

## DP-203 Dumps

### Data Engineering on Microsoft Azure

<https://www.certleader.com/DP-203-dumps.html>



**NEW QUESTION 1**

- (Exam Topic 3)

The storage account container view is shown in the Refdata exhibit. (Click the Refdata tab.) You need to configure the Stream Analytics job to pick up the new reference data. What should you configure? To answer, select the appropriate options in the answer area NOTE: Each correct selection is worth one point.

- A. Mastered
- B. Not Mastered

**Answer:** A

**Explanation:**

Answer as below

**NEW QUESTION 2**

- (Exam Topic 3)

You have two Azure Blob Storage accounts named account1 and account2?

You plan to create an Azure Data Factory pipeline that will use scheduled intervals to replicate newly created or modified blobs from account1 to account?

You need to recommend a solution to implement the pipeline. The solution must meet the following requirements:

- Ensure that the pipeline only copies blobs that were created or modified since the most recent replication event.
- Minimize the effort to create the pipeline. What should you recommend?

- A. Create a pipeline that contains a flowlet.
- B. Create a pipeline that contains a Data Flow activity.
- C. Run the Copy Data tool and select Metadata-driven copy task.
- D. Run the Copy Data tool and select Built-in copy task.

**Answer:** A

**NEW QUESTION 3**

- (Exam Topic 3)

You are designing a date dimension table in an Azure Synapse Analytics dedicated SQL pool. The date dimension table will be used by all the fact tables. Which distribution type should you recommend to minimize data movement?

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

**Answer:** B

**Explanation:**

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

Reference:

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

**NEW QUESTION 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 row for each item in the arrays.

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

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

**Answer:** B

**Explanation:**

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

Reference:

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

**NEW QUESTION 5**

- (Exam Topic 3)

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

Pool1.  
You plan to create a database named D61 in Pool1.  
You need to ensure that when tables are created in DB1, the tables are available automatically as external tables to the built-in serverless SQL pod.  
Which format should you use for the tables in DB1?

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

**Answer:** A

**Explanation:**

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

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

Reference:

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

**NEW QUESTION 6**

- (Exam Topic 3)

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

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

What should you recommend?

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

**Answer:** B

**Explanation:**

Three common analytics use cases with Microsoft Azure Databricks

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

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

Note: Recommendation engines, churn analysis, and intrusion detection are common scenarios that many organizations are solving across multiple industries.

They require machine learning, streaming analytics, and utilize massive amounts of data processing that can be difficult to scale without the right tools.

Reference:

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

**NEW QUESTION 7**

- (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 8**

- (Exam Topic 3)

You have an Azure Synapse Analytics dedicated SQL pool mat contains a table named dbo.Users.

You need to prevent a group of users from reading user email addresses from dbo.Users. What should you use?

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

**Answer:** B

**NEW QUESTION 9**

- (Exam Topic 3)

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

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

**Answer:** A

**Explanation:**

Convert nested JSON to a flattened DataFrame

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

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

Scala

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

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

**NEW QUESTION 10**

- (Exam Topic 3)

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

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

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

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

NOTE: Each correct selection is worth one point.

HubA:

▼

Stream

Reference

HubB:

▼

Stream

Reference

Database1:

▼

Stream

Reference

- A. Mastered
- B. Not Mastered

**Answer:** A

**Explanation:**

HubA: Stream HubB: Stream

Database1: Reference

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

Reference:

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

**NEW QUESTION 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 14**

- (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 19**

- (Exam Topic 3)

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

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

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

**Answer: B**

**Explanation:**

You can identify anomalies by routing data via IoT Hub to a built-in ML model in Azure Stream Analytics. Reference:

<https://docs.microsoft.com/en-us/learn/modules/data-anomaly-detection-using-azure-iot-hub/>



NEW QUESTION 23

- (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 24

- (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 28

- (Exam Topic 3)

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

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

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

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

**Answer:** CE

**Explanation:**

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

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

References:

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

**NEW QUESTION 32**

- (Exam Topic 3)

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

Which type of trigger should you use?

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

**Answer:** D

**Explanation:**

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

Reference:

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

**NEW QUESTION 33**

- (Exam Topic 3)

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

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

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

How often should you create a partition?

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

**Answer:** B

**Explanation:**

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

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

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

Reference:

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

**NEW QUESTION 37**

- (Exam Topic 3)

You have an Azure Factory instance named DF1 that contains a pipeline named PL1.PL1 includes a tumbling window trigger.

You create five clones of PL1. You configure each clone pipeline to use a different data source.

You need to ensure that the execution schedules of the clone pipeline match the execution schedule of PL1. What should you do?

- A. Add a new trigger to each cloned pipeline
- B. Associate each cloned pipeline to an existing trigger.

- C. Create a tumbling window trigger dependency for the trigger of PL1.  
D. Modify the Concurrency setting of each pipeline.

Answer: B

NEW QUESTION 38

- (Exam Topic 3)

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

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

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

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

- A. Mastered  
B. Not Mastered

Answer: A

Explanation:

Graphical user interface, text, application Description automatically generated

Box 1: Yes

A cluster mode of 'High Concurrency' is selected, unlike all the others which are 'Standard'. This results in a worker type of Standard\_DS13\_v2.

Box 2: No

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

Box 3: Yes

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

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

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

NEW QUESTION 42

- (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, 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 46**

- (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 50**

- (Exam Topic 3)

You need to design a solution that will process streaming data from an Azure Event Hub and output the data to Azure Data Lake Storage. The solution must ensure that analysts can interactively query the streaming data. What should you use?

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

**Answer:** C

**Explanation:**

Apache Spark Structured Streaming is a fast, scalable, and fault-tolerant stream processing API. You can use it to perform analytics on your streaming data in near real-time.

With Structured Streaming, you can use SQL queries to process streaming data in the same way that you would process static data.

Azure Event Hubs is a scalable real-time data ingestion service that processes millions of data in a matter of seconds. It can receive large amounts of data from multiple sources and stream the prepared data to Azure Data Lake or Azure Blob storage.

Azure Event Hubs can be integrated with Spark Structured Streaming to perform the processing of messages in near real-time. You can query and analyze the processed data as it comes by using a Structured Streaming query and Spark SQL.

Reference:

<https://k21academy.com/microsoft-azure/data-engineer/structured-streaming-with-azure-event-hubs/>

**NEW QUESTION 53**

- (Exam Topic 3)

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

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

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

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

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

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

You need to create the Databricks clusters for the workloads.

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

Does this meet the goal?

- A. Yes
- B. No

**Answer:** A

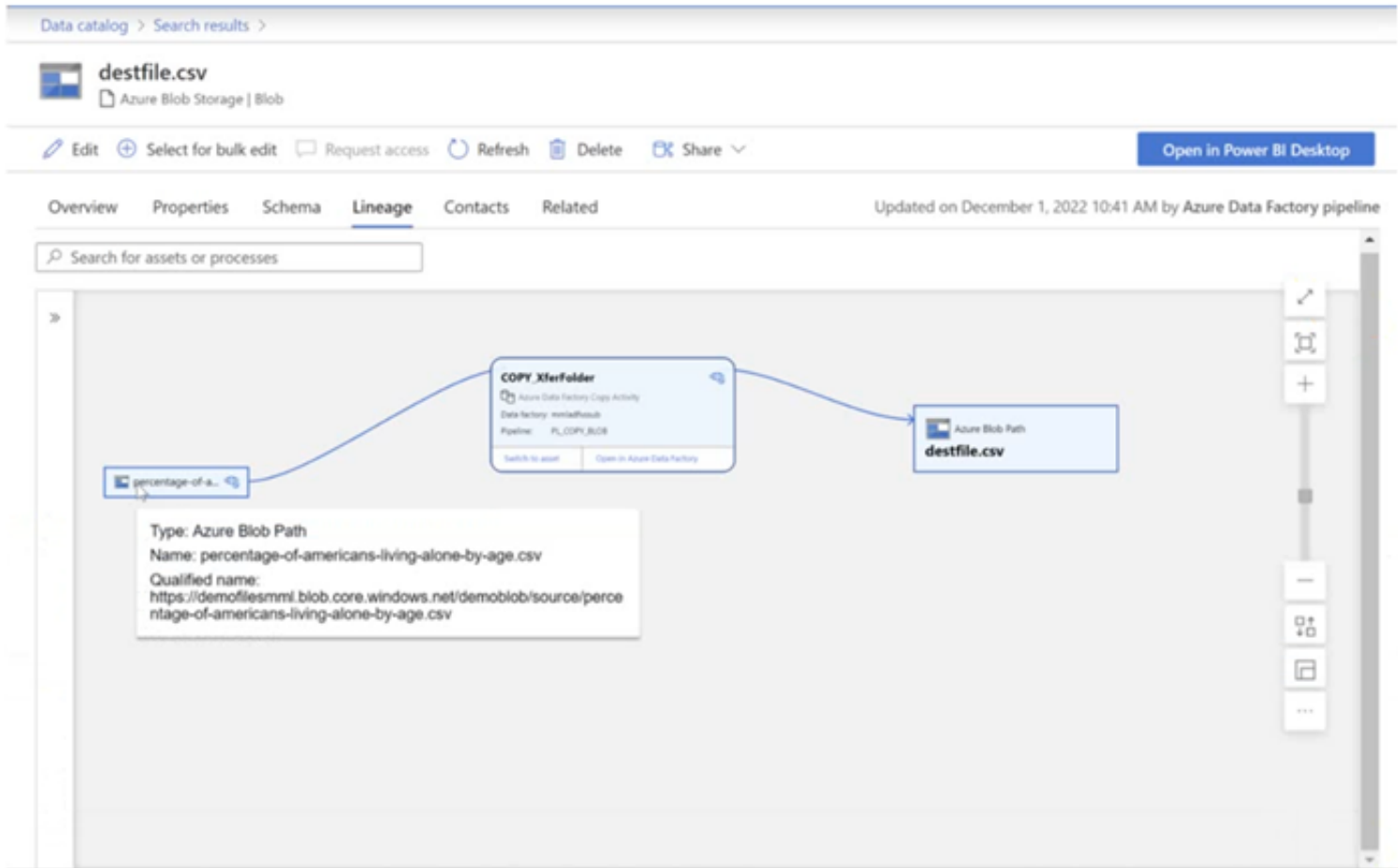
**Explanation:**

We need a High Concurrency cluster for the data engineers and the jobs. Note:

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

**NEW QUESTION 57**

- (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 guide<sup>1</sup>, data lineage in Microsoft Purview is a core platform capability that populates the Microsoft Purview Data Map with data movement and transformations across systems<sup>2</sup>. Lineage is captured as it flows in the enterprise and stitched without gaps irrespective of its source<sup>2</sup>.

**NEW QUESTION 61**

- (Exam Topic 3)  
 You are performing exploratory analysis of the bus fare data in an Azure Data Lake Storage Gen2 account by using an Azure Synapse Analytics serverless SQL pool.  
 You execute the Transact-SQL query shown in the following exhibit.

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

What do the query results include?

- A. Only CSV files in the tripdata\_2020 subfolder.
- B. All files that have file names that beginning with "tripdata\_2020".
- C. All CSV files that have file names that contain "tripdata\_2020".
- D. Only CSV that have file names that beginning with "tripdata\_2020".

**Answer: D**

**NEW QUESTION 65**

- (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 the Get-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 67**

- (Exam Topic 3)

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

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

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

You plan to copy the data from the storage account to an enterprise data warehouse in Azure Synapse Analytics.

You need to prepare the files to ensure that the data copies quickly. Solution: You convert the files to compressed delimited text files. Does this meet the goal?

- A. Yes
- B. No

**Answer:** A

**Explanation:**

All file formats have different performance characteristics. For the fastest load, use compressed delimited text files.

Reference:

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

**NEW QUESTION 68**

- (Exam Topic 3)

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

\* The output data will contain items purchased, quantity, line total sales amount, and line total tax amount.

\* Line total sales amount and line total tax amount will be aggregated in Databricks.

\* Sales transactions will never be updated. Instead, new rows will be added to adjust a sale.

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

What should you recommend?

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

**Answer:** B

**Explanation:**

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

**NEW QUESTION 71**

- (Exam Topic 3)

You have an enterprise data warehouse in Azure Synapse Analytics that contains a table named FactOnlineSales. The table contains data from the start of 2009 to the end of 2012.

You need to improve the performance of queries against FactOnlineSales by using table partitions. The solution must meet the following requirements:

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

How should you complete the T-SQL command? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

```
CREATE TABLE [dbo].[FactOnlineSales]
([OnlineSalesKey] [int] NOT NULL,
[OrderDateKey] [datetime] NOT NULL,
[StoreKey] [int] NOT NULL,
[ProductKey] [int] NOT NULL,
[CustomerKey] [int] NOT NULL,
[SalesOrderNumber] [varchar](20) NOT NULL,
[SalesQuantity] [int] NOT NULL,
[SalesAmount] [money] NOT NULL,
[UnitPrice] [money] NULL)
WITH (CLUSTERED COLUMNSTORE INDEX)
PARTITION ([OrderDateKey] RANGE 

RIGHT



LEFT

 FOR VALUES
( 

20090101,20121231



20100101,20110101,20120101



20090101,20100101,20110101,20120101

 )
```

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

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

NEW QUESTION 76

- (Exam Topic 3)  
You have an Azure subscription that contains an Azure Synapse Analytics workspace named workspace1. Workspace1 contains a dedicated SQL pool named SQL Pool and an Apache Spark pool named sparkpool. Sparkpool1 contains a DataFrame named pyspark.df.  
You need to write the contents of pyspark\_df to a tabte in SQLPooM by using a PySpark notebook. How should you complete the code? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

Answer Area

```
pyspark_df.createOrReplaceTempView("pysparkdftemptable")


%%local



%%spark



%%sql

 park.sqlContext.sql ("select * from pysparkdftemptable")
 ("sqlpool1.dbo.PySparkTable", Constants.INTERNAL)


jdbc



saveAsTable



synapsesql


```

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Answer Area

```
pyspark_df.createOrReplaceTempView("pysparkdftemptable")


%%local



%%spark



%%sql

 park.sqlContext.sql ("select * from pysparkdftemptable")
 ("sqlpool1.dbo.PySparkTable", Constants.INTERNAL)


jdbc



saveAsTable



synapsesql


```



**NEW QUESTION 80**

- (Exam Topic 3)

You use Azure Stream Analytics to receive Twitter data from Azure Event Hubs and to output the data to an Azure Blob storage account. You need to output the count of tweets from the last five minutes every minute. Which windowing function should you use?

- A. Sliding
- B. Session
- C. Tumbling
- D. Hopping

**Answer:** D

**Explanation:**

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

Reference:

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

**NEW QUESTION 85**

- (Exam Topic 3)

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

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

Which type of data redundancy should you use?

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

**Answer:** B

**Explanation:**

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

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

Reference:

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

**NEW QUESTION 86**

- (Exam Topic 3)

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

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

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

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

**Answer:** A

**Explanation:**

Each partition should have around 1 millions records. Dedicated SQL pools already have 60 partitions. We have the formula:  $\text{Records}/(\text{Partitions} \times 60) = 1 \text{ million}$   
 $\text{Partitions} = \text{Records}/(1 \text{ million} \times 60)$

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

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

Reference:

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

**NEW QUESTION 88**

- (Exam Topic 3)

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

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

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

NOTE: Each correct selection is worth one point.

**Answer Area**

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

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

- A. Mastered  
B. Not Mastered

**Answer: A**

**Explanation:**

**Answer Area**

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

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

**NEW QUESTION 89**

- (Exam Topic 3)

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

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

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

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

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

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

You need to create the Databricks clusters for the workloads.

Solution: You create a 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 94**

- (Exam Topic 3)

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

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

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

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

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

- The data engineers must share a cluster.
- The job cluster will be managed by using a request process whereby data scientists and data engineers provide packaged notebooks for deployment to the

cluster.

➤ All the data scientists must be assigned their own cluster that terminates automatically after 120 minutes of inactivity. Currently, there are three data scientists. You need to create the Databricks clusters for the workloads.

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

- A. Yes
- B. No

**Answer: B**

**Explanation:**

We would need a High Concurrency cluster for the jobs. Note:

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

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

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

**NEW QUESTION 96**

- (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?

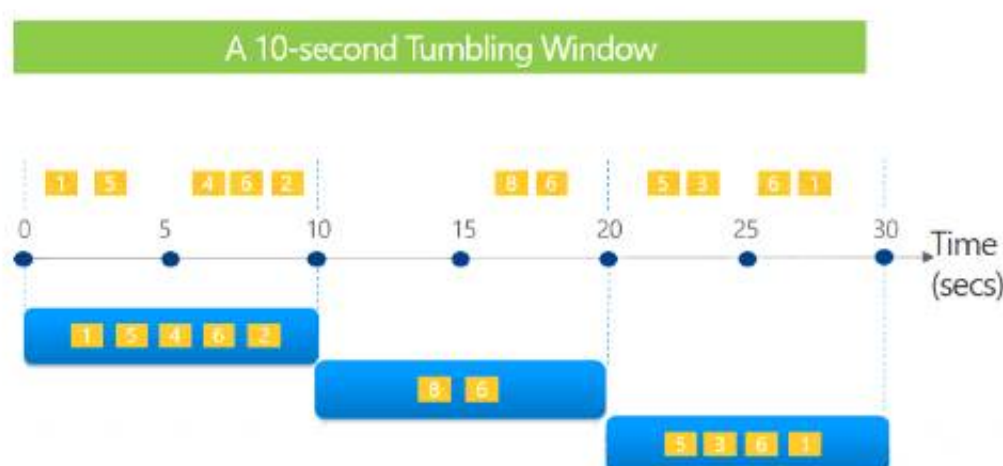
- A. snapshot
- B. tumbling
- C. hopping
- D. sliding

**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.

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 97**

- (Exam Topic 3)

You are monitoring an Azure Stream Analytics job.

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

What should you do?

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

**Answer: B**

**Explanation:**

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

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

Reference:

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

### NEW QUESTION 99

- (Exam Topic 3)

You have an Azure Synapse Analytics serverless SQ1 pool.

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

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

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

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

NOTE: Each correct selection is worth one point.

A. Mastered

B. Not Mastered

**Answer: A**

**Explanation:**

### NEW QUESTION 102

- (Exam Topic 3)

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

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

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

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

Which table distribution will minimize query times?

A. round-robin



- B. replicated
- C. hash-distributed on DateKey
- D. hash-distributed on PurchaseKey

**Answer:** D

**Explanation:**

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

Reference:

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

**NEW QUESTION 103**

- (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 be 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 106**

- (Exam Topic 3)

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

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

**Answer:** B

**Explanation:**

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

%python //or r, scala, sql Reference:

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

**NEW QUESTION 111**

- (Exam Topic 3)

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

After you answer a question in this scenario, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.  
You have an Azure Storage account that contains 100 GB of files. The files contain text and numerical values. 75% of the rows contain description data that has an average length of 1.1 MB.  
You plan to copy the data from the storage account to an Azure SQL data warehouse. You need to prepare the files to ensure that the data copies quickly.  
Solution: You modify the files to ensure that each row is less than 1 MB. Does this meet the goal?

- A. Yes
- B. No

**Answer:** A

**Explanation:**

When exporting data into an ORC File Format, you might get Java out-of-memory errors when there are large text columns. To work around this limitation, export only a subset of the columns.

References:

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

**NEW QUESTION 114**

- (Exam Topic 3)

You have an Azure Synapse Analytics dedicated SQL pool named Pool1 that contains an external table named Sales. Sales contains sales data. Each row in Sales

contains data on a single sale, including the name of the salesperson.

You need to implement row-level security (RLS). The solution must ensure that the salespeople can access only their respective sales.

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

- A. Mastered
- B. Not Mastered

**Answer:** A

**Explanation:**

Box 1: A security policy for sale

Here are the steps to create a security policy for Sales:

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

By creating a security policy for the Sales table, you ensure that each salesperson can only access their own sales data. The security policy uses a user-defined function to get the name of the current user and a security predicate function to filter the Sales table based on the current user.

Box 2: table-value function

to restrict row access by using row-level security, you need to create a table-valued function that returns a table of values that represent the rows that a user can access. You then use this function in a security policy that applies a predicate on the table.

**NEW QUESTION 118**

- (Exam Topic 3)

You use Azure Data Lake Storage Gen2.

You need to ensure that workloads can use filter predicates and column projections to filter data at the time the data is read from disk.

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

- A. Reregister the Microsoft Data Lake Store resource provider.
- B. Reregister the Azure Storage resource provider.
- C. Create a storage policy that is scoped to a container.
- D. Register the query acceleration feature.
- E. Create a storage policy that is scoped to a container prefix filter.

**Answer:** BD

**NEW QUESTION 120**

- (Exam Topic 3)

You have an Azure subscription that contains a Microsoft Purview account named MP1, an Azure data factory named DF1, and a storage account named storage. MP1 is configured

10 scan storage1. DF1 is connected to MP1 and contains 3 dataset named DS1. DS1 references 2 file in storage.

In DF1, you plan to create a pipeline that will process data from DS1.

You need to review the schema and lineage information in MP1 for the data referenced by DS1.

Which two features can you use to locate the information? Each correct answer presents a complete solution. NOTE: Each correct answer is worth one point.

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

**Answer:** CD

**Explanation:**

➤ The search bar in the Microsoft Purview governance portal: This is a feature that allows you to search for assets in your data estate using keywords, filters, and facets. You can use the search bar to find the files in storage1 that are referenced by DS1, and then view their schema and lineage information in the asset details page12.

➤ The search bar in Azure Data Factory Studio: This is a feature that allows you to search for datasets, linked services, pipelines, and other resources in your data factory. You can use the search bar to find DS1 in DF1, and then view its schema and lineage information in the dataset details page. You can also click on the Open in Purview button to open the corresponding asset in MP13.

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

The search bar in Azure Data Factory Studio allows you to search for the dataset DS1 and view its properties and lineage. This can help you locate information about the source and destination data stores, as well as the transformations that were applied to the data.

The search bar in the Microsoft Purview governance portal allows you to search for the storage account and view its metadata, including schema and lineage information. This can help you understand the different data assets that are stored in the storage account and how they are related to each other.

The Storage browser of storage1 in the Azure portal may allow you to view the files that are stored in the storage account, but it does not provide lineage or schema information for those files. Similarly, the search bar in the Azure portal may allow you to search for resources in the Azure subscription, but it does not provide detailed information about the data assets themselves.

References:

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

**NEW QUESTION 122**

- (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 125**

- (Exam Topic 1)

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

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

## Commands

## Answer Area

CREATE EXTERNAL DATA SOURCE

CREATE EXTERNAL FILE FORMAT

CREATE EXTERNAL TABLE

CREATE EXTERNAL TABLE AS SELECT

CREATE DATABASE SCOPED CREDENTIAL

- A. Mastered
- B. Not Mastered

**Answer:** A

### Explanation:

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

Box 1: CREATE EXTERNAL DATA SOURCE

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

CREATE EXTERNAL FILE FORMAT creates an external file format object that defines external data stored in Azure Blob Storage or Azure Data Lake Storage.

Creating an external file format is a prerequisite for creating an external table.

Box 3: CREATE EXTERNAL TABLE AS SELECT

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

Reference:

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

### NEW QUESTION 129

- (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 133

- (Exam Topic 1)

You need to implement versioned changes to the integration pipelines. The solution must meet the data integration requirements.

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

#### Actions

Publish changes.

Create a feature branch.

Merge changes.

Create a repository and a main branch.

Create a pull request.

#### Answer Area



- A. Mastered
- B. Not Mastered

**Answer:** A

### Explanation:

Graphical user interface, application Description automatically generated



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

Step 1: Create a repository and a main branch

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

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

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

Reference:

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

#### NEW QUESTION 138

- (Exam Topic 1)

You need to design a data retention solution for the Twitter feed data records. The solution must meet the customer sentiment analytics requirements.

Which Azure Storage functionality should you include in the solution?

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

**Answer: D**

#### NEW QUESTION 142

- (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 144

- (Exam Topic 1)

You need to design a data retention solution for the Twitter feed data records. The solution must meet the customer sentiment analytics requirements.

Which Azure Storage functionality should you include in the solution?

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

**Answer: D**

#### Explanation:

Scenario: Purge Twitter feed data records that are older than two years.

Data sets have unique lifecycles. Early in the lifecycle, people access some data often. But the need for access often drops drastically as the data ages. Some data remains idle in the cloud and is rarely accessed once stored. Some data sets expire days or months after creation, while other data sets are actively read and modified throughout their lifetimes. Azure Storage lifecycle management offers a rule-based policy that you can use to transition blob data to the appropriate access tiers or to expire data at the end of the data lifecycle.

Reference:

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

**NEW QUESTION 149**

- (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 152**

- (Exam Topic 3)

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

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

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

You plan to copy the data from the storage account to an enterprise data warehouse in Azure Synapse Analytics.

You need to prepare the files to ensure that the data copies quickly. Solution: You copy the files to a table that has a columnstore index. Does this meet the goal?

- A. Yes
- B. No

**Answer:** B

**Explanation:**

Instead convert the files to compressed delimited text files. Reference:

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

**NEW QUESTION 153**

- (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	ADFDployDemo
Repository name	ADEDployDemo
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

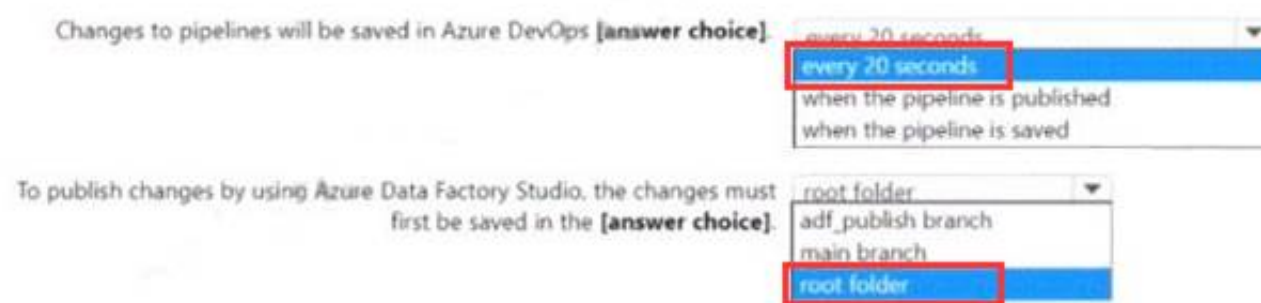
root folder

- A. Mastered  
B. Not Mastered

**Answer:** A

**Explanation:**

Answer Area



## NEW QUESTION 157

- (Exam Topic 3)

You are implementing an Azure Stream Analytics solution to process event data from devices.

The devices output events when there is a fault and emit a repeat of the event every five seconds until the fault is resolved. The devices output a heartbeat event every five seconds after a previous event if there are no faults present.

A sample of the events is shown in the following table.

DeviceID	EventType	EventTime
78cc5ht9-w357-684r-w4fr-kr16h6p9874e	HeartBeat	2020-12-01T19:00.000Z
78cc5ht9-w357-684r-w4fr-kr16h6p9874e	HeartBeat	2020-12-01T19:05.000Z
78cc5ht9-w357-684r-w4fr-kr16h6p9874e	TemperatureSensorFault	2020-12-01T19:07.000Z

You need to calculate the uptime between the faults.

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

NOTE: Each correct selection is worth one point.

```
SELECT
    DeviceID,
    MIN(EventTime) as StartTime,
    MAX(EventTime) as EndTime,
    DATEDIFF(second, MIN(EventTime), MAX(EventTime)) AS duration_in_seconds
FROM input TIMESTAMP BY EventTime
```

▼

WHERE EventType='HeartBeat'

WHERE LAG(EventType, 1) OVER (LIMIT DURATION(second,5)) <> EventType

WHERE IsFirst(second,5) = 1

GROUP BY

DeviceID

▼

,SessionWindow(second, 5, 50000) OVER (PARTITION BY DeviceID)

,TumblingWindow(second,5)

HAVING DATEDIFF(second, MIN(EventTime), MAX(EventTime)) > 5

- A. Mastered  
B. Not Mastered

**Answer:** A

**Explanation:**

Graphical user interface, text, application Description automatically generated

Box 1: WHERE EventType='HeartBeat' Box 2: ,TumblingWindow(Second, 5)

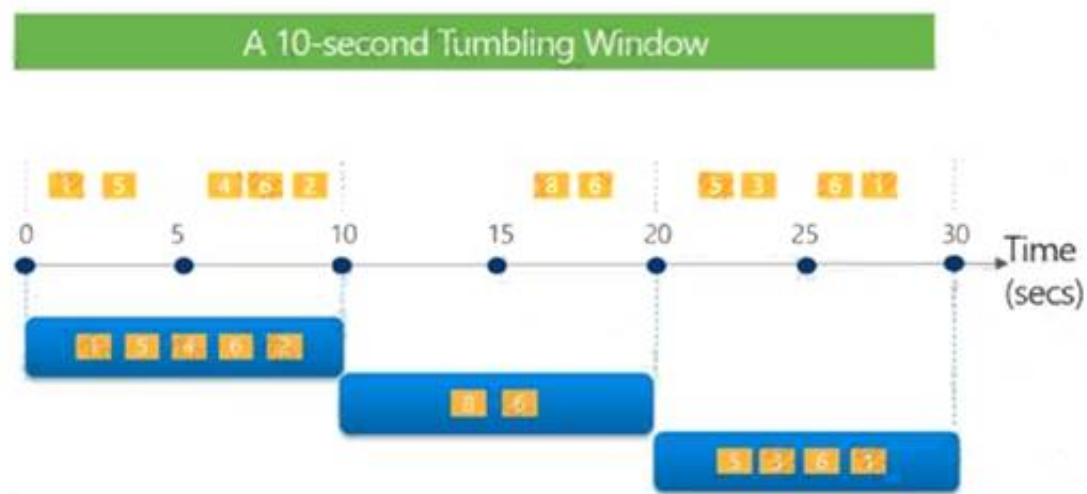
Tumbling windows are a series of fixed-sized, non-overlapping and contiguous time intervals.

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

#### NEW QUESTION 162

- (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 `abfs://` 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 167



- (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 171

- (Exam Topic 3)

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

A)

```
sys.pdw_nodes_column_store_segments
```

B)

```
sys.dm_db_column_store_row_group_operational_stats
```

C)

```
sys.pdw_nodes_column_store_row_groups
```

D)

```
sys.dm_db_column_store_row_group_physical_stats
```

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

**Answer: B**

#### NEW QUESTION 172

- (Exam Topic 3)

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

The screenshot shows the Azure Data Factory portal interface for configuring a Git repository. On the left, a navigation pane lists various settings, with 'Git configuration' highlighted. The main area, titled 'Git repository', provides details about the configured repository. It includes a 'Setting' button and a 'Disconnect' button. The configuration details are as follows:

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

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

NOTE: Each correct selection is worth one point.

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

	▼
/	
adf_publish	
main	
Parameterization template	

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

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

- A. Mastered
- B. Not Mastered

**Answer:** A

**Explanation:**

Letter Description automatically generated

Box 1: adf\_publish

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

Box 2: / dwh\_batchetl/adf\_publish/contososales

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

Reference:

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

**NEW QUESTION 173**

- (Exam Topic 3)

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

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

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

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

**Answer:** BD

**Explanation:**

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

**NEW QUESTION 177**

- (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**

**Answer Area**

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.



- 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 178**

- (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 181**

- (Exam Topic 3)

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

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

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

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

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

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

- A. Yes  
B. No

**Answer:** B



**Explanation:**

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

Reference:

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

**NEW QUESTION 182**

- (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 183**

- (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,
    [Postalcode] [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 187**

- (Exam Topic 3)  
You have an Azure subscription that contains an Azure Databricks workspace. The workspace contains a notebook named Notebook1. In Notebook1, you create an Apache Spark DataFrame named df\_sales that contains the following columns:

- Customer
- Salesperson
- Region
- Amount

You need to identify the three top performing salespersons by amount for a region named HQ.  
How should you complete the query? To answer, drag the appropriate values to the correct targets. Each value may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.

Values

agg(col('SalesPerson'))

filter(col('SalesPerson'))

groupBy(col('SalesPerson'))

groupBy(col('TotalAmount'))

orderBy(col('TotalAmount'))

orderBy(desc('TotalAmount'))

Answer Area

```
df_sales.filter(col('Region')=='HQ').  
  
    .agg(sum('Amount').alias  
    ('TotalAmount')).  
  
    
```

- A. Mastered
- B. Not Mastered

**Answer:** A

**Explanation:**

Values

agg(col('SalesPerson'))

filter(col('SalesPerson'))

groupBy(col('SalesPerson'))

groupBy(col('TotalAmount'))

orderBy(col('TotalAmount'))

orderBy(desc('TotalAmount'))

Answer Area

```
df_sales.filter(col('Region')=='HQ').  
  
    .agg(sum('Amount').alias  
    ('TotalAmount')).  
  
    
```

**NEW QUESTION 192**

- (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 196**

- (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 199**

- (Exam Topic 3)

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

- TransactionType: 40 million rows per transaction type
- CustomerSegment: 4 million per customer segment
- TransactionMonth: 65 million rows per month
- AccountType: 500 million per account type

You have the following query requirements:

- Analysts will most commonly analyze transactions for a given month.
- Transactions analysis will typically summarize transactions by transaction type, customer segment, and/or account type

You need to recommend a partition strategy for the table to minimize query times. On which column should you recommend partitioning the table?

- A. CustomerSegment
- B. AccountType
- C. TransactionType
- D. TransactionMonth

**Answer:** C

**Explanation:**

For optimal compression and performance of clustered columnstore tables, a minimum of 1 million rows per distribution and partition is needed. Before partitions are created, dedicated SQL pool already divides each table into 60 distributed databases.

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

**NEW QUESTION 203**

- (Exam Topic 3)

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

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

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

NOTE: Each correct selection is worth one point.

EventCategory:  ▼

DimChannel
DimDate
DimEvent
FactEvents

ChannelGrouping:  ▼

DimChannel
DimDate
DimEvent
FactEvents

TotalEvents:  ▼

DimChannel
DimDate
DimEvent
FactEvents

- A. Mastered
- B. Not Mastered

**Answer:** A

**Explanation:**

Table Description automatically generated

Box 1: DimEvent

Box 2: DimChannel

Box 3: FactEvents

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

Reference:

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

**NEW QUESTION 206**

- (Exam Topic 3)

You are designing a slowly changing dimension (SCD) for supplier data in an Azure Synapse Analytics dedicated SQL pool.

You plan to keep a record of changes to the available fields. The supplier data contains the following columns.

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

Which three additional columns should you add to the data to create a Type 2 SCD? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

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

**Answer:** CDF

**Explanation:**

Reference:

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

**NEW QUESTION 211**

- (Exam Topic 3)

You have two Azure Data Factory instances named ADFdev and ADFprod. ADFdev connects to an Azure DevOps Git repository. You publish changes from the main branch of the Git repository to ADFdev. You need to deploy the artifacts from ADFdev to ADFprod. What should you do first?

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

**Answer: C**

**Explanation:**

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

Note:

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

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

- Select the Empty job template. Reference:

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

**NEW QUESTION 214**

- (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 218**

- (Exam Topic 3)

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

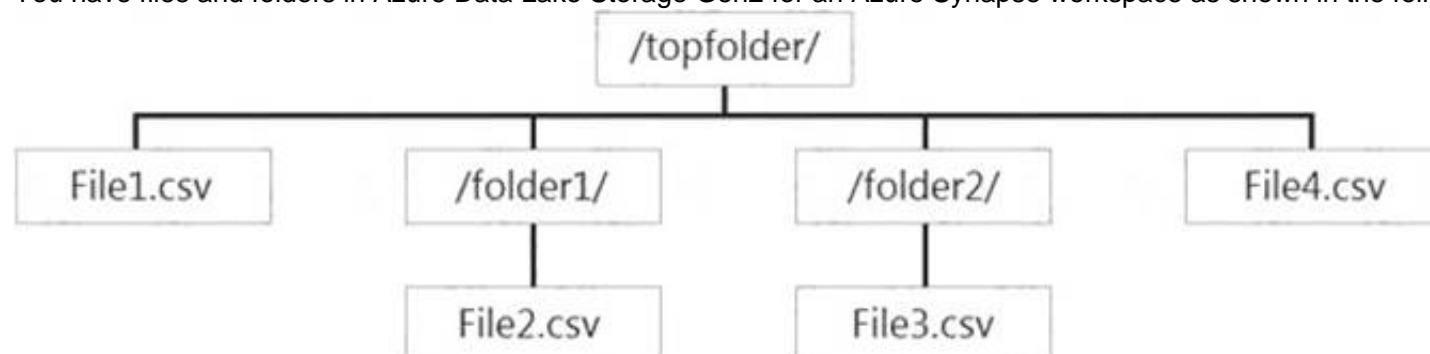
- A. JOIN
- B. WHERE
- C. DISTINCT
- D. GROUP BY

**Answer: B**

**NEW QUESTION 220**

- (Exam Topic 3)

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



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

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



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

Answer: B

Explanation:

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

NEW QUESTION 224

- (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

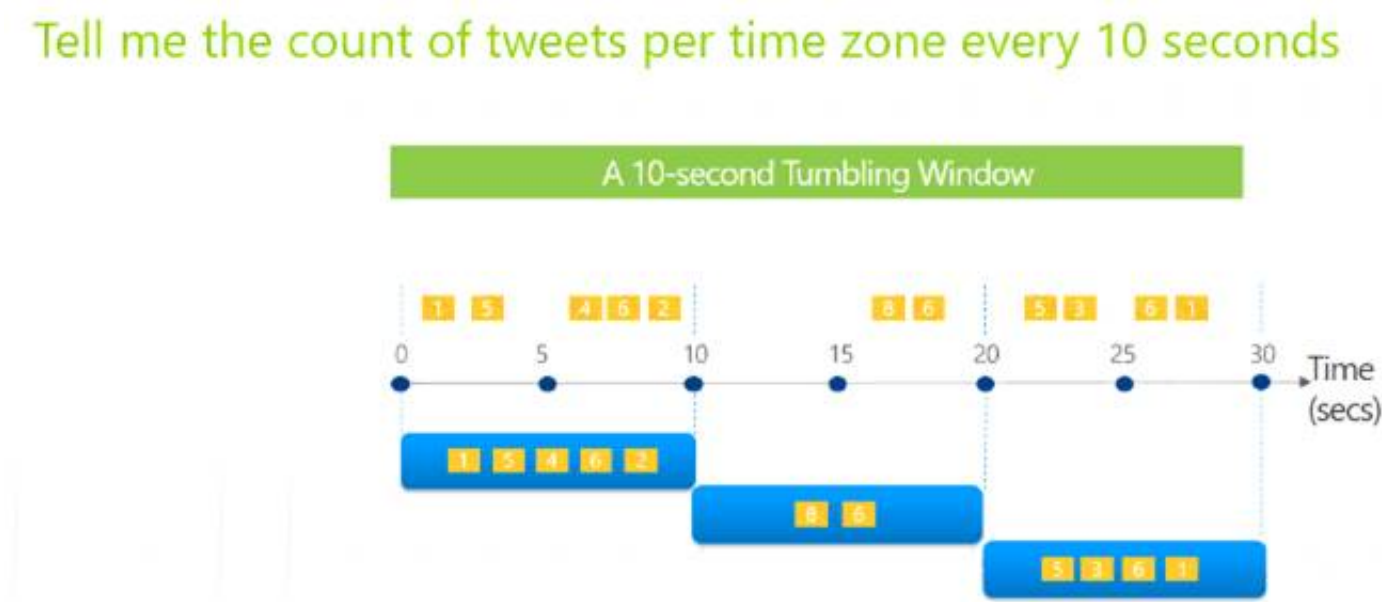
Tumbling window

- 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



```
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 226

- (Exam Topic 3)

You are batch loading a table in an Azure Synapse Analytics dedicated SQL pool. You need to load data from a staging table to the target table. The solution must ensure that if an error occurs while loading the data to the target table, all the inserts in that batch are undone. How should you complete the Transact-SQL code? To answer, drag the appropriate values to the correct

targets. Each value may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.  
NOTE Each correct selection is worth one point.

Values

BEGIN DISTRIBUTED TRANSACTION

BEGIN TRAN

COMMIT TRAN

ROLLBACK TRAN

SET RESULT\_SET\_CACHING ON

Answer Area

BEGIN TRY

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

SELECT col1, col2, col3 FROM stage.Table1;

END TRY

BEGIN CATCH

IF @@TRANCOUNT > 0

BEGIN

END

END CATCH;

IF @@TRANCOUNT >0

BEGIN

COMMIT TRAN;

END

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:  
Values

Values

BEGIN DISTRIBUTED TRANSACTION

BEGIN TRAN

COMMIT TRAN

ROLLBACK TRAN

SET RESULT\_SET\_CACHING ON

Answer Area

BEGIN TRAN

BEGIN TRY

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

SELECT col1, col2, col3 FROM stage.Table1;

END TRY

BEGIN CATCH

IF @@TRANCOUNT > 0

BEGIN

ROLLBACK TRAN

END

END CATCH;

IF @@TRANCOUNT >0

BEGIN

COMMIT TRAN;

END

NEW QUESTION 227

- (Exam Topic 3)


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

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

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

NOTE: Each correct selection is worth one point.

```
CREATE EXTERNAL DATA SOURCE source1
WITH
  ( LOCATION = 'https://account1.
  (
    PUSHDOWN = ON
    TYPE = BLOB_STORAGE
    TYPE = HADOOP
  )
```



- A. Mastered
- B. Not Mastered

**Answer:** A

**Explanation:**

Graphical user interface, diagram Description automatically generated

Box 1: blob

The following example creates an external data source for Azure Data Lake Gen2 CREATE EXTERNAL DATA SOURCE YellowTaxi

WITH ( LOCATION = 'https://azureopendatastorage.blob.core.windows.net/nyctlc/yellow/', TYPE = HADOOP)

Box 2: HADOOP

Reference:

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

**NEW QUESTION 230**

- (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 235**

- (Exam Topic 3)

You are developing an Azure Synapse Analytics pipeline that will include a mapping data flow named Dataflow1. Dataflow1 will read customer data from an external source and use a Type 1 slowly changing dimension (SCO) when loading the data into a table named DimCustomer1 in an Azure Synapse Analytics dedicated SQL pool.

You need to ensure that Dataflow1 can perform the following tasks:

\* Detect whether the data of a given customer has changed in the DimCustomer table.

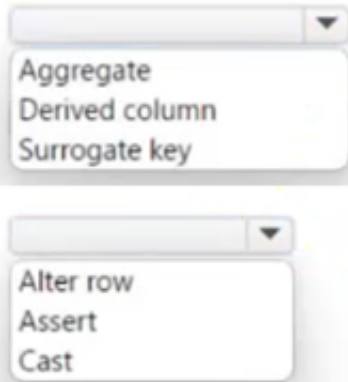
• Perform an upsert to the DimCustomer table.

Which type of transformation should you use for each task? To answer, select the appropriate options in the answer area

NOTE; Each correct selection is worth one point.

**Answer Area**

Detect whether the data of a given customer has changed in the DimCustomer table:



Perform an upsert to the DimCustomer table:

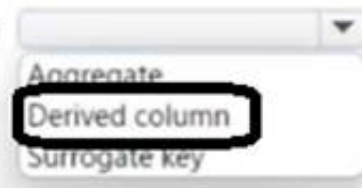
- A. Mastered
- B. Not Mastered

**Answer:** A

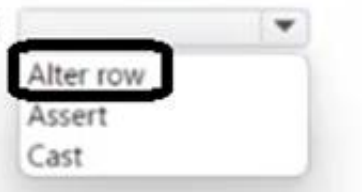
**Explanation:**

**Answer Area**

Detect whether the data of a given customer has changed in the DimCustomer table:



Perform an upsert to the DimCustomer table:



**NEW QUESTION 237**

- (Exam Topic 3)

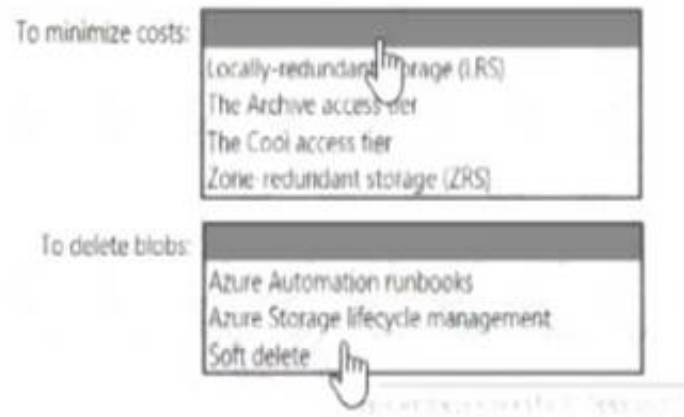
You have an Azure subscription.

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 241**

- (Exam Topic 3)

You plan to ingest streaming social media data by using Azure Stream Analytics. The data will be stored in files in Azure Data Lake Storage, and then consumed by using Azure Databricks and PolyBase in Azure Synapse Analytics.

You need to recommend a Stream Analytics data output format to ensure that the queries from Databricks and PolyBase against the files encounter the fewest possible errors. The solution must ensure that the tiles can be queried quickly and that the data type information is retained.

What should you recommend?

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

**Answer: A**

**Explanation:**

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

**NEW QUESTION 245**

- (Exam Topic 3)

You have an Azure Data Factory version 2 (V2) resource named Df1. Df1 contains a linked service. You have an Azure Key vault named vault1 that contains an encryption key named key1.

You need to encrypt Df1 by using key1. What should you do first?

- A. Add a private endpoint connection to vault 1.
- B. Enable Azure role-based access control on vault 1.
- C. Remove the linked service from Df1.
- D. Create a self-hosted integration runtime.

**Answer: C**



**Explanation:**

Linked services are much like connection strings, which define the connection information needed for Data Factory to connect to external resources.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/enable-customer-managed-key> <https://docs.microsoft.com/en-us/azure/data-factory/concepts-linked-services>

<https://docs.microsoft.com/en-us/azure/data-factory/create-self-hosted-integration-runtime>

**NEW QUESTION 247**

- (Exam Topic 3)

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

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

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

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

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

NOTE: Each correct selection is worth one point.

Values

alias

array\_union

createDataFrame

explode

select

translate

Answer Area

dbutils.fs.put("/tmp/source.json", source\_json, True)

source\_df = spark.read.option("multiline", "true").json("/tmp/source.json")

persons = source\_df. 

Value

Value

 ("persons").alias("persons")

persons\_dogs = persons.select(col("persons.name").alias("owner"), col("persons.age").alias("age"),

explode 

Value

 ("dog"))

("persons.dogs").

display(persons\_dogs)

- A. Mastered
- B. Not Mastered

**Answer:** A

**Explanation:**

Graphical user interface, text, application Description automatically generated

Box 1: select

Box 2: explode

Box 3: alias

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

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

**NEW QUESTION 249**

- (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 252**

- (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 253**

- (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 Pool1. Pool1 contains a partitioned fact table named dbo.Sales and a staging table named stg.Sales that has the matching table and partition definitions.

You need to overwrite the content of the first partition in dbo.Sales with the content of the same partition in stg.Sales. The solution must minimize load times. What should you do?

- A. Switch the first partition from dbo.Sales to stg.Sales.
- B. Switch the first partition from stg.Sales to db
- C. Sales.
- D. Update dbo.Sales from stg.Sales.
- E. Insert the data from stg.Sales into dbo.Sales.

**Answer:** A

**NEW QUESTION 258**

- (Exam Topic 3)

You have a data warehouse in Azure Synapse Analytics.

You need to ensure that the data in the data warehouse is encrypted at rest. What should you enable?

- A. Advanced Data Security for this database
- B. Transparent Data Encryption (TDE)
- C. Secure transfer required
- D. Dynamic Data Masking

**Answer:** B

**Explanation:**

Azure SQL Database currently supports encryption at rest for Microsoft-managed service side and client-side encryption scenarios.

➤ Support for server encryption is currently provided through the SQL feature called Transparent Data Encryption.

➤ Client-side encryption of Azure SQL Database data is supported through the Always Encrypted feature. Reference:

<https://docs.microsoft.com/en-us/azure/security/fundamentals/encryption-atrest>

**NEW QUESTION 263**

- (Exam Topic 3)

You are planning the deployment of Azure Data Lake Storage Gen2. You have the following two reports that will access the data lake:

➤ Report1: Reads three columns from a file that contains 50 columns.

➤ Report2: Queries a single record based on a timestamp.

You need to recommend in which format to store the data in the data lake to support the reports. The solution must minimize read times.

What should you recommend for each report? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

Report1:	<div><div></div><div>Avro</div><div>CSV</div><div>Parquet</div><div>TSV</div></div>
Report2:	<div><div></div><div>Avro</div><div>CSV</div><div>Parquet</div><div>TSV</div></div>

- A. Mastered
- B. Not Mastered

**Answer:** A

**Explanation:**

Report1: CSV

CSV: The destination writes records as delimited data. Report2: AVRO

AVRO supports timestamps.

Not Parquet, TSV: Not options for Azure Data Lake Storage Gen2. Reference:

<https://streamsets.com/documentation/datacollector/latest/help/datacollector/UserGuide/Destinations/ADLS-G2>

**NEW QUESTION 267**

- (Exam Topic 3)

You need to trigger an Azure Data Factory pipeline when a file arrives in an Azure Data Lake Storage Gen2 container. Which resource provider should you enable?

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

**Answer: C**

**Explanation:**

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

Reference:

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

**NEW QUESTION 271**

- (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	Answer Area
CustomerKey	<pre>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)                 ) )</pre>
HASH	
ROUND_ROBIN	
REPLICATE	
OrderDateKey	
SalesOrderNumber	

- 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 276**

- (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 278**

- (Exam Topic 3)

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

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

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

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

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

A. Yes

B. No

**Answer:** A

**NEW QUESTION 279**

- (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 280**

- (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 281**

- (Exam Topic 3)

You are designing an Azure Synapse Analytics dedicated SQL pool.

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

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

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

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

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

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

You are designing dynamic data masking to maintain compliance.

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

NOTE: Each correct selection is worth one point.

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

- A. Mastered
- B. Not Mastered

Answer: A

**Explanation:**  
Text Description automatically generated  
Reference:  
<https://docs.microsoft.com/en-us/azure/azure-sql/database/dynamic-data-masking-overview>

NEW QUESTION 285

- (Exam Topic 3)  
You are designing an Azure Data Lake Storage Gen2 structure for telemetry data from 25 million devices distributed across seven key geographical regions. Each minute, the devices will send a JSON payload of metrics to Azure Event Hubs.  
You need to recommend a folder structure for the data. The solution must meet the following requirements:

- > Data engineers from each region must be able to build their own pipelines for the data of their respective region only.
- > The data must be processed at least once every 15 minutes for inclusion in Azure Synapse Analytics serverless SQL pools.

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

Values

Answer Area

{deviceID}

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

{regionID}/{deviceID}

{regionID}/raw

{YYYY}/{MM}/{DD}/{HH}

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

raw/{deviceID}

raw/{regionID}

/

Value

/

Value

/

Value

.json

- A. Mastered
- B. Not Mastered

Answer: A

**Explanation:**  
Box 1: {YYYY}/{MM}/{DD}/{HH}  
Date Format [optional]: if the date token is used in the prefix path, you can select the date format in which your files are organized. Example: YYYY/MM/DD  
Time Format [optional]: if the time token is used in the prefix path, specify the time format in which your files are organized. Currently the only supported value is HH.  
Box 2: {regionID}/raw  
Data engineers from each region must be able to build their own pipelines for the data of their respective region only.  
Box 3: {deviceID} Reference:  
<https://github.com/paolosalvatori/StreamAnalyticsAzureDataLakeStore/blob/master/README.md>

NEW QUESTION 289

.....

## Thank You for Trying Our Product

\* 100% Pass or Money Back

All our products come with a 90-day Money Back Guarantee.

\* One year free update

You can enjoy free update one year. 24x7 online support.

\* Trusted by Millions

We currently serve more than 30,000,000 customers.

\* Shop Securely

All transactions are protected by VeriSign!

**100% Pass Your DP-203 Exam with Our Prep Materials Via below:**

<https://www.certleader.com/DP-203-dumps.html>