

Microsoft

Exam Questions DP-203

Data Engineering on Microsoft Azure



NEW QUESTION 1

- (Exam Topic 1)

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.

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.

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

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.

NEW QUESTION 2

- (Exam Topic 1)

You need to integrate the on-premises data sources and Azure Synapse Analytics. The solution must meet the data integration requirements. Which type of integration runtime should you use?

- A. Azure-SSIS integration runtime
- B. self-hosted integration runtime
- C. Azure integration runtime

Answer: C

NEW QUESTION 3

- (Exam Topic 1)

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 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.

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

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

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

NEW QUESTION 4

- (Exam Topic 2)

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.

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

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

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: Schedule trigger

Schedule 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.

NEW QUESTION 5

- (Exam Topic 2)

What should you do to improve high availability of the real-time data processing solution?

- A. Deploy identical Azure Stream Analytics jobs to paired regions in Azure.
- B. Deploy a High Concurrency Databricks cluster.
- C. 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.
- D. Set Data Lake Storage to use geo-redundant storage (GRS).

Answer: A

Explanation:

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>

NEW QUESTION 6

- (Exam Topic 3)

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 indexes

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.

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).

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

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-partition>

NEW QUESTION 7

- (Exam Topic 3)

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

Table	Distribution type	Distribution column
Sales:	<div><div></div><div>Hash-distributed</div><div>Round-robin</div></div>	<div><div></div><div>DateKey</div><div>ProductKey</div><div>RegionKey</div></div>
Invoices:	<div><div></div><div>Hash-distributed</div><div>Round-robin</div></div>	<div><div></div><div>DateKey</div><div>ProductKey</div><div>RegionKey</div></div>

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Box 1: Hash-distributed
Box 2: ProductKey
ProductKey is used extensively in joins.
Hash-distributed tables improve query performance on large fact tables.
Box 3: Round-robin
Box 4: RegionKey
Round-robin tables are useful for improving loading speed.
Consider using the round-robin distribution for your table in the following scenarios:

- > When getting started as a simple starting point since it is the default
- > If there is no obvious joining key
- > If there is not good candidate column for hash distributing the table
- > If the table does not share a common join key with other tables
- > If the join is less significant than other joins in the query
- > When the table is a temporary staging table

Note: A distributed table appears as a single table, but the rows are actually stored across 60 distributions. The rows are distributed with a hash or round-robin algorithm.
Reference:
<https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-tables-distribute>

NEW QUESTION 8

- (Exam Topic 3)
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 line 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. foreign key
- C. effective start date
- D. effective end date
- E. last modified date
- F. business key

Answer: BCF

NEW QUESTION 9

- (Exam Topic 3)

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

Answer: B

Explanation:

Need a High Concurrency cluster for the jobs.

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>

NEW QUESTION 10

- (Exam Topic 3)

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 HDInsight

- B. Azure Databricks
- C. Azure Machine Learning
- D. Azure Data Factory
- E. Azure Synapse Analytics

Answer: CE

NEW QUESTION 10

- (Exam Topic 3)

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.

Answer: C

Explanation:

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.

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>

NEW QUESTION 11

- (Exam Topic 3)

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.

Copy behavior:

	▼
Flatten hierarchy	
Merge files	
Preserve hierarchy	

Sink file type:

	▼
CSV	
JSON	
Parquet	
TXT	

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Box 1: Preserver herarchy

Compared to the flat namespace on Blob storage, the hierarchical namespace greatly improves the performance of directory management operations, which improves overall job performance.

Box 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>

NEW QUESTION 15

- (Exam Topic 3)

You have an Azure Data Lake Storage account that has a virtual network service endpoint configured.

You plan to use Azure Data Factory to extract data from the Data Lake Storage account. The data will then be loaded to a data warehouse in Azure Synapse Analytics by using PolyBase.

Which authentication method should you use to access Data Lake Storage?

- A. shared access key authentication
- B. managed identity authentication
- C. account key authentication
- D. service principal authentication

Answer: B

Explanation:

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/connector-azure-sql-data-warehouse#use-polybase-to-load-d>

NEW QUESTION 20

- (Exam Topic 3)

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.

Answer: ADF

Explanation:

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

Note: Managed identity for Azure resources is a feature of Azure Active Directory. The feature provides Azure services with an automatically managed identity in Azure AD Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/security/synapse-workspace-managed-identity>

NEW QUESTION 23

- (Exam Topic 3)

What should you recommend using to secure sensitive customer contact information?

- A. data labels
- B. column-level security
- C. row-level security
- D. Transparent Data Encryption (TDE)

Answer: B

Explanation:

Scenario: All cloud data must be encrypted at rest and in transit.

Always Encrypted is a feature designed to protect sensitive data stored in specific database columns from access (for example, credit card numbers, national identification numbers, or data on a need to know basis). This includes database administrators or other privileged users who are authorized to access the database to perform management tasks, but have no business need to access the particular data in the encrypted columns. The data is always encrypted, which means the encrypted data is decrypted only for processing by client applications with access to the encryption key.

References:

<https://docs.microsoft.com/en-us/azure/sql-database/sql-database-security-overview>

NEW QUESTION 25

- (Exam Topic 3)

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.

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

- A. Mastered
B. Not Mastered

Answer: A

Explanation:

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>

NEW QUESTION 29

- (Exam Topic 3)
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.

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

- A. Mastered
 B. Not Mastered

Answer: A

Explanation:

Box 1: Azure Stream Analytics Box 2: Hopping

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.

Box 3: Point within polygon Reference:

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-window-functions>

NEW QUESTION 33

- (Exam Topic 3)

You are designing a 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

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

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Reference:

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

NEW QUESTION 34

- (Exam Topic 3)

You have a C# application that process data from an Azure IoT hub and performs complex transformations. You need to replace the application with a real-time solution. The solution must reuse as much code as possible from the existing application.

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

Answer: C

Explanation:

Azure Stream Analytics on IoT Edge empowers developers to deploy near-real-time analytical intelligence closer to IoT devices so that they can unlock the full value of device-generated data. UDF are available in C# for IoT Edge jobs

Azure Stream Analytics on IoT Edge runs within the Azure IoT Edge framework. Once the job is created in Stream Analytics, you can deploy and manage it using IoT Hub.

References:

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-edge>

NEW QUESTION 36

- (Exam Topic 3)

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.

Does this meet the goal?

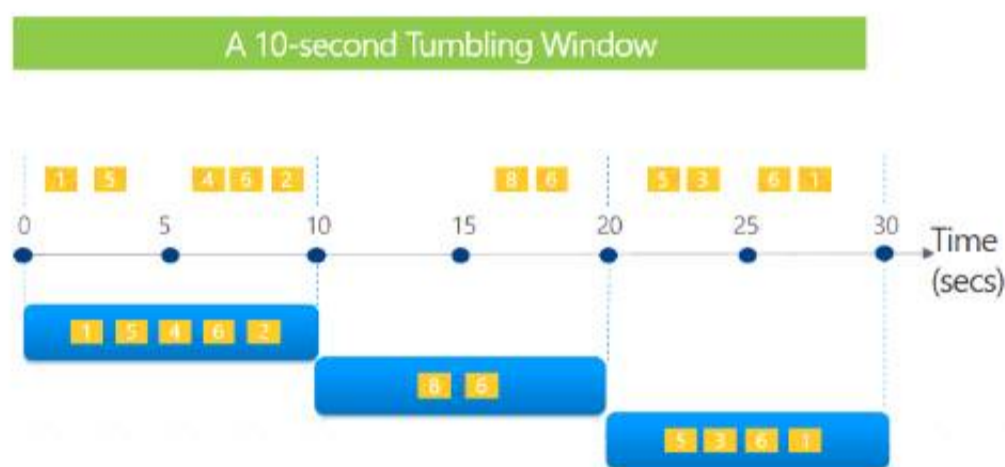
- A. Yes
- B. No

Answer: A

Explanation:

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>

NEW QUESTION 41

- (Exam Topic 3)

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.

Answer: B

Explanation:

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:

NEW QUESTION 45

- (Exam Topic 3)

You have an Azure subscription that contains the following resources:

* An Azure Active Directory (Azure AD) tenant that contains a security group named Group1.

* An Azure Synapse Analytics SQL pool named Pool1.

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 appropriate options in the answer 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

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

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

NEW QUESTION 46

- (Exam Topic 3)

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_datareader

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 AS 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.

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

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

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

NEW QUESTION 48

- (Exam Topic 3)

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

Answer: B

Explanation:

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>

NEW QUESTION 53

- (Exam Topic 3)

You plan to monitor an Azure data factory by using the Monitor & Manage app.

You need to identify the status and duration of activities that reference a table in a source database.

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

Actions

From the Data Factory monitoring app, add the Source user property to the Activity Runs table.

From the Data Factory monitoring app, add the Source user property to the Pipeline Runs table.

From the Data Factory authoring UI, publish the pipelines.

From the Data Factory monitoring app, add a linked service to the Pipeline Runs table.

From the Data Factory authoring UI, generate a user property for Source on all activities.

From the Data Factory authoring UI, generate a user property for Source on all datasets.

Answer Area

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Step 1: From the Data Factory authoring UI, generate a user property for Source on all activities. Step 2: From the Data Factory monitoring app, add the Source user property to Activity Runs table.

You can promote any pipeline activity property as a user property so that it becomes an entity that you can monitor. For example, you can promote the Source and Destination properties of the copy activity in your pipeline as user properties. You can also select Auto Generate to generate the Source and Destination user properties for a copy activity.

Step 3: From the Data Factory authoring UI, publish the pipelines

Publish output data to data stores such as Azure SQL Data Warehouse for business intelligence (BI) applications to consume.

References:

<https://docs.microsoft.com/en-us/azure/data-factory/monitor-visually>

NEW QUESTION 55

- (Exam Topic 3)

You have an Azure Synapse Analytics serverless SQL pool named Pool1 and an Azure Data Lake Storage Gen2 account named storage1. The AllowedBlobpublicAccess porperty 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 storage1 from Pool1.

What should you create first?

- A. an external resource pool
- B. a remote service binding
- C. database scoped credentials
- D. an external library

Answer: C

NEW QUESTION 59

- (Exam Topic 3)

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 SOL.
- >

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

Answer: A

Explanation:

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>

NEW QUESTION 64

- (Exam Topic 3)

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

Answer: AC

Explanation:

A: You can classify columns manually, as an alternative or in addition to the recommendation-based classification:

The screenshot shows the 'MySampleDatabase2 (mydocsamplesqlserver/MySampleDatabase2) | Data Discovery & Classification' interface. The 'Classification' tab is active, displaying '15 columns with classification recommendations'. The table below lists columns with checkboxes for selection:

	Schema	Table	Column
<input type="checkbox"/>	SalesLT	Customer	FirstName
<input type="checkbox"/>	SalesLT	Customer	LastName
<input type="checkbox"/>	SalesLT	Customer	EmailAddress
<input type="checkbox"/>	SalesLT	Customer	Phone
<input type="checkbox"/>	SalesLT	Customer	PasswordHash
<input type="checkbox"/>	SalesLT	Customer	PasswordSalt
<input type="checkbox"/>	dbo	ErrorLog	UserName
<input type="checkbox"/>	SalesLT	Address	AddressLine1
<input type="checkbox"/>	SalesLT	Address	AddressLine2
<input type="checkbox"/>	SalesLT	Address	City
<input type="checkbox"/>	SalesLT	Address	PostalCode
<input type="checkbox"/>	SalesLT	CustomerAddress	AddressType
<input type="checkbox"/>	SalesLT	SalesOrderHeader	AccountNumber
<input type="checkbox"/>	SalesLT	SalesOrderHeader	CreditCardApprovalCode
<input type="checkbox"/>	SalesLT	SalesOrderHeader	TaxAmt

- Select Add classification in the top menu of the pane.
- In the context window that opens, select the schema, table, and column that you want to classify, and the information type and sensitivity label.
- 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

NEW QUESTION 66

- (Exam Topic 3)
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

Answer: D

Explanation:
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

NEW QUESTION 67

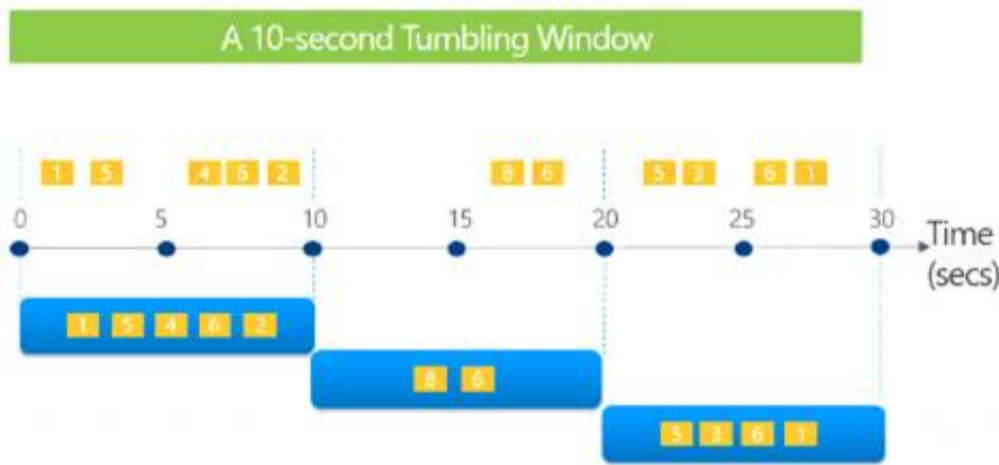
- (Exam Topic 3)
You are designing an Azure Stream Analytics job to process incoming events from sensors in retail environments.
You need to process the events to produce a running average of shopper counts during the previous 15 minutes, calculated at five-minute intervals.
Which type of window should you use?

- A. snapshot
- B. tumbling
- C. hopping
- D. sliding

Answer: B

Explanation:
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

NEW QUESTION 71

- (Exam Topic 3)
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
49488	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

Answer: B

Explanation:

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.

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>

NEW QUESTION 75

- (Exam Topic 3)

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.

Actions

Answer Area

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.

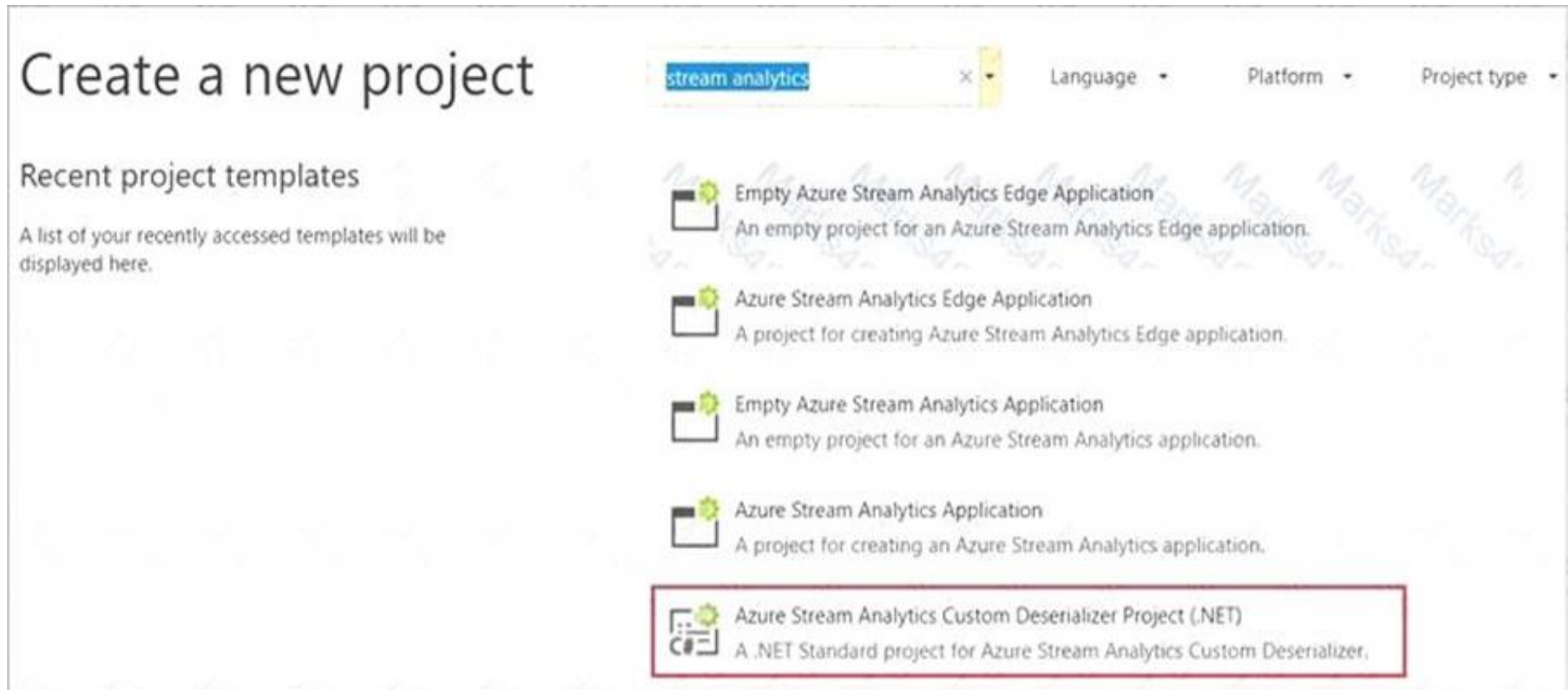
- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

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.



* 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 Protocol Buffer, Bond and other user defined formats for both cloud and edge jobs.

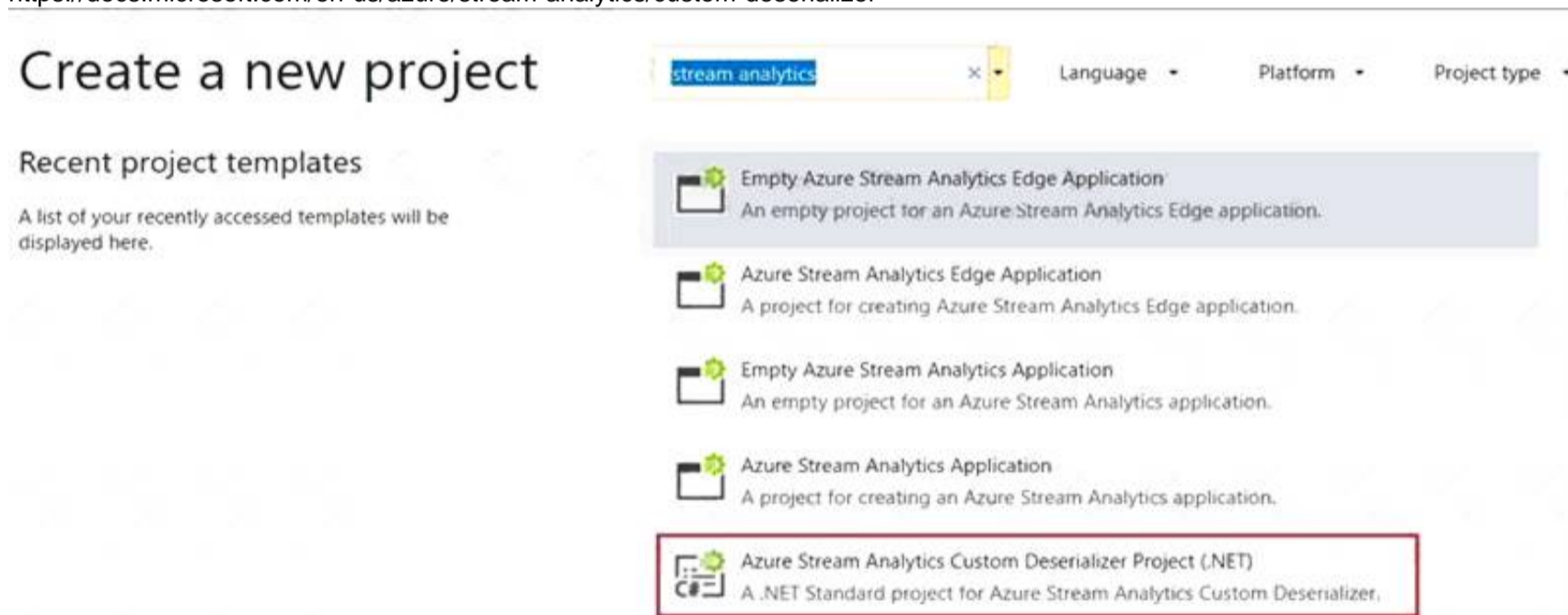
Step 3: Add an Azure Stream Analytics Application project to the solution Add an Azure Stream Analytics project

> In Solution Explorer, right-click the Protobuf Deserializer solution and select Add > New Project. Under Azure Stream Analytics > Stream Analytics, choose Azure Stream Analytics Application. Name it ProtobufCloudDeserializer and select OK.

> Right-click References under the ProtobufCloudDeserializer Azure Stream Analytics project. Under Projects, add Protobuf Deserializer. It should be automatically populated for you.

Reference:

<https://docs.microsoft.com/en-us/azure/stream-analytics/custom-deserializer>



NEW QUESTION 78

- (Exam Topic 3)

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 scenario, 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 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 Azure SQL data warehouse. 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

Answer: A

Explanation:

When exporting data into an ORC File Format, you might get Java out-of-memory errors when there are large text columns. To work around this limitation, export only a subset of the columns.

References:

<https://docs.microsoft.com/en-us/azure/sql-data-warehouse/guidance-for-loading-data>

NEW QUESTION 82

- (Exam Topic 3)

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.

Columnar format:

▼

Avro

GZip

Parquet

TXT

JSON with a timestamp:

▼

Avro

GZip

Parquet

TXT

- A. Mastered
- B. Not Mastered

Answer: A

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).

- 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>

NEW QUESTION 85

- (Exam Topic 3)

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.

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

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Answer Area

Integration runtime type:	<div><div>Azure integration runtime</div><div>Azure-SSIS integration runtime</div><div>Self-hosted integration runtime</div></div>
Trigger type:	<div><div>Event-based trigger</div><div>Schedule trigger</div><div>Tumbling window trigger</div></div>
Activity type:	<div><div>Copy activity</div><div>Lookup activity</div><div>Stored procedure activity</div></div>

NEW QUESTION 90

- (Exam Topic 3)

You use Azure Data Lake Storage Gen2.

You need to ensure that workloads can use filter predicates and column projections to filter data at the time the data is read from disk.

Which two actions should you perform? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

- A. Reregister the Microsoft Data Lake Store resource provider.
- B. Reregister the Azure Storage resource provider.
- C. Create a storage policy that is scoped to a container.
- D. Register the query acceleration feature.
- E. Create a storage policy that is scoped to a container prefix filter.

Answer: BD

NEW QUESTION 95

- (Exam Topic 3)

You are designing a sales transactions table in an Azure Synapse Analytics dedicated SQL pool. The table will contains approximately 60 million rows per month and will be partitioned by month. The table will use a clustered column store index and round-robin distribution.

Approximately how many rows will there be for each combination of distribution and partition?

- A. 1 million
- B. 5 million
- C. 20 million
- D. 60 million

Answer: D

Explanation:

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

NEW QUESTION 96

- (Exam Topic 3)

You are implementing Azure Stream Analytics windowing functions.

Which windowing function should you use for each requirement? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

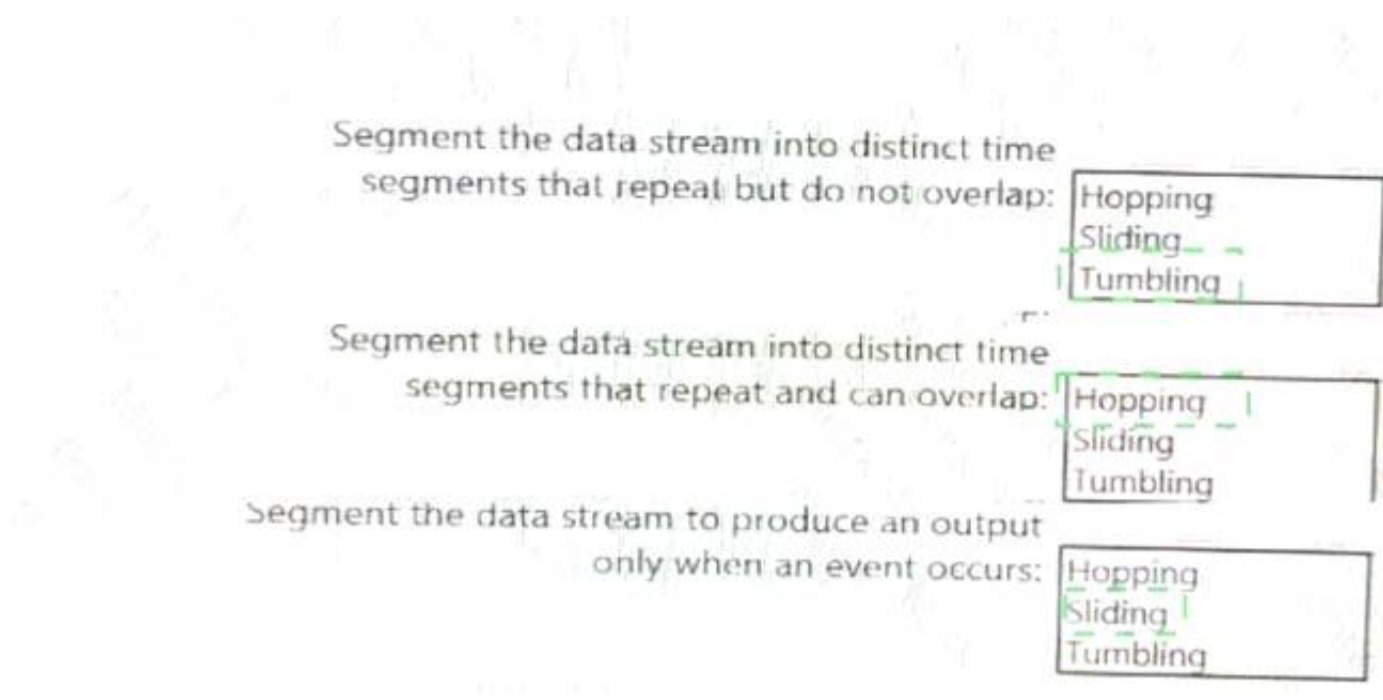
Segment the data stream into distinct time segments that repeat but do not overlap:	<div><div>Hopping</div><div>Sliding</div><div>Tumbling</div></div>
Segment the data stream into distinct time segments that repeat and can overlap:	<div><div>Hopping</div><div>Sliding</div><div>Tumbling</div></div>
Segment the data stream to produce an output only when an event occurs:	<div><div>Hopping</div><div>Sliding</div><div>Tumbling</div></div>

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Answer Area



NEW QUESTION 97

- (Exam Topic 3)

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

Answer: C

Explanation:

Cluster-scoped Init Scripts: Init scripts are shell scripts that run during the startup of each cluster node before the Spark driver or worker JVM starts. Databricks customers use init scripts for various purposes such as installing custom libraries, launching background processes, or applying enterprise security policies. Logs for Cluster-scoped init scripts are now more consistent with Cluster Log Delivery and can be found in the same root folder as driver and executor logs for the cluster.

Reference:

<https://databricks.com/blog/2018/08/30/introducing-cluster-scoped-init-scripts.html>

NEW QUESTION 101

- (Exam Topic 3)

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.

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.

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Step 1: Create a database role named Role1 and grant Role1 SELECT permissions to schema You need to grant Group1 read-only permissions to all the tables and views in schema1.

Place one or more database users into a database role and then assign permissions to the database role. Step 2: Assign Rol1 to the Group database user

Step 3: Assign the Azure role-based access control (Azure RBAC) Reader role for dw1 to Group1 Reference:

<https://docs.microsoft.com/en-us/azure/data-share/how-to-share-from-sql>

NEW QUESTION 104

- (Exam Topic 3)

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.

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

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

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>

NEW QUESTION 106

- (Exam Topic 3)

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 into Table1 and azure Data Lake Storage Gen2 container named container1.

You plan to insert data from the files 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.

- A. Yes
- B. No

Answer: B

NEW QUESTION 111

- (Exam Topic 3)

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 AverageFROM ClickStream TIMESTAMP BY CreatedAt GROUP BY Country, SlidingWindow(second, 10)

- B. SELECT Country, Count(*) AS CountFROM ClickStream TIMESTAMP BY CreatedAt GROUP BY Country, TumblingWindow(second, 10)
- C. SELECT Country, Avg(*) AS AverageFROM ClickStream TIMESTAMP BY CreatedAt GROUP BY Country, HoppingWindow(second, 10, 2)
- D. SELECT Country, Count(*) AS CountFROM ClickStream TIMESTAMP BY CreatedAt GROUP BY Country, SessionWindow(second, 5, 10)

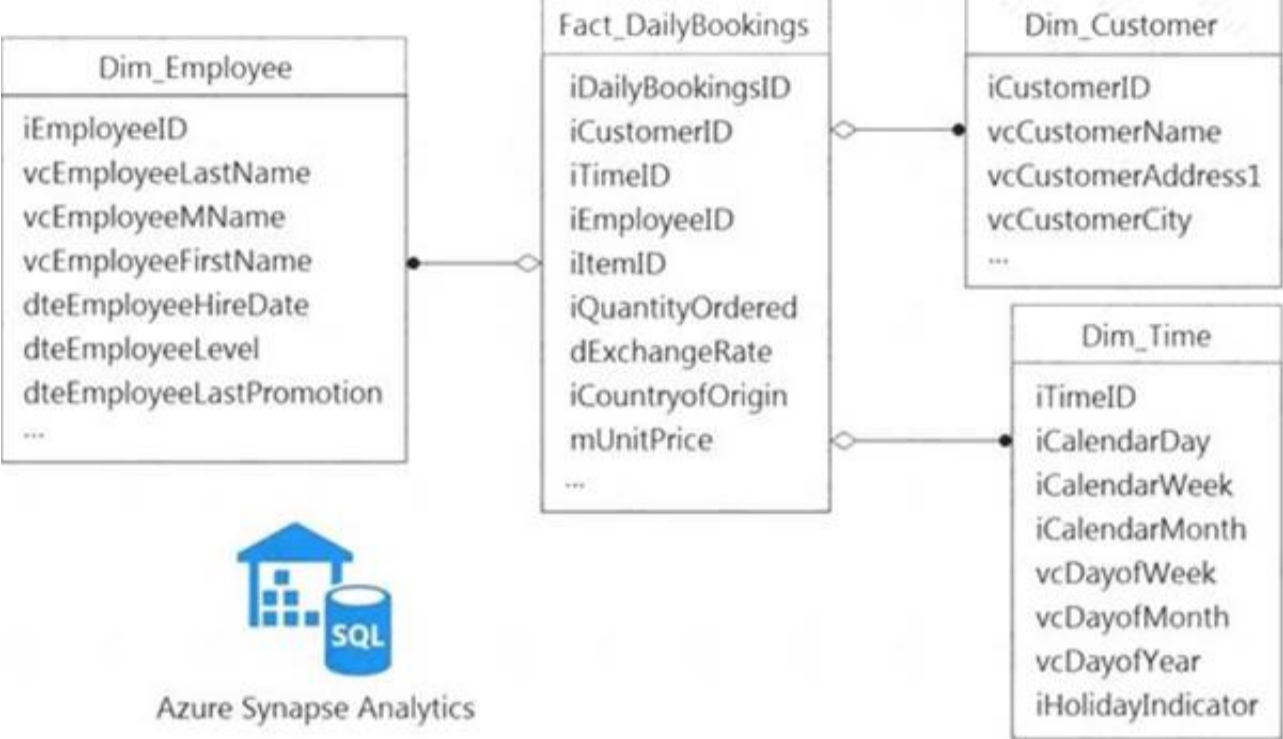
Answer: B

Explanation:

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: Reference:
<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-window-functions>

NEW QUESTION 116

- (Exam Topic 3)
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. 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.

Answer Area

Dim_Customer:

Dim_Employee:

Dim_Time:

Fact_DailyBookings:

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Answer Area

Dim_Customer:	<div>▼</div> <div> Hash distributed Round-robin Replicated </div>
Dim_Employee:	<div>▼</div> <div> Hash distributed Round-robin Replicated </div>
Dim_Time:	<div>▼</div> <div> Hash distributed Round-robin Replicated </div>
Fact_DailyBookings:	<div>▼</div> <div> Hash distributed Round-robin Replicated </div>

NEW QUESTION 118

- (Exam Topic 3)

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

Answer: C

NEW QUESTION 122

- (Exam Topic 3)

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

Answer: D

Explanation:

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>

NEW QUESTION 126

- (Exam Topic 3)

You configure monitoring for a Microsoft Azure SQL Data Warehouse implementation. The implementation uses PolyBase to load data from comma-separated value (CSV) files stored in Azure Data Lake Gen 2 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 files.'
- B. 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.'
- C. 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 regarding from an external source: 1 rows rejected out of total 1 rows processed.
- D. EXTERNAL TABLE access failed due to internal error: 'Java exception raised on call to HdfsBridge_Connect: Error [Unable to instantiate LoginClass] occurredwhile accessing external files.'

Answer: C

Explanation:

Customer Scenario:

SQL Server 2016 or SQL DW connected to Azure blob storage. The CREATE EXTERNAL TABLE DDL points to a directory (and not a specific file) and the directory contains files with different schemas.

SSMS Error:

Select query on the external table gives the following error: Msg 7320, Level 16, State 110, Line 14

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.

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.

Possible Solution:

If the data for each table consists of one file, then use the filename in the LOCATION section prepended by the directory of the external files. If there are multiple files per table, put each set of files into different directories in Azure Blob Storage and then you can point LOCATION to the directory instead of a particular file. The latter suggestion is the best practices recommended by SQLCAT even if you have one file per table.

NEW QUESTION 127

- (Exam Topic 3)

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 tiles can be queried quickly and that the data type information is retained.

What should you recommend?

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

Answer: B

Explanation:

The Avro format is great for data and message preservation. Avro schema with its support for evolution is essential for making the data robust for streaming architectures like Kafka, and with the metadata that schema provides, you can reason on the data. Having a schema provides robustness in providing meta-data about the data stored in Avro records which are self- documenting the data. References: <http://cloudurable.com/blog/avro/index.html>

NEW QUESTION 131

- (Exam Topic 3)

You are designing an Azure Databricks interactive cluster. The cluster will be used infrequently and will be configured for auto-termination.

You need to ensure that the cluster configuration is retained indefinitely after the cluster is terminated. The solution must minimize costs.

What should you do?

- A. Clone the cluster after it is terminated.
- B. Terminate the cluster manually when processing completes.
- C. Create an Azure runbook that starts the cluster every 90 days.
- D. Pin the cluster.

Answer: D

Explanation:

To keep an interactive cluster configuration even after it has been terminated for more than 30 days, an administrator can pin a cluster to the cluster list.

References:

<https://docs.azuredatabricks.net/clusters/clusters-manage.html#automatic-termination>

NEW QUESTION 136

- (Exam Topic 3)

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 use?

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

Answer: C

Explanation:

Type 2 - Creating a new additional record. In this methodology all history of dimension changes is kept in the database. You capture attribute change by adding a new row with a new surrogate key to the dimension table. Both the prior and new rows contain as attributes the natural key(or other durable identifier). Also 'effective date' and 'current indicator' columns are used in this method. There could be only one record with current indicator set to 'Y'. For 'effective date' columns, i.e. start_date and end_date, the end_date for current record usually is set to value 9999-12-31. Introducing changes to the dimensional model in type 2 could be very expensive database operation so it is not recommended to use it in dimensions where a new attribute could be added in the future.
<https://www.datawarehouse4u.info/SCD-Slowly-Changing-Dimensions.html>

NEW QUESTION 140

- (Exam Topic 3)

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 Session window
- B. a five-minute Sliding window
- C. a five-minute Tumbling window
- D. a five-minute Hopping window that has one-minute hop

Answer: C

Explanation:

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.

References:

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-window-functions>

NEW QUESTION 142

- (Exam Topic 3)

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

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.	

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

:

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-th> <https://azure.microsoft.com/en-us/blog/replicated-tables-now-generally-available-in-azure-sql-data-warehouse/>

NEW QUESTION 145

- (Exam Topic 3)

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,  
    [Postalcode] [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] [int] NULL
- B. [ManagerEmployeeID] [smallint] NULL
- C. [ManagerEmployeeKey] [int] NULL
- D. [ManagerName] [varchar](200) NULL

Answer: A

Explanation:

Use the same definition as the EmployeeID column. Reference:
<https://docs.microsoft.com/en-us/analysis-services/tabular-models/hierarchies-ssas-tabular>

NEW QUESTION 148

.....

Thank You for Trying Our Product

We offer two products:

1st - We have Practice Tests Software with Actual Exam Questions

2nd - Questions and Answers in PDF Format

DP-203 Practice Exam Features:

- * DP-203 Questions and Answers Updated Frequently
- * DP-203 Practice Questions Verified by Expert Senior Certified Staff
- * DP-203 Most Realistic Questions that Guarantee you a Pass on Your FirstTry
- * DP-203 Practice Test Questions in Multiple Choice Formats and Updatesfor 1 Year

100% Actual & Verified — Instant Download, Please Click
[Order The DP-203 Practice Test Here](#)