Microsoft

# Exam Questions DP-203

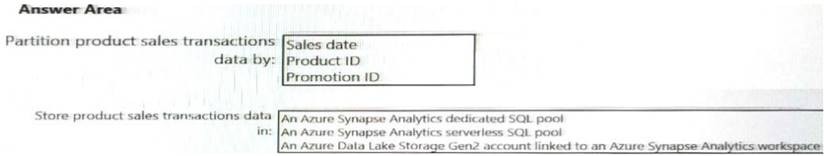
Data Engineering on Microsoft Azure

### NEW QUESTION 1

- (Exam Topic 1)

You need to design the partitions for the product sales transactions. The solution must mee the sales transaction dataset requirements.

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



1. Mastered
2. Not Mastered

### Answer: A

**Explanation:**

Box 1: Sales date

Scenario: Contoso requirements for data integration include:

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

Box 2: An Azure Synapse Analytics Dedicated SQL pool Scenario: Contoso requirements for data integration include:  Ensure that data storage costs and performance are predictable.

The size of a dedicated SQL pool (formerly SQL DW) is determined by Data Warehousing Units (DWU). Dedicated SQL pool (formerly SQL DW) stores data in relational tables with columnar storage. This format

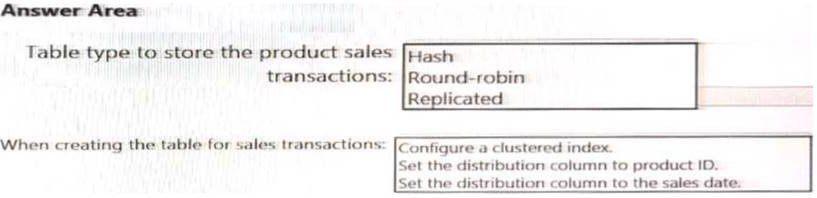
significantly reduces the data storage costs, and improves query performance. Synapse analytics dedicated sql pool Reference:

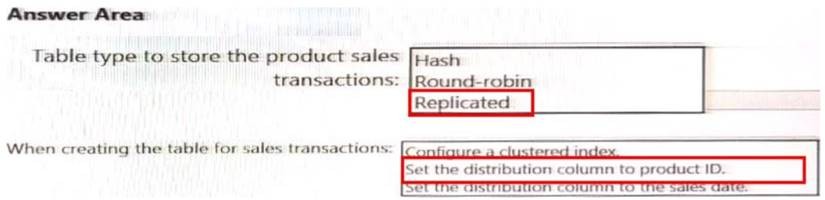
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-overview-wha

### NEW QUESTION 2

* (Exam Topic 1)

You need to design a data storage structure for the product sales transactions. The solution must meet the sales transaction dataset requirements. What should you include in the solution? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.



1. Mastered
2. Not Mastered **Answer:** A **Explanation:**

### NEW QUESTION 3

* (Exam Topic 1)

You need to integrate the on-premises data sources and Azure Synapse Analytics. The solution must meet the data integration requirements. Which type of integration runtime should you use?

1. Azure-SSIS integration runtime
2. self-hosted integration runtime
3. Azure integration runtime

### Answer: C

**NEW QUESTION 4**

* (Exam Topic 1)

You need to implement the surrogate key for the retail store table. The solution must meet the sales transaction dataset requirements. What should you create?

1. a table that has an IDENTITY property
2. a system-versioned temporal table
3. a user-defined SEQUENCE object
4. a table that has a FOREIGN KEY constraint

### Answer: A

**Explanation:**

Scenario: Implement a surrogate key to account for changes to the retail store addresses.

A surrogate key on a table is a column with a unique identifier for each row. The key is not generated from the table data. Data modelers like to create surrogate keys on their tables when they design data warehouse models. You can use the IDENTITY property to achieve this goal simply and effectively without affecting load performance.

Reference:

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

### NEW QUESTION 5

* (Exam Topic 3)

You have several Azure Data Factory pipelines that contain a mix of the following types of activities.

* Wrangling data flow
* Notebook
* Copy
* jar

Which two Azure services should you use to debug the activities? Each correct answer presents part of the solution NOTE: Each correct selection is worth one point.

1. Azure HDInsight
2. Azure Databricks
3. Azure Machine Learning
4. Azure Data Factory
5. Azure Synapse Analytics

### Answer: CE

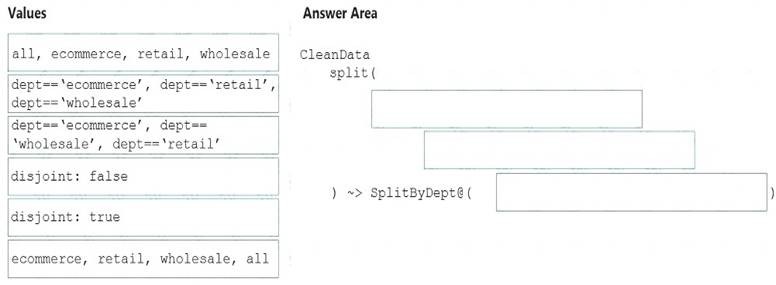
**NEW QUESTION 6**

* (Exam Topic 3)

You need to create an Azure Data Factory pipeline to process data for the following three departments at your company: Ecommerce, retail, and wholesale. The solution must ensure that data can also be processed for the entire company.

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

NOTE: Each correct selection is worth one point.



1. Mastered
2. Not Mastered

### Answer: A

**Explanation:**

The conditional split transformation routes data rows to different streams based on matching conditions. The conditional split transformation is similar to a CASE decision structure in a programming language. The transformation evaluates expressions, and based on the results, directs the data row to the specified stream. Box 1: dept=='ecommerce', dept=='retail', dept=='wholesale'

First we put the condition. The order must match the stream labeling we define in Box 3. Syntax:

<incomingStream> split(

<conditionalExpression1>

<conditionalExpression2> disjoint: {true | false}

) ~> <splitTx>@(stream1, stream2, ..., <defaultStream>)

Box 2: discount : false

disjoint is false because the data goes to the first matching condition. All remaining rows matching the third condition go to output stream all. Box 3: ecommerce, retail, wholesale, all Label the streams

Reference:

https://docs.microsoft.com/en-us/azure/data-factory/data-flow-conditional-split

### NEW QUESTION 7

* (Exam Topic 3)

You are monitoring an Azure Stream Analytics job.

The Backlogged Input Events count has been 20 for the last hour. You need to reduce the Backlogged Input Events count. What should you do?

1. Drop late arriving events from the job.
2. Add an Azure Storage account to the job.
3. Increase the streaming units for the job.
4. Stop the job.

### Answer: C

**Explanation:**

General symptoms of the job hitting system resource limits include:

 If the backlog event metric keeps increasing, it’s an indicator that the system resource is constrained (either because of output sink throttling, or high CPU). Note: Backlogged Input Events: Number of input events that are backlogged. A non-zero value for this metric implies that your job isn't able to keep up with the number of incoming events. If this value is slowly increasing or consistently non-zero, you should scale out your job: adjust Streaming Units.

Reference:

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

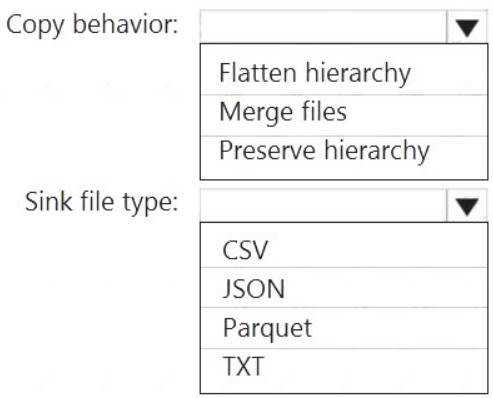
### NEW QUESTION 8

* (Exam Topic 3)

You use Azure Data Factory to prepare data to be queried by Azure Synapse Analytics serverless SQL pools. Files are initially ingested into an Azure Data Lake Storage Gen2 account as 10 small JSON files. Each file contains the same data attributes and data from a subsidiary of your company.

You need to move the files to a different folder and transform the data to meet the following requirements: Provide the fastest possible query times.  Automatically infer the schema from the underlying files.

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



1. Mastered
2. Not Mastered

### Answer: A

**Explanation:**

Box 1: Preserver herarchy

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

Box 2: Parquet

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

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

### NEW QUESTION 9

* (Exam Topic 3)

You are building an Azure Stream Analytics job to identify how much time a user spends interacting with a feature on a webpage.

The job receives events based on user actions on the webpage. Each row of data represents an event. Each event has a type of either 'start' or 'end'. You need to calculate the duration between start and end events.

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



1. Mastered
2. Not Mastered

### Answer: A

**Explanation:**

Box 1: DATEDIFF

DATEDIFF function returns the count (as a signed integer value) of the specified datepart boundaries crossed between the specified startdate and enddate. Syntax: DATEDIFF ( datepart , startdate, enddate ) Box 2: LAST

The LAST function can be used to retrieve the last event within a specific condition. In this example, the condition is an event of type Start, partitioning the search by PARTITION BY user and feature. This way, every user and feature is treated independently when searching for the Start event. LIMIT DURATION limits the search back in time to 1 hour between the End and Start events.

Example: SELECT

[user], feature, DATEDIFF( second,

LAST(Time) OVER (PARTITION BY [user], feature LIMIT DURATION(hour,

1) WHEN Event = 'start'), Time) as duration FROM input TIMESTAMP BY Time WHERE

Event = 'end' Reference:

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

### NEW QUESTION 10

* (Exam Topic 3)

You have an enterprise-wide Azure Data Lake Storage Gen2 account. The data lake is accessible only through an Azure virtual network named VNET1. You are building a SQL pool in Azure Synapse that will use data from the data lake.

Your company has a sales team. All the members of the sales team are in an Azure Active Directory group named Sales. POSIX controls are used to assign the Sales group access to the files in the data lake.

You plan to load data to the SQL pool every hour.

You need to ensure that the SQL pool can load the sales data from the data lake.

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

1. Add the managed identity to the Sales group.
2. Use the managed identity as the credentials for the data load process.
3. Create a shared access signature (SAS).
4. Add your Azure Active Directory (Azure AD) account to the Sales group.
5. Use the snared access signature (SAS) as the credentials for the data load process.
6. Create a managed identity.

**Answer:** ADF

### Explanation:

The managed identity grants permissions to the dedicated SQL pools in the workspace.

Note: Managed identity for Azure resources is a feature of Azure Active Directory. The feature provides Azure services with an automatically managed identity in Azure AD Reference:

https://docs.microsoft.com/en-us/azure/synapse-analytics/security/synapse-workspace-managed-identity

### NEW QUESTION 10

* (Exam Topic 3)

What should you recommend using to secure sensitive customer contact information?

1. data labels
2. column-level security
3. row-level security
4. Transparent Data Encryption (TDE)

### Answer: B

**Explanation:**

Scenario: All cloud data must be encrypted at rest and in transit.

Always Encrypted is a feature designed to protect sensitive data stored in specific database columns from access (for example, credit card numbers, national identification numbers, or data on a need to know basis). This includes database administrators or other privileged users who are authorized to access the database to perform management tasks, but have no business need to access the particular data in the encrypted columns. The data is always encrypted, which means the encrypted data is decrypted only for processing by client applications with access to the encryption key.

References:

https://docs.microsoft.com/en-us/azure/sql-database/sql-database-security-overview

### NEW QUESTION 15

* (Exam Topic 3)

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

You have files that are ingested and loaded into an Azure Data Lake Storage Gen2 container named container1. You plan to insert data from the files into Table1 and azure Data Lake Storage Gen2 container named container1.

You plan to insert data from the files into Table1 and transform the data. Each row of data in the files will produce one row in the serving layer of Table1. You need to ensure that when the source data files are loaded to container1, the DateTime is stored as an additional column in Table1.

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

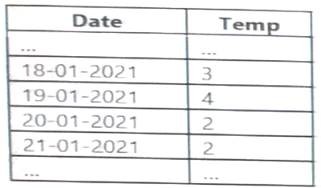
1. Yes
2. No

### Answer: A

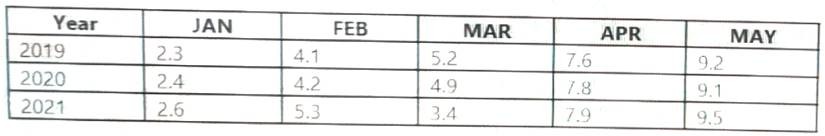
**NEW QUESTION 18**

* (Exam Topic 3)

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

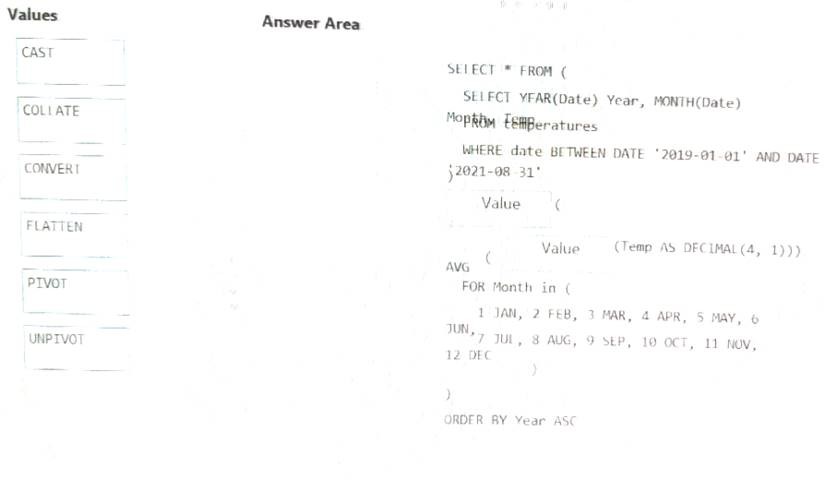


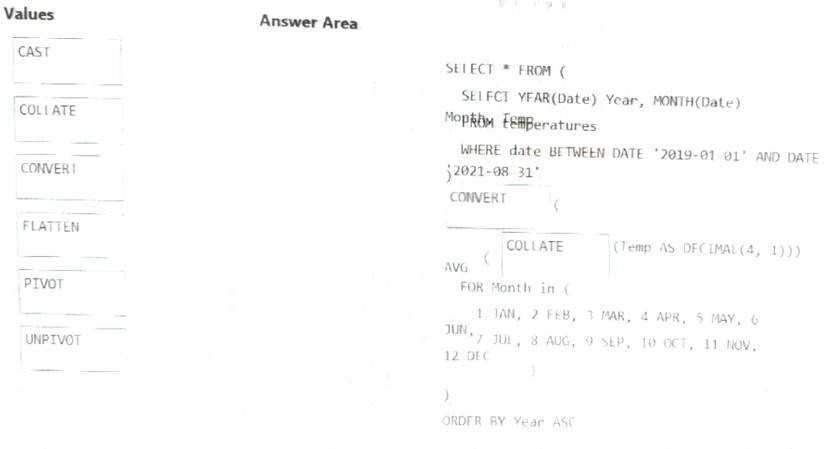
You need to produce the following table by using a Spark SQL query.



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

NOTE: Each correct selection is worth one point.



1. Mastered
2. Not Mastered **Answer:** A **Explanation:**

### NEW QUESTION 19

* (Exam Topic 3)

You have a C# application that process data from an Azure IoT hub and performs complex transformations. You need to replace the application with a real-time solution. The solution must reuse as much code as possible from the existing application.

1. Azure Databricks
2. Azure Event Grid
3. Azure Stream Analytics
4. Azure Data Factory

### Answer: C

**Explanation:**

Azure Stream Analytics on IoT Edge empowers developers to deploy near-real-time analytical intelligence closer to IoT devices so that they can unlock the full value of device-generated data. UDF are available in C# for IoT Edge jobs

Azure Stream Analytics on IoT Edge runs within the Azure IoT Edge framework. Once the job is created in Stream Analytics, you can deploy and manage it using IoT Hub.

References:

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

### NEW QUESTION 21

* (Exam Topic 3)

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

•Serve as a repository (or high volumes of large files in various formats.

•Implement optimized storage for big data analytics workloads.

•Ensure that data can be organized using a hierarchical structure. Batch processing:

•Use a managed solution for in-memory computation processing.

•Natively support Scala, Python, and R programming languages.

•Provide the ability to resize and terminate the cluster automatically. Analytical data store:

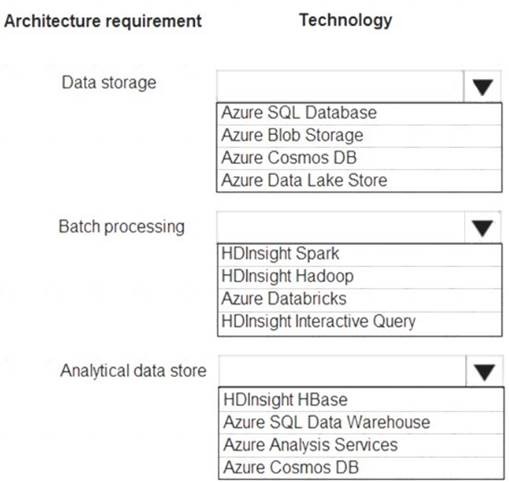
•Support parallel processing.

•Use columnar storage.

•Support SQL-based languages.

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

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



1. Mastered
2. Not Mastered

### Answer: A

**Explanation:**

Data storage: Azure Data Lake Store

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

Batch processing: HD Insight Spark

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

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

SQL Data Warehouse is a cloud-based Enterprise Data Warehouse (EDW) that uses Massively Parallel Processing (MPP). SQL Data Warehouse stores data into relational tables with columnar storage. References:

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

### NEW QUESTION 25

* (Exam Topic 3)

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

* Contain sales data for 20,000 products.
* Use hash distribution on a column named ProduclID,
* Contain 2.4 billion records for the years 20l9 and 2020.

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

A. 40

B. 240

C. 400

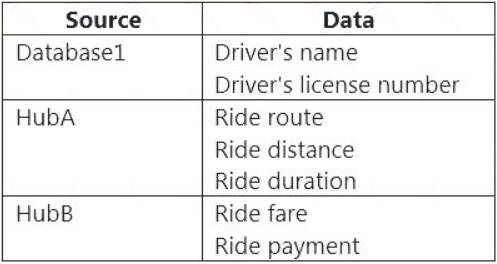
D. 2,400

### Answer: B

**NEW QUESTION 29**

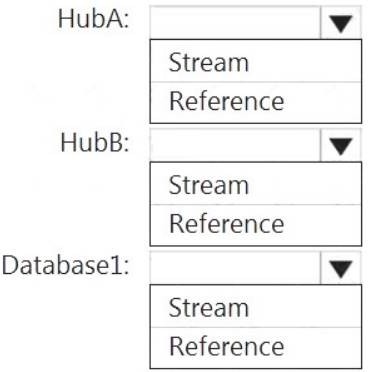
* (Exam Topic 3)

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



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

How should you configure the Stream Analytics input for each source? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.



1. Mastered
2. Not Mastered

### Answer: A

**Explanation:**

HubA: Stream HubB: Stream Database1: Reference

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

### NEW QUESTION 32

* (Exam Topic 3)

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

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

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

You plan to copy the data from the storage account to an Azure SQL data warehouse. You need to prepare the files to ensure that the data copies quickly. Solution: You modify the files to ensure that each row is more than 1 MB. Does this meet the goal?

1. Yes
2. No

### Answer: B

**Explanation:**

Instead modify the files to ensure that each row is less than 1 MB. References:

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

### NEW QUESTION 35

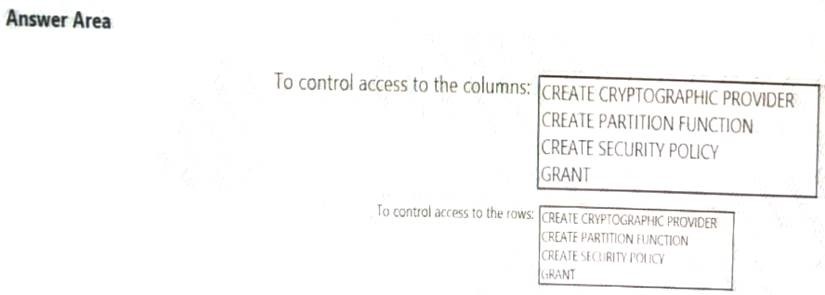
* (Exam Topic 3)

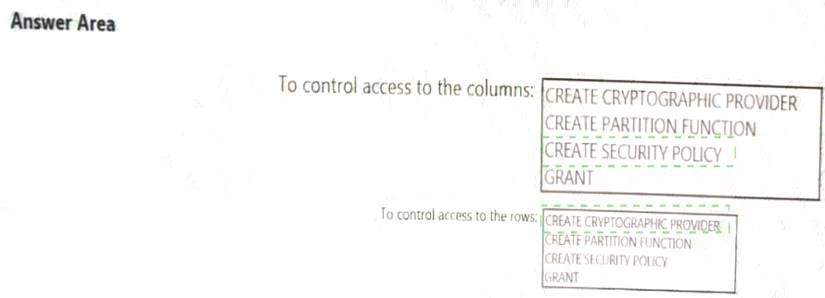
You have an Azure subscription that contains the following resources:

* An Azure Active Directory (Azure AD) tenant that contains a security group named Group1.
* An Azure Synapse Analytics SQL pool named Pool1.

You need to control the access of Group1 to specific columns and rows in a table in Pool1

Which Transact-SQL commands should you use? To answer, select the appropriate options in the answer area. NOTE: Each appropriate options in the answer area.



1. Mastered
2. Not Mastered **Answer:** A **Explanation:**

### NEW QUESTION 38

* (Exam Topic 3)

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

You need to ensure that the data lake will remain available if a data center fails in the primary Azure region. The solution must minimize costs. Which type of replication should you use for the storage account?

1. geo-redundant storage (GRS)
2. zone-redundant storage (ZRS)
3. locally-redundant storage (LRS)
4. geo-zone-redundant storage (GZRS)

### Answer: A

**Explanation:**

Geo-redundant storage (GRS) copies your data synchronously three times within a single physical location in the primary region using LRS. It then copies your data asynchronously to a single physical location in the secondary region.

Reference:

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

### NEW QUESTION 40

* (Exam Topic 3)

You are designing an Azure Synapse Analytics dedicated SQL pool.

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

1. dynamic data masking
2. row-level security (RLS)
3. sensitivity classifications
4. column-level security

### Answer: D

**NEW QUESTION 41**

* (Exam Topic 3)

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

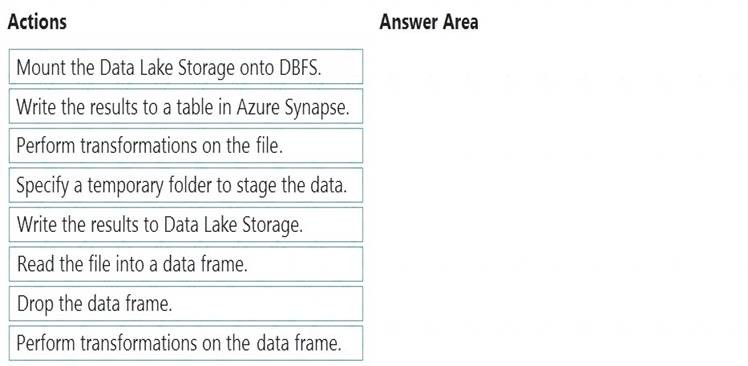
You create the following components:



A destination table in Azure Synapse An Azure Blob storage container

A service principal

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



1. Mastered
2. Not Mastered

### Answer: A

**Explanation:**

Step 1: Read the file into a data frame.

You can load the json files as a data frame in Azure Databricks. Step 2: Perform transformations on the data frame. Step 3:Specify a temporary folder to stage the data

Specify a temporary folder to use while moving data between Azure Databricks and Azure Synapse. Step 4: Write the results to a table in Azure Synapse.

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

Step 5: Drop the data frame

Clean up resources. You can terminate the cluster. From the Azure Databricks workspace, select Clusters on the left. For the cluster to terminate, under Actions, point to the ellipsis (...) and select the Terminate icon.

Reference:

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

### NEW QUESTION 44

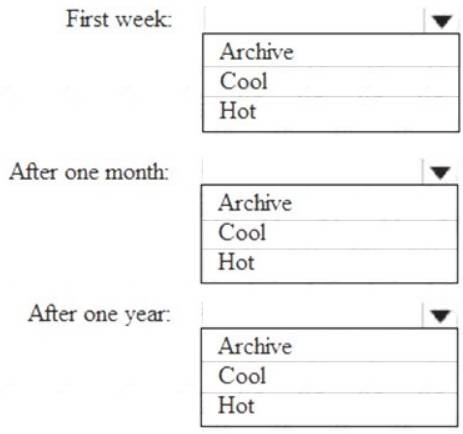
* (Exam Topic 3)

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

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

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

Which storage tier should you use for each time frame? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.



1. Mastered
2. Not Mastered

### Answer: A

**Explanation:**

First week: Hot

Hot - Optimized for storing data that is accessed frequently. After one month: Cool

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

### NEW QUESTION 46

* (Exam Topic 3)

You plan to create an Azure Synapse Analytics dedicated SQL pool.

You need to minimize the time it takes to identify queries that return confidential information as defined by the company's data privacy regulations and the users who executed the queues.

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

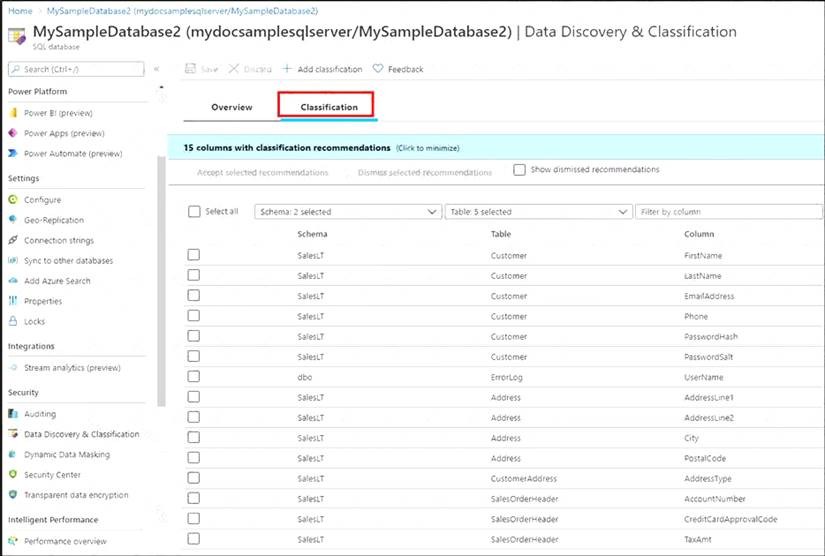
1. sensitivity-classification labels applied to columns that contain confidential information
2. resource tags for databases that contain confidential information
3. audit logs sent to a Log Analytics workspace
4. dynamic data masking for columns that contain confidential information

### Answer: AC

**Explanation:**

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

Select Add classification in the top menu of the pane.



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

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



Reference:

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

### NEW QUESTION 49

* (Exam Topic 3)

You are designing an Azure Databricks table. The table will ingest an average of 20 million streaming events per day.

You need to persist the events in the table for use in incremental load pipeline jobs in Azure Databricks. The solution must minimize storage costs and incremental load times.

What should you include in the solution?

1. Partition by DateTime fields.
2. Sink to Azure Queue storage.
3. Include a watermark column.
4. Use a JSON format for physical data storage.

### Answer: B

**Explanation:**

The Databricks ABS-AQS connector uses Azure Queue Storage (AQS) to provide an optimized file source that lets you find new files written to an Azure Blob storage (ABS) container without repeatedly listing all of the files.

This provides two major advantages:

Lower costs: no more costly LIST API requests made to ABS. Reference:

https://docs.microsoft.com/en-us/azure/databricks/spark/latest/structured-streaming/aqs

### NEW QUESTION 52

* (Exam Topic 3)

You have an Azure Synapse Analytics dedicated SQL pool that contains a large fact table. The table contains 50 columns and 5 billion rows and is a heap. Most queries against the table aggregate values from approximately 100 million rows and return only two columns.

You discover that the queries against the fact table are very slow. Which type of index should you add to provide the fastest query times?

1. nonclustered columnstore
2. clustered columnstore
3. nonclustered
4. clustered

### Answer: B

**Explanation:**

Clustered columnstore indexes are one of the most efficient ways you can store your data in dedicated SQL pool. Columnstore tables won't benefit a query unless the table has more than 60 million rows. Reference:

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

### NEW QUESTION 55

* (Exam Topic 3)

You are designing an Azure Stream Analytics job to process incoming events from sensors in retail environments.

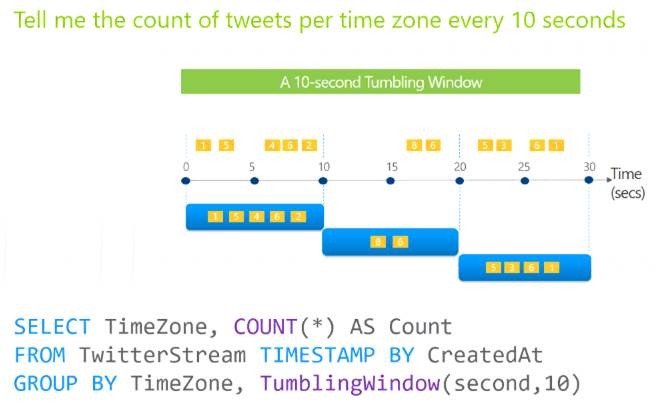
You need to process the events to produce a running average of shopper counts during the previous 15 minutes, calculated at five-minute intervals. Which type of window should you use?

1. snapshot
2. tumbling
3. hopping
4. sliding

### Answer: B

**Explanation:**

Tumbling windows are a series of fixed-sized, non-overlapping and contiguous time intervals. The following diagram illustrates a stream with a series of events and how they are mapped into 10-second tumbling windows.



Reference:

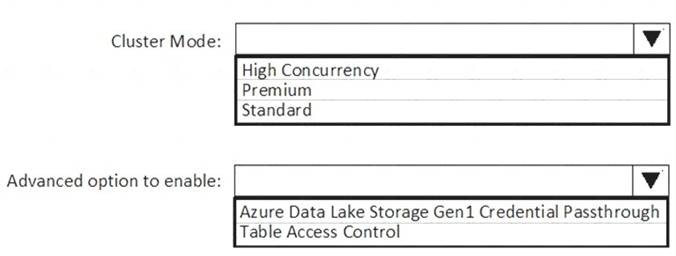
https://docs.microsoft.com/en-us/stream-analytics-query/tumbling-window-azure-stream-analytics

### NEW QUESTION 60

* (Exam Topic 3)

You need to implement an Azure Databricks cluster that automatically connects to Azure Data Lake Storage Gen2 by using Azure Active Directory (Azure AD) integration.

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



1. Mastered
2. Not Mastered

### Answer: A

**Explanation:**

Box 1: High Concurrency

Enable Azure Data Lake Storage credential passthrough for a high-concurrency cluster. Incorrect:

Support for Azure Data Lake Storage credential passthrough on standard clusters is in Public Preview.

Standard clusters with credential passthrough are supported on Databricks Runtime 5.5 and above and are limited to a single user. Box 2: Azure Data Lake Storage Gen1 Credential Passthrough

You can authenticate automatically to Azure Data Lake Storage Gen1 and Azure Data Lake Storage Gen2 from Azure Databricks clusters using the same Azure Active Directory (Azure AD) identity that you use to log into Azure Databricks. When you enable your cluster for Azure Data Lake Storage credential passthrough, commands that you run on that cluster can read and write data in Azure Data Lake Storage without requiring you to configure service principal credentials for access to storage.

References:

https://docs.azuredatabricks.net/spark/latest/data-sources/azure/adls-passthrough.html

### NEW QUESTION 63

* (Exam Topic 3)

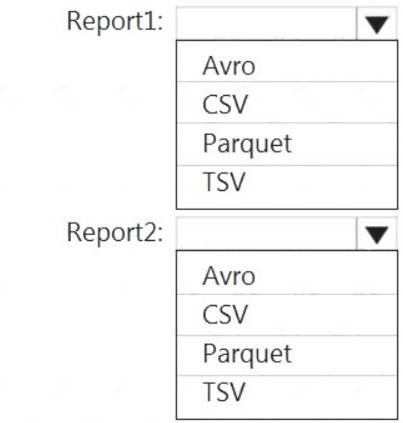
You are planning the deployment of Azure Data Lake Storage Gen2. You have the following two reports that will access the data lake: Report1: Reads three columns from a file that contains 50 columns.



Report2: Queries a single record based on a timestamp.

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

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



1. Mastered
2. Not Mastered

### Answer: A

**Explanation:**

Report1: CSV

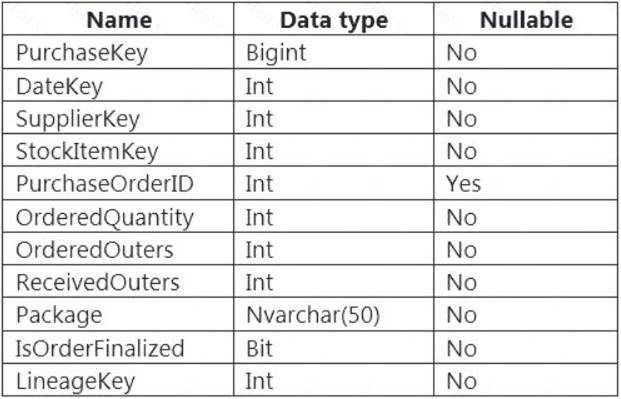
CSV: The destination writes records as delimited data. Report2: AVRO AVRO supports timestamps.

Not Parquet, TSV: Not options for Azure Data Lake Storage Gen2. Reference: https://streamsets.com/documentation/datacollector/latest/help/datacollector/UserGuide/Destinations/ADLS-G2

### NEW QUESTION 67

* (Exam Topic 3)

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



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

SELECT

SupplierKey, StockItemKey, COUNT(\*) FROM FactPurchase

WHERE DateKey >= 20210101 AND DateKey <= 20210131

GROUP By SupplierKey, StockItemKey

Which table distribution will minimize query times?

1. round-robin
2. replicated
3. hash-distributed on DateKey
4. hash-distributed on PurchaseKey

### Answer: D

**Explanation:**

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

Reference:

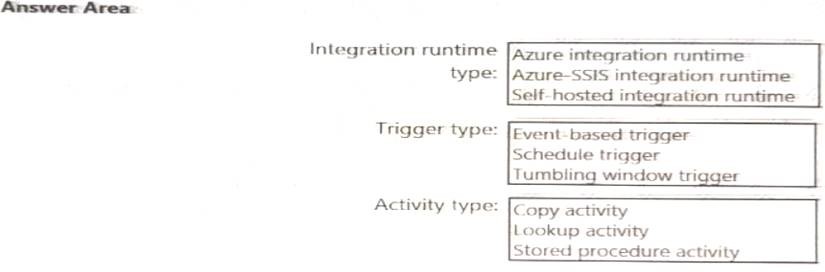
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribu

### NEW QUESTION 71

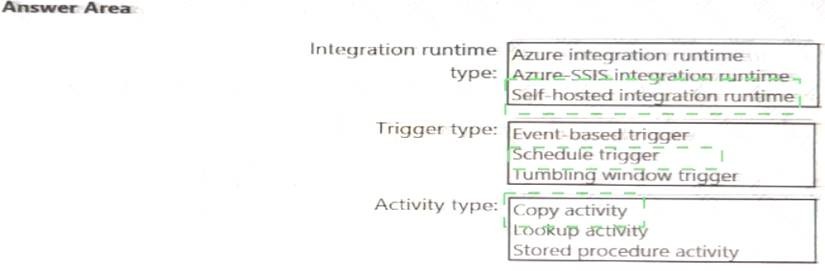
* (Exam Topic 3)

Which Azure Data Factory components should you recommend using together to import the daily inventory data from the SQL server to Azure Data Lake Storage? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.



1. Mastered
2. Not Mastered **Answer:** A **Explanation:**



### NEW QUESTION 76

* (Exam Topic 3)

You use Azure Data Lake Storage Gen2.

You need to ensure that workloads can use filter predicates and column projections to filter data at the time the data is read from disk. Which two actions should you perform? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

1. Reregister the Microsoft Data Lake Store resource provider.
2. Reregister the Azure Storage resource provider.
3. Create a storage policy that is scoped to a container.
4. Register the query acceleration feature.
5. Create a storage policy that is scoped to a container prefix filter.

### Answer: BD

**NEW QUESTION 77**

* (Exam Topic 3)

You are designing a sales transactions table in an Azure Synapse Analytics dedicated SQL pool. The table will contains approximately 60 million rows per month and will be partitioned by month. The table will use a clustered column store index and round-robin distribution.

Approximately how many rows will there be for each combination of distribution and partition?

1. 1 million
2. 5 million
3. 20 million
4. 60 million

### Answer: D

**Explanation:**

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

### NEW QUESTION 80

* (Exam Topic 3)

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

1. notebook logs
2. cluster event logs
3. global init scripts logs
4. workspace logs

### Answer: C

**Explanation:**

Cluster-scoped Init Scripts: Init scripts are shell scripts that run during the startup of each cluster node before the Spark driver or worker JVM starts. Databricks customers use init scripts for various purposes such as installing custom libraries, launching background processes, or applying enterprise security policies.

Logs for Cluster-scoped init scripts are now more consistent with Cluster Log Delivery and can be found in the same root folder as driver and executor logs for the cluster.

Reference:

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

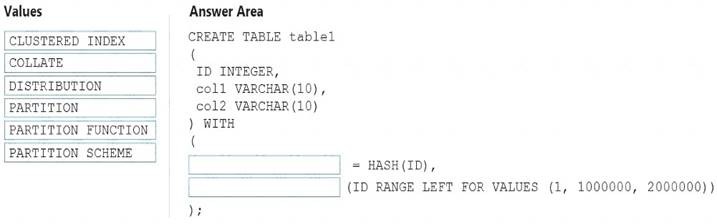
### NEW QUESTION 81

* (Exam Topic 3)

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

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

NOTE: Each correct selection is worth one point.



1. Mastered
2. Not Mastered

### Answer: A

**Explanation:**

Box 1: DISTRIBUTION

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

Table partition options. Syntax:

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

))

Reference:

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

### NEW QUESTION 82

* (Exam Topic 3)

You have an Azure Storage account and a data warehouse in Azure Synapse Analytics in the UK South region. You need to copy blob data from the storage account to the data warehouse by using Azure Data Factory. The solution must meet the following requirements:



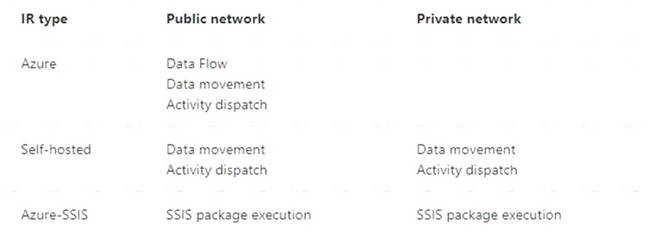
Ensure that the data remains in the UK South region at all times. Minimize administrative effort.

Which type of integration runtime should you use?

1. Azure integration runtime
2. Azure-SSIS integration runtime
3. Self-hosted integration runtime

### Answer: A

**Explanation:**



Reference:

https://docs.microsoft.com/en-us/azure/data-factory/concepts-integration-runtime

### NEW QUESTION 85

* (Exam Topic 3)

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

container1.

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

You plan to insert data from the files into Table1 and transform the data. Each row of data in the files will produce one row in the serving layer of Table1. You need to ensure that when the source data files are loaded to container1, the DateTime is stored as an additional column in Table1.

Solution: In an Azure Synapse Analytics pipeline, you use a data flow that contains a Derived Column transformation.

1. Yes
2. No

### Answer: B

**NEW QUESTION 86**

* (Exam Topic 3)

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

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

Which type of data redundancy should you use?

1. zone-redundant storage (ZRS)
2. read-access geo-redundant storage (RA-GRS)
3. locally-redundant storage (LRS)
4. geo-redundant storage (GRS)

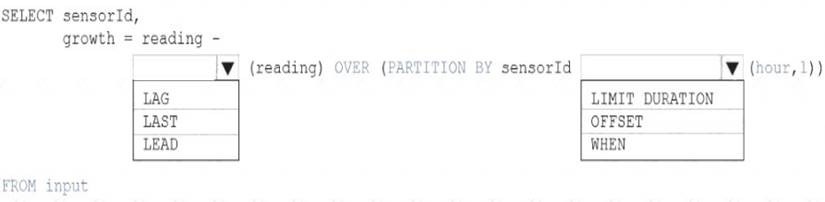
### Answer: C

**NEW QUESTION 90**

* (Exam Topic 3)

You are building an Azure Analytics query that will receive input data from Azure IoT Hub and write the results to Azure Blob storage. You need to calculate the difference in readings per sensor per hour.

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



1. Mastered
2. Not Mastered

### Answer: A

**Explanation:**

Box 1: LAG

The LAG analytic operator allows one to look up a “previous” event in an event stream, within certain constraints. It is very useful for computing the rate of growth of a variable, detecting when a variable crosses a threshold, or when a condition starts or stops being true.

Box 2: LIMIT DURATION

Example: Compute the rate of growth, per sensor: SELECT sensorId, growth = reading

LAG(reading) OVER (PARTITION BY sensorId LIMIT DURATION(hour, 1)) FROM input Reference:

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

### NEW QUESTION 94

* (Exam Topic 3)

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



Maintains the latest employee information.

Minimizes query complexity.

How should you model the employee data?

1. as a temporal table
2. as a SQL graph table
3. as a degenerate dimension table
4. as a Type 2 slowly changing dimension (SCD) table

### Answer: D

**Explanation:**

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

Reference:

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

### NEW QUESTION 98

* (Exam Topic 3)

You are designing the folder structure for an Azure Data Lake Storage Gen2 container.

Users will query data by using a variety of services including Azure Databricks and Azure Synapse Analytics serverless SQL pools. The data will be secured by subject area. Most queries will include data from the current year or current month.

Which folder structure should you recommend to support fast queries and simplified folder security?

1. /{SubjectArea}/{DataSource}/{DD}/{MM}/{YYYY}/{FileData}\_{YYYY}\_{MM}\_{DD}.csv
2. /{DD}/{MM}/{YYYY}/{SubjectArea}/{DataSource}/{FileData}\_{YYYY}\_{MM}\_{DD}.csv
3. /{YYYY}/{MM}/{DD}/{SubjectArea}/{DataSource}/{FileData}\_{YYYY}\_{MM}\_{DD}.csv
4. /{SubjectArea}/{DataSource}/{YYYY}/{MM}/{DD}/{FileData}\_{YYYY}\_{MM}\_{DD}.csv

### Answer: D

**Explanation:**

There's an important reason to put the date at the end of the directory structure. If you want to lock down certain regions or subject matters to users/groups, then you can easily do so with the POSIX permissions. Otherwise, if there was a need to restrict a certain security group to viewing just the UK data or certain planes, with the date structure in front a separate permission would be required for numerous directories under every hour directory. Additionally, having the date structure in front would exponentially increase the number of directories as time went on.

Note: In IoT workloads, there can be a great deal of data being landed in the data store that spans across numerous products, devices, organizations, and customers. It’s important to pre-plan the directory layout for organization, security, and efficient processing of the data for down-stream consumers. A general template to consider might be the following layout:

{Region}/{SubjectMatter(s)}/{yyyy}/{mm}/{dd}/{hh}/

### NEW QUESTION 102

* (Exam Topic 3)

You configure monitoring for a Microsoft Azure SQL Data Warehouse implementation. The implementation uses PolyBase to load data from comma-separated value (CSV) files stored in Azure Data Lake Gen 2 using an external table.

Files with an invalid schema cause errors to occur. You need to monitor for an invalid schema error. For which error should you monitor?

1. EXTERNAL TABLE access failed due to internal error: 'Java exception raised on call to HdfsBridge\_Connect: Error[com.microsoft.polybase.client.KerberosSecureLogin] occurred while accessing external files.'
2. EXTERNAL TABLE access failed due to internal error: 'Java exception raised on call to HdfsBridge\_Connect: Error [No FileSystem for scheme: wasbs] occurred while accessing external file.'
3. Cannot execute the query "Remote Query" against OLE DB provider "SQLNCLI11": for linked server "(null)", Query aborted- the maximum reject threshold (orows) was reached while regarding from an external source: 1 rows rejected out of total 1 rows processed.
4. EXTERNAL TABLE access failed due to internal error: 'Java exception raised on call to HdfsBridge\_Connect: Error [Unable to instantiate LoginClass] occurredwhile accessing external files.'

### Answer: C

**Explanation:**

Customer Scenario:

SQL Server 2016 or SQL DW connected to Azure blob storage. The CREATE EXTERNAL TABLE DDL points to a directory (and not a specific file) and the directory contains files with different schemas.

SSMS Error:

Select query on the external table gives the following error: Msg 7320, Level 16, State 110, Line 14

Cannot execute the query "Remote Query" against OLE DB provider "SQLNCLI11" for linked server "(null)". Query aborted-- the maximum reject threshold (0 rows) was reached while reading from an external source: 1 rows rejected out of total 1 rows processed.

Possible Reason:

The reason this error happens is because each file has different schema. The PolyBase external table DDL when pointed to a directory recursively reads all the files in that directory. When a column or data type mismatch happens, this error could be seen in SSMS.

Possible Solution:

If the data for each table consists of one file, then use the filename in the LOCATION section prepended by the directory of the external files. If there are multiple files per table, put each set of files into different directories in Azure Blob Storage and then you can point LOCATION to the directory instead of a particular

file. The latter suggestion is the best practices recommended by SQLCAT even if you have one file per table.

### NEW QUESTION 103

* (Exam Topic 3)

You are creating an Azure Data Factory data flow that will ingest data from a CSV file, cast columns to specified types of data, and insert the data into a table in an Azure Synapse Analytic dedicated SQL pool. The CSV file contains three columns named username, comment, and date.

The data flow already contains the following: A source transformation.



A Derived Column transformation to set the appropriate types of data.

A sink transformation to land the data in the pool.

You need to ensure that the data flow meets the following requirements: All valid rows must be written to the destination table.



Truncation errors in the comment column must be avoided proactively.

Any rows containing comment values that will cause truncation errors upon insert must be written to a file in blob storage.

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

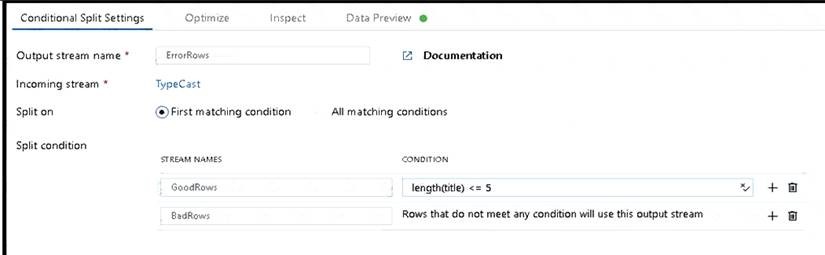
1. To the data flow, add a sink transformation to write the rows to a file in blob storage.
2. To the data flow, add a Conditional Split transformation to separate the rows that will cause truncation errors.
3. To the data flow, add a filter transformation to filter out rows that will cause truncation errors.
4. Add a select transformation to select only the rows that will cause truncation errors.

### Answer: AB

**Explanation:**

B: Example:

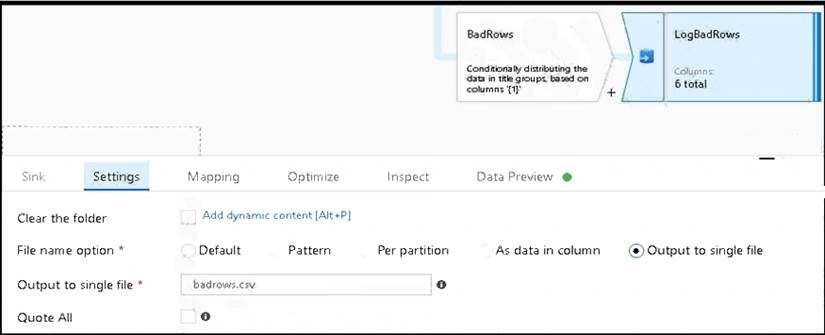
* 1. This conditional split transformation defines the maximum length of "title" to be five. Any row that is less than or equal to five will go into the GoodRows stream. Any row that is larger than five will go into the BadRows stream.



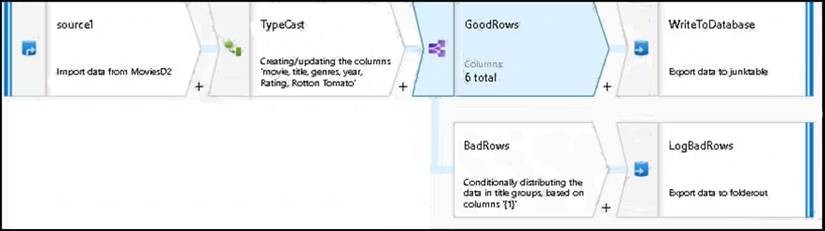
* 2. This conditional split transformation defines the maximum length of "title" to be five. Any row that is less than or equal to five will go into the GoodRows stream. Any row that is larger than five will go into the BadRows stream.

A:

* 3. Now we need to log the rows that failed. Add a sink transformation to the BadRows stream for logging. Here, we'll "auto-map" all of the fields so that we have logging of the complete transaction record. This is a text-delimited CSV file output to a single file in Blob Storage. We'll call the log file "badrows.csv".



* 4. The completed data flow is shown below. We are now able to split off error rows to avoid the SQL truncation errors and put those entries into a log file. Meanwhile, successful rows can continue to write to our target database.



Reference:

https://docs.microsoft.com/en-us/azure/data-factory/how-to-data-flow-error-rows

### NEW QUESTION 108

* (Exam Topic 3)

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

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

What should you recommend?

1. Parquet
2. Avro
3. CSV
4. JSON

### Answer: B

**Explanation:**

The Avro format is great for data and message preservation.Avro schema with its support for evolution is essential for making the data robust for streaming architectures like Kafka, and with the metadata that schema provides, you can reason on the data. Having a schema provides robustness in providing meta-data [about the data stored in Avro records which are self- documenting the data.References: http://cloudurable.com/blog/avro/index.html](http://cloudurable.com/blog/avro/index.html)

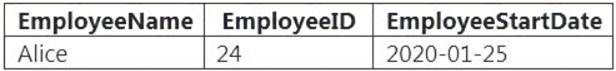
### NEW QUESTION 112

* (Exam Topic 3)

You have an Azure Synapse workspace named MyWorkspace that contains an Apache Spark database named mytestdb.

You run the following command in an Azure Synapse Analytics Spark pool in MyWorkspace. CREATE TABLE mytestdb.myParquetTable( EmployeeID int, EmployeeName string, EmployeeStartDate date) USING Parquet

You then use Spark to insert a row into mytestdb.myParquetTable. The row contains the following data.



One minute later, you execute the following query from a serverless SQL pool in MyWorkspace. SELECT EmployeeID FROM mytestdb.dbo.myParquetTable WHERE name = 'Alice';

What will be returned by the query?

1. 24
2. an error
3. a null value

### Answer: A

**Explanation:**

Once a database has been created by a Spark job, you can create tables in it with Spark that use Parquet as the storage format. Table names will be converted to lower case and need to be queried using the lower case name. These tables will immediately become available for querying by any of the Azure Synapse workspace Spark pools. They can also be used from any of the Spark jobs subject to permissions.

Note: For external tables, since they are synchronized to serverless SQL pool asynchronously, there will be a delay until they appear. Reference:

https://docs.microsoft.com/en-us/azure/synapse-analytics/metadata/table

### NEW QUESTION 117

* (Exam Topic 3)

You have an Azure Synapse Analytics job that uses Scala. You need to view the status of the job. What should you do?

1. From Azure Monitor, run a Kusto query against the AzureDiagnostics table.
2. From Azure Monitor, run a Kusto query against the SparkLogying1 Event.CL table.
3. From Synapse Studio, select the workspac
4. From Monitor, select Apache Sparks applications.
5. From Synapse Studio, select the workspac
6. From Monitor, select SQL requests.

### Answer: C

**NEW QUESTION 122**

* (Exam Topic 3)

You are designing a dimension table for a data warehouse. The table will track the value of the dimension attributes over time and preserve the history of the data by adding new rows as the data changes.

Which type of slowly changing dimension (SCD) should use?

1. Type 0
2. Type 1
3. Type 2
4. Type 3

### Answer: C

**Explanation:**

Type 2 - Creating a new additional record. In this methodology all history of dimension changes is kept in the database. You capture attribute change by adding a new row with a new surrogate key to the dimension table. Both the prior and new rows contain as attributes the natural key(or other durable identifier). Also 'effective date' and 'current indicator' columns are used in this method. There could be only one record with current indicator set to 'Y'. For 'effective date' columns,

i.e. start\_date and end\_date, the end\_date for current record usually is set to value 9999-12-31. Introducing changes to the dimensional model in type 2 could be very expensive database operation so it is not recommended to use it in dimensions where a new attribute could be added in the future. [https://www.datawarehouse4u.info/SCD-Slowly-Changing-Dimensions.html](http://www.datawarehouse4u.info/SCD-Slowly-Changing-Dimensions.html)

### NEW QUESTION 124

* (Exam Topic 3)

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

container1.

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

You plan to insert data from the files into Table1 and transform the data. Each row of data in the files will produce one row in the serving layer of Table1. You need to ensure that when the source data files are loaded to container1, the DateTime is stored as an additional column in Table1.

Solution: In an Azure Synapse Analytics pipeline, you use a Get Metadata activity that retrieves the DateTime of the files. Does this meet the goal?

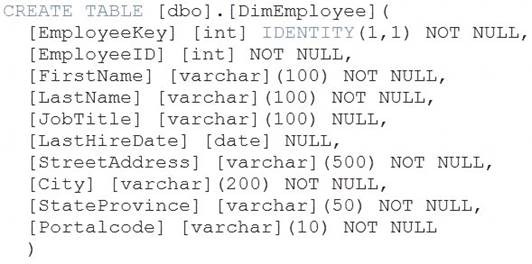
1. Yes
2. No

### Answer: B

**NEW QUESTION 128**

* (Exam Topic 3)

You have a table in an Azure Synapse Analytics dedicated SQL pool. The table was created by using the following Transact-SQL statement.



You need to alter the table to meet the following requirements: Ensure that users can identify the current manager of employees.



Support creating an employee reporting hierarchy for your entire company.

Provide fast lookup of the managers’ attributes such as name and job title.

Which column should you add to the table?

1. [ManagerEmployeeID] [int] NULL
2. [ManagerEmployeeID] [smallint] NULL
3. [ManagerEmployeeKey] [int] NULL
4. [ManagerName] [varchar](200) NULL

### Answer: A

**Explanation:**

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

### NEW QUESTION 132

......