Pytanie 1:

Pominięto

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**
- **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**

  (Poprawne)

Wyjaśnienie

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

---

Pytanie 2:

Pominięto

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**

  (Poprawne)

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

- **Make use of materialized views**

Wyjaśnienie

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

Pytanie 3:

Pominięto

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**

  (Poprawne)

- **Round Robin**

- **Replicated**

Wyjaśnienie

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

Pytanie 4:

Pominięto

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

  (Poprawne)

Wyjaśnienie

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

Pytanie 5:

Pominięto

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**
- **Replicated**

(Poprawne)

Wyjaśnienie

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

Pytanie 6:

Pominięto

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**
- **Session window**
- **Tumbling window**

  (Poprawne)
- **Hopping window**

Wyjaśnienie

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

Pytanie 7:

Pominięto

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**

  (Poprawne)

- **Session window**

- **Tumbling window**

- **Hopping window**

Wyjaśnienie

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

Pytanie 8:

Pominięto

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**
- **Session window**

  (Poprawne)
- **Tumbling window**
- **Hopping window**

Wyjaśnienie

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

Pytanie 9:

Pominięto

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

Which of the following would come in Area 1?

```
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
            )
          )
);
```

- **HASH**

  (Poprawne)

- **ROUND_ROBIN**

- **REPLICATE**

Wyjaśnienie
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

Pytanie 10:

Pominięto

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?

- **PARTITION**

  (Poprawne)

- **DateKey**

- **SPLIT**

Wyjaśnienie
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

Pytanie 11:

Pominięto

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**

  (Poprawne)
- **SPLIT**

Wyjaśnienie
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

Pytanie 12:

Pominięto

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**
- **Documents in Azure Cosmos DB**
- **Objects in Azure Data Lake Gen2**

  (Poprawne)
- **Tables in Azure SQL Databases**

Wyjaśnienie

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

Pytanie 13:

Pominięto

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**
- **CREATE EXTERNAL TABLE**
- **CREATE EXTERNAL DATA SOURCE**

  (Poprawne)

Wyjaśnienie

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

Pytanie 14:

Pominięto

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**

  (Poprawne)

- **CREATE EXTERNAL TABLE**

- **CREATE EXTERNAL DATA SOURCE**

Wyjaśnienie

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

---

Pytanie 15:

Pominięto

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**

  (Poprawne)

- **Azure-SSIS Integration runtime**

- **Self-hosted Integration runtime**

Wyjaśnienie

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

Pytanie 16:

Pominięto

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**
- **Tumbling window trigger**

  (Poprawne)

Wyjaśnienie

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

Pytanie 17:

Pominięto

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**

  (Poprawne)
- **GRANT**
- **UPDATE**

Wyjaśnienie

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

Pytanie 18:

Pominięto

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**

  (Poprawne)
- **UPDATE**

Wyjaśnienie

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

Pytanie 19:

Pominięto

Your team has an Azure Databricks workspace. They need to create two clusters. Below are the requirements for the clusters

Name	Requirement
Cluster 1	This cluster will be used by a set of Data Scientists. The cluster needs to support workloads that run on SQL and Python. The cluster should not terminate by default.
Cluster 2	This cluster will be used by a set of Data engineers. This cluster needs to run workloads that run on Scala and Python. The cluster should automatically terminate after 120 minutes.

Which of the following would you choose as the cluster mode for Cluster 1?

- **Single Node**
- **Standard**
- **High Concurrency**

  (Poprawne)

Wyjaśnienie

All of these requirements are met with the use of the High Concurrency cluster

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

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

Pytanie 20:

Pominięto

Your team has an Azure Databricks workspace. They need to create two clusters. Below are the requirements for the clusters

Name	Requirement
Cluster 1	This cluster will be used by a set of Data Scientists. The cluster needs to support workloads that run on SQL and Python. The cluster should not terminate by default.
Cluster 2	This cluster will be used by a set of Data engineers. This cluster needs to run workloads that run on Scala and Python. The cluster should automatically terminate after 120 minutes.

Which of the following would you choose as the cluster mode for Cluster 2?

- **Single Node**
- **Standard**

  (Poprawne)
- **High Concurrency**

Wyjaśnienie

We have to choose the Standard cluster because the High Concurrency cluster does not support Scala. Also, the Single Node cluster will not be effective for a set of users.

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

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

Pytanie 21:

Pominięto

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**
- **Cool**
- **Hot**

(Poprawne)

Wyjaśnienie
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

Pytanie 22:

Pominięto

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**
- **Cool**

  (Poprawne)
- **Hot**

Wyjaśnienie

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

Pytanie 23:

Pominięto

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**

  (Poprawne)

- **Cool**

- **Hot**

Wyjaśnienie

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.

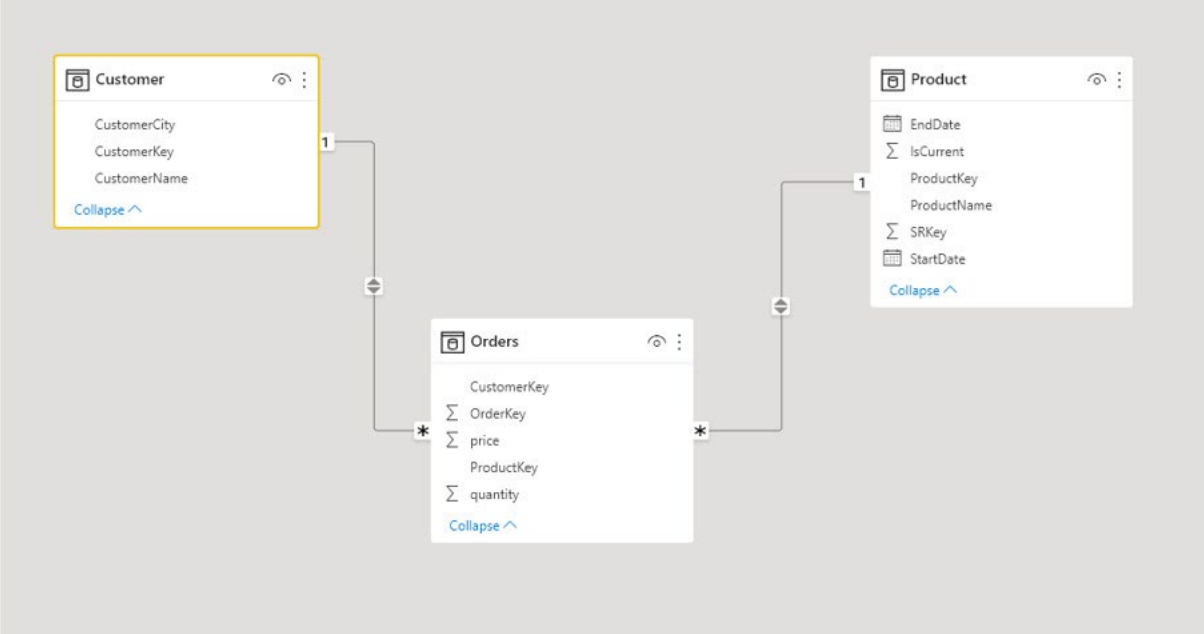
For more information on configuring access tiers for Azure Blob storage , one can visit the following URL

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

Pytanie 24:

Pominięto

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**

  (Poprawne)

- **Snowflake Schema**

Wyjaśnienie

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

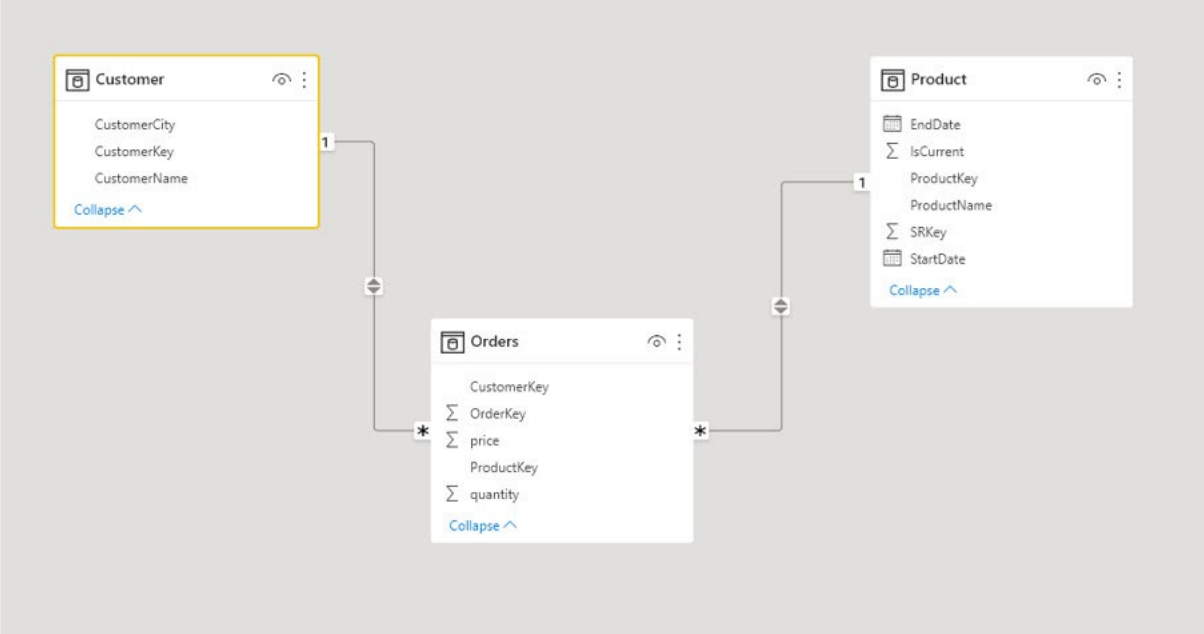
For more information on understanding the star schema , one can visit the following URL

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

Pytanie 25:

Pominięto

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**
- **Fact**

(Poprawne)

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

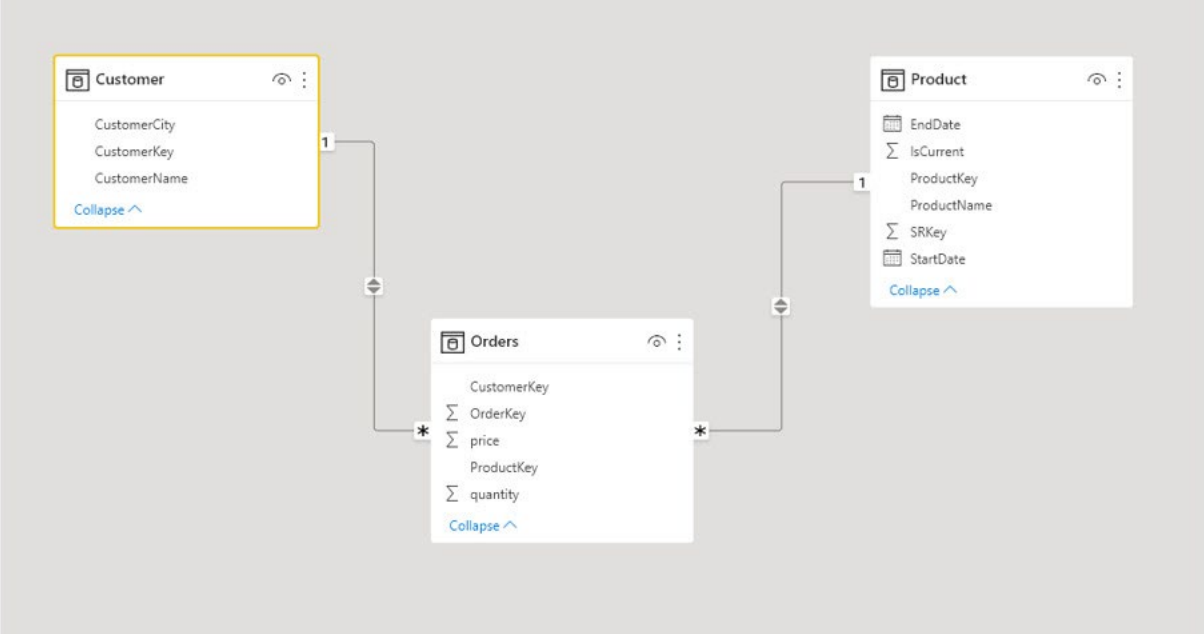
For more information on understanding the star schema , one can visit the following URL

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

Pytanie 26:

Pominięto

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**

  (Poprawne)

- **Fact**

Wyjaśnienie

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

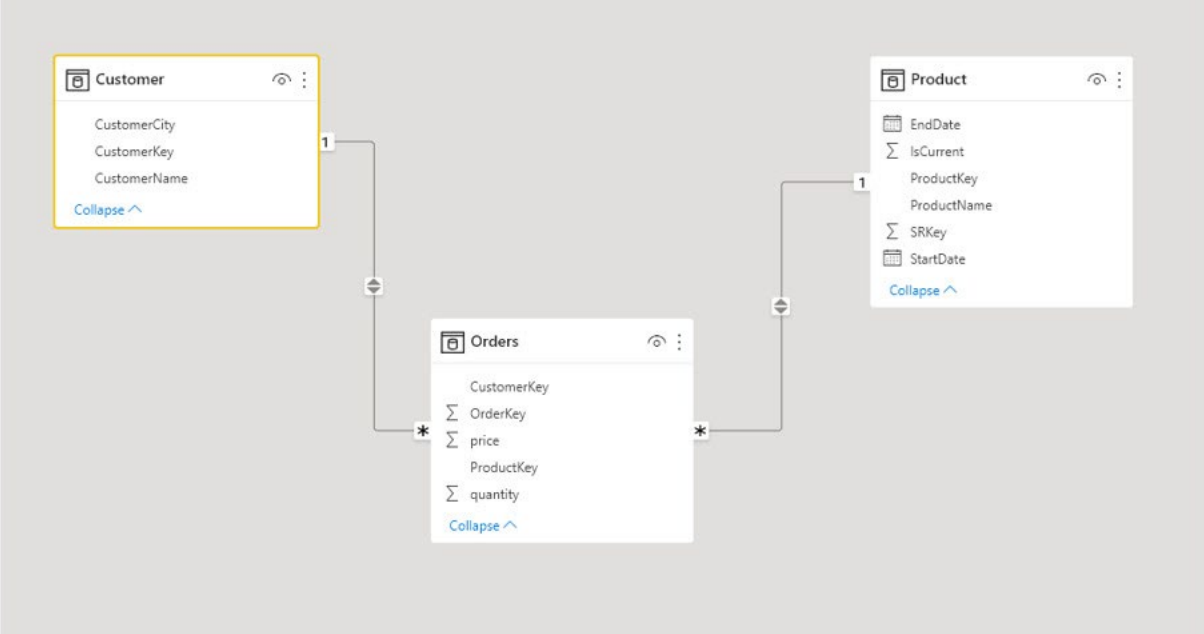
For more information on understanding the star schema , one can visit the following URL

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

Pytanie 27:

Pominięto

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

(Poprawne)

Wyjaśnienie

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

Pytanie 28:

Pominięto

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.

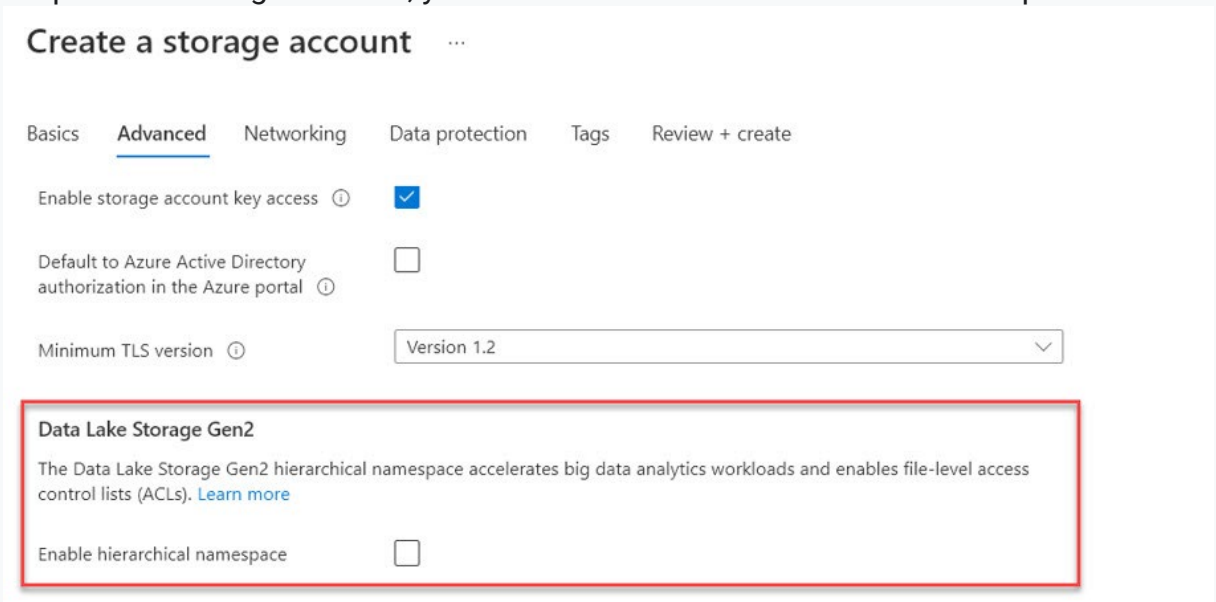
Which of the following do you need to enable when deploying an Azure General Purpose V2 Storage account to ensure that is behaves as a Data Lake Gen2 Storage account?

- **Enable storage account key access**
- **Enable hierarchical namespace**

  (Poprawne)
- **Access tier set to the Hot Access tier**
- **Enable large file shares**

Wyjaśnienie

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.



Pytanie 29:

Pominięto

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 would you choose as a redundancy option for the storage account?

- **Locally redundant storage**
- **Zone-redundant storage**
- **Geo-redundant storage**

  (Poprawne)
- **Read Access Geo-redundant storage**

Wyjaśnienie

When you set the data redundancy option to Geo-redundant storage, you can be ensured that the data in the storage account will become available in a secondary location if the primary location fails.

Using the redundancy option of Read Access Geo-redundant storage would increase the costs. And we need to minimize the costs wherever possible.

For more information on Azure Storage account redundancy , one can visit the following URL

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

Pytanie 30:

Pominięto

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 1?

- **save**
- **saveAsTable**
- **mode**
- **write**

  (Poprawne)
- **stream**

Wyjaśnienie

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

Pytanie 31:

Pominięto

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**

  (Poprawne)
- **write**
- **stream**

Wyjaśnienie

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

Pytanie 32:

Pominięto

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**

  (Poprawne)

- **saveAsTable**

- **mode**

- **write**

- **stream**

Wyjaśnienie

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

Pytanie 33:

Pominięto

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**
- **Azure Data Lake Gen2**

(Poprawne)

Wyjaśnienie

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

Pytanie 34:

Pominięto

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**

  (Poprawne)
- **Azure SQL Database**
- **Azure Data Lake Gen2**

Wyjaśnienie

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

Pytanie 35:

Pominięto

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**

  (Poprawne)

- **Azure Databricks**

- **Azure SQL Database**

- **Azure Data Lake Gen2**

Wyjaśnienie

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

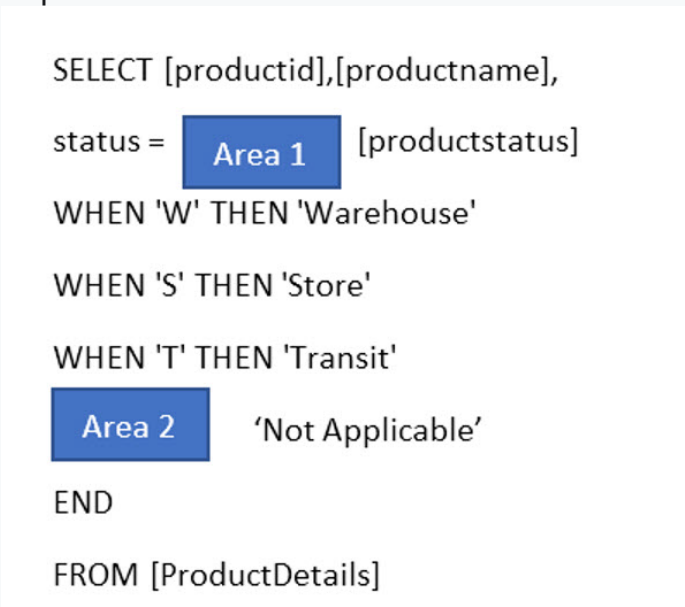
For more information on Azure Synapse , one can visit the following URL

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

Pytanie 36:

Pominięto

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**
- **CASE**

  (Poprawne)
- **ELSE**

Wyjaśnienie

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

Pytanie 37:

Pominięto

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

```
SELECT [productid],[productname],
status = Area 1 [productstatus]
WHEN 'W' THEN 'Warehouse'
WHEN 'S' THEN 'Store'
WHEN 'T' THEN 'Transit'
Area 2 'Not Applicable'
END
FROM [ProductDetails]
```

Which of the following would come in Area 2?

- **UPDATE**
- **SELECT**
- **CASE**
- **ELSE**

(Poprawne)

Wyjaśnienie

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

Pytanie 38:

Pominięto

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**
- **LogID**

(Poprawne)

Wyjaśnienie

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

Pytanie 39:

Pominięto

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 2?

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

(Poprawne)

Wyjaśnienie

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

Pytanie 40:

Pominięto

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**
- **Pin the cluster**

  (Poprawne)
- **Configure the cluster init scripts**

Wyjaśnienie

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

Pytanie 41:

Pominięto

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

  (Poprawne)
- **\DataSet\datafile_YYYY_MM_DD.csv**

Wyjaśnienie

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

Pytanie 42:

Pominięto

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**

(Poprawne)

Wyjaśnienie
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

Pytanie 43:

Pominięto

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**

  (Poprawne)
- **COLLATE**
- **DATEDIFF**

Wyjaśnienie
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

Pytanie 44:

Pominięto

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**

  (Poprawne)

- **LAST**

- **COLLATE**

- **DATEDIFF**

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

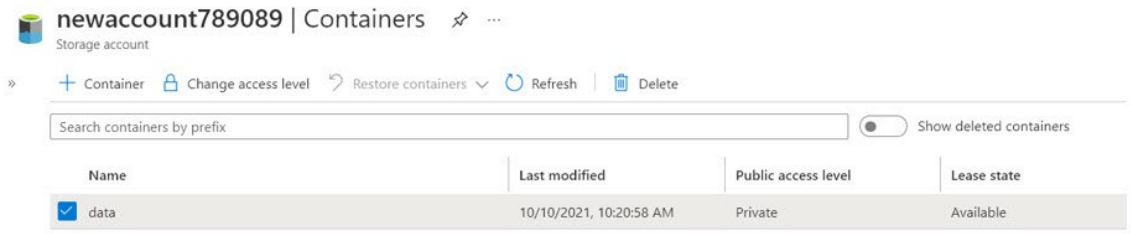
This question is based on the example in the below documentation link

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

Pytanie 45:

Pominięto

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**
- **A database scoped credential**

  (Poprawne)

Wyjaśnienie

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