

Microsoft

Exam Questions DP-203

Data Engineering on Microsoft Azure



NEW QUESTION 1

- (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 2

- (Exam Topic 3)

You are designing a date dimension table in an Azure Synapse Analytics dedicated SQL pool. The date dimension table will be used by all the fact tables.

Which distribution type should you recommend to minimize data movement?

- A. HASH
- B. REPLICATE
- C. ROUND ROBIN

Answer: B

Explanation:

A replicated table has a full copy of the table available on every Compute node. Queries run fast on replicated tables since joins on replicated tables don't require data movement. Replication requires extra storage, though, and isn't practical for large tables.

Reference:

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

NEW QUESTION 3

- (Exam Topic 3)

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

You plan to create a database named D61 in Pool1.

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

Which format should you use for the tables in DB1?

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

Answer: A

Explanation:

Serverless SQL pool can automatically synchronize metadata from Apache Spark. A serverless SQL pool database will be created for each database existing in serverless Apache Spark pools.

For each Spark external table based on Parquet or CSV and located in Azure Storage, an external table is created in a serverless SQL pool database.

Reference:

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

NEW QUESTION 4

- (Exam Topic 3)

You have an Azure subscription that contains an Azure Data Lake Storage Gen2 account named storage1. Storage1 contains a container named container1.

Container1 contains a directory named directory1. Directory1 contains a file named file1.

You have an Azure Active Directory (Azure AD) user named User1 that is assigned the Storage Blob Data Reader role for storage1.

You need to ensure that User1 can append data to file1. The solution must use the principle of least privilege. Which permissions should you grant? To answer, drag the appropriate permissions to the correct resources.

Each permission may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.

Permissions

Read

Write

Execute

Answer Area

container1: Permission

directory1: Permission

file1: Permission

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Box 1: Execute

If you are granting permissions by using only ACLs (no Azure RBAC), then to grant a security principal read or write access to a file, you'll need to give the security principal Execute permissions to the root folder of the container, and to each folder in the hierarchy of folders that lead to the file.

Box 2: Execute

On Directory: Execute (X): Required to traverse the child items of a directory Box 3: Write

On file: Write (W): Can write or append to a file. Reference:

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

NEW QUESTION 5

- (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 6

- (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 7

- (Exam Topic 3)

You are creating an Apache Spark job in Azure Databricks that will ingest JSON-formatted data. You need to convert a nested JSON string into a DataFrame that will contain multiple rows. Which Spark SQL function should you use?

- A. explode
- B. filter
- C. coalesce
- D. extract

Answer: A

Explanation:

Convert nested JSON to a flattened DataFrame

You can to flatten nested JSON, using only \$"column.*" and explode methods. Note: Extract and flatten

Use \$"column.*" and explode methods to flatten the struct and array types before displaying the flattened DataFrame.

Scala

```
display(DF.select($"id" as "main_id", $"name", $"batters", $"ppu", explode($"topping")) // Exploding the topping column using explode as it is an array type
withColumn("topping_id", $"col.id") // Extracting topping_id from col using DOT form withColumn("topping_type", $"col.type") // Extracting topping_tytpe from col
using DOT form drop($"col")
select($"*", $"batters.*") // Flattened the struct type batters tto array type which is batter drop($"batters")
select($"*", explode($"batter")) drop($"batter")
withColumn("batter_id", $"col.id") // Extracting batter_id from col using DOT form withColumn("battter_type", $"col.type") // Extracting battter_type from col using
DOT form drop($"col")
)
```

Reference: <https://learn.microsoft.com/en-us/azure/databricks/kb/scala/flatten-nested-columns-dynamically>

NEW QUESTION 8

- (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 9

- (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 10

- (Exam Topic 3)

You plan to build a structured streaming solution in Azure Databricks. The solution will count new events in five-minute intervals and report only events that arrive during the interval. The output will be sent to a Delta Lake table.

Which output mode should you use?

- A. complete
- B. update
- C. append

Answer: C

Explanation:

Append Mode: Only new rows appended in the result table since the last trigger are written to external storage. This is applicable only for the queries where existing rows in the Result Table are not expected to change.

<https://docs.databricks.com/getting-started/spark/streaming.html>

NEW QUESTION 10

- (Exam Topic 3)

You have an Azure data solution that contains an enterprise data warehouse in Azure Synapse Analytics named DW1.

Several users execute ad hoc queries to DW1 concurrently. You regularly perform automated data loads to DW1.

You need to ensure that the automated data loads have enough memory available to complete quickly and successfully when the adhoc queries run. What should you do?

- A. Hash distribute the large fact tables in DW1 before performing the automated data loads.
- B. Assign a smaller resource class to the automated data load queries.
- C. Assign a larger resource class to the automated data load queries.
- D. Create sampled statistics for every column in each table of DW1.

Answer: C

Explanation:

The performance capacity of a query is determined by the user's resource class. Resource classes are pre-determined resource limits in Synapse SQL pool that govern compute resources and concurrency for query execution.

Resource classes can help you configure resources for your queries by setting limits on the number of queries that run concurrently and on the compute-resources assigned to each query. There's a trade-off between memory and concurrency.

Smaller resource classes reduce the maximum memory per query, but increase concurrency. Larger resource classes increase the maximum memory per query, but reduce concurrency. Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/resource-classes-for-workload-ma>

NEW QUESTION 15

- (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 19

- (Exam Topic 3)

You are designing an anomaly detection solution for streaming data from an Azure IoT hub. The solution must meet the following requirements:

- > Send the output to Azure Synapse.
- > Identify spikes and dips in time series data.
- > Minimize development and configuration effort. Which should you include in the solution?

- A. Azure Databricks
- B. Azure Stream Analytics
- C. Azure SQL Database

Answer: B

Explanation:

You can identify anomalies by routing data via IoT Hub to a built-in ML model in Azure Stream Analytics. Reference:
<https://docs.microsoft.com/en-us/learn/modules/data-anomaly-detection-using-azure-iot-hub/>

NEW QUESTION 24

- (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: Preserver herarchy

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

Box 2: Parquet

Azure Data Factory parquet format is supported for Azure Data Lake Storage Gen2. Parquet supports the schema property.

Reference:

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

NEW QUESTION 26

- (Exam Topic 3)

You have an Azure Stream Analytics query. The query returns a result set that contains 10,000 distinct values for a column named clusterID.

You monitor the Stream Analytics job and discover high latency. You need to reduce the latency.

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

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

Answer: CE

Explanation:

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

E: Streaming Units (SUs) represents the computing resources that are allocated to execute a Stream Analytics job. The higher the number of SUs, the more CPU and memory resources are allocated for your job. This capacity lets you focus on the query logic and abstracts the need to manage the hardware to run your Stream Analytics job in a timely manner.

References:

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

NEW QUESTION 31

- (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 34

- (Exam Topic 3)

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

- > One billion rows
- > A clustered columnstore index
- > A hash-distributed column named Product Key
- > A column named Sales Date that is of the date data type and cannot be null Thirty million rows will be added to Table1 each month.

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

How often should you create a partition?

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

Answer: B

Explanation:

Need a minimum 1 million rows per distribution. Each table is 60 distributions. 30 millions rows is added each month. Need 2 months to get a minimum of 1 million rows per distribution in a new partition.

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

Any partitioning added to a table is in addition to the distributions created behind the scenes. Using this example, if the sales fact table contained 36 monthly partitions, and given that a dedicated SQL pool has 60 distributions, then the sales fact table should contain 60 million rows per month, or 2.1 billion rows when all months are populated. If a table contains fewer than the recommended minimum number of rows per partition, consider using fewer partitions in order to increase the number of rows per partition.

Reference:

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

NEW QUESTION 37

- (Exam Topic 3)

You have an Azure data factory that connects to a Microsoft Purview account. The data factory is registered in Microsoft Purview.

You update a Data Factory pipeline.

You need to ensure that the updated lineage is available in Microsoft Purview.

What You have an Azure subscription that contains an Azure SQL database named DB1 and a storage account named storage1. The storage1 account contains a file named File1.txt. File1.txt contains the names of selected tables in DB1.

You need to use an Azure Synapse pipeline to copy data from the selected tables in DB1 to the files in storage1. The solution must meet the following requirements:

- The Copy activity in the pipeline must be parameterized to use the data in File1.txt to identify the source and destination of the copy.
- Copy activities must occur in parallel as often as possible.

Which two pipeline activities should you include in the pipeline? Each correct answer presents part of the

solution. NOTE: Each correct selection is worth one point.

- A. If Condition
- B. ForEach
- C. Lookup
- D. Get Metadata

Answer: CD

NEW QUESTION 39

- (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 40

- (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 44

- (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. \YYYY\MM\DD\Department\DataSource\DataFile_YYYYMMDD.parquet
- B. \Department\DataSource\YYYY\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 45

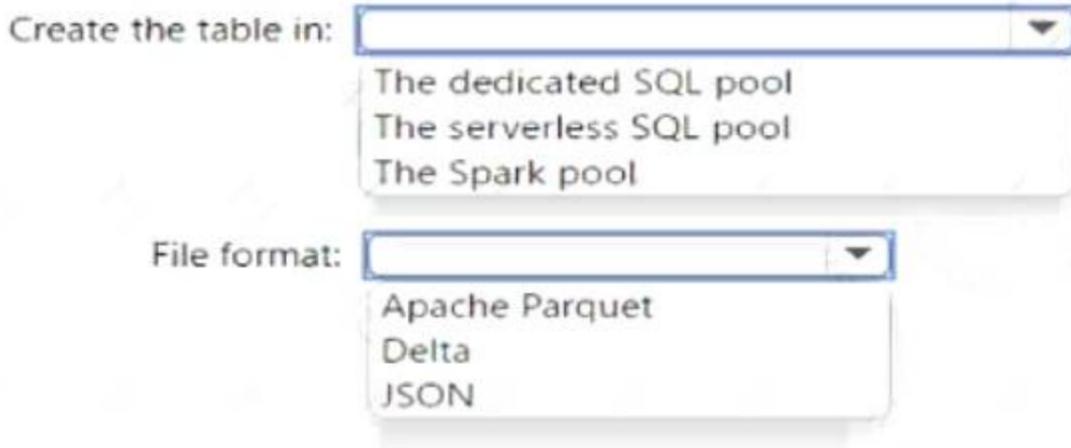
- (Exam Topic 3)

You have an Azure Synapse Analytics serverless SQL pool, an Azure Synapse Analytics dedicated SQL pool, an Apache Spark pool, and an Azure Data Lake Storage Gen2 account.

You need to create a table in a lake database. The table must be available to both the serverless SQL pool and the Spark pool.

Where should you create the table, and Which file format should you use for data in the table? 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:

The dedicated SQL pool Apache Parquet

NEW QUESTION 48

- (Exam Topic 3)

You have an Azure SQL database named DB1 and an Azure Data Factory data pipeline named pipeline. From Data Factory, you configure a linked service to DB1.

In DB1, you create a stored procedure named SP1. SP1 returns a single row of data that has four columns.

You need to add an activity to pipeline to execute SP1. The solution must ensure that the values in the columns are stored as pipeline variables.

Which two types of activities can you use to execute SP1? (Refer to Data Engineering on Microsoft Azure documents or guide for Answers explanation available at Microsoft.com)

- A. Stored Procedure
- B. Lookup
- C. Script

D. Copy

Answer: AB

Explanation:

the two types of activities that you can use to execute SP1 are Stored Procedure and Lookup.

A Stored Procedure activity executes a stored procedure on an Azure SQL Database or Azure Synapse Analytics or SQL Server1. You can specify the stored procedure name and parameters in the activity setting1s.

A Lookup activity retrieves a dataset from any data source that returns a single row of data with four columns2. You can use a query to execute a stored procedure as the source of the Lookup activit2y. You can then store the values in the columns as pipeline variables by using expressions2.

<https://learn.microsoft.com/en-us/azure/data-factory/transform-data-using-stored-procedure>

NEW QUESTION 53

- (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.

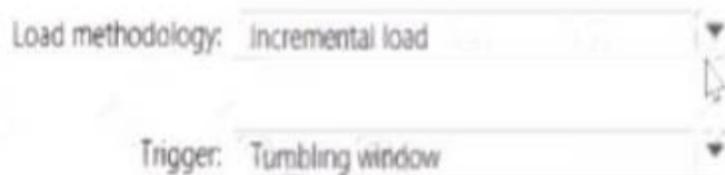
How should you configure the solution? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

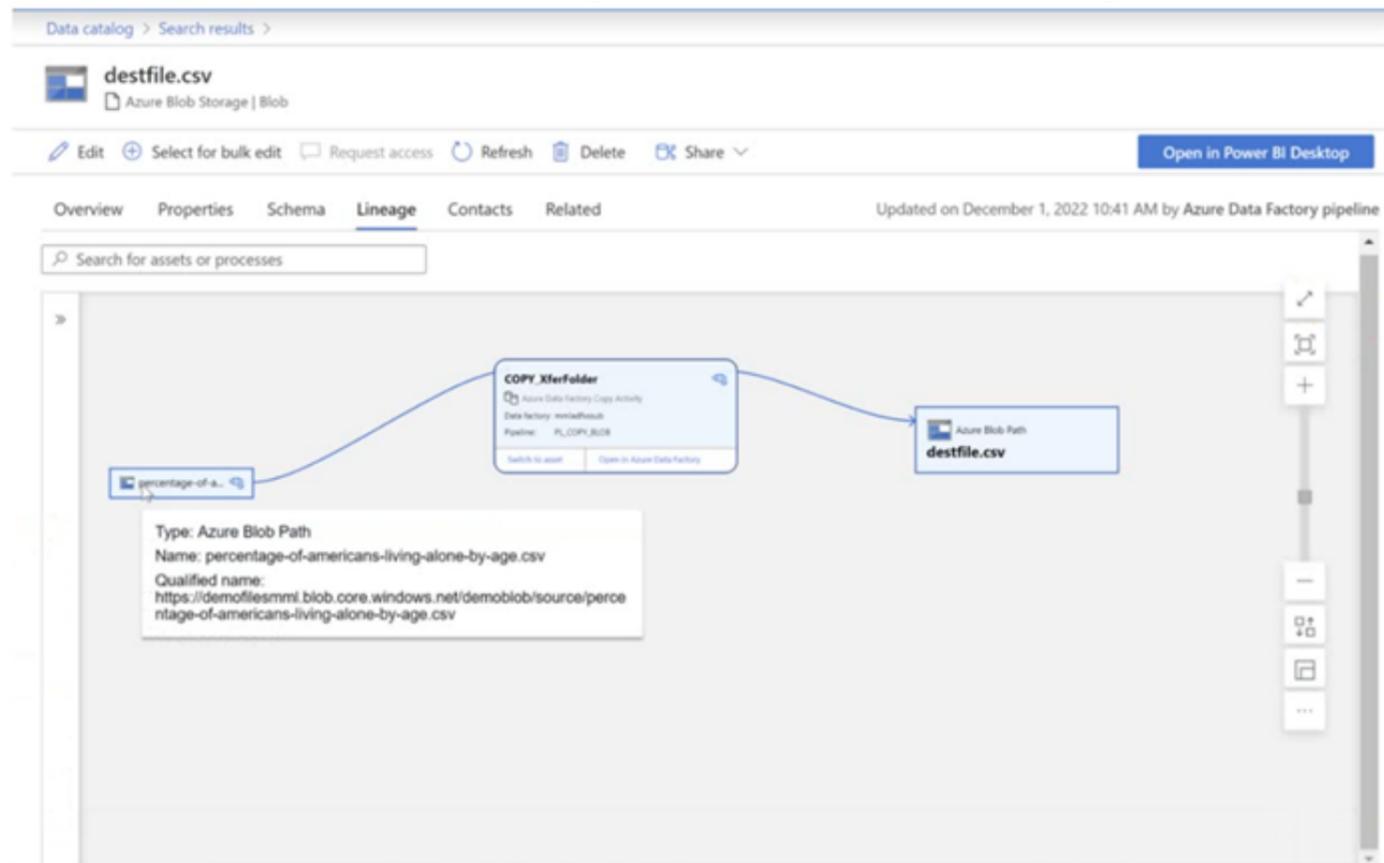
Answer Area



NEW QUESTION 54

- (Exam Topic 3)

You have a Microsoft Purview account. The Lineage view of a CSV file is shown in the following exhibit.



How is the data for the lineage populated?

- A. manually
- B. by scanning data stores
- C. by executing a Data Factory pipeline

Answer: B

Explanation:

According to Microsoft Purview Data Catalog lineage user guide1, data lineage in Microsoft Purview is a core platform capability that populates the Microsoft Purview Data Map with data movement and transformations across systems2. Lineage is captured as it flows in the enterprise and stitched without gaps irrespective of its source2.

NEW QUESTION 55

- (Exam Topic 3)

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

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

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

What do the query results include?

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

Answer: D

NEW QUESTION 60

- (Exam Topic 3)

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

What should you recommend?

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

Answer: C

Explanation:

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

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

Reference:

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

NEW QUESTION 61

- (Exam Topic 3)

You have an enterprise data warehouse in Azure Synapse Analytics named DW1 on a server named Server1. You need to determine the size of the transaction log file for each distribution of DW1.

What should you do?

- A. On DW1, execute a query against the sys.database_files dynamic management view.
- B. From Azure Monitor in the Azure portal, execute a query against the logs of DW1.
- C. Execute a query against the logs of DW1 by using theGet-AzOperationalInsightsSearchResult PowerShell cmdlet.
- D. On the master database, execute a query against the sys.dm_pdw_nodes_os_performance_counters dynamic management view.

Answer: A

Explanation:

For information about the current log file size, its maximum size, and the autogrow option for the file, you can also use the size, max_size, and growth columns for that log file in sys.database_files.

Reference:

<https://docs.microsoft.com/en-us/sql/relational-databases/logs/manage-the-size-of-the-transaction-log-file>

NEW QUESTION 64

- (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 66

- (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 69

- (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 73

- (Exam Topic 3)

You have an Azure subscription that contains an Azure Synapse Analytics workspace named workspace1. Workspace1 contains a dedicated SQL pool named SQL Pool and an Apache Spark pool named sparkpool. Sparkpool1 contains a DataFrame named pyspark.df.

You need to write the contents of pyspark_df to a table in SQLPoolM by using a PySpark notebook. 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



- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Answer Area



NEW QUESTION 78

- (Exam Topic 3)

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

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

You are designing an Azure Stream Analytics solution that will analyze Twitter data.

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

Solution: You use a hopping window that uses a hop size of 10 seconds and a window size of 10 seconds. Does this meet the goal?

- A. Yes
- B. No

Answer: B

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 79

- (Exam Topic 3)

You have an Azure Synapse Analytics dedicated SQL pool named Pool1. Pool1 contains a fact table named Table1. Table1 contains sales data. Sixty-five million rows of data are added to Table1 monthly.

At the end of each month, you need to remove data that is older than 36 months. The solution must minimize how long it takes to remove the data.

How should you partition Table1, and how should you remove the old data? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

Partition the data:
 Partition by date with one partition per day.
 Partition by date with one partition per month.
 Partition by product.

Remove the data:
 Delete the old data from Table1 by using a WHERE clause.
 Delete the old data from Table1 by using a JOIN.
 Switch the oldest partition to another table named Table2 and drop Table2.
 Truncate the oldest partition.

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:
Answer Area

Partition the data:
 Partition by date with one partition per day.
 Partition by date with one partition per month.
 Partition by product.

Remove the data:
 Delete the old data from Table1 by using a WHERE clause.
 Delete the old data from Table1 by using a JOIN.
 Switch the oldest partition to another table named Table2 and drop Table2.
 Truncate the oldest partition.

NEW QUESTION 80

- (Exam Topic 3)

You are monitoring an Azure Stream Analytics job.

You discover that the Backlogged Input Events metric is increasing slowly and is consistently non-zero. You need to ensure that the job can handle all the events. What should you do?

- A. Change the compatibility level of the Stream Analytics job.
- B. Increase the number of streaming units (SUs).
- C. Remove any named consumer groups from the connection and use \$default.
- D. Create an additional output stream for the existing input stream.

Answer: B

Explanation:

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

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

Reference:

<https://docs.microsoft.com/bs-cyrl-ba/azure/stream-analytics/stream-analytics-monitoring>

NEW QUESTION 83

- (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:

 Azure Cosmos DB
 Azure Data Lake Storage Gen2
 Azure SQL Database

Configure pool1 to use the linked service as:

 An Azure Purview account
 A Hive metastore
 A managed Hive metastore service

- 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.

- 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

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

NEW QUESTION 85

- (Exam Topic 3)

You have an Azure Synapse Analytics serverless SQ1 pool.

You have an Azure Data Lake Storage account named aols1 that contains a public container named container1 The container 1 container contains a folder named folder 1.

You need to query the top 100 rows of all the CSV files in folder 1.

How shouk1 you complete the query? To answer, drag the appropriate values to the correct targets. Each value may be used once, more than once, or not at all.

You may need to drag the split bar between panes or scroll to view content.

NOTE Each correct selection is worth one point.

A. Mastered
 B. Not Mastered

Answer: A

Explanation:

NEW QUESTION 90

- (Exam Topic 3)

You build an Azure Data Factory pipeline to move data from an Azure Data Lake Storage Gen2 container to a database in an Azure Synapse Analytics dedicated SQL pool.

Data in the container is stored in the following folder structure.

/in/{YYYY}/{MM}/{DD}/{HH}/{mm}

The earliest folder is /in/2021/01/01/00/00. The latest folder is /in/2021/01/15/01/45. You need to configure a pipeline trigger to meet the following requirements:

- Existing data must be loaded.
- Data must be loaded every 30 minutes.
- Late-arriving data of up to two minutes must he included in the load for the time at which the data should have arrived.

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

Type: ▼

Event
On-demand
Schedule
Tumbling window

Additional properties: ▼

Prefix: /in/, Event: Blob created
Recurrence: 30 minutes, Start time: 2021-01-01T00:00
Recurrence: 30 minutes, Start time: 2021-01-01T00:00, Delay: 2 minutes
Recurrence: 32 minutes, Start time: 2021-01-15T01:45

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Box 1: Tumbling window

To be able to use the Delay parameter we select Tumbling window. Box 2:

Recurrence: 30 minutes, not 32 minutes

Delay: 2 minutes.

The amount of time to delay the start of data processing for the window. The pipeline run is started after the expected execution time plus the amount of delay. The delay defines how long the trigger waits past the due time before triggering a new run. The delay doesn't alter the window startTime.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/how-to-create-tumbling-window-trigger>

NEW QUESTION 91

- (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 95

- (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 99

- (Exam Topic 1)

You need to ensure that the Twitter feed data can be analyzed in the dedicated SQL pool. The solution must meet the customer sentiment analytics requirements. Which three Transaction-SQL DDL commands should you run in sequence? To answer, move the appropriate commands from the list of commands to the answer area and arrange them in the correct order.

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

Commands

Answer Area

- CREATE EXTERNAL DATA SOURCE
- CREATE EXTERNAL FILE FORMAT
- CREATE EXTERNAL TABLE
- CREATE EXTERNAL TABLE AS SELECT
- CREATE DATABASE SCOPED CREDENTIAL

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Scenario: Allow Contoso users to use PolyBase in an Azure Synapse Analytics dedicated SQL pool to query the content of the data records that host the Twitter feeds. Data must be protected by using row-level security (RLS). The users must be authenticated by using their own Azure AD credentials.

Box 1: CREATE EXTERNAL DATA SOURCE

External data sources are used to connect to storage accounts. Box 2: CREATE EXTERNAL FILE FORMAT

CREATE EXTERNAL FILE FORMAT creates an external file format object that defines external data stored in Azure Blob Storage or Azure Data Lake Storage. Creating an external file format is a prerequisite for creating an external table.

Box 3: CREATE EXTERNAL TABLE AS SELECT

When used in conjunction with the CREATE TABLE AS SELECT statement, selecting from an external table imports data into a table within the SQL pool. In addition to the COPY statement, external tables are useful for loading data.

Reference:

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

NEW QUESTION 100

- (Exam Topic 1)

You need to implement the surrogate key for the retail store table. The solution must meet the sales transaction dataset requirements. What should you create?

- A. a table that has an IDENTITY property
- B. a system-versioned temporal table
- C. a user-defined SEQUENCE object
- D. a table that has a FOREIGN KEY constraint

Answer: A

Explanation:

Scenario: Implement a surrogate key to account for changes to the retail store addresses.

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

Reference:

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

NEW QUESTION 105

- (Exam Topic 1)

You need to implement versioned changes to the integration pipelines. The solution must meet the data integration requirements. In which order should you perform the actions? To answer, move all actions from the list of actions to the answer area and arrange them in the correct order.

Actions	Answer Area
Publish changes.	
Create a feature branch.	
Merge changes.	
Create a repository and a main branch.	
Create a pull request.	

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Graphical user interface, application Description automatically generated

Scenario: Identify a process to ensure that changes to the ingestion and transformation activities can be version-controlled and developed independently by multiple data engineers.

Step 1: Create a repository and a main branch

You need a Git repository in Azure Pipelines, TFS, or GitHub with your app. Step 2: Create a feature branch

Step 3: Create a pull request Step 4: Merge changes

Merge feature branches into the main branch using pull requests. Step 5: Publish changes

Reference:

<https://docs.microsoft.com/en-us/azure/devops/pipelines/repos/pipeline-options-for-git>

NEW QUESTION 110

- (Exam Topic 1)

You need to design a data retention solution for the Twitter feed data records. The solution must meet the customer sentiment analytics requirements. Which Azure Storage functionality should you include in the solution?

- A. time-based retention
- B. change feed
- C. soft delete
- D. lifecycle management

Answer: D

NEW QUESTION 113

- (Exam Topic 3)

You are developing a solution that will stream to Azure Stream Analytics. The solution will have both streaming data and reference data. Which input type should you use for the reference data?

- A. Azure Cosmos DB
- B. Azure Blob storage
- C. Azure IoT Hub
- D. Azure Event Hubs

Answer: B

Explanation:

Stream Analytics supports Azure Blob storage and Azure SQL Database as the storage layer for Reference Data.

Reference:

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

NEW QUESTION 116

- (Exam Topic 3)

You are designing a sales transactions table in an Azure Synapse Analytics dedicated SQL pool. The table will contains approximately 60 million rows per month and will be partitioned by month. The table will use a clustered column store index and round-robin distribution. Approximately how many rows will there be for each combination of distribution and partition?

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

Answer: D

Explanation:

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

NEW QUESTION 120

- (Exam Topic 3)

You have an Azure Synapse Analytics dedicated SQL pool named SA1 that contains a table named Table1. You need to identify tables that have a high percentage of deleted rows. What should you run?

- A) `sys.pdw_nodes_column_store_segments`
- B) `sys.dm_db_column_store_row_group_operational_stats`
- C) `sys.pdw_nodes_column_store_row_groups`
- D) `sys.dm_db_column_store_row_group_physical_stats`

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

Answer: B

NEW QUESTION 122

- (Exam Topic 3)

You need to create an Azure Data Factory pipeline to process data for the following three departments at your company: Ecommerce, retail, and wholesale. The solution must ensure that data can also be processed for the entire company.

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

NOTE: Each correct selection is worth one point.

Values

- all, ecommerce, retail, wholesale
- dept=='ecommerce', dept=='retail', dept=='wholesale'
- dept=='ecommerce', dept=='wholesale', dept=='retail'
- disjoint: false
- disjoint: true
- ecommerce, retail, wholesale, all

Answer Area

```
CleanData
split(
    [ ]
    [ ]
) ~> SplitByDept@([ ])
```

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

The conditional split transformation routes data rows to different streams based on matching conditions. The conditional split transformation is similar to a CASE decision structure in a programming language. The transformation evaluates expressions, and based on the results, directs the data row to the specified stream.

Box 1: dept=='ecommerce', dept=='retail', dept=='wholesale'

First we put the condition. The order must match the stream labeling we define in Box 3. Syntax:

```
<incomingStream> split(
<conditionalExpression1>
<conditionalExpression2> disjoint: {true | false}
) ~> <splitTx>@(stream1, stream2, ..., <defaultStream>)
```

Box 2: disjoint : false

disjoint is false because the data goes to the first matching condition. All remaining rows matching the third condition go to output stream all.

Box 3: ecommerce, retail, wholesale, all Label the streams

Reference:

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

NEW QUESTION 123

- (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 126

- (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 a dedicated 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 129

- (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 134

- (Exam Topic 3)

You have an Azure subscription that contains an Azure Databricks workspace. The workspace contains a notebook named Notebook1. In Notebook1, you create an Apache Spark DataFrame named df_sales that contains the following columns:

- Customer
- Salesperson
- Region
- Amount

You need to identify the three top performing salespersons by amount for a region named HQ.

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

Values	Answer Area
<code>agg(col('SalesPerson'))</code>	<code>df_sales.filter(col('Region')== 'HQ').</code> <input type="text"/>
<code>filter(col('SalesPerson'))</code>	<code>.agg(sum('Amount').alias('TotalAmount')).</code> <input type="text"/> <code>.limit(3)</code>
<code>groupBy(col('SalesPerson'))</code>	
<code>groupBy(col('TotalAmount'))</code>	
<code>orderBy(col('TotalAmount'))</code>	
<code>orderBy(desc('TotalAmount'))</code>	

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Values	Answer Area
<code>agg(col('SalesPerson'))</code>	<code>df_sales.filter(col('Region')== 'HQ').</code> <code>filter(col('SalesPerson'))</code>
<code>filter(col('SalesPerson'))</code>	<code>.agg(sum('Amount').alias('TotalAmount')).</code> <code>orderBy(desc('TotalAmount'))</code> <code>.limit(3)</code>
<code>groupBy(col('SalesPerson'))</code>	
<code>groupBy(col('TotalAmount'))</code>	
<code>orderBy(col('TotalAmount'))</code>	
<code>orderBy(desc('TotalAmount'))</code>	

NEW QUESTION 138

- (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 142

- (Exam Topic 3)

You are deploying a lake database by using an Azure Synapse database template.

You need to add additional tables to the database. The solution must use the same grouping method as the template tables.

Which grouping method should you use?

- A. business area
- B. size
- C. facts and dimensions
- D. partition style

Answer: A

Explanation:

➤ Business area: This is how the Azure Synapse database templates group tables by default. Each template consists of one or more enterprise templates that contain tables grouped by business areas. For example, the Retail template has business areas such as Customer, Product, Sales, and Store123. Using the same grouping method as the template tables can help you maintain consistency and compatibility with the industry-specific data model.
<https://techcommunity.microsoft.com/t5/azure-synapse-analytics-blog/database-templates-in-azure-synapse-anal>

NEW QUESTION 146

- (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 150

- (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: C

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 151

- (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 152

- (Exam Topic 3)

You are designing a streaming data solution that will ingest variable volumes of data. You need to ensure that you can change the partition count after creation. Which service should you use to ingest the data?

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

Answer: B

Explanation:

You can't change the partition count for an event hub after its creation except for the event hub in a dedicated cluster.

Reference:

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

NEW QUESTION 153

- (Exam Topic 3)

You have an Azure subscription.

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

Pool1 will contain the following tables.

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

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

- Maximize the performance of data loading operations to Staging.WebSessions.
- Minimize query times for reporting queries against the dimensional model.

Which type of table distribution should you use for each table? To answer, drag the appropriate table distribution types to the correct tables. Each table distribution type may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point.

Table distribution types

Hash

Replicated

Round-robin

Answer Area

Common.Data:

Marketing.Web.Sessions:

Staging. Web.Sessions:

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Box 1: Replicated

The best table storage option for a small table is to replicate it across all the Compute nodes. Box 2: Hash

Hash-distribution improves query performance on large fact tables. Box 3: Round-robin

Round-robin distribution is useful for improving loading speed.

Reference:

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

NEW QUESTION 158

- (Exam Topic 3)

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

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

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

Load methodology:

	▼
Full Load	
Incremental Load	
Load individual files as they arrive	

Trigger:

	▼
Fixed schedule	
New file	
Tumbling window	

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Table Description automatically generated

Box 1: Incremental load Box 2: Tumbling window

Tumbling windows are a series of fixed-sized, non-overlapping and contiguous time intervals. The following diagram illustrates a stream with a series of events and how they are mapped into 10-second tumbling windows.

Timeline Description automatically generated

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 160

- (Exam Topic 3)

You have an Azure Synapse Analytics dedicated SQL pool named Pool1. Pool1 contains a table named table1. You load 5 TB of data into table1. You need to ensure that columnstore compression is maximized for table1. Which statement should you execute?

- A. ALTER INDEX ALL on table1 REORGANIZE
- B. ALTER INDEX ALL on table1 REBUILD
- C. DBCC DBREINDEX (table1)
- D. DBCC INDEXDEFRAG (pool1,table1)

Answer: B

Explanation:

Columnstore and columnstore archive compression

Columnstore tables and indexes are always stored with columnstore compression. You can further reduce the size of columnstore data by configuring an additional compression called archival compression. To perform archival compression, SQL Server runs the Microsoft XPRESS compression algorithm on the data. Add or remove archival compression by using the following data compression types:

Use COLUMNSTORE_ARCHIVE data compression to compress columnstore data with archival compression.

Use COLUMNSTORE data compression to decompress archival compression. The resulting data continue to be compressed with columnstore compression.

To add archival compression, use ALTER TABLE (Transact-SQL) or ALTER INDEX (Transact-SQL) with the REBUILD option and DATA COMPRESSION = COLUMNSTORE_ARCHIVE.

Reference: <https://learn.microsoft.com/en-us/sql/relational-databases/data-compression/data-compression>

NEW QUESTION 163

- (Exam Topic 3)

You are batch loading a table in an Azure Synapse Analytics dedicated SQL pool.

You need to load data from a staging table to the target table. The solution must ensure that if an error occurs while loading the data to the target table, all the inserts in that batch are undone.

How should you complete the Transact-SQL 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

- BEGIN DISTRIBUTED TRANSACTION
- BEGIN TRAN
- COMMIT TRAN
- ROLLBACK TRAN
- SET RESULT_SET_CACHING ON

Answer Area

```

[ ]
BEGIN TRY
    INSERT INTO dbo.Table1 (col1, col2, col3)
    SELECT col1, col2, col3 FROM stage.Table1;
END TRY
BEGIN CATCH
    IF @@TRANCOUNT > 0
    BEGIN
        [ ] ;
    END
END CATCH;
IF @@TRANCOUNT > 0
BEGIN
    COMMIT TRAN;
END
    
```

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Values

- BEGIN DISTRIBUTED TRANSACTION
- BEGIN TRAN
- COMMIT TRAN
- ROLLBACK TRAN
- SET RESULT_SET_CACHING ON

Answer Area

```

[BEGIN TRAN]
BEGIN TRY
    INSERT INTO dbo.Table1 (col1, col2, col3)
    SELECT col1, col2, col3 FROM stage.Table1;
END TRY
BEGIN CATCH
    IF @@TRANCOUNT > 0
    BEGIN
        [ROLLBACK TRAN] ;
    END
END CATCH;
IF @@TRANCOUNT > 0
BEGIN
    COMMIT TRAN;
END
    
```

NEW QUESTION 168

- (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 172

- (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 hopping window that uses a hop size of 5 seconds and a window size 10 seconds. Does this meet the goal?

- A. Yes
- B. No

Answer: B

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 173

- (Exam Topic 3)

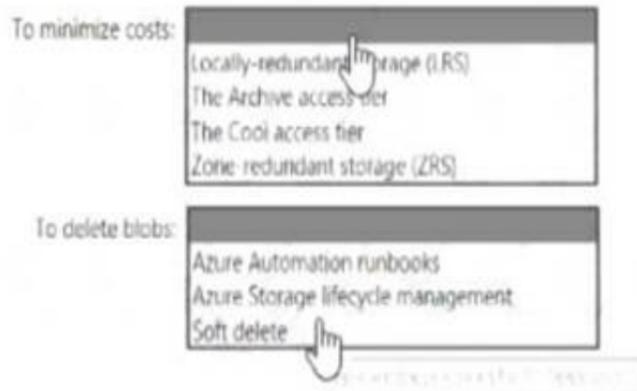
You have an Azure subscription.

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 175

- (Exam Topic 3)

You are building an Azure Analytics query that will receive input data from Azure IoT Hub and write the results to Azure Blob storage.

You need to calculate the difference in readings per sensor per hour.

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

NOTE: Each correct selection is worth one point.

```
SELECT sensorId,
       growth = reading -
           (reading) OVER (PARTITION BY sensorId
                           [ ] (hour, 1))
FROM input
```

▼
LAG
LAST
LEAD

▼
LIMIT DURATION
OFFSET
WHEN

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Box 1: LAG

The LAG analytic operator allows one to look up a “previous” event in an event stream, within certain constraints. It is very useful for computing the rate of growth of a variable, detecting when a variable crosses a threshold, or when a condition starts or stops being true.

Box 2: LIMIT DURATION

Example: Compute the rate of growth, per sensor: SELECT sensorId, growth = reading

LAG(reading) OVER (PARTITION BY sensorId LIMIT DURATION(hour, 1)) FROM input

Reference:

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

NEW QUESTION 176

- (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 Pool! 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 178

- (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	Answer Area	
Enable TDE on Pool1.		
Assign a managed identity to Server1.		
Configure key1 as the TDE protector for Server1.	⏪	⏩
Add key1 to the Azure key vault.	⏩	⏪
Create an Azure key vault and grant the managed identity permissions to the key vault.		

- 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 183

- (Exam Topic 3)

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

The data flow already contains the following:



A source transformation.

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

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

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

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

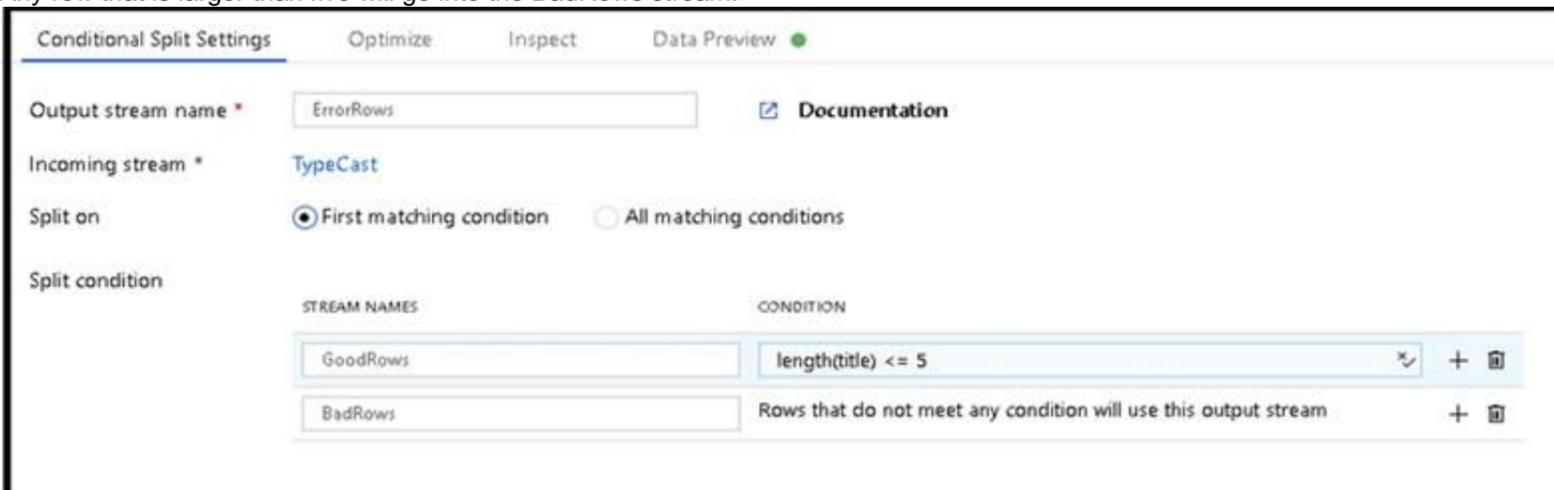
- A. To the data flow, add a sink transformation to write the rows to a file in blob storage.
- B. To the data flow, add a Conditional Split transformation to separate the rows that will cause truncation errors.
- C. To the data flow, add a filter transformation to filter out rows that will cause truncation errors.
- D. Add a select transformation to select only the rows that will cause truncation errors.

Answer: AB

Explanation:

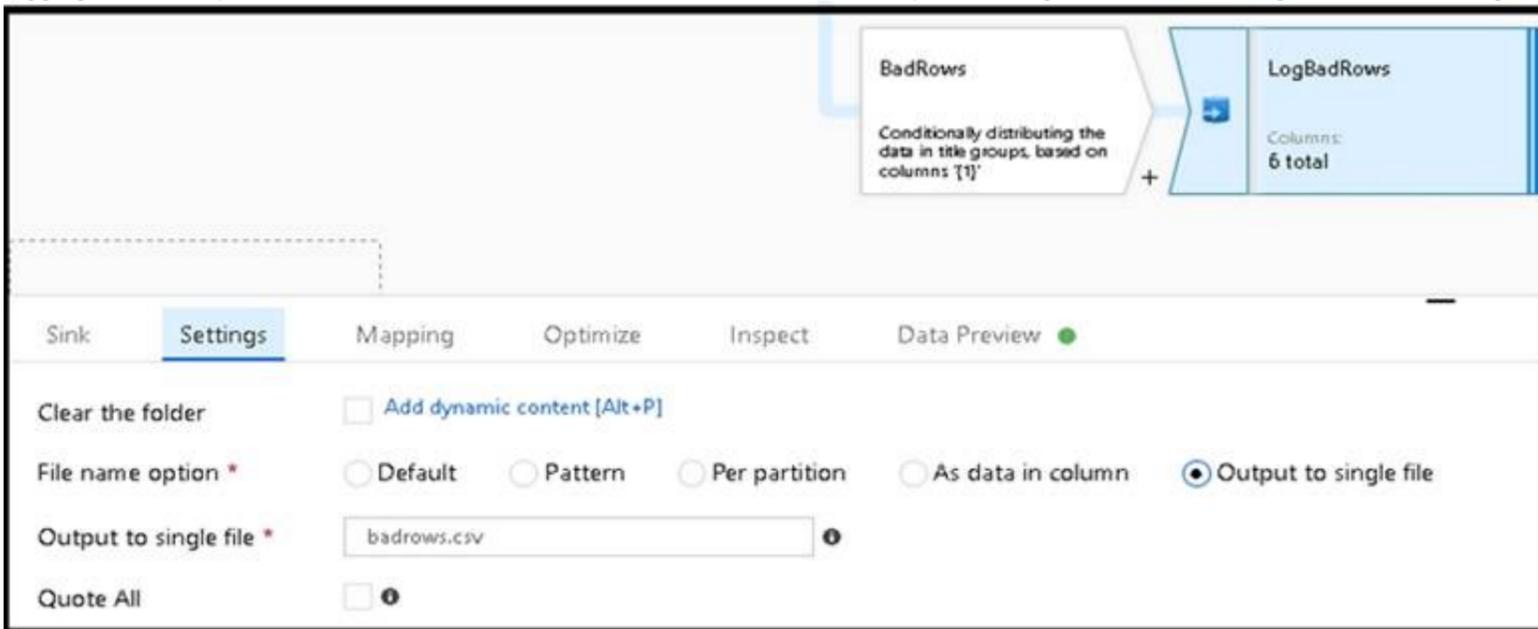
B: Example:

* 1. This conditional split transformation defines the maximum length of "title" to be five. Any row that is less than or equal to five will go into the GoodRows stream. Any row that is larger than five will go into the BadRows stream.

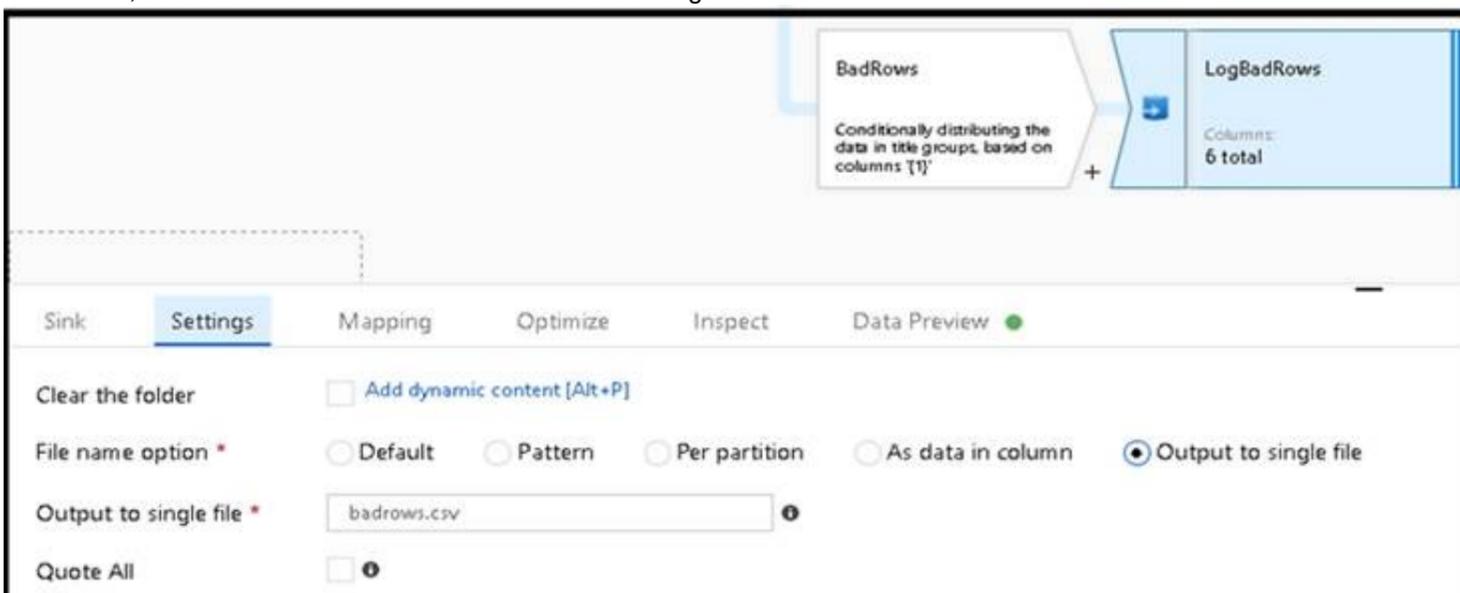


* 2. This conditional split transformation defines the maximum length of "title" to be five. Any row that is less than or equal to five will go into the GoodRows stream. Any row that is larger than five will go into the BadRows stream. A:

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



* 4. The completed data flow is shown below. We are now able to split off error rows to avoid the SQL truncation errors and put those entries into a log file. Meanwhile, successful rows can continue to write to our target database.



Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/how-to-data-flow-error-rows>

NEW QUESTION 188

- (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 192

- (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 194

- (Exam Topic 3)

You have an Azure subscription that contains an Azure Data Lake Storage account. The storage account contains a data lake named DataLake1.

You plan to use an Azure data factory to ingest data from a folder in DataLake1, transform the data, and land the data in another folder.

You need to ensure that the data factory can read and write data from any folder in the DataLake1 file system. The solution must meet the following requirements:

- > Minimize the risk of unauthorized user access.
- >

Use the principle of least privilege.

> Minimize maintenance effort.

How should you configure access to the storage account for the data factory? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Use to authenticate by using

Azure Active Directory (Azure AD)	a managed identity
a shared access signature (SAS)	a stored access policy
a shared key	an Authorization header

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Text Description automatically generated with low confidence

Box 1: Azure Active Directory (Azure AD)

On Azure, managed identities eliminate the need for developers having to manage credentials by providing an identity for the Azure resource in Azure AD and using it to obtain Azure Active Directory (Azure AD) tokens.

Box 2: a managed identity

A data factory can be associated with a managed identity for Azure resources, which represents this specific data factory. You can directly use this managed identity for Data Lake Storage Gen2 authentication, similar to using your own service principal. It allows this designated factory to access and copy data to or from your Data Lake Storage Gen2.

Note: The Azure Data Lake Storage Gen2 connector supports the following authentication types.

- > Account key authentication
- > Service principal authentication
- > Managed identities for Azure resources authentication Reference:

<https://docs.microsoft.com/en-us/azure/active-directory/managed-identities-azure-resources/overview> <https://docs.microsoft.com/en-us/azure/data-factory/connector-azure-data-lake-storage>

NEW QUESTION 196

- (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:

Stream
Reference

Function:

Aggregate
Geospatial
Windowing

- 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-analytics/geospatial-scenarios>

NEW QUESTION 198

- (Exam Topic 3)

You have an Azure Data Lake Storage Gen2 account named account1 that stores logs as shown in the following table.

Type	Designated retention period
Application	360 days
Infrastructure	60 days

You do not expect that the logs will be accessed during the retention periods.

You need to recommend a solution for account1 that meets the following requirements:

- > Automatically deletes the logs at the end of each retention period
- > Minimizes storage costs

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

NOTE: Each correct selection is worth one point.

To minimize storage costs:

▼

Store the infrastructure logs and the application logs in the Archive access tier

Store the infrastructure logs and the application logs in the Cool access tier

Store the infrastructure logs in the Cool access tier and the application logs in the Archive access tier

To delete logs automatically:

▼

Azure Data Factory pipelines

Azure Blob storage lifecycle management rules

Immutable Azure Blob storage time-based retention policies

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Table Description automatically generated

Box 1: Store the infrastructure logs in the Cool access tier and the application logs in the Archive access tier

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

For application logs: Archive tier - An offline tier optimized for storing data that is rarely accessed, and that has flexible latency requirements, on the order of hours. Data in the archive tier should be stored for a minimum of 180 days.

Box 2: Azure Blob storage lifecycle management rules

Blob storage lifecycle management offers a rule-based policy that you can use to transition your data to the desired access tier when your specified conditions are met. You can also use lifecycle management to expire data at the end of its life.

Reference:

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

NEW QUESTION 202

- (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 206

- (Exam Topic 3)

You plan to create a dimension table in Azure Synapse Analytics that will be less than 1 GB. You need to create the table to meet the following requirements:

- Provide the fastest Query time.
- Minimize data movement during queries. Which type of table should you use?

- A. hash distributed

- B. heap
- C. replicated
- D. round-robin

Answer: C

Explanation:

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

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/design-guidance-for-replicated-tab>

NEW QUESTION 210

- (Exam Topic 3)

You have an Apache Spark DataFrame named temperatures. A sample of the data is shown in the following table.

Date	Temp
...	...
18-01-2021	3
19-01-2021	4
20-01-2021	2
21-01-2021	2
...	...

You need to produce the following table by using a Spark SQL query.

Year	JAN	FEB	MAR	APR	MAY
2019	2.3	4.1	5.2	7.6	9.2
2020	2.4	4.2	4.9	7.8	9.1
2021	2.6	5.3	3.4	7.9	9.5

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

NOTE: Each correct selection is worth one point.

Values

Answer Area

- CAST
- COLLATE
- CONVERT
- FLATTEN
- PIVOT
- UNPIVOT

```

SELECT * FROM (
  SELECT YEAR(Date) Year, MONTH(Date) Month, Temp
  FROM temperatures
  WHERE date BETWEEN DATE '2019-01-01' AND DATE '2021-08-31'
)
AVG ( (Temp AS DECIMAL(4, 1)))
FOR Month in (
  1 JAN, 2 FEB, 3 MAR, 4 APR, 5 MAY, 6 JUN,
  7 JUL, 8 AUG, 9 SEP, 10 OCT, 11 NOV, 12 DEC
)
ORDER BY Year ASC
  
```

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Text Description automatically generated

Box 1: PIVOT

PIVOT rotates a table-valued expression by turning the unique values from one column in the expression into multiple columns in the output. And PIVOT runs aggregations where they're required on any remaining column values that are wanted in the final output.

Reference:

<https://learnsql.com/cookbook/how-to-convert-an-integer-to-a-decimal-in-sql-server/> <https://docs.microsoft.com/en-us/sql/t-sql/queries/from-using-pivot-and-unpivot>

NEW QUESTION 215

- (Exam Topic 3)

You are implementing Azure Stream Analytics windowing functions.

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

NOTE: Each correct selection is worth one point.

Answer Area

Segment the data stream into distinct time segments that repeat but do not overlap: Hopping Sliding Tumbling

Segment the data stream into distinct time segments that repeat and can overlap: Hopping Sliding Tumbling

Segment the data stream to produce an output only when an event occurs: Hopping Sliding Tumbling

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:
Answer Area

Segment the data stream into distinct time segments that repeat but do not overlap: Hopping Sliding Tumbling

Segment the data stream into distinct time segments that repeat and can overlap: Hopping Sliding Tumbling

Segment the data stream to produce an output only when an event occurs: Hopping Sliding Tumbling

NEW QUESTION 218

- (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 Azure Data Factory Studio, 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 222

- (Exam Topic 3)

You have an Azure subscription that contains an Azure Data Lake Storage account named myaccount1. The myaccount1 account contains two containers named container1 and contained. The subscription is linked to an Azure Active Directory (Azure AD) tenant that contains a security group named Group1.

You need to grant Group1 read access to container1. The solution must use the principle of least privilege. Which role should you assign to Group1?

- A. Storage Blob Data Reader for container1
- B. Storage Table Data Reader for container1
- C. Storage Blob Data Reader for myaccount1
- D. Storage Table Data Reader for myaccount1

Answer: A

NEW QUESTION 227

- (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	PDW_NODE_ID	DISTRIBUTION_ID
694	2776	616	48	2112	1	1
407	2704	576	48	2080	1	2
53	2376	512	16	1848	1	3
58	2376	512	16	1848	1	4
168	2632	528	32	2072	1	5
195	2696	536	32	2128	1	6
5995	3464	1424	32	2008	1	7
0	2232	496	0	1736	1	8
264	2576	544	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: C

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 232

- (Exam Topic 3)

You are designing a monitoring solution for a fleet of 500 vehicles. Each vehicle has a GPS tracking device that sends data to an Azure event hub once per minute.

You have a CSV file in an Azure Data Lake Storage Gen2 container. The file maintains the expected geographical area in which each vehicle should be.

You need to ensure that when a GPS position is outside the expected area, a message is added to another event hub for processing within 30 seconds. The solution must minimize cost.

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

NOTE: Each correct selection is worth one point.

Service: ▼

- An Azure Synapse Analytics Apache Spark pool
- An Azure Synapse Analytics serverless SQL pool
- Azure Data Factory
- Azure Stream Analytics

Window: ▼

- Hopping
- No window
- Session
- Tumbling

Analysis type: ▼

- Event pattern matching
- Lagged record comparison
- Point within polygon
- Polygon overlap

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Box 1: Azure Stream Analytics Box 2: Hopping

Hopping window functions hop forward in time by a fixed period. It may be easy to think of them as Tumbling windows that can overlap and be emitted more often than the window size. Events can belong to more than one Hopping window result set. To make a Hopping window the same as a Tumbling window, specify the hop size to be the same as the window size.

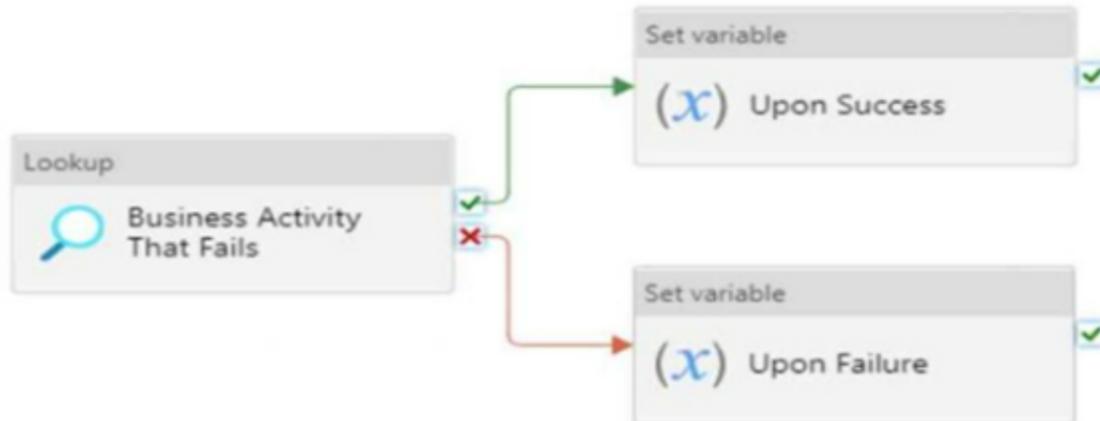
Box 3: Point within polygon Reference:

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

NEW QUESTION 233

- (Exam Topic 3)

You have the Azure Synapse Analytics pipeline shown in the following exhibit.



You need to add a set variable activity to the pipeline to ensure that after the pipeline's completion, the status of the pipeline is always successful. What should you configure for the set variable activity?

- A. a success dependency on the Business Activity That Fails activity
- B. a failure dependency on the Upon Failure activity
- C. a skipped dependency on the Upon Success activity
- D. a skipped dependency on the Upon Failure activity

Answer: A

Explanation:

A failure dependency means that the activity will run only if the previous activity fails. In this case, setting a failure dependency on the Upon Failure activity will ensure that the set variable activity will run after the pipeline fails and set the status of the pipeline to successful.

NEW QUESTION 234

- (Exam Topic 3)

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

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

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

You plan to copy the data from the storage account to an Azure SQL data warehouse. You need to prepare the files to ensure that the data copies quickly.

Solution: You modify the files to ensure that each row is more than 1 MB. Does this meet the goal?

- A. Yes
- B. No

Answer: A

Explanation:

Instead modify the files to ensure that each row is less than 1 MB. References:

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

NEW QUESTION 237

- (Exam Topic 3)

You have an Azure subscription that contains an Azure Synapse Analytics workspace named ws1 and an Azure Cosmos D6 database account named Cosmos1. Cosmos1 contains a container named container1 and ws1 contains a serverless1 SQL pool.

You need to ensure that you can Query the data in container1 by using the serverless1 SQL pool. Which three actions should you perform? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

- A. Enable Azure Synapse Link for Cosmos1
- B. Disable the analytical store for container1.
- C. In ws1, create a linked service that references Cosmos1
- D. Enable the analytical store for container1
- E. Disable indexing for container1

Answer: ACD

NEW QUESTION 242

- (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 247

- (Exam Topic 3)

You have an Azure subscription that contains the following resources:

- > An Azure Active Directory (Azure AD) tenant that contains a security group named Group1
- > An Azure Synapse Analytics SQL pool named Pool1

You need to control the access of Group1 to specific columns and rows in a table in Pool1.

Which Transact-SQL commands should you use? To answer, select the appropriate options in the answer area.

To control access to the columns:

	▼
CREATE CRYPTOGRAPHIC PROVIDER	
CREATE PARTITION FUNCTION	
CREATE SECURITY POLICY	
GRANT	

To control access to the rows:

	▼
CREATE CRYPTOGRAPHIC PROVIDER	
CREATE PARTITION FUNCTION	
CREATE SECURITY POLICY	
GRANT	

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Text Description automatically generated

Box 1: GRANT

You can implement column-level security with the GRANT T-SQL statement. Box 2: CREATE SECURITY POLICY

Implement Row Level Security by using the CREATE SECURITY POLICY Transact-SQL statement Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/column-level-security>

NEW QUESTION 251

- (Exam Topic 3)

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

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

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

Actions

Answer Area

- Create an external file format object
- Create an external data source
- Create a query that uses Create Table as Select
- Create a table
- Create an external table



- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

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

Step 1: Create an external data source

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

- > CREATE EXTERNAL DATA SOURCE to reference an external Azure storage and specify the credential that should be used to access the storage.
- > CREATE EXTERNAL FILE FORMAT to describe format of CSV or Parquet files.
- > CREATE EXTERNAL TABLE on top of the files placed on the data source with the same file format.

Step 2: Create an external file format object
 Creating an external file format is a prerequisite for creating an external table. Step 3: Create an external table

Reference:

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

NEW QUESTION 252

- (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 253

- (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 254

- (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
        COUNT
        MAX
        MIN
        TOPONE
    FROM
        Input TIMESTAMP BY Time
    GROUP BY
         (minute, 10)
        HoppingWindow
        SessionWindow
        SlidingWindow
        TumblingWindow
)
SELECT
    Input.License_plate,
    Input.Make,
    Input.Time
FROM
    Input TIMESTAMP BY Time
    INNER JOIN LastInWindow
    ON  (minute, Input, LastInWindow) BETWEEN 0 AND 10
        DATEADD
        DATEDIFF
        DATENAME
        DATEPART
    AND Input.Time = LastInWindow.LastEventTime
    
```

- 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 257

- (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

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

Search blobs by prefix (case-sensitive)

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

You need to configure the Stream Analytics job to pick up the new reference data. What should you configure? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

Path pattern:

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

Date format:

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

- 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 260

- (Exam Topic 3)

You have an Azure Synapse Analytics dedicated SQL pool.

You need to ensure that data in the pool is encrypted at rest. The solution must NOT require modifying applications that query the data.

What should you do?

- A. Enable encryption at rest for the Azure Data Lake Storage Gen2 account.
- B. Enable Transparent Data Encryption (TDE) for the pool.
- C. Use a customer-managed key to enable double encryption for the Azure Synapse workspace.
- D. Create an Azure key vault in the Azure subscription grant access to the pool.

Answer: B

Explanation:

Transparent Data Encryption (TDE) helps protect against the threat of malicious activity by encrypting and decrypting your data at rest. When you encrypt your database, associated backups and transaction log files are encrypted without requiring any changes to your applications. TDE encrypts the storage of an entire database by using a symmetric key called the database encryption key.

Reference:

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

NEW QUESTION 265

- (Exam Topic 3)

You have an Azure subscription that contains the resources shown in the following table.

Name	Type	Description
ws1	Azure Synapse Analytics workspace	None
kv1	Azure Key Vault	None
UAMI1	User-assigned managed identity	Associated with ws1
sp1	Apache Spark pool in Azure Synapse Analytics	Associated with ws1

You need to ensure that you can Spark notebooks in ws1. The solution must ensure secrets from kv1 by using UAMI1. What should you do? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

In the Azure portal: Add a role-based access control (RBAC) role to kv1. Add a role-based access control (RBAC) role to kv1. Add a role-based access control (RBAC) role to ws1. Create a linked service to kv1.

In Synapse Studio: Create a linked service to kv1. Add a role-based access control (RBAC) role to kv1. Add a role-based access control (RBAC) role to ws1. Create a linked service to kv1.

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Answer Area

In the Azure portal: Add a role-based access control (RBAC) role to kv1. Add a role-based access control (RBAC) role to kv1. Add a role-based access control (RBAC) role to ws1. Create a linked service to kv1.

In Synapse Studio: Create a linked service to kv1. Add a role-based access control (RBAC) role to kv1. Add a role-based access control (RBAC) role to ws1. Create a linked service to kv1.

NEW QUESTION 266

- (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 268

- (Exam Topic 3)

You are creating dimensions for a data warehouse in an Azure Synapse Analytics dedicated SQL pool. You create a table by using the Transact-SQL statement shown in the following exhibit.

```
CREATE TABLE [DBO].[DimProduct] (
    [ProductKey] [int] IDENTITY(1,1) NOT NULL,
    [ProductSourceID] [int] NOT NULL,
    [ProductName] [nvarchar](100) NOT NULL,
    [ProductNumber] [nvarchar](25) NOT NULL,
    [Color] [nvarchar](15) NULL,
    [Size] [nvarchar](5) NULL,
    [Weight] [decimal](8, 2) NULL,
    [ProductCategory] [nvarchar](100) NULL,
    [SellStartDate] [date] NOT NULL,
    [SellEndDate] [date] NULL,
    [RowInsertedDateTime] [datetime] NOT NULL,
    [RowUpdatedDateTime] [datetime] NOT NULL,
    [ETLAuditID] [int] NOT NULL
)
```

Use the drop-down menus to select the answer choice that completes each statement based on the information presented in the graphic.

NOTE: Each correct selection is worth one point.

DimProduct is a **[answer choice]** slowly changing dimension (SCD).

▼

Type 0

Type 1

Type 2

The ProductKey column is **[answer choice]**.

▼

a surrogate key

a business key

an audit column

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Box 1: Type 2

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

Reference:

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

NEW QUESTION 273

- (Exam Topic 3)

You have an Azure Stream Analytics job that is a Stream Analytics project solution in Microsoft Visual Studio. The job accepts data generated by IoT devices in the JSON format.

You need to modify the job to accept data generated by the IoT devices in the Protobuf format.

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

Actions	Answer Area
Change the Event Serialization Format to Protobuf in the input.json file of the job and reference the DLL.	
Add an Azure Stream Analytics Custom Deserializer Project (.NET) project to the solution.	
Add .NET deserializer code for Protobuf to the custom deserializer project.	
Add .NET deserializer code for Protobuf to the Stream Analytics project.	
Add an Azure Stream Analytics Application project to the solution.	

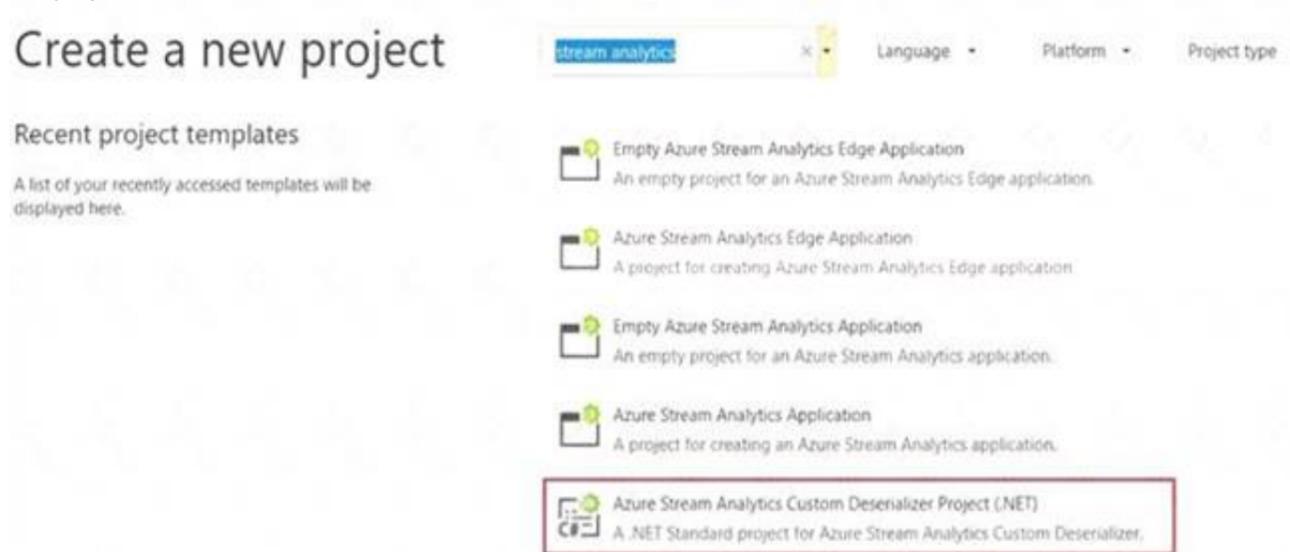
- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Step 1: Add an Azure Stream Analytics Custom Deserializer Project (.NET) project to the solution. Create a custom deserializer

* 1. Open Visual Studio and select File > New > Project. Search for Stream Analytics and select Azure Stream Analytics Custom Deserializer Project (.NET). Give the project a name, like Protobuf Deserializer.



* 2. In Solution Explorer, right-click your Protobuf Deserializer project and select Manage NuGet Packages from the menu. Then install the Microsoft.Azure.StreamAnalytics and Google.Protobuf NuGet packages.

* 3. Add the MessageBodyProto class and the MessageBodyDeserializer class to your project.

* 4. Build the Protobuf Deserializer project.

Step 2: Add .NET deserializer code for Protobuf to the custom deserializer project

Azure Stream Analytics has built-in support for three data formats: JSON, CSV, and Avro. With custom .NET deserializers, you can read data from other formats such as Protocol Buffer, Bond and other user defined formats for both cloud and edge jobs.

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

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

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

Reference:

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

NEW QUESTION 275

- (Exam Topic 3)

You are designing an enterprise data warehouse in Azure Synapse Analytics that will store website traffic analytics in a star schema.

You plan to have a fact table for website visits. The table will be approximately 5 GB.

You need to recommend which distribution type and index type to use for the table. The solution must provide the fastest query performance.

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

NOTE: Each correct selection is worth one point.

Distribution:

Index:

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Box 1: Hash

Consider using a hash-distributed table when: The table size on disk is more than 2 GB.

The table has frequent insert, update, and delete operations. Box 2: Clustered columnstore

Clustered columnstore tables offer both the highest level of data compression and the best overall query performance.

Reference:

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

NEW QUESTION 276

- (Exam Topic 3)

You have an Azure Data Factory that contains 10 pipelines.

You need to label each pipeline with its main purpose of either ingest, transform, or load. The labels must be available for grouping and filtering when using the monitoring experience in Data Factory.

What should you add to each pipeline?

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

Answer: D

Explanation:

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

Reference:

<https://www.cathrinewilhelmsen.net/annotations-user-properties-azure-data-factory/>

NEW QUESTION 280

- (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 285

.....

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](#)