Microsoft

Exam Questions DP-203

Data Engineering on Microsoft Azure

NEW QUESTION 1
- (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

Table distribution types

Hash
Replicated
Round-robin

Answer Area

Common.Data: [   ]

Marketing.Web.Sessions: [   ]

Staging. Web.Sessions: [   ]

A. Mastered
B. Not Mastered

Answer: A

Explanation:

Table distribution types

Hash
Replicated
Round-robin

Answer Area

Common.Data: [Replicated]

Marketing.Web.Sessions: [Round-robin]

Staging. Web.Sessions: [Hash]

NEW QUESTION 2
- (Exam Topic 3)
You are implementing a batch dataset in the Parquet format.
Data tiles will be produced by using Azure Data Factory and stored in Azure Data Lake Storage Gen2. The files will be consumed by an Azure Synapse Analytics serverless SQL pool.
You need to minimize storage costs for the solution. What should you do?

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

Answer: C

Explanation:
An external table points to data located in Hadoop, Azure Storage blob, or Azure Data Lake Storage. External tables are used to read data from files or write data to files in Azure Storage. With Synapse SQL, you can use external tables to read external data using dedicated SQL pool or serverless SQL pool.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables

NEW QUESTION 3
- (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 4
- (Exam Topic 3)
You plan to create an Azure Data Factory pipeline that will include a mapping data flow. You have JSON data containing objects that have nested arrays.
You need to transform the JSON-formatted data into a tabular dataset. The dataset must have one tow for each item in the arrays.
Which transformation method should you use in the mapping data flow?

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

Answer: B

Explanation:
Use the flatten transformation to take array values inside hierarchical structures such as JSON and unroll them into individual rows. This process is known as denormalization.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/data-flow-flatten

NEW QUESTION 5
- (Exam Topic 3)
You manage an enterprise data warehouse in Azure Synapse Analytics.
Users report slow performance when they run commonly used queries. Users do not report performance changes for infrequently used queries.
You need to monitor resource utilization to determine the source of the performance issues. Which metric should you monitor?

A. Data IO percentage
B. Local tempdb percentage
C. Cache used percentage
D. DWU percentage

Answer: C

Explanation:
Monitor and troubleshoot slow query performance by determining whether your workload is optimally leveraging the adaptive cache for dedicated SQL pools.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-how-to-monit

NEW QUESTION 6
- (Exam Topic 3)
You have an Azure Data Lake Storage account that contains a staging zone.
You need to design a dairy 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 copies the data to a staging table in the data warehouse, and then uses a stored procedure to execute the R script.
Does this meet the goal?

A. Yes
B. No

Answer: A

Explanation:
If you need to transform data in a way that is not supported by Data Factory, you can create a custom activity with your own data processing logic and use the activity in the pipeline.
Note: You can use data transformation activities in Azure Data Factory and Synapse pipelines to transform and process your raw data into predictions and insights at scale.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/transform-data

NEW QUESTION 7
- (Exam Topic 3)
You have an Azure subscription that contains an Azure Data Lake Storage Gen2 account named storage1. Storage1 contains a container named container1.
Container1 contains a directory named directory1. Directory1 contains a file named file1.
You have an Azure Active Directory (Azure AD) user named User1 that is assigned the Storage Blob Data Reader role for storage1.
You need to ensure that User1 can append data to file1. The solution must use the principle of least privilege. Which permissions should you grant? To answer, drag the appropriate permissions to the correct resources.
Each permission may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Box 1: Execute
If you are granting permissions by using only ACLs (no Azure RBAC), then to grant a security principal read or write access to a file, you'll need to give the security principal Execute permissions to the root folder of the container, and to each folder in the hierarchy of folders that lead to the file.
Box 2: Execute
On Directory: Execute (X): Required to traverse the child items of a directory Box 3: Write
On file: Write (W): Can write or append to a file. Reference:
https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control

NEW QUESTION 8
- (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 9
- (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 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 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 12
- (Exam Topic 3)
You have an Azure subscription that contains a logical Microsoft SQL server named Server1. Server1 hosts an Azure Synapse Analytics SQL dedicated pool named Pool1.
You need to recommend a Transparent Data Encryption (TDE) solution for Server1. The solution must meet the following requirements:

≫ Track the usage of encryption keys.

≫ Maintain the access of client apps to Pool1 in the event of an Azure datacenter outage that affects the availability of the encryption keys.
What should you include in the recommendation? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

To track encryption key usage:

Always Encrypted
TDE with customer-managed keys
TDE with platform-managed keys

To maintain client app access in the event of a datacenter outage:

Create and configure Azure key vaults in two Azure regions.
Enable Advanced Data Security on Server1.
Implement the client apps by using a Microsoft .NET Framework data provider.

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Box 1: TDE with customer-managed keys
Customer-managed keys are stored in the Azure Key Vault. You can monitor how and when your key vaults are accessed, and by whom. You can do this by enabling logging for Azure Key Vault, which saves information in an Azure storage account that you provide.
Box 2: Create and configure Azure key vaults in two Azure regions
The contents of your key vault are replicated within the region and to a secondary region at least 150 miles away, but within the same geography to maintain high durability of your keys and secrets.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/security/workspaces-encryption https://docs.microsoft.com/en-us/azure/key-vault/general/logging

NEW QUESTION 17
- (Exam Topic 3)
You are designing an anomaly detection solution for streaming data from an Azure IoT hub. The solution must meet the following requirements:
≫ Send the output to Azure Synapse.
≫ Identify spikes and dips in time series data.
≫ Minimize development and configuration effort. Which should you include in the solution?

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

Answer: B

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

NEW QUESTION 20
- (Exam Topic 3)
You use Azure Data Factory to prepare data to be queried by Azure Synapse Analytics serverless SQL pools. Files are initially ingested into an Azure Data Lake Storage Gen2 account as 10 small JSON files. Each file contains the same data attributes and data from a subsidiary of your company.
You need to move the files to a different folder and transform the data to meet the following requirements: ≫ Provide the fastest possible query times.
≫ Automatically infer the schema from the underlying files.
How should you configure the Data Factory copy activity? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

Copy behavior:

Flatten hierarchy
Merge files
Preserve hierarchy

Sink file type:

CSV
JSON
Parquet
TXT

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Box 1: Preserver 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 22
- (Exam Topic 3)
You have an Azure Synapse Analytics dedicated SQL pool named pool1.
You plan to implement a star schema in pool1 and create a new table named DimCustomer by using the following code.

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

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

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

Answer: AB

NEW QUESTION 25
- (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.

Distribution:
- Hash
- Replicated
- Round-robin

Indexing:
- Clustered
- Clustered columnstore
- Heap

Partitioning:
- Date
- None

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 Data Factory pipeline that contains a data flow. The data flow contains the following expression.

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

A. Mastered
B. Not Mastered

Answer: A

Explanation:
See below answer.
Answer Area

Number of columns: 22

Number of rows: 4

NEW QUESTION 32
- (Exam Topic 3)
You are designing an Azure Stream Analytics solution that receives instant messaging data from an Azure Event Hub.
You need to ensure that the output from the Stream Analytics job counts the number of messages per time zone every 15 seconds.
How should you complete the Stream Analytics query? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

```
Select TimeZone, count (*) AS MessageCount
```

FROM MessageStream [▼] CreatedAt

```
LAST
OVER
SYSTEM.TIMESTAMP()
TIMESTAMP BY
```

GROUP BY TimeZone, [▼] (second,15)

```
HOPPINGWINDOW
SESSIONWINDOW
SLIDINGWINDOW
TUMBLINGWINDOW
```

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Table Description automatically generated
Box 1: timestamp by
Box 2: TUMBLINGWINDOW
Tumbling window functions are used to segment a data stream into distinct time segments and perform a function against them, such as the example below. The key differentiators of a Tumbling window are that they repeat, do not overlap, and an event cannot belong to more than one tumbling window.
Timeline Description automatically generated
Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-window-functions

NEW QUESTION 37
- (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 38
- (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 40
- (Exam Topic 3)
You have an activity in an Azure Data Factory pipeline. The activity calls a stored procedure in a data warehouse in Azure Synapse Analytics and runs daily.
You need to verify the duration of the activity when it ran last. What should you use?

A. activity runs in Azure Monitor
B. Activity log in Azure Synapse Analytics
C. the sys.dm_pdw_wait_stats data management view in Azure Synapse Analytics
D. an Azure Resource Manager template

Answer: A

Explanation:
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/monitor-visually

NEW QUESTION 42
- (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_YYYMMDD.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 44
- (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.

Create the table in:	▼
The dedicated SQL pool	
The serverless SQL pool	
The Spark pool	

File format:	▼
Apache Parquet	
Delta	
JSON	

A. Mastered
B. Not Mastered

Answer: A

Explanation:
The dedicated SQL pool Apache Parquet

NEW QUESTION 46
- (Exam Topic 3)
You are building an Azure Data Factory solution to process data received from Azure Event Hubs, and then ingested into an Azure Data Lake Storage Gen2 container.
The data will be ingested every five minutes from devices into JSON files. The files have the following naming pattern.
/{deviceType}/in/{YYYY}/{MM}/{DD}/{HH}/{deviceID}_{YYYY}{MM}{DD}HH}{mm}.json
You need to prepare the data for batch data processing so that there is one dataset per hour per deviceType. The solution must minimize read times.
How should you configure the sink for the copy activity? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

Parameter:	▼
@pipeline(),TriggerTime	
@pipeline(),TriggerType	
@trigger().outputs.windowStartTime	
@trigger().startTime	

Naming pattern:	▼
/{deviceID}/out/{YYYY}/{MM}/{DD}/{HH}.json	
/{YYYY}/{MM}/{DD}/{deviceType}.json	
/{YYYY}/{MM}/{DD}/{HH}.json	
/{YYYY}/{MM}/{DD}/{HH}_{deviceType}.json	

Copy behavior:	▼
Add dynamic content	
Flatten hierarchy	
Merge files	

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Box 1: @trigger().startTime
startTime: A date-time value. For basic schedules, the value of the startTime property applies to the first occurrence. For complex schedules, the trigger starts no

sooner than the specified startTime value.

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

Box 3: Flatten hierarchy

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

Reference:

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

NEW QUESTION 49

- (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
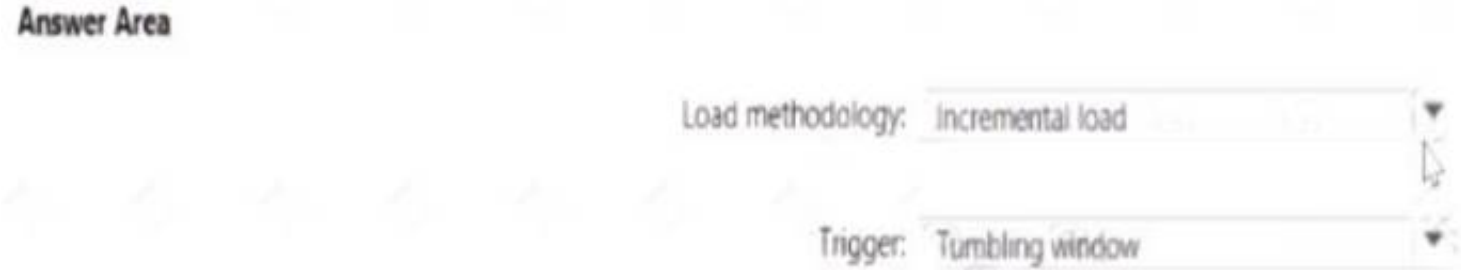
NEW QUESTION 54

- (Exam Topic 3)

You have an Azure Storage account that generates 200.000 new files daily. The file names have a format of (YYY)/(MM)/(DD)/|HH])/(CustornerID).csv.

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

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

A. Mastered
B. Not Mastered

Answer: A

Explanation:

Answer Area

Load methodology:	Incremental load ▼
Trigger:	Tumbling window ▼

NEW QUESTION 55

- (Exam Topic 3)

You are designing the folder structure for an Azure Data Lake Storage Gen2 account. You identify the following usage patterns:

• Users will query data by using Azure Synapse Analytics serverless SQL pools and Azure Synapse Analytics serverless Apache Spark pods.

• Most queries will include a filter on the current year or week.

• Data will be secured by data source.

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

• Supports the usage patterns

• Simplifies folder security

• Minimizes query times

Which folder structure should you recommend?

A)

\YYYYY\WW\DataSource\SubjectArea\FileData_YYYY_MM_DD.parquet

B)

DataSource\SubjectArea\WW\YYYY\FileData_YYYY_MM_DD.parquet

C)

\DataSource\SubjectArea\YYYY\WW\FileData_YYYY_MM_DD.parquet

D)

\DataSource\SubjectArea\YYYY-WW\FileData_YYYY_MM_DD.parquet

E)

```
WW\YYYY\SubjectArea\DataSource\FileData_YYYY_MM_DD.parquet
```

A. Option A
B. Option B
C. Option C
D. Option D
E. Option E

Answer: C

Explanation:
Data will be secured by data source. -> Use DataSource as top folder.
Most queries will include a filter on the current year or week -> Use \YYYY\WW\ as subfolders. Common Use Cases
A common use case is to filter data stored in a date (and possibly time) folder structure such as
/YYYY/MM/DD/ or /YYYY/MM/YYYY-MM-DD/. As new data is generated/sent/copied/moved to the storage account, a new folder is created for each specific time period. This strategy organises data into a maintainable folder structure.
Reference: https://www.serverlesssql.com/optimisation/azurestoragefilteringusingfilepath/

NEW QUESTION 59
- (Exam Topic 3)
You have a Microsoft Purview account. The Lineage view of a CSV file is shown in the following exhibit.
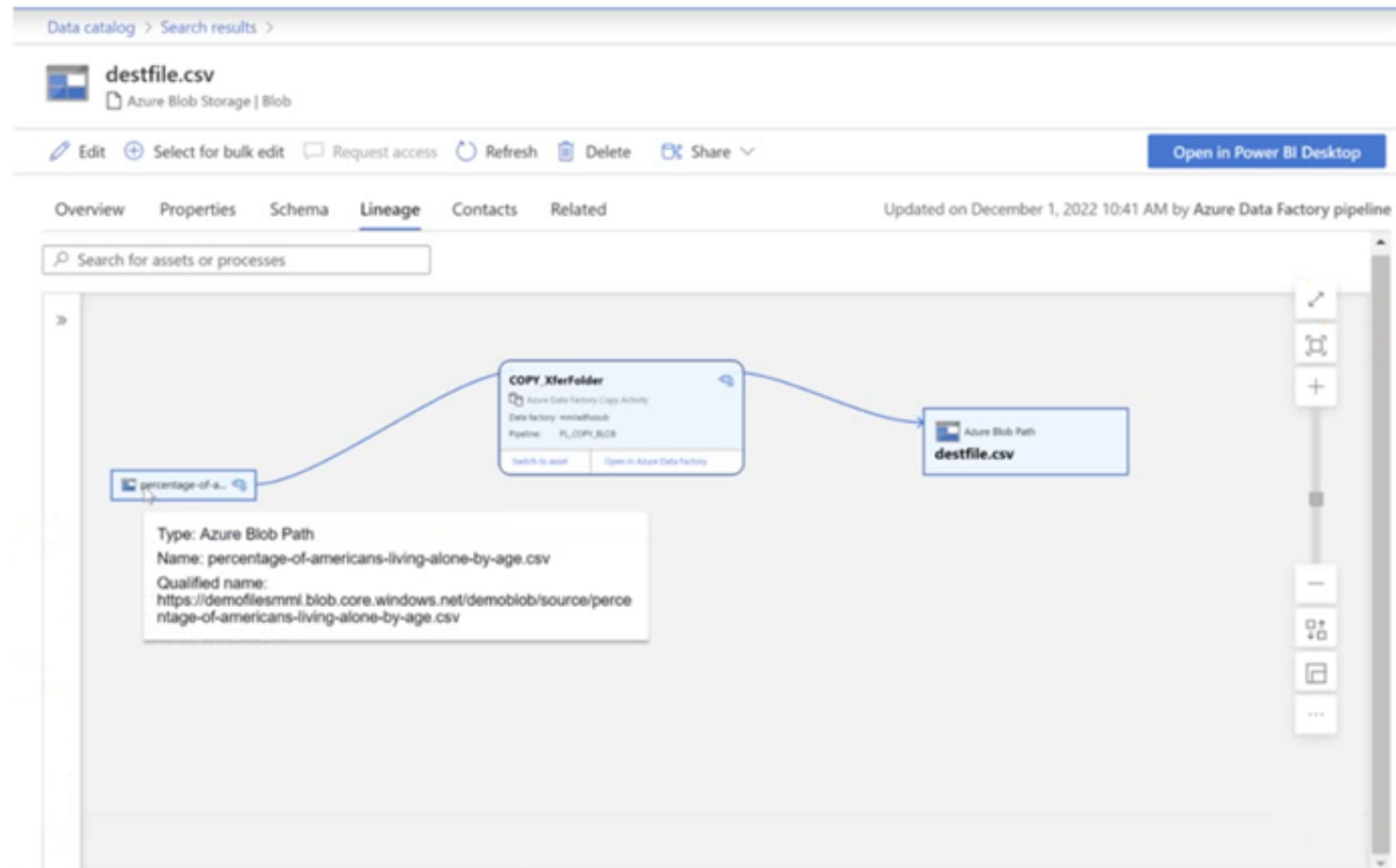
How is the data for the lineage populated?

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

Answer: B

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

NEW QUESTION 63
- (Exam Topic 3)
You are developing a solution using a Lambda architecture on Microsoft Azure. The data at test layer must meet the following requirements:
Data storage:
•Serve as a repository (or high volumes of large files in various formats.
•Implement optimized storage for big data analytics workloads.
•Ensure that data can be organized using a hierarchical structure. Batch processing:
•Use a managed solution for in-memory computation processing.
•Natively support Scala, Python, and R programming languages.
•Provide the ability to resize and terminate the cluster automatically. Analytical data store:
•Support parallel processing.
•Use columnar storage.
•Support SQL-based languages.
You need to identify the correct technologies to build the Lambda architecture.
Which technologies should you use? To answer, select the appropriate options in the answer area NOTE: Each correct selection is worth one point.

Architecture requirement   **Technology**

Data storage

Azure SQL Database
Azure Blob Storage
Azure Cosmos DB
Azure Data Lake Store

Batch processing

HDInsight Spark
HDInsight Hadoop
Azure Databricks
HDInsight Interactive Query

Analytical data store

HDInsight HBase
Azure SQL Data Warehouse
Azure Analysis Services
Azure Cosmos DB

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 65
- (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.	○	○
The stream scheme key and count must match the output scheme.	○	○
Providing 60 streaming units will optimize the performance of the query.	○	○

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 68
- (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 71
- (Exam Topic 3)
You have an enterprise data warehouse in Azure Synapse Analytics named DW1 on a server named Server1. You need to determine the size of the transaction log file for each distribution of DW1.
What should you do?

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

Answer: A

Explanation:
For information about the current log file size, its maximum size, and the autogrow option for the file, you can also use the size, max_size, and growth columns for that log file in sys.database_files.
Reference:
https://docs.microsoft.com/en-us/sql/relational-databases/logs/manage-the-size-of-the-transaction-log-file

NEW QUESTION 73
- (Exam Topic 3)
You have an Azure Databricks workspace named workspace! in the Standard pricing tier. Workspace1 contains an all-purpose cluster named cluster). You need to

reduce the time it takes for cluster 1 to start and scale up. The solution must minimize costs. What should you do first?

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

Answer: C

Explanation:
You can use Databricks Pools to Speed up your Data Pipelines and Scale Clusters Quickly.
Databricks Pools, a managed cache of virtual machine instances that enables clusters to start and scale 4 times faster.
Reference:
https://databricks.com/blog/2019/11/11/databricks-pools-speed-up-data-pipelines.html

NEW QUESTION 74
- (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

Architecture requirement	Technology
Ingest	Logic Apps / Azure Data Factory / Azure Automation
Store	Azure Data Lake Storage / Azure Blob storage / Azure files
Prepare and Train	HDInsight Apache Spark cluster / Azure Databricks / HDInsight Apache Storm cluster
Model and Serve	HDInsight Apache Kafka cluster / Azure Synapse Analytics / Azure Data Lake Storage

A. Mastered
B. Not Mastered

Answer: A

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

NEW QUESTION 79
- (Exam Topic 3)

You develop a dataset named DBTBL1 by using Azure Databricks. DBTBL1 contains the following columns:

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

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

```
df.write
```

▼
.bucketBy
.format
.partitionBy
.sortBy

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

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

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

A. Mastered
B. Not Mastered

Answer: A

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

NEW QUESTION 83
- (Exam Topic 3)
You have the following table named Employees.

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

You need to calculate the employee_type value based on the hire_date value.
How should you complete the Transact-SQL statement? To answer, drag the appropriate values to the correct targets. Each value may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.
NOTE: Each correct selection is worth one point.

Values

CASE

ELSE

OVER

PARTITION BY

ROW_NUMBER

Answer Area

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

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Graphical user interface, text, application Description automatically generated
Box 1: CASE
CASE evaluates a list of conditions and returns one of multiple possible result expressions.
CASE can be used in any statement or clause that allows a valid expression. For example, you can use CASE in statements such as SELECT, UPDATE, DELETE and SET, and in clauses such as select_list, IN, WHERE, ORDER BY, and HAVING.
Syntax: Simple CASE expression: CASE input_expression
WHEN when_expression THEN result_expression [...n] [ELSE else_result_expression]
END
Box 2: ELSE
Reference:
https://docs.microsoft.com/en-us/sql/t-sql/language-elements/case-transact-sql

NEW QUESTION 87
- (Exam Topic 3)
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You are designing an Azure Stream Analytics solution that will analyze Twitter data.
You need to count the tweets in each 10-second window. The solution must ensure that each tweet is counted only once.
Solution: You use a hopping window that uses a hop size of 10 seconds and a window size of 10 seconds. Does this meet the goal?

A. Yes
B. No

Answer: B

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

NEW QUESTION 90
- (Exam Topic 3)
You have an Azure subscription that contains an Azure Synapse Analytics dedicated SQL pool named Pool1. Pool1 receives new data once every 24 hours.
You have the following function.

```
create function dbo.udfFtoC(F decimal)
return decimal
as
begin
return (F - 32) * 5.0 / 9
end
```

You have the following query.

```
select avg_date, sensorid, avg_f, dbo.udfFtoC(avg_temperature) as avg_c from SensorTemps
where avg_date = @parameter
```

The query is executed once every 15 minutes and the @parameter value is set to the current date. You need to minimize the time it takes for the query to return results.
Which two actions should you perform? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

A. Create an index on the avg_f column.
B. Convert the avg_c column into a calculated column.
C. Create an index on the sensorid column.
D. Enable result set caching.
E. Change the table distribution to replicate.

Answer: BD

Explanation:
https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/performance-tuning-result-set-cac

NEW QUESTION 95
- (Exam Topic 3)
You are designing 2 solution that will use tables in Delta Lake on Azure Databricks. You need to minimize how long it takes to perform the following:
*Queries against non-partitioned tables
* Joins on non-partitioned columns
Which two options should you include in the solution? Each correct answer presents part of the solution. (Choose Correct Answer and Give explanation and References to Support the answers based from Data
Engineering on Microsoft Azure)

A. Z-Ordering
B. Apache Spark caching
C. dynamic file pruning (DFP)
D. the clone command

Answer: AC

Explanation:
According to the information I found on the web, two options that you should include in the solution to minimize how long it takes to perform queries and joins on non-partitioned tables are:

≫ Z-Ordering: This is a technique to colocate related information in the same set of files. This co-locality
is automatically used by Delta Lake in data-skipping algorithms. This behavior dramatically reduces the
amount of data that Delta Lake on Azure Databricks needs to read123.

≫ Apache Spark caching: This is a feature that allows you to cache data in memory or on disk for faster access. Caching can improve the performance of
repeated queries and joins on the same data. You can cache Delta tables using the CACHE TABLE or CACHE LAZY commands.
To minimize the time it takes to perform queries against non-partitioned tables and joins on non-partitioned columns in Delta Lake on Azure Databricks, the following options should be included in the solution:
* A. Z-Ordering: Z-Ordering improves query performance by co-locating data that share the same column values in the same physical partitions. This reduces the need for shuffling data across nodes during query execution. By using Z-Ordering, you can avoid full table scans and reduce the amount of data processed.
* B. Apache Spark caching: Caching data in memory can improve query performance by reducing the amount of data read from disk. This helps to speed up subsequent queries that need to access the same data. When you cache a table, the data is read from the data source and stored in memory. Subsequent queries can then read the data from memory, which is much faster than reading it from disk.
References:

≫ Delta Lake on Databricks: https://docs.databricks.com/delta/index.html

≫ Best Practices for Delta Lake on
Databricks: https://databricks.com/blog/2020/05/14/best-practices-for-delta-lake-on-databricks.html

NEW QUESTION 98
- (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 Standard cluster for the data engineers, and a High Concurrency cluster for the jobs.
Does this meet the goal?

A. Yes
B. No

Answer: B

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 101
- (Exam Topic 3)
You are building an Azure Stream Analytics job to identify how much time a user spends interacting with a feature on a webpage.
The job receives events based on user actions on the webpage. Each row of data represents an event. Each event has a type of either 'start' or 'end'.
You need to calculate the duration between start and end events.
How should you complete the query? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

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

A. Mastered
B. Not Mastered

Answer: A

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

NEW QUESTION 106
- (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 110
- (Exam Topic 3)
You build an Azure Data Factory pipeline to move data from an Azure Data Lake Storage Gen2 container to a database in an Azure Synapse Analytics dedicated SQL pool.
Data in the container is stored in the following folder structure.
/in/{YYYY}/{MM}/{DD}/{HH}/{mm}
The earliest folder is /in/2021/01/01/00/00. The latest folder is /in/2021/01/15/01/45. You need to configure a pipeline trigger to meet the following requirements:

≫ Existing data must be loaded.

≫ Data must be loaded every 30 minutes.

≫ Late-arriving data of up to two minutes must he included in the load for the time at which the data should have arrived.
How should you configure the pipeline trigger? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

Type: ▼

Event
On-demand
Schedule
Tumbling window

Additional properties: ▼

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

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Box 1: Tumbling window
To be able to use the Delay parameter we select Tumbling window. Box 2:
Recurrence: 30 minutes, not 32 minutes
Delay: 2 minutes.
The amount of time to delay the start of data processing for the window. The pipeline run is started after the expected execution time plus the amount of delay. The delay defines how long the trigger waits past the due time before triggering a new run. The delay doesn't alter the window startTime.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/how-to-create-tumbling-window-trigger

NEW QUESTION 115
- (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.

Create: ▼

A materialized view in Pool1
A security policy for Sales
Database scoped credentials in Pool1

Restrict row access by using: ▼

A masking rule
A table-valued function
The CONTAINS predicate

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 118
- (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 123
- (Exam Topic 1)
You need to design the partitions for the product sales transactions. The solution must meet the sales transaction dataset requirements.
What should you include in the solution? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

Partition product sales transactions data by:

| Sales date |
| Product ID |
| Promotion ID |

Store product sales transactions data in:

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

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Box 1: Sales date
Scenario: Contoso requirements for data integration include:
> Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month. Boundary values must belong to the partition on the right.
Box 2: An Azure Synapse Analytics Dedicated SQL pool Scenario: Contoso requirements for data integration include:
> Ensure that data storage costs and performance are predictable.
The size of a dedicated SQL pool (formerly SQL DW) is determined by Data Warehousing Units (DWU). Dedicated SQL pool (formerly SQL DW) stores data in relational tables with columnar storage. This format
significantly reduces the data storage costs, and improves query performance.
Synapse analytics dedicated sql pool Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-overview-wha

NEW QUESTION 126
- (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.

Transact-SQL DDL command to use:	▼
	CREATE EXTERNAL TABLE
	CREATE TABLE
	CREATE VIEW

Partitioning option to use in the WITH clause of the DDL statement:	▼
	FORMAT_OPTIONS
	FORMAT_TYPE
	RANGE LEFT FOR VALUES
	RANGE RIGHT FOR VALUES

A. Mastered
B. Not Mastered

Answer: A

Explanation:
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 127
- (Exam Topic 1)
You need to implement the surrogate key for the retail store table. The solution must meet the sales transaction dataset requirements.
What should you create?

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

Answer: A

Explanation:
Scenario: Implement a surrogate key to account for changes to the retail store addresses.
A surrogate key on a table is a column with a unique identifier for each row. The key is not generated from the table data. Data modelers like to create surrogate keys on their tables when they design data warehouse models. You can use the IDENTITY property to achieve this goal simply and effectively without affecting load performance.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-identity

NEW QUESTION 132
- (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.

Table type to store retail store data:

| Hash |
| Replicated |
| Round-robin |

Table type to store promotional data:

| Hash |
| Replicated |
| Round-robin |

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 136
- (Exam Topic 1)
You need to design a data ingestion and storage solution for the Twitter feeds. The solution must meet the customer sentiment analytics requirements.
What should you include in the solution? To answer, select the appropriate options in the answer area NOTE: Each correct selection b worth one point.

Answer Area

To increase the throughput of ingesting the Twitter feeds:

| Configure Event Hubs partitions. |
| Enable Auto-Inflate in Event Hubs. |
| Use Event Hubs Dedicated. |

To store the Twitter feed data, use:

| An Azure Data Lake Storage Gen2 account |
| An Azure Databricks high concurrency cluster |
| An Azure General-purpose v2 storage account in the Premium tier |

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Graphical user interface, text Description automatically generated
Box 1: Configure Evegent Hubs partitions
Scenario: Maximize the throughput of ingesting Twitter feeds from Event Hubs to Azure Storage without purchasing additional throughput or capacity units.
Event Hubs is designed to help with processing of large volumes of events. Event Hubs throughput is scaled by using partitions and throughput-unit allocations.
Event Hubs traffic is controlled by TUs (standard tier). Auto-inflate enables you to start small with the minimum required TUs you choose. The feature then scales automatically to the maximum limit of TUs you need, depending on the increase in your traffic.
Box 2: An Azure Data Lake Storage Gen2 account
Scenario: Ensure that the data store supports Azure AD-based access control down to the object level. Azure Data Lake Storage Gen2 implements an access control model that supports both Azure role-based
access control (Azure RBAC) and POSIX-like access control lists (ACLs).
Reference:
https://docs.microsoft.com/en-us/azure/event-hubs/event-hubs-features https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control

NEW QUESTION 141
- (Exam Topic 3)
You have an Azure Active Directory (Azure AD) tenant that contains a security group named Group1. You have an Azure Synapse Analytics dedicated SQL pool named dw1 that contains a schema named schema1.
You need to grant Group1 read-only permissions to all the tables and views in schema1. The solution must use the principle of least privilege.
Which three actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.
NOTE: More than one order of answer choices is correct. You will receive credit for any of the correct orders you select.

Actions   **Answer Area**

Create a database role named Role1 and grant Role1 `SELECT` permissions to schema1.
Create a database role named Role1 and grant Role1 `SELECT` permissions to dw1.
Assign the Azure role-based access control (Azure RBAC) Reader role for dw1 to Group1.
Create a database user in dw1 that represents Group1 and uses the `FROM EXTERNAL PROVIDER` clause.
Assign Role1 to the Group1 database user.

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Step 1: Create a database role named Role1 and grant Role1 SELECT permissions to schema You need to grant Group1 read-only permissions to all the tables and views in schema1.
Place one or more database users into a database role and then assign permissions to the database role. Step 2: Assign Rol1 to the Group database user
Step 3: Assign the Azure role-based access control (Azure RBAC) Reader role for dw1 to Group1 Reference:
https://docs.microsoft.com/en-us/azure/data-share/how-to-share-from-sql

NEW QUESTION 144
- (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 148
- (Exam Topic 3)
You are creating an Azure Data Factory data flow that will ingest data from a CSV file, cast columns to specified types of data, and insert the data into a table in an Azure Synapse Analytics dedicated SQL pool. The CSV file contains columns named username, comment and date.
The data flow already contains the following:
• A source transformation
• A Derived Column transformation to set the appropriate types of data
• A sink transformation to land the data in the pool
You need to ensure that the data flow meets the following requirements;
• All valid rows must be written to the destination table.
• Truncation errors in the comment column must be avoided proactively.
• Any rows containing comment values that will cause truncation errors upon insert must be written to a file in blob storage.
Which two actions should you perform? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point

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

Answer: BD

NEW QUESTION 149
- (Exam Topic 3)
You have an Azure data factory that has the Git repository settings shown in the following exhibit.

Git repository

Git repository information associated with your data factory. CI/CD best practices ⬚

✎ Edit   ⟳ Overwrite live mode   ⮒ Disconnect   ⬆ Import resources

Repository type	Azure DevOps Git
Azure DevOps Account	
Project name	ADFDeployDemo
Repository name	ADEDeployDemo
Collaboration branch	main
Publish branch	adf_publish
Root folder	/
Last published commit	23b144ac4aa7daf16f2fe7c2ab0eb303a8e4ed65
Publish (from ADF Studio)	Enabled

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

Changes to pipelines will be saved in Azure DevOps [answer choice].

every 20 seconds ▼
every 20 seconds
when the pipeline is published
when the pipeline is saved

To publish changes by using Azure Data Factory Studio, the changes must first be saved in the [answer choice].

root folder ▼
adf_publish branch
main branch
root folder

A. Mastered
B. Not Mastered

Answer: A

Explanation:

Answer Area

Changes to pipelines will be saved in Azure DevOps [answer choice].

every 20 seconds ▼
every 20 seconds
when the pipeline is published
when the pipeline is saved

To publish changes by using Azure Data Factory Studio, the changes must first be saved in the [answer choice].

root folder ▼
adf_publish branch
main branch
root folder

NEW QUESTION 151
- (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.

Databricks: ▼

Azure Active Directory credential passthrough
Azure Key Vault secrets
Personal access tokens

Data Lake Storage: ▼

Azure Active Directory credential passthrough
Shared access keys
Shared access signatures

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 156
- (Exam Topic 3)
You have an Azure Synapse Analytics serverless SQL pool named Pool1 and an Azure Data Lake Storage Gen2 account named storage1. The AllowedBlobpublicAccess porperty is disabled for storage1.
You need to create an external data source that can be used by Azure Active Directory (Azure AD) users to access storage1 from Pool1.
What should you create first?

A. an external resource pool
B. a remote service binding
C. database scoped credentials
D. an external library

Answer: C

Explanation:
 Security
User must have SELECT permission on an external table to read the data. External tables access underlying Azure storage using the database scoped credential defined in data source.
Note: A database scoped credential is a record that contains the authentication information that is required to connect to a resource outside SQL Server. Most credentials include a Windows user and password.
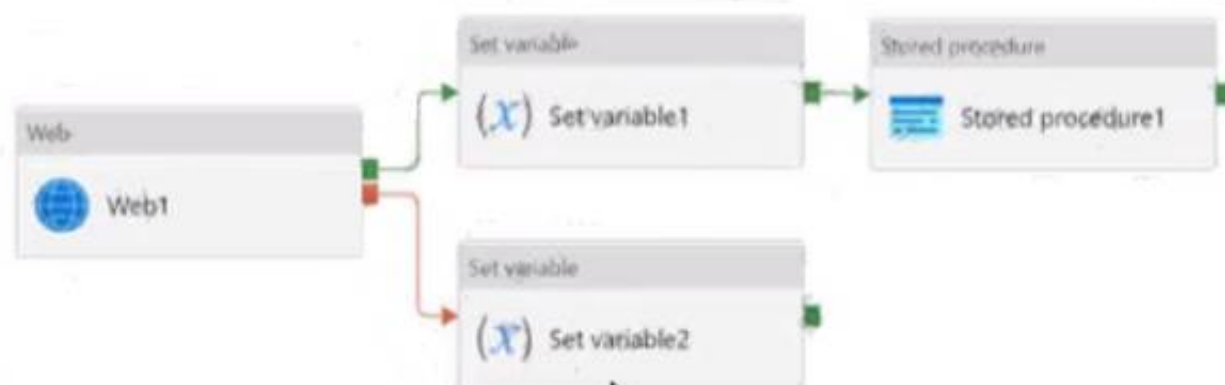Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables https://docs.microsoft.com/en-us/sql/t-sql/statements/create-database-scoped-credential-transact-sql

NEW QUESTION 158
- (Exam Topic 3)
You have an Azure Data Factory pipeline that has the activity shown in the following exhibit.

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

A. Mastered
B. Not Mastered

Answer: A

Explanation:

Answer Area

Stored procedure1 will execute if Web1 and Set variable1 [answer choice].   succeed   ▼

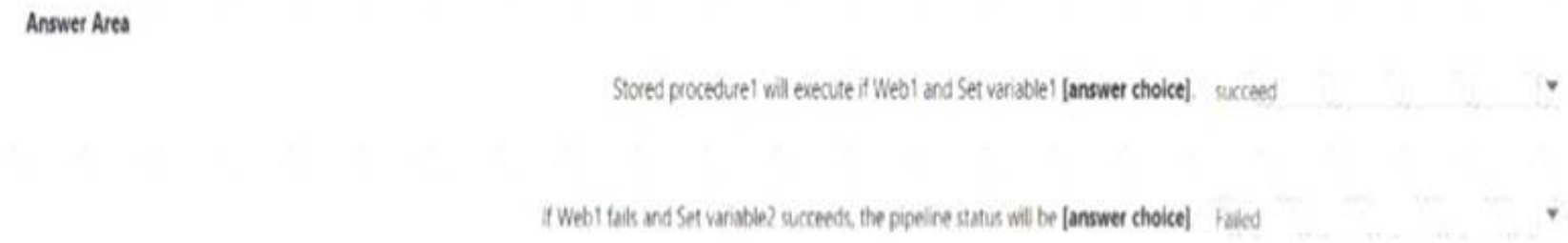If Web1 fails and Set variable2 succeeds, the pipeline status will be [answer choice]   Failed   ▼

NEW QUESTION 162
- (Exam Topic 3)
You have an enterprise data warehouse in Azure Synapse Analytics.
You need to monitor the data warehouse to identify whether you must scale up to a higher service level to accommodate the current workloads
Which is the best metric to monitor?
More than one answer choice may achieve the goal. Select the BEST answer.

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

Answer: C

NEW QUESTION 165
- (Exam Topic 3)
You plan to monitor an Azure data factory by using the Monitor & Manage app.
You need to identify the status and duration of activities that reference a table in a source database.
Which three actions should you perform in sequence? To answer, move the actions from the list of actions to the answer are and arrange them in the correct order.

Actions

From the Data Factory monitoring app, add the Source user property to the Activity Runs table.

From the Data Factory monitoring app, add the Source user property to the Pipeline Runs table.

From the Data Factory authoring UI, publish the pipelines.

From the Data Factory monitoring app, add a linked service to the Pipeline Runs table.

From the Data Factory authoring UI, generate a user property for Source on all activities.

From the Data Factory authoring UI, generate a user property for Source on all datasets.

Answer Area

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Step 1: From the Data Factory authoring UI, generate a user property for Source on all activities. Step 2: From the Data Factory monitoring app, add the Source user property to Activity Runs table.
You can promote any pipeline activity property as a user property so that it becomes an entity that you can
monitor. For example, you can promote the Source and Destination properties of the copy activity in your pipeline as user properties. You can also select Auto Generate to generate the Source and Destination user properties for a copy activity.
Step 3: From the Data Factory authoring UI, publish the pipelines
Publish output data to data stores such as Azure SQL Data Warehouse for business intelligence (BI) applications to consume.
References:
https://docs.microsoft.com/en-us/azure/data-factory/monitor-visually

NEW QUESTION 168
- (Exam Topic 3)
You need to create an Azure Data Factory pipeline to process data for the following three departments at your company: Ecommerce, retail, and wholesale. The solution must ensure that data can also be processed for the entire company.
How should you complete the Data Factory data flow script? To answer, drag the appropriate values to the correct targets. Each value may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.
NOTE: Each correct selection is worth one point.

Values

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

Answer Area

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

A. Mastered
B. Not Mastered

Answer: A

Explanation:
The conditional split transformation routes data rows to different streams based on matching conditions. The conditional split transformation is similar to a CASE decision structure in a programming language. The transformation evaluates expressions, and based on the results, directs the data row to the specified stream.
Box 1: dept=='ecommerce', dept=='retail', dept=='wholesale'
First we put the condition. The order must match the stream labeling we define in Box 3. Syntax:
<incomingStream> split(
<conditionalExpression1>
<conditionalExpression2> disjoint: {true | false}
) ~> <splitTx>@(stream1, stream2, ..., <defaultStream>)
Box 2: discount : false
disjoint is false because the data goes to the first matching condition. All remaining rows matching the third condition go to output stream all.
Box 3: ecommerce, retail, wholesale, all Label the streams
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/data-flow-conditional-split

NEW QUESTION 170
- (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 171
- (Exam Topic 3)
You are designing an Azure Synapse Analytics dedicated SQL pool.
You need to ensure that you can audit access to Personally Identifiable information (PII). What should you include in the solution?

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

Answer: C

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

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

≫ Helping to meet standards for data privacy and requirements for regulatory compliance.

≫ Various security scenarios, such as monitoring (auditing) access to sensitive data.

≫ Controlling access to and hardening the security of databases that contain highly sensitive data.

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

NEW QUESTION 175
- (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.

Library:
| Azure Databricks Monitoring Library |
| Microsoft Azure Management Monitoring Library |
| PyTorch |
| TensorFlow |

Workspace:
| Azure Databricks |
| Azure Log Analytics |
| Azure Machine Learning |

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 178
- (Exam Topic 3)
You have two Azure Storage accounts named Storage1 and Storage2. Each account holds one container and has the hierarchical namespace enabled. The system has files that contain data stored in the Apache Parquet format.
You need to copy folders and files from Storage1 to Storage2 by using a Data Factory copy activity. The
solution must meet the following requirements:

≫ No transformations must be performed.

≫ The original folder structure must be retained.

≫ Minimize time required to perform the copy activity.

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

Source dataset type:
| Binary |
| Parquet |
| Delimited text |

Copy activity copy behavior:
| FlattenHierarchy |
| MergeFiles |
| PreserveHierarchy |

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Graphical user interface, text, application, chat or text message Description automatically generated
Box 1: Parquet

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

NEW QUESTION 182
- (Exam Topic 3)
You have a table in an Azure Synapse Analytics dedicated SQL pool. The table was created by using the following Transact-SQL statement.

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

You need to alter the table to meet the following requirements:
≫ Ensure that users can identify the current manager of employees.
≫ Support creating an employee reporting hierarchy for your entire company.
≫ Provide fast lookup of the managers' attributes such as name and job title.
Which column should you add to the table?

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

Answer: A

Explanation:
Use the same definition as the EmployeeID column. Reference:
https://docs.microsoft.com/en-us/analysis-services/tabular-models/hierarchies-ssas-tabular

NEW QUESTION 185
- (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 190
- (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 195
- (Exam Topic 3)
You are designing a highly available Azure Data Lake Storage solution that will induce geo-zone-redundant storage (GZRS).
You need to monitor for replication delays that can affect the recovery point objective (RPO). What should you include m the monitoring solution?

A. Last Sync Time
B. Average Success Latency
C. Error errors
D. availability

Answer: A

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 197
- (Exam Topic 3)
You are deploying a lake database by using an Azure Synapse database template.
You need to add additional tables to the database. The solution must use the same grouping method as the template tables.
'Which grouping method should you use?

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

Answer: A

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

NEW QUESTION 200
- (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 201
- (Exam Topic 3)
You have an on-premises data warehouse that includes the following fact tables. Both tables have the following columns: DateKey, ProductKey, RegionKey. There are 120 unique product keys and 65 unique region keys.

Table	Comments
Sales	The table is 600 GB in size. DateKey is used extensively in the WHERE clause in queries. ProductKey is used extensively in join operations. RegionKey is used for grouping. Severity-five percent of records relate to one of 40 regions.
Invoice	The table is 6 GB in size. DateKey and ProductKey are used extensively in the WHERE clause in queries. RegionKey is used for grouping.

Queries that use the data warehouse take a long time to complete.
You plan to migrate the solution to use Azure Synapse Analytics. You need to ensure that the Azure-based solution optimizes query performance and minimizes processing skew.
What should you recommend? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point

Table	Distribution type	Distribution column
Sales:	Hash-distributed / Round-robin	DateKey / ProductKey / RegionKey
Invoices:	Hash-distributed / Round-robin	DateKey / ProductKey / RegionKey

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Box 1: Hash-distributed
Box 2: ProductKey
ProductKey is used extensively in joins.
Hash-distributed tables improve query performance on large fact tables. Box 3: Round-robin
Box 4: RegionKey
Round-robin tables are useful for improving loading speed.
Consider using the round-robin distribution for your table in the following scenarios:

> When getting started as a simple starting point since it is the default

> If there is no obvious joining key

> If there is not good candidate column for hash distributing the table

> If the table does not share a common join key with other tables

> If the join is less significant than other joins in the query

> When the table is a temporary staging table

Note: A distributed table appears as a single table, but the rows are actually stored across 60 distributions. The rows are distributed with a hash or round-robin algorithm.
Reference:
https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-tables-distribute

NEW QUESTION 203
- (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 205
- (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 206
- (Exam Topic 3)
You need to implement a Type 3 slowly changing dimension (SCD) for product category data in an Azure Synapse Analytics dedicated SQL pool.
You have a table that was created by using the following Transact-SQL statement.

```
CREATE TABLE [DBO].[DimProduct] (
[ProductKey] [int] IDENTITY(1,1) NOT NULL,
[ProductSourceID] [int] NOT NULL,
[ProductName] [nvarchar] (100) NULL,
[Color] [nvarchar] (15) NULL,
[SellStartDate] [date] NOT NULL,
[SellEndDate] [date] NULL,
[RowInsertedDateTime] [datetime] NOT NULL,
[RowUpdatedDateTime] [datetime] NOT NULL,
[ETLAuditID] [int] NOT NULL
)
```

Which two columns should you add to the table? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.

A. [EffectiveScarcDate] [datetime] NOT NULL,
B. [CurrentProduccCacegory] [nvarchar] (100) NOT NULL,
C. [EffectiveEndDace] [dacecime] NULL,
D. [ProductCategory] [nvarchar] (100) NOT NULL,
E. [OriginalProduccCacegory] [nvarchar] (100) NOT NULL,

Answer: BE

Explanation:
A Type 3 SCD supports storing two versions of a dimension member as separate columns. The table includes a column for the current value of a member plus either the original or previous value of the member. So Type 3 uses additional columns to track one key instance of history, rather than storing additional rows to track each change like in a Type 2 SCD.
This type of tracking may be used for one or two columns in a dimension table. It is not common to use it for many members of the same table. It is often used in combination with Type 1 or Type 2 members.
Graphical user interface, application, email Description automatically generated

CustomerID	FirstName	LastName	CurrentEmail	OriginalEmail	CompanyName	InsertedDate	ModifiedDate
2	Keith	Harris	keith0@aw.com	keith0@aw.com	Progressive Sports	2021-03-20	2021-03-20
3	Donna	Carreras	donna0@aw.com	donna0@aw.com	A Bike Store	2021-03-20	2021-03-20

CustomerID	FirstName	LastName	CurrentEmail	OriginalEmail	CompanyName	InsertedDate	ModifiedDate
2	Keith	Harris	keith0@aw.com	keith0@aw.com	Progressive Sports	2021-03-20	2021-03-20
3	Donna	Carreras	dc3@aw.com	donna0@aw.com	A Bike Store	2021-03-20	2021-03-22

Reference:
https://k21academy.com/microsoft-azure/azure-data-engineer-dp203-q-a-day-2-live-session-review/
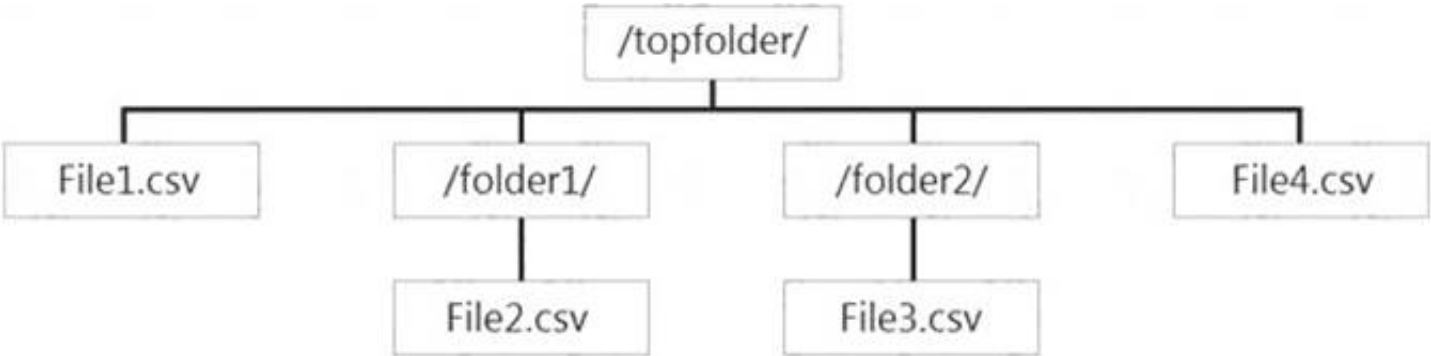

NEW QUESTION 208
- (Exam Topic 3)
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct 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 213
- (Exam Topic 3)
You have files and folders in Azure Data Lake Storage Gen2 for an Azure Synapse workspace as shown in the following exhibit.

```
                    /topfolder/
        ┌──────────┬──────────┬──────────┐
   File1.csv   /folder1/   /folder2/   File4.csv
                    │           │
               File2.csv   File3.csv
```

You create an external table named ExtTable that has LOCATION='/topfolder/'.
When you query ExtTable by using an Azure Synapse Analytics serverless SQL pool, which files are returned?

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

Answer: B

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

NEW QUESTION 217
- (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.

Table distribution types	Answer Area
Hash	Common.Data: [ ]
Replicated	Marketing.Web.Sessions: [ ]
Round-robin	Staging. Web.Sessions: [ ]

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Box 1: Replicated
The best table storage option for a small table is to replicate it across all the Compute nodes. Box 2: Hash
Hash-distribution improves query performance on large fact tables. Box 3: Round-robin
Round-robin distribution is useful for improving loading speed.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribu

NEW QUESTION 219
- (Exam Topic 3)
You have an Azure Storage account that generates 200,000 new files daily. The file names have a format of {YYYY}/{MM}/{DD}/{HH}/{CustomerID}.csv.
You need to design an Azure Data Factory solution that will load new data from the storage account to an
Azure Data Lake once hourly. The solution must minimize load times and costs.
How should you configure the solution? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

Load methodology:	▼
Full Load	
Incremental Load	
Load individual files as they arrive	

| Trigger:	▼
Fixed schedule	
New file	
Tumbling window	

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Table Description automatically generated
Box 1: Incremental load Box 2: Tumbling window
Tumbling windows are a series of fixed-sized, non-overlapping and contiguous time intervals. The following diagram illustrates a stream with a series of events and how they are mapped into 10-second tumbling windows.
Timeline Description automatically generated

Tell me the count of tweets per time zone every 10 seconds
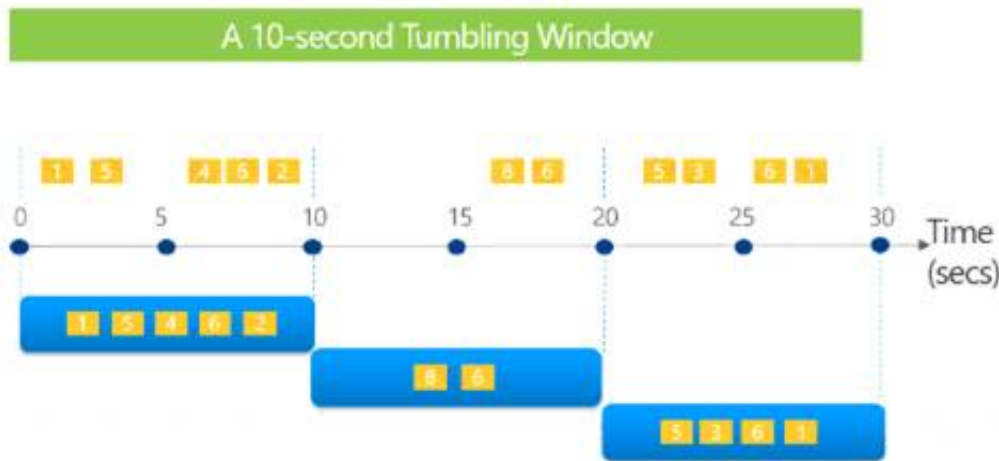
```
SELECT TimeZone, COUNT(*) AS Count
FROM TwitterStream TIMESTAMP BY CreatedAt
GROUP BY TimeZone, TumblingWindow(second,10)
```

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

NEW QUESTION 220
- (Exam Topic 3)
You need to output files from Azure Data Factory.
Which file format should you use for each type of output? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

Columnar format:

Avro
GZip
Parquet
TXT

JSON with a timestamp:

Avro
GZip
Parquet
TXT

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Box 1: Parquet
Parquet stores data in columns, while Avro stores data in a row-based format. By their very nature,
column-oriented data stores are optimized for read-heavy analytical workloads, while row-based databases are best for write-heavy transactional workloads.
Box 2: Avro
An Avro schema is created using JSON format. AVRO supports timestamps.
Note: Azure Data Factory supports the following file formats (not GZip or TXT).

≫ Avro format
≫ Binary format
≫ Delimited text format
≫ Excel format
≫ JSON format
≫ ORC format
≫ Parquet format
≫ XML format
Reference:
https://www.datanami.com/2018/05/16/big-data-file-formats-demystified

NEW QUESTION 222
- (Exam Topic 3)
You are batch loading a table in an Azure Synapse Analytics dedicated SQL pool.
You need to load data from a staging table to the target table. The solution must ensure that if an error occurs while loading the data to the target table, all the inserts in that batch are undone.
How should you complete the Transact-SQL code? To answer, drag the appropriate values to the correct
targets. Each value may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.
NOTE Each correct selection is worth one point.

Values

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

Answer Area

```
[                    ]

BEGIN TRY

    INSERT INTO dbo.Table1 (col1, col2, col3)

    SELECT col1, col2, col3 FROM stage.Table1;

END TRY

    BEGIN CATCH

        IF @@TRANCOUNT > 0

        BEGIN

            [                    ];

        END

    END CATCH;

IF @@TRANCOUNT >0

BEGIN

    COMMIT TRAN;

END
```

A. Mastered
B. Not Mastered

Answer: A

Explanation:

NEW QUESTION 227
- (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 230
- (Exam Topic 3)
You have an Azure data factory named ADM that contains a pipeline named Pipelwe1 Pipeline! must execute every 30 minutes with a 15-minute offset.
Vou need to create a trigger for Pipehne1. The trigger must meet the following requirements:
• Backfill data from the beginning of the day to the current time.
• If Pipeline1 fairs, ensure that the pipeline can re-execute within the same 30-mmute period.
• Ensure that only one concurrent pipeline execution can occur.
• Minimize de4velopment and configuration effort Which type of trigger should you create?

A. schedule
B. event-based
C. manual
D. tumbling window
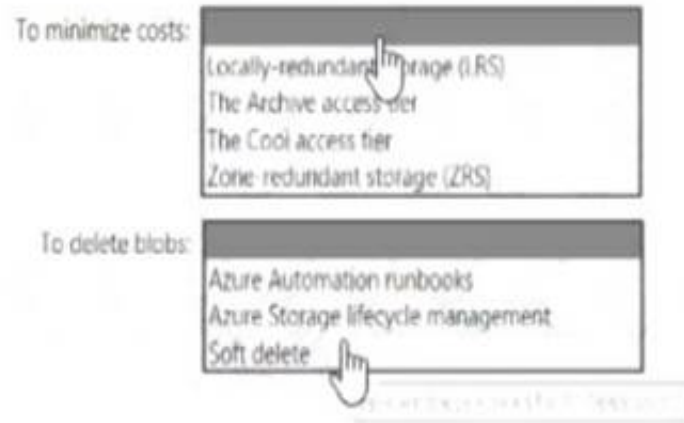
Answer: A

NEW QUESTION 232
- (Exam Topic 3)
You have an Azure subscription.
You need to deploy an Azure Data Lake Storage Gen2 Premium account. The solution must meet the following requirements:
• Blobs that are older than 365 days must be deleted.
• Administrator efforts must be minimized.
• Costs must be minimized
What should you use? To answer, select the appropriate options in the answer area. NOTE Each correct selection is worth one point.

Answer Area

A. Mastered
B. Not Mastered

Answer: A

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

NEW QUESTION 236
- (Exam Topic 3)
You have an Azure Synapse Analytics dedicated SQL pool named Pool1 and a database named DB1. DB1 contains a fact table named Table1.
You need to identify the extent of the data skew in Table1. What should you do in Synapse Studio?

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

Answer: D

Explanation:
Microsoft recommends use of sys.dm_pdw_nodes_db_partition_stats to analyze any skewness in the data. Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/cheat-sheet

NEW QUESTION 241
- (Exam Topic 3)
You have an Azure Data Lake Storage Gen2 account that contains two folders named Folder and Folder2. You use Azure Data Factory to copy multiple files from Folder1 to Folder2.

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

You receive the following error.
What should you do to resolve the error.

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

Answer: A

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

NEW QUESTION 242
- (Exam Topic 3)
You are designing an Azure Data Lake Storage solution that will transform raw JSON files for use in an analytical workload.
You need to recommend a format for the transformed files. The solution must meet the following requirements:

➢ Contain information about the data types of each column in the files.

➢ Support querying a subset of columns in the files.

➢ Support read-heavy analytical workloads.

➢ Minimize the file size.
What should you recommend?

A. JSON
B. CSV
C. Apache Avro

D. Apache Parquet

Answer: D

Explanation:
Parquet, an open-source file format for Hadoop, stores nested data structures in a flat columnar format. Compared to a traditional approach where data is stored in a row-oriented approach, Parquet file format is
more efficient in terms of storage and performance.
It is especially good for queries that read particular columns from a "wide" (with many columns) table since only needed columns are read, and IO is minimized.
Reference: https://www.clairvoyant.ai/blog/big-data-file-formats

NEW QUESTION 247
- (Exam Topic 3)
You have an Azure Synapse Analytics SQL pool named Pool1 on a logical Microsoft SQL server named Server1.
You need to implement Transparent Data Encryption (TDE) on Pool1 by using a custom key named key1. Which five actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

Actions

Answer Area

Enable TDE on Pool1.

Assign a managed identity to Server1.

Configure key1 as the TDE protector for Server1.

Add key1 to the Azure key vault.

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

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Graphical user interface, text, application Description automatically generated
Step 1: Assign a managed identity to Server1
You will need an existing Managed Instance as a prerequisite.
Step 2: Create an Azure key vault and grant the managed identity permissions to the vault Create Resource and setup Azure Key Vault.
Step 3: Add key1 to the Azure key vault
The recommended way is to import an existing key from a .pfx file or get an existing key from the vault. Alternatively, generate a new key directly in Azure Key Vault.
Step 4: Configure key1 as the TDE protector for Server1 Provide TDE Protector key
Step 5: Enable TDE on Pool1 Reference:
https://docs.microsoft.com/en-us/azure/azure-sql/managed-instance/scripts/transparent-data-encryption-byok-po

NEW QUESTION 252
- (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 256
- (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 261
- (Exam Topic 3)
You are responsible for providing access to an Azure Data Lake Storage Gen2 account.
Your user account has contributor access to the storage account, and you have the application ID and access key.
You plan to use PolyBase to load data into an enterprise data warehouse in Azure Synapse Analytics. You need to configure PolyBase to connect the data warehouse to storage account.
Which three components should you create in sequence? To answer, move the appropriate components from the list of components to the answer area and arrange them in the correct order.

Components

```
a database scoped credential

an asymmetric key

an external data source

a database encryption key

an external file format
```

Answer Area

A. Mastered
B. Not Mastered

Answer: A

Explanation:

Components

| a database scoped credential |
| an asymmetric key |
| an external data source |
| a database encryption key |
| an external file format |

Answer Area

| a database scoped credential |
| an external data source |
| an external file format |

NEW QUESTION 262

- (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.

Values

| CustomerKey |
| HASH |
| ROUND_ROBIN |
| REPLICATE |
| OrderDateKey |
| SalesOrderNumber |

Answer Area

```
CREATE TABLE [dbo].[FactSales]
(
    [ProductKey]        int        NOT NULL
,   [OrderDateKey]      int        NOT NULL
,   [CustomerKey]       int        NOT NULL
,   [SalesOrderNumber]  nvarchar ( 20 )   NOT NULL
,   [OrderQuantity]         smallint      NOT NULL
,   [UnitPrice]             money         NOT NULL
)
WITH
(   CLUSTERED       COLUMNSTORE      INDEX
,   DISTRIBUTION =  [  Value  ]  ([ProductKey])

,   PARTITION   (  [  Value  ]  ) RANGE RIGHT FOR VALUES
                (20170101,20180101,20190101,20200101,20210101)
                    )
)
```

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 264

- (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 267
- (Exam Topic 3)
You have an Azure Databricks workspace that contains a Delta Lake dimension table named Tablet. Table1 is a Type 2 slowly changing dimension (SCD) table.
You need to apply updates from a source table to Table1. Which Apache Spark SQL operation should you use?

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

Answer: C

Explanation:
The Delta provides the ability to infer the schema for data input which further reduces the effort required in managing the schema changes. The Slowly Changing Data(SCD) Type 2 records all the changes made to each key in the dimensional table. These operations require updating the existing rows to mark the previous values of the keys as old and then inserting new rows as the latest values. Also, Given a source table with the updates and the target table with dimensional data, SCD Type 2 can be expressed with the merge.
Example:
// Implementing SCD Type 2 operation using merge function customersTable
as("customers") merge(
stagedUpdates.as("staged_updates"), "customers.customerId = mergeKey")
whenMatched("customers.current = true AND customers.address <> staged_updates.address") updateExpr(Map(
"current" -> "false",
"endDate" -> "staged_updates.effectiveDate")) whenNotMatched()
insertExpr(Map(
"customerid" -> "staged_updates.customerId", "address" -> "staged_updates.address", "current" -> "true",
"effectiveDate" -> "staged_updates.effectiveDate",
"endDate" -> "null")) execute()
}
Reference:
https://www.projectpro.io/recipes/what-is-slowly-changing-data-scd-type-2-operation-delta-table-databricks

NEW QUESTION 270
- (Exam Topic 3)
You are incrementally loading data into fact tables in an Azure Synapse Analytics dedicated SQL pool. Each batch of incoming data is staged before being loaded into the fact tables. |
You need to ensure that the incoming data is staged as quickly as possible. |
How should you configure the staging tables? To answer, select the appropriate options in the answer area.

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Round-robin distribution is recommended for staging tables because it distributes data evenly across all the distributions without requiring a hash column. This can improve the speed of data loading and avoid data skew. Heap tables are recommended for staging tables because they do not have any indexes or partitions that

can slow down the data loading process. Heap tables are also easier to truncate and reload than clustered index or columnstore index tables.

NEW QUESTION 271
- (Exam Topic 3)
You have an enterprise data warehouse in Azure Synapse Analytics.
Using PolyBase, you create an external table named [Ext].[Items] to query Parquet files stored in Azure Data Lake Storage Gen2 without importing the data to the data warehouse.
The external table has three columns.
You discover that the Parquet files have a fourth column named ItemID.
Which command should you run to add the ItemID column to the external table?

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

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

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

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

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

Answer: C

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

NEW QUESTION 273
- (Exam Topic 3)
You have an Azure subscription that is linked to a hybrid Azure Active Directory (Azure AD) tenant. The subscription contains an Azure Synapse Analytics SQL pool named Pool1.
You need to recommend an authentication solution for Pool1. The solution must support multi-factor authentication (MFA) and database-level authentication.
Which authentication solution or solutions should you include in the recommendation? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

MFA:
Azure AD authentication
Microsoft SQL Server authentication
Passwordless authentication
Windows authentication

Database-level authentication:
Application roles
Contained database users
Database roles
Microsoft SQL Server logins

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Graphical user interface, text, application, chat or text message Description automatically generated
Box 1: Azure AD authentication
Azure Active Directory authentication supports Multi-Factor authentication through Active Directory Universal Authentication.
Box 2: Contained database users
Azure Active Directory Uses contained database users to authenticate identities at the database level. Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-authentication

NEW QUESTION 275
- (Exam Topic 3)
You are designing a security model for an Azure Synapse Analytics dedicated SQL pool that will support multiple companies. You need to ensure that users from each company can view only the data of their respective company. Which two objects should you include in the solution? Each correct answer presents part of the solution
NOTE: Each correct selection it worth one point.

A. a custom role-based access control (RBAC) role.
B. asymmetric keys
C. a predicate function
D. a column encryption key
E. a security policy

Answer: AE

Explanation:
Reference:
https://docs.microsoft.com/en-us/sql/relational-databases/security/row-level-security https://docs.microsoft.com/en-us/azure/synapse-analytics/security/synapse-workspace-access-control-overview

NEW QUESTION 277
......

Thank You for Trying Our Product

We offer two products:

1st - We have Practice Tests Software with Actual Exam Questions

2nd - Questons and Answers in PDF Format

DP-203 Practice Exam Features:

* DP-203 Questions and Answers Updated Frequently

* DP-203 Practice Questions Verified by Expert Senior Certified Staff

* DP-203 Most Realistic Questions that Guarantee you a Pass on Your FirstTry

* DP-203 Practice Test Questions in Multiple Choice Formats and Updatesfor 1 Year

100% Actual & Verified — Instant Download, Please Click
Order The DP-203 Practice Test Here