

Exam Questions DP-203

Data Engineering on Microsoft Azure

<https://www.2passeasy.com/dumps/DP-203/>



NEW QUESTION 1

- (Exam Topic 3)

You have an Azure Data Lake Storage Gen2 account that contains a JSON file for customers. The file contains two attributes named FirstName and LastName. You need to copy the data from the JSON file to an Azure Synapse Analytics table by using Azure Databricks. A new column must be created that concatenates the FirstName and LastName values.

You create the following components:

- > A destination table in Azure Synapse
- > An Azure Blob storage container
- > A service principal

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

Actions

Answer Area

Mount the Data Lake Storage onto DBFS.

Write the results to a table in Azure Synapse.

Perform transformations on the file.

Specify a temporary folder to stage the data.

Write the results to Data Lake Storage.

Read the file into a data frame.

Drop the data frame.

Perform transformations on the data frame.

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

- 1) mount onto DBFS
- 2) read into data frame
- 3) transform data frame
- 4) specify temporary folder
- 5) write the results to table in in Azure Synapse <https://docs.databricks.com/data/data-sources/azure/azure-datalake-gen2.html>
<https://docs.microsoft.com/en-us/azure/databricks/scenarios/databricks-extract-load-sql-data-warehouse>

NEW QUESTION 2

- (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 3

- (Exam Topic 3)

You are designing a fact table named FactPurchase in an Azure Synapse Analytics dedicated SQL pool. The table contains purchases from suppliers for a retail store. FactPurchase will contain the following columns.

Name	Data type	Nullable
PurchaseKey	Bigint	No
DateKey	Int	No
SupplierKey	Int	No
StockItemKey	Int	No
PurchaseOrderID	Int	Yes
OrderedQuantity	Int	No
OrderedOuters	Int	No
ReceivedOuters	Int	No
Package	Nvarchar(50)	No
IsOrderFinalized	Bit	No
LineageKey	Int	No

FactPurchase will have 1 million rows of data added daily and will contain three years of data. Transact-SQL queries similar to the following query will be executed daily.

```
SELECT
SupplierKey, StockItemKey, COUNT(*) FROM FactPurchase
WHERE DateKey >= 20210101 AND DateKey <= 20210131
GROUP By SupplierKey, StockItemKey
```

Which table distribution will minimize query times?

- A. round-robin
- B. replicated
- C. hash-distributed on DateKey
- D. hash-distributed on PurchaseKey

Answer: D

Explanation:

Hash-distributed tables improve query performance on large fact tables, and are the focus of this article. Round-robin tables are useful for improving loading speed.

Reference:

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

NEW QUESTION 4

- (Exam Topic 3)

You are processing streaming data from vehicles that pass through a toll booth.

You need to use Azure Stream Analytics to return the license plate, vehicle make, and hour the last vehicle passed during each 10-minute window.

How should you complete the query? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

```
WITH LastInWindow AS
(
    SELECT
        (Time) AS LastEventTime
    FROM
        Input TIMESTAMP BY Time
    GROUP BY
        (minute, 10)
)
SELECT
    Input.License_plate,
    Input.Make,
    Input.Time
FROM
    Input TIMESTAMP BY Time
    INNER JOIN LastInWindow
    ON (minute, Input, LastInWindow) BETWEEN 0 AND 10
    AND Input.Time = LastInWindow.LastEventTime
```

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Graphical user interface, text, application Description automatically generated

Box 1: MAX

The first step on the query finds the maximum time stamp in 10-minute windows, that is the time stamp of the last event for that window. The second step joins the results of the first query with the original stream to find the event that match the last time stamps in each window.

Query:

```
WITH LastInWindow AS (
SELECT
MAX(Time) AS LastEventTime FROM
Input TIMESTAMP BY Time GROUP BY
TumblingWindow(minute, 10)
) SELECT
Input.License_plate, Input.Make, Input.Time
FROM
Input TIMESTAMP BY Time INNER JOIN LastInWindow
ON DATEDIFF(minute, Input, LastInWindow) BETWEEN 0 AND 10 AND Input.Time = LastInWindow.LastEventTime
```

Box 2: TumblingWindow

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

Box 3: DATEDIFF

DATEDIFF is a date-specific function that compares and returns the time difference between two DateTime fields, for more information, refer to date functions.

Reference:

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

NEW QUESTION 5

- (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 6

- (Exam Topic 3)

You have an Azure Databricks workspace named workspace1 in the Standard pricing tier.

You need to configure workspace1 to support autoscaling all-purpose clusters. The solution must meet the following requirements:

- > Automatically scale down workers when the cluster is underutilized for three minutes.
- > Minimize the time it takes to scale to the maximum number of workers.
- > Minimize costs. What should you do first?

- A. Enable container services for workspace1.
- B. Upgrade workspace1 to the Premium pricing tier.
- C. Set Cluster Mode to High Concurrency.
- D. Create a cluster policy in workspace1.

Answer: B

Explanation:

For clusters running Databricks Runtime 6.4 and above, optimized autoscaling is used by all-purpose clusters in the Premium plan

Optimized autoscaling:

Scales up from min to max in 2 steps.

Can scale down even if the cluster is not idle by looking at shuffle file state. Scales down based on a percentage of current nodes.

On job clusters, scales down if the cluster is underutilized over the last 40 seconds.

On all-purpose clusters, scales down if the cluster is underutilized over the last 150 seconds.

The spark.databricks.aggressiveWindowDownS Spark configuration property specifies in seconds how often a cluster makes down-scaling decisions. Increasing the value causes a cluster to scale down more slowly. The maximum value is 600.

Note: Standard autoscaling

Starts with adding 8 nodes. Thereafter, scales up exponentially, but can take many steps to reach the max. You can customize the first step by setting the spark.databricks.autoscaling.standardFirstStepUp Spark configuration property.

Scales down only when the cluster is completely idle and it has been underutilized for the last 10 minutes. Scales down exponentially, starting with 1 node.

Reference: <https://docs.databricks.com/clusters/configure.html>

NEW QUESTION 7

- (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 8

- (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 9

- (Exam Topic 3)

You are designing an Azure Synapse Analytics dedicated SQL pool.

You need to ensure that you can audit access to Personally Identifiable information (PII). What should you include in the solution?

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

Answer: C

Explanation:

Data Discovery & Classification is built into Azure SQL Database, Azure SQL Managed Instance, and Azure Synapse Analytics. It provides basic capabilities for discovering, classifying, labeling, and reporting the sensitive data in your databases.

Your most sensitive data might include business, financial, healthcare, or personal information. Discovering and classifying this data can play a pivotal role in your organization's information-protection approach. It can serve as infrastructure for:

- > Helping to meet standards for data privacy and requirements for regulatory compliance.
- > Various security scenarios, such as monitoring (auditing) access to sensitive data.
- > Controlling access to and hardening the security of databases that contain highly sensitive data.

Reference:

<https://docs.microsoft.com/en-us/azure/azure-sql/database/data-discovery-and-classification-overview>

NEW QUESTION 10

- (Exam Topic 3)

You have an Azure data factory.

You need to examine the pipeline failures from the last 180 flays. What should you use?

- A. the Activity tog blade for the Data Factory resource
- B. Azure Data Factory activity runs in Azure Monitor
- C. Pipeline runs in the Azure Data Factory user experience
- D. the Resource health blade for the Data Factory resource

Answer: B

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 10

- (Exam Topic 3)

You have an Azure Synapse Analytics dedicated SQL pool mat 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 (TDD)

Answer: B

NEW QUESTION 11

- (Exam Topic 3)

You are designing an application that will use an Azure Data Lake Storage Gen 2 account to store petabytes of license plate photos from toll booths. The account will use zone-redundant storage (ZRS).

You identify the following usage patterns:

- The data will be accessed several times a day during the first 30 days after the data is created. The data must meet an availability SU of 99.9%.
- After 90 days, the data will be accessed infrequently but must be available within 30 seconds.
- After 365 days, the data will be accessed infrequently but must be available within five minutes.

First 30 days:

After 90 days:

After 365 days:

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Box 1: Hot

The data will be accessed several times a day during the first 30 days after the data is created. The data must meet an availability SLA of 99.9%.

Box 2: Cool

After 90 days, the data will be accessed infrequently but must be available within 30 seconds. Data in the Cool tier should be stored for a minimum of 30 days.

When your data is stored in an online access tier (either Hot or Cool), users can access it immediately. The Hot tier is the best choice for data that is in active use, while the Cool tier is ideal for data that is accessed less frequently, but that still must be available for reading and writing.

Box 3: Cool

After 365 days, the data will be accessed infrequently but must be available within five minutes. Reference: <https://docs.microsoft.com/en-us/azure/storage/blobs/access-tiers-overview> <https://docs.microsoft.com/en-us/azure/storage/blobs/archive-rehydrate-overview>

NEW QUESTION 15

- (Exam Topic 3)

You are implementing a batch dataset in the Parquet format.

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

You need to minimize storage costs for the solution. What should you do?

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

Answer: C

Explanation:

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

Reference:

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

NEW QUESTION 17

- (Exam Topic 3)

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

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

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

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

Answer: A

Explanation:

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

We have the formula: $\text{Records}/(\text{Partitions} \times 60) = 1 \text{ million}$ $\text{Partitions} = \text{Records}/(1 \text{ million} \times 60)$

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

Note: Having too many partitions can reduce the effectiveness of clustered columnstore indexes if each partition has fewer than 1 million rows. Dedicated SQL pools automatically partition your data into 60 databases. So, if you create a table with 100 partitions, the result will be 6000 partitions.

Reference:

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

NEW QUESTION 20

- (Exam Topic 3)

You have two fact tables named Flight and Weather. Queries targeting the tables will be based on the join between the following columns.

Table	Column
Flight	ArrivalAirportID ArrivalDateTime
Weather	AirportID ReportDateTime

You need to recommend a solution that maximum query performance. What should you include in the recommendation?

- A. In each table, create a column as a composite of the other two columns in the table.
- B. In each table, create an IDENTITY column.
- C. In the tables, use a hash distribution of ArriveDateTime and ReportDateTime.
- D. In the tables, use a hash distribution of ArriveAirPortID and AirportID.

Answer: D

NEW QUESTION 25

- (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
<input type="text" value="CASE"/>	*,
<input type="text" value="ELSE"/>	WHEN hire_date >= '2019-01-01' THEN 'New'
<input type="text" value="OVER"/>	<input type="text" value="Standard"/>
<input type="text" value="PARTITION BY"/>	END AS employee_type
<input type="text" value="ROW_NUMBER"/>	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 26

- (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 28

- (Exam Topic 3)

You are developing an application that uses Azure Data Lake Storage Gen 2.

You need to recommend a solution to grant permissions to a specific application for a limited time period. What should you include in the recommendation?

- A. Azure Active Directory (Azure AD) identities
- B. shared access signatures (SAS)
- C. account keys
- D. role assignments

Answer: B

Explanation:

A shared access signature (SAS) provides secure delegated access to resources in your storage account. With a SAS, you have granular control over how a client can access your data. For example:

What resources the client may access.

What permissions they have to those resources. How long the SAS is valid.

Reference:

<https://docs.microsoft.com/en-us/azure/storage/common/storage-sas-overview>

NEW QUESTION 32

- (Exam Topic 3)

From a website analytics system, you receive data extracts about user interactions such as downloads, link clicks, form submissions, and video plays. The data contains the following columns.

Name	Sample value
Date	15 Jan 2021
EventCategory	Videos
EventAction	Play
EventLabel	Contoso Promotional
ChannelGrouping	Social
TotalEvents	150
UniqueEvents	120
SessionWithEvents	99

You need to design a star schema to support analytical queries of the data. The star schema will contain four tables including a date dimension. To which table should you add each column? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

EventCategory:

▼

DimChannel
DimDate
DimEvent
FactEvents

ChannelGrouping:

▼

DimChannel
DimDate
DimEvent
FactEvents

TotalEvents:

▼

DimChannel
DimDate
DimEvent
FactEvents

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Table Description automatically generate

Box 1: DimEvent

Box 2: DimChannel

Box 3: FactEvents

Fact tables store observations or events, and can be sales orders, stock balances, exchange rates, temperatures, etc

Reference:

<https://docs.microsoft.com/en-us/power-bi/guidance/star-schema>

NEW QUESTION 36

- (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 39

- (Exam Topic 3)

You have an Azure Data Factory instance named ADF1 and two Azure Synapse Analytics workspaces named WS1 and WS2.

ADF1 contains the following pipelines:

- > P1: Uses a copy activity to copy data from a nonpartitioned table in a dedicated SQL pool of WS1 to an Azure Data Lake Storage Gen2 account
- > P2: Uses a copy activity to copy data from text-delimited files in an Azure Data Lake Storage Gen2 account to a nonpartitioned table in a dedicated SQL pool of WS2

You need to configure P1 and P2 to maximize parallelism and performance.

Which dataset settings should you configure for the copy activity if each pipeline? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

P1:

	▼
Set the Copy method to Bulk insert	
Set the Copy method to PolyBase	
Set the Isolation level to Repeatable read	
Set the Partition option to Dynamic range	

P2:

	▼
Set the Copy method to Bulk insert	
Set the Copy method to PolyBase	
Set the Isolation level to Repeatable read	
Set the Partition option to Dynamic range	

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Box 1: Set the Copy method to PolyBase

While SQL pool supports many loading methods including non-Polybase options such as BCP and SQL BulkCopy API, the fastest and most scalable way to load data is through PolyBase. PolyBase is a technology that accesses external data stored in Azure Blob storage or Azure Data Lake Store via the T-SQL language.

Box 2: Set the Copy method to Bulk insert

Polybase not possible for text files. Have to use Bulk insert. Reference:

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

NEW QUESTION 40

- (Exam Topic 3)

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

Which resource provider should you enable?

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

Answer: C

Explanation:

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

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/how-to-create-event-trigger> <https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipeline-execution-triggers>

NEW QUESTION 45

- (Exam Topic 3)

You have an Azure Data Lake Storage Gen2 account that contains two folders named Folder 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 50

- (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 52

- (Exam Topic 3)
You need to implement an Azure Databricks cluster that automatically connects to Azure Data lake Storage Gen2 by using Azure Active Directory (Azure AD) integration. How should you configure the new clutter? To answer, select the appropriate options in the answers area. NOTE: Each correct selection is worth one point.



- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

<https://docs.azuredatabricks.net/spark/latest/data-sources/azure/adls-passthrough.html>

NEW QUESTION 53

- (Exam Topic 3)
You have an Azure Synapse Analytics SQL pool named Pool1 on a logical Microsoft SQL server named Server1.
You need to implement Transparent Data Encryption (TDE) on Pool1 by using a custom key named key1. Which five actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

Actions

Enable TDE on Pool1.

Assign a managed identity to Server1.

Configure key1 as the TDE protector for Server1.

Add key1 to the Azure key vault.

Create an Azure key vault and grant the managed identity permissions to the key vault.

Answer Area

<

>

⬆

⬇

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Graphical user interface, text, application Description automatically generated

Step 1: Assign a managed identity to Server1

You will need an existing Managed Instance as a prerequisite.

Step 2: Create an Azure key vault and grant the managed identity permissions to the vault Create Resource and setup Azure Key Vault.

Step 3: Add key1 to the Azure key vault

The recommended way is to import an existing key from a .pfx file or get an existing key from the vault. Alternatively, generate a new key directly in Azure Key Vault.

Step 4: Configure key1 as the TDE protector for Server1 Provide TDE Protector key

Step 5: Enable TDE on Pool1 Reference:

<https://docs.microsoft.com/en-us/azure/azure-sql/managed-instance/scripts/transparent-data-encryption-byok-po>

NEW QUESTION 57

- (Exam Topic 3)

You have a table in an Azure Synapse Analytics dedicated SQL pool. The table was created by using the following Transact-SQL statement.

```
CREATE TABLE [dbo].[DimEmployee] (
    [EmployeeKey] [int] IDENTITY(1,1) NOT NULL,
    [EmployeeID] [int] NOT NULL,
    [FirstName] [varchar](100) NOT NULL,
    [LastName] [varchar](100) NOT NULL,
    [JobTitle] [varchar](100) NULL,
    [LastHireDate] [date] NULL,
    [StreetAddress] [varchar](500) NOT NULL,
    [City] [varchar](200) NOT NULL,
    [StateProvince] [varchar](50) NOT NULL,
    [Portalcode] [varchar](10) NOT NULL
)
```

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

- Ensure that users can identify the current manager of employees.
- Support creating an employee reporting hierarchy for your entire company.
- Provide fast lookup of the managers' attributes such as name and job title.

Which column should you add to the table?

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

Answer: A

Explanation:

Use the same definition as the EmployeeID column. Reference:

<https://docs.microsoft.com/en-us/analysis-services/tabular-models/hierarchies-ssas-tabular>

NEW QUESTION 58

- (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 an Azure Databricks notebook, 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 an Azure Databricks notebook, 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 63

- (Exam Topic 3)

A company has a real-time data analysis solution that is hosted on Microsoft Azure. The solution uses Azure Event Hub to ingest data and an Azure Stream Analytics cloud job to analyze the data. The cloud job is configured to use 120 Streaming Units (SU).

You need to optimize performance for the Azure Stream Analytics job.

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

NOTE: Each correct selection is worth one point.

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

Answer: DF

Explanation:

Reference:

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

NEW QUESTION 68

- (Exam Topic 3)

You have an Azure Data Lake Storage Gen2 container that contains 100 TB of data.

You need to ensure that the data in the container is available for read workloads in a secondary region if an outage occurs in the primary region. The solution must minimize costs.

Which type of data redundancy should you use?

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

Answer: B

Explanation:

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

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

Reference:

<https://docs.microsoft.com/en-us/azure/storage/common/storage-redundancy>

NEW QUESTION 71

- (Exam Topic 3)

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

The data will be loaded daily to the data lake and will use a folder structure of {Year}/{Month}/{Day}/. You need to design a daily Azure Data Factory data load to minimize the data transfer between the two accounts.

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

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

Answer: BD

Explanation:

Copy data from one place to another. The requirements are : 1- need to minimize transfert and 2- need to adapte data to the destination folder structure. Filter on LastModifiedDate will copy everything that have changed since the latest load while minimizing the data transfert. Specifying the file naming pattern allows to copy data at the right place to the destination Data Lake.

NEW QUESTION 73

- (Exam Topic 3)

You have several Azure Data Factory pipelines that contain a mix of the following types of activities.

- * Wrangling data flow
- * Notebook
- * Copy
- * jar

Which two Azure services should you use to debug the activities? Each correct answer presents part of the solution NOTE: Each correct selection is worth one point.

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

Answer: CE

NEW QUESTION 74

- (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 he 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
Bop 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 76

- (Exam Topic 3)
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You plan to create an Azure Databricks workspace that has a tiered structure. The workspace will contain the following three workloads:
➤ A workload for data engineers who will use Python and SQL.
➤ A workload for jobs that will run notebooks that use Python, Scala, and SOL.
➤ A workload that data scientists will use to perform ad hoc analysis in Scala and R.
The enterprise architecture team at your company identifies the following standards for Databricks environments:
➤ The data engineers must share a cluster.
➤ The job cluster will be managed by using a request process whereby data scientists and data engineers provide packaged notebooks for deployment to the cluster.
➤ All the data scientists must be assigned their own cluster that terminates automatically after 120 minutes of inactivity. Currently, there are three data scientists.
You need to create the Databricks clusters for the workloads.
Solution: You create a Standard cluster for each data scientist, a High Concurrency cluster for the data engineers, and a Standard cluster for the jobs.
Does this meet the goal?

- A. Yes
- B. No

Answer: B

Explanation:
We would need a High Concurrency cluster for the jobs. Note:
Standard clusters are recommended for a single user. Standard can run workloads developed in any language: Python, R, Scala, and SQL.

A high concurrency cluster is a managed cloud resource. The key benefits of high concurrency clusters are that they provide Apache Spark-native fine-grained sharing for maximum resource utilization and minimum query latencies.
 Reference: <https://docs.azuredatabricks.net/clusters/configure.html>

NEW QUESTION 80

- (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 82

- (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 86

- (Exam Topic 3)

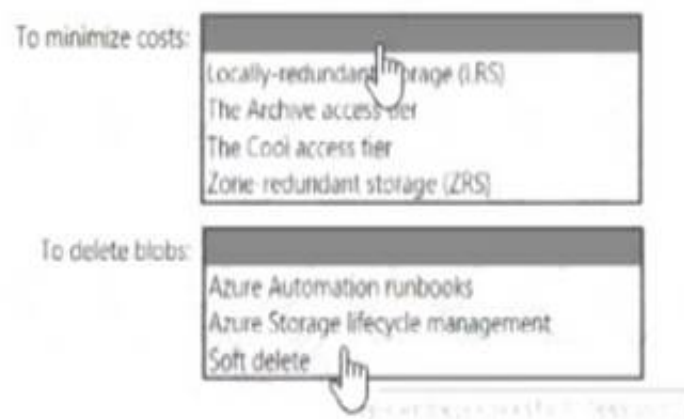
You have an Azure subscription.

You need to deploy an Azure Data Lake Storage Gen2 Premium account. The solution must meet the following requirements:

- Blobs that are older than 365 days must be deleted.
- Administrator efforts must be minimized.
- Costs must be minimized

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

Answer Area



- A. Mastered
- B. Not Mastered

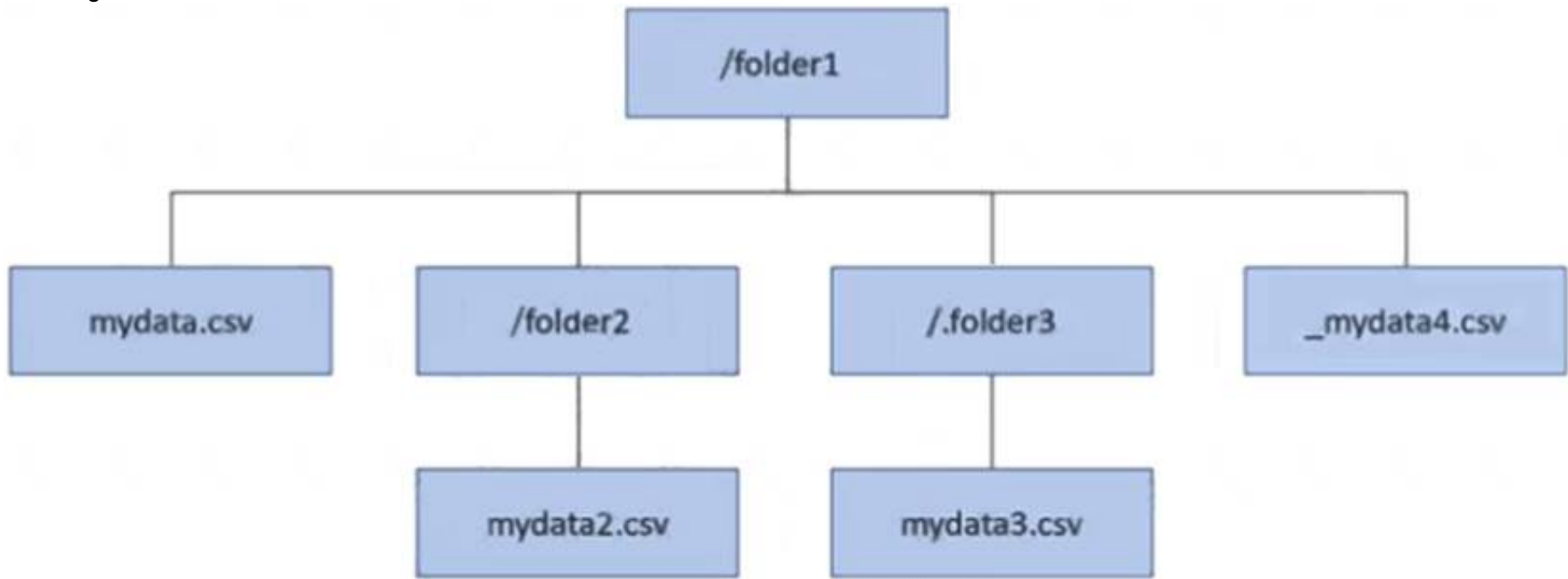
Answer: A

Explanation:

<https://learn.microsoft.com/en-us/azure/storage/blobs/premium-tier-for-data-lake-storage>

NEW QUESTION 88

- (Exam Topic 3)
You have an Azure Data Lake Storage Gen2 account that contains a container named container1. You have an Azure Synapse Analytics serverless SQL pool that contains a native external table named dbo.Table1. The source data for dbo.Table1 is stored in container1. The folder structure of container1 is shown in the following exhibit.



The external data source is defined by using the following statement.

```
CREATE EXTERNAL DATA SOURCE DataLake
WITH
(
    LOCATION          = 'https://mydatalake.dfs.core.windows.net/container1/folder1/**'
    , CREDENTIAL = DataLakeCred
);
```

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
When selecting all the rows in dbo.Table1, data from the mydata2.csv file will be returned.	<input type="radio"/>	<input type="radio"/>
When selecting all the rows in dbo.Table1, data from the mydata3.csv file will be returned.	<input type="radio"/>	<input type="radio"/>
When selecting all the rows in dbo.Table1, data from the _mydata4.csv file will be returned.	<input type="radio"/>	<input type="radio"/>

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Box 1: Yes
In the serverless SQL pool you can also use recursive wildcards /logs/** to reference Parquet or CSV files in any sub-folder beneath the referenced folder.
Box 2: Yes
Box 3: No
Reference: <https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables>

NEW QUESTION 89

- (Exam Topic 3)
You are designing an Azure Data Lake Storage Gen2 structure for telemetry data from 25 million devices distributed across seven key geographical regions. Each minute, the devices will send a JSON payload of metrics to Azure Event Hubs.
You need to recommend a folder structure for the data. The solution must meet the following requirements:
> Data engineers from each region must be able to build their own pipelines for the data of their respective region only.
> The data must be processed at least once every 15 minutes for inclusion in Azure Synapse Analytics serverless SQL pools.
How should you recommend completing the structure? To answer, drag the appropriate values to the correct targets. Each value may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.
NOTE: Each correct selection is worth one point.

Values	Answer Area
{deviceId}	/ Value / Value / Value .json
{mm}/{HH}/{DD}/{MM}/{YYYY}	
{regionID}/{deviceId}	
{regionID}/raw	
{YYYY}/{MM}/{DD}/{HH}	
{YYYY}/{MM}/{DD}/{HH}/{mm}	
raw/{deviceId}	
raw/{regionID}	

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Box 1: {YYYY}/{MM}/{DD}/{HH}

Date Format [optional]: if the date token is used in the prefix path, you can select the date format in which your files are organized. Example: YYYY/MM/DD

Time Format [optional]: if the time token is used in the prefix path, specify the time format in which your files are organized. Currently the only supported value is HH.

Box 2: {regionID}/raw

Data engineers from each region must be able to build their own pipelines for the data of their respective region only.

Box 3: {deviceId} Reference:

<https://github.com/paolosalvatori/StreamAnalyticsAzureDataLakeStore/blob/master/README.md>

NEW QUESTION 92

- (Exam Topic 3)

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

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

What should you do first?

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

Answer: C

Explanation:

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

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

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

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

NEW QUESTION 94

- (Exam Topic 3)

You have an Azure Databricks resource.

You need to log actions that relate to changes in compute for the Databricks resource. Which Databricks services should you log?

- A. clusters
- B. workspace
- C. DBFS
- D. SSHE lobs

Answer: B

Explanation:

Cloud Provider Infrastructure Logs.Databricks logging allows security and admin teams to demonstrate conformance to data governance standards within or from a Databricks workspace. Customers, especially in the regulated industries, also need records on activities like:– User access control to cloud data storage– Cloud Identity and Access Management roles– User access to cloud network and compute

Azure Databricks offers three distinct workloads on several VM Instances tailored for your data analytics workflow—the Jobs Compute and Jobs Light Compute workloads make it easy for data engineers to build and execute jobs, and the All-Purpose Compute workload makes it easy for data scientists to explore, visualize, manipulate, and share data and insights interactively.

NEW QUESTION 96

- (Exam Topic 3)

You are designing an Azure Data Lake Storage Gen2 container to store data for the human resources (HR) department and the operations department at your company. You have the following data access requirements:

- After initial processing, the HR department data will be retained for seven years.
- The operations department data will be accessed frequently for the first six months, and then accessed once per month.

You need to design a data retention solution to meet the access requirements. The solution must minimize storage costs.

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:



NEW QUESTION 101

- (Exam Topic 3)

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

You plan to insert data from the files into Table1 and azure Data Lake Storage Gen2 container named container1.

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

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

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

- A. Yes
- B. No

Answer: A

NEW QUESTION 102

- (Exam Topic 3)

You are designing an inventory updates table in an Azure Synapse Analytics dedicated SQL pool. The table will have a clustered columnstore index and will include the following columns:

Table	Comment
EventDate	One million records are added to the table each day
EventTypeID	The table contains 10 million records for each event type.
WarehouseID	The table contains 100 million records for each warehouse.
ProductCategoryTypeID	The table contains 25 million records for each product category type.

You identify the following usage patterns:

- Analysts will most commonly analyze transactions for a warehouse.
- Queries will summarize by product category type, date, and/or inventory event type. You need to recommend a partition strategy for the table to minimize query times.

On which column should you partition the table?

- A. ProductCategoryTypeID
- B. EventDate
- C. WarehouseID
- D. EventTypeID

Answer: C

Explanation:

The number of records for each warehouse is big enough for a good partitioning.

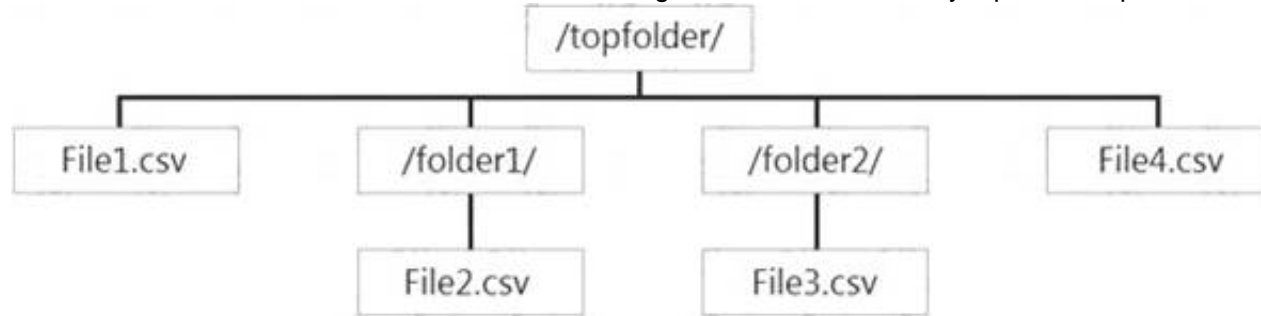
Note: Table partitions enable you to divide your data into smaller groups of data. In most cases, table partitions are created on a date column.

When creating partitions on clustered columnstore tables, it is important to consider how many rows belong to each partition. For optimal compression and performance of clustered columnstore tables, a minimum of 1 million rows per distribution and partition is needed. Before partitions are created, dedicated SQL pool already divides each table into 60 distributed databases.

NEW QUESTION 106

- (Exam Topic 3)

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



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

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

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

Answer: B

Explanation:

To run a T-SQL query over a set of files within a folder or set of folders while treating them as a single entity or rowset, provide a path to a folder or a pattern (using wildcards) over a set of files or folders. Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/query-data-storage#query-multiple-files-or-folders>

NEW QUESTION 110

- (Exam Topic 3)

You are designing an enterprise data warehouse in Azure Synapse Analytics that will contain a table named Customers. Customers will contain credit card information.

You need to recommend a solution to provide salespeople with the ability to view all the entries in Customers. The solution must prevent all the salespeople from viewing or inferring the credit card information.

What should you include in the recommendation?

- A. data masking
- B. Always Encrypted
- C. column-level security
- D. row-level security

Answer: A

Explanation:

SQL Database dynamic data masking limits sensitive data exposure by masking it to non-privileged users. The Credit card masking method exposes the last four digits of the designated fields and adds a constant string as a prefix in the form of a credit card.

Example: XXXX-XXXX-XXXX-1234

Reference:

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

NEW QUESTION 111

- (Exam Topic 3)

You have an Azure Synapse Analytics job that uses Scala. You need to view the status of the job.

What should you do?

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

Answer: C

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 116

- (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	POH_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	48	2096	1	51
192	2632	528	32	2072	1	52
1127	2768	680	48	2040	1	53
1244	3032	704	64	2264	1	54
409	2632	568	32	2032	1	55
0	2232	496	0	1736	1	56
1417	2832	728	48	2064	1	57
0	2232	496	0	1736	1	58
384	2632	560	32	2040	1	59
225	2768	544	48	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 120

- (Exam Topic 3)

You have two Azure Storage accounts named Storage1 and Storage2. Each account holds one container and has the hierarchical namespace enabled. The system has files that contain data stored in the Apache Parquet format.

You need to copy folders and files from Storage1 to Storage2 by using a Data Factory copy activity. The solution must meet the following requirements:

- No transformations must be performed.
- The original folder structure must be retained.
- Minimize time required to perform the copy activity.

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

NOTE: Each correct selection is worth one point.

Source dataset type:

▼

Binary
Parquet
Delimited text

Copy activity copy behavior:

▼

FlattenHierarchy
MergeFiles
PreserveHierarchy

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Graphical user interface, text, application, chat or text message Description automatically generated

Box 1: Parquet

For Parquet datasets, the type property of the copy activity source must be set to ParquetSource. Box 2: PreserveHierarchy

PreserveHierarchy (default): Preserves the file hierarchy in the target folder. The relative path of the source file to the source folder is identical to the relative path of the target file to the target folder.

Reference:

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

NEW QUESTION 125

- (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 127

- (Exam Topic 3)

You need to design an Azure Synapse Analytics dedicated SQL pool that meets the following requirements: ➤ Can return an employee record from a given point in time.

- Maintains the latest employee information.
- Minimizes query complexity.

How should you model the employee data?

A. as a temporal table

B. as a SQL graph table

C. as a degenerate dimension table

D. as a Type 2 slowly changing dimension (SCD) table

Answer: D

Explanation:

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

Reference:

<https://docs.microsoft.com/en-us/learn/modules/populate-slowly-changing-dimensions-azure-synapse-analytics>

NEW QUESTION 132

- (Exam Topic 3)

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

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

You plan to create an Azure Databricks workspace that has a tiered structure. The workspace will contain the following three workloads:

- A workload for data engineers who will use Python and SQL.
- A workload for jobs that will run notebooks that use Python, Scala, and SQL.
- A workload that data scientists will use to perform ad hoc analysis in Scala and R.

The enterprise architecture team at your company identifies the following standards for Databricks environments:

- The data engineers must share a cluster.
- The job cluster will be managed by using a request process whereby data scientists and data engineers provide packaged notebooks for deployment to the cluster.
- All the data scientists must be assigned their own cluster that terminates automatically after 120 minutes of inactivity. Currently, there are three data scientists.

You need to create the Databricks clusters for the workloads.

Solution: You create a High Concurrency cluster for each data scientist, a High Concurrency cluster for the data engineers, and a Standard cluster for the jobs.

Does this meet the goal?

- A. Yes
- B. No

Answer: B

Explanation:

Need a High Concurrency cluster for the jobs.

Standard clusters are recommended for a single user. Standard can run workloads developed in any language: Python, R, Scala, and SQL.

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

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

NEW QUESTION 135

- (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 137

- (Exam Topic 3)

You configure version control for an Azure Data Factory instance as shown in the following exhibit.

Connections

Linked services

Integration runtimes

Source control

Git configuration

ARM template

Parameterization template

Author

Triggers

Global parameters

Security

Customer managed key

Managed private endpoints

Git repository

Git repository information associated with your data factory. [CI/CD best practices](#)

Setting Disconnect

Repository type	Azure DevOps Git
Azure DevOps Account	CONTOSO
Project name	Data
Repository name	dwh_batchetl
Collaboration branch	main
Publish branch	adf_publish
Root folder	/

Use the drop-down menus to select the answer choice that completes each statement based on the information presented in the graphic.
NOTE: Each correct selection is worth one point.

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 140

- (Exam Topic 3)
You have an Azure Data Factory pipeline named Pipeline1!. Pipelinel 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 Pipelinel 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 144

- (Exam Topic 3)

You plan to implement an Azure Data Lake Storage Gen2 container that will contain CSV files. The size of the files will vary based on the number of events that occur per hour.

File sizes range from 4.KB to 5 GB.

You need to ensure that the files stored in the container are optimized for batch processing. What should you do?

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

Answer: D

Explanation:

Avro supports batch and is very relevant for streaming.

Note: Avro is framework developed within Apache's Hadoop project. It is a row-based storage format which is widely used as a serialization process. AVRO stores its schema in JSON format making it easy to read and interpret by any program. The data itself is stored in binary format by doing it compact and efficient.

Reference:

<https://www.adaltas.com/en/2020/07/23/benchmark-study-of-different-file-format/>

NEW QUESTION 146

- (Exam Topic 3)

You are creating an Azure Data Factory data flow that will ingest data from a CSV file, cast columns to specified types of data, and insert the data into a table in an Azure Synapse Analytics dedicated SQL pool. The CSV file contains columns named username, comment and date.

The data flow already contains the following:

- A source transformation
- A Derived Column transformation to set the appropriate types of data
- A sink transformation to land the data in the pool

You need to ensure that the data flow meets the following requirements;

- All valid rows must be written to the destination table.
- Truncation errors in the comment column must be avoided proactively.
- Any rows containing comment values that will cause truncation errors upon insert must be written to a file in blob storage.

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

- A. Add a select transformation that selects only the rows which will cause truncation errors.
- B. Add a sink transformation that writes the rows to a file in blob storage.
- C. Add a filter transformation that filters out rows which will cause truncation errors.
- D. Add a Conditional Split transformation that separates the rows which will cause truncation errors.

Answer: BD

NEW QUESTION 147

- (Exam Topic 3)

You have an Azure subscription that contains an Azure Databricks workspace named databricks1 and an Azure Synapse Analytics workspace named synapse1.

The synapse1 workspace contains an Apache Spark pool named pool1.

You need to share an Apache Hive catalog of pool1 with databricks1.

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

NOTE: Each correct selection is worth one point.

From synapse1, create a linked service to:

<input type="checkbox"/> Azure Cosmos DB
<input checked="" type="checkbox"/> Azure Data Lake Storage Gen2
<input type="checkbox"/> Azure SQL Database

Configure pool1 to use the linked service as:

<input type="checkbox"/> An Azure Purview account
<input checked="" type="checkbox"/> A Hive metastore
<input checked="" type="checkbox"/> A managed Hive metastore service

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Box 1: Azure SQL Database

Use external Hive Metastore for Synapse Spark Pool

Azure Synapse Analytics allows Apache Spark pools in the same workspace to share a managed HMS (Hive Metastore) compatible metastore as their catalog.

Set up linked service to Hive Metastore

Follow below steps to set up a linked service to the external Hive Metastore in Synapse workspace.

- > Open Synapse Studio, go to Manage > Linked services at left, click New to create a new linked service.
- > Set up Hive Metastore linked service
- > Choose Azure SQL Database or Azure Database for MySQL based on your database type, click Continue.
- > Provide Name of the linked service. Record the name of the linked service, this info will be used to configure Spark shortly.
- > You can either select Azure SQL Database/Azure Database for MySQL for the external Hive Metastore from Azure subscription list, or enter the info manually.
- > Provide User name and Password to set up the connection.
- > Test connection to verify the username and password.
- > Click Create to create the linked service.

Box 2: A Hive Metastore

Reference: https://docs.microsoft.com/en-us/azure/synapse-analytics/spark/apache-spark-external-metastore

NEW QUESTION 152

- (Exam Topic 3)

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

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

Input Details

products

Test

Delete

Container

☐ Create new
 ☒ Use existing

refdata

Path pattern

product.csv

Date format

YYYY/MM/DD

Time format

HH

Event serialization format *

CSV

Delimiter

comma (,)

Encoding

UTF-8

Save

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

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

refdata

Container

Search (Ctrl + /)

Upload

Add Directory

Refresh

Rename

Delete

Overview

Access Control (IAM)

Settings

Access policy

Properties

Metadata

Authentication method: Access key (Switch to Azure AD User Account)

Location: refdata / 2020-03-20

Search blobs by prefix (case-sensitive)

Name

☐ [..]
 ☐ product.csv

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

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

NOTE: Each correct selection is worth one point.

Path pattern:
Date format:

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

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

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Graphical user interface, application, table Description automatically generated

Box 1: {date}/product.csv

In the 2nd exhibit we see: Location: refdata / 2020-03-20

Note: Path Pattern: This is a required property that is used to locate your blobs within the specified container. Within the path, you may choose to specify one or more instances of the following 2 variables:

{date}, {time}

Example 1: products/{date}/{time}/product-list.csv

Example 2: products/{date}/product-list.csv

Example 3: product-list.csv

Box 2: YYYY-MM-DD

Note: Date Format [optional]: If you have used {date} within the Path Pattern that you specified, then you can select the date format in which your blobs are organized from the drop-down of supported formats.

Example: YYYY/MM/DD, MM/DD/YYYY, etc. Reference:

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

NEW QUESTION 156

- (Exam Topic 3)

You have a table named SalesFact in an enterprise data warehouse in Azure Synapse Analytics. SalesFact contains sales data from the past 36 months and has the following characteristics:

- > Is partitioned by month
- > Contains one billion rows
- > Has clustered columnstore indexes

At the beginning of each month, you need to remove data from SalesFact that is older than 36 months as quickly as possible.

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

Actions

Answer Area

Switch the partition containing the stale data from SalesFact to SalesFact_Work.
Truncate the partition containing the stale data.
Drop the SalesFact_Work table.
Create an empty table named SalesFact_Work that has the same schema as SalesFact.
Execute a DELETE statement where the value in the Date column is more than 36 months ago.
Copy the data to a new table by using CREATE TABLE AS SELECT (CTAS).

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Step 1: Create an empty table named SalesFact_work that has the same schema as SalesFact. Step 2: Switch the partition containing the stale data from SalesFact to SalesFact_Work.

SQL Data Warehouse supports partition splitting, merging, and switching. To switch partitions between two tables, you must ensure that the partitions align on their respective boundaries and that the table definitions match.

Loading data into partitions with partition switching is a convenient way stage new data in a table that is not visible to users the switch in the new data.

Step 3: Drop the SalesFact_Work table. Reference:

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

NEW QUESTION 158

- (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 162

- (Exam Topic 3)

You have an Azure Stream Analytics job that receives clickstream data from an Azure event hub.

You need to define a query in the Stream Analytics job. The query must meet the following requirements: ➤ Count the number of clicks within each 10-second window based on the country of a visitor.

➤ Ensure that each click is NOT counted more than once. How should you define the Query?

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

Answer: B

Explanation:

Tumbling window functions are used to segment a data stream into distinct time segments and perform a function against them, such as the example below. The key differentiators of a Tumbling window are that they repeat, do not overlap, and an event cannot belong to more than one tumbling window.

Example: Reference:

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

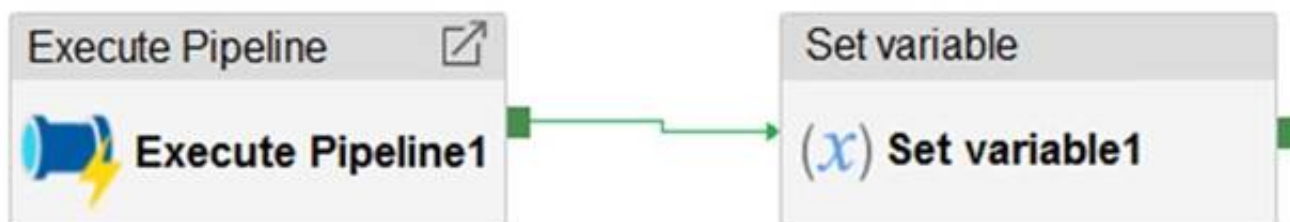
NEW QUESTION 166

- (Exam Topic 3)

You have an Azure Data Factory instance that contains two pipelines named Pipeline1 and Pipeline2. Pipeline1 has the activities shown in the following exhibit.



Pipeline2 has the activities shown in the following exhibit.



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

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

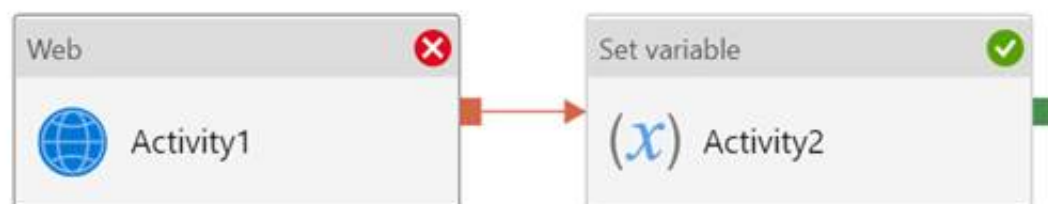
Answer: A

Explanation:

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

Consider Pipeline1:

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



The failure dependency means this pipeline reports success. Note:

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

Reference:

<https://datasavvy.me/category/azure-data-factory/>

NEW QUESTION 170

- (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 175

- (Exam Topic 3)

You plan to create a real-time monitoring app that alerts users when a device travels more than 200 meters away from a designated location.

You need to design an Azure Stream Analytics job to process the data for the planned app. The solution must minimize the amount of code developed and the number of technologies used.

What should you include in the Stream Analytics job? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Input type:

Function:

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Diagram, table Description automatically generated

Input type: Stream

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

With built-in geospatial functions, you can use Azure Stream Analytics to build applications for scenarios such as fleet management, ride sharing, connected cars, and asset tracking.

Note: In a real-world scenario, you could have hundreds of these sensors generating events as a stream. Ideally, a gateway device would run code to push these events to Azure Event Hubs or Azure IoT Hubs.

Reference:

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-get-started-with-azure-stream-analytic> <https://docs.microsoft.com/en-us/azure/stream->

NEW QUESTION 180

- (Exam Topic 3)

You have an Azure subscription that contains an Azure Synapse Analytics dedicated SQL pool. You plan to deploy a solution that will analyze sales data and include the following:

- A table named Country that will contain 195 rows
- A table named Sales that will contain 100 million rows
- A query to identify total sales by country and customer from the past 30 days

You need to create the tables. The solution must maximize query performance.

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

Answer Area

```
CREATE TABLE [dbo].[Sales]
(
    [OrderDate]          date           NOT NULL
,   [CustomerId] int NOT NULL
,   [CountryId] int NOT NULL
,   [Total] money NOT NULL
)

WITH
(
    DISTRIBUTION = HASH([CustomerId])
    CLUSTERED COLUMNSTORE INDEX
    ON HASH([CustomerId])
    REPLICATE
    ROUND_ROBIN
)
CREATE TABLE [dbo].[Customer]
(
```

- A. Mastered
B. Not Mastered

Answer: A

Explanation:

Answer Area

```
CREATE TABLE [dbo].[Sales]
(
    [OrderDate]          date           NOT NULL
,   [CustomerId] int NOT NULL
,   [CountryId] int NOT NULL
,   [Total] money NOT NULL
)

WITH
(
    DISTRIBUTION = HASH([CustomerId])
    CLUSTERED COLUMNSTORE INDEX ON HASH([CustomerId])
    REPLICATE
    ROUND_ROBIN
)
CREATE TABLE [dbo].[Customer]
```

NEW QUESTION 185

- (Exam Topic 3)

You have a partitioned table in an Azure Synapse Analytics dedicated SQL pool. You need to design queries to maximize the benefits of partition elimination. What should you include in the Transact-SQL queries?

- A. JOIN
B. WHERE
C. DISTINCT
D. GROUP BY

Answer: B

NEW QUESTION 190

- (Exam Topic 2)

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

- A. Deploy identical Azure Stream Analytics jobs to paired regions in Azure.
- B. Deploy a High Concurrency Databricks cluster.
- C. Deploy an Azure Stream Analytics job and use an Azure Automation runbook to check the status of the job and to start the job if it stops.
- D. Set Data Lake Storage to use geo-redundant storage (GRS).

Answer: A

Explanation:

Guarantee Stream Analytics job reliability during service updates

Part of being a fully managed service is the capability to introduce new service functionality and improvements at a rapid pace. As a result, Stream Analytics can have a service update deploy on a weekly (or more frequent) basis. No matter how much testing is done there is still a risk that an existing, running job may break due to the introduction of a bug. If you are running mission critical jobs, these risks need to be avoided. You can reduce this risk by following Azure's paired region model.

Scenario: The application development team will create an Azure event hub to receive real-time sales data, including store number, date, time, product ID, customer loyalty number, price, and discount amount, from the point of sale (POS) system and output the data to data storage in Azure

Reference:

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

NEW QUESTION 193

- (Exam Topic 1)

You need to integrate the on-premises data sources and Azure Synapse Analytics. The solution must meet the data integration requirements.

Which type of integration runtime should you use?

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

Answer: C

NEW QUESTION 198

- (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 203

.....

THANKS FOR TRYING THE DEMO OF OUR PRODUCT

Visit Our Site to Purchase the Full Set of Actual DP-203 Exam Questions With Answers.

We Also Provide Practice Exam Software That Simulates Real Exam Environment And Has Many Self-Assessment Features. Order the DP-203 Product From:

<https://www.2passeasy.com/dumps/DP-203/>

Money Back Guarantee

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