azure/storage/common/storage-sas-overview>

Question 178

CertyIQ

HOTSPOT -

You use Azure Data Lake Storage Gen2 to store data that data scientists and data engineers will query by using Azure Databricks interactive notebooks. Users will have access only to the Data Lake Storage folders that relate to the projects on which they work.

You need to recommend which authentication methods to use for Databricks and Data Lake Storage to provide the users with the appropriate access. The solution must minimize administrative effort and development effort.

Which authentication method should you recommend for each Azure service? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Databricks:

- Azure Active Directory credential passthrough
- Azure Key Vault secrets
- Personal access tokens



Data Lake Storage:

- Azure Active Directory credential passthrough
- Shared access keys
- Shared access signatures



Answer Area

Databricks:

- Azure Active Directory credential passthrough
- Azure Key Vault secrets
- Personal access tokens



Suggested Answer:

Data Lake Storage:

- Azure Active Directory credential passthrough
- Shared access keys
- Shared access signatures



Explanation:

Correct Answer:

Accessing the ADLS via Databricks should be using Azure Active Directory with Passthrough. Accessing the files in ADLS should be SAS, based on the options provided.

The question is about how to authenticate the ADLS gen2 dataset both in Databricks and ADLSGen2... Its not about how you authenticate the Databricks.

Question 1: A Azure Active Directory with Passthrough

Access ADLS Gen2 from Databricks by running query interactively from notebooks.

Question 2: C 'Shared access signatures'

Users also need directly access to the Data Lake Storage for specific folders.

Question 179

CertyIQ

You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Contacts. Contacts contains a column named Phone.

You need to ensure that users in a specific role only see the last four digits of a phone number when querying the Phone column.

What should you include in the solution?

- A. table partitions
- B. a default value
- C. row-level security (RLS)
- D. column encryption
- E. dynamic data masking

Explanation:

Correct Answer:

Dynamic data masking helps prevent unauthorized access to sensitive data by enabling customers to designate how much of the sensitive data to reveal with minimal impact on the application layer. It's a policy-based security feature that hides the sensitive data in the result set of a query over designated database fields, while the data in the database is not changed.

Reference:

<https://docs.microsoft.com/en-us/azure/azure-sql/database/dynamic-data-masking-overview>

Question 180

CertyIQ

You are designing database for an Azure Synapse Analytics dedicated SQL pool to support workloads for detecting ecommerce transaction fraud.

Data will be combined from multiple ecommerce sites and can include sensitive financial information such as credit card numbers.

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

Users must be able to identify potentially fraudulent transactions.

⇒ Users must be able to use credit cards as a potential feature in models.

- ⇒ Users must NOT be able to access the actual credit card numbers.

What should you include in the recommendation?

- A. Transparent Data Encryption (TDE)
- B. row-level security (RLS)
- C. column-level encryption**
- D. Azure Active Directory (Azure AD) pass-through authentication

Explanation:

Correct Answer: C

By discard, is C, you can create a symmetric key to encrypt a data, for example one column, and then use this data as feature of the model

<https://docs.microsoft.com/en-us/sql/relational-databases/security/encryption/encrypt-a-column-of-data?view=sql-server-ver15>

The other options that not meet the requirements:

- TDE encrypt data, but decrypt when you query <https://docs.microsoft.com/en-us/azure/azure-sql/database/transparent-data-encryption-tde-overview?tabs=azure-portal>
- RLS is for row restriction, not meet the requirement
- Azure AD pass-through is for authentication

Use Always Encrypted to secure the required columns. You can configure Always Encrypted for individual database columns containing your sensitive data.

Always Encrypted is a feature designed to protect sensitive data, such as credit card numbers or national identification numbers (for example, U.S. social security numbers), stored in Azure SQL Database or SQL Server databases.

Reference:

<https://docs.microsoft.com/en-us/sql/relational-databases/security/encryption/always-encrypted-database-engine>

Question 181

CertyIQ

You have an Azure subscription linked to an Azure Active Directory (Azure AD) tenant that contains a service principal named ServicePrincipal1. The subscription contains an Azure Data Lake Storage account named adls1. Adls1 contains a folder named Folder2 that has a URI of <https://adls1.dfs.core.windows.net/container1/Folder1/Folder2/>.

ServicePrincipal1 has the access control list (ACL) permissions shown in the following table.

Resource	Permission
container1	Access – Execute
Folder1	Access – Execute
Folder2	Access – Read

You need to ensure that ServicePrincipal1 can perform the following actions:

- ⇒ Traverse child items that are created in Folder2.

☞ Read files that are created in Folder2.

The solution must use the principle of least privilege.

Which two permissions should you grant to ServicePrincipal1 for Folder2? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. Access "Read"
- B. Access "Write"
- C. Access "Execute"
- D. Default "Read"**
- E. Default "Write"
- F. Default "Execute"**

Explanation:

Correct Answer: DF

Execute (X) permission is required to traverse the child items of a folder.

There are two kinds of access control lists (ACLs), Access ACLs and Default ACLs.

Access ACLs: These control access to an object. Files and folders both have Access ACLs.

Default ACLs: A "template" of ACLs associated with a folder that determine the Access ACLs for any child items that are created under that folder. Files do not have Default ACLs.

Reference:

<https://docs.microsoft.com/en-us/azure/data-lake-store/data-lake-store-access-control>

Question 182

CertyIQ

HOTSPOT -

You have an Azure subscription that is linked to a hybrid Azure Active Directory (Azure AD) tenant. The subscription contains an Azure Synapse Analytics SQL pool named Pool1.

You need to recommend an authentication solution for Pool1. The solution must support multi-factor authentication (MFA) and database-level authentication.

Which authentication solution or solutions should you include in the recommendation? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

MFA:

Azure AD authentication
Microsoft SQL Server authentication
Passwordless authentication
Windows authentication

Database-level authentication:

Application roles
Contained database users
Database roles
Microsoft SQL Server logins

Answer Area

MFA:

Azure AD authentication
Microsoft SQL Server authentication
Passwordless authentication
Windows authentication

Suggested Answer:

Database-level authentication:

Application roles
Contained database users
Database roles
Microsoft SQL Server logins

Explanation:

Correct Answer:

Box 1: Azure AD authentication -

Azure AD authentication has the option to include MFA.

Box 2: Contained database users -

Azure AD authentication uses contained database users to authenticate identities at the database level.

Reference:

<https://docs.microsoft.com/en-us/azure/azure-sql/database/authentication-mfa-ssms-overview>

<https://docs.microsoft.com/en-us/azure/azure-sql/database/authentication-aad-overview>

Question 183

CertyIQ

DRAG DROP -

You have an Azure data factory.

You need to ensure that pipeline-run data is retained for 120 days. The solution must ensure that you can query the data by using the Kusto query language.

Which four actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

NOTE: More than one order of answer choices is correct. You will receive credit for any of the correct orders you select.

Select and Place:

Actions

Answer Area

Select the PipelineRuns category.

Create a Log Analytics workspace that has Data Retention set to 120 days.

Stream to an Azure event hub.

Create an Azure Storage account that has a lifecycle policy.

From the Azure portal, add a diagnostic setting.

Send the data to a Log Analytics workspace.

Select the TriggerRuns category.

Actions

Answer Area

Select the PipelineRuns category.

Create a Log Analytics workspace that has Data Retention set to 120 days.

Create a Log Analytics workspace that has Data Retention set to 120 days.

From the Azure portal, add a diagnostic setting.

Stream to an Azure event hub.

Select the PipelineRuns category.

Suggested Answer:

Create an Azure Storage account that has a lifecycle policy.

From the Azure portal, add a diagnostic setting.

Send the data to a Log Analytics workspace.

Select the TriggerRuns category.

Explanation:

Correct Answer:

Step 1: Create a Log Analytics workspace that has Data Retention set to 120 days.

Step 2: From Azure Portal, add a diagnostic setting.

Step 3: Select the PipelineRuns Category

Step 4: Send the data to a Log Analytics workspace.

Question 184

CertyIQ

You have an Azure Synapse Analytics dedicated SQL pool.

You need to ensure that data in the pool is encrypted at rest. The solution must NOT require modifying applications that query the data.

What should you do?

A. Enable encryption at rest for the Azure Data Lake Storage Gen2 account.

B. Enable Transparent Data Encryption (TDE) for the pool.

C. Use a customer-managed key to enable double encryption for the Azure Synapse workspace.

D. Create an Azure key vault in the Azure subscription grant access to the pool.

Explanation:

Correct Answer: B

Transparent Data Encryption (TDE) helps protect against the threat of malicious activity by encrypting and decrypting your data at rest. When you encrypt your database, associated backups and transaction log files are encrypted without requiring any changes to your applications. TDE encrypts the storage of an entire database by using a symmetric key called the database encryption key.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-overview-manage-security>

Question 185

CertyIQ

DRAG DROP -

You have an Azure subscription that contains an Azure Data Lake Storage Gen2 account named storage1. Storage1 contains a container named container1.

Container1 contains a directory named directory1. Directory1 contains a file named file1.

You have an Azure Active Directory (Azure AD) user named User1 that is assigned the Storage Blob Data Reader role for storage1.

You need to ensure that User1 can append data to file1. The solution must use the principle of least privilege. Which permissions should you grant? To answer, drag the appropriate permissions to the correct resources. Each permission 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.

Select and Place:

Permissions	Answer Area
Read	container1: <input type="text"/>
Write	directory1: <input type="text"/>
Execute	file1: <input type="text"/>

Suggested Answer:	Permissions	Answer Area
	Read	container1: <input type="text"/> Execute
	Write	directory1: <input type="text"/> Execute
	Execute	file1: <input type="text"/> Write

Explanation:

Correct Answer:

Box 1: Execute -

If you are granting permissions by using only ACLs (no Azure RBAC), then to grant a security principal read or write access to a file, you'll need to give the security principal Execute permissions to the root folder of the container, and to each folder in the hierarchy of folders that lead to the file.

Box 2: Execute -

On Directory: Execute (X): Required to traverse the child items of a directory

Box 3: Write -

On file: Write (W): Can write or append to a file.

Reference:

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

Question 186

CertyIQ

HOTSPOT -

You have an Azure subscription that contains an Azure Databricks workspace named databricks1 and an Azure Synapse Analytics workspace named synapse1.

The synapse1 workspace contains an Apache Spark pool named pool1.
You need to share an Apache Hive catalog of pool1 with databricks1.
What should you do? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

Hot Area:

From synapse1, create a linked service to:

- Azure Cosmos DB
- Azure Data Lake Storage Gen2
- Azure SQL Database

Configure pool1 to use the linked service as:

- An Azure Purview account
- A Hive metastore
- A managed Hive metastore service

Suggested Answer:

From synapse1, create a linked service to:

- Azure Cosmos DB
- Azure Data Lake Storage Gen2
- Azure SQL Database

Configure pool1 to use the linked service as:

- An Azure Purview account
- A Hive metastore
- A managed Hive metastore service

Explanation:

Correct Answer:

Box 1: Azure SQL Database -

Use external Hive Metastore for Synapse Spark Pool

Azure Synapse Analytics allows Apache Spark pools in the same workspace to share a managed HMS (Hive Metastore) compatible metastore as their catalog.

Set up linked service to Hive Metastore

Follow below steps to set up a linked service to the external Hive Metastore in Synapse workspace.

1. Open Synapse Studio, go to Manage > Linked services at left, click New to create a new linked service.
2. Set up Hive Metastore linked service
3. Choose Azure SQL Database or Azure Database for MySQL based on your database type, click Continue.
4. Provide Name of the linked service. Record the name of the linked service, this info will be used to configure Spark shortly.
5. You can either select Azure SQL Database/Azure Database for MySQL for the external Hive Metastore from Azure subscription list, or enter the info manually.
6. Provide User name and Password to set up the connection.
7. Test connection to verify the username and password.

8. Click Create to create the linked service.

Box 2: A Hive Metastore -

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/spark/apache-spark-external-metastore>

Question 187

CertyIQ

HOTSPOT -

You have an Azure subscription.

You need to deploy an Azure Data Lake Storage Gen2 Premium account. The solution must meet the following requirements:

- * Blobs that are older than 365 days must be deleted.
- * Administrative effort must be minimized.
- * Costs must be minimized.

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

NOTE: Each correct selection is worth one point.

Hot Area:

To minimize costs:

Locally-redundant storage (LRS)
The Archive access tier
The Cool access tier
Zone-redundant storage (ZRS)

To delete blobs:

Azure Automation runbooks
Azure Storage lifecycle management
Soft delete

To minimize costs:

Locally-redundant storage (LRS)

The Archive access tier

The Cool access tier

Zone-redundant storage (ZRS)

Suggested Answer:

To delete blobs:

Azure Automation runbooks

Azure Storage lifecycle management

Soft delete

Explanation:

Correct Answer:

Box 1: Locally-redundant storage (LRS)

If you choose premium storage account, there is no possibility to choose tiers (hot, cool, archive), it's always hot, so LRS

Box 2: Azure Storage lifecycle management

With the lifecycle management policy, you can:

- * Delete current versions of a blob, previous versions of a blob, or blob snapshots at the end of their lifecycles.

Transition blobs from cool to hot immediately when they're accessed, to optimize for performance.

Transition current versions of a blob, previous versions of a blob, or blob snapshots to a cooler storage tier if these objects haven't been accessed or modified for a period of time, to optimize for cost. In this scenario, the lifecycle management policy can move objects from hot to cool, from hot to archive, or from cool to archive.

Etc.

Reference:

<https://docs.microsoft.com/en-us/azure/storage/blobs/access-tiers-overview>

<https://docs.microsoft.com/en-us/azure/storage/blobs/lifecycle-management-overview>

Question 188

CertyIQ

HOTSPOT -

You are designing an application that will use an Azure Data Lake Storage Gen 2 account to store petabytes of license plate photos from toll booths. The account will use zone-redundant storage (ZRS).

You identify the following usage patterns:

- * The data will be accessed several times a day during the first 30 days after the data is created. The data must meet an availability SLA of 99.9%.
- * After 90 days, the data will be accessed infrequently but must be available within 30 seconds.
- * After 365 days, the data will be accessed infrequently but must be available within five minutes.

You need to recommend a data retention solution. The solution must minimize costs.

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

NOTE: Each correct selection is worth one point.

Hot Area:

First 30 days:

Archive
Cool
Hot

After 90 days:

Archive
Cool
Hot

After 365 days:

Archive
Cool
Hot

First 30 days:

Archive
Cool
Hot

After 90 days:

Archive
Cool
Hot

Suggested Answer:

After 365 days:

Archive
Cool
Hot

Explanation:

Correct Answer:

Box 1: Hot -

The data will be accessed several times a day during the first 30 days after the data is created. The data must meet an availability SLA of 99.9%.

Box 2: Cool -

After 90 days, the data will be accessed infrequently but must be available within 30 seconds.

Data in the Cool tier should be stored for a minimum of 30 days.

When your data is stored in an online access tier (either Hot or Cool), users can access it immediately. The Hot tier is the best choice for data that is in active use, while the Cool tier is ideal for data that is accessed less frequently, but that still must be available for reading and writing.

Box 3: Cool -

After 365 days, the data will be accessed infrequently but must be available within five minutes.

Incorrect:

Not Archive:

While a blob is in the Archive access tier, it's considered to be offline and can't be read or modified. In order to read or modify data in an archived blob, you must first rehydrate the blob to an online tier, either the Hot or Cool tier.

Rehydration priority -

When you rehydrate a blob, you can set the priority for the rehydration operation via the optional `x-ms-rehydrate-priority` header on a Set Blob Tier or Copy Blob operation. Rehydration priority options include:

Standard priority: The rehydration request will be processed in the order it was received and may take up to 15 hours.

High priority: The rehydration request will be prioritized over standard priority requests and may complete in less than one hour for objects under 10 GB in size.

Reference:

<https://docs.microsoft.com/en-us/azure/storage/blobs/access-tiers-overview>

<https://docs.microsoft.com/en-us/azure/storage/blobs/archive-rehydrate-overview>

Question 189

CertyIQ

You implement an enterprise data warehouse in Azure Synapse Analytics.

You have a large fact table that is 10 terabytes (TB) in size.

Incoming queries use the primary key SaleKey column to retrieve data as displayed in the following table:

SaleKey	CityKey	CustomerKey	StockItemKey	InvoiceDateKey	Quantity	UnitPrice	TotalExcludingTax
49309	90858	70	69	10/22/13	8	16	128
49313	55710	126	69	10/22/13	2	16	32
49343	44710	234	68	10/22/13	10	16	160
49352	66109	163	70	10/22/13	4	16	64
49448	65312	230	70	10/22/13	8	16	128
49646	85877	271	70	10/24/13	1	16	16
49798	41238	288	69	10/24/13	1	16	16

You need to distribute the large fact table across multiple nodes to optimize performance of the table.

Which technology should you use?

- A. hash distributed table with clustered index
- B. hash distributed table with clustered Columnstore index**
- C. round robin distributed table with clustered index
- D. round robin distributed table with clustered Columnstore index
- E. heap table with distribution replicate

Explanation:

Correct Answer: B

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

Columnstore indexes can achieve up to 100x better performance on analytics and data warehousing workloads and up to 10x better data compression than traditional rowstore indexes.

Clustered columnstore indexes are the most efficient way you can store your data in Azure SQL Data Warehouse. Storing your data in tables that have a clustered columnstore index are the fastest way to query your data. It will give you the greatest data compression and lower your storage costs.

Hash-distributed tables work well for large fact tables in a star schema. They can have very large numbers of rows and still achieve high performance.

Consider using a hash-distributed table when:

The table size on disk is more than 2 GB.

The table has frequent insert, update, and delete operations.

Incorrect Answers:

C, D: Round-robin tables are useful for improving loading speed.

Reference:

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

<https://docs.microsoft.com/en-us/sql/relational-databases/indexes/columnstore-indexes-query-performance>

Question 190

CertyIQ

You have an Azure Synapse Analytics dedicated SQL pool that contains a large fact table. The table contains 50 columns and 5 billion rows and is a heap.

Most queries against the table aggregate values from approximately 100 million rows and return only two columns.

You discover that the queries against the fact table are very slow.

Which type of index should you add to provide the fastest query times?

- A. nonclustered columnstore
- B. clustered columnstore**
- C. nonclustered
- D. clustered

Explanation:

Correct Answer: B

Clustered columnstore indexes are one of the most efficient ways you can store your data in dedicated SQL pool.

Columnstore tables won't benefit a query unless the table has more than 60 million rows.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/best-practices-dedicated-sql-pool>

Question 191

CertyIQ

You create an Azure Databricks cluster and specify an additional library to install.

When you attempt to load the library to a notebook, the library is not found.

You need to identify the cause of the issue.

What should you review?

- A. notebook logs
- B. cluster event logs**
- C. global init scripts logs
- D. workspace logs

Explanation:

Correct Answer: B Cluster Event logs:

Azure Databricks provides three kinds of logging of cluster-related activity:

Cluster event logs, which capture cluster lifecycle events, like creation, termination, configuration edits, and so on.

Apache Spark driver and worker logs, which you can use for debugging.

Cluster init-script logs, valuable for debugging init scripts.

<https://docs.microsoft.com/en-us/azure/databricks/clusters/clusters-manage#event-log>

Question 192

CertyIQ

You have an Azure data factory.

You need to examine the pipeline failures from the last 60 days.

What should you use?

- A. the Activity log blade for the Data Factory resource
- B. the Monitor & Manage app in Data Factory
- C. the Resource health blade for the Data Factory resource
- D. Azure Monitor**

Explanation:

Correct Answer: D

Data Factory stores pipeline-run data for only 45 days. Use Azure Monitor if you want to keep that data for a longer time.

Activity logs show only activities, e.g., trigger the pipeline, stop the pipeline, ...

Resource health check shows only the healthiness of the resource.

The monitor app indeed contains the pipeline run failure information. But it keeps the data only for 45 days.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/monitor-using-azure-monitor>

Question 193

CertyIQ

You are monitoring an Azure Stream Analytics job.

The Backlogged Input Events count has been 20 for the last hour.

You need to reduce the Backlogged Input Events count.

What should you do?

- A. Drop late arriving events from the job.
- B. Add an Azure Storage account to the job.
- C. Increase the streaming units for the job.**
- D. Stop the job.

Explanation:

Correct Answer: C

General symptoms of the job hitting system resource limits include:

☞ If the backlog event metric keeps increasing, it's an indicator that the system resource is constrained (either because of output sink throttling, or high CPU).

Note: Backlogged Input Events: Number of input events that are backlogged. A non-zero value for this metric implies that your job isn't able to keep up with the number of incoming events. If this value is slowly increasing or consistently non-zero, you should scale out your job: adjust Streaming Units.

"Backlogged Input Events Number of input events that are backlogged. A non-zero value for this metric implies that your job isn't able to keep up with the number of incoming events. If this value is slowly increasing or consistently non-zero, you should scale out your job."

Reference:

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-scale-jobs>

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-monitoring>

Question 194

CertyIQ

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. Pin the cluster.
- B. Create an Azure runbook that starts the cluster every 90 days.
- C. Terminate the cluster manually when processing completes.
- D. Clone the cluster after it is terminated.

Explanation:

Correct Answer: A

Azure Databricks retains cluster configuration information for up to 70 all-purpose clusters terminated in the last 30 days and up to 30 job clusters recently terminated by the job scheduler. To keep an all-purpose 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.microsoft.com/en-us/azure/databricks/clusters/>

Question 195

CertyIQ

You have an Azure data solution that contains an enterprise data warehouse in Azure Synapse Analytics named DW1.

Several users execute ad hoc queries to DW1 concurrently.

You regularly perform automated data loads to DW1.

You need to ensure that the automated data loads have enough memory available to complete quickly and successfully when the adhoc queries run.

What should you do?

- A. Hash distribute the large fact tables in DW1 before performing the automated data loads.
- B. Assign a smaller resource class to the automated data load queries.
- C. Assign a larger resource class to the automated data load queries.
- D. Create sampled statistics for every column in each table of DW1.

Explanation:

Correct Answer: C

The performance capacity of a query is determined by the user's resource class. Resource classes are pre-determined resource limits in Synapse SQL pool that govern compute resources and concurrency for query execution.

Resource classes can help you configure resources for your queries by setting limits on the number of queries that run concurrently and on the compute- resources assigned to each query. There's a trade-off between memory and concurrency.

Smaller resource classes reduce the maximum memory per query, but increase concurrency.

Larger resource classes increase the maximum memory per query, but reduce concurrency.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/resource-classes-for-workload-management>

Question 196

CertyIQ

You have an Azure Synapse Analytics dedicated SQL pool named Pool1 and a database named DB1. DB1 contains a fact table named Table1.

You need to identify the extent of the data skew in Table1.

What should you do in Synapse Studio?

- A. Connect to the built-in pool and run DBCC PDW_SHOWSPACEUSED.
- B. Connect to the built-in pool and run DBCC CHECKALLOC.
- C. Connect to Pool1 and query sys.dm_pdw_node_status.

D. Connect to Pool1 and query sys.dm_pdw_nodes_db_partition_stats.

Explanation:

Correct Answer: D

Microsoft recommends use of sys.dm_pdw_nodes_db_partition_stats to analyze any skewness in the data.

Use sys.dm_pdw_nodes_db_partition_stats to analyze any skewness in the data.

ref: <https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/cheat-sheet>

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/cheat-sheet>

Question 197

CertyIQ

HOTSPOT -

You need to collect application metrics, streaming query events, and application log messages for an Azure Databrick cluster.

Which type of library and workspace should you implement? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Library:

- Azure Databricks Monitoring Library
- Microsoft Azure Management Monitoring Library
- PyTorch
- TensorFlow

Workspace:

- Azure Databricks
- Azure Log Analytics
- Azure Machine Learning

Answer Area

Library:

- Azure Databricks Monitoring Library
- Microsoft Azure Management Monitoring Library
- PyTorch
- TensorFlow

Suggested Answer:

Workspace:

- Azure Databricks
- Azure Log Analytics
- Azure Machine Learning

Explanation:

Correct Answer:

You can send application logs and metrics from Azure Databricks to a Log Analytics workspace. It uses the Azure Databricks Monitoring Library, which is available on GitHub.

Reference:

<https://docs.microsoft.com/en-us/azure/architecture/databricks-monitoring/application-logs>

Question 198

CertyIQ

You have a SQL pool in Azure Synapse.

You discover that some queries fail or take a long time to complete.

You need to monitor for transactions that have rolled back.

Which dynamic management view should you query?

A. sys.dm_pdw_request_steps

B. sys.dm_pdw_nodes_tran_database_transactions

C. sys.dm_pdw_waits

D. sys.dm_pdw_exec_sessions

Explanation:

Correct Answer: B

You can use Dynamic Management Views (DMVs) to monitor your workload including investigating query execution in SQL pool.

If your queries are failing or taking a long time to proceed, you can check and monitor if you have any transactions rolling back.

Example:

```
-- Monitor rollback
```

```
SELECT -
```

```
SUM(CASE WHEN t.database_transaction_next_undo_lsn IS NOT NULL THEN 1 ELSE 0 END), t.pdw_node_id, nod.[type]  
FROM sys.dm_pdw_nodes_tran_database_transactions t  
JOIN sys.dm_pdw_nodes nod ON t.pdw_node_id = nod.pdw_node_id  
GROUP BY t.pdw_node_id, nod.[type]
```

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-manage-monitor#monitor-transaction-log-rollback>

Question 199

CertyIQ

You are monitoring an Azure Stream Analytics job.

You discover that the Backlogged Input Events metric is increasing slowly and is consistently non-zero.

You need to ensure that the job can handle all the events.

What should you do?

A. Change the compatibility level of the Stream Analytics job.

B. Increase the number of streaming units (SUs).

C. Remove any named consumer groups from the connection and use \$default.

D. Create an additional output stream for the existing input stream.

Explanation:

Correct Answer: B

Backlogged Input Events: Number of input events that are backlogged. A non-zero value for this metric implies that your job isn't able to keep up with the number of incoming events. If this value is slowly increasing or consistently non-zero, you should scale out your job. You should increase the Streaming Units.

Note: Streaming Units (SUs) represents the computing resources that are allocated to execute a Stream Analytics job. The higher the number of SUs, the more

CPU and memory resources are allocated for your job.

Reference:

<https://docs.microsoft.com/bs-cyrl-ba/azure/stream-analytics/stream-analytics-monitoring>

Question 200

CertyIQ

You are designing an inventory updates table in an Azure Synapse Analytics dedicated SQL pool. The table will have a clustered columnstore index and will include the following columns:

Table	Comment
EventDate	One million records are added to the table each day
EventTypeID	The table contains 10 million records for each event type.
WarehouseID	The table contains 100 million records for each warehouse.
ProductCategoryTypeID	The table contains 25 million records for each product category type.

You identify the following usage patterns:

- ⌚ Analysts will most commonly analyze transactions for a warehouse.
- ⌚ Queries will summarize by product category type, date, and/or inventory event type.

You need to recommend a partition strategy for the table to minimize query times.

On which column should you partition the table?

- A. EventTypeID
- B. ProductCategoryTypeID
- C. EventDate
- D. WarehouseID**

Explanation:

Correct Answer: D

The number of records for each warehouse is big enough for a good partitioning.

Note: Table partitions enable you to divide your data into smaller groups of data. In most cases, table partitions are created on a date column.

When creating partitions on clustered columnstore tables, it is important to consider how many rows belong to each partition. For optimal compression and performance of clustered columnstore tables, a minimum of 1 million rows per distribution and partition is needed. Before partitions are created, dedicated SQL pool already divides each table into 60 distributed databases.

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-partition>

End of Part 5



Please find the videos of this **AZ-900/AI-900/AZ-305/
AZ-104 /DP-900/ SC-900 and other Microsoft exam**
series on

CertyIQ Official YouTube channel (**FREE PDFs**): -

Please [Subscribe](#) to CertyIQ YouTube Channel to get notified for latest exam dumps by clicking on the below image, it will redirect to the [CertyIQ](#) YouTube page.



Connect with us @ [LinkedIn](#) [Telegram](#)

Contact us for other dumps: - certyiq@gmail.com

For any other enquiry, please drop us a mail at
certyiqofficial@gmail.com

DP-203

Real Exam Questions & Answers



Latest Exam Questions
All Topics Covered-2023
Added New Que's



Get Certified Today!

New Course Covered

Subscribe

Question 201

CertyIQ

You are designing a star schema for a dataset that contains records of online orders. Each record includes an order date, an order due date, and an order ship date.

You need to ensure that the design provides the fastest query times of the records when querying for arbitrary date ranges and aggregating by fiscal calendar attributes.

Which two actions should you perform? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

A. Create a date dimension table that has a DateTime key.

B. Use built-in SQL functions to extract date attributes.

C. Create a date dimension table that has an integer key in the format of YYYYMMDD.

D. In the fact table, use integer columns for the date fields.

E. Use DateTime columns for the date fields.

Explanation:

Correct Answer: CD

The question has many clues, it states fiscal calendar year and then star schema which hints we need proper fact and dim tables and appropriate date keys to link these.

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 Analysis Services using Azure Portal
- B. Azure Analysis Services using Azure PowerShell
- C. Azure Stream Analytics cloud job using Azure Portal**
- D. Azure Data Factory instance using Microsoft Visual Studio

Explanation:

Correct Answer: C

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. Your Stream Analytics job would ingest these events from Event Hubs and run real-time analytics queries against the streams.

Create a Stream Analytics job:

In the Azure portal, select + Create a resource from the left navigation menu. Then, select Stream Analytics job from Analytics.

Reference:

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-get-started-with-azure-stream-analytics-to-process-data-from-iot-devices>

You have a SQL pool in Azure Synapse.

A user reports that queries against the pool take longer than expected to complete. You determine that the issue relates to queried columnstore segments.

You need to add monitoring to the underlying storage to help diagnose the issue.

Which two metrics should you monitor? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. Snapshot Storage Size
- B. Cache used percentage**
- C. DWU Limit
- D. Cache hit percentage**

Explanation:

Correct Answer: BD

D: Cache hit percentage: $(\text{cache hits} / \text{cache miss}) * 100$ where cache hits is the sum of all columnstore segments hits in the local SSD cache and cache miss is the columnstore segments misses in the local SSD cache summed across all nodes

B: $(\text{cache used} / \text{cache capacity}) * 100$ where cache used is the sum of all bytes in the local SSD cache across all nodes and cache capacity is the sum of the storage capacity of the local SSD cache across all nodes

Incorrect Answers:

C: DWU limit: Service level objective of the data warehouse.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-concept-resource-utilization-query-activity>

Question 204

CertyIQ

You manage an enterprise data warehouse in Azure Synapse Analytics.

Users report slow performance when they run commonly used queries. Users do not report performance changes for infrequently used queries.

You need to monitor resource utilization to determine the source of the performance issues.

Which metric should you monitor?

A. DWU percentage

B. Cache hit percentage

C. DWU limit

D. Data IO percentage

Explanation:

Correct Answer: B

Monitor and troubleshoot slow query performance by determining whether your workload is optimally leveraging the adaptive cache for dedicated SQL pools.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-how-to-monitor-cache>

Question 205

CertyIQ

You have an Azure Databricks resource.

You need to log actions that relate to changes in compute for the Databricks resource.

Which Databricks services should you log?

A. clusters

- B. workspace
- C. DBFS
- D. SSH
- E. jobs

Explanation:

Correct Answer: A

workspace logs does not have any cluster related resource change.

<https://docs.microsoft.com/en-us/azure/databricks/administration-guide/account-settings/azure-diagnostic-logs#configure-diagnostic-log-delivery>

Question 206

CertyIQ

You are designing a highly available Azure Data Lake Storage solution that will include geo-zone-redundant storage (GZRS).

You need to monitor for replication delays that can affect the recovery point objective (RPO).

What should you include in the monitoring solution?

- A. 5xx: Server Error errors
- B. Average Success E2E Latency
- C. availability
- D. Last Sync Time**

Explanation:

Correct Answer: D

Because geo-replication is asynchronous, it is possible that data written to the primary region has not yet been written to the secondary region at the time an outage occurs. The Last Sync Time property indicates the last time that data from the primary region was written successfully to the secondary region. All writes made to the primary region before the last sync time are available to be read from the secondary location. Writes made to the primary region after the last sync time property may or may not be available for reads yet.

Reference:

<https://docs.microsoft.com/en-us/azure/storage/common/last-sync-time-get>

Question 207

CertyIQ

You configure monitoring for an Azure Synapse Analytics implementation. The implementation uses PolyBase to load data from comma-separated value (CSV) files stored in Azure Data Lake Storage Gen2 using an external table.

Files with an invalid schema cause errors to occur.

You need to monitor for an invalid schema error.

For which error should you monitor?

- A. EXTERNAL TABLE access failed due to internal error: 'Java exception raised on call to HdfsBridge_Connect: Error [com.microsoft.polybase.client.KerberosSecureLogin] occurred while accessing external file.'
- B. Cannot execute the query "Remote Query" against OLE DB provider "SQLNCLI11" for linked server "(null)".
Query aborted- the maximum reject threshold (0 rows) was reached while reading from an external source: 1 rows rejected out of total 1 rows processed.**
- C. EXTERNAL TABLE access failed due to internal error: 'Java exception raised on call to HdfsBridge_Connect: Error [Unable to instantiate LoginClass] occurred while accessing external file.'
- D. EXTERNAL TABLE access failed due to internal error: 'Java exception raised on call to HdfsBridge_Connect: Error [No FileSystem for scheme: wasbs] occurred while accessing external file.'

Explanation:

Correct Answer: B

Error message: Cannot execute the query "Remote Query"

Possible Reason:

The reason this error happens is because each file has different schema. The PolyBase external table DDL when pointed to a directory recursively reads all the files in that directory. When a column or data type mismatch happens, this error could be seen in SSMS.

Reference:

<https://docs.microsoft.com/en-us/sql/relational-databases/polybase/polybase-errors-and-possible-solutions>

Question 208

CertyIQ

You have an Azure Synapse Analytics dedicated SQL pool.

You run PDW_SHOWSPACEUSED('dbo.FactInternetSales'); and get the results shown in the following table.

ROWS	RESERVED_SPACE	DATA_SPACE	INDEX_SPACE	UNUSED_SPACE	PDW_NODE_ID	DISTRIBUTION_ID
694	2776	616	48	2112	1	1
407	2704	576	48	2080	1	2
53	2376	512	16	1848	1	3
58	2376	512	16	1848	1	4
168	2632	528	32	2072	1	5
195	2696	536	32	2128	1	6
5995	3464	1424	32	2008	1	7
0	2232	496	0	1736	1	8
264	2576	544	40	1992	1	9
3008	3016	960	32	2024	1	10
...
1550	2832	752	48	2032	1	50
1238	2832	696	40	2096	1	51
192	2632	528	32	2072	1	52
1127	2768	680	48	2040	1	53
1244	3032	704	64	2264	1	54
409	2632	568	32	2032	1	55
0	2232	496	0	1736	1	56
1437	2832	728	40	2064	1	57
0	2232	496	0	1736	1	58
384	2632	560	32	2040	1	59
225	2768	544	40	2184	1	60

Which statement accurately describes the dbo.FactInternetSales table?

- A. All distributions contain data.
- B. The table contains less than 10,000 rows.
- C. The table uses round-robin distribution.
- D. The table is skewed.**

Explanation:

Correct Answer: D

Data skew means the data is not distributed evenly across the distributions.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute>

Question 209

CertyIQ

You have two fact tables named Flight and Weather. Queries targeting the tables will be based on the join between the following columns.

Table	Column
Flight	ArrivalAirportID ArrivalDateTime
Weather	AirportID ReportDateTime

You need to recommend a solution that maximizes query performance.
What should you include in the recommendation?

- A. In the tables use a hash distribution of ArrivalDateTime and ReportDateTime.
- B. In the tables use a hash distribution of ArrivalAirportID and AirportID.**
- C. In each table, create an IDENTITY column.
- D. In each table, create a column as a composite of the other two columns in the table.

Explanation:

Correct Answer: B

Hash-distribution improves query performance on large fact tables.

Incorrect Answers:

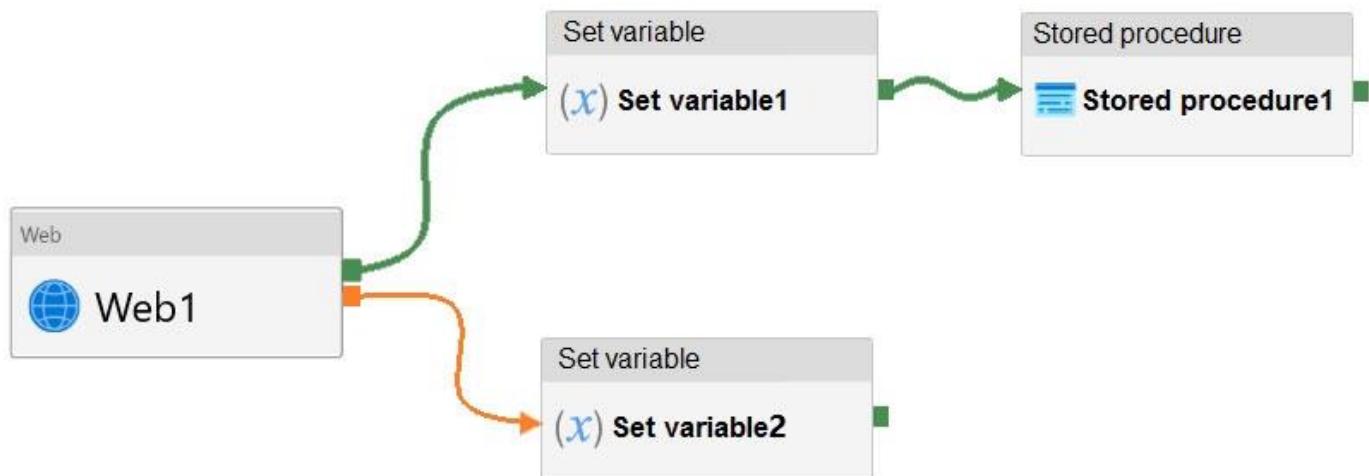
A: Do not use a date column for hash distribution. All data for the same date lands in the same distribution. If several users are all filtering on the same date, then only 1 of the 60 distributions do all the processing work.

Question 210

CertyIQ

HOTSPOT -

You have an Azure Data Factory pipeline that has the activities shown in the following exhibit.



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

Stored procedure1 will execute Web1 and Set variable1 [answer choice]

▼
complete
fail
succeed

If Web1 fails and Set variable2 succeeds, the pipeline status will be [answer choice]

▼
Canceled
Failed
Succeeded

Suggested Answer:

Answer Area

Stored procedure1 will execute Web1 and Set variable1 [answer choice]

▼
complete
fail
succeed

If Web1 fails and Set variable2 succeeds, the pipeline status will be [answer choice]

▼
Canceled
Failed
Succeeded

Explanation:

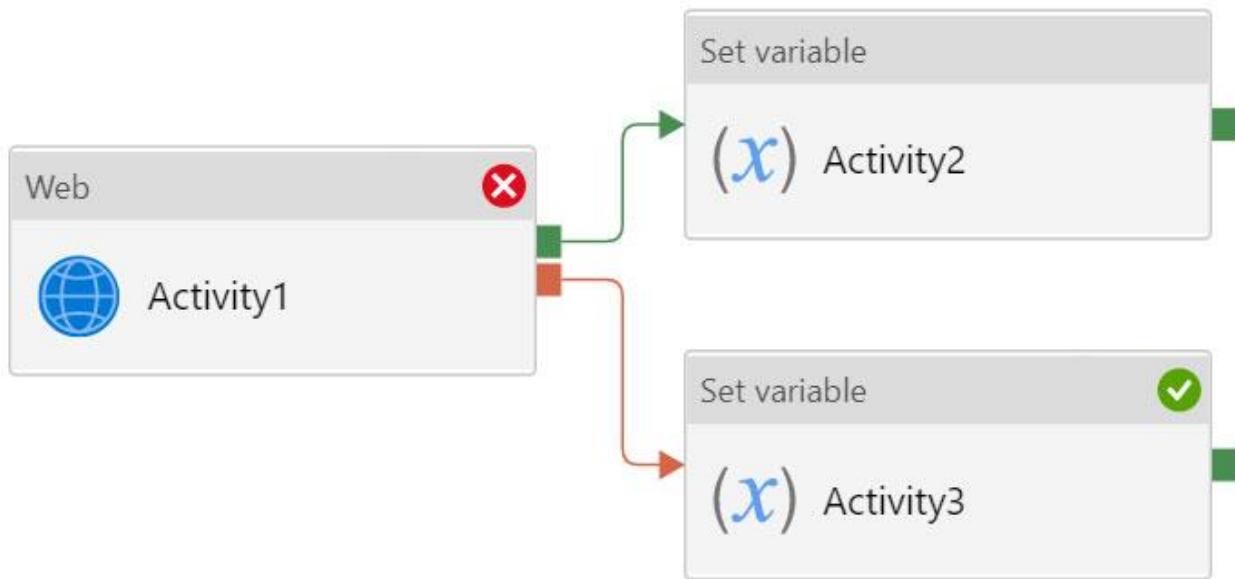
Correct Answer:

Box 1: succeed -

Box 2: failed -

Example:

Now let's say we have a pipeline with 3 activities, where Activity1 has a success path to Activity2 and a failure path to Activity3. If Activity1 fails and Activity3 succeeds, the pipeline will fail. The presence of the success path alongside the failure path changes the outcome reported by the pipeline, even though the activity executions from the pipeline are the same as the previous scenario.



Activity1 fails, Activity2 is skipped, and Activity3 succeeds. The pipeline reports failure.

Reference:

<https://datasavvy.me/2021/02/18/azure-data-factory-activity-failures-and-pipeline-outcomes/>

Question 211

CertyIQ

You have several Azure Data Factory pipelines that contain a mix of the following types of activities:

- ⇒ Wrangling data flow
- ⇒ Notebook
- ⇒ Copy
- ⇒ Jar

Which two Azure services should you use to debug the activities? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point

- A. Azure Synapse Analytics
- B. Azure HDInsight
- C. Azure Machine Learning
- D. Azure Data Factory**
- E. Azure Databricks**

Explanation:

Correct Answer: DE

Wrangling and Copy = ADF

Jar and Notebooks = Databricks

Question 212

CertyIQ

You have an Azure Synapse Analytics dedicated SQL pool named Pool1 and a database named DB1. DB1 contains a fact table named Table1.

You need to identify the extent of the data skew in Table1.

What should you do in Synapse Studio?

- A. Connect to the built-in pool and run sys.dm_pdw_nodes_db_partition_stats.
- B. Connect to Pool1 and run DBCC CHECKALLOC.
- C. Connect to the built-in pool and run DBCC CHECKALLOC.
- D. Connect to Pool1 and query sys.dm_pdw_nodes_db_partition_stats**

Explanation:

Correct Answer: D

Microsoft recommends use of sys.dm_pdw_nodes_db_partition_stats to analyze any skewness in the data.

Reference:

<https://docs.microsoft.com/en-us/sql/relational-databases/system-dynamic-management-views/sys-dm-db-partition-stats-transact-sql>

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/cheat-sheet>

Question 213

CertyIQ

You manage an enterprise data warehouse in Azure Synapse Analytics.

Users report slow performance when they run commonly used queries. Users do not report performance changes for infrequently used queries.

You need to monitor resource utilization to determine the source of the performance issues.

Which metric should you monitor?

- A. Local tempdb percentage
- B. Cache used percentage**
- C. Data IO percentage
- D. CPU percentage

Explanation:

Correct Answer: B

Monitor and troubleshoot slow query performance by determining whether your workload is optimally leveraging the adaptive cache for dedicated SQL pools.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-how-to-monitor-cache>

Question 214

CertyIQ

You have an Azure data factory.

You need to examine the pipeline failures from the last 180 days.

What should you use?

- A. the Activity log blade for the Data Factory resource
- B. Pipeline runs in the Azure Data Factory user experience
- C. the Resource health blade for the Data Factory resource
- D. Azure Data Factory activity runs in Azure Monitor**

Explanation:

Correct Answer: D

Data Factory stores pipeline-run data for only 45 days. Use Azure Monitor if you want to keep that data for a longer time.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/monitor-using-azure-monitor>

Question 215

CertyIQ

You have an Azure Synapse Analytics dedicated SQL pool named SA1 that contains a table named Table1.

You need to identify tables that have a high percentage of deleted rows.

What should you run?

- A. sys.pdw_nodes_column_store_segments
- B. sys.dm_db_column_store_row_group_operational_stats
- C. sys.pdw_nodes_column_store_row_groups**
- D. sys.dm_db_column_store_row_group_physical_stats

Explanation:

Correct Answer: C

Use sys.pdw_nodes_column_store_row_groups to determine which row groups have a high percentage of deleted rows and should be rebuilt.

Note: sys.pdw_nodes_column_store_row_groups provides clustered columnstore index information on a per-segment basis to help the administrator make system management decisions in Azure Synapse Analytics. sys.pdw_nodes_column_store_row_groups has a column for the total number of rows physically stored (including those marked as deleted) and a column for the number of rows marked as deleted.

Incorrect:

Not A: You can join sys.pdw_nodes_column_store_segments with other system tables to determine the number of columnstore segments per logical table.

Not B: Use sys.dm_db_column_store_row_group_operational_stats to track the length of time a user query must wait to read or write to a compressed rowgroup or partition of a columnstore index, and identify rowgroups that are encountering significant I/O activity or hot spots.

Question 216

CertyIQ

You have an enterprise data warehouse in Azure Synapse Analytics.

You need to monitor the data warehouse to identify whether you must scale up to a higher service level to accommodate the current workloads.

Which is the best metric to monitor?

More than one answer choice may achieve the goal. Select the BEST answer.

- A. DWU used
- B. CPU percentage
- C. DWU percentage
- D. Data IO percentage

Explanation:

Correct Answer: A

DWU used: DWU limit * DWU percentage

DWU used represents only a high-level representation of usage across the SQL pool and is not meant to be a comprehensive indicator of utilization. To determine whether to scale up or down, consider all factors which can be impacted by DWU such as concurrency, memory, tempdb, and adaptive cache capacity. We recommend running your workload at different DWU settings to determine what works best to meet your business objectives.

Azure Synapse Analytics monitor metric "DWU used"

Incorrect:

* CPU percentage: CPU utilization across all nodes for the data warehouse.

* DWU percentage: Maximum between CPU percentage and Data IO percentage

* Data IO percentage: IO Utilization across all nodes for the data warehouse

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-concept-resource-utilization-query-activity>

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	

Answer Area

Suggested Answer:

Number of partitions:

1
8
16
32

Partition key:

Fraud indicator
Fraud score
Individual line items
Payment details
Transaction ID

Explanation:

Correct Answer:

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 -

Event Hub -> Event Hub: x:x partitions

Event Hub -> Blob Storage: x:1 partitions or x:y partitions

Blob Storage -> Event Hub: x:x partitions

Blob Storage -> Blob Storage: x:1 partitions

Reference:

<https://docs.microsoft.com/en-us/azure/event-hubs/event-hubs-features#partitions>

Question 218

CertyIQ

HOTSPOT -

You have an on-premises data warehouse that includes the following fact tables. Both tables have the following columns: DateKey, 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. Severity-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:	<input type="checkbox"/> Hash-distributed <input type="checkbox"/> Round-robin	<input type="checkbox"/> DateKey <input type="checkbox"/> ProductKey <input type="checkbox"/> RegionKey
Invoices:	<input type="checkbox"/> Hash-distributed <input type="checkbox"/> Round-robin	<input type="checkbox"/> DateKey <input type="checkbox"/> ProductKey <input type="checkbox"/> RegionKey

Answer Area

Suggested Answer:	Table	Distribution type	Distribution column
	Sales:	<input checked="" type="checkbox"/> Hash-distributed <input type="checkbox"/> Round-robin	<input type="checkbox"/> DateKey <input checked="" type="checkbox"/> ProductKey <input type="checkbox"/> RegionKey
	Invoices:	<input checked="" type="checkbox"/> Hash-distributed <input type="checkbox"/> Round-robin	<input type="checkbox"/> DateKey <input type="checkbox"/> ProductKey <input checked="" type="checkbox"/> RegionKey

Explanation:

Correct Answer:

1. Hash Distributed, ProductKey because >2GB and ProductKey is extensively used in joins
2. Hash Distributed, RegionKey because "The table size on disk is more than 2 GB." and you have to chose a distribution column which: "Is not used in WHERE clauses. This could narrow the query to not run on all the distributions."

source: <https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute#choosing-a-distribution-column>

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: Hash-distributed -

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>

Question 219

CertyIQ

You have a partitioned table in an Azure Synapse Analytics dedicated SQL pool.

You need to design queries to maximize the benefits of partition elimination.

What should you include in the Transact-SQL queries?

- A. JOIN
- B. WHERE**
- C. DISTINCT
- D. GROUP BY

Explanation:

Correct Answer: B

Think of it this way, you have 36 partitions over Month column for a table. You are interested in a specific month. so in WHERE clause of your select statement, you will give specific month to "eliminate" other 35 partitions scan.

<https://stackoverflow.com/questions/51677471/what-is-a-difference-between-table-distribution-and-table-partition-in-sql/51677595>

Question 220

CertyIQ

You have an Azure Synapse Analytics dedicated SQL pool named Pool1 and a database named DB1. DB1 contains a fact table named Table1.

You need to identify the extent of the data skew in Table1.

What should you do in Synapse Studio?

- A. Connect to the built-in pool and query sys.dmv_nodes_db_partition_stats.
- B. Connect to the built-in pool and run DBCC CHECKALLOC.
- C. Connect to Pool1 and query sys.dmv_node_status.
- D. Connect to Pool1 and query sys.dmv_nodes_db_partition_stats**

Explanation:

Correct Answer: D

built-in pool comes from a Synapse Serverless pool and here it says Dedicated

Question 221

CertyIQ

You have an Azure Synapse Analytics dedicated SQL pool named Pool1. Pool1 contains a fact table named Table1.

You need to identify the extent of the data skew in Table1.

What should you do in Synapse Studio?

- A. Connect to Pool1 and DBCC PDW_SHOWSPACEUSED.**
- B. Connect to the built-in pool and run DBCC PDW_SHOWSPACEUSED.
- C. Connect to the built-in pool and run DBCC CHECKALLOC.
- D. Connect to the built-in pool and query sys.dmv_sys_info.

Explanation:

Correct Answer: A

A quick way to check for data skew is to use DBCC PDW_SHOWSPACEUSED. The following SQL code returns the number of table rows that are stored in each of the 60 distributions. For balanced performance, the rows in your distributed table should be spread evenly across all the distributions.

<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute>

Question 222

CertyIQ

Contoso, Ltd. is a clothing retailer based in Seattle. The company has 2,000 retail stores across the United States and an emerging online presence.

The network contains an Active Directory forest named contoso.com. The forest is integrated with an Azure Active Directory (Azure AD) tenant named contoso.com. Contoso has an Azure subscription associated to the contoso.com Azure AD tenant.

Existing Environment

Transactional Data

Contoso has three years of customer, transactional, operational, sourcing, and supplier data comprised of 10 billion records stored across multiple on-premises Microsoft SQL Server servers. The SQL Server instances contain data from various operational systems. The data is loaded into the instances by using SQL Server Integration Services (SSIS) packages.

You estimate that combining all product sales transactions into a company-wide sales transactions dataset will result in a single table that contains 5 billion rows, with one row per transaction.

Most queries targeting the sales transactions data will be used to identify which products were sold in retail stores and which products were sold online during different time periods. Sales transaction data that is older than three years will be removed monthly.

You plan to create a retail store table that will contain the address of each retail store. The table will be approximately 2 MB. Queries for retail store sales will include the retail store addresses.

You plan to create a promotional table that will contain a promotion ID. The promotion ID will be associated to a specific product. The product will be identified by a product ID. The table will be approximately 5 GB.

Streaming Twitter Data

The ecommerce department at Contoso develops an Azure logic app that captures trending Twitter feeds referencing the company's products and pushes the products to Azure Event Hubs.

Planned Changes and Requirements

Planned Changes

Contoso plans to implement the following changes:

Load the sales transaction dataset to Azure Synapse Analytics.

Integrate on-premises data stores with Azure Synapse Analytics by using SSIS packages.

Use Azure Synapse Analytics to analyze Twitter feeds to assess customer sentiments about products.

Sales Transaction Dataset Requirements

Contoso identifies the following requirements for the sales transaction dataset:

Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month. Boundary values must belong to the partition on the right.

Ensure that queries joining and filtering sales transaction records based on product ID complete as quickly as possible.

Implement a surrogate key to account for changes to the retail store addresses.

Ensure that data storage costs and performance are predictable.

Minimize how long it takes to remove old records.

Customer Sentiment Analytics Requirements

Contoso identifies the following requirements for customer sentiment analytics:

Allow Contoso users to use PolyBase in an Azure Synapse Analytics dedicated SQL pool to query the content of the data records that host the Twitter feeds.

Data must be protected by using row-level security (RLS). The users must be authenticated by using their own Azure AD credentials.

Maximize the throughput of ingesting Twitter feeds from Event Hubs to Azure Storage without purchasing additional throughput or capacity units.

Store Twitter feeds in Azure Storage by using Event Hubs Capture. The feeds will be converted into Parquet files.

Ensure that the data store supports Azure AD-based access control down to the object level.

Minimize administrative effort to maintain the Twitter feed data records.

Purge Twitter feed data records that are older than two years.

Data Integration Requirements

Contoso identifies the following requirements for data integration:

Use an Azure service that leverages the existing SSIS packages to ingest on-premises data into datasets stored in a dedicated SQL pool of Azure Synapse Analytics and transform the data.

Identify a process to ensure that changes to the ingestion and transformation activities can be version-controlled and developed independently by multiple data engineers.

Question:

HOTSPOT -

You need to design a data storage structure for the product sales transactions. The solution must meet the sales transaction dataset requirements.

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Table type to store the product sales transactions:

Hash
Round-robin
Replicated

When creating the table for sales transactions:

Configure a clustered index.
Set the distribution column to product ID.
Set the distribution column to the sales date.

Suggested Answer:

Answer Area

Table type to store the product sales transactions:

Hash
Round-robin
Replicated

When creating the table for sales transactions:

Configure a clustered index.
Set the distribution column to product ID.
Set the distribution column to the sales date.

Explanation:

Correct Answer:

Box 1: Hash -

Scenario:

Ensure that queries joining and filtering sales transaction records based on product ID complete as quickly as possible.

A hash distributed table can deliver the highest query performance for joins and aggregations on large tables.

Box 2: Set distribution to Product ID

Question 223

CertyIQ

DRAG DROP -

You need to ensure that the Twitter feed data can be analyzed in the dedicated SQL pool. The solution must

meet the customer sentiment analytics requirements.

Which three Transact-SQL DDL commands should you run in sequence? To answer, move the appropriate commands from the list of commands to the answer area and arrange them in the correct order.

NOTE: More than one order of answer choices is correct. You will receive credit for any of the correct orders you select.

Select and Place:

Commands

Answer Area

CREATE EXTERNAL DATA SOURCE
CREATE EXTERNAL FILE FORMAT
CREATE EXTERNAL TABLE
CREATE EXTERNAL TABLE AS SELECT
CREATE DATABASE SCOPED CREDENTIAL

Suggested Answer:

Commands

Answer Area

CREATE EXTERNAL DATA SOURCE
CREATE EXTERNAL FILE FORMAT
CREATE EXTERNAL TABLE
CREATE EXTERNAL TABLE AS SELECT
CREATE DATABASE SCOPED CREDENTIAL

CREATE EXTERNAL DATA SOURCE
CREATE EXTERNAL FILE FORMAT
CREATE EXTERNAL TABLE AS SELECT

Explanation:

Correct Answer:

Scenario: Allow Contoso users to use PolyBase in an Azure Synapse Analytics dedicated SQL pool to query the content of the data records that host the Twitter feeds. Data must be protected by using row-level security (RLS). The users must be authenticated by using their own Azure AD credentials.

Box 1: CREATE EXTERNAL DATA SOURCE

External data sources are used to connect to storage accounts.

Box 2: CREATE EXTERNAL FILE FORMAT

CREATE EXTERNAL FILE FORMAT creates an external file format object that defines external data stored in Azure Blob Storage or Azure Data Lake Storage.

Creating an external file format is a prerequisite for creating an external table.

Box 3: CREATE EXTERNAL TABLE AS SELECT

When used in conjunction with the CREATE TABLE AS SELECT statement, selecting from an external table imports data into a table within the SQL pool. In addition to the COPY statement, external tables are useful for loading data.

Incorrect Answers:

CREATE EXTERNAL TABLE -

The CREATE EXTERNAL TABLE command creates an external table for Synapse SQL to access data stored in Azure Blob Storage or Azure Data Lake Storage.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables>

Question 224

CertyIQ

Contoso, Ltd. is a clothing retailer based in Seattle. The company has 2,000 retail stores across the United States and an emerging online presence.

The network contains an Active Directory forest named contoso.com. The forest is integrated with an Azure Active Directory (Azure AD) tenant named contoso.com. Contoso has an Azure subscription associated to the contoso.com Azure AD tenant.

Existing Environment

Transactional Data

Contoso has three years of customer, transactional, operational, sourcing, and supplier data comprised of 10 billion records stored across multiple on-premises Microsoft SQL Server servers. The SQL Server instances contain data from various operational systems. The data is loaded into the instances by using SQL Server Integration Services (SSIS) packages.

You estimate that combining all product sales transactions into a company-wide sales transactions dataset will result in a single table that contains 5 billion rows, with one row per transaction.

Most queries targeting the sales transactions data will be used to identify which products were sold in retail stores and which products were sold online during different time periods. Sales transaction data that is older than three years will be removed monthly.

You plan to create a retail store table that will contain the address of each retail store. The table will be approximately 2 MB. Queries for retail store sales will include the retail store addresses.

You plan to create a promotional table that will contain a promotion ID. The promotion ID will be associated to a specific product. The product will be identified by a product ID. The table will be approximately 5 GB.

Streaming Twitter Data

The ecommerce department at Contoso develops an Azure logic app that captures trending Twitter feeds referencing the company's products and pushes the products to Azure Event Hubs.

Planned Changes and Requirements

Planned Changes

Contoso plans to implement the following changes:

Load the sales transaction dataset to Azure Synapse Analytics.

Integrate on-premises data stores with Azure Synapse Analytics by using SSIS packages.

Use Azure Synapse Analytics to analyze Twitter feeds to assess customer sentiments about products.

Sales Transaction Dataset Requirements

Contoso identifies the following requirements for the sales transaction dataset:

Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month. Boundary values must belong to the partition on the right.

Ensure that queries joining and filtering sales transaction records based on product ID complete as quickly as possible.

Implement a surrogate key to account for changes to the retail store addresses.

Ensure that data storage costs and performance are predictable.

Minimize how long it takes to remove old records.

Customer Sentiment Analytics Requirements

Contoso identifies the following requirements for customer sentiment analytics:

Allow Contoso users to use PolyBase in an Azure Synapse Analytics dedicated SQL pool to query the content of the data records that host the Twitter feeds.

Data must be protected by using row-level security (RLS). The users must be authenticated by using their own Azure AD credentials.

Maximize the throughput of ingesting Twitter feeds from Event Hubs to Azure Storage without purchasing additional throughput or capacity units.

Store Twitter feeds in Azure Storage by using Event Hubs Capture. The feeds will be converted into Parquet files.

Ensure that the data store supports Azure AD-based access control down to the object level.

Minimize administrative effort to maintain the Twitter feed data records.

Purge Twitter feed data records that are older than two years.

Data Integration Requirements

Contoso identifies the following requirements for data integration:

Use an Azure service that leverages the existing SSIS packages to ingest on-premises data into datasets stored in a dedicated SQL pool of Azure Synapse Analytics and transform the data.

Identify a process to ensure that changes to the ingestion and transformation activities can be version-controlled and developed independently by multiple data engineers.

Question:

HOTSPOT -

You need to design the partitions for the product sales transactions. The solution must meet the sales

transaction dataset requirements.

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Partition product sales transactions data by:

Sales date
Product ID
Promotion ID

Store product sales transactions data in:

An Azure Synapse Analytics dedicated SQL pool
An Azure Synapse Analytics serverless SQL pool
An Azure Data Lake Storage Gen2 account linked to an Azure Synapse Analytics workspace

Answer Area

Partition product sales transactions data by:

Sales date
Product ID
Promotion ID

Suggested Answer:

Store product sales transactions data in:

An Azure Synapse Analytics dedicated SQL pool
An Azure Synapse Analytics serverless SQL pool
An Azure Data Lake Storage Gen2 account linked to an Azure Synapse Analytics workspace
An Azure Synapse Analytics workspace

Explanation:

Correct Answer:

Box 1: Sales date -

Scenario: Contoso requirements for data integration include:

☞ Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month. Boundary values must belong to the partition on the right.

Box 2: An Azure Synapse Analytics Dedicated SQL pool

Scenario: Contoso requirements for data integration include:

⇒ Ensure that data storage costs and performance are predictable.

The size of a dedicated SQL pool (formerly SQL DW) is determined by Data Warehousing Units (DWU).

Dedicated SQL pool (formerly SQL DW) stores data in relational tables with columnar storage. This format significantly reduces the data storage costs, and improves query performance.

Synapse analytics dedicated sql pool

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-overview-what-is>

Question 225

CertyIQ

Contoso, Ltd. is a clothing retailer based in Seattle. The company has 2,000 retail stores across the United States and an emerging online presence.

The network contains an Active Directory forest named contoso.com. The forest is integrated with an Azure Active Directory (Azure AD) tenant named contoso.com. Contoso has an Azure subscription associated to the contoso.com Azure AD tenant.

Existing Environment

Transactional Data

Contoso has three years of customer, transactional, operational, sourcing, and supplier data comprised of 10 billion records stored across multiple on-premises Microsoft SQL Server servers. The SQL Server instances contain data from various operational systems. The data is loaded into the instances by using SQL Server Integration Services (SSIS) packages.

You estimate that combining all product sales transactions into a company-wide sales transactions dataset will result in a single table that contains 5 billion rows, with one row per transaction.

Most queries targeting the sales transactions data will be used to identify which products were sold in retail stores and which products were sold online during different time periods. Sales transaction data that is older than three years will be removed monthly.

You plan to create a retail store table that will contain the address of each retail store. The table will be approximately 2 MB. Queries for retail store sales will include the retail store addresses.

You plan to create a promotional table that will contain a promotion ID. The promotion ID will be associated to a specific product. The product will be identified by a product ID. The table will be approximately 5 GB.

Streaming Twitter Data

The ecommerce department at Contoso develops an Azure logic app that captures trending Twitter feeds referencing the company's products and pushes the products to Azure Event Hubs.

Planned Changes and Requirements

Planned Changes

Contoso plans to implement the following changes:

Load the sales transaction dataset to Azure Synapse Analytics.

Integrate on-premises data stores with Azure Synapse Analytics by using SSIS packages.

Use Azure Synapse Analytics to analyze Twitter feeds to assess customer sentiments about products.

Sales Transaction Dataset Requirements

Contoso identifies the following requirements for the sales transaction dataset:

Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month. Boundary values must belong to the partition on the right.

Ensure that queries joining and filtering sales transaction records based on product ID complete as quickly as possible.

Implement a surrogate key to account for changes to the retail store addresses.

Ensure that data storage costs and performance are predictable.

Minimize how long it takes to remove old records.

Customer Sentiment Analytics Requirements

Contoso identifies the following requirements for customer sentiment analytics:

Allow Contoso users to use PolyBase in an Azure Synapse Analytics dedicated SQL pool to query the content of the data records that host the Twitter feeds.

Data must be protected by using row-level security (RLS). The users must be authenticated by using their own Azure AD credentials.

Maximize the throughput of ingesting Twitter feeds from Event Hubs to Azure Storage without purchasing additional throughput or capacity units.

Store Twitter feeds in Azure Storage by using Event Hubs Capture. The feeds will be converted into Parquet files.

Ensure that the data store supports Azure AD-based access control down to the object level.

Minimize administrative effort to maintain the Twitter feed data records.

Purge Twitter feed data records that are older than two years.

Data Integration Requirements

Contoso identifies the following requirements for data integration:

Use an Azure service that leverages the existing SSIS packages to ingest on-premises data into datasets stored in a dedicated SQL pool of Azure Synapse Analytics and transform the data.

Identify a process to ensure that changes to the ingestion and transformation activities can be version-controlled and developed independently by multiple data engineers.

Question

You need to implement the surrogate key for the retail store table. The solution must meet the sales transaction dataset requirements.

What should you create?

- A. a table that has an IDENTITY property
- B. a system-versioned temporal table
- C. a user-defined SEQUENCE object
- D. a table that has a FOREIGN KEY constraint

Explanation:

Correct Answer: A

Scenario: Implement a surrogate key to account for changes to the retail store addresses.

A surrogate key on a table is a column with a unique identifier for each row. The key is not generated from the table data. Data modelers like to create surrogate keys on their tables when they design data warehouse models. You can use the IDENTITY property to achieve this goal simply and effectively without affecting load performance.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-identity>

Question 226

CertyIQ

Overview

Contoso, Ltd. is a clothing retailer based in Seattle. The company has 2,000 retail stores across the United States and an emerging online presence.

The network contains an Active Directory forest named contoso.com. The forest is integrated with an Azure Active Directory (Azure AD) tenant named contoso.com. Contoso has an Azure subscription associated to the contoso.com Azure AD tenant.

Existing Environment -

Transactional Data -

Contoso has three years of customer, transactional, operational, sourcing, and supplier data comprised of 10 billion records stored across multiple on-premises

Microsoft SQL Server servers. The SQL Server instances contain data from various operational systems. The data is loaded into the instances by using SQL

Server Integration Services (SSIS) packages.

You estimate that combining all product sales transactions into a company-wide sales transactions dataset will result in a single table that contains 5 billion rows, with one row per transaction.

Most queries targeting the sales transactions data will be used to identify which products were sold in retail stores and which products were sold online during different time periods. Sales transaction data that is older than three years will be removed monthly.

You plan to create a retail store table that will contain the address of each retail store. The table will be approximately 2 MB. Queries for retail store sales will include the retail store addresses.

You plan to create a promotional table that will contain a promotion ID. The promotion ID will be associated to a specific product. The product will be identified by a product ID. The table will be approximately 5 GB.

Streaming Twitter Data -

The ecommerce department at Contoso develops an Azure logic app that captures trending Twitter feeds referencing the company's products and pushes the products to Azure Event Hubs.

Planned Changes and Requirements

Planned Changes -

Contoso plans to implement the following changes:

Load the sales transaction dataset to Azure Synapse Analytics.

Integrate on-premises data stores with Azure Synapse Analytics by using SSIS packages.

Use Azure Synapse Analytics to analyze Twitter feeds to assess customer sentiments about products.

Sales Transaction Dataset Requirements

Contoso identifies the following requirements for the sales transaction dataset:

Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month. Boundary values must belong to the partition on the right.

Ensure that queries joining and filtering sales transaction records based on product ID complete as quickly as possible.

Implement a surrogate key to account for changes to the retail store addresses.

Ensure that data storage costs and performance are predictable.

Minimize how long it takes to remove old records.

Customer Sentiment Analytics Requirements

Contoso identifies the following requirements for customer sentiment analytics:

Allow Contoso users to use PolyBase in an Azure Synapse Analytics dedicated SQL pool to query the content of the data records that host the Twitter feeds.

Data must be protected by using row-level security (RLS). The users must be authenticated by using their own Azure AD credentials.

Maximize the throughput of ingesting Twitter feeds from Event Hubs to Azure Storage without purchasing additional throughput or capacity units.

Store Twitter feeds in Azure Storage by using Event Hubs Capture. The feeds will be converted into Parquet files.

Ensure that the data store supports Azure AD-based access control down to the object level.

Minimize administrative effort to maintain the Twitter feed data records.

Purge Twitter feed data records that are older than two years.

Data Integration Requirements -

Contoso identifies the following requirements for data integration:

Use an Azure service that leverages the existing SSIS packages to ingest on-premises data into datasets stored in a dedicated SQL pool of Azure Synapse

Analytics and transform the data.

Identify a process to ensure that changes to the ingestion and transformation activities can be version-controlled and developed independently by multiple data engineers.

Question

HOTSPOT -

You need to design an analytical storage solution for the transactional data. The solution must meet the sales transaction dataset requirements.

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Table type to store retail store data:

Hash
Replicated
Round-robin

Table type to store promotional data:

Hash
Replicated
Round-robin

Answer Area

Table type to store retail store data:

Hash
Replicated
Round-robin

Suggested Answer:

Table type to store promotional data:

Hash
Replicated
Round-robin

Explanation:

Correct Answer:

Box 1: Replicated

Box 2: Hash -

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

Scenario:

☞ You plan to create a promotional table that will contain a promotion ID. The promotion ID will be associated to a specific product. The product will be identified by a product ID. The table will be approximately 5 GB.

☞ Ensure that queries joining and filtering sales transaction records based on product ID complete as quickly as possible.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute>

Question 227

CertyIQ

Overview

Contoso, Ltd. is a clothing retailer based in Seattle. The company has 2,000 retail stores across the United States and an emerging online presence.

The network contains an Active Directory forest named contoso.com. The forest is integrated with an Azure Active Directory (Azure AD) tenant named contoso.com. Contoso has an Azure subscription associated to the contoso.com Azure AD tenant.

Existing Environment -

Transactional Data -

Contoso has three years of customer, transactional, operational, sourcing, and supplier data comprised of 10 billion records stored across multiple on-premises

Microsoft SQL Server servers. The SQL Server instances contain data from various operational systems. The data is loaded into the instances by using SQL

Server Integration Services (SSIS) packages.

You estimate that combining all product sales transactions into a company-wide sales transactions dataset will result in a single table that contains 5 billion rows, with one row per transaction.

Most queries targeting the sales transactions data will be used to identify which products were sold in retail stores and which products were sold online during different time periods. Sales transaction data that is older than three years will be removed monthly.

You plan to create a retail store table that will contain the address of each retail store. The table will be approximately 2 MB. Queries for retail store sales will include the retail store addresses.

You plan to create a promotional table that will contain a promotion ID. The promotion ID will be associated to a specific product. The product will be identified by a product ID. The table will be approximately 5 GB.

Streaming Twitter Data -

The ecommerce department at Contoso develops an Azure logic app that captures trending Twitter feeds referencing the company's products and pushes the products to Azure Event Hubs.

Planned Changes and Requirements

Planned Changes -

Contoso plans to implement the following changes:

Load the sales transaction dataset to Azure Synapse Analytics.

Integrate on-premises data stores with Azure Synapse Analytics by using SSIS packages.

Use Azure Synapse Analytics to analyze Twitter feeds to assess customer sentiments about products.

Sales Transaction Dataset Requirements

Contoso identifies the following requirements for the sales transaction dataset:

Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month. Boundary values must belong to the partition on the right.

Ensure that queries joining and filtering sales transaction records based on product ID complete as quickly as possible.

Implement a surrogate key to account for changes to the retail store addresses.

Ensure that data storage costs and performance are predictable.

Minimize how long it takes to remove old records.

Customer Sentiment Analytics Requirements

Contoso identifies the following requirements for customer sentiment analytics:

Allow Contoso users to use PolyBase in an Azure Synapse Analytics dedicated SQL pool to query the content of the data records that host the Twitter feeds.

Data must be protected by using row-level security (RLS). The users must be authenticated by using their own Azure AD credentials.

Maximize the throughput of ingesting Twitter feeds from Event Hubs to Azure Storage without purchasing additional throughput or capacity units.

Store Twitter feeds in Azure Storage by using Event Hubs Capture. The feeds will be converted into Parquet files.

Ensure that the data store supports Azure AD-based access control down to the object level.

Minimize administrative effort to maintain the Twitter feed data records.

Purge Twitter feed data records that are older than two years.

Data Integration Requirements -

Contoso identifies the following requirements for data integration:

Use an Azure service that leverages the existing SSIS packages to ingest on-premises data into datasets stored in a dedicated SQL pool of Azure Synapse

Analytics and transform the data.

Identify a process to ensure that changes to the ingestion and transformation activities can be version-controlled and developed independently by multiple data engineers.

Question

HOTSPOT -

You need to implement an Azure Synapse Analytics database object for storing the sales transactions data. The solution must meet the sales transaction dataset requirements.

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Transact-SQL DDL command to use:

▼
CREATE EXTERNAL TABLE
CREATE TABLE
CREATE VIEW

Partitioning option to use in the WITH clause of the DDL statement:

▼
FORMAT_OPTIONS
FORMAT_TYPE
RANGE LEFT FOR VALUES
RANGE RIGHT FOR VALUES

Suggested Answer:

Answer Area

Transact-SQL DDL command to use:

▼
CREATE EXTERNAL TABLE
CREATE TABLE
CREATE VIEW

Partitioning option to use in the WITH clause of the DDL statement:

▼
FORMAT_OPTIONS
FORMAT_TYPE
RANGE LEFT FOR VALUES
RANGE RIGHT FOR VALUES

Explanation:

Correct Answer:

Box 1: Create table -

Scenario: Load the sales transaction dataset to Azure Synapse Analytics

Box 2: RANGE RIGHT FOR VALUES -

Scenario: Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month. Boundary values must belong to the partition on the right.

RANGE RIGHT: Specifies the boundary value belongs to the partition on the right (higher values).

FOR VALUES (boundary_value [,...n]): Specifies the boundary values for the partition.

Scenario: Load the sales transaction dataset to Azure Synapse Analytics.

Contoso identifies the following requirements for the sales transaction dataset:

⇒ Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month. Boundary values must belong to the partition on the right.

⇒ Ensure that queries joining and filtering sales transaction records based on product ID complete as quickly as possible.

⇒ Implement a surrogate key to account for changes to the retail store addresses.

⇒ Ensure that data storage costs and performance are predictable.

⇒ Minimize how long it takes to remove old records.

Reference:

<https://docs.microsoft.com/en-us/sql/t-sql/statements/create-table-azure-sql-data-warehouse>

Question 228

CertyIQ

Overview -

Contoso, Ltd. is a clothing retailer based in Seattle. The company has 2,000 retail stores across the United States and an emerging online presence.

The network contains an Active Directory forest named contoso.com. The forest is integrated with an Azure Active Directory (Azure AD) tenant named contoso.com. Contoso has an Azure subscription associated to the contoso.com Azure AD tenant.

Existing Environment -

Transactional Data -

Contoso has three years of customer, transactional, operational, sourcing, and supplier data comprised of 10 billion records stored across multiple on-premises

Microsoft SQL Server servers. The SQL Server instances contain data from various operational systems. The data is loaded into the instances by using SQL

Server Integration Services (SSIS) packages.

You estimate that combining all product sales transactions into a company-wide sales transactions dataset will result in a single table that contains 5 billion rows, with one row per transaction.

Most queries targeting the sales transactions data will be used to identify which products were sold in retail stores and which products were sold online during different time periods. Sales transaction data that is older than three years will be removed monthly.

You plan to create a retail store table that will contain the address of each retail store. The table will be approximately 2 MB. Queries for retail store sales will include the retail store addresses.

You plan to create a promotional table that will contain a promotion ID. The promotion ID will be associated to a specific product. The product will be identified by a product ID. The table will be approximately 5 GB.

Streaming Twitter Data -

The ecommerce department at Contoso develops an Azure logic app that captures trending Twitter feeds referencing the company's products and pushes the products to Azure Event Hubs.

Planned Changes and Requirements

Planned Changes -

Contoso plans to implement the following changes:

Load the sales transaction dataset to Azure Synapse Analytics.

Integrate on-premises data stores with Azure Synapse Analytics by using SSIS packages.

Use Azure Synapse Analytics to analyze Twitter feeds to assess customer sentiments about products.

Sales Transaction Dataset Requirements

Contoso identifies the following requirements for the sales transaction dataset:

Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month. Boundary values must belong to the partition on the right.

Ensure that queries joining and filtering sales transaction records based on product ID complete as quickly as possible.

Implement a surrogate key to account for changes to the retail store addresses.

Ensure that data storage costs and performance are predictable.

Minimize how long it takes to remove old records.

Customer Sentiment Analytics Requirements

Contoso identifies the following requirements for customer sentiment analytics:

Allow Contoso users to use PolyBase in an Azure Synapse Analytics dedicated SQL pool to query the content of the data records that host the Twitter feeds.

Data must be protected by using row-level security (RLS). The users must be authenticated by using their own Azure AD credentials.

Maximize the throughput of ingesting Twitter feeds from Event Hubs to Azure Storage without purchasing additional throughput or capacity units.

Store Twitter feeds in Azure Storage by using Event Hubs Capture. The feeds will be converted into Parquet files.

Ensure that the data store supports Azure AD-based access control down to the object level.

Minimize administrative effort to maintain the Twitter feed data records.

Purge Twitter feed data records that are older than two years.

Data Integration Requirements -

Contoso identifies the following requirements for data integration:

Use an Azure service that leverages the existing SSIS packages to ingest on-premises data into datasets stored in a dedicated SQL pool of Azure Synapse

Analytics and transform the data.

Identify a process to ensure that changes to the ingestion and transformation activities can be version-controlled and developed independently by multiple data engineers.

Question

You need to design a data retention solution for the Twitter feed data records. The solution must meet the customer sentiment analytics requirements.

Which Azure Storage functionality should you include in the solution?

- A. change feed
- B. soft delete
- C. time-based retention
- D. lifecycle management**

Explanation:

Correct Answer: D

Scenario: Purge Twitter feed data records that are older than two years.

Data sets have unique lifecycles. Early in the lifecycle, people access some data often. But the need for access often drops drastically as the data ages. Some data remains idle in the cloud and is rarely accessed once stored. Some data sets expire days or months after creation, while other data sets are actively read and modified throughout their lifetimes. Azure Storage lifecycle management offers a rule-based policy that you can use to transition blob data to the appropriate access tiers or to expire data at the end of the data lifecycle.

Reference:

<https://docs.microsoft.com/en-us/azure/storage/blobs/lifecycle-management-overview>

Question 229

CertyIQ

Case study -

This is a case study. Case studies are not timed separately. You can use as much exam time as you would like to complete each case. However, there may be additional case studies and sections on this exam. You must manage your time to ensure that you are able to complete all questions included on this exam in the time provided.

To answer the questions included in a case study, you will need to reference information that is provided in the case study. Case studies might contain exhibits and other resources that provide more information about the scenario that is described in the case study. Each question is independent of the other questions in this case study.

At the end of this case study, a review screen will appear. This screen allows you to review your answers and to make changes before you move to the next section of the exam. After you begin a new section, you cannot return to this section.

To start the case study -

To display the first question in this case study, click the Next button. Use the buttons in the left pane to explore the content of the case study before you answer the questions. Clicking these buttons displays information such as business requirements, existing environment, and problem statements. If the case study has an All Information tab, note that the information displayed is identical to the information displayed on the subsequent tabs. When you are ready to answer a question, click the Question button to return to the question.

Overview -

Litware, Inc. owns and operates 300 convenience stores across the US. The company sells a variety of packaged foods and drinks, as well as a variety of prepared foods, such as sandwiches and pizzas.

Litware has a loyalty club whereby members can get daily discounts on specific items by providing their membership number at checkout.

Litware employs business analysts who prefer to analyze data by using Microsoft Power BI, and data scientists who prefer analyzing data in Azure Databricks notebooks.

Requirements -

Business Goals -

Litware wants to create a new analytics environment in Azure to meet the following requirements:

See inventory levels across the stores. Data must be updated as close to real time as possible.

Execute ad hoc analytical queries on historical data to identify whether the loyalty club discounts increase sales of the discounted products.

Every four hours, notify store employees about how many prepared food items to produce based on historical demand from the sales data.

Technical Requirements -

Litware identifies the following technical requirements:

Minimize the number of different Azure services needed to achieve the business goals.

Use platform as a service (PaaS) offerings whenever possible and avoid having to provision virtual machines that must be managed by Litware.

Ensure that the analytical data store is accessible only to the company's on-premises network and Azure services.

Use Azure Active Directory (Azure AD) authentication whenever possible.

Use the principle of least privilege when designing security.

Stage Inventory data in Azure Data Lake Storage Gen2 before loading the data into the analytical data store. Litware wants to remove transient data from Data

Lake Storage once the data is no longer in use. Files that have a modified date that is older than 14 days must be removed.

Limit the business analysts' access to customer contact information, such as phone numbers, because this type of data is not analytically relevant.

Ensure that you can quickly restore a copy of the analytical data store within one hour in the event of corruption or accidental deletion.

Planned Environment -

Litware plans to implement the following environment:

The application development team will create an Azure event hub to receive real-time sales data, including store number, date, time, product ID, customer loyalty number, price, and discount amount, from the point of sale (POS) system and output the data to data storage in Azure.

Customer data, including name, contact information, and loyalty number, comes from Salesforce, a SaaS application, and can be imported into Azure once every eight hours. Row modified dates are not trusted in the source table.

Product data, including product ID, name, and category, comes from Salesforce and can be imported into Azure once every eight hours. Row modified dates are not trusted in the source table.

Daily inventory data comes from a Microsoft SQL server located on a private network.

Litware currently has 5 TB of historical sales data and 100 GB of customer data. The company expects approximately 100 GB of new data per month for the next year.

Litware will build a custom application named FoodPrep to provide store employees with the calculation results of how many prepared food items to produce every four hours.

Litware does not plan to implement Azure ExpressRoute or a VPN between the on-premises network and Azure.

Question

HOTSPOT -

Which Azure Data Factory components should you recommend using together to import the daily inventory data from the SQL server to Azure Data Lake Storage?

To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Integration runtime type:

Azure integration runtime
Azure-SSIS integration runtime
Self-hosted integration runtime

Trigger type:

Event-based trigger
Schedule trigger
Tumbling window trigger

Activity type:

Copy activity
Lookup activity
Stored procedure activity

Answer Area

Integration runtime type:

Azure integration runtime
Azure-SSIS integration runtime
Self-hosted integration runtime

Trigger type:

Event-based trigger
Schedule trigger
Tumbling window trigger

Activity type:

Copy activity
Lookup activity
Stored procedure activity

Suggested Answer:

Explanation:

Correct Answer:

Box 1: Self-hosted integration runtime

A self-hosted IR is capable of running copy activity between a cloud data stores and a data store in private network.

Box 2: Tumbling Window

Since Inventory data should be updated in real time as close as possible. Only Customer & Product data are available every 8 hours.

Box 3: Copy activity -

Scenario:

⇒ Customer data, including name, contact information, and loyalty number, comes from Salesforce and can be imported into Azure once every eight hours. Row modified dates are not trusted in the source table.

⇒ Product data, including product ID, name, and category, comes from Salesforce and can be imported into Azure once every eight hours. Row modified dates are not trusted in the source table.

Question 230

CertyIQ

Overview -

Contoso, Ltd. is a clothing retailer based in Seattle. The company has 2,000 retail stores across the United States and an emerging online presence.

The network contains an Active Directory forest named contoso.com. The forest is integrated with an Azure Active Directory (Azure AD) tenant named contoso.com. Contoso has an Azure subscription associated to the contoso.com Azure AD tenant.

Existing Environment -

Transactional Data -

Contoso has three years of customer, transactional, operational, sourcing, and supplier data comprised of 10 billion records stored across multiple on-premises

Microsoft SQL Server servers. The SQL Server instances contain data from various operational systems. The data is loaded into the instances by using SQL

Server Integration Services (SSIS) packages.

You estimate that combining all product sales transactions into a company-wide sales transactions dataset will result in a single table that contains 5 billion rows, with one row per transaction.

Most queries targeting the sales transactions data will be used to identify which products were sold in retail stores and which products were sold online during different time periods. Sales transaction data that is older than three years will be removed monthly.

You plan to create a retail store table that will contain the address of each retail store. The table will be approximately 2 MB. Queries for retail store sales will include the retail store addresses.

You plan to create a promotional table that will contain a promotion ID. The promotion ID will be associated to a specific product. The product will be identified by a product ID. The table will be approximately 5 GB.

Streaming Twitter Data -

The ecommerce department at Contoso develops an Azure logic app that captures trending Twitter feeds referencing the company's products and pushes the products to Azure Event Hubs.

Planned Changes and Requirements

Planned Changes -

Contoso plans to implement the following changes:

Load the sales transaction dataset to Azure Synapse Analytics.

Integrate on-premises data stores with Azure Synapse Analytics by using SSIS packages.

Use Azure Synapse Analytics to analyze Twitter feeds to assess customer sentiments about products.

Sales Transaction Dataset Requirements

Contoso identifies the following requirements for the sales transaction dataset:

Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month. Boundary values must belong to the partition on the right.

Ensure that queries joining and filtering sales transaction records based on product ID complete as quickly as possible.

Implement a surrogate key to account for changes to the retail store addresses.

Ensure that data storage costs and performance are predictable.

Minimize how long it takes to remove old records.

Customer Sentiment Analytics Requirements

Contoso identifies the following requirements for customer sentiment analytics:

Allow Contoso users to use PolyBase in an Azure Synapse Analytics dedicated SQL pool to query the content of the data records that host the Twitter feeds.

Data must be protected by using row-level security (RLS). The users must be authenticated by using their own Azure AD credentials.

Maximize the throughput of ingesting Twitter feeds from Event Hubs to Azure Storage without purchasing additional throughput or capacity units.

Store Twitter feeds in Azure Storage by using Event Hubs Capture. The feeds will be converted into Parquet files.

Ensure that the data store supports Azure AD-based access control down to the object level.

Minimize administrative effort to maintain the Twitter feed data records.

Purge Twitter feed data records that are older than two years.

Data Integration Requirements -

Contoso identifies the following requirements for data integration:

Use an Azure service that leverages the existing SSIS packages to ingest on-premises data into datasets stored in a dedicated SQL pool of Azure Synapse

Analytics and transform the data.

Identify a process to ensure that changes to the ingestion and transformation activities can be version-controlled and developed independently by multiple data engineers.

Question

DRAG DROP -

You need to implement versioned changes to the integration pipelines. The solution must meet the data integration requirements.

In which order should you perform the actions? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

Select and Place:

Actions	Answer Area
Merge changes	
Create a pull request	
Create a feature branch	
Publish changes	
Create a repository and a main branch	

Suggested Answer:

Actions	Answer Area
Merge changes	Create a repository and a main branch
Create a pull request	Create a feature branch
Create a feature branch	Create a pull request
Publish changes	Merge changes
Create a repository and a main branch	Publish changes

Explanation:

Correct Answer:

Scenario: Identify a process to ensure that changes to the ingestion and transformation activities can be version-controlled and developed independently by multiple data engineers.

Step 1: Create a repository and a main branch

You need a Git repository in Azure Pipelines, TFS, or GitHub with your app.

Step 2: Create a feature branch -

Step 3: Create a pull request -

Step 4: Merge changes -

Merge feature branches into the main branch using pull requests.

Step 5: Publish changes -

Reference:

<https://docs.microsoft.com/en-us/azure/devops/pipelines/repos/pipeline-options-for-git>

Question 231

CertyIQ

Litware, Inc. owns and operates 300 convenience stores across the US. The company sells a variety of packaged foods and drinks, as well as a variety of prepared foods, such as sandwiches and pizzas.

Litware has a loyalty club whereby members can get daily discounts on specific items by providing their membership number at checkout.

Litware employs business analysts who prefer to analyze data by using Microsoft Power BI, and data scientists who prefer analyzing data in Azure Databricks notebooks.

Requirements

Business Goals

Litware wants to create a new analytics environment in Azure to meet the following requirements:

See inventory levels across the stores. Data must be updated as close to real time as possible.

Execute ad hoc analytical queries on historical data to identify whether the loyalty club discounts increase sales of the discounted products.

Every four hours, notify store employees about how many prepared food items to produce based on historical demand from the sales data.

Technical Requirements

Litware identifies the following technical requirements:

Minimize the number of different Azure services needed to achieve the business goals.

Use platform as a service (PaaS) offerings whenever possible and avoid having to provision virtual machines that must be managed by Litware.

Ensure that the analytical data store is accessible only to the company's on-premises network and Azure services.

Use Azure Active Directory (Azure AD) authentication whenever possible.

Use the principle of least privilege when designing security.

Stage Inventory data in Azure Data Lake Storage Gen2 before loading the data into the analytical data store. Litware wants to remove transient data from Data

Lake Storage once the data is no longer in use. Files that have a modified date that is older than 14 days must be removed.

Limit the business analysts' access to customer contact information, such as phone numbers, because this type of data is not analytically relevant.

Ensure that you can quickly restore a copy of the analytical data store within one hour in the event of corruption or accidental deletion.

Planned Environment

Litware plans to implement the following environment:

The application development team will create an Azure event hub to receive real-time sales data, including store number, date, time, product ID, customer loyalty number, price, and discount amount, from the point of sale (POS) system and output the data to data storage in Azure.

Customer data, including name, contact information, and loyalty number, comes from Salesforce, a SaaS application, and can be imported into Azure once every eight hours. Row modified dates are not trusted in the source table.

Product data, including product ID, name, and category, comes from Salesforce and can be imported into Azure once every eight hours. Row modified dates are not trusted in the source table.

Daily inventory data comes from a Microsoft SQL server located on a private network.

Litware currently has 5 TB of historical sales data and 100 GB of customer data. The company expects approximately 100 GB of new data per month for the next year.

Litware will build a custom application named FoodPrep to provide store employees with the calculation results of how many prepared food items to produce every four hours.

Litware does not plan to implement Azure ExpressRoute or a VPN between the on-premises network and Azure.

Question

What should you recommend using to secure sensitive customer contact information?

- A. Transparent Data Encryption (TDE)
- B. row-level security
- C. column-level security**
- D. data sensitivity labels

Explanation:

Correct Answer: C

Scenario: Limit the business analysts' access to customer contact information, such as phone numbers, because this type of data is not analytically relevant.

<https://azure.microsoft.com/en-ca/updates/column-level-security-is-now-supported-in-azure-sql-data-warehouse/>

You can use CLS to manage user access to specific columns in your tables in a simpler manner, without having to redesign your data warehouse. CLS eliminates the need to maintain access restriction logic away from the data in another application or introduce views for filtering out sensitive columns for a subset of users.

Question 232

CertyIQ

Litware, Inc. owns and operates 300 convenience stores across the US. The company sells a variety of packaged foods and drinks, as well as a variety of prepared foods, such as sandwiches and pizzas.

Litware has a loyalty club whereby members can get daily discounts on specific items by providing their membership number at checkout.

Litware employs business analysts who prefer to analyze data by using Microsoft Power BI, and data scientists who prefer analyzing data in Azure Databricks notebooks.

Requirements

Business Goals

Litware wants to create a new analytics environment in Azure to meet the following requirements:

See inventory levels across the stores. Data must be updated as close to real time as possible.

Execute ad hoc analytical queries on historical data to identify whether the loyalty club discounts increase sales of the discounted products.

Every four hours, notify store employees about how many prepared food items to produce based on historical demand from the sales data.

Technical Requirements

Litware identifies the following technical requirements:

Minimize the number of different Azure services needed to achieve the business goals.

Use platform as a service (PaaS) offerings whenever possible and avoid having to provision virtual machines that must be managed by Litware.

Ensure that the analytical data store is accessible only to the company's on-premises network and Azure services.

Use Azure Active Directory (Azure AD) authentication whenever possible.

Use the principle of least privilege when designing security.

Stage Inventory data in Azure Data Lake Storage Gen2 before loading the data into the analytical data store.
Litware wants to remove transient data from Data

Lake Storage once the data is no longer in use. Files that have a modified date that is older than 14 days must be removed.

Limit the business analysts' access to customer contact information, such as phone numbers, because this type of data is not analytically relevant.

Ensure that you can quickly restore a copy of the analytical data store within one hour in the event of corruption or accidental deletion.

Planned Environment

Litware plans to implement the following environment:

The application development team will create an Azure event hub to receive real-time sales data, including store number, date, time, product ID, customer loyalty number, price, and discount amount, from the point of sale (POS) system and output the data to data storage in Azure.

Customer data, including name, contact information, and loyalty number, comes from Salesforce, a SaaS application, and can be imported into Azure once every eight hours. Row modified dates are not trusted in the source table.

Product data, including product ID, name, and category, comes from Salesforce and can be imported into Azure once every eight hours. Row modified dates are not trusted in the source table.

Daily inventory data comes from a Microsoft SQL server located on a private network.

Litware currently has 5 TB of historical sales data and 100 GB of customer data. The company expects approximately 100 GB of new data per month for the next year.

Litware will build a custom application named FoodPrep to provide store employees with the calculation results of how many prepared food items to produce every four hours.

Litware does not plan to implement Azure ExpressRoute or a VPN between the on-premises network and Azure.

Question

What should you do to improve high availability of the real-time data processing solution?

- A. Deploy a High Concurrency Databricks cluster.
- B. Deploy an Azure Stream Analytics job and use an Azure Automation runbook to check the status of the job and to start the job if it stops.
- C. Set Data Lake Storage to use geo-redundant storage (GRS).
- D. Deploy identical Azure Stream Analytics jobs to paired regions in Azure.**

Explanation:

Correct Answer: D

Guarantee Stream Analytics job reliability during service updates

Part of being a fully managed service is the capability to introduce new service functionality and improvements at a rapid pace. As a result, Stream Analytics can have a service update deploy on a weekly (or more frequent) basis. No matter how much testing is done there is still a risk that an existing, running job may break due to the introduction of a bug. If you are running mission critical jobs, these risks need to be avoided. You can reduce this risk by following Azure's paired region model.

Scenario: The application development team will create an Azure event hub to receive real-time sales data, including store number, date, time, product ID, customer loyalty number, price, and discount amount, from the point of sale (POS) system and output the data to data storage in Azure

Reference:

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-job-reliability>

Question 233

CertyIQ

Litware, Inc. owns and operates 300 convenience stores across the US. The company sells a variety of packaged foods and drinks, as well as a variety of prepared foods, such as sandwiches and pizzas.

Litware has a loyalty club whereby members can get daily discounts on specific items by providing their membership number at checkout.

Litware employs business analysts who prefer to analyze data by using Microsoft Power BI, and data scientists who prefer analyzing data in Azure Databricks notebooks.

Requirements

Business Goals

Litware wants to create a new analytics environment in Azure to meet the following requirements:

See inventory levels across the stores. Data must be updated as close to real time as possible.

Execute ad hoc analytical queries on historical data to identify whether the loyalty club discounts increase sales of the discounted products.

Every four hours, notify store employees about how many prepared food items to produce based on historical demand from the sales data.

Technical Requirements

Litware identifies the following technical requirements:

Minimize the number of different Azure services needed to achieve the business goals.

Use platform as a service (PaaS) offerings whenever possible and avoid having to provision virtual machines that must be managed by Litware.

Ensure that the analytical data store is accessible only to the company's on-premises network and Azure services.

Use Azure Active Directory (Azure AD) authentication whenever possible.

Use the principle of least privilege when designing security.

Stage Inventory data in Azure Data Lake Storage Gen2 before loading the data into the analytical data store. Litware wants to remove transient data from Data

Lake Storage once the data is no longer in use. Files that have a modified date that is older than 14 days must be removed.

Limit the business analysts' access to customer contact information, such as phone numbers, because this type of data is not analytically relevant.

Ensure that you can quickly restore a copy of the analytical data store within one hour in the event of corruption or accidental deletion.

Planned Environment

Litware plans to implement the following environment:

The application development team will create an Azure event hub to receive real-time sales data, including store number, date, time, product ID, customer loyalty number, price, and discount amount, from the point of sale (POS) system and output the data to data storage in Azure.

Customer data, including name, contact information, and loyalty number, comes from Salesforce, a SaaS application, and can be imported into Azure once every eight hours. Row modified dates are not trusted in the source table.

Product data, including product ID, name, and category, comes from Salesforce and can be imported into Azure once every eight hours. Row modified dates are not trusted in the source table.

Daily inventory data comes from a Microsoft SQL server located on a private network.

Litware currently has 5 TB of historical sales data and 100 GB of customer data. The company expects approximately 100 GB of new data per month for the next year.

Litware will build a custom application named FoodPrep to provide store employees with the calculation results of how many prepared food items to produce every four hours.

Litware does not plan to implement Azure ExpressRoute or a VPN between the on-premises network and Azure.

What should you recommend to prevent users outside the Litware on-premises network from accessing the analytical data st

- A. a server-level firewall IP rule
- B. a database-level virtual network rule
- C. a server-level virtual network rule
- D. a database-level firewall IP rule

Explanation:

Correct Answer: A

The company doesn't want any virtual network stuff and server-level is more comprehensive, thus safer than just database-level rule.

Since there is no VPN between on-premises machines and Azure SQL server, communications use a public endpoint. You can limit the public access to databases through a Server Level IP Firewall rules.

Reference:

<https://docs.microsoft.com/en-us/azure/azure-sql/database/network-access-controls-overview>

Question 234

CertyIQ

Contoso, Ltd. is a clothing retailer based in Seattle. The company has 2,000 retail stores across the United States and an emerging online presence.

The network contains an Active Directory forest named contoso.com. The forest it integrated with an Azure Active Directory (Azure AD) tenant named contoso.com. Contoso has an Azure subscription associated to the contoso.com Azure AD tenant.

Existing Environment

Transactional Data

Contoso has three years of customer, transactional, operational, sourcing, and supplier data comprised of 10 billion records stored across multiple on-premises Microsoft SQL Server servers. The SQL Server instances contain data from various operational systems. The data is loaded into the instances by using SQL Server Integration Services (SSIS) packages.

You estimate that combining all product sales transactions into a company-wide sales transactions dataset will result in a single table that contains 5 billion rows, with one row per transaction.

Most queries targeting the sales transactions data will be used to identify which products were sold in retail stores and which products were sold online during different time periods. Sales transaction data that is older than three years will be removed monthly.

You plan to create a retail store table that will contain the address of each retail store. The table will be approximately 2 MB. Queries for retail store sales will include the retail store addresses.

You plan to create a promotional table that will contain a promotion ID. The promotion ID will be associated to a specific product. The product will be identified by a product ID. The table will be approximately 5 GB.

Streaming Twitter Data

The ecommerce department at Contoso develops an Azure logic app that captures trending Twitter feeds referencing the company's products and pushes the products to Azure Event Hubs.

Planned Changes and Requirements

Planned Changes

Contoso plans to implement the following changes:

Load the sales transaction dataset to Azure Synapse Analytics.

Integrate on-premises data stores with Azure Synapse Analytics by using SSIS packages.

Use Azure Synapse Analytics to analyze Twitter feeds to assess customer sentiments about products.

Sales Transaction Dataset Requirements

Contoso identifies the following requirements for the sales transaction dataset:

Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month. Boundary values must belong to the partition on the right.

Ensure that queries joining and filtering sales transaction records based on product ID complete as quickly as possible.

Implement a surrogate key to account for changes to the retail store addresses.

Ensure that data storage costs and performance are predictable.

Minimize how long it takes to remove old records.

Customer Sentiment Analytics Requirements

Contoso identifies the following requirements for customer sentiment analytics:

- Allow Contoso users to use PolyBase in an Azure Synapse Analytics dedicated SQL pool to query the content of the data records that host the Twitter feeds.
- Data must be protected by using row-level security (RLS). The users must be authenticated by using their own Azure AD credentials.
- Maximize the throughput of ingesting Twitter feeds from Event Hubs to Azure Storage without purchasing additional throughput or capacity units.
- Store Twitter feeds in Azure Storage by using Event Hubs Capture. The feeds will be converted into Parquet files.
- Ensure that the data store supports Azure AD-based access control down to the object level.
- Minimize administrative effort to maintain the Twitter feed data records.
- Purge Twitter feed data records that are older than two years.

Data Integration Requirements

Contoso identifies the following requirements for data integration:

Use an Azure service that leverages the existing SSIS packages to ingest on-premises data into datasets stored in a dedicated SQL pool of Azure Synapse Analytics and transform the data.

Identify a process to ensure that changes to the ingestion and transformation activities can be version-controlled and developed independently by multiple data engineers.

Question:

HOTSPOT -

You need to design a data ingestion and storage solution for the Twitter feeds. The solution must meet the customer sentiment analytics requirements.

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

To increase the throughput of ingesting the Twitter feeds:

Configure Event Hubs partitions.
Enable Auto-Inflate in Event Hubs.
Use Event Hubs Dedicated.

To store the Twitter feed data, use:

An Azure Data Lake Storage Gen2 account
An Azure Databricks high concurrency cluster
An Azure General-purpose v2 storage account in the Premium tier

Answer Area

To increase the throughput of ingesting the Twitter feeds:

Configure Event Hubs partitions.
Enable Auto-Inflate in Event Hubs.
Use Event Hubs Dedicated.

To store the Twitter feed data, use:

An Azure Data Lake Storage Gen2 account
An Azure Databricks high concurrency cluster
An Azure General-purpose v2 storage account in the Premium tier

Explanation:

Correct Answer:

Box 1: Configure Event Hubs partitions

Scenario: Maximize the throughput of ingesting Twitter feeds from Event Hubs to Azure Storage without purchasing additional throughput or capacity units.

Event Hubs is designed to help with processing of large volumes of events. Event Hubs throughput is scaled by using partitions and throughput-unit allocations.

Incorrect Answers:

☞ Event Hubs Dedicated: Event Hubs clusters offer single-tenant deployments for customers with the most demanding streaming needs. This single-tenant offering has a guaranteed 99.99% SLA and is available only on our Dedicated pricing tier.

☞ Auto-Inflate: The Auto-inflate feature of Event Hubs automatically scales up by increasing the number of TUs, to meet usage needs.

Event Hubs traffic is controlled by TUs (standard tier). Auto-inflate enables you to start small with the minimum required TUs you choose. The feature then scales automatically to the maximum limit of TUs you need, depending on the increase in your traffic.

Box 2: An Azure Data Lake Storage Gen2 account

Scenario: Ensure that the data store supports Azure AD-based access control down to the object level.

Azure Data Lake Storage Gen2 implements an access control model that supports both Azure role-based access control (Azure RBAC) and POSIX-like access control lists (ACLs).

Incorrect Answers:

☞ Azure Databricks: An Azure administrator with the proper permissions can configure Azure Active Directory conditional access to control where and when users are permitted to sign in to Azure Databricks.

☞ Azure Storage supports using Azure Active Directory (Azure AD) to authorize requests to blob data.

You can scope access to Azure blob resources at the following levels, beginning with the narrowest scope:

- An individual container. At this scope, a role assignment applies to all of the blobs in the container, as well as container properties and metadata.

- The storage account. At this scope, a role assignment applies to all containers and their blobs.

- The resource group. At this scope, a role assignment applies to all of the containers in all of the storage accounts in the resource group.

- The subscription. At this scope, a role assignment applies to all of the containers in all of the storage accounts in all of the resource groups in the subscription.

- A management group.

Reference:

<https://docs.microsoft.com/en-us/azure/event-hubs/event-hubs-features>

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

End of Part 6



Please find the videos of this **AZ-900/AI-900/AZ-305/
AZ-104 /DP-900/ SC-900 and other Microsoft exam
series on**

CertyIQ Official YouTube channel (FREE PDFs): -

Please **Subscribe** to CertyIQ YouTube Channel to get notified for latest exam dumps by clicking on the below image, it will redirect to the **CertyIQ** YouTube page.



Connect with us @ [LinkedIn](#) [Telegram](#)

Contact us for other dumps: - certyiq@gmail.com

For any other enquiry, please drop us a mail at
certyiqofficial@gmail.com