Microsoft

Exam Questions DP-203

Data Engineering on Microsoft Azure

About Exambible

Your Partner of IT Exam

Found in 1998

Exambible is a company specialized on providing high quality IT exam practice study materials, especially Cisco CCNA, CCDA, CCNP, CCIE, Checkpoint CCSE, CompTIA A+, Network+ certification practice exams and so on. We guarantee that the candidates will not only pass any IT exam at the first attempt but also get profound understanding about the certificates they have got. There are so many alike companies in this industry, however, Exambible has its unique advantages that other companies could not achieve.

Our Advances

* 99.9% Uptime

All examinations will be up to date.

* 24/7 Quality Support

We will provide service round the clock.

* 100% Pass Rate

Our guarantee that you will pass the exam.

* Unique Gurantee

If you do not pass the exam at the first time, we will not only arrange FULL REFUND for you, but also provide you another exam of your claim, ABSOLUTELY FREE!

NEW QUESTION 1
- (Exam Topic 3)
You are designing a data mart for the human resources (MR) department at your company. The data mart will contain information and employee transactions.
From a source system you have a flat extract that has the following fields:
• EmployeeID
• FirstName
• LastName
• Recipient
• GrossArnount
• TransactionID
• GovernmentID
• NetAmountPaid
• TransactionDate
You need to design a start schema data model in an Azure Synapse analytics dedicated SQL pool for the data mart.
Which two tables should you create? Each Correct answer present part of the solution.

A. a dimension table for employee
B. a fabric for Employee
C. a dimension table far EmployeeTransaction
D. a dimension table for Transaction
E. a fact table for Transaction

Answer: AE

Explanation:
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-overvie

NEW QUESTION 2
- (Exam Topic 3)
A company has a real-time data analysis solution that is hosted on Microsoft Azure. The solution uses Azure Event Hub to ingest data and an Azure Stream Analytics cloud job to analyze the data. The cloud job is configured to use 120 Streaming Units (SU).
You need to optimize performance for the Azure Stream Analytics job.
Which two actions should you perform? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.

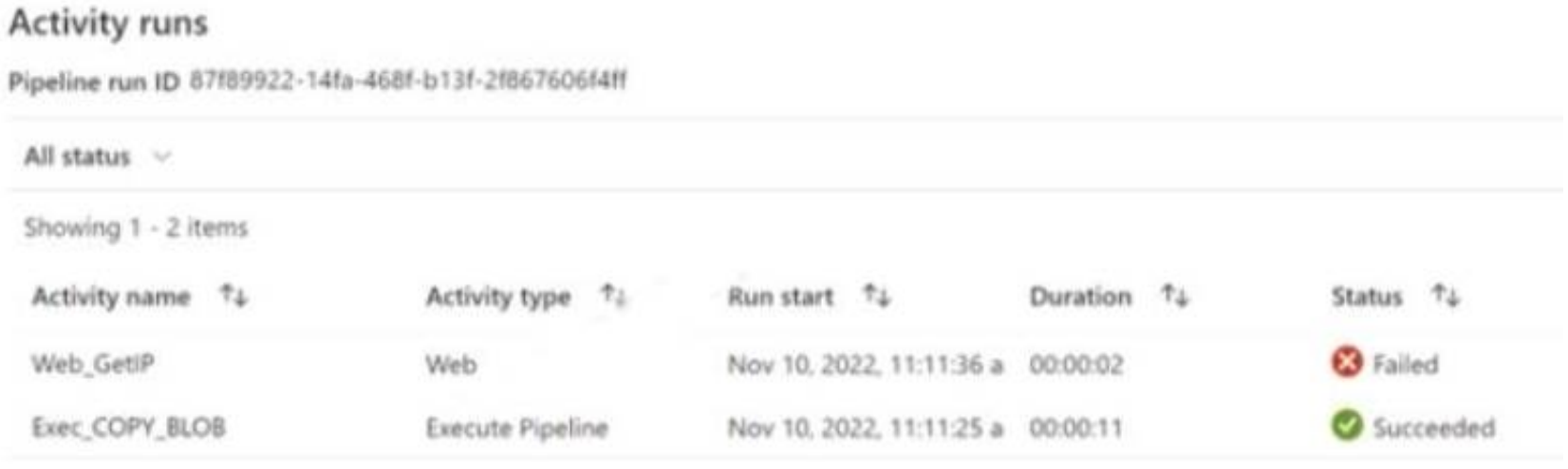
A. Implement event ordering.
B. Implement Azure Stream Analytics user-defined functions (UDF).
C. Implement query parallelization by partitioning the data output.
D. Scale the SU count for the job up.
E. Scale the SU count for the job down.
F. Implement query parallelization by partitioning the data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

Answer: DF

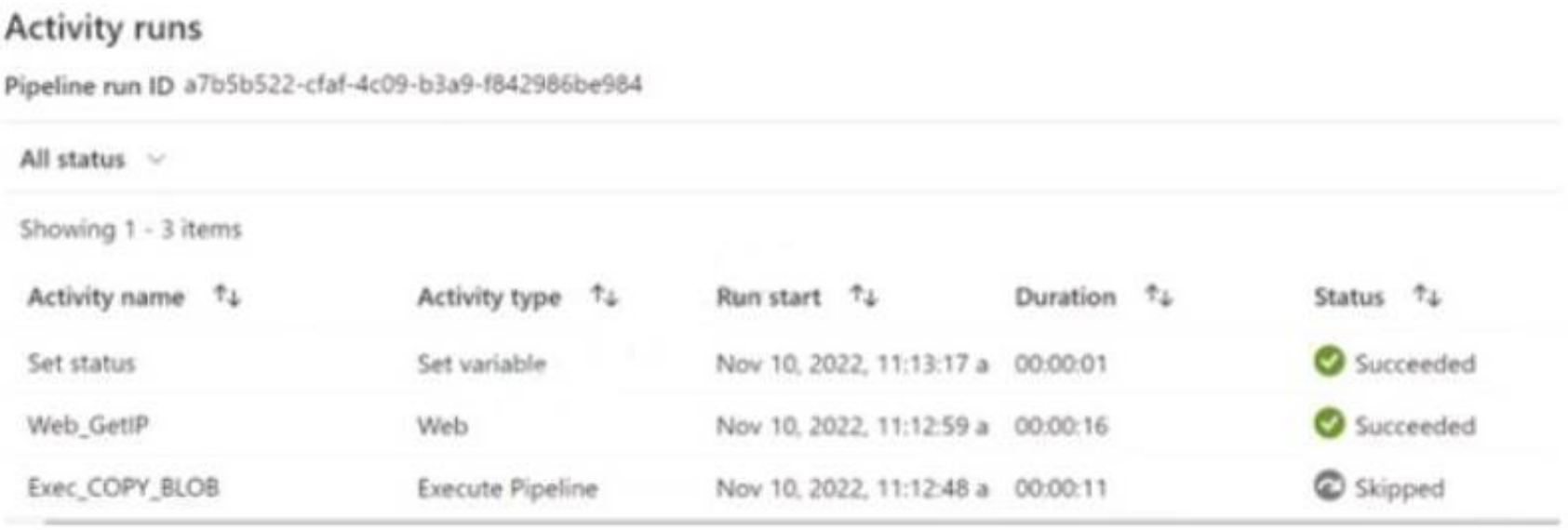
Explanation:
Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-parallelization

NEW QUESTION 3
- (Exam Topic 3)
You have an Azure Data Factory pipeline shown the following exhibit.

The execution log for the first pipeline run is shown in the following exhibit.

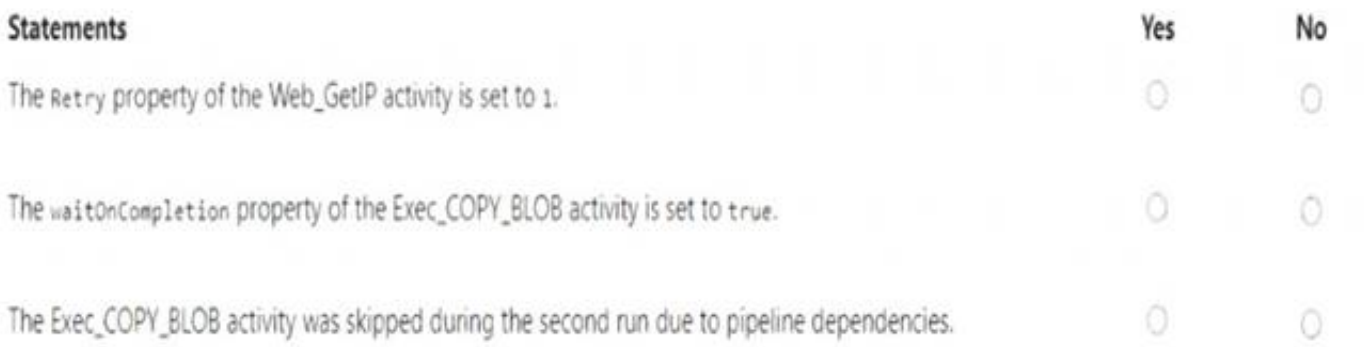


The execution log for the second pipeline run is shown in the following exhibit.



For each of the following statements, select Yes if the statement is true. Otherwise, select No. NOTE: Each correct selection is worth one point.

Answer Area

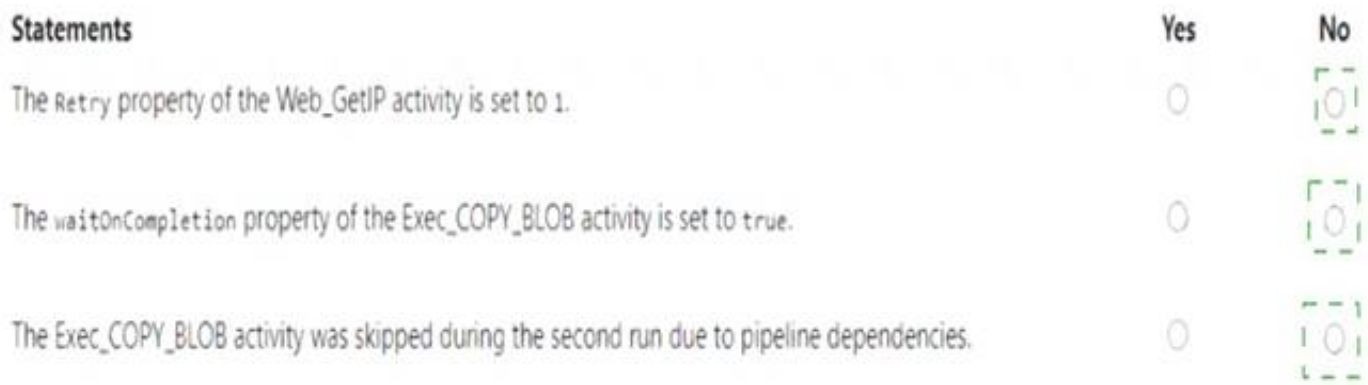


A. Mastered
B. Not Mastered

Answer: A

Explanation:

Answer Area



NEW QUESTION 4
- (Exam Topic 3)
You develop data engineering solutions for a company.
A project requires the deployment of data to Azure Data Lake Storage.
You need to implement role-based access control (RBAC) so that project members can manage the Azure Data Lake Storage resources.
Which three actions should you perform? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

A. Assign Azure AD security groups to Azure Data Lake Storage.
B. Configure end-user authentication for the Azure Data Lake Storage account.
C. Configure service-to-service authentication for the Azure Data Lake Storage account.
D. Create security groups in Azure Active Directory (Azure AD) and add project members.
E. Configure access control lists (ACL) for the Azure Data Lake Storage account.

Answer: ADE

Explanation:
References:
https://docs.microsoft.com/en-us/azure/data-lake-store/data-lake-store-secure-data

NEW QUESTION 5
- (Exam Topic 3)
You implement an enterprise data warehouse in Azure Synapse Analytics. You have a large fact table that is 10 terabytes (TB) in size.
Incoming queries use the primary key SaleKey column to retrieve data as displayed in the following table:

SaleKey	CityKey	CustomerKey	StockItemKey	InvoiceDateKey	Quantity	UnitPrice	TotalExcludingTax
49309	90858	70	69	10/22/13	8	16	128
49313	55710	126	69	10/22/13	2	16	32
49343	44710	234	68	10/22/13	10	16	160
49352	66109	163	70	10/22/13	4	16	64
49488	65312	230	70	10/22/13	8	16	128
49646	85877	271	70	10/24/13	1	16	16
49798	41238	288	69	10/24/13	1	16	16

You need to distribute the large fact table across multiple nodes to optimize performance of the table. Which technology should you use?

A. hash distributed table with clustered index
B. hash distributed table with clustered Columnstore index
C. round robin distributed table with clustered index
D. round robin distributed table with clustered Columnstore index
E. heap table with distribution replicate

Answer: B

Explanation:
Hash-distributed tables improve query performance on large fact tables.
Columnstore indexes can achieve up to 100x better performance on analytics and data warehousing workloads and up to 10x better data compression than traditional rowstore indexes.
Reference:
https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-tables-distribute https://docs.microsoft.com/en-us/sql/relational-databases/indexes/columnstore-indexes-query-performance

NEW QUESTION 6
- (Exam Topic 3)
You have an Azure Synapse Analytics dedicated SQL pool named Pool1 that contains a table named Sales. Sales has row-level security (RLS) applied. RLS uses the following predicate filter.

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

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

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

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

Answer: A

NEW QUESTION 7
- (Exam Topic 3)
You have several Azure Data Factory pipelines that contain a mix of the following types of activities.
* Wrangling data flow
* Notebook
* Copy
* jar
Which two Azure services should you use to debug the activities? Each correct answer presents part of the solution NOTE: Each correct selection is worth one point.

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

Answer: CE

NEW QUESTION 8
- (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 9
- (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 10
- (Exam Topic 3)
You are designing an Azure Databricks cluster that runs user-defined local processes. You need to recommend a cluster configuration that meets the following requirements:
• Minimize query latency.
• Maximize the number of users that can run queues on the cluster at the same time « Reduce overall costs without compromising other requirements
Which cluster type should you recommend?

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

Answer: C

Explanation:
A High Concurrency cluster is a managed cloud resource. The key benefits of High Concurrency clusters are that they provide fine-grained sharing for maximum resource utilization and minimum query latencies.
Databricks chooses the appropriate number of workers required to run your job. This is referred to as autoscaling. Autoscaling makes it easier to achieve high cluster utilization, because you don’t need to provision the cluster to match a workload.
Reference:
https://docs.microsoft.com/en-us/azure/databricks/clusters/configure

NEW QUESTION 10
- (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 12
- (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 16
- (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

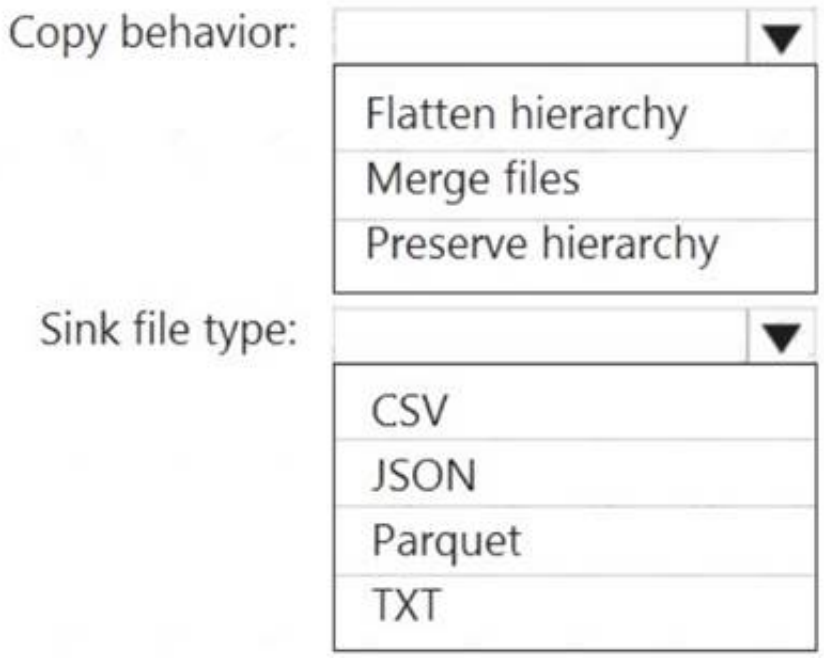
NEW QUESTION 21
- (Exam Topic 3)
You are designing a star schema for a dataset that contains records of online orders. Each record includes an order date, an order due date, and an order ship date.
You need to ensure that the design provides the fastest query times of the records when querying for arbitrary 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 yyyymmdd.
D. In the fact table, use integer columns for the date fields.
E. Use DateTime columns for the date fields.

Answer: BD

NEW QUESTION 25
- (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.

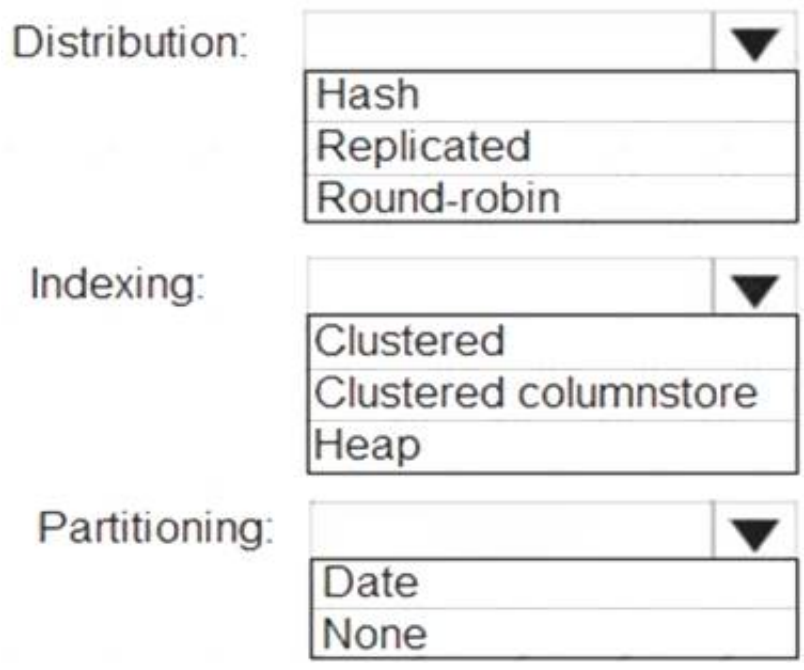


A. Mastered
B. Not Mastered

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 28
- (Exam Topic 3)
You have a SQL pool in Azure Synapse.
You plan to load data from Azure Blob storage to a staging table. Approximately 1 million rows of data will be loaded daily. The table will be truncated before each daily load.
You need to create the staging table. The solution must minimize how long it takes to load the data to the staging table.
How should you configure 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:
Graphical user interface, application, table Description automatically generated
Box 1: Hash
Hash-distributed tables improve query performance on large fact tables. They can have very large numbers of rows and still achieve high performance.
Box 2: Clustered columnstore
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.
Box 3: Date
Table partitions enable you to divide your data into smaller groups of data. In most cases, table partitions are created on a date column.
Partition switching can be used to quickly remove or replace a section of a table. Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-partitio https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribu

NEW QUESTION 30
- (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.

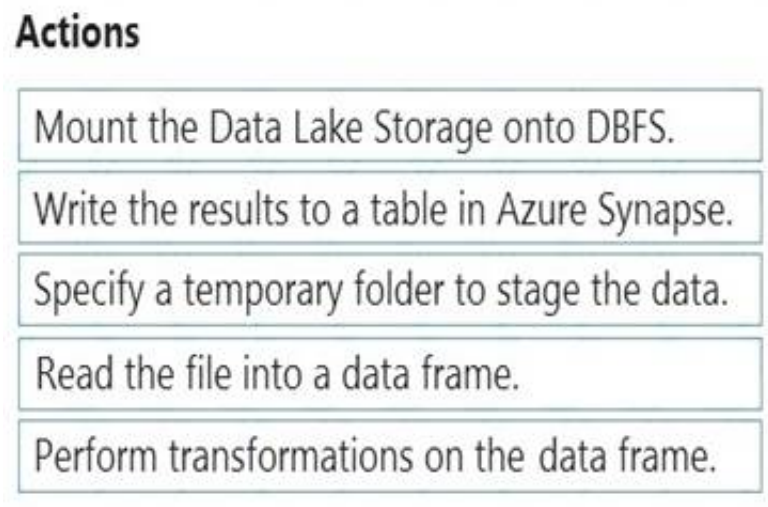
A. Add a pass-through query.
B. Add a temporal analytic function.
C. Scale out the query by using PARTITION BY.
D. Convert the query to a reference query.
E. Increase the number of streaming units.

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 34
- (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.



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 39
- (Exam Topic 3)
You need to schedule an Azure Data Factory pipeline to execute when a new file arrives in an Azure Data Lake Storage Gen2 container.
Which type of trigger should you use?

A. on-demand
B. tumbling window
C. schedule
D. storage event

Answer: D

Explanation:
Event-driven architecture (EDA) is a common data integration pattern that involves production, detection, consumption, and reaction to events. Data integration scenarios often require Data Factory customers to trigger pipelines based on events happening in storage account, such as the arrival or deletion of a file in Azure Blob Storage account.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/how-to-create-event-trigger

NEW QUESTION 43
- (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 46
- (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 47
- (Exam Topic 3)
You have an Azure Factory instance named DF1 that contains a pipeline named PL1.PL1 includes a tumbling window trigger.
You create five clones of PL1. You configure each clone pipeline to use a different data source.
You need to ensure that the execution schedules of the clone pipeline match the execution schedule of PL1. What should you do?

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

Answer: B

NEW QUESTION 52
- (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.	○	○
The Databricks cluster minimizes costs when running scheduled jobs that execute notebooks.	○	○
The Databricks cluster supports the creation of a Delta Lake table.	○	○

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 57
- (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 info 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,5 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

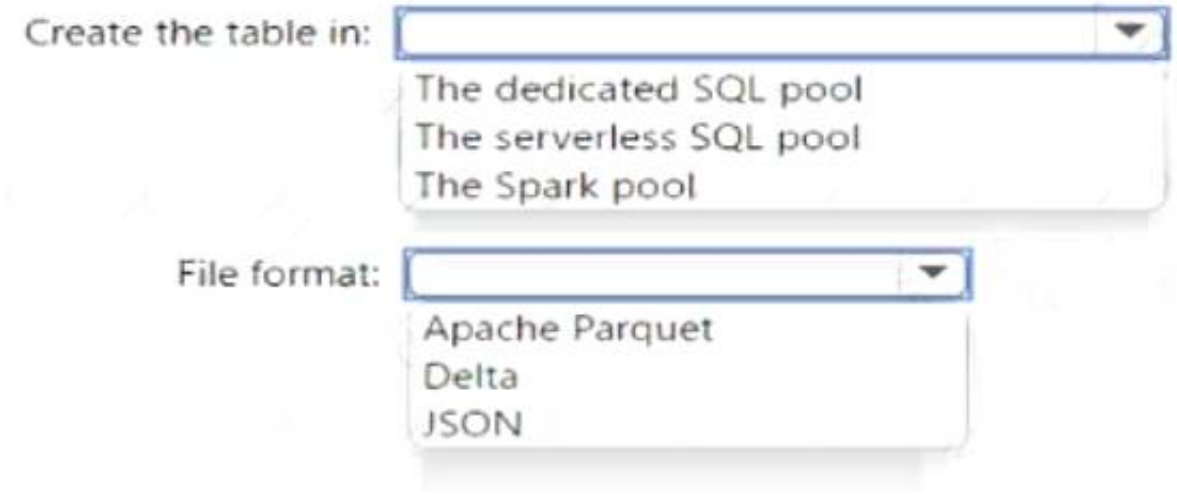
NEW QUESTION 59
- (Exam Topic 3)
You are designing a folder structure for the files m an Azure Data Lake Storage Gen2 account. The account has one container that contains three years of data.
You need to recommend a folder structure that meets the following requirements:
• Supports partition elimination for queries by Azure Synapse Analytics serverless SQL pooh
• Supports fast data retrieval for data from the current month
• Simplifies data security management by department Which folder structure should you recommend?

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

Answer: B

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

NEW QUESTION 63
- (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.

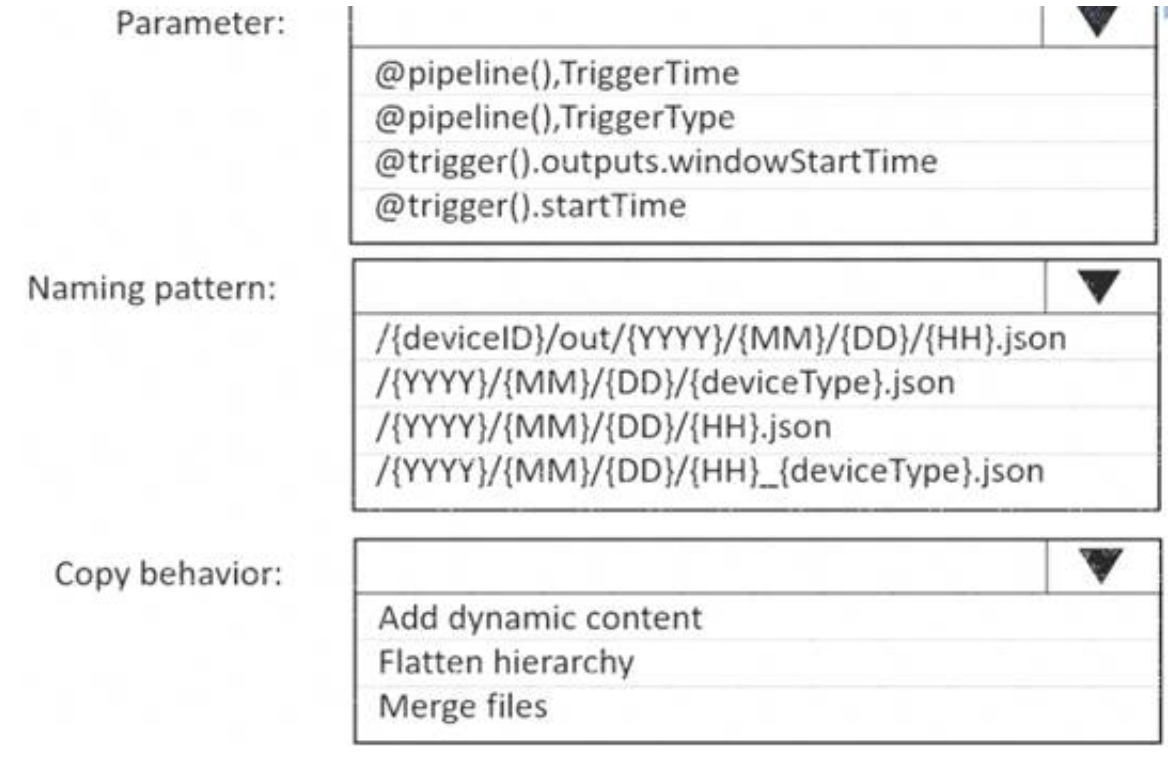


A. Mastered
B. Not Mastered

Answer: A

Explanation:
The dedicated SQL pool Apache Parquet

NEW QUESTION 65
- (Exam Topic 3)
You are building an Azure Data Factory solution to process data received from Azure Event Hubs, and then ingested into an Azure Data Lake Storage Gen2 container.
The data will be ingested every five minutes from devices into JSON files. The files have the following naming pattern.
/{deviceType}/in/{YYYY}/{MM}/{DD}/{HH}/{deviceID}_{YYYY}{MM}{DD}HH}{mm}.json
You need to prepare the data for batch data processing so that there is one dataset per hour per deviceType. The solution must minimize read times.
How should you configure the sink for the copy activity? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



A. Mastered
B. Not Mastered

Answer: A

Explanation:
Box 1: @trigger().startTime
startTime: A date-time value. For basic schedules, the value of the startTime property applies to the first occurrence. For complex schedules, the trigger starts no sooner than the specified startTime value.
Box 2: /{YYYY}/{MM}/{DD}/{HH}_{deviceType}.json One dataset per hour per deviceType.
Box 3: Flatten hierarchy
- FlattenHierarchy: All files from the source folder are in the first level of the target folder. The target files have autogenerated names.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipeline-execution-triggers https://docs.microsoft.com/en-us/azure/data-factory/connector-file-system

NEW QUESTION 68
- (Exam Topic 3)
You need to design a solution that will process streaming data from an Azure Event Hub and output the data to Azure Data Lake Storage. The solution must ensure that analysts can interactively query the streaming data.
What should you use?

A. event triggers in Azure Data Factory
B. Azure Stream Analytics and Azure Synapse notebooks
C. Structured Streaming in Azure Databricks
D. Azure Queue storage and read-access geo-redundant storage (RA-GRS)

Answer: C

Explanation:
Apache Spark Structured Streaming is a fast, scalable, and fault-tolerant stream processing API. You can use it to perform analytics on your streaming data in near real-time.
With Structured Streaming, you can use SQL queries to process streaming data in the same way that you would process static data.
Azure Event Hubs is a scalable real-time data ingestion service that processes millions of data in a matter of seconds. It can receive large amounts of data from multiple sources and stream the prepared data to Azure Data Lake or Azure Blob storage.
Azure Event Hubs can be integrated with Spark Structured Streaming to perform the processing of messages in near real-time. You can query and analyze the processed data as it comes by using a Structured Streaming query and Spark SQL.
Reference:
https://k21academy.com/microsoft-azure/data-engineer/structured-streaming-with-azure-event-hubs/

NEW QUESTION 70
- (Exam Topic 3)
You have an Azure Synapse Analytics Apache Spark pool named Pool1.
You plan to load JSON files from an Azure Data Lake Storage Gen2 container into the tables in Pool1. The structure and data types vary by file.
You need to load the files into the tables. The solution must maintain the source data types. What should you do?

A. Use a Get Metadata activity in Azure Data Factory.
B. Use a Conditional Split transformation in an Azure Synapse data flow.
C. Load the data by using the OPEHROwset Transact-SQL command in an Azure Synapse Anarytics serverless SQL pool.
D. Load the data by using PySpark.

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.
Serverless SQL pool enables you to query data in your data lake. It offers a T-SQL query surface area that accommodates semi-structured and unstructured data queries.

To support a smooth experience for in place querying of data that's located in Azure Storage files, serverless SQL pool uses the OPENROWSET function with additional capabilities.
The easiest way to see to the content of your JSON file is to provide the file URL to the OPENROWSET function, specify csv FORMAT.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/query-json-files https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/query-data-storage

NEW QUESTION 73
- (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 76
- (Exam Topic 3)
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You plan to create an Azure Databricks workspace that has a tiered structure. The workspace will contain the following three workloads:
- A workload for data engineers who will use Python and SQL.
- A workload for jobs that will run notebooks that use Python, Scala, and SOL.
- A workload that data scientists will use to perform ad hoc analysis in Scala and R.

The enterprise architecture team at your company identifies the following standards for Databricks environments:
- The data engineers must share a cluster.
- The job cluster will be managed by using a request process whereby data scientists and data engineers provide packaged notebooks for deployment to the cluster.
- All the data scientists must be assigned their own cluster that terminates automatically after 120 minutes of inactivity. Currently, there are three data scientists.

You need to create the Databricks clusters for the workloads.
Solution: You create a Standard cluster for each data scientist, a High Concurrency cluster for the data engineers, and a High Concurrency cluster for the jobs.
Does this meet the goal?

A. Yes
B. No

Answer: A

Explanation:
We need a High Concurrency cluster for the data engineers and the jobs. Note:
Standard clusters are recommended for a single user. Standard can run workloads developed in any language: Python, R, Scala, and SQL.
A high concurrency cluster is a managed cloud resource. The key benefits of high concurrency clusters are that they provide Apache Spark-native fine-grained sharing for maximum resource utilization and minimum query latencies.
Reference: https://docs.azuredatabricks.net/clusters/configure.html

NEW QUESTION 77
- (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.

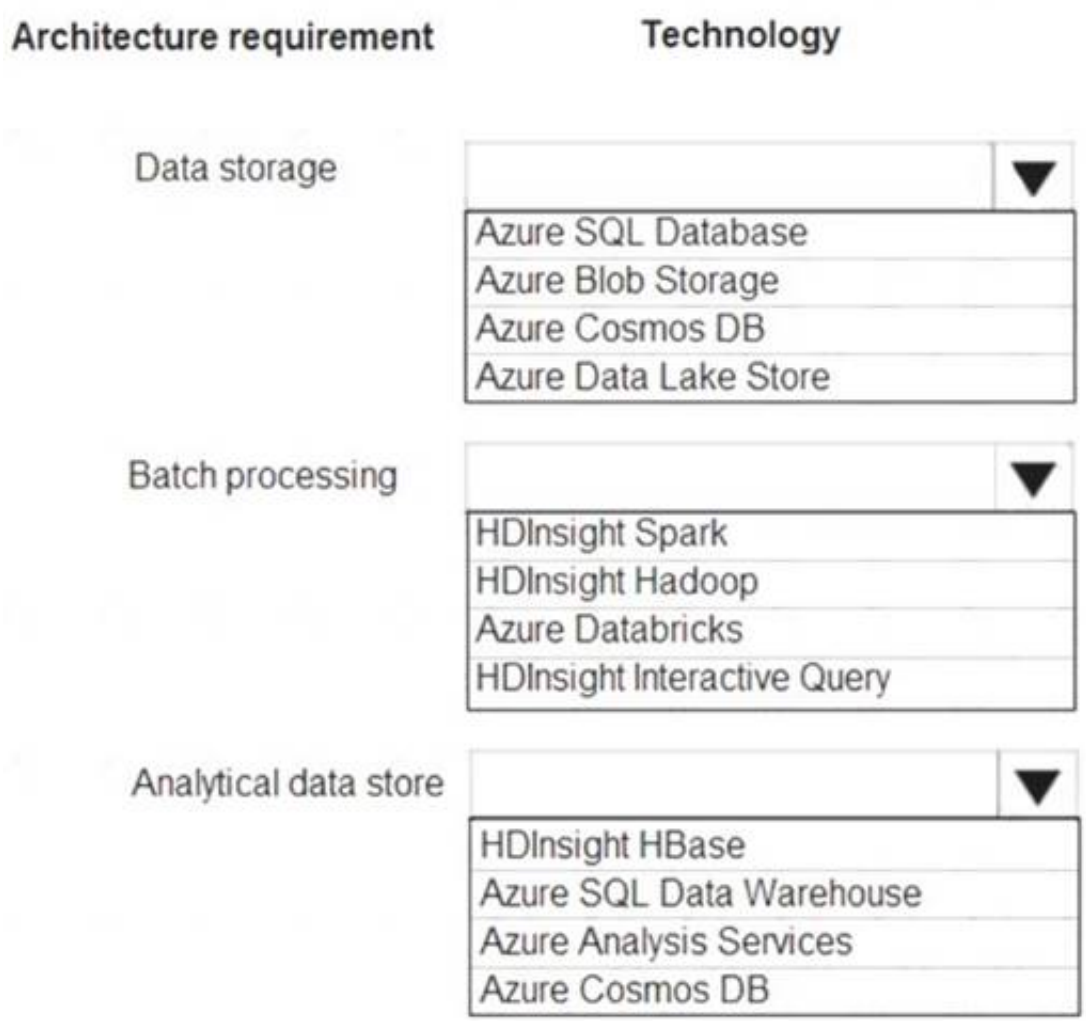
```
SELECT
    payment_type,
    SUM(fare_amount) AS fare_total
FROM OPENROWSET(
        BULK 'csv/busfare/tripdata_2020*.csv',
        DATA_SOURCE = 'BusData',
        FORMAT = 'CSV', PARSER_VERSION = '2.0',
        FIRSTROW = 2
    )
    WITH (
        payment_type INT 10,
        fare_amount FLOAT 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 80
- (Exam Topic 3)
You are developing a solution using a Lambda architecture on Microsoft Azure. The data at test layer must meet the following requirements:
Data storage:
•Serve as a repository (or high volumes of large files in various formats.
•Implement optimized storage for big data analytics workloads.
•Ensure that data can be organized using a hierarchical structure. Batch processing:
•Use a managed solution for in-memory computation processing.
•Natively support Scala, Python, and R programming languages.
•Provide the ability to resize and terminate the cluster automatically. Analytical data store:
•Support parallel processing.
•Use columnar storage.
•Support SQL-based languages.
You need to identify the correct technologies to build the Lambda architecture.
Which technologies should you use? To answer, select the appropriate options in the answer area NOTE: Each correct selection is worth one point.



A. Mastered
B. Not Mastered

Answer: A

Explanation:
Data storage: Azure Data Lake Store
A key mechanism that allows Azure Data Lake Storage Gen2 to provide file system performance at object storage scale and prices is the addition of a hierarchical

namespace. This allows the collection of objects/files within an account to be organized into a hierarchy of directories and nested subdirectories in the same way that the file system on your computer is organized. With the hierarchical namespace enabled, a storage account becomes capable of providing the scalability and cost-effectiveness of object storage, with file system semantics that are familiar to analytics engines and frameworks.
Batch processing: HD Insight Spark
Aparch Spark is an open-source, parallel-processing framework that supports in-memory processing to boost the performance of big-data analysis applications. HDInsight is a managed Hadoop service. Use it deploy and manage Hadoop clusters in Azure. For batch processing, you can use Spark, Hive, Hive LLAP, MapReduce.
Languages: R, Python, Java, Scala, SQL Analytic data store: SQL Data Warehouse
SQL Data Warehouse is a cloud-based Enterprise Data Warehouse (EDW) that uses Massively Parallel Processing (MPP).
SQL Data Warehouse stores data into relational tables with columnar storage. References:
https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-namespace https://docs.microsoft.com/en-us/azure/architecture/data-guide/technology-choices/batch-processing https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-overview-what-is

NEW QUESTION 85
- (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 90
- (Exam Topic 3)
You have the following Azure Stream Analytics query.

```
WITH

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

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

For each of the following statements, select Yes if the statement is true. Otherwise, select No.
NOTE: Each correct selection is worth one point.

Statements	Yes	No
The query combines two streams of partitioned data.	O	O
The stream scheme key and count must match the output scheme.	O	O
Providing 60 streaming units will optimize the performance of the query.	O	O

A. Mastered
B. Not Mastered

Answer: A

Explanation:

Box 1: No
Note: You can now use a new extension of Azure Stream Analytics SQL to specify the number of partitions of a stream when reshuffling the data.
The outcome is a stream that has the same partition scheme. Please see below for an example: WITH step1 AS (SELECT * FROM [input1] PARTITION BY DeviceID INTO 10),
step2 AS (SELECT * FROM [input2] PARTITION BY DeviceID INTO 10)
SELECT * INTO [output] FROM step1 PARTITION BY DeviceID UNION step2 PARTITION BY DeviceID Note: The new extension of Azure Stream Analytics SQL includes a keyword INTO that allows you to specify the number of partitions for a stream when performing reshuffling using a PARTITION BY statement.
Box 2: Yes
When joining two streams of data explicitly repartitioned, these streams must have the same partition key and partition count. Box 3: Yes
Streaming Units (SUs) represents the computing resources that are allocated to execute a Stream Analytics job. The higher the number of SUs, the more CPU and memory resources are allocated for your job.
In general, the best practice is to start with 6 SUs for queries that don't use PARTITION BY. Here there are 10 partitions, so 6x10 = 60 SUs is good.
Note: Remember, Streaming Unit (SU) count, which is the unit of scale for Azure Stream Analytics, must be adjusted so the number of physical resources available to the job can fit the partitioned flow. In general, six SUs is a good number to assign to each partition. In case there are insufficient resources assigned to the job, the system will only apply the repartition if it benefits the job.
Reference:
https://azure.microsoft.com/en-in/blog/maximize-throughput-with-repartitioning-in-azure-stream-analytics/ https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-streaming-unit-consumption

NEW QUESTION 92
- (Exam Topic 3)
You are designing an Azure Synapse solution that will provide a query interface for the data stored in an Azure Storage account. The storage account is only accessible from a virtual network.
You need to recommend an authentication mechanism to ensure that the solution can access the source data.
What should you recommend?

A. a managed identity
B. anonymous public read access
C. a shared key

Answer: A

Explanation:
Managed Identity authentication is required when your storage account is attached to a VNet. Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/quickstart-bulk-load-copy-tsql-exa

NEW QUESTION 97
- (Exam Topic 3)
You have an enterprise data warehouse in Azure Synapse Analytics that contains a table named FactOnlineSales. The table contains data from the start of 2009 to the end of 2012.
You need to improve the performance of queries against FactOnlineSales by using table partitions. The solution must meet the following requirements:

- Create four partitions based on the order date.
- Ensure that each partition contains all the orders places during a given calendar year.

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

```
CREATE TABLE [dbo].FactOnlineSales
([OnlineSalesKey] [int] NOT NULL,
[OrderDateKey] [datetime]    NOT NULL,
[StoreKey] [int]         NOT NULL,
[ProductKey] [int]       NOT NULL,
[CustomerKey] [int]      NOT NULL,
[SalesOrderNumber] [varchar](20) NOT NULL,
[SalesQuantity] [int]    NOT NULL,
[SalesAmount] [money]    NOT NULL,
[UnitPrice]    [money]   NULL)
WITH (CLUSTERED COLUMNSTORE INDEX)
PARTITION ([OrderDateKey] RANGE [ RIGHT / LEFT ] FOR VALUES
(
  [ 20090101,20121231 / 20100101,20110101,20120101 / 20090101,20100101,20110101,20120101 ]
)
```

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Text Description automatically generated

Range Left or Right, both are creating similar partition but there is difference in comparison For example: in this scenario, when you use LEFT and 20100101,20110101,20120101
Partition will be, datecol<=20100101, datecol>20100101 and datecol<=20110101, datecol>20110101 and datecol<=20120101, datecol>20120101
But if you use range RIGHT and 20100101,20110101,20120101
Partition will be, datecol<20100101, datecol>=20100101 and datecol<20110101, datecol>=20110101 and datecol<20120101, datecol>=20120101
In this example, Range RIGHT will be suitable for calendar comparison Jan 1st to Dec 31st Reference:
https://docs.microsoft.com/en-us/sql/t-sql/statements/create-partition-function-transact-sql?view=sql-server-ver1

NEW QUESTION 99
- (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 tabte in SQLPooM 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 100
- (Exam Topic 3)
You have an Azure Data Lake Storage Gen2 container that contains 100 TB of data.
You need to ensure that the data in the container is available for read workloads in a secondary region if an outage occurs in the primary region. The solution must minimize costs.
Which type of data redundancy should you use?
A. zone-redundant storage (ZRS)
B. read-access geo-redundant storage (RA-GRS)
C. locally-redundant storage (LRS)
D. geo-redundant storage (GRS)

Answer: B

Explanation:
Geo-redundant storage (with GRS or GZRS) replicates your data to another physical location in the secondary region to protect against regional outages. However, that data is available to be read only if the customer or Microsoft initiates a failover from the primary to secondary region. When you enable read access to the secondary region, your data is available to be read at all times, including in a situation where the primary region becomes unavailable.
Reference:
https://docs.microsoft.com/en-us/azure/storage/common/storage-redundancy

NEW QUESTION 105
- (Exam Topic 3)
You have an Azure Synapse Analytics dedicated SQL pool that contains the users shown in the following table.

Name	Role
User1	Server admin
User2	db_datereader

User1 executes a query on the database, and the query returns the results shown in the following exhibit.

```
SELECT c.name,
    tbl.name as table_name,
    typ.name as datatype,
    c.is_masked,
    c.masking_function
FROM sys.masked_columns AS c
INNER JOIN sys.tables AS tbl ON c.[object_id] = tbl.[object_id]
INNER JOIN sys.types typ ON c.user_type_id = typ.user_type_id
WHERE is_masked = 1;
```

Results Messages

	name	table_name	datatype	is_masked	masking_function
1	BirthDate	DimCustomer	date	1	default()
2	Gender	DimCustomer	nvarchar	1	default()
3	EmailAddress	DimCustomer	nvarchar	1	email()
4	YearlyIncome	DimCustomer	money	1	default()

User1 is the only user who has access to the unmasked data.
Use the drop-down menus to select the answer choice that completes each statement based on the information presented in the graphic.
NOTE: Each correct selection is worth one point.

When User2 queries the YearlyIncome column, the values returned will be **[answer choice]**.

- a random number
- the values stored in the database
- XXXX
- 0

When User1 queries the BirthDate column, the values returned will be **[answer choice]**.

- a random date
- the values stored in the database
- XXXX
- 1900-01-01

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Graphical user interface, text, application, email Description automatically generated
Box 1: 0
The YearlyIncome column is of the money data type.
The Default masking function: Full masking according to the data types of the designated fields

- Use a zero value for numeric data types (bigint, bit, decimal, int, money, numeric, smallint, smallmoney, tinyint, float, real).

Box 2: the values stored in the database
Users with administrator privileges are always excluded from masking, and see the original data without any mask.
Reference:
https://docs.microsoft.com/en-us/azure/azure-sql/database/dynamic-data-masking-overview

NEW QUESTION 107
- (Exam Topic 3)
A company plans to use Platform-as-a-Service (PaaS) to create the new data pipeline process. The process must meet the following requirements:
Ingest:

- Access multiple data sources.
- Provide the ability to orchestrate workflow.
- Provide the capability to run SQL Server Integration Services packages.

Store:
Optimize storage for big data workloads. Provide encryption of data at rest. Operate with no size limits.
Prepare and Train:

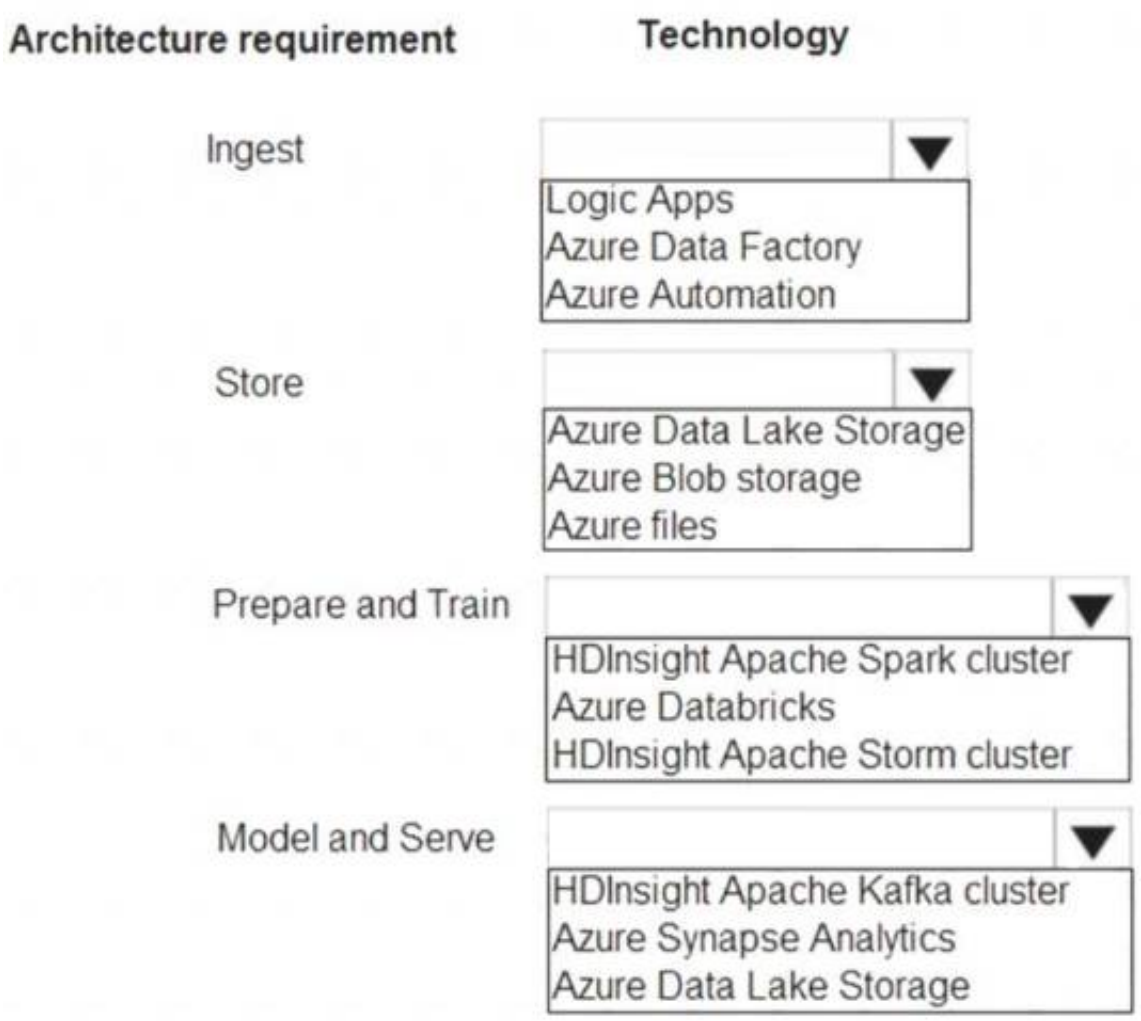
- Provide a fully-managed and interactive workspace for exploration and visualization.
- Provide the ability to program in R, SQL, Python, Scala, and Java.
-

Provide seamless user authentication with Azure Active Directory. Model & Serve:

- Implement native columnar storage.
- Support for the SQL language
- Provide support for structured streaming. You need to build the data integration pipeline.

Which technologies 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

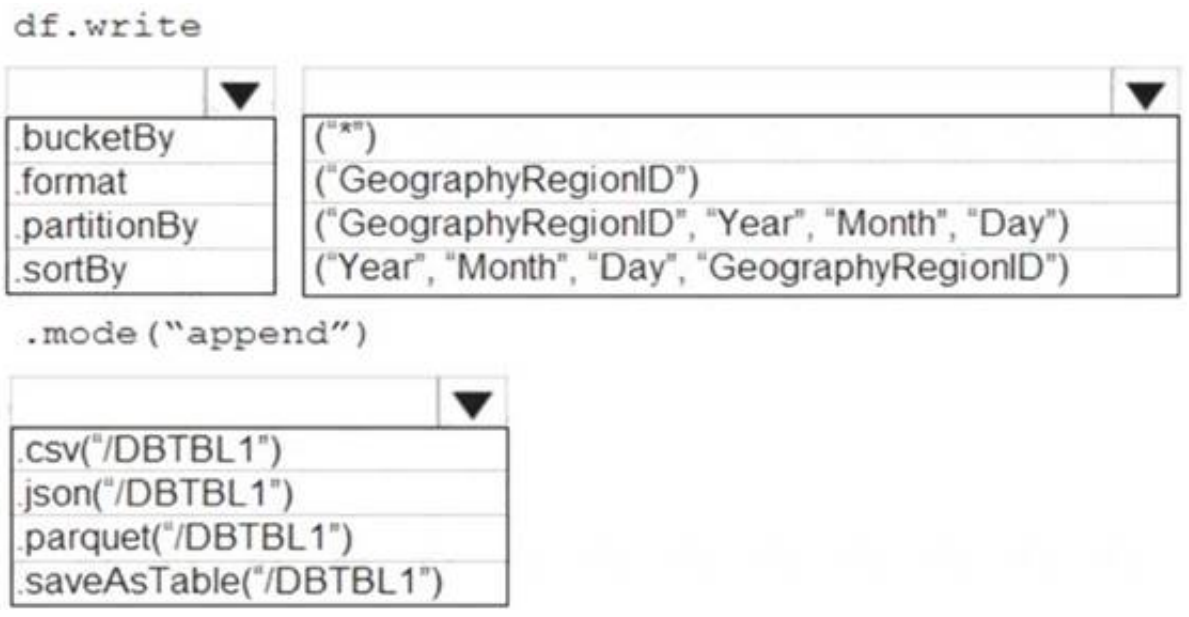
Answer: A

Explanation:
Graphical user interface, application, table, email Description automatically generated

NEW QUESTION 110
- (Exam Topic 3)
You develop a dataset named DBTBL1 by using Azure Databricks. DBTBL1 contains the following columns:

- SensorTypeID
- GeographyRegionID
- Year
- Month
- Day
- Hour
- Minute
- Temperature
- WindSpeed
- Other

You need to store the data to support daily incremental load pipelines that vary for each GeographyRegionID. The solution must minimize storage costs.
How should you complete the code? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



A. Mastered
B. Not Mastered

Answer: A

Explanation:
Graphical user interface, text, application Description automatically generated

NEW QUESTION 111
- (Exam Topic 3)
You have an Azure Stream Analytics job.
You need to ensure that the job has enough streaming units provisioned. You configure monitoring of the SU % Utilization metric.
Which two additional metrics should you monitor? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.

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

Answer: AB

Explanation:
To react to increased workloads and increase streaming units, consider setting an alert of 80% on the SU Utilization metric. Also, you can use watermark delay and backlogged events metrics to see if there is an impact.
Note: Backlogged Input Events: Number of input events that are backlogged. A non-zero value for this metric implies that your job isn't able to keep up with the number of incoming events. If this value is slowly increasing or consistently non-zero, you should scale out your job, by increasing the SUs.
Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-monitoring

NEW QUESTION 115
- (Exam Topic 3)
You are designing a partition strategy for a fact table in an Azure Synapse Analytics dedicated SQL pool. The table has the following specifications:
• Contain sales data for 20,000 products.
• Use hash distribution on a column named ProductID,
• Contain 2.4 billion records for the years 2019 and 2020.
Which number of partition ranges provides optimal compression and performance of the clustered columnstore index?

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

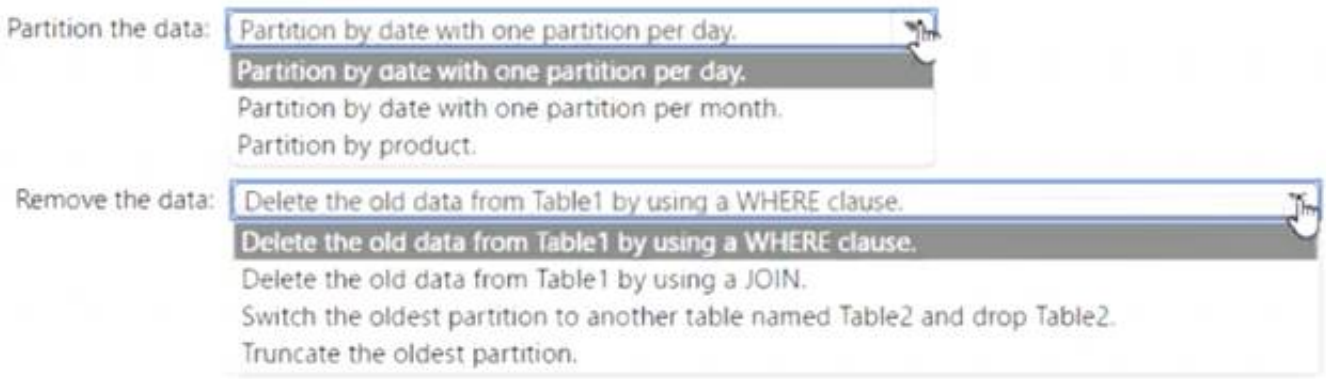
Answer: A

Explanation:
Each partition should have around 1 millions records. Dedication SQL pools already have 60 partitions. We have the formula: Records/(Partitions*60)= 1 million
Partitions= Records/(1 million * 60)
Partitions= 2.4 x 1,000,000,000/(1,000,000 * 60) = 40
Note: Having too many partitions can reduce the effectiveness of clustered columnstore indexes if each partition has fewer than 1 million rows. Dedicated SQL pools automatically partition your data into 60 databases. So, if you create a table with 100 partitions, the result will be 6000 partitions.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/best-practices-dedicated-sql-pool

NEW QUESTION 117
- (Exam Topic 3)
You have an Azure Synapse Analytics dedicated SQL pool named Pool1. Pool1 contains a fact table named Tablet. 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

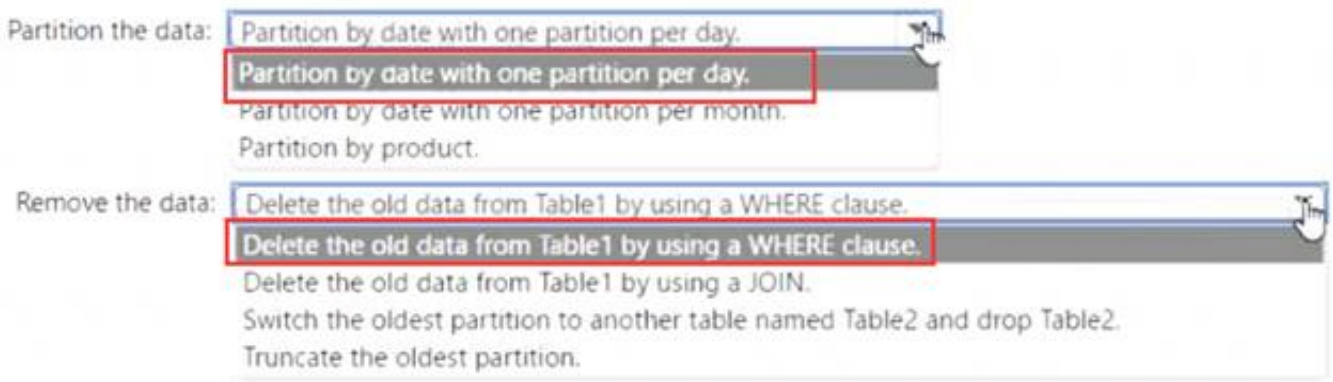


A. Mastered
B. Not Mastered

Answer: A

Explanation:

Answer Area



NEW QUESTION 119
- (Exam Topic 3)
You are building an Azure Stream Analytics job to identify how much time a user spends interacting with a feature on a webpage.
The job receives events based on user actions on the webpage. Each row of data represents an event. Each event has a type of either 'start' or 'end'.
You need to calculate the duration between start and end events.
How should you complete the query? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Box 1: DATEDIFF
DATEDIFF function returns the count (as a signed integer value) of the specified datepart boundaries crossed between the specified startdate and enddate.
Syntax: DATEDIFF (datepart , startdate, enddate) Box 2: LAST
The LAST function can be used to retrieve the last event within a specific condition. In this example, the condition is an event of type Start, partitioning the search by PARTITION BY user and feature. This way, every user and feature is treated independently when searching for the Start event. LIMIT DURATION limits the search back in time to 1 hour between the End and Start events.
Example: SELECT
[user], feature, DATEDIFF(
second,
LAST(Time) OVER (PARTITION BY [user], feature LIMIT DURATION(hour,
1) WHEN Event = 'start'), Time) as duration
FROM input TIMESTAMP BY Time

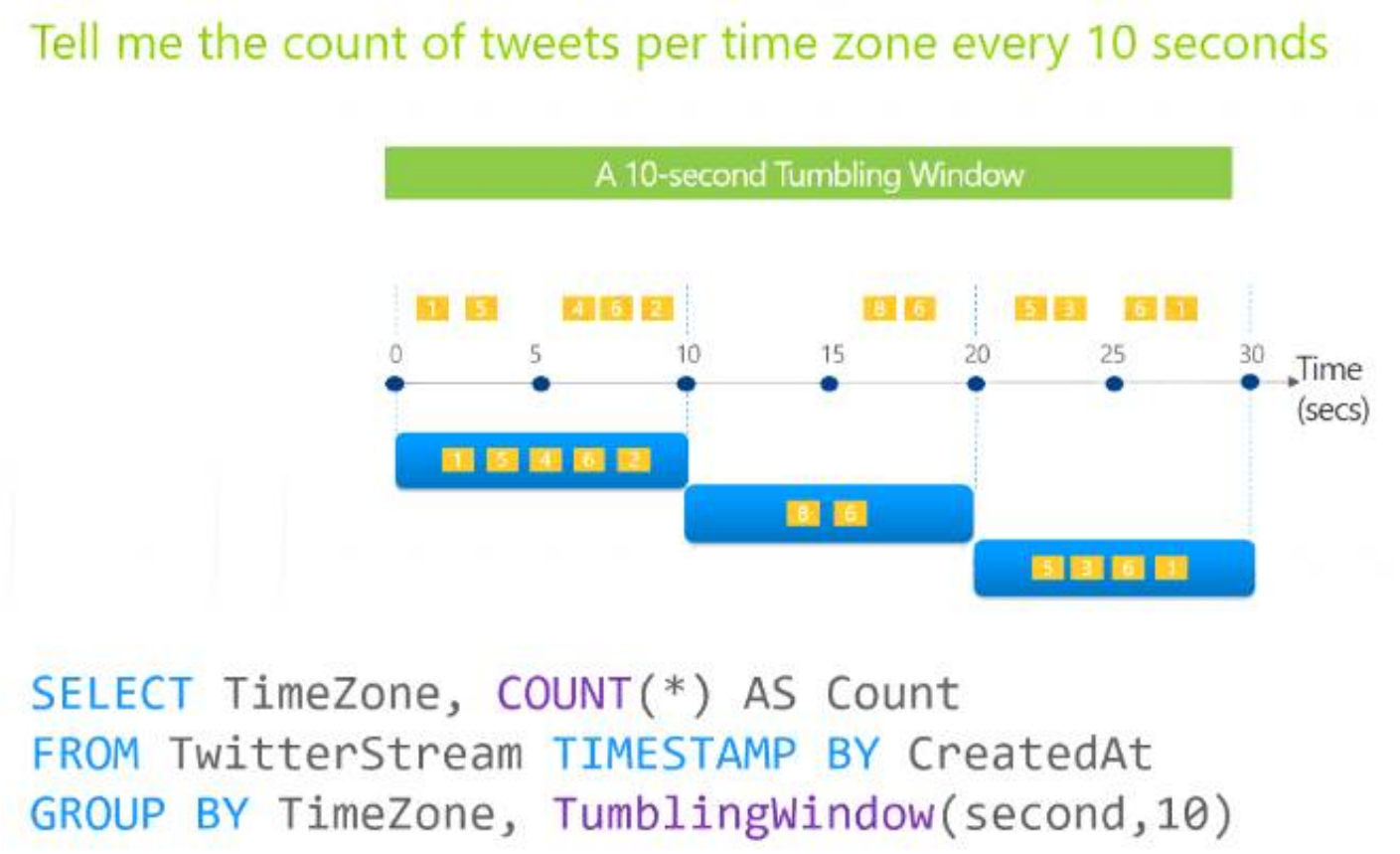
WHERE
Event = 'end' Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-stream-analytics-query-patterns

NEW QUESTION 123
- (Exam Topic 3)
You are designing an Azure Stream Analytics job to process incoming events from sensors in retail environments.
You need to process the events to produce a running average of shopper counts during the previous 15 minutes, calculated at five-minute intervals.
Which type of window should you use?

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

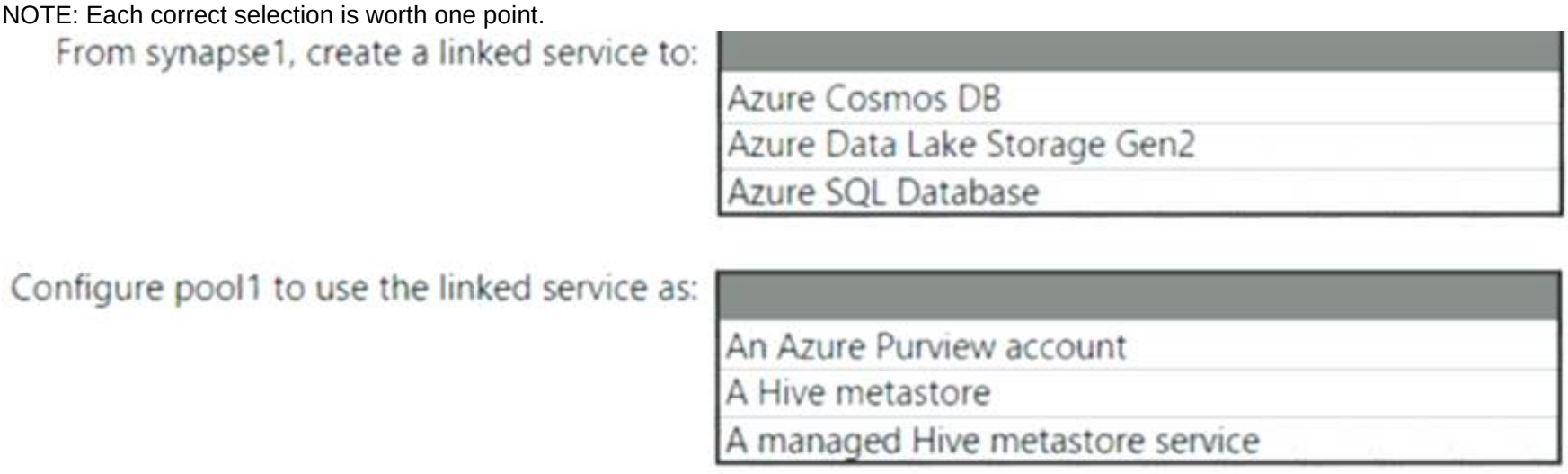
Answer: B

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



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

NEW QUESTION 128
- (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.



A. Mastered
B. Not Mastered

Answer: A

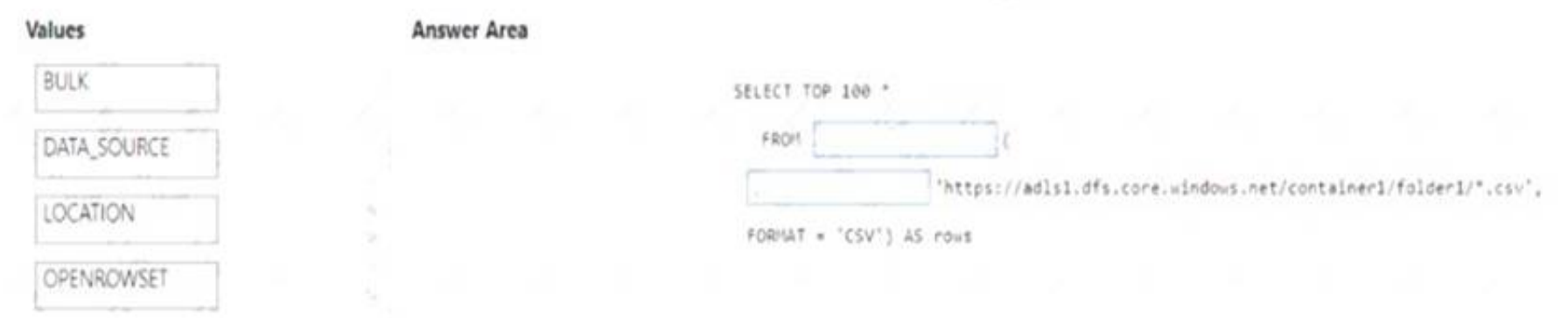
Explanation:
Box 1: Azure SQL Database
Use external Hive Metastore for Synapse Spark Pool
Azure Synapse Analytics allows Apache Spark pools in the same workspace to share a managed HMS (Hive Metastore) compatible metastore as their catalog.

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 133
- (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 137
- (Exam Topic 3)
You are designing a fact table named FactPurchase in an Azure Synapse Analytics dedicated SQL pool. The table contains purchases from suppliers for a retail store. FactPurchase will contain the following columns.

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

FactPurchase will have 1 million rows of data added daily and will contain three years of data. Transact-SQL queries similar to the following query will be executed daily.
SELECT
SupplierKey, StockItemKey, COUNT(*) FROM FactPurchase
WHERE DateKey >= 20210101 AND DateKey <= 20210131
GROUP By SupplierKey, StockItemKey
Which table distribution will minimize query times?

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

Answer: D

Explanation:
Hash-distributed tables improve query performance on large fact tables, and are the focus of this article. Round-robin tables are useful for improving loading speed.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribu

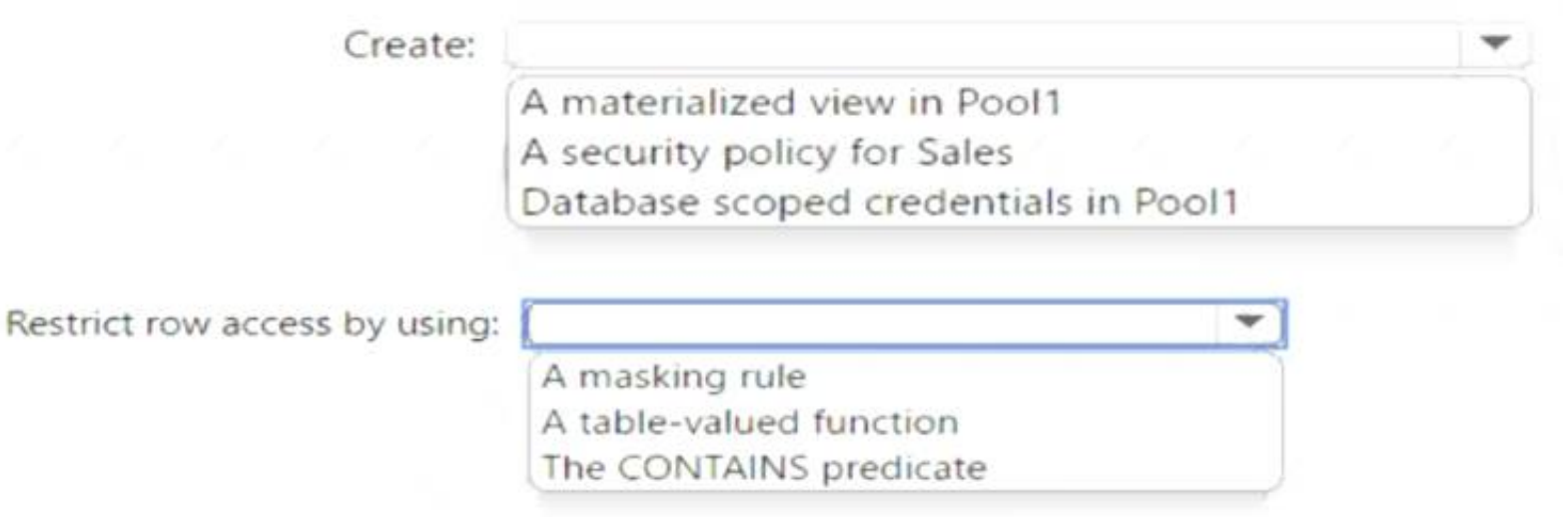
NEW QUESTION 140
- (Exam Topic 3)
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this scenario, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You have an Azure Storage account that contains 100 GB of files. The files contain text and numerical values. 75% of the rows contain description data that has an average length of 1.1 MB.
You plan to copy the data from the storage account to an Azure SQL data warehouse. You need to prepare the files to ensure that the data copies quickly.
Solution: You modify the files to ensure that each row is less than 1 MB. Does this meet the goal?

A. Yes
B. No

Answer: A

Explanation:
When exporting data into an ORC File Format, you might get Java out-of-memory errors when there are large text columns. To work around this limitation, export only a subset of the columns.
References:
https://docs.microsoft.com/en-us/azure/sql-data-warehouse/guidance-for-loading-data

NEW QUESTION 141
- (Exam Topic 3)
You have an Azure Synapse Analytics dedicated SQL pool named Pool1 that contains an external table named Sales. Sales contains sales data. Each row in Sales
contains data on a single sale, including the name of the salesperson.
You need to implement row-level security (RLS). The solution must ensure that the salespeople can access only their respective sales.
What should you do? 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:
Box 1: A security policy for sale
Here are the steps to create a security policy for Sales:

- Create a user-defined function that returns the name of the current user:
- CREATE FUNCTION dbo.GetCurrentUser()
- RETURNS NVARCHAR(128)
- AS
- BEGIN
- RETURN SUSER_SNAME();
- END;
- Create a security predicate function that filters the Sales table based on the current user:
- CREATE FUNCTION dbo.SalesPredicate(@salesperson NVARCHAR(128))
- RETURNS TABLE
- WITH SCHEMABINDING
- AS
- RETURN SELECT 1 AS access_result
- WHERE @salesperson = SalespersonName;
- Create a security policy on the Sales table that uses the SalesPredicate function to filter the data:
- CREATE SECURITY POLICY SalesFilter
- ADD FILTER PREDICATE dbo.SalesPredicate(dbo.GetCurrentUser()) ON dbo.Sales
- WITH (STATE = ON);

By creating a security policy for the Sales table, you ensure that each salesperson can only access their own sales data. The security policy uses a user-defined function to get the name of the current user and a security predicate function to filter the Sales table based on the current user.
Box 2: table-value function
to restrict row access by using row-level security, you need to create a table-valued function that returns a table of values that represent the rows that a user can access. You then use this function in a security
policy that applies a predicate on the table.

NEW QUESTION 144
- (Exam Topic 3)
You have an Azure subscription that contains a Microsoft Purview account named MP1, an Azure data factory named DF1, and a storage account named storage. MP1 is configured
10 scan storage1. DF1 is connected to MP1 and contains 3 dataset named DS1. DS1 references 2 file in storage.
In DF1, you plan to create a pipeline that will process data from DS1.
You need to review the schema and lineage information in MP1 for the data referenced by DS1.
Which two features can you use to locate the information? Each correct answer presents a complete solution. NOTE: Each correct answer is worth one point.

A. the Storage browser of storage1 in the Azure portal
B. the search bar in the Azure portal
C. the search bar in Azure Data Factory Studio
D. the search bar in the Microsoft Purview governance portal

Answer: CD

Explanation:

- The search bar in the Microsoft Purview governance portal: This is a feature that allows you to search for assets in your data estate using keywords, filters, and facets. You can use the search bar to find the files in storage1 that are referenced by DS1, and then view their schema and lineage information in the asset details page12.
- The search bar in Azure Data Factory Studio: This is a feature that allows you to search for datasets, linked services, pipelines, and other resources in your data factory. You can use the search bar to find DS1 in DF1, and then view its schema and lineage information in the dataset details page. You can also click on the Open in Purview button to open the corresponding asset in MP13.

The two features that can be used to locate the schema and lineage information for the data referenced by DS1 are the search bar in Azure Data Factory Studio and the search bar in the Microsoft Purview governance portal.

The search bar in Azure Data Factory Studio allows you to search for the dataset DS1 and view its properties and lineage. This can help you locate information about the source and destination data stores, as well as the transformations that were applied to the data.
The search bar in the Microsoft Purview governance portal allows you to search for the storage account and view its metadata, including schema and lineage information. This can help you understand the different data assets that are stored in the storage account and how they are related to each other.
The Storage browser of storage1 in the Azure portal may allow you to view the files that are stored in the storage account, but it does not provide lineage or schema information for those files. Similarly, the search bar in the Azure portal may allow you to search for resources in the Azure subscription, but it does not provide detailed information about the data assets themselves.
References:

- What is Azure Purview?
- Use Azure Data Factory Studio

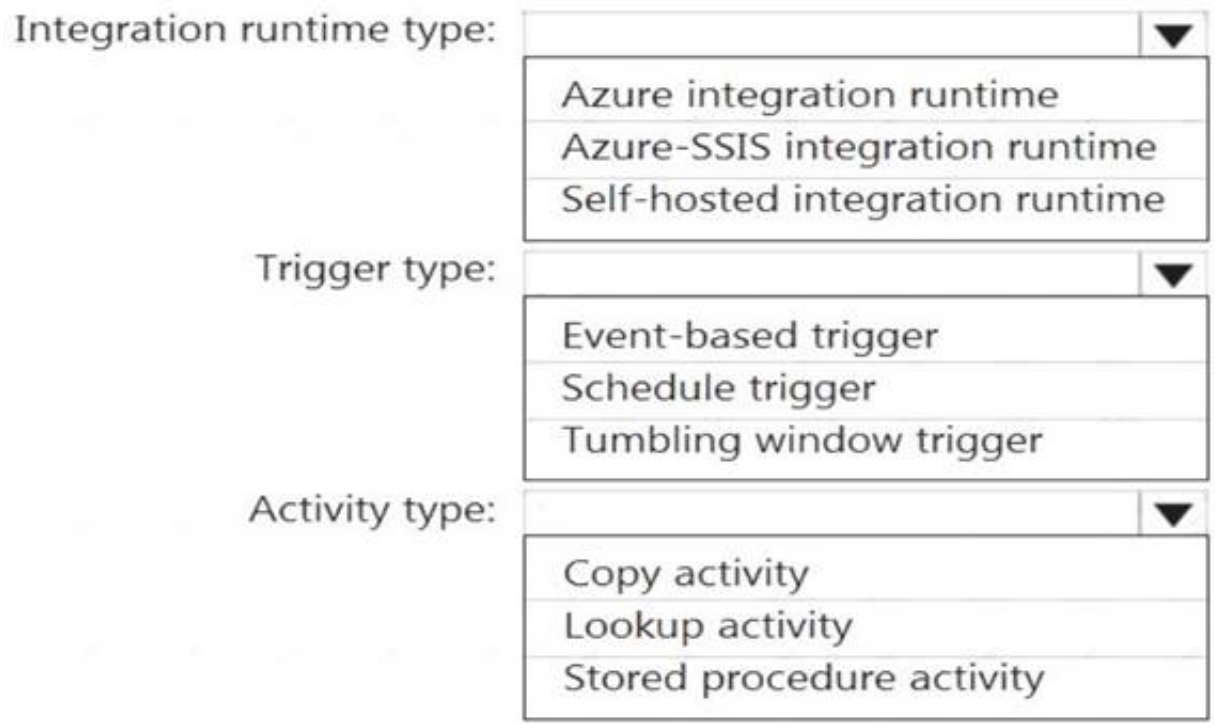
NEW QUESTION 147
- (Exam Topic 2)
What should you recommend to prevent users outside the Litware on-premises network from accessing the analytical data store?

A. a server-level virtual network rule
B. a database-level virtual network rule
C. a database-level firewall IP rule
D. a server-level firewall IP rule

Answer: A

Explanation:
Virtual network rules are one firewall security feature that controls whether the database server for your single databases and elastic pool in Azure SQL Database or for your databases in SQL Data Warehouse accepts communications that are sent from particular subnets in virtual networks.
Server-level, not database-level: Each virtual network rule applies to your whole Azure SQL Database server, not just to one particular database on the server. In other words, virtual network rule applies at the serverlevel, not at the database-level.
References:
https://docs.microsoft.com/en-us/azure/sql-database/sql-database-vnet-service-endpoint-rule-overview

NEW QUESTION 151
- (Exam Topic 2)
Which Azure Data Factory components should you recommend using together to import the daily inventory data from the SQL server to Azure Data Lake Storage?
To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



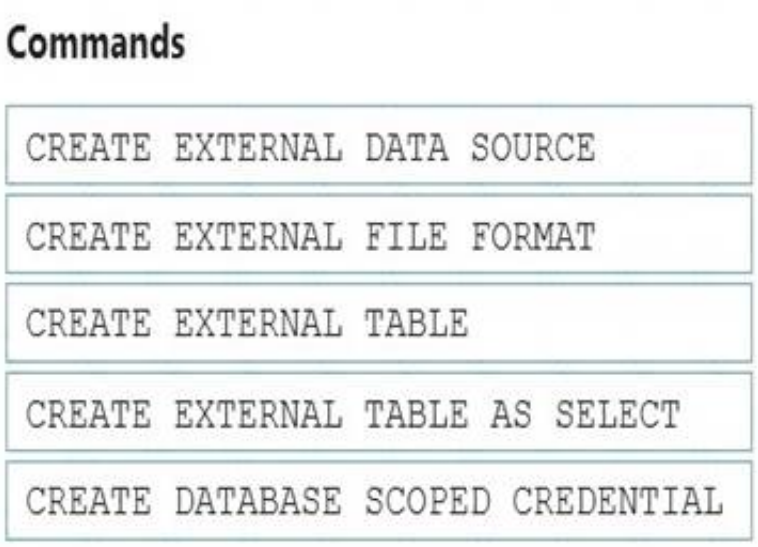
A. Mastered
B. Not Mastered

Answer: A

Explanation:
Box 1: Self-hosted integration runtime
A self-hosted IR is capable of running copy activity between a cloud data stores and a data store in private network.
Box 2: Schedule trigger Schedule every 8 hours Box 3: Copy activity Scenario:

- Customer data, including name, contact information, and loyalty number, comes from Salesforce and can be imported into Azure once every eight hours. Row modified dates are not trusted in the source table.
- Product data, including product ID, name, and category, comes from Salesforce and can be imported into Azure once every eight hours. Row modified dates are not trusted in the source table.

NEW QUESTION 153
- (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.

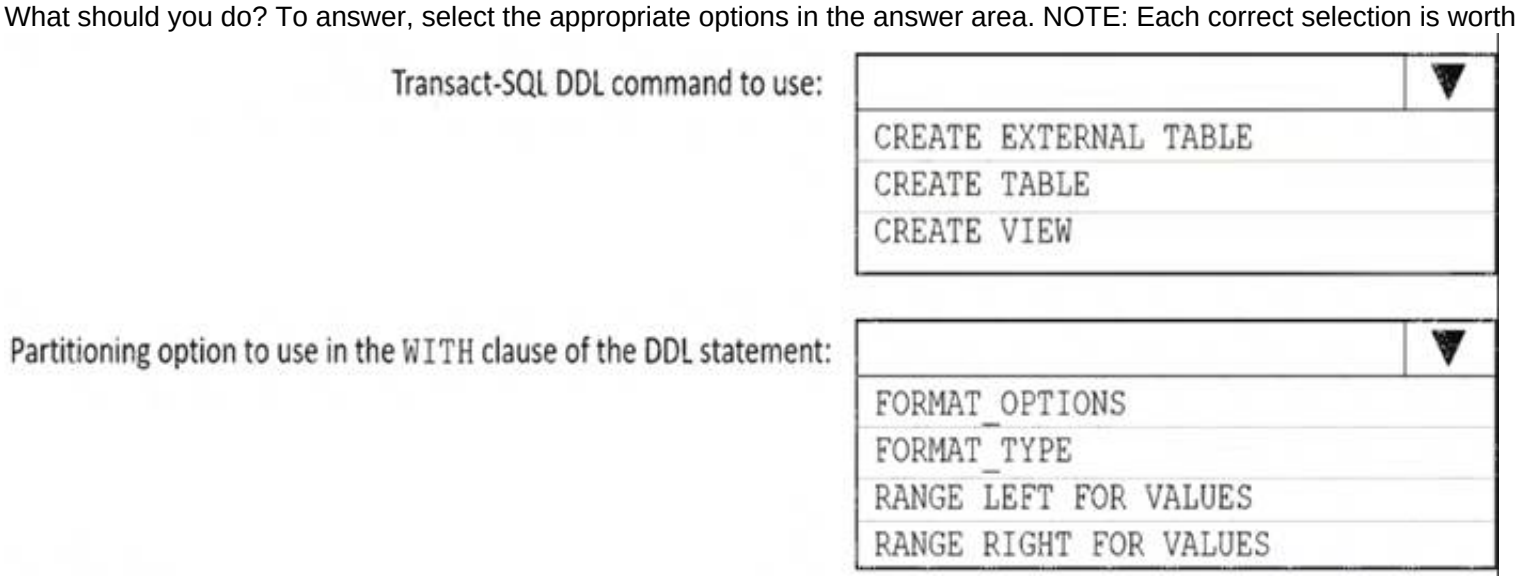


A. Mastered
B. Not Mastered

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 157
- (Exam Topic 1)
You need to implement an Azure Synapse Analytics database object for storing the sales transactions data. The solution must meet the sales transaction dataset requirements.
What solution must meet the sales transaction dataset requirements.
What should you do? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.



A. Mastered
B. Not Mastered

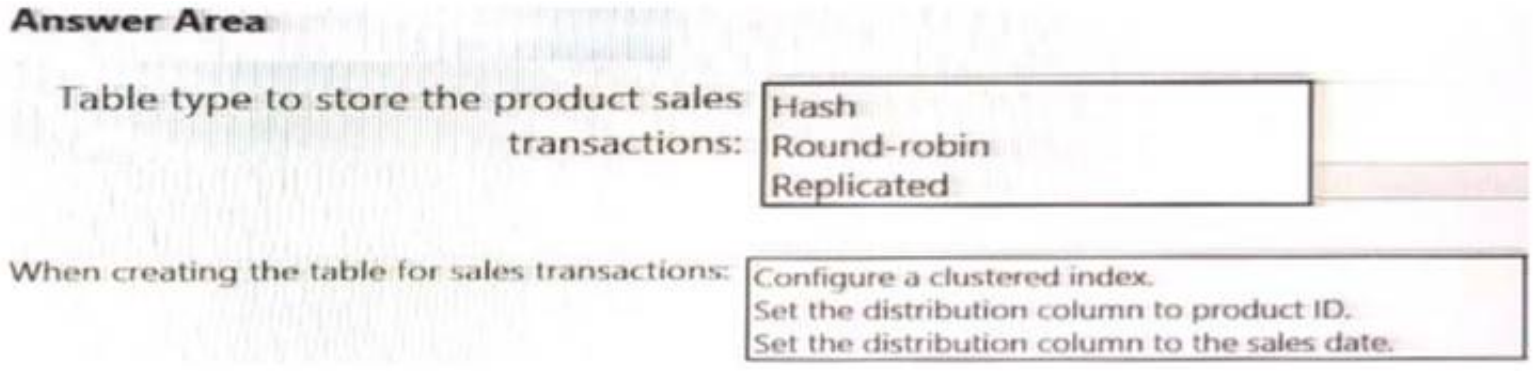
Answer: A

Explanation:
Graphical user interface, text, application, table Description automatically generated
Box 1: Create table
Scenario: Load the sales transaction dataset to Azure Synapse Analytics Box 2: RANGE RIGHT FOR VALUES
Scenario: 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.
RANGE RIGHT: Specifies the boundary value belongs to the partition on the right (higher values). FOR VALUES (boundary_value [,...n]): Specifies the boundary values for the partition.
Scenario: Load the sales transaction dataset to Azure Synapse Analytics. Contoso identifies the following requirements for the sales transaction dataset:

- 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.
- Ensure that queries joining and filtering sales transaction records based on product ID complete as quickly as possible.
- Implement a surrogate key to account for changes to the retail store addresses.
- Ensure that data storage costs and performance are predictable.
- Minimize how long it takes to remove old records. Reference:

https://docs.microsoft.com/en-us/sql/t-sql/statements/create-table-azure-sql-data-warehouse

NEW QUESTION 158
- (Exam Topic 1)
You need to design a data storage structure for the product sales transactions. The solution must meet the sales transaction dataset requirements.
What should you include in the solution? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.



A. Mastered
B. Not Mastered

Answer: A

Explanation:
Graphical user interface, text, application, chat or text message Description automatically generated
Box 1: Hash Scenario:
Ensure that queries joining and filtering sales transaction records based on product ID complete as quickly as possible.
A hash distributed table can deliver the highest query performance for joins and aggregations on large tables. Box 2: Set the distribution column to the sales date.
Scenario: 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.
Reference:
https://rajanieshkaushikk.com/2020/09/09/how-to-choose-right-data-distribution-strategy-for-azure-synapse/

NEW QUESTION 159
- (Exam Topic 1)
You need to design an analytical storage solution for the transactional data. 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.

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Graphical user interface, text, application, table Description automatically generated
Box 1: Round-robin
Round-robin tables are useful for improving loading speed.
Scenario: Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month.
Box 2: Hash
Hash-distributed tables improve query performance on large fact tables. Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribu

NEW QUESTION 163
- (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

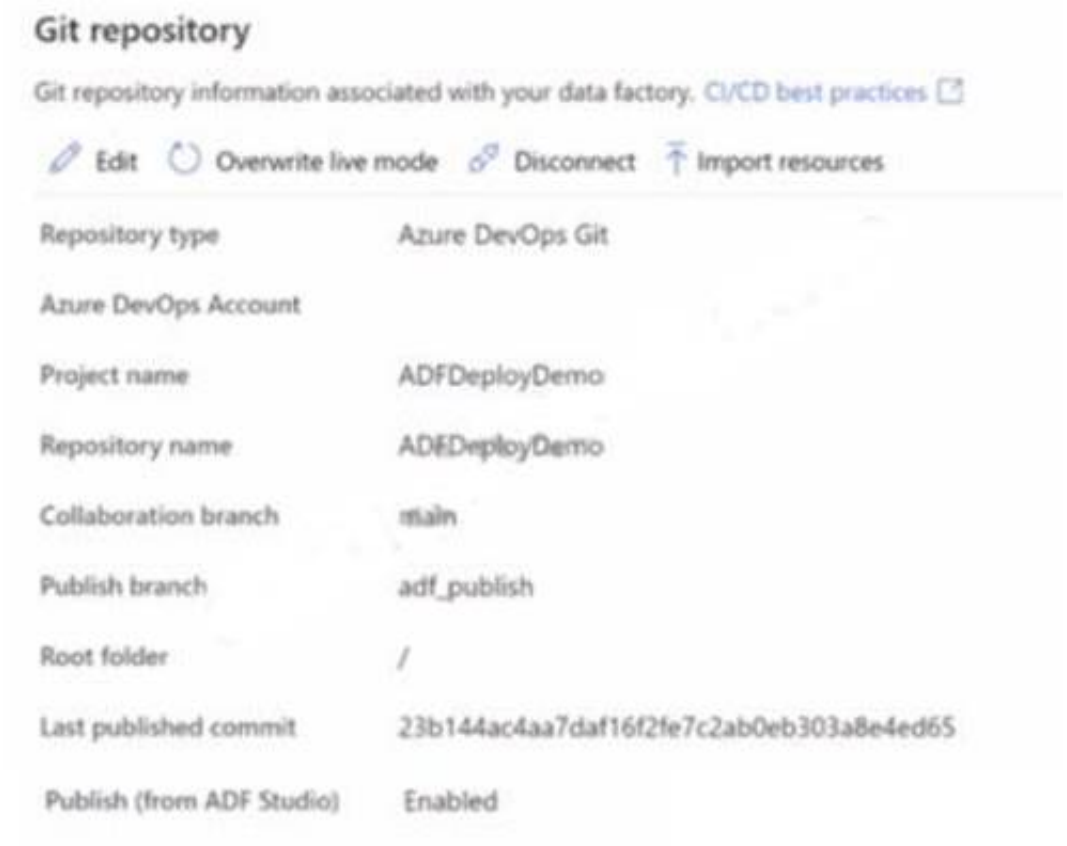
NEW QUESTION 166
- (Exam Topic 3)
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more
than one correct solution, while others might not have a correct 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 Storage account that contains 100 GB of files. The files contain rows of 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 copy the files to a table that has a columnstore index. Does this meet the goal?

A. Yes
B. No

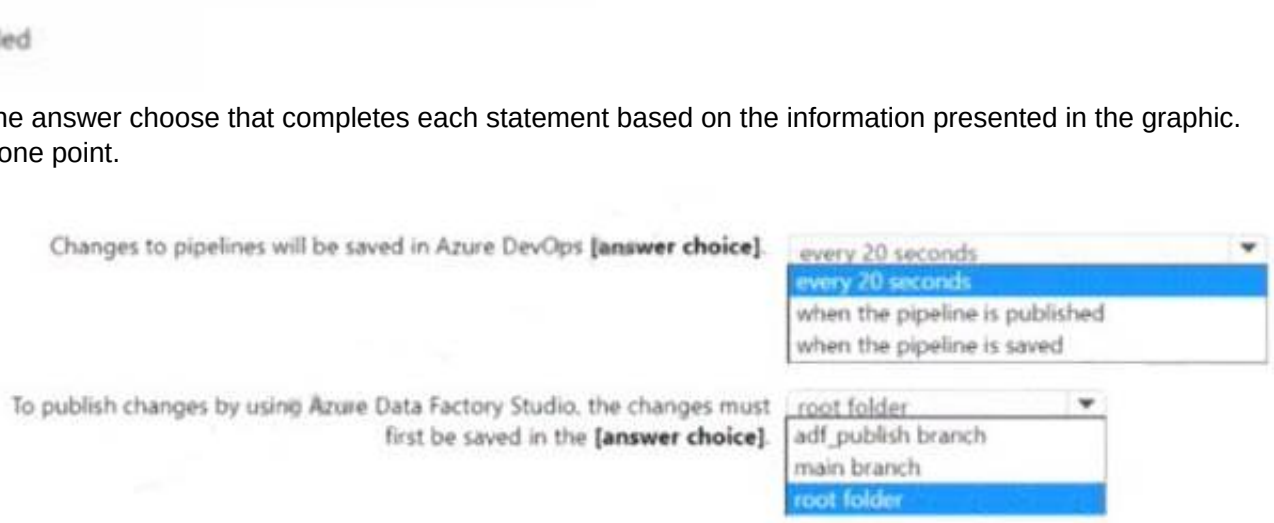
Answer: B

Explanation:
Instead convert the files to compressed delimited text files. Reference:
https://docs.microsoft.com/en-us/azure/sql-data-warehouse/guidance-for-loading-data

NEW QUESTION 169
- (Exam Topic 3)
You have an Azure data factory that has the Git repository settings shown in the following exhibit.



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

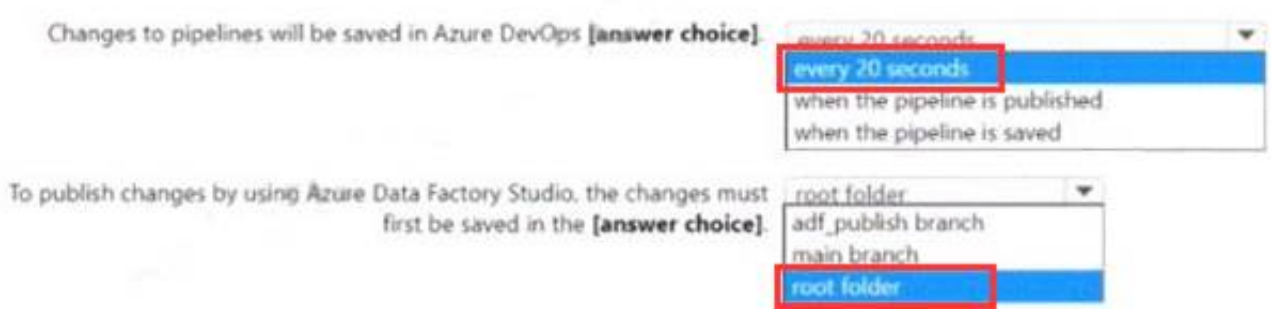


A. Mastered
B. Not Mastered

Answer: A

Explanation:

Answer Area



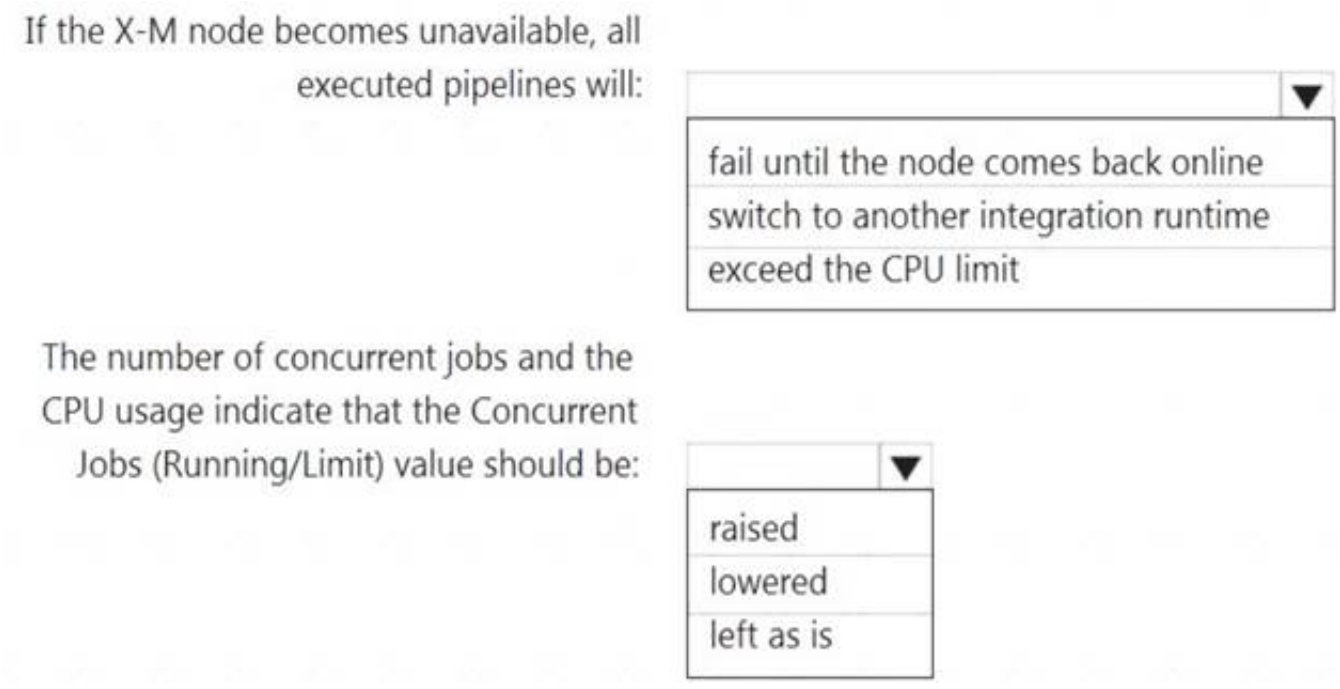
NEW QUESTION 174
- (Exam Topic 3)
You have a self-hosted integration runtime in Azure Data Factory.
The current status of the integration runtime has the following configurations:

- Status: Running
- Type: Self-Hosted
- Version: 4.4.7292.1
- Running / Registered Node(s): 1/1
- High Availability Enabled: False
- Linked Count: 0
- Queue Length: 0
- Average Queue Duration. 0.00s

The integration runtime has the following node details:

- Name: X-M
- Status: Running
- Version: 4.4.7292.1
- Available Memory: 7697MB
- CPU Utilization: 6%
- Network (In/Out): 1.21KBps/0.83KBps
- Concurrent Jobs (Running/Limit): 2/14
- Role: Dispatcher/Worker
- Credential Status: In Sync

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



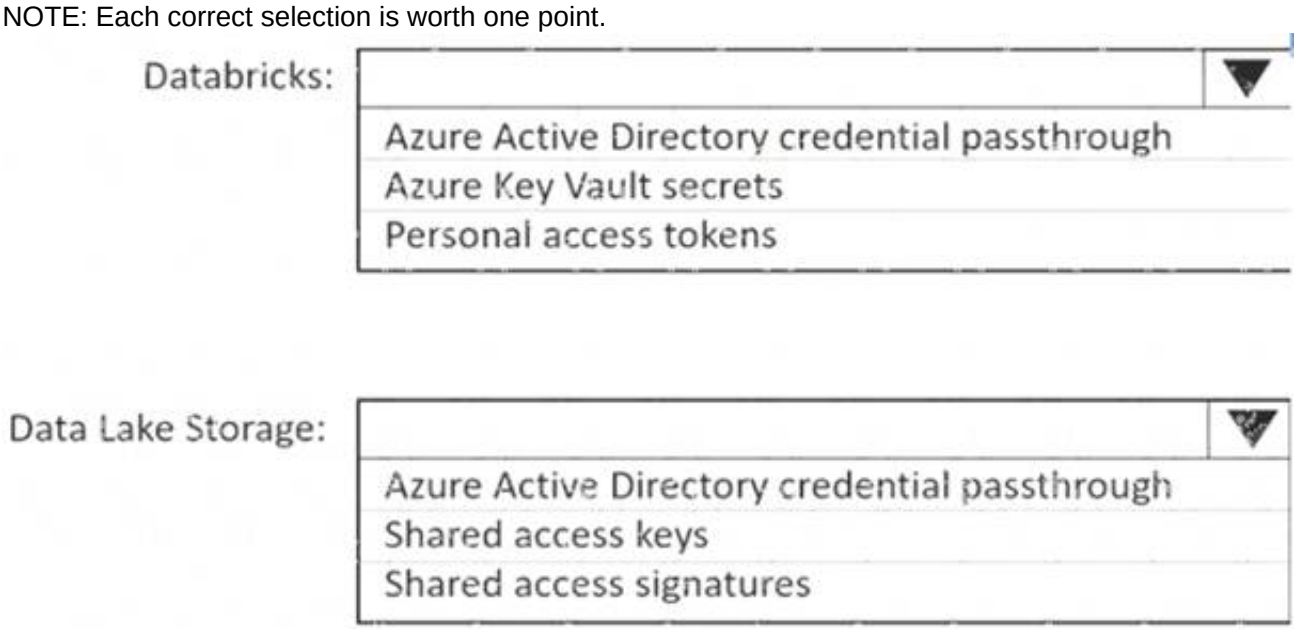
A. Mastered
B. Not Mastered

Answer: A

Explanation:
Box 1: fail until the node comes back online We see: High Availability Enabled: False
Note: Higher availability of the self-hosted integration runtime so that it's no longer the single point of failure in your big data solution or cloud data integration with Data Factory.
Box 2: lowered We see:
Concurrent Jobs (Running/Limit): 2/14 CPU Utilization: 6%
Note: When the processor and available RAM aren't well utilized, but the execution of concurrent jobs reaches a node's limits, scale up by increasing the number of concurrent jobs that a node can run
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/create-self-hosted-integration-runtime

NEW QUESTION 176
- (Exam Topic 3)

You use Azure Data Lake Storage Gen2 to store data that data scientists and data engineers will query by using Azure Databricks interactive notebooks. Users will have access only to the Data Lake Storage folders that relate to the projects on which they work.
You need to recommend which authentication methods to use for Databricks and Data Lake Storage to provide the users with the appropriate access. The solution must minimize administrative effort and development effort.
Which authentication method should you recommend for each Azure service? 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:
Table Description automatically generated
Box 1: Personal access tokens
You can use storage shared access signatures (SAS) to access an Azure Data Lake Storage Gen2 storage account directly. With SAS, you can restrict access to a storage account using temporary tokens with fine-grained access control.
You can add multiple storage accounts and configure respective SAS token providers in the same Spark session.
Box 2: Azure Active Directory credential passthrough
You can authenticate automatically to Azure Data Lake Storage Gen1 (ADLS Gen1) and Azure Data Lake Storage Gen2 (ADLS Gen2) from Azure Databricks clusters using the same Azure Active Directory (Azure AD) identity that you use to log into Azure Databricks. When you enable your cluster for Azure Data Lake Storage credential passthrough, commands that you run on that cluster can read and write data in Azure Data Lake Storage without requiring you to configure service principal credentials for access to storage.
After configuring Azure Data Lake Storage credential passthrough and creating storage containers, you can access data directly in Azure Data Lake Storage Gen1 using an adl:// path and Azure Data Lake Storage Gen2 using an abfss:// path:
Reference:
https://docs.microsoft.com/en-us/azure/databricks/data/data-sources/azure/adls-gen2/azure-datalake-gen2-sas-ac https://docs.microsoft.com/en-us/azure/databricks/security/credential-passthrough/adls-passthrough

NEW QUESTION 179
- (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 184
- (Exam Topic 3)
You are designing a solution that will copy Parquet files stored in an Azure Blob storage account to an Azure Data Lake Storage Gen2 account.
The data will be loaded daily to the data lake and will use a folder structure of {Year}/{Month}/{Day}/. You need to design a daily Azure Data Factory data load to minimize the data transfer between the two accounts.
Which two configurations should you include in the design? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

A. Delete the files in the destination before loading new data.
B. Filter by the last modified date of the source files.

C. Delete the source files after they are copied.
D. Specify a file naming pattern for the destination.

Answer: BD

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

NEW QUESTION 185
- (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

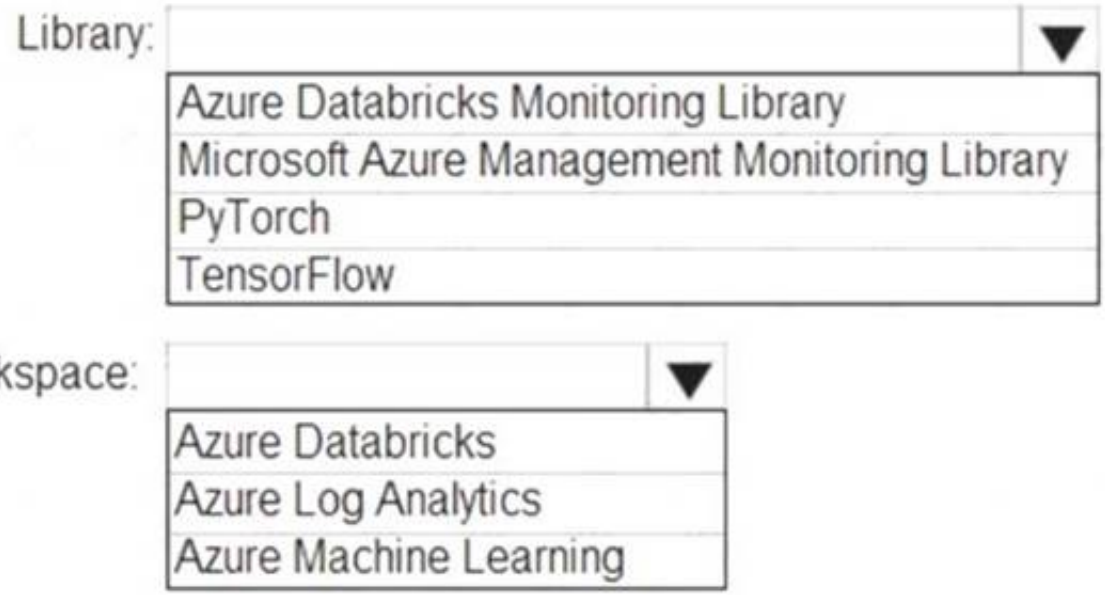
NEW QUESTION 188
- (Exam Topic 3)
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct 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 190
- (Exam Topic 3)
You need to collect application metrics, streaming query events, and application log messages for an Azure Databrick cluster.
Which type of library and workspace should you implement? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



A. Mastered
B. Not Mastered

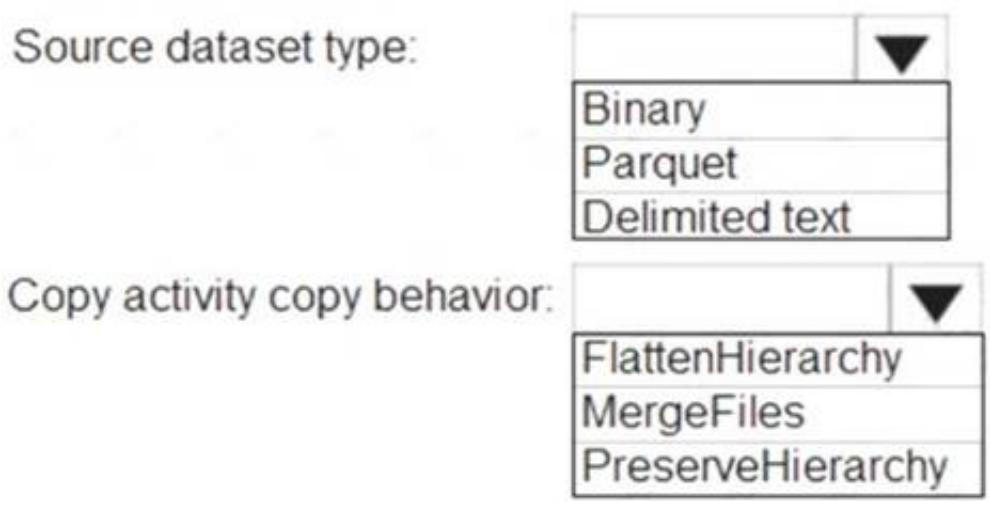
Answer: A

Explanation:
You can send application logs and metrics from Azure Databricks to a Log Analytics workspace. It uses the Azure Databricks Monitoring Library, which is available on GitHub.
References:
https://docs.microsoft.com/en-us/azure/architecture/databricks-monitoring/application-logs

NEW QUESTION 191
- (Exam Topic 3)
You have two Azure Storage accounts named Storage1 and Storage2. Each account holds one container and has the hierarchical namespace enabled. The system has files that contain data stored in the Apache Parquet format.
You need to copy folders and files from Storage1 to Storage2 by using a Data Factory copy activity. The solution must meet the following requirements:

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

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



A. Mastered
B. Not Mastered

Answer: A

Explanation:
Graphical user interface, text, application, chat or text message Description automatically generated
Box 1: Parquet
For Parquet datasets, the type property of the copy activity source must be set to ParquetSource. Box 2: PreserveHierarchy
PreserveHierarchy (default): Preserves the file hierarchy in the target folder. The relative path of the source file to the source folder is identical to the relative path of the target file to the target folder.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/format-parquet https://docs.microsoft.com/en-us/azure/data-factory/connector-azure-data-lake-storage

NEW QUESTION 195
- (Exam Topic 3)
You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Contacts. Contacts contains a column named Phone.
You need to ensure that users in a specific role only see the last four digits of a phone number when querying the Phone column.
What should you include in the solution?

A. a default value
B. dynamic data masking

C. row-level security (RLS)
D. column encryption
E. table partitions

Answer: B

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

NEW QUESTION 200
- (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 205
- (Exam Topic 3)
You are designing an Azure Databricks interactive cluster. The cluster will be used infrequently and will be configured for auto-termination.
You need to ensure that the cluster configuration is retained indefinitely after the cluster is terminated. The solution must minimize costs.
What should you do?

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

Answer: D

Explanation:
To keep an interactive cluster configuration even after it has been terminated for more than 30 days, an administrator can pin a cluster to the cluster list.
References:
https://docs.azuredatabricks.net/clusters/clusters-manage.html#automatic-termination

NEW QUESTION 206
- (Exam Topic 3)
You have an Azure Synapse Analytics dedicated SQL pool named pool1.
You need to perform a monthly audit of SQL statements that affect sensitive data. The solution must minimize administrative effort.
What should you include in the solution?

A. Microsoft Defender for SQL
B. dynamic data masking
C. sensitivity labels
D. workload management

Answer: B

NEW QUESTION 210
- (Exam Topic 3)
You are designing a slowly changing dimension (SCD) for supplier data in an Azure Synapse Analytics dedicated SQL pool.
You plan to keep a record of changes to the available fields. The supplier data contains the following columns.

Name	Description
SupplierSystemID	Unique supplier ID in an enterprise resource planning (ERP) system
SupplierName	Name of the supplier company
SupplierAddress1	Address of the supplier company
SupplierAddress2	Second address line of the supplier company
SupplierCity	City of the supplier company
SupplierStateProvince	State or province of the supplier company
SupplierCountry	Country of the supplier company
SupplierPostalCode	Postal code of the supplier company
SupplierDescription	Free-text description of the supplier company
SupplierCategory	Category of goods provided by the supplier company

Which three additional columns should you add to the data to create a Type 2 SCD? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.

A. surrogate primary key
B. foreign key
C. effective start date
D. effective end date
E. last modified date
F. business key

Answer: CDF

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

NEW QUESTION 215
- (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 216
- (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 220
- (Exam Topic 3)
You are designing a statistical analysis solution that will use custom proprietary1 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 225
- (Exam Topic 3)
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You are designing an Azure Stream Analytics solution that will analyze Twitter data.
You need to count the tweets in each 10-second window. The solution must ensure that each tweet is counted only once.
Solution: You use a session window that uses a timeout size of 10 seconds. Does this meet the goal?

A. Yes
B. No

Answer: A

Explanation:
Instead use a tumbling window. Tumbling windows are a series of fixed-sized, non-overlapping and contiguous time intervals. Reference:
https://docs.microsoft.com/en-us/stream-analytics-query/tumbling-window-azure-stream-analytics

NEW QUESTION 230
- (Exam Topic 3)
You plan to implement an Azure Data Lake Gen2 storage account.
You need to ensure that the data lake will remain available if a data center fails in the primary Azure region. The solution must minimize costs.
Which type of replication should you use for the storage account?

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

Answer: C

Explanation:
Locally redundant storage (LRS) copies your data synchronously three times within a single physical location in the primary region. LRS is the least expensive replication option
Reference:
https://docs.microsoft.com/en-us/azure/storage/common/storage-redundancy

NEW QUESTION 232
- (Exam Topic 3)
You have an Azure Databricks workspace and an Azure Data Lake Storage Gen2 account named storage! New files are uploaded daily to storage1.
• Incrementally process new files as they are upkorage1 as a structured streaming source. The solution must meet the following requirements:
• Minimize implementation and maintenance effort.
• Minimize the cost of processing millions of files.
• Support schema inference and schema drift. Which should you include in the recommendation?

A. Auto Loader
B. Apache Spark FileStreamSource
C. COPY INTO
D. Azure Data Factory

Answer: D

NEW QUESTION 234
- (Exam Topic 3)
You have a partitioned table in an Azure Synapse Analytics dedicated SQL pool. You need to design queries to maximize the benefits of partition elimination. What

should you include in the Transact-SQL queries?

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

Answer: B

NEW QUESTION 235
- (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

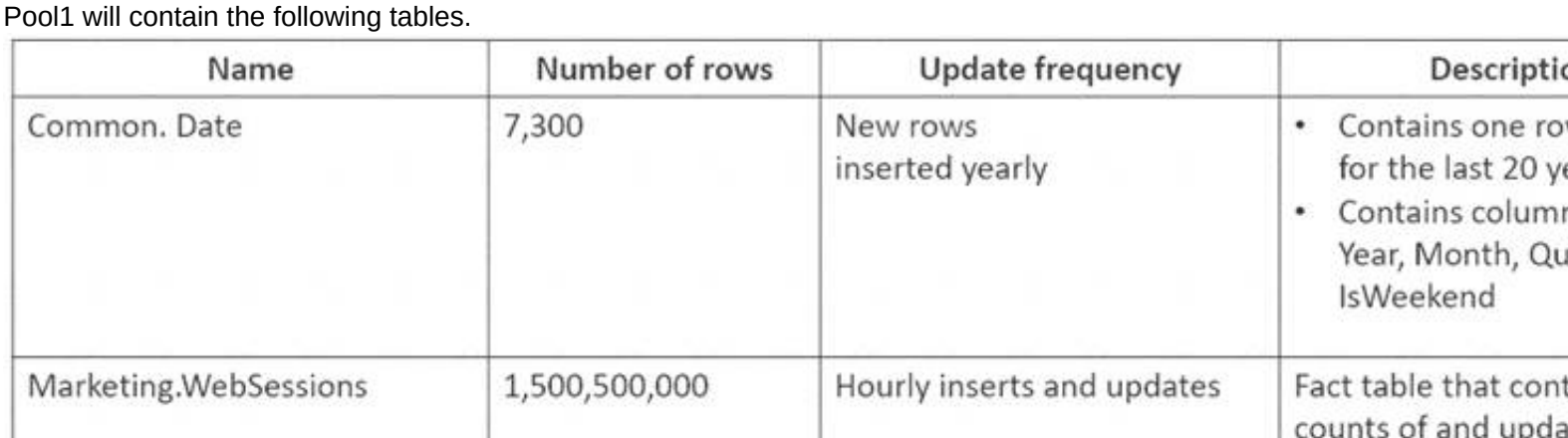
Explanation:
You can't change the partition count for an event hub after its creation except for the event hub in a dedicated cluster.
Reference:
https://docs.microsoft.com/en-us/azure/event-hubs/event-hubs-features

NEW QUESTION 236
- (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	• 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.



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 239
- (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 intotable1.
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 DBREINOEX (table1)
D. DBCC INDEXDEFRAG (pool1,tablel)

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 241
- (Exam Topic 3)
You are developing an application that uses Azure Data Lake Storage Gen 2.
You need to recommend a solution to grant permissions to a specific application for a limited time period. What should you include in the recommendation?

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

Answer: B

Explanation:
A shared access signature (SAS) provides secure delegated access to resources in your storage account. With a SAS, you have granular control over how a client can access your data. For example:
What resources the client may access.
What permissions they have to those resources. How long the SAS is valid.
Reference:
https://docs.microsoft.com/en-us/azure/storage/common/storage-sas-overview

NEW QUESTION 245
- (Exam Topic 3)
You use PySpark in Azure Databricks to parse the following JSON input.

```
{
    "persons":[
        {
            "name":"Keith",
            "age":30,
            "dogs":["Fido", "Fluffy"]

        },
        {
            "name":"Donna",
            "age":46,
            "dogs":["Spot"]


        }
    ]
}
```

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

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

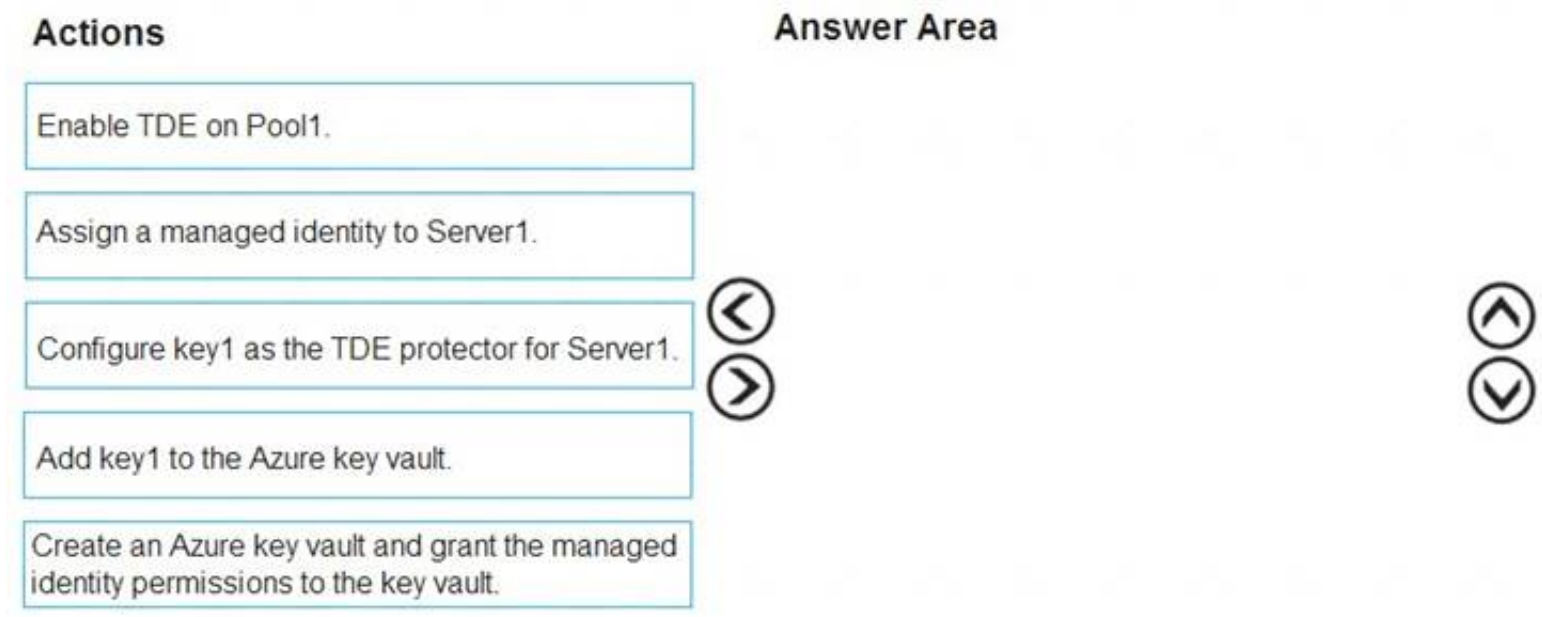
How should you complete the PySpark code? To answer, drag the appropriate values to he correct targets. Each value may be used once, more than once or not at all. You may need to drag the split bar between panes or scroll to view content.
NOTE: Each correct selection is worth one point.

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Graphical user interface, text, application Description automatically generated
Box 1: select
Box 2: explode
Bop 3: alias
pyspark.sql.Column.alias returns this column aliased with a new name or names (in the case of expressions that return more than one column, such as explode).
Reference: https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.sql.Column.alias.html https://docs.microsoft.com/en-us/azure/databricks/sql/language-manual/functions/explode

NEW QUESTION 248
- (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.



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 250
- (Exam Topic 3)
You have an Azure Storage account and a data warehouse in Azure Synapse Analytics in the UK South region. You need to copy blob data from the storage account to the data warehouse by using Azure Data Factory. The solution must meet the following requirements:
- Ensure that the data remains in the UK South region at all times.
- Minimize administrative effort.

Which type of integration runtime should you use?

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

Answer: A

Explanation:

IR type	Public network	Private network
Azure	Data Flow Data movement Activity dispatch	
Self-hosted	Data movement Activity dispatch	Data movement Activity dispatch
Azure-SSIS	SSIS package execution	SSIS package execution

Reference:
https://docs.microsoft.com/en-us/azure/data-factory/concepts-integration-runtime

NEW QUESTION 254
- (Exam Topic 3)
You have an Azure Synapse Analytics dedicated SQL pod.
You need to create a pipeline that will execute a stored procedure in the dedicated SQL pool and use the returned result set as the input (or a downstream activity. The solution must minimize development effort.
Which Type of activity should you use in the pipeline?

A. Notebook
B. U-SQL
C. Script
D. Stored Procedure

Answer: D

NEW QUESTION 259
- (Exam Topic 3)
You are planning a solution to aggregate streaming data that originates in Apache Kafka and is output to Azure Data Lake Storage Gen2. The developers who will implement the stream processing solution use Java, Which service should you recommend using to process the streaming data?

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

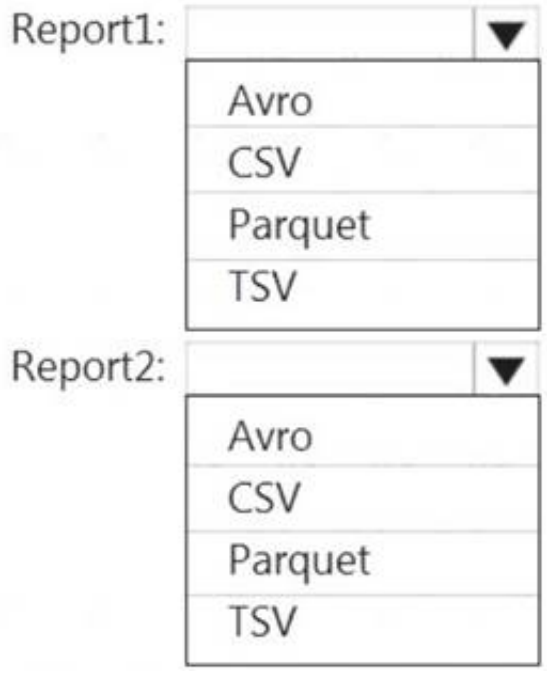
Answer: C

Explanation:
https://docs.microsoft.com/en-us/azure/architecture/data-guide/technology-choices/stream-processing

NEW QUESTION 261
- (Exam Topic 3)
You are planning the deployment of Azure Data Lake Storage Gen2. You have the following two reports that will access the data lake:

- Report1: Reads three columns from a file that contains 50 columns.
- Report2: Queries a single record based on a timestamp.

You need to recommend in which format to store the data in the data lake to support the reports. The solution must minimize read times.
What should you recommend for each report? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

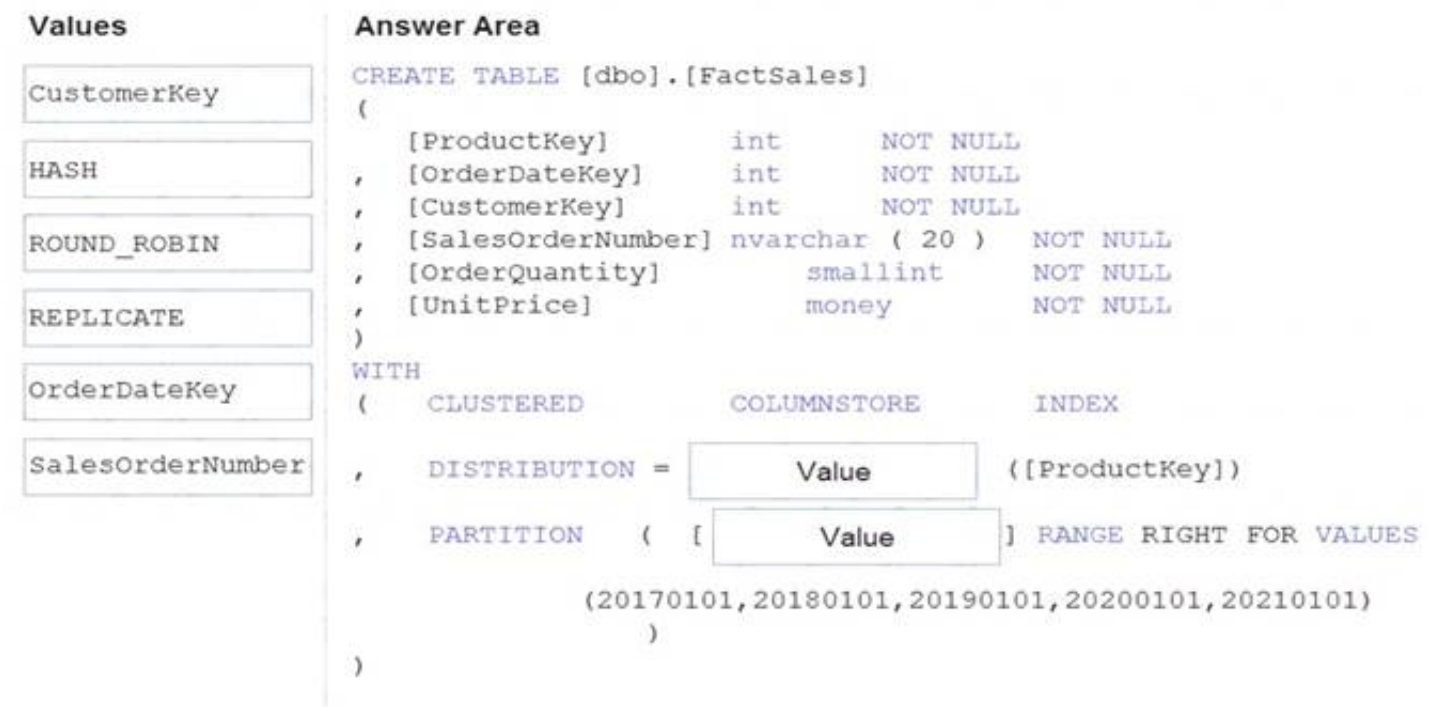


A. Mastered
B. Not Mastered

Answer: A

Explanation:
Report1: CSV
CSV: The destination writes records as delimited data. Report2: AVRO
AVRO supports timestamps.
Not Parquet, TSV: Not options for Azure Data Lake Storage Gen2. Reference:
https://streamsets.com/documentation/datacollector/latest/help/datacollector/UserGuide/Destinations/ADLS-G2

NEW QUESTION 265
- (Exam Topic 3)
You plan to create a table in an Azure Synapse Analytics dedicated SQL pool.
Data in the table will be retained for five years. Once a year, data that is older than five years will be deleted. You need to ensure that the data is distributed evenly across partitions. The solution must minimize the
amount of time required to delete old data.
How should you complete the Transact-SQL statement? To answer, drag the appropriate values to the correct targets. Each value may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.
NOTE: Each correct selection is worth one point.



A. Mastered
B. Not Mastered

Answer: A

Explanation:
Box 1: HASH
Box 2: OrderDateKey
In most cases, table partitions are created on a date column.
A way to eliminate rollbacks is to use Metadata Only operations like partition switching for data management. For example, rather than execute a DELETE statement to delete all rows in a table where the order_date was in October of 2001, you could partition your data early. Then you can switch out the partition with data for an empty partition from another table.
Reference:
https://docs.microsoft.com/en-us/sql/t-sql/statements/create-table-azure-sql-data-warehouse https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/best-practices-dedicated-sql-pool

NEW QUESTION 269
- (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 272
.....

Relate Links

100% Pass Your DP-203 Exam with Exambible Prep Materials

https://www.exambible.com/DP-203-exam/

Contact us

We are proud of our high-quality customer service, which serves you around the clock 24/7.

Viste - https://www.exambible.com/