

Microsoft

Exam Questions DP-203

Data Engineering on Microsoft Azure



NEW QUESTION 1

- (Exam Topic 1)

You need to design the partitions for the product sales transactions. The solution must mee 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

Partition product sales transactions data by:	<div>Sales date Product ID Promotion ID</div>
Store product sales transactions data in:	<div>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</div>

- A. Mastered
B. Not Mastered

Answer: A

Explanation:

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-wha>

NEW QUESTION 2

- (Exam Topic 1)

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 Transaction-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.

Commands

Answer Area

CREATE EXTERNAL DATA SOURCE

CREATE EXTERNAL FILE FORMAT

CREATE EXTERNAL TABLE

CREATE EXTERNAL TABLE AS SELECT

CREATE DATABASE SCOPED CREDENTIAL

- A. Mastered
B. Not Mastered

Answer: A

Explanation:

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.

Reference:

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

NEW QUESTION 3

- (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 4

- (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 5

- (Exam Topic 1)

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

Answer: A

Explanation:

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>

NEW QUESTION 6

- (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 7

- (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 8

- (Exam Topic 3)

A company purchases IoT devices to monitor manufacturing machinery. The company uses an IoT appliance to communicate with the IoT devices. The company must be able to monitor the devices in real-time. You need to design the solution.

What should you recommend?

- A. Azure Stream Analytics cloud job using Azure PowerShell
- B. Azure Analysis Services using Azure Portal
- C. Azure Data Factory instance using Azure Portal
- D. Azure Analysis Services using Azure PowerShell

Answer: A

Explanation:

Stream Analytics is a cost-effective event processing engine that helps uncover real-time insights from devices, sensors, infrastructure, applications and data quickly and easily.

Monitor and manage Stream Analytics resources with Azure PowerShell cmdlets and powershell scripting that execute basic Stream Analytics tasks.

Reference:

<https://cloudblogs.microsoft.com/sqlserver/2014/10/29/microsoft-adds-iot-streaming-analytics-data-production-a>

NEW QUESTION 9

- (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 10
- (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 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

Answer: A

Explanation:
All file formats have different performance characteristics. For the fastest load, use compressed delimited text files.
Reference:
<https://docs.microsoft.com/en-us/azure/sql-data-warehouse/guidance-for-loading-data>

NEW QUESTION 10
- (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>▼</div> <div>Hash-distributed</div> <div>Round-robin</div> </div>	<div> <div></div> <div>▼</div> <div>DateKey</div> <div>ProductKey</div> <div>RegionKey</div> </div>
Invoices:	<div> <div></div> <div>▼</div> <div>Hash-distributed</div> <div>Round-robin</div> </div>	<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 14
 - (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 19

- (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.

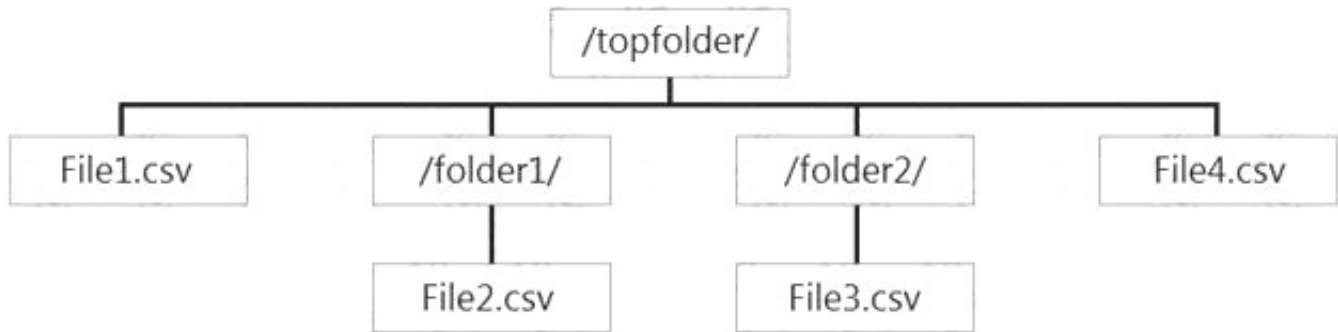
Reference:

<https://docs.azuredatabricks.net/clusters/configure.html>

NEW QUESTION 21

- (Exam Topic 3)

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

Answer: C

Explanation:

To run a T-SQL query over a set of files within a folder or set of folders while treating them as a single entity or rowset, provide a path to a folder or a pattern (using wildcards) over a set of files or folders. Reference:
<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/query-data-storage#query-multiple-files-or-folders>

NEW QUESTION 26

- (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 27

- (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 hierarchy

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 29

- (Exam Topic 3)

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.

```
SELECT
[user],
feature,
[ ]
DATEADD(
DATEDIFF(
DATEPART(
second,
[ ]
(Time) OVER (PARTITION BY [user], feature LIMIT DURATION(hour, 1) WHEN Event = 'start'),
ISFIRST
LAST
TOPONE
Time) as duration
FROM input TIMESTAMP BY Time
WHERE
Event = 'end'
```

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

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>

NEW QUESTION 31

- (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 snared 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 36

- (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 37

- (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: You use a dedicated SQL pool to create an external table that has a additional DateTime column. Does this meet the goal?

- A. Yes
- B. No

Answer: A

NEW QUESTION 40

- (Exam Topic 3)

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.

Answer Area

```
df.write
  .bucketBy
  .format
  .partitionBy
  .sortBy
  .csv("/DBTBL1")
  .json("/DBTBL1")
  .parquet("/DBTBL1")
  .saveAsTable("/DBTBL1")
```

Options for bucketBy:

- (*)
- ("GeographyRegionID")
- ("GeographyRegionID", "Year", "Month", "Day")
- ("Year", "Month", "Day", "GeographyRegionID")

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Answer Area

```
df.write
  .bucketBy
  .format
  .partitionBy
  .sortBy

  ("")
  ("GeographyRegionID")
  ("GeographyRegionID", "Year", "Month", "Day")
  ("Year", "Month", "Day", "GeographyRegionID")

.csv("/DBTBL1")
.json("/DBTBL1")
.parquet("/DBTBL1")
.saveAsTable("/DBTBL1")
```

NEW QUESTION 44

- (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 49

- (Exam Topic 3)

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

Answer: A

NEW QUESTION 53

- (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:	<div><div></div><div><div>An Azure Synapse Analytics Apache Spark pool</div><div>An Azure Synapse Analytics serverless SQL pool</div><div>Azure Data Factory</div><div>Azure Stream Analytics</div></div></div>
Window:	<div><div></div><div><div>Hopping</div><div>No window</div><div>Session</div><div>Tumbling</div></div></div>
Analysis type:	<div><div></div><div><div>Event pattern matching</div><div>Lagged record comparison</div><div>Point within polygon</div><div>Polygon overlap</div></div></div>

- 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 57

- (Exam Topic 3)

You are designing a solution that will copy Parquet files stored in an Azure Blob storage account to an Azure Data Lake Storage Gen2 account.

The data will be loaded daily to the data lake and will use a folder structure of {Year}/{Month}/{Day}/.

You need to design a daily Azure Data Factory data load to minimize the data transfer between the two accounts.

Which two configurations should you include in the design? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

- A. Delete the files in the destination before loading new data.
- B. Filter by the last modified date of the source files.
- C. Delete the source files after they are copied.
- D. Specify a file naming pattern for the destination.

Answer: BC

Explanation:

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/connector-azure-data-lake-storage>

NEW QUESTION 58

- (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?

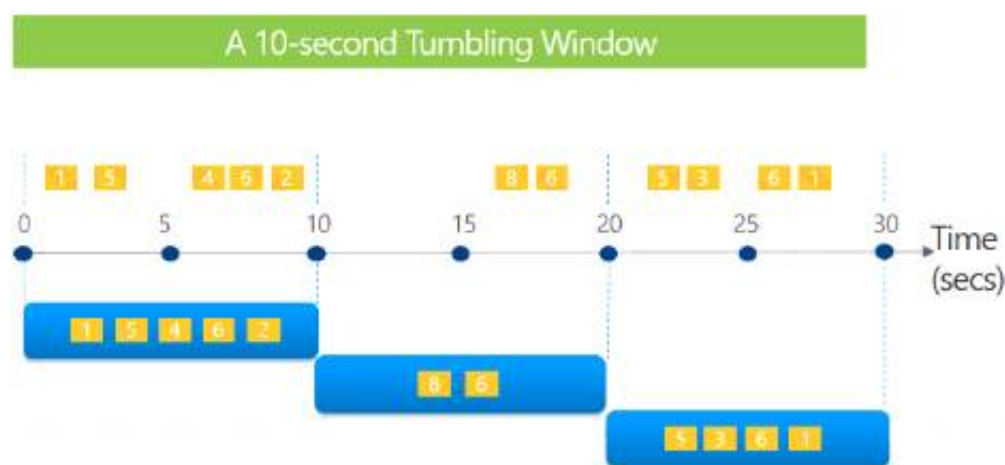
- A. Yes
- B. No

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 61

- (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 Standard cluster for the jobs.

Does this meet the goal?

- A. Yes
- B. No

Answer: B

Explanation:

We would need a High Concurrency cluster for 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 implement an Azure Data Lake Gen2 storage account.
 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. zone-redundant storage (ZRS)
- C. locally-redundant storage (LRS)
- D. geo-zone-redundant storage (GZRS)

Answer: A

Explanation:

Geo-redundant storage (GRS) copies your data synchronously three times within a single physical location in the primary region using LRS. It then copies your data asynchronously to a single physical location in the secondary region.
 Reference:
<https://docs.microsoft.com/en-us/azure/storage/common/storage-redundancy>

NEW QUESTION 67

- (Exam Topic 3)
 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. dynamic data masking
- B. row-level security (RLS)
- C. sensitivity classifications
- D. column-level security

Answer: D

NEW QUESTION 70

- (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 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 72

- (Exam Topic 3)

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 is Databricks notebook? 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

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.

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Step 1: Read the file into a data frame.
You can load the json files as a data frame in Azure Databricks. Step 2: Perform transformations on the data frame.
Step 3:Specify a temporary folder to stage the data
Specify a temporary folder to use while moving data between Azure Databricks and Azure Synapse. Step 4: 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.
Step 5: Drop the data frame
Clean up resources. You can terminate the cluster. From the Azure Databricks workspace, select Clusters on the left. For the cluster to terminate, under Actions, point to the ellipsis (...) and select the Terminate icon.
Reference:
<https://docs.microsoft.com/en-us/azure/azure-databricks/databricks-extract-load-sql-data-warehouse>

NEW QUESTION 77

- (Exam Topic 3)
You store files in an Azure Data Lake Storage Gen2 container. The container has the storage policy 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.

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/contoso1.csv
container1/docs/contoso.json
container1/mycontoso/contoso.csv

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

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/contoso1.csv
container1/docs/contoso.json
container1/mycontoso/contoso.csv

NEW QUESTION 79

- (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 81

- (Exam Topic 3)

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.

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.

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

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>

NEW QUESTION 86

- (Exam Topic 3)

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.

First week:

▼

Archive
Cool
Hot

After one month:

▼

Archive
Cool
Hot

After one year:

▼

Archive
Cool
Hot

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

First week: Hot

Hot - Optimized for storing data that is accessed frequently. After one month: Cool

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

After one year: Cool

NEW QUESTION 87

- (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 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 91

- (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:

Home > MySampleDatabase2 (mydocsamplesqlserver/MySampleDatabase2)

MySampleDatabase2 (mydocsamplesqlserver/MySampleDatabase2) | Data Discovery & Classification

SQL database

Search (Ctrl+/) Save Discard + Add classification Feedback

Power Platform

- Power BI (preview)
- Power Apps (preview)
- Power Automate (preview)

Settings

- Configure
- Geo-Replication
- Connection strings
- Sync to other databases
- Add Azure Search
- Properties
- Locks

Integrations

- Stream analytics (preview)

Security

- Auditing
- Data Discovery & Classification
- Dynamic Data Masking
- Security Center
- Transparent data encryption

Intelligent Performance

- Performance overview

Overview **Classification**

15 columns with classification recommendations (Click to minimize)

Accept selected recommendations Dismiss selected recommendations ☐ Show dismissed recommendations

☐ Select all Schema: 2 selected Table: 5 selected Filter by column

	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 96

- (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?

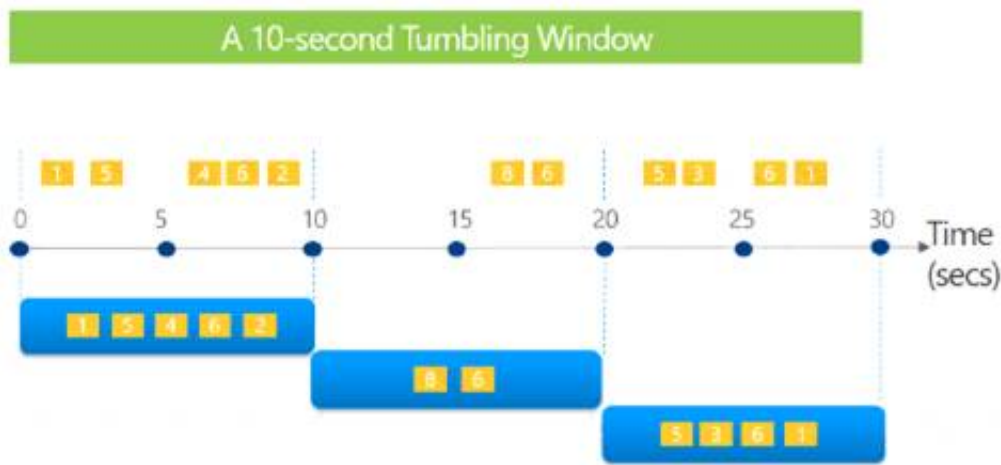
- A. snapshot
- B. tumbling
- C. hopping
- D. sliding

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 99

- (Exam Topic 3)
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.

Cluster Mode:

High Concurrency

Premium

Standard

Advanced option to enable:

Azure Data Lake Storage Gen1 Credential Passthrough

Table Access Control

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:
Box 1: High Concurrency
Enable Azure Data Lake Storage credential passthrough for a high-concurrency cluster. Incorrect: Support for Azure Data Lake Storage credential passthrough on standard clusters is in Public Preview. Standard clusters with credential passthrough are supported on Databricks Runtime 5.5 and above and are limited to a single user.
Box 2: Azure Data Lake Storage Gen1 Credential Passthrough
You can authenticate automatically to Azure Data Lake Storage Gen1 and Azure Data Lake Storage Gen2 from Azure Databricks clusters using the same Azure Active Directory (Azure AD) identity that you use to log into Azure Databricks. 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.
References:
<https://docs.azuredatabricks.net/spark/latest/data-sources/azure/adls-passthrough.html>

NEW QUESTION 102

- (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 105

- (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.

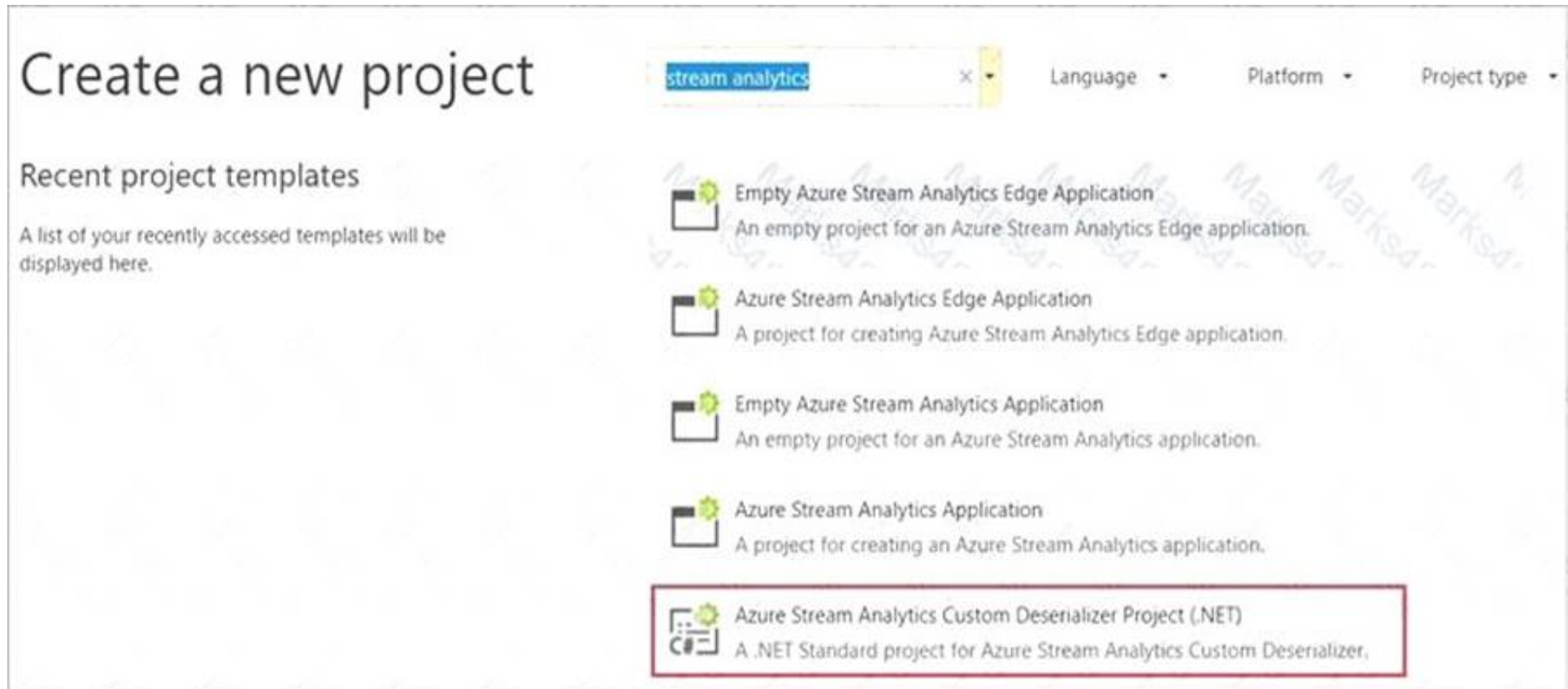
- A. Mastered
- B. Not Mastered

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.

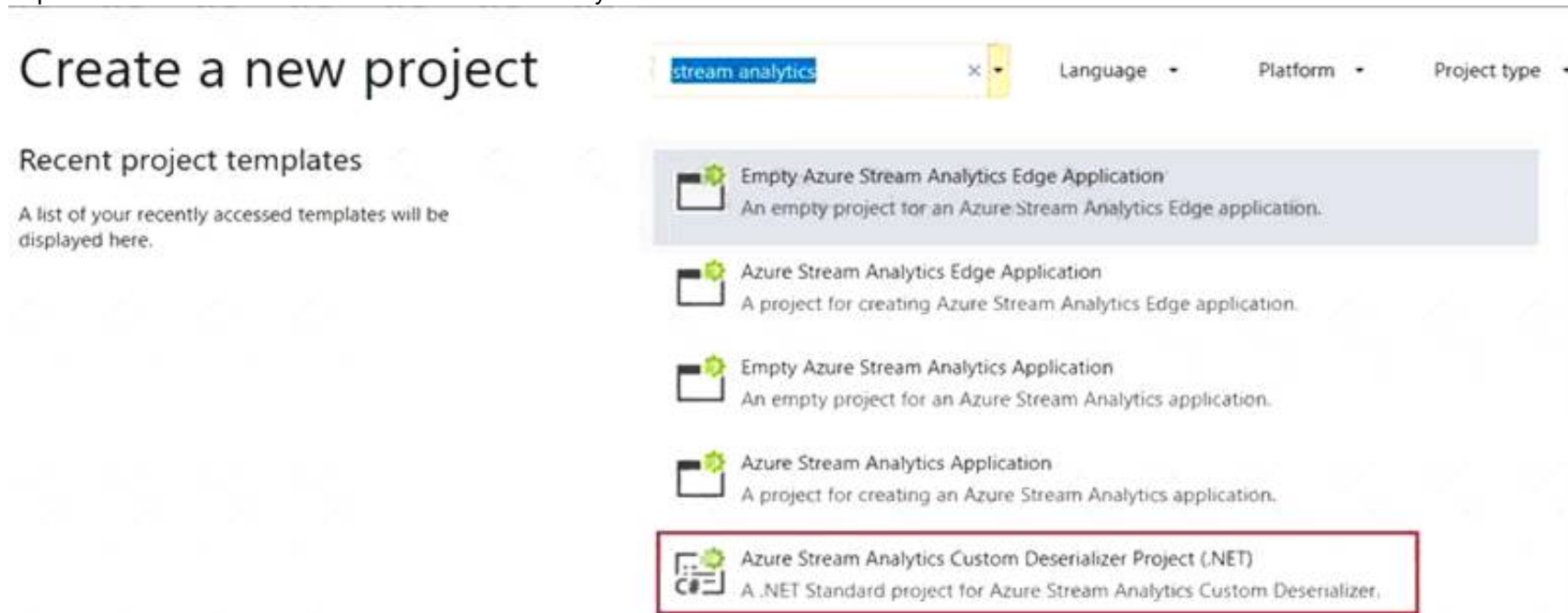
Step 3: Add an Azure Stream Analytics Application project to the solution Add an Azure Stream Analytics project

> 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 108

- (Exam Topic 3)

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.

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.

- A. Mastered
B. Not Mastered

Answer: A

Explanation:

Step 1: Create an Azure Storage account that has a lifecycle policy

To automate common data management tasks, Microsoft created a solution based on Azure Data Factory. The service, Data Lifecycle Management, makes frequently accessed data available and archives or purges other data according to retention policies. Teams across the company use the service to reduce storage costs, improve app performance, and comply with data retention policies.

Step 2: Create a Log Analytics workspace that has Data Retention set to 120 days.

Data Factory stores pipeline-run data for only 45 days. Use Azure Monitor if you want to keep that data for a longer time. With Monitor, you can route diagnostic logs for analysis to multiple different targets, such as a Storage Account: Save your diagnostic logs to a storage account for auditing or manual inspection. You can use the diagnostic settings to specify the retention time in days.

Step 3: From Azure Portal, add a diagnostic setting. Step 4: Send the data to a log Analytics workspace,

Event Hub: A pipeline that transfers events from services to Azure Data Explorer. Keeping Azure Data Factory metrics and pipeline-run data.

Configure diagnostic settings and workspace.

Create or add diagnostic settings for your data factory.

- In the portal, go to Monitor. Select Settings > Diagnostic settings.
- Select the data factory for which you want to set a diagnostic setting.
- If no settings exist on the selected data factory, you're prompted to create a setting. Select Turn on diagnostics.
- Give your setting a name, select Send to Log Analytics, and then select a workspace from Log Analytics Workspace.
- Select Save. Reference:

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

NEW QUESTION 113

- (Exam Topic 3)

You have an Azure Factory instance named DF1 that contains a pipeline named PL1.PL1 includes a tumbling window trigger.

You create five clones of PL1. You configure each clone pipeline to use a different data source.

You need to ensure that the execution schedules of the clone pipeline match the execution schedule of PL1. What should you do?

- A. Add a new trigger to each cloned pipeline
B. Associate each cloned pipeline to an existing trigger.
C. Create a tumbling window trigger dependency for the trigger of PL1.
D. Modify the Concurrency setting of each pipeline.

Answer: B

NEW QUESTION 116

- (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 118

- (Exam Topic 3)

You have two Azure Data Factory instances named ADFdev and ADFprod. ADFdev connects to an Azure DevOps Git repository.

You publish changes from the main branch of the Git repository to ADFdev. You need to deploy the artifacts from ADFdev to ADFprod.

What should you do first?

- A. From ADFdev, modify the Git configuration.
- B. From ADFdev, create a linked service.
- C. From Azure DevOps, create a release pipeline.
- D. From Azure DevOps, update the main branch.

Answer: C

Explanation:

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

Note:

The following is a guide for setting up an Azure Pipelines release that automates the deployment of a data factory to multiple environments.

- In Azure DevOps, open the project that's configured with your data factory.
 - On the left side of the page, select Pipelines, and then select Releases.
 - Select New pipeline, or, if you have existing pipelines, select New and then New release pipeline.
 - In the Stage name box, enter the name of your environment.
 - Select Add artifact, and then select the git repository configured with your development data factory.
- Select the publish branch of the repository for the Default branch. By default, this publish branch is adf_publish.
- Select the Empty job template. Reference:

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

NEW QUESTION 123

- (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 126

- (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

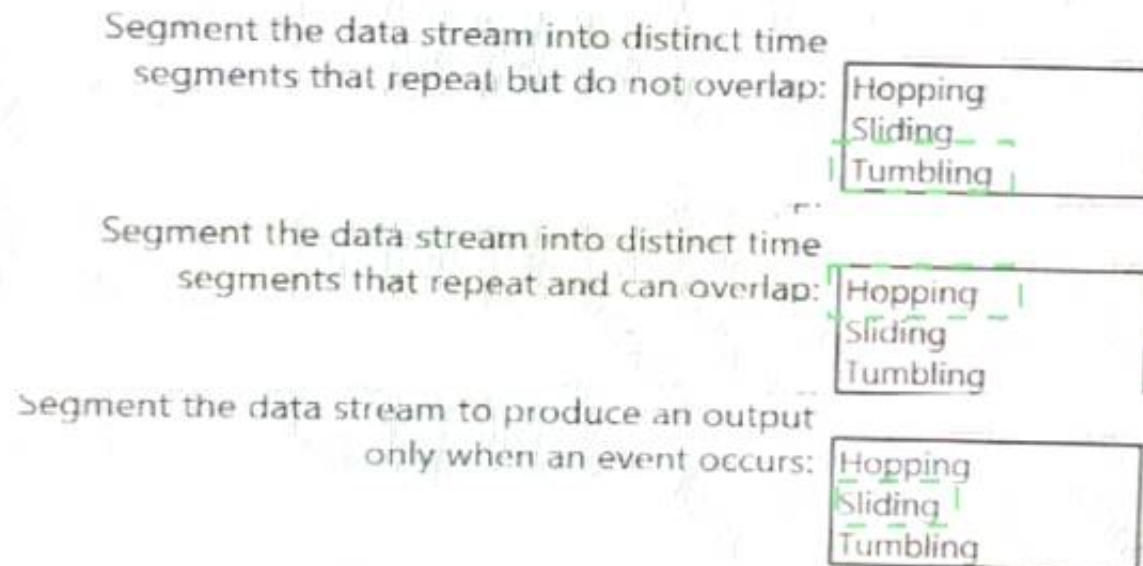
Segment the data stream into distinct time segments that repeat but do not overlap:	<input type="checkbox"/> Hopping <input type="checkbox"/> Sliding <input type="checkbox"/> Tumbling
Segment the data stream into distinct time segments that repeat and can overlap:	<input type="checkbox"/> Hopping <input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

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Answer Area



NEW QUESTION 127

- (Exam Topic 3)

You have an Azure Stream Analytics query. The query returns a result set that contains 10,000 distinct values for a column named clusterID. You monitor the Stream Analytics job and discover high latency. You need to reduce the latency. Which two actions should you perform? Each correct answer presents a complete solution. NOTE: Each correct selection is worth one point.

- A. Add a pass-through query.
- B. Add a temporal analytic function.
- C. Scale out the query by using PARTITION BY.
- D. Convert the query to a reference query.
- E. Increase the number of streaming units.

Answer: CE

Explanation:

C: Scaling a Stream Analytics job takes advantage of partitions in the input or output. Partitioning lets you divide data into subsets based on a partition key. A process that consumes the data (such as a Streaming Analytics job) can consume and write different partitions in parallel, which increases throughput.

E: 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. This capacity lets you focus on the query logic and abstracts the need to manage the hardware to run your Stream Analytics job in a timely manner.

References:

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

NEW QUESTION 131

- (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 134

- (Exam Topic 3)

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

Answer: D

NEW QUESTION 137

- (Exam Topic 3)

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 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.

SELECT sensorId,

growth = reading -

LAG

LAST

LEAD

▼

(reading) OVER (PARTITION BY sensorId

▼

(hour, 1))

LIMIT DURATION

OFFSET

WHEN

FROM input

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

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>

NEW QUESTION 141

- (Exam Topic 3)

You have an Azure Synapse Analytics dedicated SQL 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. Switch the first partition from dbo.Sales to stg.Sales.
- B. Switch the first partition from stg.Sales to db
- C. Sales.
- D. Update dbo.Sales from stg.Sales.
- E. Insert the data from stg.Sales into dbo.Sales.

Answer: D

NEW QUESTION 143

- (Exam Topic 3)

You have the following Azure Stream Analytics query.

WITH

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

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

Statements	Yes	No
The query joins 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 type="radio"/>	<input type="radio"/>

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Box 1: Yes

You can now use a new extension of Azure Stream Analytics SQL to specify the number of partitions of a stream when reshuffling the data.

The outcome is a stream that has the same partition scheme. Please see below for an example: WITH step1 AS (SELECT * FROM [input1] PARTITION BY DeviceID INTO 10),

step2 AS (SELECT * FROM [input2] PARTITION BY DeviceID INTO 10)

SELECT * INTO [output] FROM step1 PARTITION BY DeviceID UNION step2 PARTITION BY DeviceID Note: The new extension of Azure Stream Analytics SQL includes a keyword INTO that allows you to specify

the number of partitions for a stream when performing reshuffling using a PARTITION BY statement.

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

10 partitions x six SUs = 60 SUs is fine.

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/>

NEW QUESTION 145

- (Exam Topic 3)

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.

Answer: AB

Explanation:

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.

STREAM NAMES	CONDITION
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:

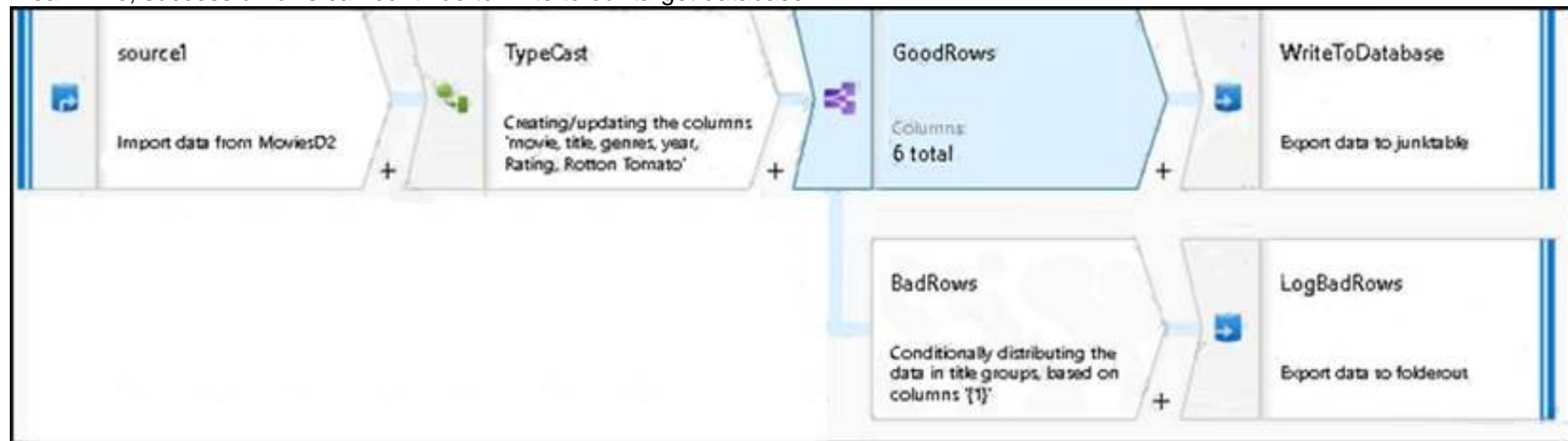
* 3. Now we need to log the rows that failed. Add a sink transformation to the BadRows stream for logging. Here, we'll "auto-map" all of the fields so that we have logging of the complete transaction record. This is a text-delimited CSV file output to a single file in Blob Storage. We'll call the log file "badrows.csv".

File name option * ☐ Default ☐ Pattern ☐ Per partition ☐ As data in column ☒ Output to single file

Output to single file * badrows.csv

Quote All ☐

* 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>

NEW QUESTION 146

- (Exam Topic 3)

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 Azure Monitor, run a Kusto query against the AzureDiagnostics table.
- B. From Azure Monitor, run a Kusto query against the SparkLogging1 Event.CL table.
- C. From Synapse Studio, select the workspace
- D. From Monitor, select Apache Sparks applications.
- E. From Synapse Studio, select the workspace
- F. From Monitor, select SQL requests.

Answer: C

NEW QUESTION 149

- (Exam Topic 3)

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

Answer: B

NEW QUESTION 151

- (Exam Topic 3)

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. a default value
- B. dynamic data masking
- C. row-level security (RLS)
- D. column encryption
- E. table partitions

Answer: C

NEW QUESTION 155

- (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,
    [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] [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 156

.....

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](#)