

# Microsoft

## Exam Questions DP-203

Data Engineering on Microsoft Azure



### 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

Answer Area

Path pattern:

Date format:

### NEW QUESTION 2

- (Exam Topic 3)

You are designing a data mart for the human resources (MR) department at your company. The data mart will contain information and employee transactions. From a source system you have a flat extract that has the following fields:

- EmployeeID
- FirstName
- LastName
- Recipient
- GrossAmount
- TransactionID
- GovernmentID
- NetAmountPaid
- TransactionDate

You need to design a start schema data model in an Azure Synapse analytics dedicated SQL pool for the data mart. Which two tables should you create? Each Correct answer present part of the solution.

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

**Answer:** AE

#### Explanation:

Reference:

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

### NEW QUESTION 3

- (Exam Topic 3)

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

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

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

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

**Answer:** B

#### Explanation:

Hash-distributed tables improve query performance on large fact tables.

Columnstore indexes can achieve up to 100x better performance on analytics and data warehousing workloads and up to 10x better data compression than

traditional rowstore indexes.

Reference:

<https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-tables-distribute> <https://docs.microsoft.com/en-us/sql/relational-databases/indexes/columnstore-indexes-query-performance>

#### NEW QUESTION 4

- (Exam Topic 3)

You 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 5

- (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 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 have an Azure Data Lake Storage account that contains a staging zone.

You need to design a data process to ingest incremental data from the staging zone, transform the data by executing an R script, and then insert the transformed data into a data warehouse in Azure Synapse Analytics.

Solution: You use an Azure Data Factory schedule trigger to execute a pipeline that copies the data to a staging table in the data warehouse, and then uses a stored procedure to execute the R script.

Does this meet the goal?

- A. Yes
- B. No

**Answer:** A

**Explanation:**

If you need to transform data in a way that is not supported by Data Factory, you can create a custom activity with your own data processing logic and use the activity in the pipeline.

Note: You can use data transformation activities in Azure Data Factory and Synapse pipelines to transform and process your raw data into predictions and insights at scale.

Reference:

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

**NEW QUESTION 8**

- (Exam Topic 3)

You are designing an Azure Databricks cluster that runs user-defined local processes. You need to recommend a cluster configuration that meets the following requirements:

- Minimize query latency.
- Maximize the number of users that can run queues on the cluster at the same time « Reduce overall costs without compromising other requirements

Which cluster type should you recommend?

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

**Answer:** C

**Explanation:**

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

Databricks chooses the appropriate number of workers required to run your job. This is referred to as autoscaling. Autoscaling makes it easier to achieve high cluster utilization, because you don't need to provision the cluster to match a workload.

Reference:

<https://docs.microsoft.com/en-us/azure/databricks/clusters/configure>

**NEW QUESTION 9**

- (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<br>Driver's license number     |
| HubA      | Ride route<br>Ride distance<br>Ride duration |
| HubB      | Ride fare<br>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 10

- (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 11

- (Exam Topic 3)

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

Which output mode should you use?

- A. complete



- B. update
- C. append

**Answer:** C

**Explanation:**

Append Mode: Only new rows appended in the result table since the last trigger are written to external storage. This is applicable only for the queries where existing rows in the Result Table are not expected to change.  
<https://docs.databricks.com/getting-started/spark/streaming.html>

**NEW QUESTION 14**

- (Exam Topic 3)

You have an Azure data solution that contains an enterprise data warehouse in Azure Synapse Analytics named DW1. Several users execute ad hoc queries to DW1 concurrently. You regularly perform automated data loads to DW1. You need to ensure that the automated data loads have enough memory available to complete quickly and successfully when the adhoc queries run. What should you do?

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

**Answer:** C

**Explanation:**

The performance capacity of a query is determined by the user's resource class. Resource classes are pre-determined resource limits in Synapse SQL pool that govern compute resources and concurrency for query execution. Resource classes can help you configure resources for your queries by setting limits on the number of queries that run concurrently and on the compute-resources assigned to each query. There's a trade-off between memory and concurrency. Smaller resource classes reduce the maximum memory per query, but increase concurrency. Larger resource classes increase the maximum memory per query, but reduce concurrency. Reference:  
<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/resource-classes-for-workload-ma>

**NEW QUESTION 17**

- (Exam Topic 3)

You use Azure Data Factory to prepare data to be queried by Azure Synapse Analytics serverless SQL pools. Files are initially ingested into an Azure Data Lake Storage Gen2 account as 10 small JSON files. Each file contains the same data attributes and data from a subsidiary of your company. You need to move the files to a different folder and transform the data to meet the following requirements: ➤ Provide the fastest possible query times.

➤ Automatically infer the schema from the underlying files.

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

NOTE: Each correct selection is worth one point.

Copy behavior:

|                    |
|--------------------|
| Flatten hierarchy  |
| Merge files        |
| Preserve hierarchy |

Sink file type:

|         |
|---------|
| CSV     |
| JSON    |
| Parquet |
| TXT     |

- A. Mastered
- B. Not Mastered

**Answer:** A

**Explanation:**

Box 1: Preserver herarchy

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

Box 2: Parquet

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

Reference:

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

**NEW QUESTION 22**

- (Exam Topic 3)

You are building an Azure Stream Analytics job to retrieve game data.

You need to ensure that the job returns the highest scoring record for each five-minute time interval of each game.  
How should you complete the Stream Analytics query? To answer, select the appropriate options in the answer area.  
NOTE: Each correct selection is worth one point.

SELECT

Collect(Score)

CollectTop(1) OVER(ORDER BY Score Desc)

Game, MAX(Score)

TopOne() OVER(PARTITION BY Game ORDER BY Score Desc)

as HighestScore

FROM input

TIMESTAMP BY CreatedAt

GROUP BY

Game

Hopping(minute,5)

Tumbling(minute,5)

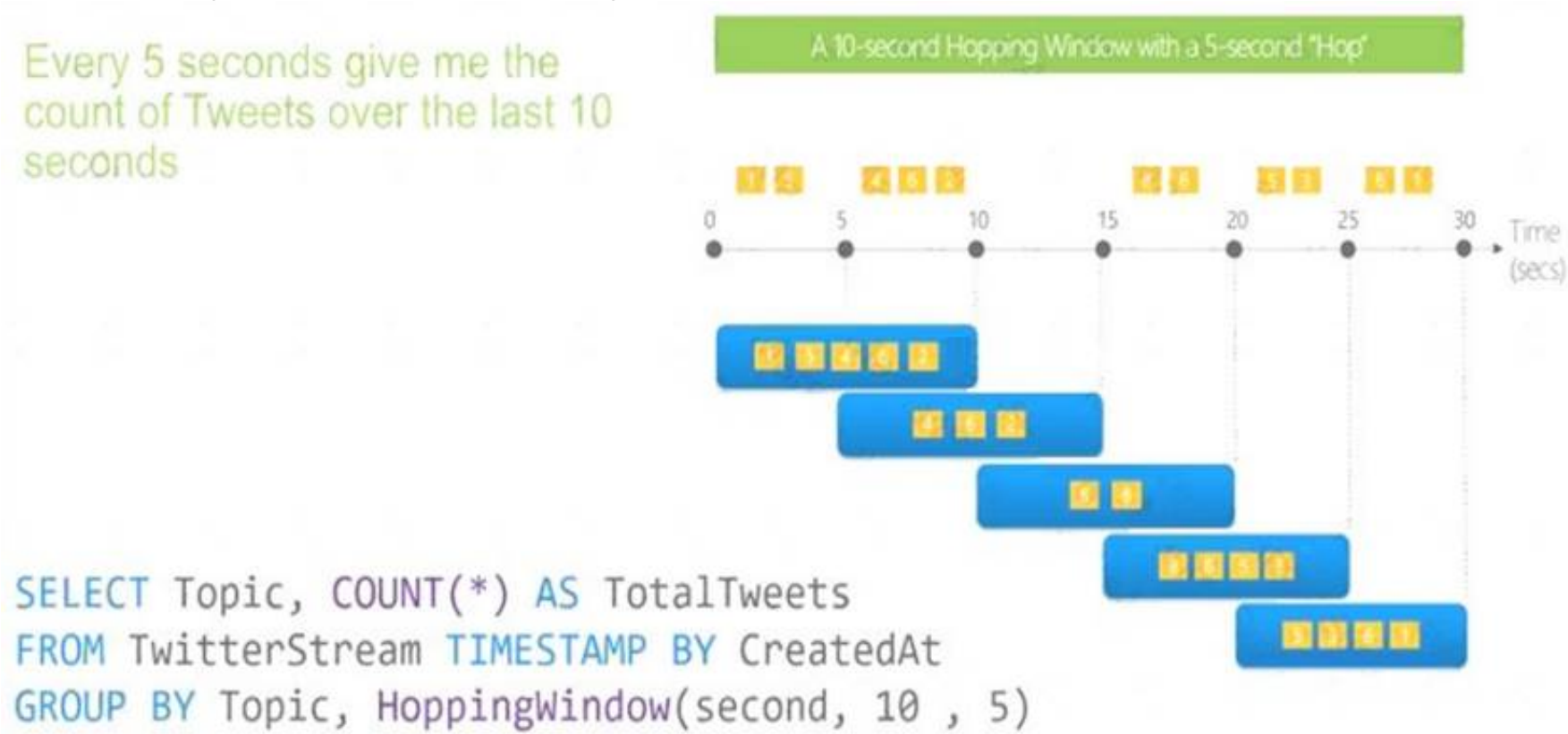
Windows(TumblingWindow(minute,5),Hopping(minute,5))

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Box 1: TopOne OVER(PARTITION BY Game ORDER BY Score Desc)  
TopOne returns the top-rank record, where rank defines the ranking position of the event in the window according to the specified ordering. Ordering/ranking is based on event columns and can be specified in ORDER BY clause.  
Box 2: Hopping(minute,5)  
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.  
A picture containing timeline Description automatically generated



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

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

You create the following components:

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

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

#### Actions

#### Answer Area

Mount the Data Lake Storage onto DBFS.

Write the results to a table in Azure Synapse.

Specify a temporary folder to stage the data.

Read the file into a data frame.

Perform transformations on the data frame.

- A. Mastered  
B. Not Mastered

**Answer:** A

#### Explanation:

Table Description automatically generated

Step 1: Mount the Data Lake Storage onto DBFS

Begin with creating a file system in the Azure Data Lake Storage Gen2 account. Step 2: Read the file into a data frame.

You can load the json files as a data frame in Azure Databricks. Step 3: Perform transformations on the data frame.

Step 4: Specify a temporary folder to stage the data

Specify a temporary folder to use while moving data between Azure Databricks and Azure Synapse. Step 5: Write the results to a table in Azure Synapse.

You upload the transformed data frame into Azure Synapse. You use the Azure Synapse connector for Azure Databricks to directly upload a dataframe as a table in a Azure Synapse.

Reference:

<https://docs.microsoft.com/en-us/azure/azure-databricks/databricks-extract-load-sql-data-warehouse>

#### NEW QUESTION 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 34**

- (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 38**

- (Exam Topic 3)  
You have an Azure Data Lake Storage account that contains a staging zone.  
You need to design a daily process to ingest incremental data from the staging zone, transform the data by executing an R script, and then insert the transformed

data into a data warehouse in Azure Synapse Analytics.

Solution: You use an Azure Data Factory schedule trigger to execute a pipeline that executes mapping data Flow, and then inserts the data into the data warehouse.

Does this meet the goal?

- A. Yes
- B. No

**Answer: B**

**Explanation:**

If you need to transform data in a way that is not supported by Data Factory, you can create a custom activity, not a mapping flow, with your own data processing logic and use the activity in the pipeline. You can create a custom activity to run R scripts on your HDInsight cluster with R installed.

Reference:

<https://docs.microsoft.com/en-US/azure/data-factory/transform-data>

**NEW QUESTION 42**

- (Exam Topic 3)

DRAG DROP

You need to create a partitioned table in an Azure Synapse Analytics dedicated SQL pool.

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                                      |
|--------------------|--|
| CLUSTERED INDEX    | CREATE TABLE table1                              |
| COLLATE            | (  |
| DISTRIBUTION       | ID INTEGER,                                      |
| PARTITION          | col1 VARCHAR(10),                                |
| PARTITION FUNCTION | col2 VARCHAR(10)                                 |
| PARTITION SCHEME   | ) WITH   |
|                    | (  |
|                    | = HASH(ID),                                      |
|                    | (ID RANGE LEFT FOR VALUES (1, 1000000, 2000000)) |
|                    | );   |

- A. Mastered
- B. Not Mastered

**Answer: A**

**Explanation:**

Box 1: DISTRIBUTION

Table distribution options include DISTRIBUTION = HASH ( distribution\_column\_name ), assigns each row to one distribution by hashing the value stored in distribution\_column\_name. Box 2: PARTITION

Table partition options. Syntax:

PARTITION ( partition\_column\_name RANGE [ LEFT | RIGHT ] FOR VALUES ( [ boundary\_value [...n] ] ) )

Reference:

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

**NEW QUESTION 43**

- (Exam Topic 3)

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

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

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

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

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

**Answer: AB**

**Explanation:**

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

A Stored Procedure activity executes a stored procedure on an Azure SQL Database or Azure Synapse Analytics or SQL Server1. You can specify the stored procedure name and parameters in the activity settings1s.

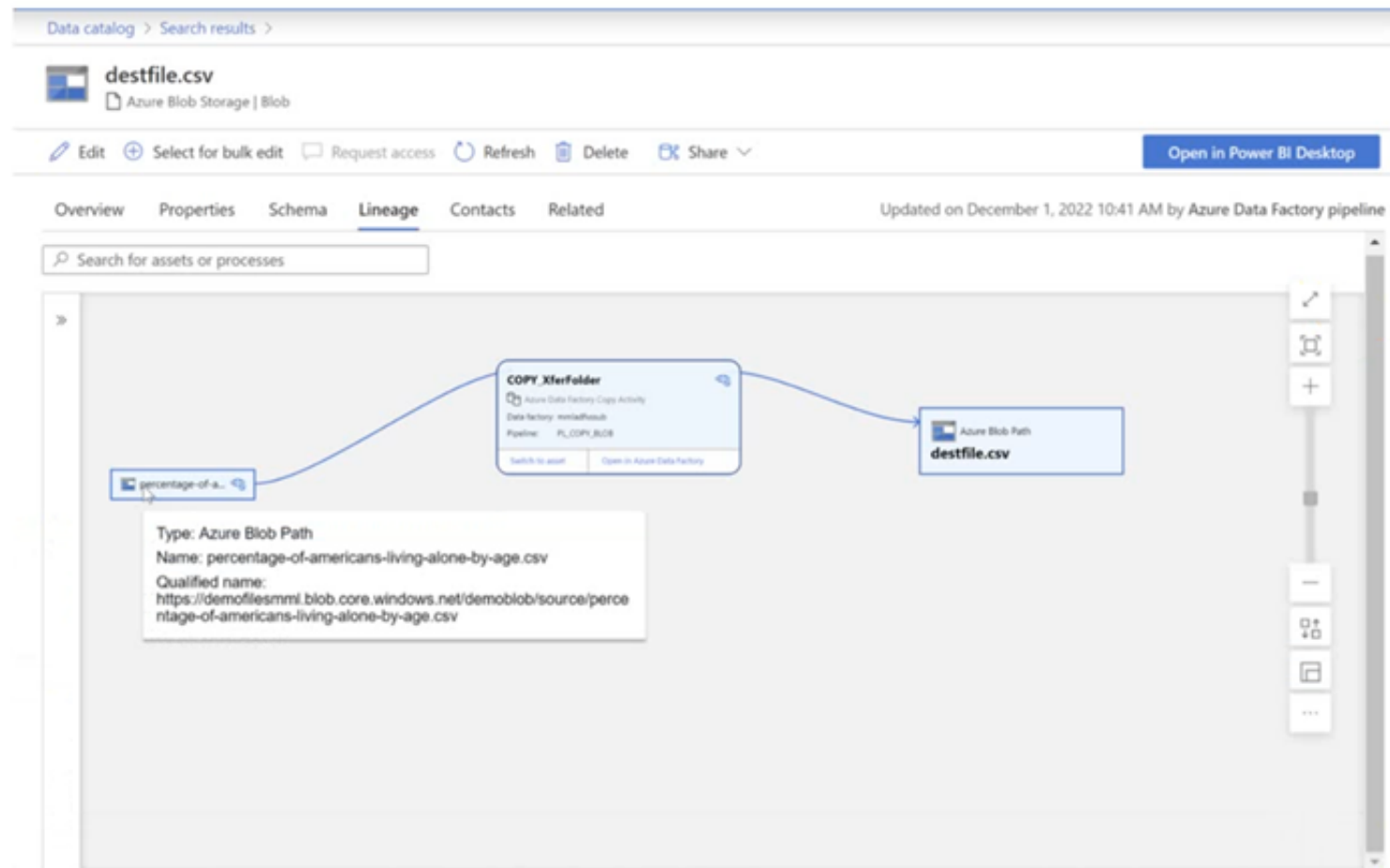
A Lookup activity retrieves a dataset from any data source that returns a single row of data with four columns2. You can use a query to execute a stored procedure as the source of the Lookup activity2y. You can then store the values in the columns as pipeline variables by using expressions2.

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

**NEW QUESTION 44**

- (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 49**

- (Exam Topic 3)

You are developing a solution using a Lambda architecture on Microsoft Azure. The data at test layer must meet the following requirements:

Data storage:

- Serve as a repository (or high volumes of large files in various formats.
- Implement optimized storage for big data analytics workloads.
- Ensure that data can be organized using a hierarchical structure. Batch processing:
- Use a managed solution for in-memory computation processing.
- Natively support Scala, Python, and R programming languages.
- Provide the ability to resize and terminate the cluster automatically. Analytical data store:
- Support parallel processing.
- Use columnar storage.
- Support SQL-based languages.

You need to identify the correct technologies to build the Lambda architecture.

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

| Architecture requirement | Technology  |
|--------------------------|---|
| Data storage             | <div>▼</div> <div>                     Azure SQL Database<br/>                     Azure Blob Storage<br/>                     Azure Cosmos DB<br/>                     Azure Data Lake Store                 </div>      |
| Batch processing         | <div>▼</div> <div>                     HDInsight Spark<br/>                     HDInsight Hadoop<br/>                     Azure Databricks<br/>                     HDInsight Interactive Query                 </div>    |
| Analytical data store    | <div>▼</div> <div>                     HDInsight HBase<br/>                     Azure SQL Data Warehouse<br/>                     Azure Analysis Services<br/>                     Azure Cosmos DB                 </div> |

- A. Mastered  
 B. Not Mastered

**Answer:** A

**Explanation:**

Data storage: Azure Data Lake Store

A key mechanism that allows Azure Data Lake Storage Gen2 to provide file system performance at object storage scale and prices is the addition of a hierarchical namespace. This allows the collection of objects/files within an account to be organized into a hierarchy of directories and nested subdirectories in the same way that the file system on your computer is organized. With the hierarchical namespace enabled, a storage account becomes capable of providing the scalability and cost-effectiveness of object storage, with file system semantics that are familiar to analytics engines and frameworks.

Batch processing: HD Insight Spark

Apache Spark is an open-source, parallel-processing framework that supports in-memory processing to boost the performance of big-data analysis applications. HDInsight is a managed Hadoop service. Use it to deploy and manage Hadoop clusters in Azure. For batch processing, you can use Spark, Hive, Hive LLAP, MapReduce.

Languages: R, Python, Java, Scala, SQL Analytic data store: SQL Data Warehouse

SQL Data Warehouse is a cloud-based Enterprise Data Warehouse (EDW) that uses Massively Parallel Processing (MPP).

SQL Data Warehouse stores data into relational tables with columnar storage. References:

<https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-namespaces> <https://docs.microsoft.com/en-us/azure/architecture/data-guide/technology-choices/batch-processing> <https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-overview-what-is>

**NEW QUESTION 50**

- (Exam Topic 3)

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

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

**Answer:** C

**Explanation:**

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

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

Reference:

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

**NEW QUESTION 54**

- (Exam Topic 3)

You are designing an Azure Synapse solution that will provide a query interface for the data stored in an Azure Storage account. The storage account is only accessible from a virtual network.

You need to recommend an authentication mechanism to ensure that the solution can access the source data.

What should you recommend?

- A. a managed identity



- B. anonymous public read access
- C. a shared key

Answer: A

**Explanation:**

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

**NEW QUESTION 55**

- (Exam Topic 3)

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

You identify the following usage patterns:

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

First 30 days: 

ArchiveCoolHot

After 90 days: 

ArchiveCoolHot

After 365 days: 

ArchiveCoolHot

- A. Mastered
- B. Not Mastered

Answer: A

**Explanation:**

Box 1: Hot

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

Box 2: Cool

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

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

Box 3: Cool

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

**NEW QUESTION 56**

- (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

%%local

%%spark

%%sql

jdbc

saveAsTable

synapsesql

pyspark\_df.createOrReplaceTempView("pysparkdftemptable")

park.sqlContext.sql ("select \* from pysparkdftemptable")

("sqlpool1.dbo.PySparkTable", Constants.INTERNAL)

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Answer Area



NEW QUESTION 61

- (Exam Topic 3)

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

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

Store:

Optimize storage for big data workloads. Provide encryption of data at rest. Operate with no size limits.

Prepare and Train:

- > Provide a fully-managed and interactive workspace for exploration and visualization.
- > Provide the ability to program in R, SQL, Python, Scala, and Java.
- > Provide seamless user authentication with Azure Active Directory.

Model & Serve:

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

Which technologies should you use? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

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

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

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

### NEW QUESTION 65

- (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:   
  
  
  
Remove the data:

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Answer Area

Partition the data:   
  
  
  
Remove the data:

### NEW QUESTION 68

- (Exam Topic 3)

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

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

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

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

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

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

You need to create the Databricks clusters for the workloads.

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

Does this meet the goal?

- A. Yes
- B. No

Answer: B

Explanation:

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

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

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

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

### NEW QUESTION 72

- (Exam Topic 3)

You plan to use an Apache Spark pool in Azure Synapse Analytics to load data to an Azure Data Lake Storage Gen2 account.

You need to recommend which file format to use to store the data in the Data Lake Storage account. The solution must meet the following requirements:

- Column names and data types must be defined within the files loaded to the Data Lake Storage account.
- Data must be accessible by using queries from an Azure Synapse Analytics serverless SQL pool.

- Partition elimination must be supported without having to specify a specific partition. What should you recommend?

A. Delta Lake  
B. JSON  
C. CSV  
D. ORC

**Answer:** D

#### NEW QUESTION 76

- (Exam Topic 3)

You are monitoring an Azure Stream Analytics job by using metrics in Azure.

You discover that during the last 12 hours, the average watermark delay is consistently greater than the configured late arrival tolerance.

What is a possible cause of this behavior?

A. Events whose application timestamp is earlier than their arrival time by more than five minutes arrive as inputs.  
B. There are errors in the input data.  
C. The late arrival policy causes events to be dropped.  
D. The job lacks the resources to process the volume of incoming data.

**Answer:** D

#### Explanation:

Watermark Delay indicates the delay of the streaming data processing job.

There are a number of resource constraints that can cause the streaming pipeline to slow down. The watermark delay metric can rise due to:

- Not enough processing resources in Stream Analytics to handle the volume of input events. To scale up resources, see Understand and adjust Streaming Units.
- Not enough throughput within the input event brokers, so they are throttled. For possible solutions, see Automatically scale up Azure Event Hubs throughput units.
- Output sinks are not provisioned with enough capacity, so they are throttled. The possible solutions vary widely based on the flavor of output service being used.

Reference:

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

#### NEW QUESTION 79

- (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 82

- (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 83

- (Exam Topic 2)

Which Azure Data Factory components should you recommend using together to import the daily inventory data from the SQL server to Azure Data Lake Storage?

To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.



Integration runtime type:

|                                 |   |
|---------------------------------|---|
|                                 | ▼ |
| Azure integration runtime       |   |
| Azure-SSIS integration runtime  |   |
| Self-hosted integration runtime |   |

Trigger type:

|                         |   |
|-------------------------|---|
|                         | ▼ |
| Event-based trigger     |   |
| Schedule trigger        |   |
| Tumbling window trigger |   |

Activity type:

|                           |   |
|---------------------------|---|
|                           | ▼ |
| Copy activity             |   |
| Lookup activity           |   |
| Stored procedure activity |   |

- A. Mastered  
B. Not Mastered

Answer: A

**Explanation:**

Box 1: Self-hosted integration runtime

A self-hosted IR is capable of running copy activity between a cloud data stores and a data store in private network.

Box 2: Schedule trigger Schedule every 8 hours Box 3: Copy activity Scenario:

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

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

**NEW QUESTION 85**

- (Exam Topic 1)

You need to design a data storage structure for the product sales transactions. The solution must meet the sales transaction dataset requirements.

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

**Answer Area**

Table type to store the product sales transactions:

|             |
|-------------|
| Hash        |
| Round-robin |
| Replicated  |

When creating the table for sales transactions:

|  |
|--|
| Configure a clustered index.                   |
| Set the distribution column to product ID.     |
| Set the distribution column to the sales date. |

- A. Mastered  
B. Not Mastered

Answer: A

**Explanation:**

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

Box 1: Hash Scenario:

Ensure that queries joining and filtering sales transaction records based on product ID complete as quickly as possible.

A hash distributed table can deliver the highest query performance for joins and aggregations on large tables. Box 2: Set the distribution column to the sales date.

Scenario: Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month. Boundary values must belong to the partition on the right.

Reference:

<https://rajanieshkaushikk.com/2020/09/09/how-to-choose-right-data-distribution-strategy-for-azure-synapse/>

**NEW QUESTION 88**

- (Exam Topic 1)

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

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

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

- A. Mastered  
B. Not Mastered

**Answer: A**

**Explanation:**

Graphical user interface, application Description automatically generated

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

## Step 1: Create a repository and a main branch

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

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

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

Reference:

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

## NEW QUESTION 91

- (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 96

- (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 97

- (Exam Topic 3)

You have an Azure Active Directory (Azure AD) tenant that contains a security group named Group1. You have an Azure Synapse Analytics dedicated SQL pool named dw1 that contains a schema named schema1.

You need to grant Group1 read-only permissions to all the tables and views in schema1. The solution must use the principle of least privilege.

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

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

Actions

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

Answer Area

- A. Mastered  
B. Not Mastered





Answer: A

Explanation:

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

NEW QUESTION 99

- (Exam Topic 3)  
You configure version control for an Azure Data Factory instance as shown in the following exhibit.



Connections

Linked services

Integration runtimes

Source control

**Git configuration**

ARM template

Parameterization template

Author

Triggers

Global parameters



Security

Customer managed key

Managed private endpoints

### Git repository

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

 Setting  Disconnect

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

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

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

▼

/

adf\_publish

main

Parameterization template

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

▼

/

/contososales

/dwh\_batchetl/adf\_publish/contososales

/main

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Letter Description automatically generated

Box 1: adf\_publish

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

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

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

Reference:

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

NEW QUESTION 103

- (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 104

- (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 108**

- (Exam Topic 3)

You have an Azure Data Factory pipeline named Pipeline1!. Pipelinel contains a copy activity that sends data to an Azure Data Lake Storage Gen2 account.

Pipeline 1 is executed by a schedule trigger.

You change the copy activity sink to a new storage account and merge the changes into the collaboration branch.

After Pipelinel executes, you discover that data is NOT copied to the new storage account. You need to ensure that the data is copied to the new storage account.

What should you do?

- A. Publish from the collaboration branch.
- B. Configure the change feed of the new storage account.
- C. Create a pull request.
- D. Modify the schedule trigger.

**Answer:** A

**Explanation:**

CI/CD lifecycle

➤ A development data factory is created and configured with Azure Repos Git. All developers should have permission to author Data Factory resources like pipelines and datasets.

➤ A developer creates a feature branch to make a change. They debug their pipeline runs with their most recent changes

➤ After a developer is satisfied with their changes, they create a pull request from their feature branch to the main or collaboration branch to get their changes reviewed by peers.

➤ After a pull request is approved and changes are merged in the main branch, the changes get published to the development factory.

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

**NEW QUESTION 111**

- (Exam Topic 3)

You are designing an Azure Synapse Analytics dedicated SQL pool.

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

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

**Answer:** C

**Explanation:**

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

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

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

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

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

Reference:

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

**NEW QUESTION 112**

- (Exam Topic 3)

You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Contacts. Contacts contains a column named Phone.

You need to ensure that users in a specific role only see the last four digits of a phone number when querying the Phone column.

What should you include in the solution?

- A. a default value
- B. dynamic data masking
- C. row-level security (RLS)
- D. column encryption
- E. table partitions

**Answer:** B

**Explanation:**

Dynamic data masking helps prevent unauthorized access to sensitive data by enabling customers to designate how much of the sensitive data to reveal with minimal impact on the application layer. It's a policy-based security feature that hides the sensitive data in the result set of a query over designated database fields, while the data in the database is not changed.

Reference:

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

**NEW QUESTION 116**

- (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 117**

- (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:

ChannelGrouping:

TotalEvents:

- 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 120**

- (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 125**

- (Exam Topic 3)

You are designing a statistical analysis solution that will use custom proprietary Python functions on near real-time data from Azure Event Hubs.

You need to recommend which Azure service to use to perform the statistical analysis. The solution must minimize latency.

What should you recommend?

A. Azure Stream Analytics

B. Azure SQL Database

C. Azure Databricks

D. Azure Synapse Analytics

**Answer:** A

**Explanation:**

Reference:

<https://docs.microsoft.com/en-us/azure/event-hubs/process-data-azure-stream-analytics>

**NEW QUESTION 128**

- (Exam Topic 3)

You plan to implement an Azure Data Lake Gen2 storage account.

You need to ensure that the data lake will remain available if a data center fails in the primary Azure region. The solution must minimize costs.

Which type of replication should you use for the storage account?

A. geo-redundant storage (GRS)

B. zone-redundant storage (ZRS)

C. locally-redundant storage (LRS)

D. geo-zone-redundant storage (GZRS)

**Answer:** C

**Explanation:**

Locally redundant storage (LRS) copies your data synchronously three times within a single physical location in the primary region. LRS is the least expensive replication option

Reference:

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

**NEW QUESTION 129**

- (Exam Topic 3)

You have an Azure Data Lake Storage Gen 2 account named storage1.

You need to recommend a solution for accessing the content in storage1. The solution must meet the following requirements:

➤ List and read permissions must be granted at the storage account level.

➤ Additional permissions can be applied to individual objects in storage1.

➤ Security principals from Microsoft Azure Active Directory (Azure AD), part of Microsoft Entra, must be used for authentication.

What should you use? To answer, drag the appropriate components to the correct requirements. Each component 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.

Components

Access control lists (ACLs)

Role-based access control (RBAC) roles

Shared access signatures (SAS)

Shared account keys

Answer Area

To grant permissions at the storage account level:

To grant permissions at the object level:

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Box 1: Role-based access control (RBAC) roles  
List and read permissions must be granted at the storage account level.  
Security principals from Microsoft Azure Active Directory (Azure AD), part of Microsoft Entra, must be used for authentication.  
Role-based access control (Azure RBAC)  
Azure RBAC uses role assignments to apply sets of permissions to security principals. A security principal is an object that represents a user, group, service principal, or managed identity that is defined in Azure Active Directory (AD). A permission set can give a security principal a "coarse-grain" level of access such as read or write access to all of the data in a storage account or all of the data in a container.  
Box 2: Access control lists (ACLs)  
Additional permissions can be applied to individual objects in storage1. Access control lists (ACLs)  
ACLs give you the ability to apply "finer grain" level of access to directories and files. An ACL is a permission construct that contains a series of ACL entries. Each ACL entry associates security principal with an access level.  
Reference: <https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control-model>

NEW QUESTION 133

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

Load methodology:

Full Load

Incremental Load

Load individual files as they arrive

Trigger:

Fixed schedule

New file

Tumbling window

- A. Mastered
- B. Not Mastered

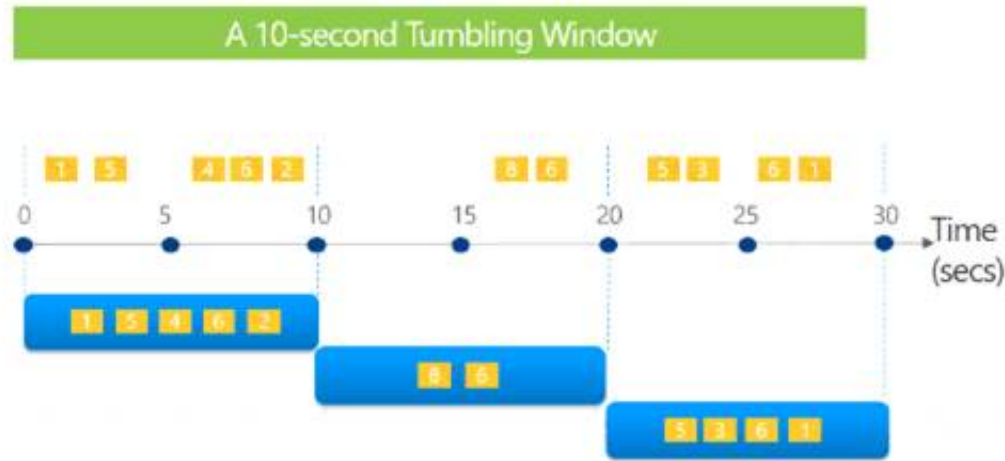
Answer: A

Explanation:

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



Tell me the count of tweets per time zone every 10 seconds



```
SELECT TimeZone, COUNT(*) AS Count
FROM TwitterStream TIMESTAMP BY CreatedAt
GROUP BY TimeZone, TumblingWindow(second,10)
```

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

#### NEW QUESTION 138

- (Exam Topic 3)

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

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

**Answer:** B

#### Explanation:

Columnstore and columnstore archive compression

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

Use COLUMNSTORE\_ARCHIVE data compression to compress columnstore data with archival compression.

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

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

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

#### NEW QUESTION 139

- (Exam Topic 3)

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

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

How should you complete the Transact-SQL code? To answer, drag the appropriate values to the correct

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

NOTE Each correct selection is worth one point.

Values

BEGIN DISTRIBUTED TRANSACTION

BEGIN TRAN

COMMIT TRAN

ROLLBACK TRAN

SET RESULT\_SET\_CACHING ON

Answer Area

BEGIN TRY

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

SELECT col1, col2, col3 FROM stage.Table1;

END TRY

BEGIN CATCH

IF @@TRANCOUNT > 0

BEGIN

END

END CATCH;

IF @@TRANCOUNT >0

BEGIN

COMMIT TRAN;

END

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Values

BEGIN DISTRIBUTED TRANSACTION

BEGIN TRAN

COMMIT TRAN

ROLLBACK TRAN

SET RESULT\_SET\_CACHING ON

Answer Area

BEGIN TRAN

BEGIN TRY

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

SELECT col1, col2, col3 FROM stage.Table1;

END TRY

BEGIN CATCH

IF @@TRANCOUNT > 0

BEGIN

ROLLBACK TRAN

END

END CATCH;

IF @@TRANCOUNT >0

BEGIN

COMMIT TRAN;

END

NEW QUESTION 144

- (Exam Topic 3)

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution. After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen. You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Table1. You have files that are ingested and loaded into an Azure Data Lake Storage Gen2 container named container1. You plan to insert data from the files in container1 into Table1 and transform the data. Each row of data in the files will produce one row in the serving layer of Table1. You need to ensure that when the source data files are loaded to container1, the DateTime is stored as an additional column in Table1. Solution: You use an Azure Synapse Analytics serverless SQL pool to create an external table that has an additional DateTime column. Does this meet the goal?

- A. Yes
- B. No

Answer: B

Explanation:

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

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

#### NEW QUESTION 148

- (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 151

- (Exam Topic 3)

You have an Azure Databricks resource.

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

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

**Answer:** B

#### Explanation:

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

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

#### NEW QUESTION 154

- (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 158

- (Exam Topic 3)

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

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

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

NOTE: Each correct selection is worth one point.

SELECT sensorId,  
growth = reading -

LAG

LAST

LEAD

(reading) OVER (PARTITION BY sensorId

LIMIT DURATION

OFFSET

WHEN

(hour,1))

FROM input

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Box 1: LAG  
The LAG analytic operator allows one to look up a “previous” event in an event stream, within certain constraints. It is very useful for computing the rate of growth of a variable, detecting when a variable crosses a threshold, or when a condition starts or stops being true.  
Box 2: LIMIT DURATION  
Example: Compute the rate of growth, per sensor: SELECT sensorId,  
growth = reading  
LAG(reading) OVER (PARTITION BY sensorId LIMIT DURATION(hour, 1)) FROM input  
Reference:  
https://docs.microsoft.com/en-us/stream-analytics-query/lag-azure-stream-analytics

NEW QUESTION 161

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

- A. Connect to the built-in pool and query sysdm\_pdw\_sys\_info.
- B. Connect to Pool1 and run DBCC CHECKALLOC.
- C. Connect to the built-in pool and run DBCC CHECKALLOC.
- D. Connect to Pool! and query sys.dm\_pdw\_nodes\_db\_partition\_stats.

Answer: D

Explanation:

Microsoft recommends use of sys.dm\_pdw\_nodes\_db\_partition\_stats to analyze any skewness in the data. Reference:  
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/cheat-sheet

NEW QUESTION 165

- (Exam Topic 3)  
You have an Azure Synapse serverless SQL pool.  
You need to read JSON documents from a file by using the OPENROWSET function.  
How should you complete the query? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

Answer Area

SELECT \*

FROM OPENROWSET

(

BULK

'https://sourcedatalake.blob.core.windows.net/public/docs.json',

FORMAT =

'JSON'

'CSV'

'DELTA'

'JSON'

'PARQUET'

FIELDTERMINATOR = '0x0b',

FIELDQUOTE =

'0x0b'

'0x09'

'0x0a'

'0x0b'

'0x0c'

)

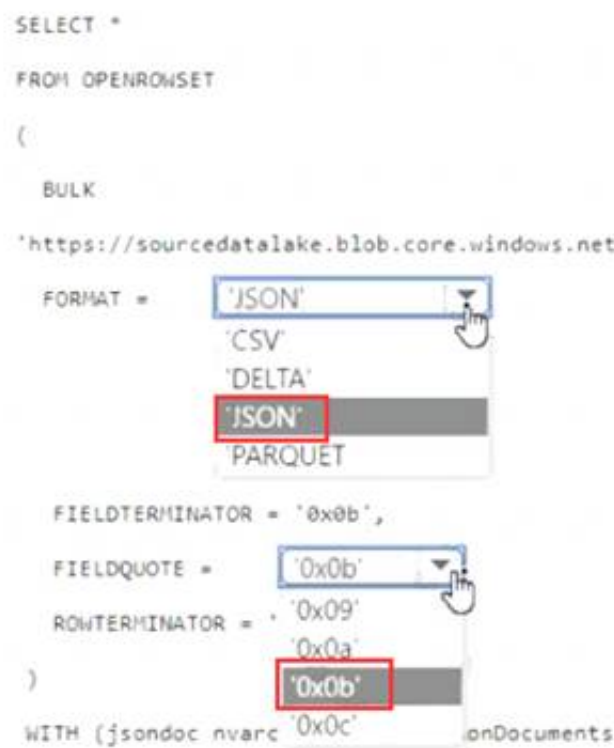
WITH (jsondoc nvarchar(1024) onDocuments

- A. Mastered
- B. Not Mastered



Answer: A

Explanation:  
Answer Area



**NEW QUESTION 170**

- (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 172**

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

Components

a database scoped credential

an asymmetric key

an external data source

a database encryption key

an external file format

>

<

Answer Area

⬆

⬇

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Components

a database scoped credential

an asymmetric key

an external data source

a database encryption key

an external file format

Answer Area

a database scoped credential

an external data source

an external file format

⬆

⬇

NEW QUESTION 173

- (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 177

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

Values

CustomerKey

HASH

ROUND\_ROBIN

REPLICATE

OrderDateKey

SalesOrderNumber

Answer Area

```
CREATE TABLE [dbo].[FactSales]
(
    [ProductKey]          int          NOT NULL
,   [OrderDateKey]       int          NOT NULL
,   [CustomerKey]        int          NOT NULL
,   [SalesOrderNumber]   nvarchar ( 20 ) NOT NULL
,   [OrderQuantity]      smallint     NOT NULL
,   [UnitPrice]          money        NOT NULL
)
WITH
(
    CLUSTERED COLUMNSTORE INDEX
,   DISTRIBUTION = Value ([ProductKey])
,   PARTITION ( [ Value ] RANGE RIGHT FOR VALUES
                (20170101,20180101,20190101,20200101,20210101)
            )
)
```

- A. Mastered
- B. Not Mastered

Answer: A

**Explanation:**  
Box 1: HASH  
Box 2: OrderDateKey

In most cases, table partitions are created on a date column.

A way to eliminate rollbacks is to use Metadata Only operations like partition switching for data management. For example, rather than execute a DELETE statement to delete all rows in a table where the order\_date was in October of 2001, you could partition your data early. Then you can switch out the partition with data for an empty partition from another table.

Reference:

<https://docs.microsoft.com/en-us/sql/t-sql/statements/create-table-azure-sql-data-warehouse> <https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/best-practices-dedicated-sql-pool>

#### NEW QUESTION 181

- (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 185

- (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   |
|---------------------------------------|---|
| <div>{deviceID}</div>                 | <div>/ <div>Value</div> / <div>Value</div> / <div>Value</div> .json</div> |
| <div>{mm}/{HH}/{DD}/{MM}/{YYYY}</div> |   |
| <div>{regionID}/{deviceID}</div>      |   |
| <div>{regionID}/raw</div>             |   |
| <div>{YYYY}/{MM}/{DD}/{HH}</div>      |   |
| <div>{YYYY}/{MM}/{DD}/{HH}/{mm}</div> |   |
| <div>raw/{deviceID}</div>             |   |
| <div>raw/{regionID}</div>             |   |

- 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/paolosavatori/StreamAnalyticsAzureDataLakeStore/blob/master/README.md>

#### NEW QUESTION 188

- (Exam Topic 3)

You have a SQL pool in Azure Synapse.

You discover that some queries fail or take a long time to complete. You need to monitor for transactions that have rolled back.

Which dynamic management view should you query?

- A. sys.dm\_pdw\_request\_steps
- B. sys.dm\_pdw\_nodes\_tran\_database\_transactions
- C. sys.dm\_pdw\_waits
- D. sys.dm\_pdw\_exec\_sessions

**Answer: B**

#### Explanation:

You can use Dynamic Management Views (DMVs) to monitor your workload including investigating query execution in SQL pool.

If your queries are failing or taking a long time to proceed, you can check and monitor if you have any transactions rolling back.

Example:

-- Monitor rollback SELECT

SUM(CASE WHEN t.database\_transaction\_next\_undo\_lsn IS NOT NULL THEN 1 ELSE 0 END), t.pdw\_node\_id,  
 nod.[type]

FROM sys.dm\_pdw\_nodes\_tran\_database\_transactions t

JOIN sys.dm\_pdw\_nodes nod ON t.pdw\_node\_id = nod.pdw\_node\_id GROUP BY t.pdw\_node\_id, nod.[type]

Reference:

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

#### NEW QUESTION 192

- (Exam Topic 3)

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

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

What should you include in the recommendation?

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

**Answer: A**

#### Explanation:

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

Example: XXXX-XXXX-XXXX-1234

Reference:

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

#### NEW QUESTION 193

- (Exam Topic 3)

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

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

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

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

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

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

NOTE: Each correct selection is worth one point.



Values

Answer Area

CAST

COLLATE

CONVERT

FLATTEN

PIVOT

UNPIVOT

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

- A. Mastered  
 B. Not Mastered

Answer: A

**Explanation:**

Text Description automatically generated

Box 1: PIVOT

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

Reference:

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

**NEW QUESTION 198**

- (Exam Topic 3)


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


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

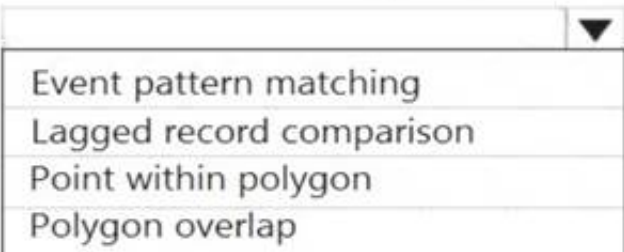
You need to ensure that when a GPS position is outside the expected area, a message is added to another event hub for processing within 30 seconds. The solution must minimize cost.

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

NOTE: Each correct selection is worth one point.

Service: 

Window: 

Analysis type: 

- A. Mastered  
 B. Not Mastered

Answer: A

**Explanation:**

Box 1: Azure Stream Analytics Box 2: Hopping

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

Box 3: Point within polygon Reference:

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

**NEW QUESTION 200**

- (Exam Topic 3)

You create an Azure Databricks cluster and specify an additional library to install. When you attempt to load the library to a notebook, the library is not found. You need to identify the cause of the issue. What should you review?

- A. notebook logs
- B. cluster event logs
- C. global init scripts logs
- D. workspace logs

**Answer:** C

**Explanation:**

Cluster-scoped Init Scripts: Init scripts are shell scripts that run during the startup of each cluster node before the Spark driver or worker JVM starts. Databricks customers use init scripts for various purposes such as installing custom libraries, launching background processes, or applying enterprise security policies. Logs for Cluster-scoped init scripts are now more consistent with Cluster Log Delivery and can be found in the same root folder as driver and executor logs for the cluster.

Reference:

<https://databricks.com/blog/2018/08/30/introducing-cluster-scoped-init-scripts.html>

**NEW QUESTION 202**

- (Exam Topic 3)

You have an Azure subscription that contains an Azure Synapse Analytics workspace named workspace1. Workspace1 connects to an Azure DevOps repository named repo1. Repo1 contains a collaboration branch named main and a development branch named branch1. Branch1 contains an Azure Synapse pipeline named pipeline1.

In workspace1, you complete testing of pipeline1. You need to schedule pipeline1 to run daily at 6 AM.

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

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

Actions

Answer Area

Create a new branch in Repo1.

Merge the changes from branch1 into main.

Associate the schedule trigger with pipeline1.

Switch to Synapse live mode.

Create a schedule trigger.

Publish the contents of main.

>

<

- A. Mastered
- B. Not Mastered

**Answer:** A

**Explanation:**

Timeline Description automatically generated

**NEW QUESTION 204**

- (Exam Topic 3)

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

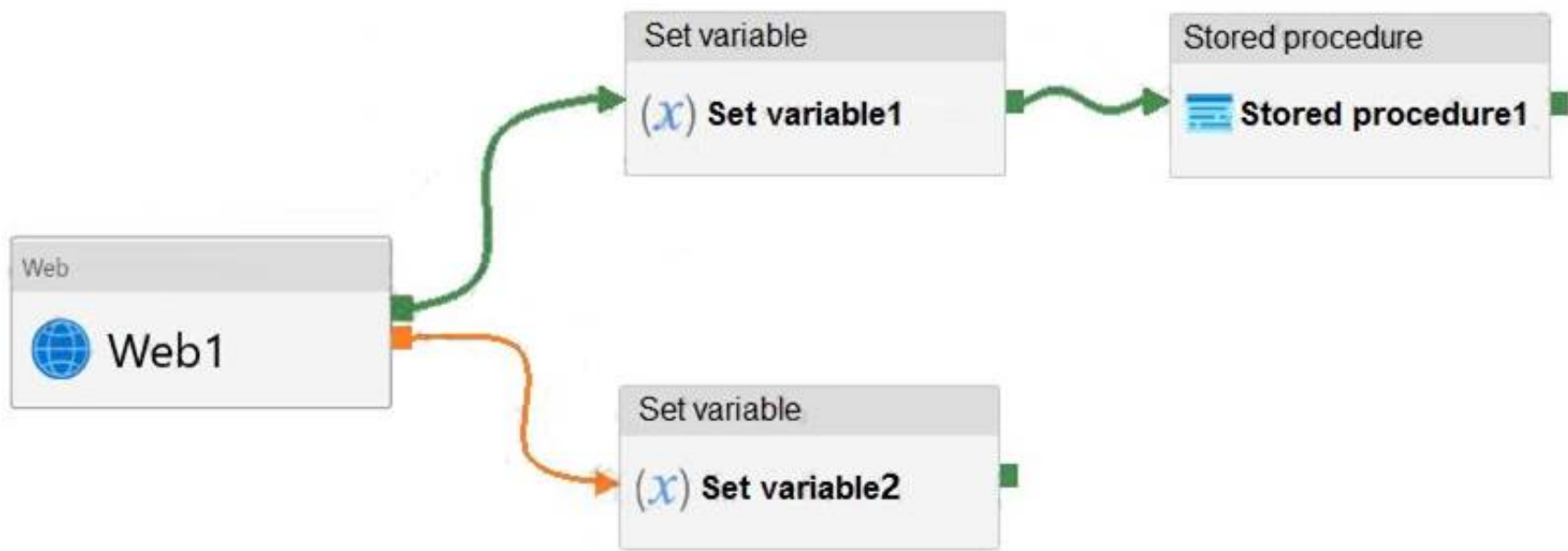
You need to ensure that you can query the data in container1 by using the serverless1 SQL pool. Which three actions should you perform? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

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

**Answer:** ACD

**NEW QUESTION 209**

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



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

Stored procedure1 will execute Web1 and Set variable1 [answer choice]

|          |   |
|----------|---|
|          | ▼ |
| complete |   |
| fail     |   |
| succeed  |   |

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

|           |   |
|-----------|---|
|           | ▼ |
| Canceled  |   |
| Failed    |   |
| Succeeded |   |

- A. Mastered
- B. Not Mastered

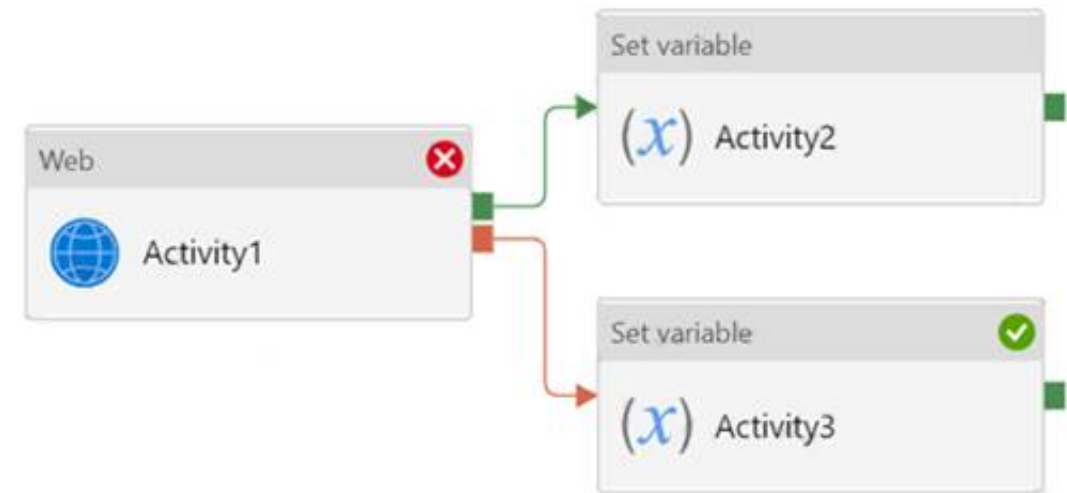
Answer: A

Explanation:

Box 1: succeed

Box 2: failed Example:

Now let's say we have a pipeline with 3 activities, where Activity1 has a success path to Activity2 and a failure path to Activity3. If Activity1 fails and Activity3 succeeds, the pipeline will fail. The presence of the success path alongside the failure path changes the outcome reported by the pipeline, even though the activity executions from the pipeline are the same as the previous scenario.



Activity1 fails, Activity2 is skipped, and Activity3 succeeds. The pipeline reports failure. Reference:  
<https://datasavvy.me/2021/02/18/azure-data-factory-activity-failures-and-pipeline-outcomes/>

NEW QUESTION 213

- (Exam Topic 3)

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

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

How should you model the employee data?

- A. as a temporal table
- B. as a SQL graph table

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

**Answer:** D

**Explanation:**

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

Reference:

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

**NEW QUESTION 215**

- (Exam Topic 3)

You are designing an application that will store petabytes of medical imaging data

When the data is first created, the data will be accessed frequently during the first week. After one month, the data must be accessible within 30 seconds, but files will be accessed infrequently. After one year, the data will be accessed infrequently but must be accessible within five minutes.

You need to select a storage strategy for the data. The solution must minimize costs.

Which storage tier should you use for each time frame? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

|                  |  |
|------------------|--|
| First week:      | <div><div>▼</div><div>Archive</div><div>Cool</div><div>Hot</div></div> |
| After one month: | <div><div>▼</div><div>Archive</div><div>Cool</div><div>Hot</div></div> |
| After one year:  | <div><div>▼</div><div>Archive</div><div>Cool</div><div>Hot</div></div> |

- A. Mastered
- B. Not Mastered

**Answer:** A

**Explanation:**

First week: Hot

Hot - Optimized for storing data that is accessed frequently.

After one month: Cool

Cool - Optimized for storing data that is infrequently accessed and stored for at least 30 days.

After one year: Cool

**NEW QUESTION 219**

- (Exam Topic 3)

You have an Azure event hub named retailhub that has 16 partitions. Transactions are posted to retailhub. Each transaction includes the transaction ID, the individual line items, and the payment details. The transaction ID is used as the partition key.

You are designing an Azure Stream Analytics job to identify potentially fraudulent transactions at a retail store. The job will use retailhub as the input. The job will output the transaction ID, the individual line items, the payment details, a fraud score, and a fraud indicator.

You plan to send the output to an Azure event hub named fraudhub.

You need to ensure that the fraud detection solution is highly scalable and processes transactions as quickly as possible.

How should you structure the output of the Stream Analytics job? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.



Number of partitions:

Partition key:

- A. Mastered
- B. Not Mastered

**Answer:** A

**Explanation:**

Box 1: 16

For Event Hubs you need to set the partition key explicitly.

An embarrassingly parallel job is the most scalable scenario in Azure Stream Analytics. It connects one partition of the input to one instance of the query to one partition of the output. Box 2: Transaction ID

Reference:

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

**NEW QUESTION 223**

- (Exam Topic 3)

You have an Azure Synapse Analytics dedicated SQL pool.

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

What should you do?

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

**Answer:** B

**Explanation:**

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

Reference:

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

**NEW QUESTION 228**

- (Exam Topic 3)

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

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

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

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

Type 0

Type 1

Type 2

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

a surrogate key

a business key

an audit column

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Box 1: Type 2  
A Type 2 SCD supports versioning of dimension members. Often the source system doesn't store versions, so the data warehouse load process detects and manages changes in a dimension table. In this case, the dimension table must use a surrogate key to provide a unique reference to a version of the dimension member. It also includes columns that define the date range validity of the version (for example, StartDate and EndDate) and possibly a flag column (for example, IsCurrent) to easily filter by current dimension members.  
Reference:  
<https://docs.microsoft.com/en-us/learn/modules/populate-slowly-changing-dimensions-azure-synapse-analytics>

NEW QUESTION 232

- (Exam Topic 3)  
You have an Azure Stream Analytics job that is a Stream Analytics project solution in Microsoft Visual Studio. The job accepts data generated by IoT devices in the JSON format.  
You need to modify the job to accept data generated by the IoT devices in the Protobuf format.  
Which three actions should you perform from Visual Studio on sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

Actions

Change the Event Serialization Format to Protobuf in the input.json file of the job and reference the DLL.

Add an Azure Stream Analytics Custom Deserializer Project (.NET) project to the solution.

Add .NET deserializer code for Protobuf to the custom deserializer project.

Add .NET deserializer code for Protobuf to the Stream Analytics project.

Add an Azure Stream Analytics Application project to the solution.

Answer Area

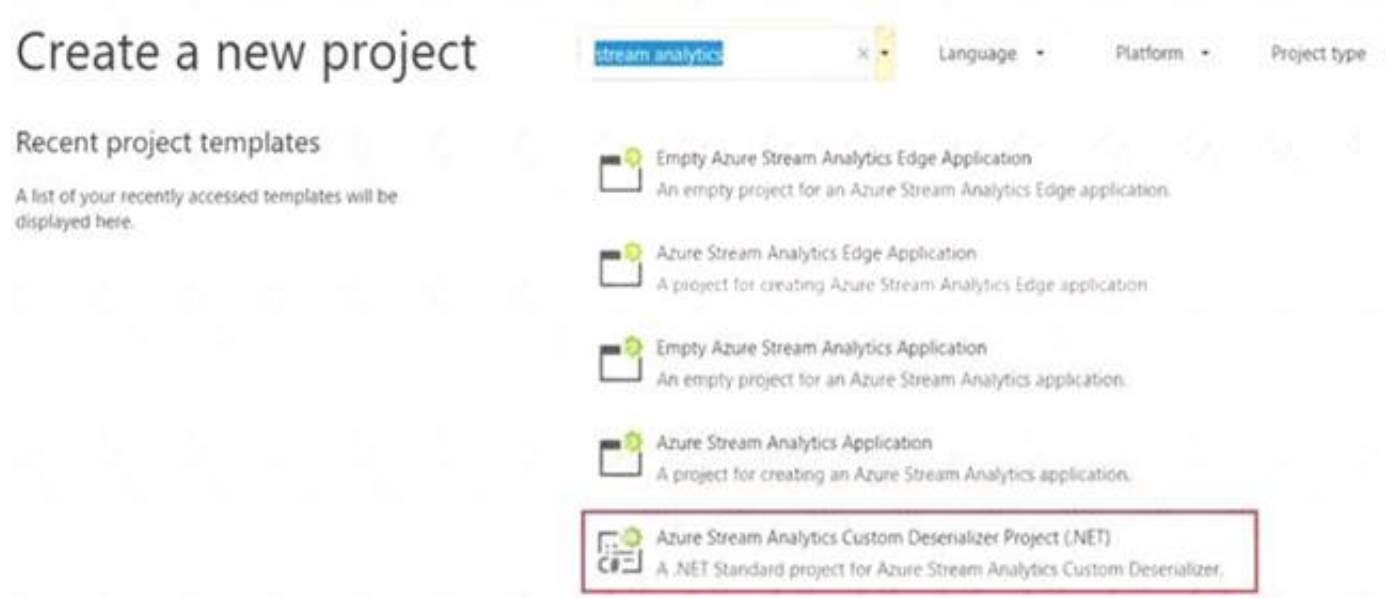
- A. Mastered
- B. Not Mastered

**Answer:** A

**Explanation:**

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

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



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

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

\* 4. Build the Protobuf Deserializer project.

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

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

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

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

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

Reference:

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

**NEW QUESTION 233**

- (Exam Topic 3)

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

File sizes range from 4.KB to 5 GB.

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

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

**Answer:** D

**Explanation:**

Avro supports batch and is very relevant for streaming.

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

Reference:

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

**NEW QUESTION 236**

- (Exam Topic 3)

You are building an Azure Synapse Analytics dedicated SQL pool that will contain a fact table for transactions from the first half of the year 2020.

You need to ensure that the table meets the following requirements:

- > Minimizes the processing time to delete data that is older than 10 years
- > Minimizes the I/O for queries that use year-to-date values

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.

Answer Area

```
CREATE TABLE [dbo].[FactTransaction]
(
    [TransactionTypeID] int NOT NULL
,   [TransactionDateID] int NOT NULL
,   [CustomerID] int NOT NULL
,   [RecipientID] int NOT NULL
,   [Amount] money NOT NU::
)
WITH
(
    CLUSTERED COLUMNSTORE INDEX
    DISTRIBUTION
    PARTITION
    TRUNCATE_TARGET
    (
        [TransactionDateID]
        [TransactionDateID], [TransactionTypeID]
        HASH([TransactionTypeID])
        ROUND_ROBIN
        (20200101,20200201,20200301,20200401,20200501,20200601)
    )
    RANGE RIGHT FOR VALUES
```

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Table Description automatically generated  
Box 1: PARTITION  
RANGE RIGHT FOR VALUES is used with PARTITION.  
Part 2: [TransactionDateID] Partition on the date column.  
Example: Creating a RANGE RIGHT partition function on a datetime column  
The following partition function partitions a table or index into 12 partitions, one for each month of a year's worth of values in a datetime column.  
CREATE PARTITION FUNCTION [myDateRangePF1] (datetime)  
AS RANGE RIGHT FOR VALUES ('20030201', '20030301', '20030401',  
'20030501', '20030601', '20030701', '20030801',  
'20030901', '20031001', '20031101', '20031201');  
Reference:  
<https://docs.microsoft.com/en-us/sql/t-sql/statements/create-partition-function-transact-sql>

NEW QUESTION 241

- (Exam Topic 3)  
You have an Azure Synapse Analytics workspace named WS1.  
You have an Azure Data Lake Storage Gen2 container that contains JSON-formatted files in the following format.



```
{
  "id": "66532691-ab20-11ea-8b1d-936b3ec64e54",
  "context": {
    "data": {
      "eventTime": "2020-06-10T13:43:34.553Z",
      "samplingRate": "100.0",
      "isSynthetic": "false"
    },
    "session": {
      "isFirst": "false",
      "id": "38619c14-7a23-4687-8268-95862c5326b1"
    },
    "custom": {
      "dimensions": [
        {
          "customerInfo": {
            "ProfileType": "ExpertUser",
            "RoomName": "",
            "CustomerName": "diamond",
            "UserName": "XXXX@yahoo.com"
          }
        }
      ]
    }
  },
  {
    "customerInfo": {
      "ProfileType": "Novice",
      "RoomName": "",
      "CustomerName": "topaz",
      "UserName": "XXXX@outlook.com"
    }
  }
]
```

You need to use the serverless SQL pool in WS1 to read the files.

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

NOTE: Each correct selection is worth one point.

#### Values

#### Answer Area

select\*

FROM

(

BULK 'https://contoso.blob.core.windows.net/contosodw',  
 FORMAT= 'CSV',  
 fieldterminator = '0x0b',  
 fieldquote = '0x0b',  
 rowterminator = '0x0b'

)

with (id varchar(50),  
 contextdateeventTime varchar(50) '\$.context.data.eventTime',  
 contextdatasamplingRate varchar(50) '\$.context.data.samplingRate',  
 contextdataisSynthetic varchar(50) '\$.context.data.isSynthetic',  
 contextsessionisFirst varchar(50) '\$.context.session.isFirst',  
 contextsession varchar(50) '\$.context.session.id',  
 contextcustomdimensions varchar(max) '\$.context.custom.dimensions'

) as q

cross apply  (contextcustomdimensions)

with ( ProfileType varchar(50) '\$.customerInfo.ProfileType',  
 RoomName varchar(50) '\$.customerInfo.RoomName',  
 CustomerName varchar(50) '\$.customerInfo.CustomerName',  
 UserName varchar(50) '\$.customerInfo.UserName'

)

opendatasource

openjson

openquery

openrowset

- A. Mastered
- B. Not Mastered

**Answer: A**

#### Explanation:

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

Box 1: openrowset

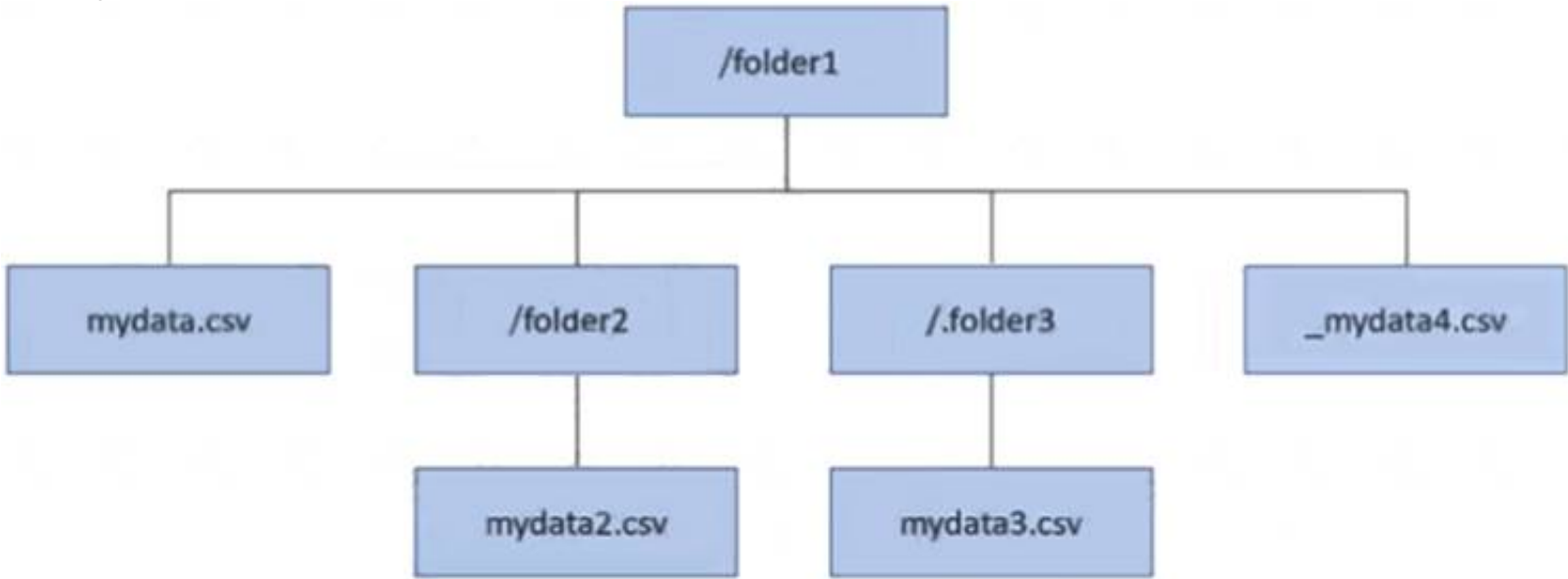
The easiest way to see to the content of your CSV file is to provide file URL to OPENROWSET function, specify csv FORMAT.

Example: SELECT \*

FROM OPENROWSET(  
BULK 'csv/population/population.csv', DATA\_SOURCE = 'SqlOnDemandDemo', FORMAT = 'CSV', PARSER\_VERSION = '2.0', FIELDTERMINATOR = ',',  
ROWTERMINATOR = '\n'  
Box 2: openjson  
You can access your JSON files from the Azure File Storage share by using the mapped drive, as shown in the following example:  
SELECT book.\* FROM  
OPENROWSET(BULK N't:\books\books.json', SINGLE\_CLOB) AS json CROSS APPLY OPENJSON(BulkColumn)  
WITH( id nvarchar(100), name nvarchar(100), price float, pages\_i int, author nvarchar(100)) AS book  
Reference:  
<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/query-single-csv-file> <https://docs.microsoft.com/en-us/sql/relational-databases/json/import-json-documents-into-sql-server>

NEW QUESTION 245

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



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

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

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

| Statements   | Yes                   | No                    |
|--|-----------------------|-----------------------|
| When selecting all the rows in dbo.Table1, data from the mydata2.csv file will be returned.  | <input type="radio"/> | <input type="radio"/> |
| When selecting all the rows in dbo.Table1, data from the mydata3.csv file will be returned.  | <input type="radio"/> | <input type="radio"/> |
| When selecting all the rows in dbo.Table1, data from the _mydata4.csv file will be returned. | <input type="radio"/> | <input type="radio"/> |

- A. Mastered
- B. Not Mastered

Answer: A

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

NEW QUESTION 250

- (Exam Topic 3)  
You have an Azure Data Lake Storage account that has a virtual network service endpoint configured.  
You plan to use Azure Data Factory to extract data from the Data Lake Storage account. The data will then be loaded to a data warehouse in Azure Synapse Analytics by using PolyBase.  
Which authentication method should you use to access Data Lake Storage?

- A. shared access key authentication
- B. managed identity authentication
- C. account key authentication

D. service principal authentication

**Answer:** B

**Explanation:**

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/connector-azure-sql-data-warehouse#use-polybase-to-load-d>

**NEW QUESTION 252**

- (Exam Topic 3)

You plan to create an Azure Data Lake Storage Gen2 account

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

- Provides the highest degree of data resiliency
- Ensures that content remains available for writes if a primary data center fails

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

**Answer Area**

Replication mechanism:

| Change feed                                     |
|---|
| Zone-redundant storage (ZRS)                    |
| Read-access geo-redundant storage (RA-GRS)      |
| Read-access geo-zone-redundant storage (RA-GRS) |

Failover process:

| Failover initiated by Microsoft                             |
|---|
| Failover manually initiated by the customer                 |
| Failover automatically initiated by an Azure Automation job |

A. Mastered

B. Not Mastered

**Answer:** A

**Explanation:**

Graphical user interface, text, application Description automatically generated

Availability : "Microsoft recommends RA-GZRS for maximum availability and durability for your applications."

Failover: "The customer initiates the account failover to the secondary endpoint. " <https://docs.microsoft.com/en-us/azure/storage/common/storage-disaster-recovery-guidance?toc=/azure/storage/>

<https://docs.microsoft.com/en-us/answers/questions/32583/azure-data-lake-gen2-disaster-recoverystorage-acco.h>

**NEW QUESTION 255**

- (Exam Topic 3)

You are building a data flow in Azure Data Factory that upserts data into a table in an Azure Synapse Analytics dedicated SQL pool.

You need to add a transformation to the data flow. The transformation must specify logic indicating when a row from the input data must be upserted into the sink.

Which type of transformation should you add to the data flow?

A. join

B. select

C. surrogate key

D. alter row

**Answer:** D

**Explanation:**

The alter row transformation allows you to specify insert, update, delete, and upsert policies on rows based on expressions. You can use the alter row transformation to perform upserts on a sink table by matching on a key column and setting the appropriate row policy

**NEW QUESTION 260**

- (Exam Topic 3)

You have an Azure data factory.

You need to examine the pipeline failures from the last 60 days. What should you use?

A. the Activity log blade for the Data Factory resource

B. the Monitor & Manage app in Data Factory

C. the Resource health blade for the Data Factory resource

D. Azure Monitor

**Answer:** D

**Explanation:**

Data Factory stores pipeline-run data for only 45 days. Use Azure Monitor if you want to keep that data for a longer time.  
Reference:  
<https://docs.microsoft.com/en-us/azure/data-factory/monitor-using-azure-monitor>

#### NEW QUESTION 262

- (Exam Topic 3)

You plan to develop a dataset named Purchases by using Azure databricks Purchases will contain the following columns:

- ProductID
- ItemPrice
- lineTotal
- Quantity
- StoreID
- Minute
- Month
- Hour
- Year
- Day

You need to store the data to support hourly incremental load pipelines that will vary for each StoreID. the solution must minimize storage costs. How should you complete the rode? To answer, select the appropriate options In the answer area.

NOTE: Each correct selection is worth one point.

df.write

|              |   |
|--------------|---|
| .bucketBy    | ▼ |
| .partitionBy |   |
| .range       |   |
| .sortBy      |   |

|   |   |
|---|---|
| (***)                                       | ▼ |
| ("StoreID", "Hour")                         |   |
| ("StoreID", "Year", "Month", "Day", "Hour") |   |

.mode("append")

|                            |   |
|----------------------------|---|
| .csv("/Purchases")         | ▼ |
| .json("/Purchases")        |   |
| .parquet("/Purchases")     |   |
| .saveAsTable("/Purchases") |   |

- A. Mastered  
B. Not Mastered

**Answer:** A

#### Explanation:

Box 1: partitionBy

We should overwrite at the partition level. Example: df.write.partitionBy("y","m","d") mode(SaveMode.Append)

parquet("/data/hive/warehouse/db\_name.db/" + tableName) Box 2: ("StoreID", "Year", "Month", "Day", "Hour", "StoreID") Box 3: parquet("/Purchases")

Reference:

<https://intellipaat.com/community/11744/how-to-partition-and-write-dataframe-in-spark-without-deleting-partiti>

#### NEW QUESTION 265

- (Exam Topic 3)

You have an enterprise data warehouse in Azure Synapse Analytics named DW1 on a server named Server1. You need to verify whether the size of the transaction log file for each distribution of DW1 is smaller than 160 GB.

What should you do?

- A. On the master database, execute a query against the sys.dm\_pdw\_nodes\_os\_performance\_counters dynamic management view.  
B. From Azure Monitor in the Azure portal, execute a query against the logs of DW1.  
C. On DW1, execute a query against the sys.database\_files dynamic management view.  
D. Execute a query against the logs of DW1 by using theGet-AzOperationalInsightSearchResult PowerShell cmdlet.

**Answer:** A

#### Explanation:

The following query returns the transaction log size on each distribution. If one of the log files is reaching 160 GB, you should consider scaling up your instance or limiting your transaction size.

-- Transaction log size SELECT

instance\_name as distribution\_db, cntr\_value\*1.0/1048576 as log\_file\_size\_used\_GB, pdw\_node\_id

FROM sys.dm\_pdw\_nodes\_os\_performance\_counters WHERE

instance\_name like 'Distribution\_%'

AND counter\_name = 'Log File(s) Used Size (KB)'

References:

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

#### NEW QUESTION 267

- (Exam Topic 3)

You have an Azure Synapse Analytics dedicated SQL pool named Pool1 and a database named DB1. DB1 contains a fact table named Table1.

You need to identify the extent of the data skew in Table1. What should you do in Synapse Studio?



- A. Connect to the built-in pool and run dbcc pdw\_showspaceused.
- B. Connect to the built-in pool and run dbcc checkalloc.
- C. Connect to Pool1 and query sys.dm\_pdw\_node\_scacus.
- D. Connect to Pool1 and query sys.dm\_pdw\_nodes\_db\_partition\_scacs.

**Answer:** A

**Explanation:**

A quick way to check for data skew is to use DBCC PDW\_SHOWSPACEUSED. The following SQL code returns the number of table rows that are stored in each of the 60 distributions. For balanced performance, the rows in your distributed table should be spread evenly across all the distributions.  
 DBCC PDW\_SHOWSPACEUSED('dbo.FactInternetSales'); Reference:  
<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribu>

**NEW QUESTION 271**

- (Exam Topic 3)

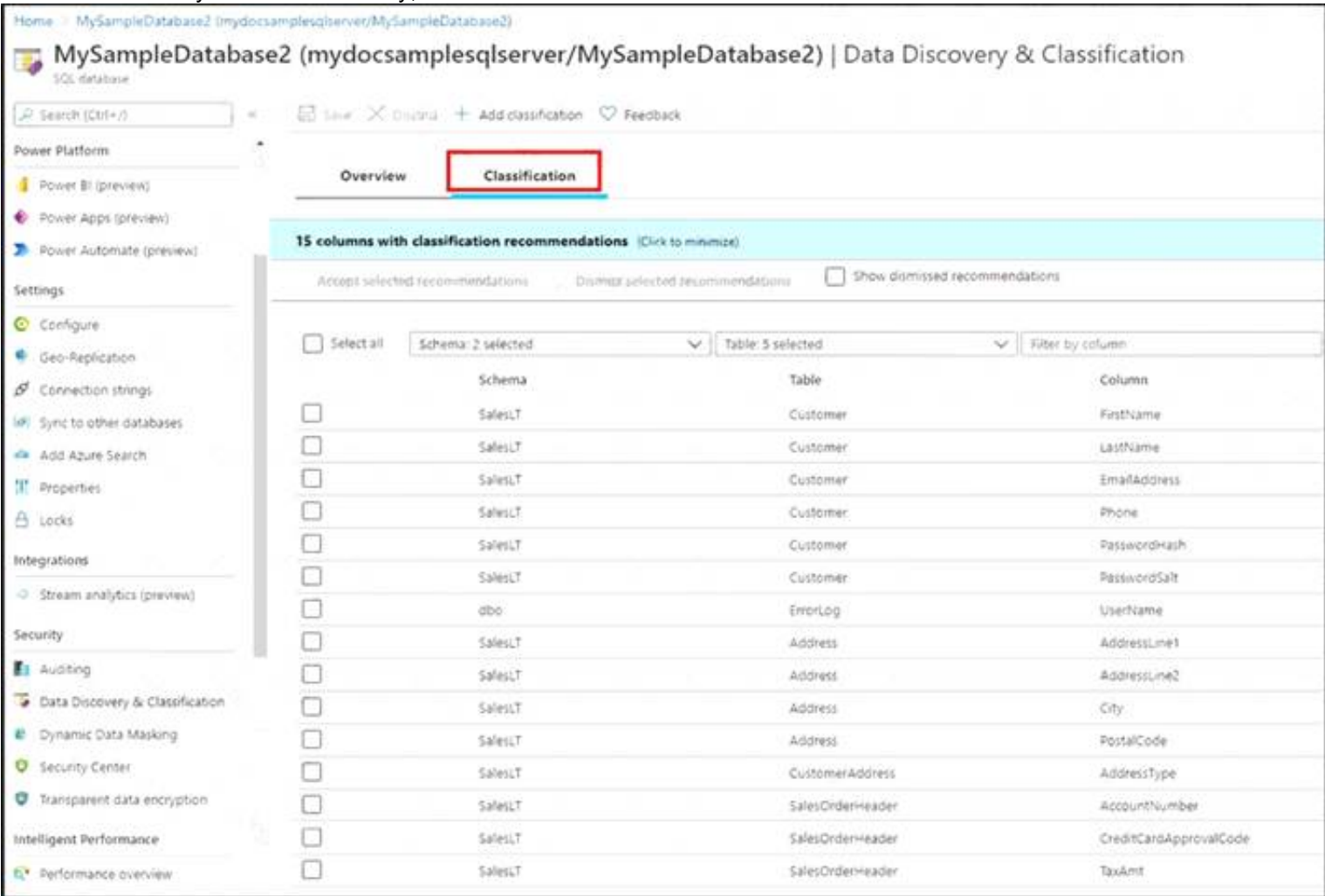
You plan to create an Azure Synapse Analytics dedicated SQL pool.  
 You need to minimize the time it takes to identify queries that return confidential information as defined by the company's data privacy regulations and the users who executed the queues.  
 Which two components should you include in the solution? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

- A. sensitivity-classification labels applied to columns that contain confidential information
- B. resource tags for databases that contain confidential information
- C. audit logs sent to a Log Analytics workspace
- D. dynamic data masking for columns that contain confidential information

**Answer:** AC

**Explanation:**

A: You can classify columns manually, as an alternative or in addition to the recommendation-based classification:



- > Select Add classification in the top menu of the pane.
- > In the context window that opens, select the schema, table, and column that you want to classify, and the information type and sensitivity label.
- > Select Add classification at the bottom of the context window.

C: An important aspect of the information-protection paradigm is the ability to monitor access to sensitive data. Azure SQL Auditing has been enhanced to include a new field in the audit log called data\_sensitivity\_information. This field logs the sensitivity classifications (labels) of the data that was returned by a query. Here's an example:

| d | client_ip | application_name                               | duration_milliseconds | response_rows | affected_rows | connection_id     | data_sensitivity_information      |
|---|-----------|--|-----------------------|---------------|---------------|-------------------|-----------------------------------|
|   | 7.125     | Microsoft SQL Server Management Studio - Query | 1                     | 847           | 847           | C244A066-2271-... | Confidential - GDPR               |
|   | 7.125     | Microsoft SQL Server Management Studio - Query | 2                     | 32            | 32            | C244A066-2271-... | Confidential                      |
|   | 7.125     | Microsoft SQL Server Management Studio - Query | 41                    | 32            | 32            | A7088FD4-759E-... | Confidential, Confidential - GDPR |

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

**NEW QUESTION 272**

- (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 during the last five minutes every five minutes. Each tweet must only be counted once.

Which windowing function should you use?

- A. a five-minute Session window
- B. a five-minute Sliding window
- C. a five-minute Tumbling window
- D. a five-minute Hopping window that has one-minute hop

**Answer:** C

**Explanation:**

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

References:

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

**NEW QUESTION 276**

.....

## Thank You for Trying Our Product

### We offer two products:

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

2nd - Questions and Answers in PDF Format

### DP-203 Practice Exam Features:

- \* DP-203 Questions and Answers Updated Frequently
- \* DP-203 Practice Questions Verified by Expert Senior Certified Staff
- \* DP-203 Most Realistic Questions that Guarantee you a Pass on Your FirstTry
- \* DP-203 Practice Test Questions in Multiple Choice Formats and Updatesfor 1 Year

**100% Actual & Verified — Instant Download, Please Click**  
**[Order The DP-203 Practice Test Here](#)**