

Microsoft

Exam Questions DP-203

Data Engineering on Microsoft Azure



NEW QUESTION 1

- (Exam Topic 3)

A company plans to use Apache Spark analytics to analyze intrusion detection data.

You need to recommend a solution to analyze network and system activity data for malicious activities and policy violations. The solution must minimize administrative efforts.

What should you recommend?

- A. Azure Data Lake Storage
- B. Azure Databricks
- C. Azure HDInsight
- D. Azure Data Factory

Answer: B

Explanation:

Three common analytics use cases with Microsoft Azure Databricks

Recommendation engines, churn analysis, and intrusion detection are common scenarios that many organizations are solving across multiple industries. They require machine learning, streaming analytics, and utilize massive amounts of data processing that can be difficult to scale without the right tools.

Recommendation engines, churn analysis, and intrusion detection are common scenarios that many organizations are solving across multiple industries. They require machine learning, streaming analytics, and utilize massive amounts of data processing that can be difficult to scale without the right tools.

Note: Recommendation engines, churn analysis, and intrusion detection are common scenarios that many organizations are solving across multiple industries. They require machine learning, streaming analytics, and utilize massive amounts of data processing that can be difficult to scale without the right tools.

Reference:

<https://azure.microsoft.com/es-es/blog/three-critical-analytics-use-cases-with-microsoft-azure-databricks/>

NEW QUESTION 2

- (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 3

- (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:

<input type="checkbox"/> Flatten hierarchy <input type="checkbox"/> Merge files <input type="checkbox"/> Preserve hierarchy

Sink file type:

<input type="checkbox"/> CSV <input type="checkbox"/> JSON <input type="checkbox"/> Parquet <input type="checkbox"/> TXT

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Box 1: Preserve 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 4

- (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 have an Azure Synapse Analytics dedicated SQL pool that contains a table named Table1. You have files that are ingested and loaded into an Azure Data Lake Storage Gen2 container named container1.

You plan to insert data from the files in container1 into Table1 and transform the data. Each row of data in the files will produce one row in the serving layer of Table1.

You need to ensure that when the source data files are loaded to container1, the DateTime is stored as an additional column in Table1.

Solution: You use an Azure Synapse Analytics serverless SQL pool to create an external table that has an additional DateTime column.

Does this meet the goal?

- A. Yes
- B. No

Answer: B

Explanation:

Instead use the derived column transformation to generate new columns in your data flow or to modify existing fields.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/data-flow-derived-column>

NEW QUESTION 5

- (Exam Topic 3)

You have an Azure Databricks workspace that contains a Delta Lake dimension table named Table1. Table1 is a Type 2 slowly changing dimension (SCD) table. You need to apply updates from a source table to Table1. Which Apache Spark SQL operation should you use?

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

Answer: C

Explanation:

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

Example:

```
// Implementing SCD Type 2 operation using merge function customersTable
as("customers") merge(
stagedUpdates.as("staged_updates"), "customers.customerId = mergeKey")
whenMatched("customers.current = true AND customers.address <> staged_updates.address") updateExpr(Map(
"current" -> "false",
"endDate" -> "staged_updates.effectiveDate")) whenNotMatched()
insertExpr(Map(
"customerid" -> "staged_updates.customerId", "address" -> "staged_updates.address", "current" -> "true",
"effectiveDate" -> "staged_updates.effectiveDate", "endDate" -> "null"))
execute()
}
```

Reference:

<https://www.projectpro.io/recipes/what-is-slowly-changing-data-scd-type-2-operation-delta-table-databricks>

NEW QUESTION 6

- (Exam Topic 3)

You have an Azure event hub named retailhub that has 16 partitions. Transactions are posted to retailhub. Each transaction includes the transaction ID, the individual line items, and the payment details. The transaction ID is used as the partition key.

You are designing an Azure Stream Analytics job to identify potentially fraudulent transactions at a retail store. The job will use retailhub as the input. The job will output the transaction ID, the individual line items, the payment details, a fraud score, and a fraud indicator.

You plan to send the output to an Azure event hub named fraudhub.

You need to ensure that the fraud detection solution is highly scalable and processes transactions as quickly as possible.

How should you structure the output of the Stream Analytics job? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Number of partitions:

Partition key:

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Box 1: 16

For Event Hubs you need to set the partition key explicitly.

An embarrassingly parallel job is the most scalable scenario in Azure Stream Analytics. It connects one partition of the input to one instance of the query to one partition of the output.

Box 2: Transaction ID Reference:

<https://docs.microsoft.com/en-us/azure/event-hubs/event-hubs-features#partitions>

NEW QUESTION 7

- (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 8

- (Exam Topic 3)

You have an Azure Synapse Analytics dedicated SQL pool named Pool1 and a database named DB1. DB1 contains a fact table named Table1. You need to identify the extent of the data skew in Table1. What should you do in Synapse Studio?

- A. Connect to the built-in pool and run dbcc pdw_showpaceused.
- B. Connect to the built-in pool and run dbcc checkalloc.
- C. Connect to Pool1 and query sys.dm_pdw_node_scacus.
- D. Connect to Pool1 and query sys.dm_pdw_nodes_db_partition_scacs.

Answer: A

Explanation:

A quick way to check for data skew is to use DBCC PDW_SHOWSPACEUSED. The following SQL code returns the number of table rows that are stored in each of the 60 distributions. For balanced performance, the rows in your distributed table should be spread evenly across all the distributions.

DBCC PDW_SHOWSPACEUSED('dbo.FactInternetSales'); Reference:

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

NEW QUESTION 9

- (Exam Topic 3)

You have an Azure Storage account that generates 200,000 new files daily. The file names have a format of (YYY)/(MM)/(DD)/(HH)/(CustomerID).csv.

You need to design an Azure Data Factory solution that will load new data from the storage account to an Azure Data lake once hourly. The solution must minimize load times and costs.

How should you configure the solution? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:
Answer Area



NEW QUESTION 10

- (Exam Topic 3)

You have an Azure Synapse Analytics dedicated SQL pool that contains a table named dbo.Users. You need to prevent a group of users from reading user email addresses from dbo.Users. What should you use?

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

Answer: B

NEW QUESTION 10

- (Exam Topic 3)

You are implementing a star schema in an Azure Synapse Analytics dedicated SQL pool. You plan to create a table named DimProduct. DimProduct must be a Type 3 slowly changing dimension (SCD) table that meets the following requirements:

- The values in two columns named ProductKey and ProductSourceID will remain the same.
- The values in three columns named ProductName, ProductDescription, and Color can change. You need to add additional columns to complete the following table definition.

```
CREATE TABLE [dbo].[dimproduct]
(
    [ProductKey] INT NOT NULL,
    [ProductSourceID] INT NOT NULL,
    [ProductName] NVARCHAR(100) NOT NULL,
    [ProductDescription] NVARCHAR(2000) NOT NULL,
    [Color] NVARCHAR(50) NOT NULL
)
WITH
(
    DISTRIBUTION = REPLICATE,
    CLUSTERED COLUMNSTORE INDEX
);
```

- A) [OriginalProductDescription] NVARCHAR(2000) NOT NULL
- B) [IsCurrentRow] [bit] NOT NULL
- C) [EffectiveStartDate] [datetime] NOT NULL
- D) [EffectiveEndDate] [datetime] NOT NULL
- E) [OriginalProductName] NVARCHAR(100) NULL
- F) [OriginalColor] NVARCHAR(50) NOT NULL

- A. Option A
- B. Option B
- C. Option C
- D. Option D
- E. Option E
- F. Option F

Answer: ABC

NEW QUESTION 12

- (Exam Topic 3)

You have an Azure subscription linked to an Azure Active Directory (Azure AD) tenant that contains a service principal named ServicePrincipal1. The subscription contains an Azure Data Lake Storage account named adls1. Adls1 contains a folder named Folder2 that has a URI of

<https://adls1.dfs.core.windows.net/container1/Folder1/Folder2/>.
 ServicePrincipal1 has the access control list (ACL) permissions shown in the following table.

Resource	Permission
container1	Access – Execute
Folder1	Access – Execute
Folder2	Access – Read

You need to ensure that ServicePrincipal1 can perform the following actions:

- > Traverse child items that are created in Folder2.
- > Read files that are created in Folder2.

The solution must use the principle of least privilege.

Which two permissions should you grant to ServicePrincipal1 for Folder2? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. Access - Read
- B. Access - Write
- C. Access - Execute
- D. Default-Read
- E. Default - Write
- F. Default - Execute

Answer: DF

Explanation:

Execute (X) permission is required to traverse the child items of a folder.

There are two kinds of access control lists (ACLs), Access ACLs and Default ACLs. Access ACLs: These control access to an object. Files and folders both have Access ACLs.

Default ACLs: A "template" of ACLs associated with a folder that determine the Access ACLs for any child items that are created under that folder. Files do not have Default ACLs.

Reference:

<https://docs.microsoft.com/en-us/azure/data-lake-store/data-lake-store-access-control>

NEW QUESTION 17

- (Exam Topic 3)

You plan to create an Azure Data Lake Storage Gen2 account

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

- Provides the highest degree of data resiliency
- Ensures that content remains available for writes if a primary data center fails

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

Answer Area

Replication mechanism:

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

Failover process:

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

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Graphical user interface, text, application Description automatically generated

Availability : "Microsoft recommends RA-GZRS for maximum availability and durability for your applications."

Failover: "The customer initiates the account failover to the secondary endpoint. " <https://docs.microsoft.com/en-us/azure/storage/common/storage-disaster-recovery-guidance?toc=/azure/storage/>

<https://docs.microsoft.com/en-us/answers/questions/32583/azure-data-lake-gen2-disaster-recovery-storage-acco.h>

NEW QUESTION 22

- (Exam Topic 3)

You are building an Azure Data Factory solution to process data received from Azure Event Hubs, and then ingested into an Azure Data Lake Storage Gen2 container.

The data will be ingested every five minutes from devices into JSON files. The files have the following naming pattern.

`/{deviceType}/in/{YYYY}/{MM}/{DD}/{HH}/{deviceID}_{YYYY}{MM}{DD}{HH}{mm}.json`

You need to prepare the data for batch data processing so that there is one dataset per hour per deviceType. The solution must minimize read times.

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

NOTE: Each correct selection is worth one point.

Parameter:

- @pipeline(),TriggerTime
- @pipeline(),TriggerType
- @trigger().outputs.windowStartTime
- @trigger().startTime

Naming pattern:

- /{deviceID}/out/{YYYY}/{MM}/{DD}/{HH}.json
- /{YYYY}/{MM}/{DD}/{deviceType}.json
- /{YYYY}/{MM}/{DD}/{HH}.json
- /{YYYY}/{MM}/{DD}/{HH}_{deviceType}.json

Copy behavior:

- Add dynamic content
- Flatten hierarchy
- Merge files

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Box 1: @trigger().startTime

startTime: A date-time value. For basic schedules, the value of the startTime property applies to the first occurrence. For complex schedules, the trigger starts no sooner than the specified startTime value.

Box 2: /{YYYY}/{MM}/{DD}/{HH}_{deviceType}.json One dataset per hour per deviceType.

Box 3: Flatten hierarchy

- FlattenHierarchy: All files from the source folder are in the first level of the target folder. The target files have autogenerated names.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipeline-execution-triggers> <https://docs.microsoft.com/en-us/azure/data-factory/connector-file-system>

NEW QUESTION 27

- (Exam Topic 3)

You have a SQL pool in Azure Synapse.

You discover that some queries fail or take a long time to complete. You need to monitor for transactions that have rolled back.

Which dynamic management view should you query?

- A. sys.dm_pdw_request_steps
- B. sys.dm_pdw_nodes_tran_database_transactions
- C. sys.dm_pdw_waits
- D. sys.dm_pdw_exec_sessions

Answer: B

Explanation:

You can use Dynamic Management Views (DMVs) to monitor your workload including investigating query execution in SQL pool.

If your queries are failing or taking a long time to proceed, you can check and monitor if you have any transactions rolling back.

Example:

-- Monitor rollback SELECT

SUM(CASE WHEN t.database_transaction_next_undo_lsn IS NOT NULL THEN 1 ELSE 0 END), t.pdw_node_id, nod.[type]

FROM sys.dm_pdw_nodes_tran_database_transactions t

JOIN sys.dm_pdw_nodes nod ON t.pdw_node_id = nod.pdw_node_id GROUP BY t.pdw_node_id, nod.[type]

Reference:

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

NEW QUESTION 31

- (Exam Topic 3)

You have the following table named Employees.

first_name	last_name	hire_date	employee_type
Jane	Doe	2019-08-23	new
Ben	Smith	2017-12-15	Standard

You need to calculate the employee_type value based on the hire_date value.

How should you complete the Transact-SQL statement? To answer, drag the appropriate values to the correct targets. Each value may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point.

Values

Answer Area

-
-
-
-
-

```
SELECT
    *,
    WHEN hire_date >= '2019-01-01' THEN 'New'
    'Standard'
END AS employee_type
FROM
    employees
```

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Graphical user interface, text, application Description automatically generated

Box 1: CASE

CASE evaluates a list of conditions and returns one of multiple possible result expressions.

CASE can be used in any statement or clause that allows a valid expression. For example, you can use CASE in statements such as SELECT, UPDATE, DELETE and SET, and in clauses such as select_list, IN, WHERE, ORDER BY, and HAVING.

Syntax: Simple CASE expression: CASE input_expression

WHEN when_expression THEN result_expression [...n] [ELSE else_result_expression]

END

Box 2: ELSE

Reference:

<https://docs.microsoft.com/en-us/sql/t-sql/language-elements/case-transact-sql>

NEW QUESTION 35

- (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 38

- (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 41

- (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.

Solution: You use a tumbling window, and you set the window size to 10 seconds. Does this meet the goal?

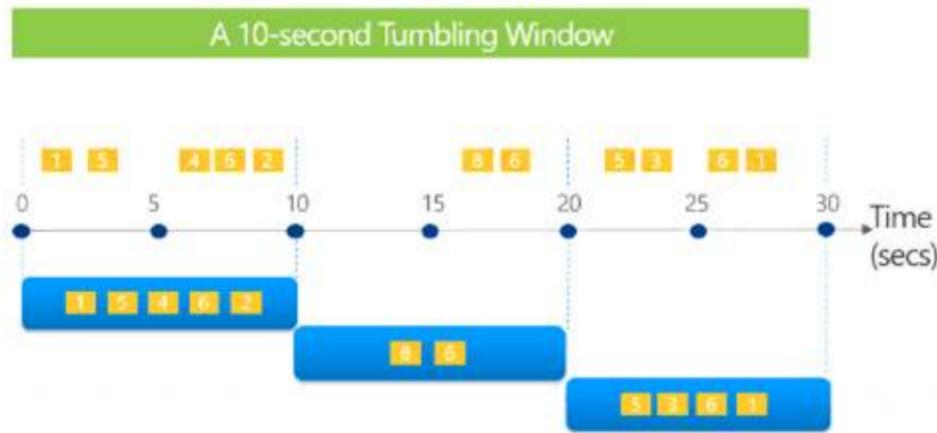
- A. Yes
- B. No

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 44

- (Exam Topic 3)

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

You run the following command in an Azure Synapse Analytics Spark pool in MyWorkspace. CREATE TABLE mytestdb.myParquetTable(EmployeeID int, EmployeeName string, EmployeeStartDate date) USING Parquet

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

EmployeeName	EmployeeID	EmployeeStartDate
Alice	24	2020-01-25

One minute later, you execute the following query from a serverless SQL pool in MyWorkspace. SELECT EmployeeID FROM mytestdb.dbo.myParquetTable WHERE name = 'Alice';

What will be returned by the query?

- A. 24
- B. an error
- C. a null value

Answer: B

Explanation:

Once a database has been created by a Spark job, you can create tables in it with Spark that use Parquet as the storage format. Table names will be converted to lower case and need to be queried using the lower case name. These tables will immediately become available for querying by any of the Azure Synapse workspace Spark pools. They can also be used from any of the Spark jobs subject to permissions.

Note: For external tables, since they are synchronized to serverless SQL pool asynchronously, there will be a delay until they appear.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/metadata/table>

NEW QUESTION 47

- (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

Explanation:

Reference:

<https://docs.microsoft.com/en-us/azure/event-hubs/process-data-azure-stream-analytics>

NEW QUESTION 50

- (Exam Topic 3)

You have an Azure data factory named ADF1.

You currently publish all pipeline authoring changes directly to ADF1.

You need to implement version control for the changes made to pipeline artifacts. The solution must ensure that you can apply version control to the resources currently defined in the UX Authoring canvas for ADF1.

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

- A. Create an Azure Data Factory trigger
- B. From the UX Authoring canvas, select Set up code repository
- C. Create a GitHub action
- D. From the UX Authoring canvas, run Publish All.
- E. Create a Git repository
- F. From the UX Authoring canvas, select Publish

Answer: DE

Explanation:

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/source-control>

NEW QUESTION 53

- (Exam Topic 3)

You have an Azure Data Lake Storage Gen2 account that contains two folders named Folder1 and Folder2. You use Azure Data Factory to copy multiple files from Folder1 to Folder2.

```
Operation on target Copy_sks failed: Failure happened on 'Sink' side.
ErrorCode=DelimitedTextMoreColumnsThanDefined,
'Type=Microsoft.DataTransfer.Common.Shared.HybridDeliveryException,
Message=Error found when processing 'Csv/Tsv Format Text' source
'0_2020_11_09_11_43_32.avro' with row number 53: found more columns
than expected column count 27., Source=Microsoft.DataTransfer.Common,'
```

You receive the following error.

What should you do to resolve the error.

- A. Add an explicit mapping.
- B. Enable fault tolerance to skip incompatible rows.
- C. Lower the degree of copy parallelism
- D. Change the Copy activity setting to Binary Copy

Answer: A

Explanation:

Reference:

<https://knowledge.informatica.com/s/article/Microsoft-Azure-Data-Lake-Store-Gen2-target-file-names-not-gene>

NEW QUESTION 58

- (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

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

Actions

Answer Area

- Mount the Data Lake Storage onto DBFS.
- Write the results to a table in Azure Synapse.
- Specify a temporary folder to stage the data.
- Read the file into a data frame.
- Perform transformations on the data frame.

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Table Description automatically generated

Step 1: Mount the Data Lake Storage onto DBFS

Begin with creating a file system in the Azure Data Lake Storage Gen2 account. Step 2: Read the file into a data frame.

You can load the json files as a data frame in Azure Databricks. Step 3: Perform transformations on the data frame.

Step 4: Specify a temporary folder to stage the data

Specify a temporary folder to use while moving data between Azure Databricks and Azure Synapse. Step 5: Write the results to a table in Azure Synapse.

You upload the transformed data frame into Azure Synapse. You use the Azure Synapse connector for Azure Databricks to directly upload a dataframe as a table in a Azure Synapse.

Reference:

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

NEW QUESTION 61

- (Exam Topic 3)

You have an Azure Stream Analytics job.

You need to ensure that the job has enough streaming units provisioned. You configure monitoring of the SU % Utilization metric.

Which two additional metrics should you monitor? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. Backlogged Input Events
- B. Watermark Delay
- C. Function Events
- D. Out of order Events
- E. Late Input Events

Answer: AB

Explanation:

To react to increased workloads and increase streaming units, consider setting an alert of 80% on the SU Utilization metric. Also, you can use watermark delay and backlogged events metrics to see if there is an impact.

Note: Backlogged Input Events: Number of input events that are backlogged. A non-zero value for this metric implies that your job isn't able to keep up with the number of incoming events. If this value is slowly increasing or consistently non-zero, you should scale out your job, by increasing the SUs.

Reference:

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

NEW QUESTION 62

- (Exam Topic 3)

You have an Azure Synapse Analytics dedicated SQL pool named Pool1 that contains a table named Sales. Sales has row-level security (RLS) applied. RLS uses the following predicate filter.

```
CREATE FUNCTION Security.fn_securitypredicate(@SalesRep AS sysname)
    RETURNS TABLE
WITH SCHEMABINDING
AS
    RETURN SELECT 1 AS fn_securitypredicate_result
WHERE @SalesRep = USER_NAME() OR USER_NAME() = 'Manager';
```

A user named SalesUser1 is assigned the db_datareader role for Pool1.

A user named SalesUser1 is assigned the db_datareader role for Pool1. Which rows in the Sales table are returned when SalesUser1 queries the table?

- A. only the rows for which the value in the User_Name column is SalesUser1
- B. all the rows
- C. only the rows for which the value in the SalesRep column is Manager
- D. only the rows for which the value in the SalesRep column is SalesUser1

Answer: C

NEW QUESTION 66

- (Exam Topic 3)

You are developing a solution using a Lambda architecture on Microsoft Azure. The data at test layer must meet the following requirements:

Data storage:

- Serve as a repository (or high volumes of large files in various formats).
- Implement optimized storage for big data analytics workloads.
- Ensure that data can be organized using a hierarchical structure. Batch processing:
- Use a managed solution for in-memory computation processing.
- Natively support Scala, Python, and R programming languages.
- Provide the ability to resize and terminate the cluster automatically. Analytical data store:
- Support parallel processing.
- Use columnar storage.
- Support SQL-based languages.

You need to identify the correct technologies to build the Lambda architecture.

Which technologies should you use? To answer, select the appropriate options in the answer area NOTE: Each correct selection is worth one point.

Architecture requirement	Technology
Data storage	<div style="border: 1px solid black; padding: 5px;"> <div style="border-bottom: 1px solid black; padding-bottom: 5px;">▼</div> <div style="padding: 5px;"> Azure SQL Database Azure Blob Storage Azure Cosmos DB Azure Data Lake Store </div> </div>
Batch processing	<div style="border: 1px solid black; padding: 5px;"> <div style="border-bottom: 1px solid black; padding-bottom: 5px;">▼</div> <div style="padding: 5px;"> HDInsight Spark HDInsight Hadoop Azure Databricks HDInsight Interactive Query </div> </div>
Analytical data store	<div style="border: 1px solid black; padding: 5px;"> <div style="border-bottom: 1px solid black; padding-bottom: 5px;">▼</div> <div style="padding: 5px;"> HDInsight HBase Azure SQL Data Warehouse Azure Analysis Services Azure Cosmos DB </div> </div>

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Data storage: Azure Data Lake Store

A key mechanism that allows Azure Data Lake Storage Gen2 to provide file system performance at object storage scale and prices is the addition of a hierarchical namespace. This allows the collection of objects/files within an account to be organized into a hierarchy of directories and nested subdirectories in the same way that the file system on your computer is organized. With the hierarchical namespace enabled, a storage account becomes capable of providing the scalability and cost-effectiveness of object storage, with file system semantics that are familiar to analytics engines and frameworks.

Batch processing: HD Insight Spark

Apache Spark is an open-source, parallel-processing framework that supports in-memory processing to boost the performance of big-data analysis applications. HDInsight is a managed Hadoop service. Use it to deploy and manage Hadoop clusters in Azure. For batch processing, you can use Spark, Hive, Hive LLAP, MapReduce.

Languages: R, Python, Java, Scala, SQL Analytic data store: SQL Data Warehouse

SQL Data Warehouse is a cloud-based Enterprise Data Warehouse (EDW) that uses Massively Parallel Processing (MPP).

SQL Data Warehouse stores data into relational tables with columnar storage. References:

<https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-namespace> <https://docs.microsoft.com/en-us/azure/architecture/data-guide/technology-choices/batch-processing> <https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-overview-what-is>

NEW QUESTION 67

- (Exam Topic 3)

You have an Azure Synapse Analytics dedicated SQL pool named Pool1 and a database named DB1. DB1 contains a fact table named Table1.

You need to identify the extent of the data skew in Table1. What should you do in Synapse Studio?

- A. Connect to the built-in pool and query sysdm_pdw_sys_info.
- B. Connect to Pool1 and run DBCC CHECKALLOC.
- C. Connect to the built-in pool and run DBCC CHECKALLOC.
- D. Connect to Pool1 and query sys.dm_pdw_nodes_db_partition_stats.

Answer: D

Explanation:

Microsoft recommends use of sys.dm_pdw_nodes_db_partition_stats to analyze any skewness in the data. Reference:

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

NEW QUESTION 68

- (Exam Topic 3)

You have an Azure Data Factory pipeline that contains a data flow. The data flow contains the following expression.

```
source(output(
    License_plate as string,
    Make as string,
    Time as string
),
allowSchemaDrift: true,
```

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

See the answer in
See below answer.

Answer Area

Number of columns: 22

Number of rows: 4

NEW QUESTION 70

- (Exam Topic 3)

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

You are implementing a pattern that batch loads the files daily into an enterprise data warehouse in Azure Synapse Analytics by using PolyBase.

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

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

NOTE: Each correct selection is worth one point

Actions

Answer Area

- Create a database scoped credential that uses Azure Active Directory Application and a Service Principal Key
- Create an external data source that uses the abfs location
- Use CREATE EXTERNAL TABLE AS SELECT (CETAS) and configure the reject options to specify reject values or percentages
- Create an external file format and set the First_Row option



- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

A picture containing timeline Description automatically generated

Step 1: Create an external data source that uses the abfs location

Create External Data Source to reference Azure Data Lake Store Gen 1 or 2 Step 2: Create an external file format and set the First_Row option.

Create External File Format.

Step 3: Use CREATE EXTERNAL TABLE AS SELECT (CETAS) and configure the reject options to specify reject values or percentages

To use PolyBase, you must create external tables to reference your external data. Use reject options.

Note: REJECT options don't apply at the time this CREATE EXTERNAL TABLE AS SELECT statement is run. Instead, they're specified here so that the database

can use them at a later time when it imports data from the external table. Later, when the CREATE TABLE AS SELECT statement selects data from the external table, the database will use the reject options to determine the number or percentage of rows that can fail to import before it stops the import.

Reference:

<https://docs.microsoft.com/en-us/sql/relational-databases/polybase/polybase-t-sql-objects> <https://docs.microsoft.com/en-us/sql/t-sql/statements/create-external-table-as-select-transact-sql>

NEW QUESTION 73

- (Exam Topic 3)

You are creating a new notebook in Azure Databricks that will support R as the primary language but will also support Scala and SQL. Which switch should you use to switch between languages?

- A. @<Language>
- B. %<Language>
- C. \(<Language>)
- D. \(<Language>)

Answer: B

Explanation:

To change the language in Databricks' cells to either Scala, SQL, Python or R, prefix the cell with '%', followed by the language.

%python //or r, scala, sql

Reference:

<https://www.theta.co.nz/news-blogs/tech-blog/enhancing-digital-twins-part-3-predictive-maintenance-with-azur>

NEW QUESTION 74

- (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 75

- (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 77

- (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.

Solution: You use a session window that uses a timeout size of 10 seconds. Does this meet the goal?

- A. Yes
- B. No

Answer: A

Explanation:

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

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

NEW QUESTION 81

- (Exam Topic 3)

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

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

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

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

How should you complete the PySpark code? To answer, drag the appropriate values to the correct targets. Each value may be used once, more than once or not at all. You may need to drag the split bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point.

Values

- alias
- array_union
- createDataFrame
- explode
- select
- translate

Answer Area

```

@utils.fs.put("/tmp/source.json", source_json, True)
source_df = spark.read.option("multiline", "true").json("/tmp/source.json")
persons = source_df. [Value] [Value] ("persons").alias("persons")
persons_dogs = persons.select(col("persons.name").alias("owner"), col("persons.age").alias("age"),
explode [Value] ("dog")
("persons.dogs"),
display(persons_dogs)
                    
```

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Graphical user interface, text, application Description automatically generated

Box 1: select

Box 2: explode

Box 3: alias

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

Reference: <https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.sql.Column.alias.html> <https://docs.microsoft.com/en-us/azure/databricks/sql/language-manual/functions/explode>

NEW QUESTION 83

- (Exam Topic 3)

You have an Azure SQL database named Database1 and two Azure event hubs named HubA and HubB. The data consumed from each source is shown in the following table.

Source	Data
Database1	Driver's name Driver's license number
HubA	Ride route Ride distance Ride duration
HubB	Ride fare Ride payment

You need to implement Azure Stream Analytics to calculate the average fare per mile by driver.

How should you configure the Stream Analytics input for each source? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

HubA: ▼
 Stream
 Reference

HubB: ▼
 Stream
 Reference

Database1: ▼
 Stream
 Reference

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

HubA: Stream HubB: Stream
 Database1: Reference

Reference data (also known as a lookup table) is a finite data set that is static or slowly changing in nature, used to perform a lookup or to augment your data streams. For example, in an IoT scenario, you could store metadata about sensors (which don't change often) in reference data and join it with real time IoT data streams. Azure Stream Analytics loads reference data in memory to achieve low latency stream processing

Reference:

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-use-reference-data>

NEW QUESTION 87

- (Exam Topic 3)

You have an Azure Databricks workspace named workspace1 in the Standard pricing tier. Workspace1 contains an all-purpose cluster named cluster1. You need to reduce the time it takes for cluster 1 to start and scale up. The solution must minimize costs. What should you do first?

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

Answer: C

Explanation:

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

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

Reference:

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

NEW QUESTION 91

- (Exam Topic 3)

You plan to create an Azure Data Factory pipeline that will include a mapping data flow. You have JSON data containing objects that have nested arrays.

You need to transform the JSON-formatted data into a tabular dataset. The dataset must have one row for each item in the arrays.

Which transformation method should you use in the mapping data flow?

- A. unpivot
- B. flatten
- C. new branch
- D. alter row

Answer: B

Explanation:

Use the flatten transformation to take array values inside hierarchical structures such as JSON and unroll them into individual rows. This process is known as denormalization.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/data-flow-flatten>

NEW QUESTION 92

- (Exam Topic 3)

You are designing a highly available Azure Data Lake Storage solution that will include geo-zone-redundant storage (GZRS).

You need to monitor for replication delays that can affect the recovery point objective (RPO). What should you include in the monitoring solution?

- A. availability

- B. Average Success E2E Latency
- C. 5xx: Server Error errors
- D. Last Sync Time

Answer: D

Explanation:

Because geo-replication is asynchronous, it is possible that data written to the primary region has not yet been written to the secondary region at the time an outage occurs. The Last Sync Time property indicates the last time that data from the primary region was written successfully to the secondary region. All writes made to the primary region before the last sync time are available to be read from the secondary location. Writes made to the primary region after the last sync time property may or may not be available for reads yet.

Reference:

<https://docs.microsoft.com/en-us/azure/storage/common/last-sync-time-get>

NEW QUESTION 97

- (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 102

- (Exam Topic 3)

You are designing an Azure Synapse Analytics dedicated SQL pool.

Groups will have access to sensitive data in the pool as shown in the following table.

Name	Enhanced access
Executives	No access to sensitive data
Analysts	Access to in-region sensitive data
Engineers	Access to all numeric sensitive data

You have policies for the sensitive data. The policies vary by region as shown in the following table.

Region	Data considered sensitive
RegionA	Financial, Personally Identifiable Information (PII)
RegionB	Financial, Personally Identifiable Information (PII), medical
RegionC	Financial, medical

You have a table of patients for each region. The tables contain the following potentially sensitive columns.

Name	Sensitive data	Description
CardOnFile	Financial	Debit/credit card number for charges
Height	Medical	Patient's height in cm
ContactEmail	PII	Email address for secure communications

You are designing dynamic data masking to maintain compliance.

For each of the following statements, select Yes if the statement is true. Otherwise, select No.

NOTE: Each correct selection is worth one point.

Statements	Yes	No
Analysts in RegionA require dynamic data masking rules for [Patients_RegionA].	<input type="radio"/>	<input type="radio"/>
Engineers in RegionC require a dynamic data masking rule for [Patients_RegionA], [Height]	<input type="radio"/>	<input type="radio"/>
Engineers in RegionB require a dynamic data masking rule for [Patients_RegionB], [Height]	<input type="radio"/>	<input type="radio"/>

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Text Description automatically generated

Reference:

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

NEW QUESTION 106

- (Exam Topic 3)

You are designing an Azure Databricks cluster that runs user-defined local processes. You need to recommend a cluster configuration that meets the following requirements:

- Minimize query latency.
- Maximize the number of users that can run queues on the cluster at the same time « Reduce overall costs without compromising other requirements

Which cluster type should you recommend?

- A. Standard with Auto termination
- B. Standard with Autoscaling
- C. High Concurrency with Autoscaling
- D. High Concurrency with Auto Termination

Answer: C

Explanation:

A High Concurrency cluster is a managed cloud resource. The key benefits of High Concurrency clusters are that they provide fine-grained sharing for maximum resource utilization and minimum query latencies.

Databricks chooses the appropriate number of workers required to run your job. This is referred to as autoscaling. Autoscaling makes it easier to achieve high cluster utilization, because you don't need to provision the cluster to match a workload.

Reference:

<https://docs.microsoft.com/en-us/azure/databricks/clusters/configure>

NEW QUESTION 109

- (Exam Topic 3)

You are planning a streaming data solution that will use Azure Databricks. The solution will stream sales transaction data from an online store. The solution has the following specifications:

- * The output data will contain items purchased, quantity, line total sales amount, and line total tax amount.
- * Line total sales amount and line total tax amount will be aggregated in Databricks.
- * Sales transactions will never be updated. Instead, new rows will be added to adjust a sale.

You need to recommend an output mode for the dataset that will be processed by using Structured Streaming. The solution must minimize duplicate data.

What should you recommend?

- A. Append
- B. Update
- C. Complete

Answer: B

Explanation:

By default, streams run in append mode, which adds new records to the table. <https://docs.databricks.com/delta/delta-streaming.html>

NEW QUESTION 113

- (Exam Topic 3)

You are designing an Azure Synapse Analytics workspace.

You need to recommend a solution to provide double encryption of all the data at rest.

Which two components should you include in the recommendation? Each correct answer presents part of the solution

NOTE: Each correct selection is worth one point.

- A. an X509 certificate
- B. an RSA key
- C. an Azure key vault that has purge protection enabled
- D. an Azure virtual network that has a network security group (NSG)
- E. an Azure Policy initiative

Answer: BC

Explanation:

Synapse workspaces encryption uses existing keys or new keys generated in Azure Key Vault. A single key is used to encrypt all the data in a workspace.

Synapse workspaces support RSA 2048 and 3072 byte-sized keys, and RSA-HSM keys.

The Key Vault itself needs to have purge protection enabled. Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/security/workspaces-encryption>

NEW QUESTION 116

- (Exam Topic 3)

You are designing a folder structure for the files in an Azure Data Lake Storage Gen2 account. The account has one container that contains three years of data.

You need to recommend a folder structure that meets the following requirements:

- Supports partition elimination for queries by Azure Synapse Analytics serverless SQL pool
- Supports fast data retrieval for data from the current month

• Simplifies data security management by department Which folder structure should you recommend?

- A. \YYY\MM\DD\Department\DataSource\DataFile_YYYYMMDD.parquet
- B. \Depdfment\DataSource\YYY\MM\DataFile_YYYYMMDD.parquet
- C. \DD\MM\YYYY\Department\DataSource\DataFile_DDMMYY.parquet
- D. \DataSource\Department\YYYYMM\DataFile_YYYYMMDD.parquet

Answer: B

Explanation:

Department top level in the hierarchy to simplify security management.
 Month (MM) at the leaf/bottom level to support fast data retrieval for data from the current month.

NEW QUESTION 117

- (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 Get Metadata activity that retrieves the DateTime of the files.

Does this meet the goal?

- A. Yes
- B. No

Answer: B

Explanation:

Instead use a serverless SQL pool to create an external table with the extra column. Reference:
<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/create-use-external-tables>

NEW QUESTION 118

- (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 123

- (Exam Topic 3)

You have an enterprise data warehouse in Azure Synapse Analytics.

You need to monitor the data warehouse to identify whether you must scale up to a higher service level to accommodate the current workloads

Which is the best metric to monitor?

More than one answer choice may achieve the goal. Select the BEST answer.

- A. Data 10 percentage
- B. CPU percentage
- C. DWU used
- D. DWU percentage

Answer: D

NEW QUESTION 128

- (Exam Topic 3)

You have an Azure Data Lake Storage account that contains a staging zone.

You need to design a daily process to ingest incremental data from the staging zone, transform the data by executing an R script, and then insert the transformed data into a data warehouse in Azure Synapse Analytics.

Solution: You use an Azure Data Factory schedule trigger to execute a pipeline that executes mapping data Flow, and then inserts the data into the data warehouse.

Does this meet the goal?

- A. Yes
- B. No

Answer: B

Explanation:

If you need to transform data in a way that is not supported by Data Factory, you can create a custom activity, not a mapping flow, with your own data processing logic and use the activity in the pipeline. You can create a custom activity to run R scripts on your HDInsight cluster with R installed.

Reference:

<https://docs.microsoft.com/en-US/azure/data-factory/transform-data>

NEW QUESTION 132

- (Exam Topic 3)

You are designing a star schema for a dataset that contains records of online orders. Each record includes an order date, an order due date, and an order ship date.

You need to ensure that the design provides the fastest query times of the records when querying for arbitrary date ranges and aggregating by fiscal calendar attributes.

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

NOTE: Each correct selection is worth one point.

- A. Create a date dimension table that has a DateTime key.
- B. Use built-in SQL functions to extract date attributes.
- C. Create a date dimension table that has an integer key in the format of yyyyymmdd.
- D. In the fact table, use integer columns for the date fields.
- E. Use DateTime columns for the date fields.

Answer: BD

NEW QUESTION 137

- (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 138

- (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 workspac
- D. From Monitor, select Apache Sparks applications.
- E. From Synapse Studio, select the workspac
- F. From Monitor, select SQL requests.

Answer: C

Explanation:

Use Synapse Studio to monitor your Apache Spark applications. To monitor running Apache Spark application Open Monitor, then select Apache Spark applications. To view the details about the Apache Spark applications that are running, select the submitting Apache Spark application and view the details. If the Apache Spark application is still running, you can monitor the progress.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/monitoring/apache-spark-applications>

NEW QUESTION 142

- (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 143

- (Exam Topic 3)

You have an Azure Synapse Analytics dedicated SQL pool.

You run PDW_SHOWSPACEUSED(dbo,FactInternetSales'); and get the results shown in the following table.

ROWS	RESERVED_SPACE	DATA_SPACE	INDEX_SPACE	UNUSED_SPACE	PDH_NODE_ID	DISTRIBUTION_ID
694	2776	616	48	2112	1	1
407	2704	576	48	2080	1	2
53	2376	512	16	1848	1	3
58	2376	512	16	1848	1	4
168	2632	528	32	2072	1	5
195	2696	536	32	2128	1	6
5995	3464	1424	32	2008	1	7
0	2232	496	0	1736	1	8
264	2576	544	48	1992	1	9
3008	3016	960	32	2024	1	10
--	--	--	--	--	--	--
1550	2832	752	48	2032	1	50
1238	2832	696	40	2096	1	51
192	2632	528	32	2072	1	52
1127	2768	680	48	2040	1	53
1244	3032	784	64	2264	1	54
409	2632	568	32	2032	1	55
0	2232	496	0	1736	1	56
1417	2832	728	40	2064	1	57
0	2232	496	0	1736	1	58
384	2632	560	32	2040	1	59
225	2768	544	40	2184	1	60

Which statement accurately describes the dbo,FactInternetSales table?

- A. The table contains less than 1,000 rows.
- B. All distributions contain data.
- C. The table is skewed.
- D. The table uses round-robin distribution.

Answer: B

Explanation:

Data skew means the data is not distributed evenly across the distributions. Reference:
<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribu>

NEW QUESTION 148

- (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 150

- (Exam Topic 3)

You have an enterprise data warehouse in Azure Synapse Analytics.

Using PolyBase, you create an external table named [Ext].[Items] to query Parquet files stored in Azure Data Lake Storage Gen2 without importing the data to the data warehouse.

The external table has three columns.

You discover that the Parquet files have a fourth column named ItemID.

Which command should you run to add the ItemID column to the external table?

- A.

```
ALTER EXTERNAL TABLE [Ext].[Items]
  ADD [ItemID] int;
```
- B.

```
DROP EXTERNAL FILE FORMAT parquetfile1;
CREATE EXTERNAL FILE FORMAT parquetfile1
WITH (
  FORMAT_TYPE = PARQUET,
  DATA_COMPRESSION = 'org.apache.hadoop.io.compress.SnappyCodec'
);
```
- C.

```
DROP EXTERNAL TABLE [Ext].[Items]
CREATE EXTERNAL TABLE [Ext].[Items]
([ItemID] [int] NULL,
 [ItemName] nvarchar(50) NULL,
 [ItemType] nvarchar(20) NULL,
 [ItemDescription] nvarchar(250))
WITH
(
  LOCATION= '/Items/',
  DATA_SOURCE = AzureDataLakeStore,
  FILE_FORMAT = PARQUET,
  REJECT_TYPE = VALUE,
  REJECT_VALUE = 0
);
```
- D.

```
ALTER TABLE [Ext].[Items]
  ADD [ItemID] int;
```

- A. Option A
- B. Option B
- C. Option C
- D. Option D

Answer: C

Explanation:

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

NEW QUESTION 152

- (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: A

NEW QUESTION 157

- (Exam Topic 3)

You have an Azure Synapse Analytics pipeline named Pipeline1 that contains a data flow activity named Dataflow1.

Pipeline1 retrieves files from an Azure Data Lake Storage Gen 2 account named storage1.
 Dataflow1 uses the AutoResolveIntegrationRuntime integration runtime configured with a core count of 128. You need to optimize the number of cores used by Dataflow1 to accommodate the size of the files in storage1. What should you configure? To answer, select the appropriate options in the answer area.

To Pipeline1, add:

A custom activity
A Get Metadata activity
An If Condition activity

For Dataflow1, set the core count by using:

Dynamic content
Parameters
User properties

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Box 1: A Get Metadata activity

Dynamically size data flow compute at runtime

The Core Count and Compute Type properties can be set dynamically to adjust to the size of your incoming source data at runtime. Use pipeline activities like Lookup or Get Metadata in order to find the size of the source dataset data. Then, use Add Dynamic Content in the Data Flow activity properties.

Box 2: Dynamic content

Reference: <https://docs.microsoft.com/en-us/azure/data-factory/control-flow-execute-data-flow-activity>

NEW QUESTION 160

- (Exam Topic 3)

You are designing an Azure Databricks table. The table will ingest an average of 20 million streaming events per day.

You need to persist the events in the table for use in incremental load pipeline jobs in Azure Databricks. The solution must minimize storage costs and incremental load times.

What should you include in the solution?

- A. Partition by DateTime fields.
- B. Sink to Azure Queue storage.
- C. Include a watermark column.
- D. Use a JSON format for physical data storage.

Answer: A

Explanation:

The Databricks ABS-AQS connector uses Azure Queue Storage (AQS) to provide an optimized file source that lets you find new files written to an Azure Blob storage (ABS) container without repeatedly listing all of the files.

This provides two major advantages:

- > Lower costs: no more costly LIST API requests made to ABS.

Reference:

<https://docs.microsoft.com/en-us/azure/databricks/spark/latest/structured-streaming/aqs>

NEW QUESTION 163

- (Exam Topic 3)

You have an Azure Synapse Analytics dedicated SQL pool named pool1.

You plan to implement a star schema in pool1 and create a new table named DimCustomer by using the following code.

```
CREATE TABLE dbo.[DimCustomer](
    [CustomerKey] int NOT NULL,
    [CustomerSourceID] [int] NOT NULL,
    [Title] [nvarchar](8) NULL,
    [FirstName] [nvarchar](50) NOT NULL,
    [MiddleName] [nvarchar](50) NULL,
    [LastName] [nvarchar](50) NOT NULL,
    [Suffix] [nvarchar](10) NULL,
    [CompanyName] [nvarchar](128) NULL,
    [SalesPerson] [nvarchar](256) NULL,
    [EmailAddress] [nvarchar](50) NULL,
    [Phone] [nvarchar](25) NULL,
    [InsertedDate] [datetime] NOT NULL,
    [ModifiedDate] [datetime] NOT NULL,
    [HashKey] [varchar](100) NOT NULL,
    [IsCurrentRow] [bit] NOT NULL
)
WITH
(
    DISTRIBUTION = REPLICATE,
    CLUSTERED COLUMNSTORE INDEX
);
GO
```

You need to ensure that DimCustomer has the necessary columns to support a Type 2 slowly changing dimension (SCD). Which two columns should you add? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

- A. [HistoricalSalesPerson] [nvarchar] (256) NOT NULL
- B. [EffectiveEndDate] [datetime] NOT NULL
- C. [PreviousModifiedDate] [datetime] NOT NULL
- D. [RowID] [bigint] NOT NULL
- E. [EffectiveStartDate] [datetime] NOT NULL

Answer: AB

NEW QUESTION 168

- (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.

```
df.write
```

<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/> .bucketBy	<input type="checkbox"/> ("*")
<input type="checkbox"/> .format	<input type="checkbox"/> ("GeographyRegionID")
<input type="checkbox"/> .partitionBy	<input type="checkbox"/> ("GeographyRegionID", "Year", "Month", "Day")
<input type="checkbox"/> .sortBy	<input type="checkbox"/> ("Year", "Month", "Day", "GeographyRegionID")

```
.mode ("append")
```

<input type="checkbox"/>
<input type="checkbox"/> .csv("/DBTBL1")
<input type="checkbox"/> .json("/DBTBL1")
<input type="checkbox"/> .parquet("/DBTBL1")
<input type="checkbox"/> .saveAsTable("/DBTBL1")

- A. Mastered
- B. Not Mastered

Answer: A

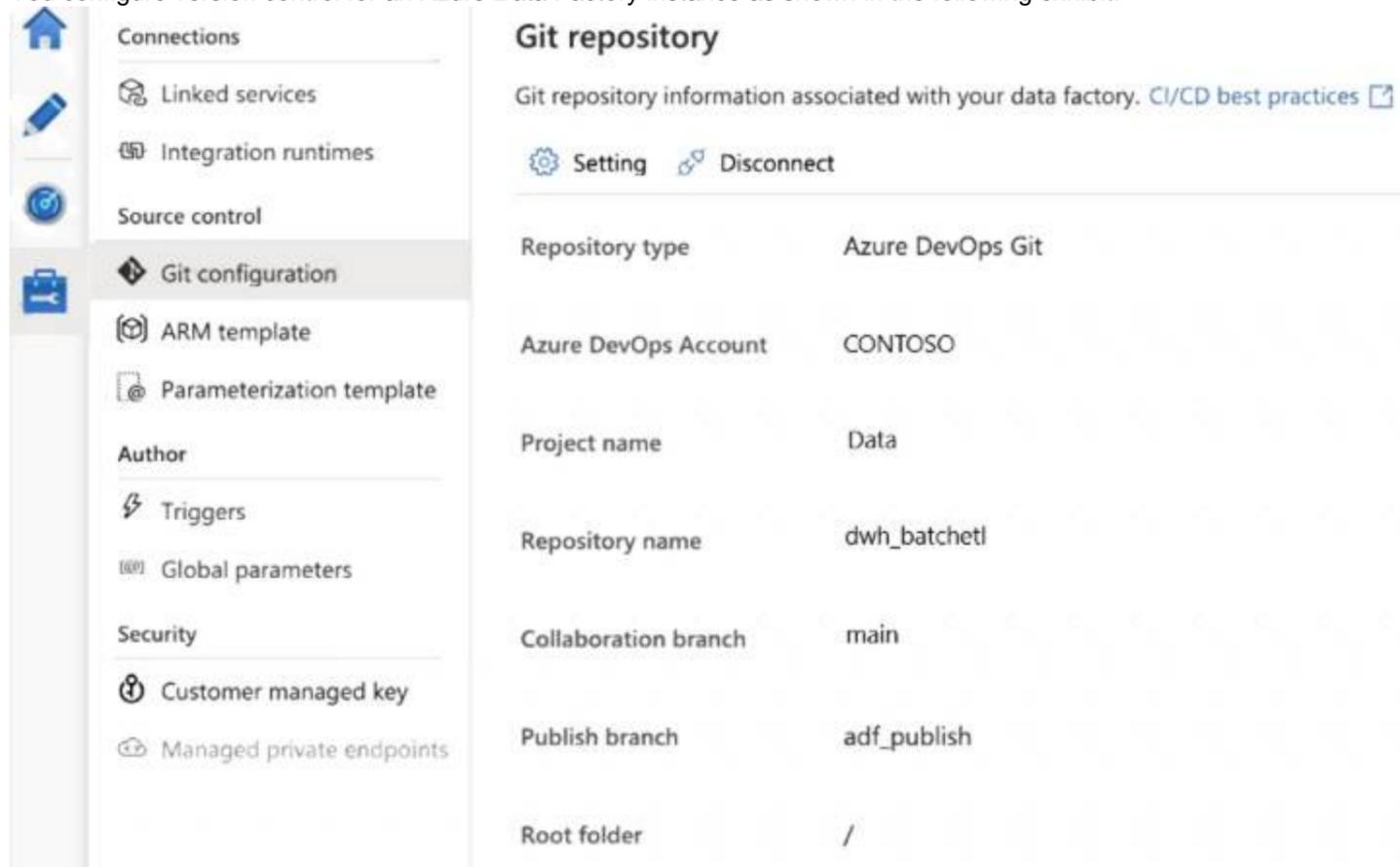
Explanation:

Graphical user interface, text, application Description automatically generated

NEW QUESTION 171

- (Exam Topic 3)

You configure version control for an Azure Data Factory instance as 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.

Azure Resource Manager (ARM) templates for the pipeline assets are stored in [answer choice]

	▼
/	
adf_publish	
main	
Parameterization template	

A Data Factory Azure Resource Manager (ARM) template named contososales can be found in [answer choice]

	▼
/	
/contososales	
/dwh_batchetl/adf_publish/contososales	
/main	

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Letter Description automatically generated

Box 1: adf_publish

The Publish branch is the branch in your repository where publishing related ARM templates are stored and updated. By default, it's adf_publish.

Box 2: / dwh_batchetl/adf_publish/contososales

Note: RepositoryName (here dwh_batchetl): Your Azure Repos code repository name. Azure Repos projects contain Git repositories to manage your source code as your project grows. You can create a new repository or use an existing repository that's already in your project.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/source-control>

NEW QUESTION 174

- (Exam Topic 3)

You have an Azure Data Factory pipeline named Pipeline1!. Pipeline1 contains a copy activity that sends data to an Azure Data Lake Storage Gen2 account.

Pipeline 1 is executed by a schedule trigger.

You change the copy activity sink to a new storage account and merge the changes into the collaboration branch.

After Pipeline1 executes, you discover that data is NOT copied to the new storage account. You need to ensure that the data is copied to the new storage account.

What should you do?

- A. Publish from the collaboration branch.
- B. Configure the change feed of the new storage account.
- C. Create a pull request.
- D. Modify the schedule trigger.

Answer: A

Explanation:

CI/CD lifecycle

- > A development data factory is created and configured with Azure Repos Git. All developers should have permission to author Data Factory resources like pipelines and datasets.
- > A developer creates a feature branch to make a change. They debug their pipeline runs with their most recent changes
- > After a developer is satisfied with their changes, they create a pull request from their feature branch to the main or collaboration branch to get their changes reviewed by peers.
- > After a pull request is approved and changes are merged in the main branch, the changes get published to the development factory.

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

NEW QUESTION 176

- (Exam Topic 3)

The following code segment is used to create an Azure Databricks cluster.

```
{
  "num_workers": null,
  "autoscale": {
    "min_workers": 2,
    "max_workers": 8
  },
  "cluster_name": "MyCluster",
  "spark_version": "latest-stable-scala2.11",
  "spark_conf": {
    "spark.databricks.cluster.profile": "serverless",
    "spark.databricks.repl.allowedLanguages": "sql,python,r"
  },
  "node_type_id": "Standard_DS13_v2",
  "ssh_public_keys": [],
  "custom_tags": {
    "ResourceClass": "Serverless"
  },
  "spark_env_vars": {
    "PYSPARK_PYTHON": "/databricks/python3/bin/python3"
  },
  "autotermination_minutes": 90,
  "enable_elastic_disk": true,
  "init_scripts": []
}
```

For each of the following statements, select Yes if the statement is true. Otherwise, select No.

NOTE: Each correct selection is worth one point.

Statements	Yes	No
The Databricks cluster supports multiple concurrent users.	<input type="radio"/>	<input type="radio"/>
The Databricks cluster minimizes costs when running scheduled jobs that execute notebooks.	<input type="radio"/>	<input type="radio"/>
The Databricks cluster supports the creation of a Delta Lake table.	<input type="radio"/>	<input type="radio"/>

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Graphical user interface, text, application Description automatically generated

Box 1: Yes

A cluster mode of 'High Concurrency' is selected, unlike all the others which are 'Standard'. This results in a worker type of Standard_DS13_v2.

Box 2: No

When you run a job on a new cluster, the job is treated as a data engineering (job) workload subject to the job workload pricing. When you run a job on an existing cluster, the job is treated as a data analytics (all-purpose) workload subject to all-purpose workload pricing.

Box 3: Yes

Delta Lake on Databricks allows you to configure Delta Lake based on your workload patterns. Reference:

<https://adatis.co.uk/databricks-cluster-sizing/> <https://docs.microsoft.com/en-us/azure/databricks/jobs>

<https://docs.databricks.com/administration-guide/capacity-planning/cmbp.html> <https://docs.databricks.com/delta/index.html>

NEW QUESTION 177

- (Exam Topic 3)

You have the following Azure Stream Analytics query.

WITH

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

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

For each of the following statements, select Yes if the statement is true. Otherwise, select No.

NOTE: Each correct selection is worth one point.

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

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Box 1: No

Note: 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

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

In general, the best practice is to start with 6 SUs for queries that don't use PARTITION BY. Here there are 10 partitions, so 6x10 = 60 SUs is good.

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

Reference:

<https://azure.microsoft.com/en-in/blog/maximize-throughput-with-repartitioning-in-azure-stream-analytics/> <https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-streaming-unit-consumption>

NEW QUESTION 178

- (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 181

- (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: Move to cool storage Box 2: Move to archive storage

Archive - Optimized for storing data that is rarely accessed and stored for at least 180 days with flexible latency requirements, on the order of hours. The following table shows a comparison of premium performance block blob storage, and the hot, cool, and archive access tiers.

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

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

NEW QUESTION 182

- (Exam Topic 3)

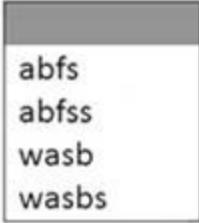
You have an Azure subscription that contains an Azure Synapse Analytics dedicated SQL pool named Pool1 and an Azure Data Lake Storage account named storage1. Storage1 requires secure transfers.

You need to create an external data source in Pool1 that will be used to read .orc files in storage1. How should you complete the code? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

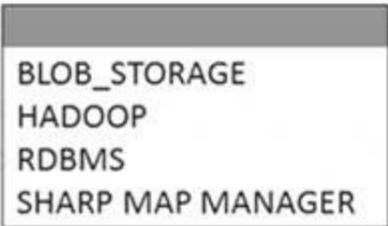
Answer Area

```
CREATE EXTERNAL DATA SOURCE AzureDataLakeStore
```

```
WITH
```

```
( Location1 '  ://data@newyorktaxidataset.dfs.core.windows.net' ,
```

```
credential = ADLS_credential ,
```

```
TYPE - 
```

```
);
```

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Graphical user interface, text, application, email Description automatically generated

Reference:

<https://docs.microsoft.com/en-us/sql/t-sql/statements/create-external-data-source-transact-sql?view=azure-sqldw>

NEW QUESTION 185

- (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

Explanation:

Reference:

<https://docs.microsoft.com/en-us/sql/integration-services/data-flow/transformations/slowly-changing-dimension>

NEW QUESTION 189

- (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 190

- (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: B

Explanation:

Dynamic data masking helps prevent unauthorized access to sensitive data by enabling customers to designate how much of the sensitive data to reveal with minimal impact on the application layer. It's a policy-based security feature that hides the sensitive data in the result set of a query over designated database fields, while the data in the database is not changed.
 Reference:
<https://docs.microsoft.com/en-us/azure/azure-sql/database/dynamic-data-masking-overview>

NEW QUESTION 193

- (Exam Topic 3)
 You have an Azure Data Lake Storage Gen 2 account named storage1.
 You need to recommend a solution for accessing the content in storage1. The solution must meet the following requirements:

- > List and read permissions must be granted at the storage account level.
- > Additional permissions can be applied to individual objects in storage1.
- > Security principals from Microsoft Azure Active Directory (Azure AD), part of Microsoft Entra, must be used for authentication.

What should you use? To answer, drag the appropriate components to the correct requirements. Each component may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point.

Components	Answer Area
Access control lists (ACLs)	To grant permissions at the storage account level: <input style="width: 100%;" type="text"/> To grant permissions at the object level: <input style="width: 100%;" type="text"/>
Role-based access control (RBAC) roles	
Shared access signatures (SAS)	
Shared account keys	

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Box 1: Role-based access control (RBAC) roles
 List and read permissions must be granted at the storage account level.
 Security principals from Microsoft Azure Active Directory (Azure AD), part of Microsoft Entra, must be used for authentication.
 Role-based access control (Azure RBAC)
 Azure RBAC uses role assignments to apply sets of permissions to security principals. A security principal is an object that represents a user, group, service principal, or managed identity that is defined in Azure Active Directory (AD). A permission set can give a security principal a "coarse-grain" level of access such as read or write access to all of the data in a storage account or all of the data in a container.
 Box 2: Access control lists (ACLs)
 Additional permissions can be applied to individual objects in storage1. Access control lists (ACLs)
 ACLs give you the ability to apply "finer grain" level of access to directories and files. An ACL is a permission construct that contains a series of ACL entries. Each ACL entry associates security principal with an access level.
 Reference: <https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control-model>

NEW QUESTION 196

- (Exam Topic 3)

You are designing a sales transactions table in an Azure Synapse Analytics dedicated SQL pool. The table will contain 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 200

- (Exam Topic 3)

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

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

You need to create a data source in Pool1 that you can reference when you create the external table. How should you complete the Transact-SQL statement? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

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

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Graphical user interface, diagram Description automatically generated

Box 1: blob

The following example creates an external data source for Azure Data Lake Gen2 CREATE EXTERNAL DATA SOURCE YellowTaxi WITH (LOCATION = 'https://azureopendatastorage.blob.core.windows.net/nyctlc/yellow/', TYPE = HADOOP)

Box 2: HADOOP

Reference:

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

NEW QUESTION 205

- (Exam Topic 3)

You plan to perform batch processing in Azure Databricks once daily. Which type of Databricks cluster should you use?

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

Answer: B

Explanation:

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

Example: Scheduled batch workloads (data engineers running ETL jobs)

This scenario involves running batch job JARs and notebooks on a regular cadence through the Databricks platform.

The suggested best practice is to launch a new cluster for each run of critical jobs. This helps avoid any issues (failures, missing SLA, and so on) due to an existing workload (noisy neighbor) on a shared cluster.

Reference:

<https://docs.databricks.com/administration-guide/cloud-configurations/aws/cmbp.html#scenario-3-scheduled-bat>

NEW QUESTION 206

- (Exam Topic 2)

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 208

- (Exam Topic 1)

You need to design the partitions for the product sales transactions. The solution must meet the sales transaction dataset requirements.

What should you include in the solution? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

Partition product sales

transactions data by:

	▼
Sales date	
Product ID	
Promotion ID	

Store product sales

transactions data in:

	▼
An Azure Synapse Analytics dedicated SQL pool	
An Azure Synapse Analytics serverless SQL pool	
An Azure Data Lake Storage Gen2 account linked to an Azure Synapse Analytics workspace	

- 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 211

.....

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](#)