Question 1: **Incorrect**
Your data engineering team has an Azure Stream Analytics job in place. Currently the job is configured to take in events from an Azure Event Hub. It then outputs data to an Azure Dedicated SQL pool within Azure Synapse Analytics. The engineers have been reviewing the metrics. They are seeing a high number of Backlogged input events. Which of the following can be done to ensure the Backlogged input events are kept in check?

- **Add another output to the Stream Analytics job**

  (Incorrect)

- **Change the partition key of the incoming stream**

- **Add another input to the Stream Analytics job**

- **Increase the number of streaming units assigned to the job**

  (Correct)

Explanation
One reason for a high value of Backlogged Inputs events could be because the job is not able to keep up with the incoming stream. By adding more streaming units , it can help to add more resources to ensure the job can keep up with the incoming streams.

For more information on monitoring for a Stream Analytics job, one can visit the following URL

https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-monitoring

Question 2: **Correct**
Your data engineering team is planning on setting up a dedicated SQL pool in an Azure Synapse Analytics workspace. A separate set of users will be responsible for loading data into the SQL pool. And another set of users will be responsible for querying of data from the SQL pool. You have to ensure that the loading process has enough resources assigned to it. Which of the following can be implemented for this requirement?

- **Assign more resources via workload classification**

  (Correct)

- **Make sure to use the COPY statement while loading the data**

- **Make use of materialized views**

Explanation
You need to make use of Workload Classifiers to ensure that more resources are allocated to the users who will be performing the load process.

For more information on Workload Classifiers, one can visit the following URL

https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-workload-classification

Question 3: **Correct**
Your team is designing the tables for a data warehouse. The data warehouse is going to be hosted in a Dedicated SQL pool in Azure Synapse Analytics. The following tables are going to be hosted initially in the pool

Table name	Description
Sales	This is a Fact table. The size of table is around 6 GB.
Customer	This is a dimension table that will be used along with the fact table in queries
Date	This is a dimension table that will be used along with the fact table in queries

You have to choose the right distribution for each table. You have to ensure data movement is minimized across tables

Which of the following distribution type would you choose for the Sales table?

- **Hash**

  (Correct)

- **Round Robin**
- **Replicated**

Explanation

Here since this is a large fact table, you should use Hash distributed tables

For more information on table distribution, one can visit the following URL

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

Question 4: **Incorrect**

Your team is designing the tables for a data warehouse. The data warehouse is going to be hosted in a Dedicated SQL pool in Azure Synapse Analytics. The following tables are going to be hosted initially in the pool

Table name	Description
Sales	This is a Fact table. The size of table is around 6 GB.
Customer	This is a dimension table that will be used along with the fact table in queries
Date	This is a dimension table that will be used along with the fact table in queries

You have to choose the right distribution for each table. You have to ensure data movement is minimized across tables

Which of the following distribution type would you choose for the Customer table?

- **Hash**

  (Incorrect)
- **Round Robin**
-

Replicated

(Correct)

Explanation
Here since this is a dimension table and you need to ensure data movement is minimized , you should choose replicated tables so that the data is available across all nodes in the SQL pool.

For more information on replicated table design, one can visit the following URL

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

Question 5: **Incorrect**
Your team is designing the tables for a data warehouse. The data warehouse is going to be hosted in a Dedicated SQL pool in Azure Synapse Analytics. The following tables are going to be hosted initially in the pool

Table name	Description
Sales	This is a Fact table. The size of table is around 6 GB.
Customer	This is a dimension table that will be used along with the fact table in queries
Date	This is a dimension table that will be used along with the fact table in queries

You have to choose the right distribution for each table. You have to ensure data movement is minimized across tables

Which of the following distribution type would you choose for the Date table?

- **Hash**
- **Round Robin**

  (Incorrect)
- **Replicated**

(Correct)

Explanation
Here since this is a dimension table and you need to ensure data movement is minimized , you should choose replicated tables so that the data is available across all nodes in the SQL pool.

For more information on replicated table design, one can visit the following URL

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

Question 6: **Incorrect**
Your team has several Azure Stream Analytics jobs in place. They need to make use of several windowing functions based on the needed requirement. Which of the following windowing function can be used for the below requirement?

"Ensure that the data stream is segmented into distinct time segments and ensure that events don't overlap."

- **Sliding window**

  (Incorrect)

- **Session window**

- **Tumbling window**

  (Correct)

- **Hopping window**

Explanation
Here we need to use the Tumbling window.

For more information on Azure Stream Analytics windowing functions, one can visit the following URL

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

Question 7: **Incorrect**
Your team is going to make use of Azure Data Lake Gen 2 storage accounts for storage of data. Data will be uploaded to Azure Data Lake Gen 2 storage account via a pipeline in Azure Data Factory. The pipeline will run once every day.

You have to design the storage access for the storage account based on the following requirements

1) During the first 2 weeks, the data in the storage account will be accessed frequently

2) After 2 weeks, the data will be accessed less frequently. But the data needs to be accessed immediately whenever required.

3) After 3 months the data will be rarely accessed. Whenever an object is required , an SLA of one day is in place to have the object in place.

You have to ensure data storage costs are minimized

Which of the following access tier would you use for the objects during the first 2 weeks?

- **Archive**

  (Incorrect)

- **Cool**

- **Hot**

  (Correct)

Explanation
Here since the objects need to be accessed immediately, then we need to opt for the Hot access tier.

For more information on configuring access tiers for Azure Blob storage , one can visit the following URL

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

Question 8: **Incorrect**
Your team is going to make use of Azure Data Lake Gen 2 storage accounts for storage of data. Data will be uploaded to Azure Data Lake Gen 2 storage account via a pipeline in Azure Data Factory. The pipeline will run once every day.

You have to design the storage access for the storage account based on the following requirements

1) During the first 2 weeks, the data in the storage account will be accessed frequently

2) After 2 weeks, the data will be accessed less frequently. But the data needs to be accessed immediately whenever required.

3) After 3 months the data will be rarely accessed. Whenever an object is required , an SLA of one day is in place to have the object in place.

You have to ensure data storage costs are minimized

Which of the following access tier would you use for the objects after the first 2 weeks and before 3 months?

- **Archive**

  (Incorrect)

- **Cool**

  (Correct)

- **Hot**

Explanation

Here since the objects are not accessed that frequently, we can choose the Cool Access tier. Here we will not choose the Archive Access tier because the objects need to be accessed immediately whenever required.

For more information on configuring access tiers for Azure Blob storage , one can visit the following URL

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

Question 9: **Correct**
Your team is going to make use of Azure Data Lake Gen 2 storage accounts for storage of data. Data will be uploaded to Azure Data Lake Gen 2 storage account via a pipeline in Azure Data Factory. The pipeline will run once every day.

You have to design the storage access for the storage account based on the following requirements

1) During the first 2 weeks, the data in the storage account will be accessed frequently

2) After 2 weeks, the data will be accessed less frequently. But the data needs to be accessed immediately whenever required.

3) After 3 months the data will be rarely accessed. Whenever an object is required , an SLA of one day is in place to have the object in place.

You have to ensure data storage costs are minimized

Which of the following access tier would you use for the objects after 3 months?

- **Archive**

  (Correct)

- **Cool**

- **Hot**

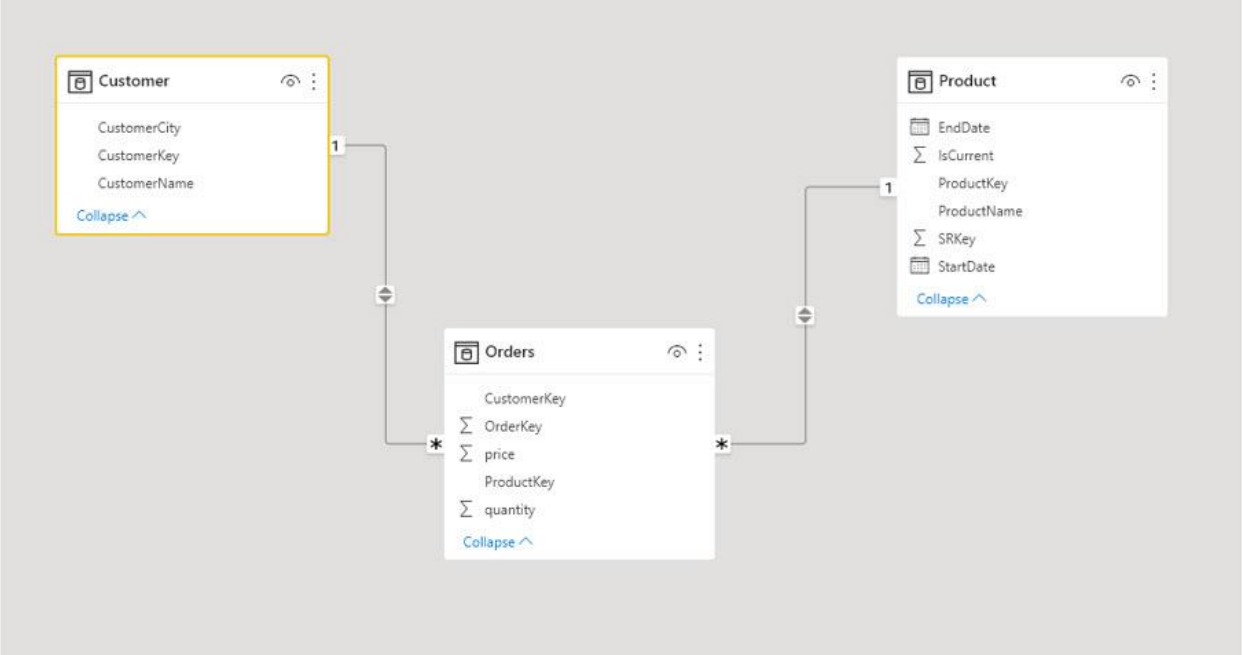
Explanation

Here we can opt for the Archive Access tier since the objects are rarely accessed. Also, with the SLA of one day, that can be taken to rehydrate the object whenever required.

For more information on configuring access tiers for Azure Blob storage , one can visit the following URL

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

Question 10: **Correct**
Your data engineering team is developing a data analytics solution. Part of the solution is to develop a data warehousing environment. Initially the below table design has been proposed.



Which of the following design is this tending towards?

- **Star Schema**

  (Correct)

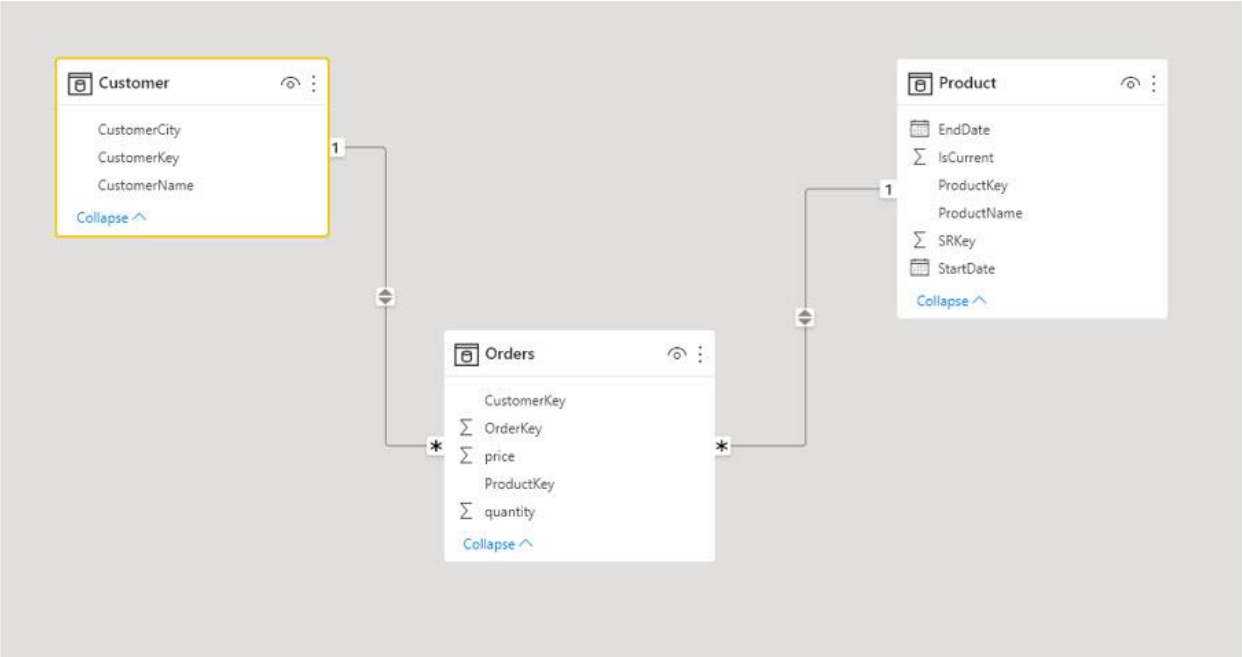
- **Snowflake Schema**

Explanation
Here this is going towards a star schema where you will be having Fact and single Dimension tables

For more information on understanding the star schema , one can visit the following URL

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

Question 11: **Incorrect**
Your data engineering team is developing a data analytics solution. Part of the solution is to develop a data warehousing environment. Initially the below table design has been proposed.



What type of table is the Orders table going to be?

- **Dimension**

  (Incorrect)

- **Fact**

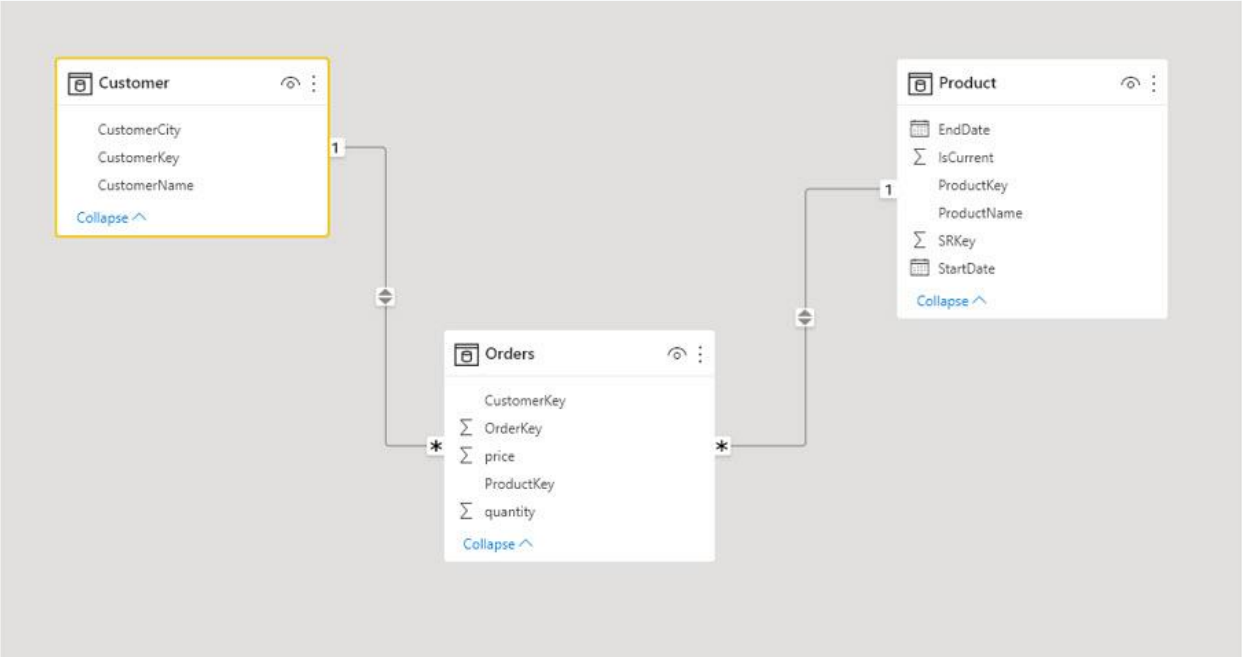
  (Correct)

Explanation
As per the Star schema design, the Orders table is going to be a Fact table

For more information on understanding the star schema , one can visit the following URL

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

Question 12: **Correct**
Your data engineering team is developing a data analytics solution. Part of the solution is to develop a data warehousing environment. Initially the below table design has been proposed.



What type of table is the Customers table going to be?

- **Dimension**

  (Correct)

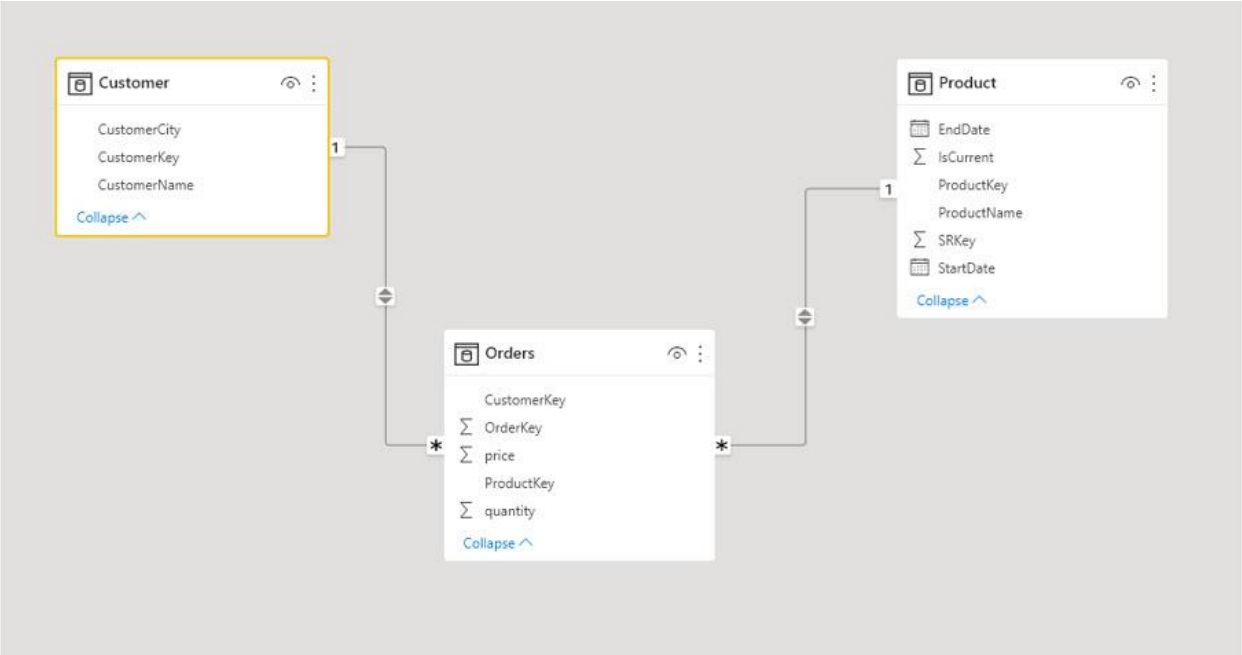
- **Fact**

Explanation
As per the Star schema design, the Customers table is going to be a Dimension table

For more information on understanding the star schema , one can visit the following URL

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

Question 13: **Incorrect**
Your data engineering team is developing a data analytics solution. Part of the solution is to develop a data warehousing environment. Initially the below table design has been proposed.



What type of dimension is the Product Table designed to be?

- **Type 0**

  (Incorrect)

- **Type 1**

- **Type 2**

  (Correct)

Explanation

The Product table is designed to be a Type 2 Slowly changing dimension. Here the table has the additional columns of StartDate, EndDate and IsCurrent. This design suggests that it is going to be a Type 2 Slowly Changing Dimension.

For more information on understanding the star schema , one can visit the following URL

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

Question 14: **Incorrect**
Your team needs to deploy an Azure Data Lake Gen2 storage account. You have to ensure that the Storage account remains available even if there is a region-level failure. Costs need to be minimized wherever possible.

Which of the following do you need to enable when deploying an Azure General Purpose V2 Storage account to ensure that is behaves as a Data Lake Gen2 Storage account?

- **Enable storage account key access**

  (Incorrect)

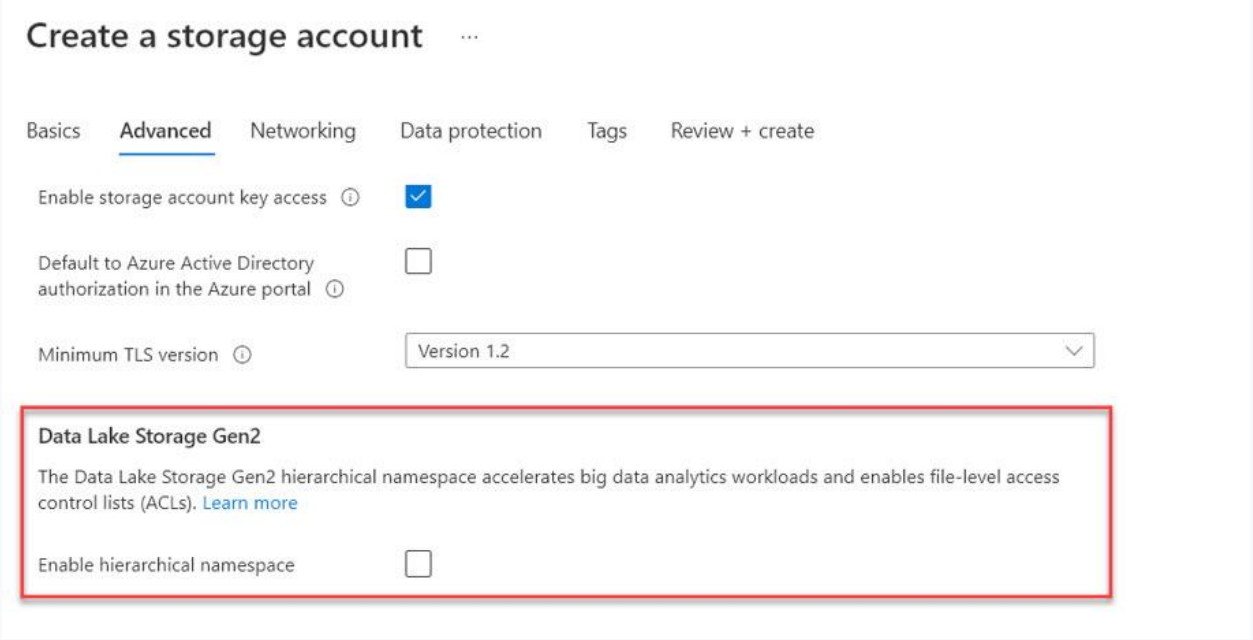
- **Enable hierarchical namespace**

  (Correct)

- **Access tier set to the Hot Access tier**

- **Enable large file shares**

Explanation
When creating an Azure Data Lake Gen 2 storage account, for a normal General Purpose V2 storage account, you need to enable the hierarchical namespace.



Question 15: **Incorrect**
Your team currently has the following resources defined on Azure

1) An Azure Data Lake Gen2 Storage account

2) An Azure Databricks cluster

A Notebook is being developed in Scala in Azure Databricks. The Notebook will take data from the Azure Data Lake Gen2 storage account as batch updates and save the data onto a delta table.

Below is a snippet of the code that needs to be completed

df. Area 1 .format("delta"). Area 2 ("append"). Area 3 ("/mnt/delta/data")

Which of the following would go into Area 1?

- **save**
- **saveAsTable**

(Incorrect)

- **mode**
- **write**

  (Correct)
- **stream**

Explanation
Here since we are performing batch updates, we can make use of the write method.

For more information on working with batch workloads in delta tables , one can visit the following URL

https://docs.microsoft.com/en-us/azure/databricks/delta/delta-batch

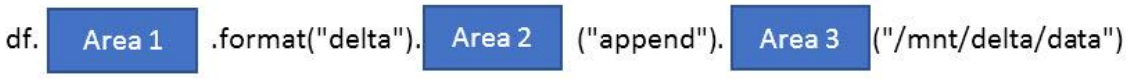
Question 16: **Correct**
Your team currently has the following resources defined on Azure

1) An Azure Data Lake Gen2 Storage account

2) An Azure Databricks cluster

A Notebook is being developed in Scala in Azure Databricks. The Notebook will take data from the Azure Data Lake Gen2 storage account as batch updates and save the data onto a delta table.

Below is a snippet of the code that needs to be completed



Which of the following would go into Area 2?

- save
- saveAsTable
- mode

  (Correct)
- write
- stream

Explanation
Here the append keyword is being used , that means the mode is append. We are appending data onto the table.

For more information on working with batch workloads in delta tables , one can visit the following URL

https://docs.microsoft.com/en-us/azure/databricks/delta/delta-batch

Question 17: **Incorrect**
Your team currently has the following resources defined on Azure

1) An Azure Data Lake Gen2 Storage account

2) An Azure Databricks cluster

A Notebook is being developed in Scala in Azure Databricks. The Notebook will take data from the Azure Data Lake Gen2 storage account as batch updates and save the data onto a delta table.

Below is a snippet of the code that needs to be completed

Which of the following would go into Area 3?

- **save**

  (Correct)

- **saveAsTable**

- **mode**

- **write**

  (Incorrect)

- **stream**

Explanation
Here we are saving it onto a location, so we need to use the save option.

For more information on working with batch workloads in delta tables , one can visit the following URL

https://docs.microsoft.com/en-us/azure/databricks/delta/delta-batch

Question 18: **Incorrect**
You have to design a Data Analytics solution for your company. You need to decide on the services that are going to be used based on the below requirements

Service	Requirement
Service 1	Be able to hold a large number of files. The files will be in different formats that includes CSV, JSON and Parquet-based files.
Service 2	Provide Spark clusters for processing of data. This needs to be a managed service without the need to manage the clusters. Have support for a variety of programming languages that includes Python and R. Have the ability to terminate clusters based on cluster inactivity.
Service 3	Provide an environment to support an Analytical data store.

Which of the following would you consider for Service 1?

- **Azure Synapse Analytics**
- **Azure Databricks**
- **Azure SQL Database**

  (Incorrect)
- **Azure Data Lake Gen2**

  (Correct)

Explanation
We can use Azure Data Lake Gen2 storage accounts as a repository for the data. Here the storage scales on demand. And you can store files in various formats.

For more information on Azure Data Lake Gen2 Storage accounts , one can visit the following URL

https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-introduction

Question 19: **Incorrect**

You have to design a Data Analytics solution for your company. You need to decide on the services that are going to be used based on the below requirements

Service	Requirement
Service 1	Be able to hold a large number of files. The files will be in different formats that includes CSV, JSON and Parquet-based files.
Service 2	Provide Spark clusters for processing of data. This needs to be a managed service without the need to manage the clusters. Have support for a variety of programming languages that includes Python and R. Have the ability to terminate clusters based on cluster inactivity.
Service 3	Provide an environment to support an Analytical data store.

Which of the following would you consider for Service 2?

- **Azure Synapse Analytics**
- **Azure Databricks**

  (Correct)
- **Azure SQL Database**

  (Incorrect)
- **Azure Data Lake Gen2**

Explanation

For the processing needs we can use Azure Databricks. Here you can provision Spark clusters in which you can develop Notebooks in a variety of languages. It also has support for auto-termination of clusters.

For more information on Azure Databricks , one can visit the following URL

https://docs.microsoft.com/en-us/azure/databricks/scenarios/what-is-azure-databricks

Question 20: **Incorrect**

You have to design a Data Analytics solution for your company. You need to decide on the services that are going to be used based on the below requirements

Service	Requirement
Service 1	Be able to hold a large number of files. The files will be in different formats that includes CSV, JSON and Parquet-based files.
Service 2	Provide Spark clusters for processing of data. This needs to be a managed service without the need to manage the clusters. Have support for a variety of programming languages that includes Python and R. Have the ability to terminate clusters based on cluster inactivity.
Service 3	Provide an environment to support an Analytical data store.

Which of the following would you consider for Service 3?

- **Azure Synapse Analytics**

  (Correct)

- **Azure Databricks**

- **Azure SQL Database**

  (Incorrect)

- **Azure Data Lake Gen2**

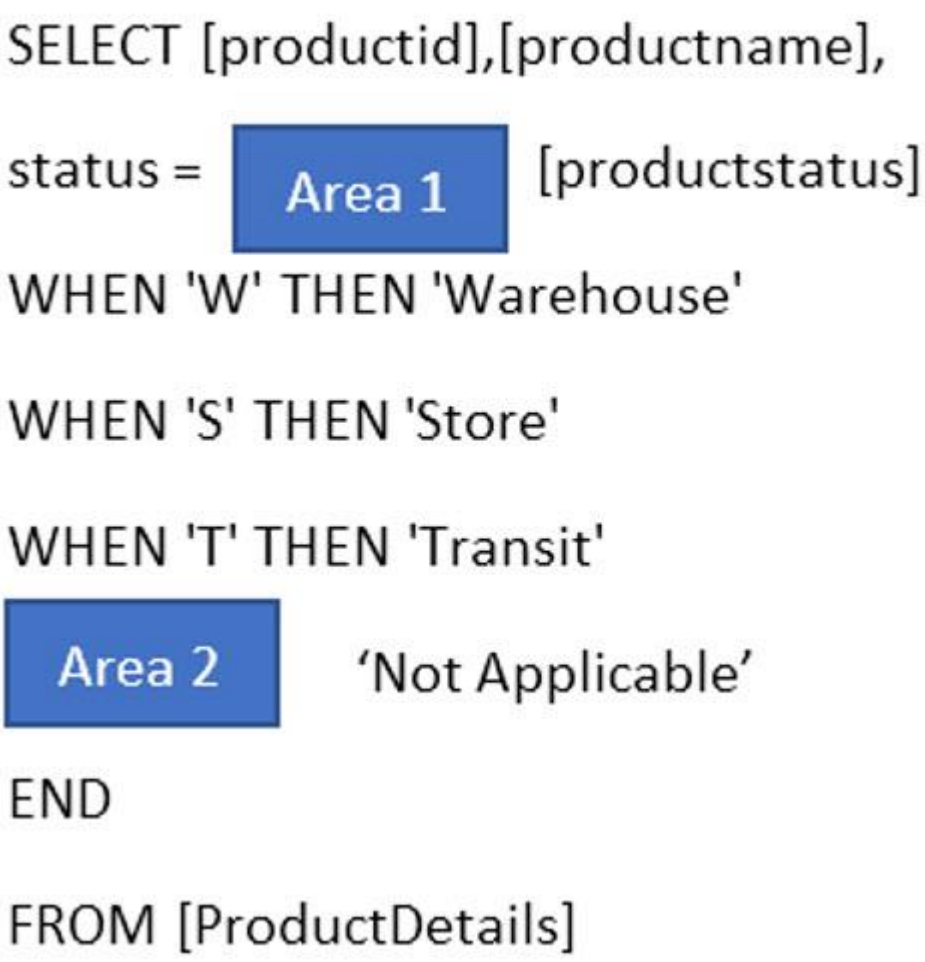
Explanation

As the Analytical data store, we can provision a Dedicated SQL Pool to serve as a SQL data warehouse in Azure Synapse.

For more information on Azure Synapse , one can visit the following URL

https://docs.microsoft.com/en-us/azure/synapse-analytics/overview-what-is

Question 21: **Incorrect**
You have a table named ProductDetails hosted in a Dedicated SQL Pool in an Azure Synapse Analytics workspace. You have to segregate the status of each product in the table. Below is the SQL statement that needs to be completed for this requirement



Which of the following would come in Area 1?

- UPDATE
- SELECT

  (Incorrect)
-

CASE

(Correct)

- **ELSE**

Explanation

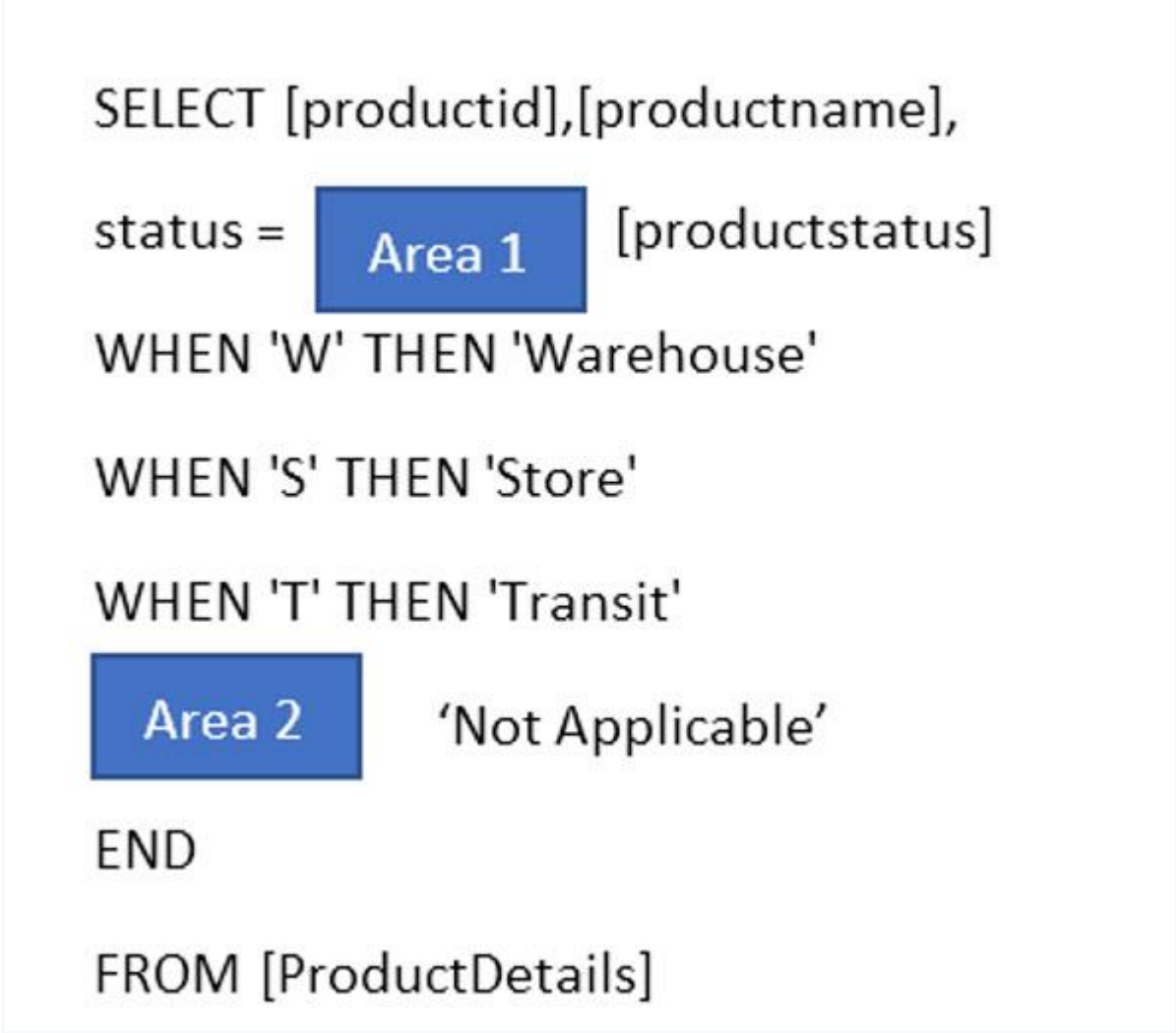
Here we can compare the different product status with the use of the CASE statement

For more information on the CASE statement , one can visit the following URL

https://docs.microsoft.com/en-us/sql/t-sql/language-elements/case-transact-sql

Question 22: **Incorrect**

You have a table named ProductDetails hosted in a Dedicated SQL Pool in an Azure Synapse Analytics workspace. You have to segregate the status of each product in the table. Below is the SQL statement that needs to be completed for this requirement



Which of the following would come in Area 2?

- UPDATE
- SELECT
- CASE

  (Incorrect)
-

ELSE

(Correct)

Explanation
Here we put the final condition with the use of the ELSE clause

For more information on the CASE statement , one can visit the following URL

https://docs.microsoft.com/en-us/sql/t-sql/language-elements/case-transact-sql

Question 23: **Incorrect**
You have to develop the SQL statement for an Azure Stream Analytics Job. The Job will take inputs from two separate Azure Event Hubs. And then write the data to a table in a Dedicated SQL pool in an Azure Synapse Analytics workspace.

Below is the script that needs to be completed

```
WITH step1 AS (SELECT * FROM EventHubInput1 PARTITION BY LogID),
step2 AS (SELECT * FROM EventHubInput1 PARTITION BY LogID)

SELECT * INTO TableOutput FROM step1 PARTITION BY [Area 1] UNION step2 PARTITION BY
[Area 2]
```

Which of the following would go into Area 1?

- **TIMESTAMP**
- **CreatedAt**

  (Incorrect)
- **LogID**

(Correct)

Explanation
Here the data is being repartitioned based on the LogID that could be part of the input data stream. You have to ensure that the stream scheme key and count of each stream in the same. The output scheme is then matching the input stream scheme key

For more information on repartitioning data , one can visit the following URL

https://docs.microsoft.com/en-us/azure/stream-analytics/repartition

Question 24: **Incorrect**
You have to develop the SQL statement for an Azure Stream Analytics Job. The Job will take inputs from two separate Azure Event Hubs. And then write the data to a table in a Dedicated SQL pool in an Azure Synapse Analytics workspace.

Below is the script that needs to be completed

Which of the following would go into Area 2?

- **TIMESTAMP**
- **CreatedAt**

  (Incorrect)
- **LogID**

(Correct)

Explanation
Here the data is being repartitioned based on the LogID that could be part of the input data stream. You have to ensure that the stream scheme key and count of each stream in the same. The output scheme is then matching the input stream scheme key

For more information on repartitioning data , one can visit the following URL

https://docs.microsoft.com/en-us/azure/stream-analytics/repartition

Question 25: **Incorrect**
You have an Azure Databricks cluster. You want to keep the configuration of the cluster even after it is terminated. Which of the following can you do for this requirement?

- **Create a notebook in the cluster with the cluster configuration**

  (Incorrect)

- **Pin the cluster**

  (Correct)

- **Configure the cluster init scripts**

Explanation
If you want to maintain the configuration of the cluster, you just need to Pin the cluster.

For more information on managing clusters , one can visit the following URL

https://docs.microsoft.com/en-us/azure/databricks/clusters/clusters-manage

Question 26: **Incorrect**
Your team has an Azure Data Lake Gen2 storage account. Continuous Time series-based data is going to be streamed into the Data Lake Gen2 storage account. Which of the following is the right design pattern to follow when it comes to the folder structure and file naming convention for the streaming data?

\YYYY\MM\DD\DataSet\datafile_YYYY_MM_DD.csv

(Incorrect)

- **\DataSet\YYYY\MM\DD\datafile_YYYY_MM_DD.csv**

  (Correct)

- **\DataSet\datafile_YYYY_MM_DD.csv**

Explanation
The recommendation is to ensure to have a parent folder that could specify the data source or the data set for the data. Then the format of the child folders needs to be in the form of the year , then the month and then the day. And then finally you have the file.

For more information on the best practices for Azure Data Lake Storage , one can visit the following URL

https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-best-practices

Question 27: **Correct**
Your team currently has an Azure Stream Analytics job in place. The job is used to take in data being streamed via Azure Event Hubs. Here log-based metrics from an application is being streamed from Azure Event Hubs onto the Stream Analytics job.

You have to find the difference in time between the First and the Final Event in the stream over a 2-hour duration.

You have to complete the below script for this requirement

```
SELECT
        metricName,
    Area 1 (
            second,
                Area 2 (Time) OVER (PARTITION BY metricName, Area 3 DURATION(hour,
2)
WHEN Event = 'First'),
                Time) as duration
FROM LogInput TIMESTAMP BY Time
WHERE
        Event = 'Final'
```

Which of the following should come in Area 1?

- **LIMIT**
- **LAST**
- **COLLATE**
- **DATEDIFF**

(Correct)

Explanation
Here we need to use the DATEDIFF function to find the time difference

This question is based on the example in the below documentation link

https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-stream-analytics-query-patterns

Question 28: **Incorrect**
Your team currently has an Azure Stream Analytics job in place. The job is used to take in data being streamed via Azure Event Hubs. Here log-based metrics from an application is being streamed from Azure Event Hubs onto the Stream Analytics job.

You have to find the difference in time between the First and the Final Event in the stream over a 2-hour duration.

You have to complete the below script for this requirement

```
SELECT
        metricName,
    Area 1 (
            second,
            Area 2 (Time) OVER (PARTITION BY metricName, Area 3 DURATION(hour,
2)
WHEN Event = 'First'),
            Time) as duration
FROM LogInput TIMESTAMP BY Time
WHERE
        Event = 'Final'
```

Which of the following should come in Area 2?

- LIMIT
- LAST

  (Correct)
-

COLLATE

(Incorrect)

- **DATEDIFF**

Explanation
Here we use the LAST function to retrieve the last event

This question is based on the example in the below documentation link

https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-stream-analytics-query-patterns

Question 29: **Incorrect**
Your team currently has an Azure Stream Analytics job in place. The job is used to take in data being streamed via Azure Event Hubs. Here log-based metrics from an application is being streamed from Azure Event Hubs onto the Stream Analytics job.

You have to find the difference in time between the First and the Final Event in the stream over a 2-hour duration.

You have to complete the below script for this requirement

```
SELECT
        metricName,
    Area 1 (
            second,
            Area 2 (Time) OVER (PARTITION BY metricName, Area 3 DURATION(hour,
2)
WHEN Event = 'First'),
            Time) as duration
FROM LogInput TIMESTAMP BY Time
WHERE
        Event = 'Final'
```

Which of the following should come in Area 3?

- **LIMIT**

  (Correct)

- **LAST**

- **COLLATE**

  (Incorrect)

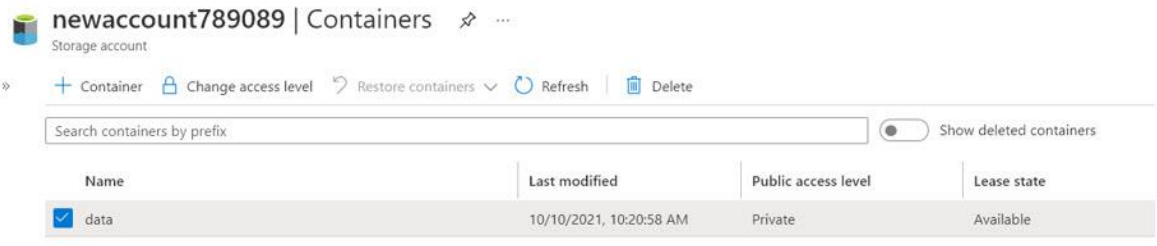
- **DATEDIFF**

Explanation
Since we need to search within the last 2 hours , we use the LIMIT function here.

This question is based on the example in the below documentation link

https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-stream-analytics-query-patterns

Question 30: **Incorrect**
Your team needs to create an external table in an Azure Synapse Serverless SQL pool. The table will be used to query parquet-based files in an Azure Data Lake Gen2 storage account. Currently the storage account container is configured as shown below



You have to ensure the Serverless SQL Pool has the right authorization to query the data in the storage account. Which of the following would you create for this requirement?

An encryption key

- **An Azure Databricks scoped secret**

  (Incorrect)

- **A database scoped credential**

  (Correct)

Explanation
Here we need to create a database scoped credential that would have the right authorization such as Shared Access Signatures. This would allow the external table to query the data in the Azure Data Lake Gen2 storage account.

For more information on working with external tables , one can visit the below URL

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

Question 31: **Correct**
Your team has several Azure Stream Analytics jobs in place. They need to make use of several windowing functions based on the needed requirement. Which of the following windowing function can be used for the below requirement?

"Ensure to output events only for points in time when the content of the window actually changes"

- **Sliding window**

  (Correct)

- **Session window**

-

Tumbling window

- **Hopping window**

Explanation
Here we need to use the Sliding window.

For more information on Azure Stream Analytics windowing functions, one can visit the following URL

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

Question 32: **Incorrect**
Your team has several Azure Stream Analytics jobs in place. They need to make use of several windowing functions based on the needed requirement. Which of the following windowing function can be used for the below requirement?

"Ensure to group events that arrive at similar times"

- **Sliding window**

  (Incorrect)

- **Session window**

  (Correct)

- **Tumbling window**

- **Hopping window**

Explanation
Here we need to use the Sliding window.

For more information on Azure Stream Analytics windowing functions, one can visit the following URL

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

Question 33: **Incorrect**
You have to design a Fact table . The table will be used to store Orders-based data. The size of the table will be around 10 GB on disk. The table needs to be partitioned based on date values.

Below is the uncompleted table definition

```
CREATE TABLE [dbo].[FactOrders]
(
   [ProductKey]      int        NOT NULL
,  [DateKey]         int        NOT NULL
,  [CustomerKey]     int        NOT NULL
,  [Number]          nvarchar(20) NOT NULL
,  [Quantity]        smallint   NOT NULL
,  [Price]           money      NOT NULL
)
WITH
(  CLUSTERED COLUMNSTORE INDEX
,  DISTRIBUTION = Area 1 ([ProductKey])

,  Area 2 ( Area 3 RANGE RIGHT FOR VALUES
            (20000101,20010101,20020101
            ,20030101,20040101,20050101
            )
          )
);
```

Which of the following would come in Area 1?

- ○

HASH

(Correct)

- **ROUND_ROBIN**

  (Incorrect)

- **REPLICATE**

Explanation
Since this is a Fact table with a large size , the preferred way for the distribution should be a hash-based distribution.

For more information on table distribution, one can visit the following URL

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

Question 34: **Incorrect**
You have to design a Fact table . The table will be used to store Orders-based data. The size of the table will be around 10 GB on disk. The table needs to be partitioned based on date values.

Below is the uncompleted table definition

```
CREATE TABLE [dbo].[FactOrders]
(
   [ProductKey]      int        NOT NULL
,  [DateKey]         int        NOT NULL
,  [CustomerKey]     int        NOT NULL
,  [Number]          nvarchar(20) NOT NULL
,  [Quantity]        smallint   NOT NULL
,  [Price]           money      NOT NULL
)
WITH
(  CLUSTERED COLUMNSTORE INDEX
,  DISTRIBUTION = Area 1 ([ProductKey])

,  Area 2 ( Area 3 RANGE RIGHT FOR VALUES
            (20000101,20010101,20020101
            ,20030101,20040101,20050101
            )
          )
);
```

Which of the following would come in Area 2?

- C

- **PARTITION**

  (Correct)

- **DateKey**

  (Incorrect)

- **SPLIT**

Explanation
Here we will use the PARTITION BY clause since we need to partition the data by dates.

For more information on table partitioning , one can visit the following URL

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

Question 35: **Correct**
You have to design a Fact table . The table will be used to store Orders-based data. The size of the table will be around 10 GB on disk. The table needs to be partitioned based on date values.

Below is the uncompleted table definition

```
CREATE TABLE [dbo].[FactOrders]
(
   [ProductKey]      int        NOT NULL
,  [DateKey]         int        NOT NULL
,  [CustomerKey]     int        NOT NULL
,  [Number]          nvarchar(20) NOT NULL
,  [Quantity]        smallint   NOT NULL
,  [Price]           money      NOT NULL
)
WITH
(  CLUSTERED COLUMNSTORE INDEX
,  DISTRIBUTION = Area 1 ([ProductKey])

,  Area 2 ( Area 3 RANGE RIGHT FOR VALUES
            (20000101,20010101,20020101
            ,20030101,20040101,20050101
            )
          )
);
```

Which of the following would come in Area 3?

- ○

PARTITION

- **DateKey**

  (Correct)

- **SPLIT**

Explanation
Here we specify the column we want to partition the table by. Since its by the dates , we mention the column name as DateKey.

For more information on table partitioning , one can visit the following URL

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

Question 36: **Incorrect**
Your team is planning on using External tables in Azure Synapse Analytics. Which of the following can be queried via the use of External tables?

- **Files in Azure File shares**

  (Incorrect)

- **Documents in Azure Cosmos DB**

- **Objects in Azure Data Lake Gen2**

  (Correct)

- **Tables in Azure SQL Databases**

Explanation

With external tables you can query for data in Azure Blob Storage or Azure Data Lake Gen2

For more information on External tables , one can visit the following URL

https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables?tabs=hadoop

Question 37: **Incorrect**
Your team is planning on using External tables in Azure Synapse Analytics. The team will be using a set of Parquet-based files hosted in an Azure Data Lake Gen2 Storage account.

Which of the following statement is used to reference the Azure Data Lake Gen2 Storage account and the associated credentials to access the account?

- **CREATE EXTERNAL FILE FORMAT**

  (Incorrect)

- **CREATE EXTERNAL TABLE**

- **CREATE EXTERNAL DATA SOURCE**

  (Correct)

Explanation
With the CREATE EXTERNAL DATA SOURCE statement, you will mention the source of data which is in Azure Data Lake Gen2 Storage account. You will also specify the credentials which will be used to access the storage account.

For more information on External tables , one can visit the following URL

https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables?tabs=hadoop

Question 38: **Correct**

Your team is planning on using External tables in Azure Synapse Analytics. The team will be using a set of Parquet-based files hosted in an Azure Data Lake Gen2 Storage account.

Which of the following statement is used to describe the format of the files?

- **CREATE EXTERNAL FILE FORMAT**

  (Correct)

- **CREATE EXTERNAL TABLE**

- **CREATE EXTERNAL DATA SOURCE**

Explanation
The CREATE EXTERNAL FILE FORMAT statement is used to specify the format of the files.

For more information on External tables , one can visit the following URL

https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables?tabs=hadoop

Question 39: **Incorrect**
Your data engineering team currently has the following resources defined in Azure

1) An Azure Event Hub – This is used to stream events from external data sources onto Azure.

2) An Azure Data Lake Gen2 Storage account – This is used to store the events streamed via Azure Event Hubs

3) An Azure Data Factory instance – This is used to build various ETL pipelines

4) An Azure Synapse Analytics workspace – This is used to host a dedicated SQL pool.

You have to build a pipeline in Azure Data Factory to copy data at regular time intervals from the Azure Data Lake Gen2 Storage account onto tables in the dedicated SQL pool.

You have to ensure that only data within a specified time window is copied onto tables in the dedicated SQL pool.

Which of the following would you choose the Integration runtime type for the pipeline?

- **Azure Integration runtime**

  (Correct)

- **Azure-SSIS Integration runtime**

  (Incorrect)

- **Self-hosted Integration runtime**

Explanation
Here since the source and destination of the data are Azure-based resources ,we can make use of the Azure Integration runtime itself.

For more information on the Azure Integration runtime , one can visit the following URL

https://docs.microsoft.com/en-us/azure/data-factory/create-azure-integration-runtime?tabs=data-factory

Question 40: **Incorrect**
Your data engineering team currently has the following resources defined in Azure

1) An Azure Event Hub – This is used to stream events from external data sources onto Azure.

2) An Azure Data Lake Gen2 Storage account – This is used to store the events streamed via Azure Event Hubs

3) An Azure Data Factory instance – This is used to build various ETL pipelines

4) An Azure Synapse Analytics workspace – This is used to host a dedicated SQL pool.

You have to build a pipeline in Azure Data Factory to copy data at regular time intervals from the Azure Data Lake Gen2 Storage account onto tables in the dedicated SQL pool.

You have to ensure that only data within a specified time window is copied onto tables in the dedicated SQL pool.

Which of the following should be used as the trigger type?

- **Event-based trigger**
- **Schedule trigger**

  (Incorrect)
- **Tumbling window trigger**

  (Correct)

Explanation
Here since we need to ensure that jobs are executed within a particular time frame, and each window is independent of the other, we should look towards using the Tumbling window trigger.

For more information on the tumbling window trigger , one can visit the following URL

https://docs.microsoft.com/en-us/azure/data-factory/how-to-create-tumbling-window-trigger?tabs=data-factory

Question 41: **Incorrect**
Your data engineering team has a table that has the following structure in a dedicated SQL pool in an Azure Synapse Analytics workspace

Column name	Column type
salesid	int
salesperson	varchar(200)
price	decimal
quantity	int

Which of the following statement can be used to implement row level security in the table?

- **CREATE DYNAMIC MASK**
- **CREATE SECURITY POLICY**

  (Correct)
- **GRANT**

  (Incorrect)
- **UPDATE**

Explanation
You can implement row-level security with the use of the statement CREATE SECURITY POLICY

For more information on row-level security , one can visit the following URL

https://docs.microsoft.com/en-us/sql/relational-databases/security/row-level-security?view=sql-server-ver15

Question 42: **Correct**
Your data engineering team has a table that has the following structure in a dedicated SQL pool in an Azure Synapse Analytics workspace

Column name	Column type
salesid	int
salesperson	varchar(200)
price	decimal
quantity	int

Which of the following statement can be used to implement column level security in the table?

- **CREATE DYNAMIC MASK**

CREATE SECURITY POLICY

- **GRANT**

  (Correct)

- **UPDATE**

Explanation
You can implement column-level security with the use of the GRANT statement

For more information on column-level security , one can visit the following URL

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

Question 43: **Correct**
Your team has created the following dimension table in a dedicated SQL pool

```
CREATE TABLE [dbo].[DimProduct](

[ProductKey] [int] IDENTITY(1,1) NOT NULL,

[ProductID] [int] NOT NULL,

[ProductModelID] [int] NOT NULL,

[ProductcategoryID] [int] NOT NULL,

[ProductName] varchar(50) NOT NULL,

[ProductNameStartDate] [date] NOT NULL,

[ProductNameEndDate] [date] NOT NULL,

[IsCurrent] [int] NOT NULL,

[ProductModelName] varchar(50) NULL,

[ProductCategoryName] varchar(50) NULL

)
```

Based on the above table structure, what type of Slowly changing dimension is the above table?

- **Type 0**
- **Type 1**
- **Type 2**

  (Correct)

Explanation
The above table is a Type 2 Slowly changing dimension table. This is because it has the additional columns of [ProductNameStartDate], [ProductNameEndDate] and [IsCurrent] to signify the row changes in the table.

Question 44: **Incorrect**
Your team has created the following dimension table in a dedicated SQL pool

```
CREATE TABLE [dbo].[DimProduct](

[ProductKey] [int] IDENTITY(1,1) NOT NULL,

[ProductID] [int] NOT NULL,

[ProductModelID] [int] NOT NULL,

[ProductcategoryID] [int] NOT NULL,

[ProductName] varchar(50) NOT NULL,

[ProductNameStartDate] [date] NOT NULL,

[ProductNameEndDate] [date] NOT NULL,

[IsCurrent] [int] NOT NULL,
```

[ProductModelName] varchar(50) NULL,

[ProductCategoryName] varchar(50) NULL

)

Which of the following is the surrogate key in the table?

- **[ProductKey]**

  (Correct)

- **[ProductID]**

- **[ProductNameStartDate]**

- **[IsCurrent]**

  (Incorrect)

Explanation
This is the [ProductKey] column. It's based on an IDENTITY which means we want to generate a number for each row entered the table. This would behave as our surrogate key.

Question 45: **Incorrect**
Your team needs to create External tables in Azure Synapse that would read data from an Azure Storage account. Most of the queries fired against the EXTERNAL table would only target a few columns. Which of the following would be the best ideal file format for the data stored in Azure Storage account when it comes to read efficiency and less file storage?

- **CSV**

  (Incorrect)

- **JSON**
- **Parquet**

  (Correct)
- **Avro**

Explanation
Here since only a few columns need to be queried and Parquet is a column-based file format, this would be best suited. It's also very effective when it comes to file storage when compared to CSV and JSON. Avro is a row-based file format and would not be the best suited for this scenario.

Question 46: **Incorrect**
Your team has created the following table

```
CREATE TABLE [logdata]

(

[Correlation id] [varchar](200) NULL,

[Operation name] [varchar](200) NULL,

[Status] [varchar](100) NULL,

[Event category] [varchar](100) NULL,

[Level] [varchar](100) NULL,

[Time] [datetime] NULL,

[Subscription] [varchar](200) NULL,
```

```
[Event initiated by] [varchar](1000) NULL,

[Resource type] [varchar](1000) NULL,

[Resource group] [varchar](1000) NULL,

[Resource] [varchar](2000) NULL

)

WITH

(

DISTRIBUTION = HASH ([Operation name]),

PARTITION ( [Time] RANGE RIGHT FOR VALUES

('2023-01-01','2023-02-01','2023-03-01','2023-04-01'))

)
```

Which of the following in the query would make use of the partitions defined in the table?

- **ORDER BY**
- **GROUP BY**

  (Incorrect)
- **JOIN**
- **WHERE**

(Correct)

Explanation
The WHERE clause in the SQL statements would make use of the partitions.
Question 47: **Incorrect**
Your team currently has a set of tables in dedicated SQL pool in Azure Synapse. Data is loaded on a weekly basis to a table within the pool. Around 1 million rows of data are added during the loading time. You have to create a staging table to facilitate the loading process. The load time needs to be minimized in the entire implementation.

Which of the following would you consider when creating the table?

- **Creating a table with a clustered column store index**

  (Incorrect)

- **Creating a table with a clustered index**

- **Creating a table with a clustered and non-clustered index**

- **Creating a heap table**

  (Correct)

Explanation
Since this is a staging table, it's more efficient to create a heap table to minimize on the data load times.
Question 48: **Incorrect**
Your team currently has a set of tables in dedicated SQL pool in Azure Synapse. Data is loaded on a weekly basis to a table within the pool. Around 1 million rows of data are added during the loading time. You have to create a staging table to facilitate the loading process. The load time needs to be minimized in the entire implementation.

Which of the following would you consider as the Distribution type for the table?

-

Hash

(Incorrect)

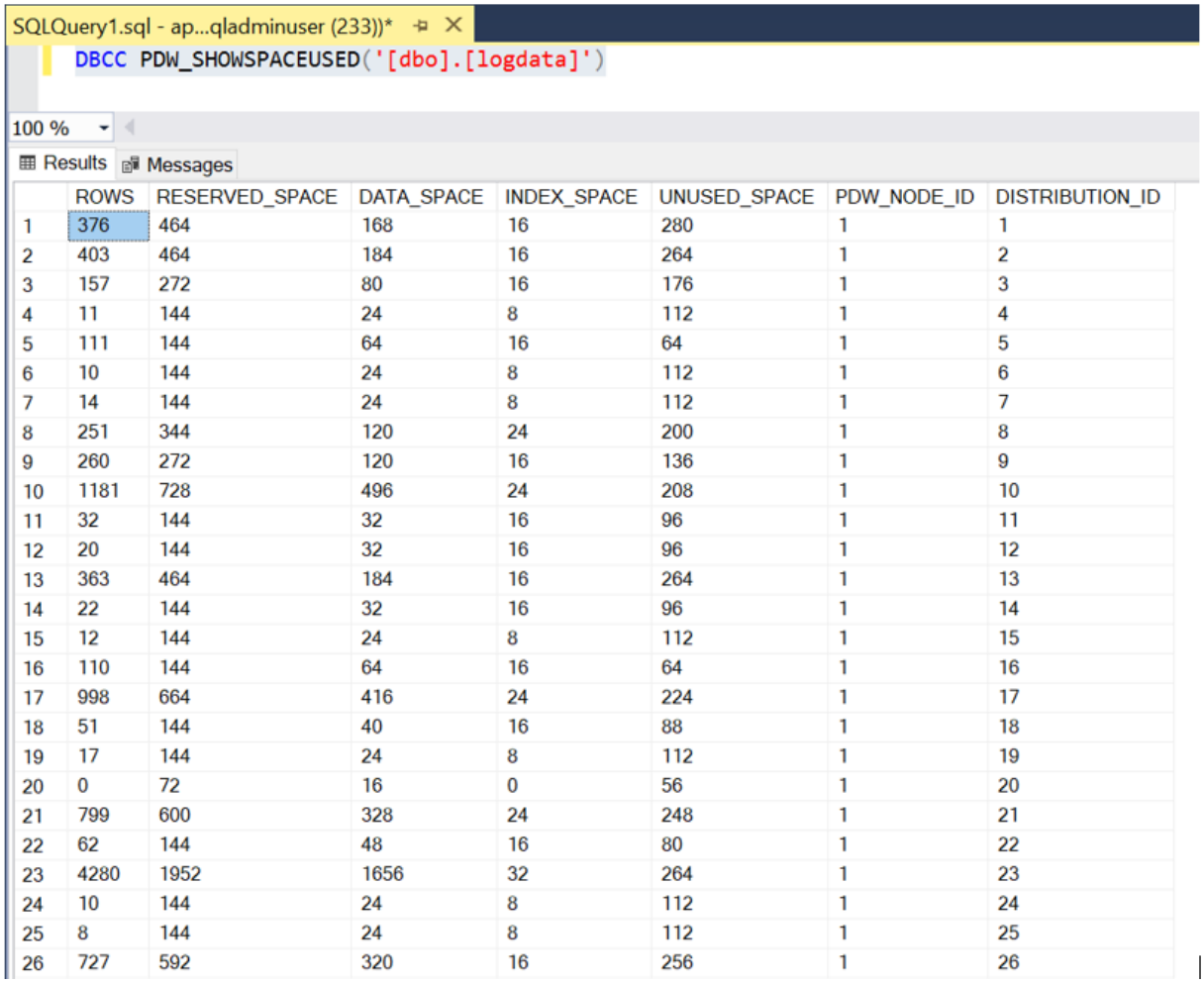
- **Replicated**

- **Round-Robin**

  (Correct)

Explanation
When it comes to fast loading of data into a table, you need to consider the Round-Robin distribution.

Question 49: **Correct**
You've run and got the following output on one of the tables in a dedicated SQL pool



	ROWS	RESERVED_SPACE	DATA_SPACE	INDEX_SPACE	UNUSED_SPACE	PDW_NODE_ID	DISTRIBUTION_ID
1	376	464	168	16	280	1	1
2	403	464	184	16	264	1	2
3	157	272	80	16	176	1	3
4	11	144	24	8	112	1	4
5	111	144	64	16	64	1	5
6	10	144	24	8	112	1	6
7	14	144	24	8	112	1	7
8	251	344	120	24	200	1	8
9	260	272	120	16	136	1	9
10	1181	728	496	24	208	1	10
11	32	144	32	16	96	1	11
12	20	144	32	16	96	1	12
13	363	464	184	16	264	1	13
14	22	144	32	16	96	1	14
15	12	144	24	8	112	1	15
16	110	144	64	16	64	1	16
17	998	664	416	24	224	1	17
18	51	144	40	16	88	1	18
19	17	144	24	8	112	1	19
20	0	72	16	0	56	1	20
21	799	600	328	24	248	1	21
22	62	144	48	16	80	1	22
23	4280	1952	1656	32	264	1	23
24	10	144	24	8	112	1	24
25	8	144	24	8	112	1	25
26	727	592	320	16	256	1	26

Does the table contain more than 5000 rows?

- **Yes**

  (Correct)

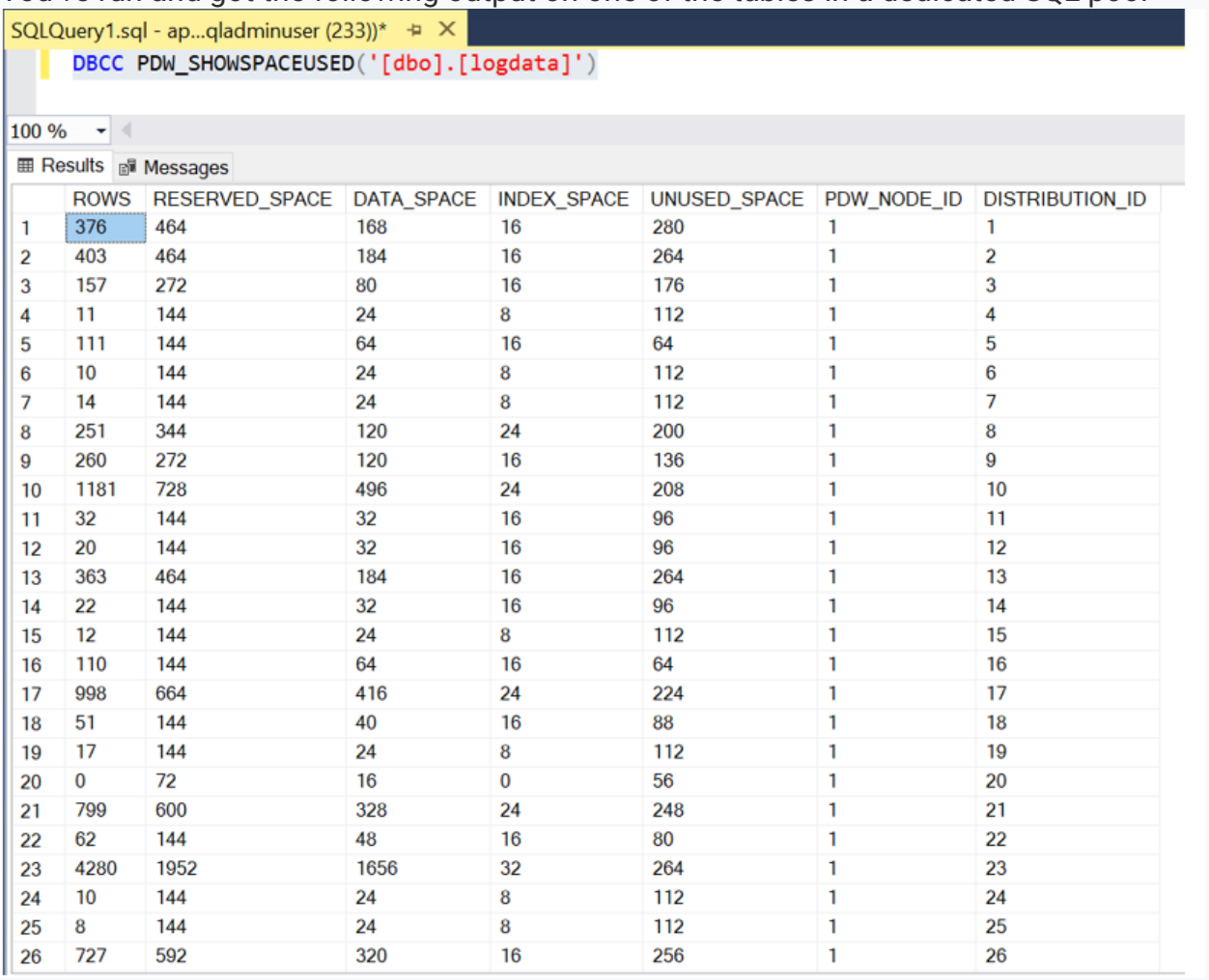
- **No**

Explanation

Yes, because if you just add the rows in the distributions shown above, it should add up to more than 5000.

Question 50: **Incorrect**

You've run and got the following output on one of the tables in a dedicated SQL pool



	ROWS	RESERVED_SPACE	DATA_SPACE	INDEX_SPACE	UNUSED_SPACE	PDW_NODE_ID	DISTRIBUTION_ID
1	376	464	168	16	280	1	1
2	403	464	184	16	264	1	2
3	157	272	80	16	176	1	3
4	11	144	24	8	112	1	4
5	111	144	64	16	64	1	5
6	10	144	24	8	112	1	6
7	14	144	24	8	112	1	7
8	251	344	120	24	200	1	8
9	260	272	120	16	136	1	9
10	1181	728	496	24	208	1	10
11	32	144	32	16	96	1	11
12	20	144	32	16	96	1	12
13	363	464	184	16	264	1	13
14	22	144	32	16	96	1	14
15	12	144	24	8	112	1	15
16	110	144	64	16	64	1	16
17	998	664	416	24	224	1	17
18	51	144	40	16	88	1	18
19	17	144	24	8	112	1	19
20	0	72	16	0	56	1	20
21	799	600	328	24	248	1	21
22	62	144	48	16	80	1	22
23	4280	1952	1656	32	264	1	23
24	10	144	24	8	112	1	24
25	8	144	24	8	112	1	25
26	727	592	320	16	256	1	26

Is there any sort of data skew in the table?

- **Yes**

  (Correct)

- **No**

  (Incorrect)

Explanation
Yes, the rows of data do not seem to uniformly distributed across the table. Also there seems to be one distribution with 0 rows.

Question 51: **Incorrect**
Your team is going to create a table with the following statement

CREATE TABLE [logdata]

(

[Correlation id] [varchar](200) NULL,

[Operation name] [varchar](200) NULL,

[Status] [varchar](100) NULL,

[Event category] [varchar](100) NULL,

[Level] [varchar](100) NULL,

[Time] [datetime] NULL,

[Subscription] [varchar](200) NULL,

[Event initiated by] [varchar](1000) NULL,

[Resource type] [varchar](1000) NULL,

[Resource group] [varchar](1000) NULL,

[Resource] [varchar](2000) NULL

)

The table is going to ingest around 30 million rows of data per month. Most of the queries fired against the table would be used to analyze the data for a given month. Which of the following would be the ideal column to partition the table by?

- **[Operation name]**
- **[Event category]**

(Incorrect)

- **[Time]**

  (Correct)

- **[Event initiated by]**

Explanation
Since the data is going to be analyzed month-wise, it's best to partition the data by the Time column.

Question 52: **Incorrect**
Your team needs to build a table in an Azure Synapse Serverless SQL pool. The table needs to reference the data stored in parquet files in an Azure Data Lake Gen2 storage account. The table needs to be built on certain columns in the data stored in Parquet files.

Which table command would you create for this requirement?

- **CREATE TABLE**

  (Incorrect)

- **CREATE EXTERNAL TABLE**

  (Correct)

Explanation
Only External tables can be created in a Serverless SQL pool.

Question 53: **Incorrect**
Your team needs to build a table in an Azure Synapse Serverless SQL pool. The table needs to reference the data stored in parquet files in an Azure Data Lake Gen2 storage account. The table needs to be built on certain columns in the data stored in Parquet files.

You need to complete the below partial SQL code that would be used to referencing the data

```
SELECT *
FROM [Area 1] (BULK 'https://datalake2000233.dfs.core.windows.net/parquet/*.parquet',
    FORMAT = 'PARQUET') AS [logdata_temp]
```

Which of the following would go into Area 1?

- **OPENROWSET**

  (Correct)

- **INSERT**

- **UPDATE**

  (Incorrect)

- **TABLE**

Explanation
Here we can use the OPENROWSET command to reference the data in the parquet files.

Question 54: **Correct**
You need to read data from a JSON file stored in an Azure Data Lake Gen2 storage account. The data is being read from a serverless SQL pool. You need to complete the below script for this requirement?

```
SELECT TOP 100
  jsonContent
FROM
  OPENROWSET(
    BULK 'https://datalake2000233.dfs.core.windows.net/json/Log.json',
    FORMAT = Area 1,
    FIELDQUOTE = '0x0b',
    FIELDTERMINATOR ='0x0b',
    ROWTERMINATOR = '0x0b'
  )
  WITH (
    jsonContent varchar(MAX)
  ) AS [rows]
```

Which of the following would go into Area 1?

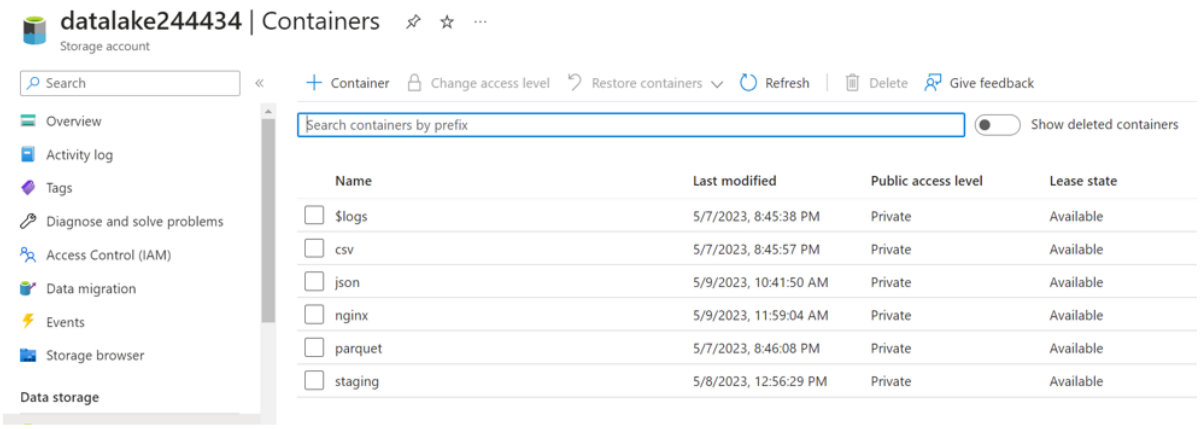
- JSON
- CSV

  (Correct)
- XML

Explanation

To read the JSON-based files, in the OPENROWSET command, you need to mention the format as CSV.

Question 55: **Correct**
You currently have the following containers within an Azure Data Lake Gen2 storage account



You need to transfer CSV-based files from the csv container to the parquet folder. The destination files need to be in parquet-format. No transformations need to be carried out on the files. The folder structure needs to be retained when it comes to transferring the files.

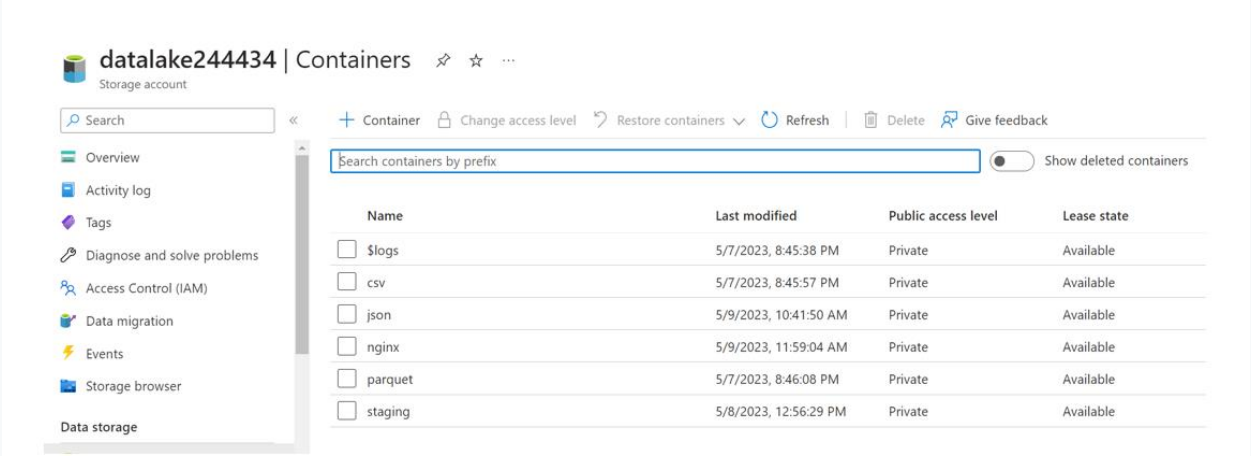
Which of the following would you specify as the Copy Behavior for this requirement?

- **Flatten hierarchy**
- **Merge files**
- **Preserve hierarchy**

  (Correct)

Explanation
Since we need to retain the folder structure , we need to choose the Preserve hierarchy option.

Question 56: **Incorrect**
You currently have the following containers within an Azure Data Lake Gen2 storage account



You need to transfer JSON-based files to a table in Azure Synapse. The JSON file contains objects of data. Each object also has a nested array structure. You have to ensure that each value in the array appears as a separate row in the table in Azure Synapse.

Which of the following transformation would you consider including in your mapping data flow for the transformation?

- **Conditional Split**
- **Aggregate**

  (Incorrect)
- **Flatten**

  (Correct)
- **Alter Row**

Explanation

Here you can use the Flatten transformation to unroll the values in the array in the JSON objects.

Question 57: **Incorrect**

Your team needs to transfer data from a parquet-based file to a table in Azure Synapse. Which of the following can be configured for the copy activity in Azure Data Factory to ensure that higher performance is achieved along with a higher degree of parallelism in the copy process?

- **Make use of Partitions in the copy process by setting the Dynamic range**

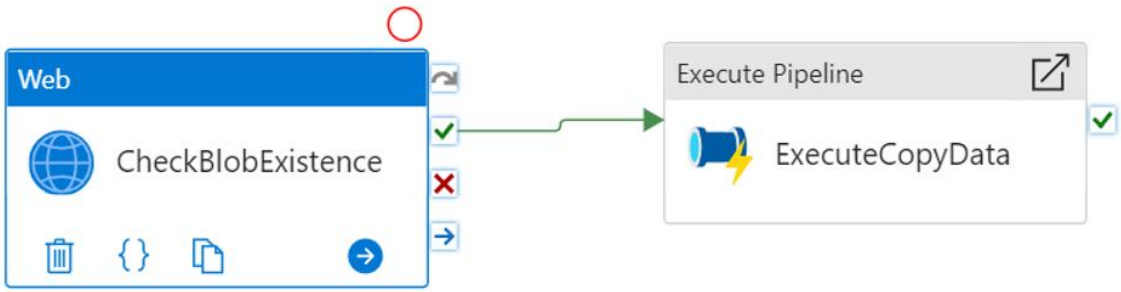
  (Incorrect)

- **Set the copy process to Bulk Insert**

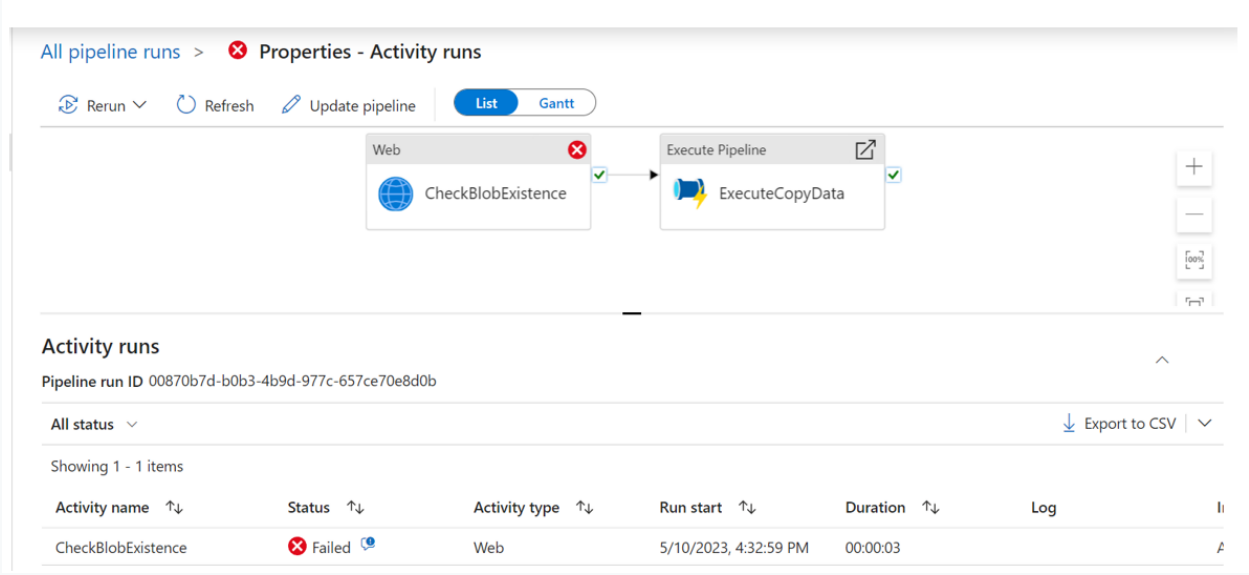
- **Set the copy process to use PolyBase**

  (Correct)

Explanation
To achieve the best performance when it comes to transferring data to Azure Synapse, you need to use PolyBase.
Question 58: Incorrect
Your team has the following pipeline in place in Azure Data Factory



Following is the pipeline run



What is the value set for the Retry option for the 'CheckBlobExistence' activity?

- **0**

  (Correct)

- **1**

- **2**

  (Incorrect)

Explanation

Here since the activity failed on the first time around, that is why the Retry was only set to 0.

Question 59: **Correct**

You need to design a pipeline in Azure Data Factory. The pipeline would need to run a stored procedure in an Azure SQL database. The stored procedure would return a list of rows which would then need to be passed onto subsequent activities in the pipeline. Which of the following activity could you use for calling the Stored Procedure?

- **Lookup**

  (Correct)

- **Web**

- **Get Metadata**

- **Filter**

Explanation

You can use the Lookup activity to call the Stored Procedure and get the output accordingly.

Question 60: **Incorrect**

Your team has set up an Azure Stream Analytics job that reads metric data from an Azure Web App resource. You have to formulate a query that would find the growth in readings for a particular metric. You need to complete the below query for this requirement.

```
SELECT
   time,
   growth=average - Area 1 (average) OVER (PARTITION BY resourceId LIMIT
DURATION(second,10))
INTO
   [dbpool]
FROM
   [webhub]
WHERE
   metricName='MemoryWorkingSet'
```

Which of the following would go into Area 1?

- TIMESTAMP BY
- LAG

  (Correct)
- LAST
- SET

  (Incorrect)

Explanation
Here we can use the LAG function to find the difference between prior events send to the Stream Analytics job.