

Microsof.DP-203.vJul-2024.by.Aien.166q

Number: DP-203  
Passing Score: 800  
Time Limit: 120  
File Version: 31.0

Exam Code: DP-203  
Exam Name: Data Engineering on Microsoft Azure



## Case 01-Design and implement data storage

### Case study

This is a case study. Case studies are not timed separately. You can use as much exam time as you would like to complete each case. However, there may be additional case studies and sections on this exam. You must manage your time to ensure that you are able to complete all questions included on this exam in the time provided.

To answer the questions included in a case study, you will need to reference information that is provided in the case study. Case studies might contain exhibits and other resources that provide more information about the scenario that is described in the case study. Each question is independent of the other questions in this case study.

At the end of this case study, a review screen will appear. This screen allows you to review your answers and to make changes before you move to the next section of the exam. After you begin a new section, you cannot return to this section.

To start the case study

To display the first question in this case study, click the Next button. Use the buttons in the left pane to explore the content of the case study before you answer the questions. Clicking these buttons displays information such as business requirements, existing environment, and problem statements. If the case study has an All Information tab, note that the information displayed is identical to the information displayed on the subsequent tabs.

When you are ready to answer a question, click the Question button to return to the question.

### Overview

Contoso, Ltd. is a clothing retailer based in Seattle. The company has 2,000 retail stores across the United States and an emerging online presence.

The network contains an Active Directory forest named contoso.com. The forest is integrated with an Azure Active Directory (Azure AD) tenant named contoso.com. Contoso has an Azure subscription associated to the contoso.com Azure AD tenant.

### Existing Environment

#### Transactional Data

Contoso has three years of customer, transactional, operational, sourcing, and supplier data comprised of 10 billion records stored across multiple on-premises Microsoft SQL Server servers. The SQL Server instances contain data from various operational systems. The data is loaded into the instances by using SQL Server Integration Services (SSIS) packages.

You estimate that combining all product sales transactions into a company-wide sales transactions dataset will result in a single table that contains 5 billion rows, with one row per transaction.

Most queries targeting the sales transactions data will be used to identify which products were sold in retail stores and which products were sold online during different time periods. Sales transaction data that is older than three years will be removed monthly.

You plan to create a retail store table that will contain the address of each retail store. The table will be approximately 2 MB. Queries for retail store sales will include the retail store addresses.

You plan to create a promotional table that will contain a promotion ID. The promotion ID will be associated to a specific product. The product will be identified by a product ID. The table will be approximately 5 GB.

#### Streaming Twitter Data

The ecommerce department at Contoso develops an Azure logic app that captures trending Twitter feeds referencing the company's products and pushes the products to Azure Event Hubs.

#### Planned Changes and Requirements

##### Planned Changes

Contoso plans to implement the following changes:

- Load the sales transaction dataset to Azure Synapse Analytics.

- Integrate on-premises data stores with Azure Synapse Analytics by using SSIS packages.

- Use Azure Synapse Analytics to analyze Twitter feeds to assess customer sentiments about products.

##### Sales Transaction Dataset Requirements

Contoso identifies the following requirements for the sales transaction dataset:

- 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.

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

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

- Ensure that data storage costs and performance are predictable.

- Minimize how long it takes to remove old records.

##### Customer Sentiment Analytics Requirements

Contoso identifies the following requirements for customer sentiment analytics:

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

- Maximize the throughput of ingesting Twitter feeds from Event Hubs to Azure Storage without purchasing additional throughput or capacity units.

- Store Twitter feeds in Azure Storage by using Event Hubs Capture. The feeds will be converted into Parquet files.

- Ensure that the data store supports Azure AD-based access control down to the object level.

Minimize administrative effort to maintain the Twitter feed data records.

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

Data Integration Requirements

Contoso identifies the following requirements for data integration:

Use an Azure service that leverages the existing SSIS packages to ingest on-premises data into datasets stored in a dedicated SQL pool of Azure Synapse Analytics and transform the data.

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

### QUESTION 1

HOTSPOT

You need to design an analytical storage solution for the transactional data. The solution must meet the sales transaction dataset requirements.

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

NOTE: Each correct selection is worth one point.

Hot Area:

**Answer Area**

Table type to store retail store data: 

▼
Hash
Replicated
Round-robin

Table type to store promotional data: 

▼
Hash
Replicated
Round-robin

Vdumps

Answer Area:

**Answer Area**

Table type to store retail store data: 

▼
Hash
Replicated
Round-robin

Table type to store promotional data: 

▼
Hash
Replicated
Round-robin

Section:

**Explanation:**

<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/design-guidance-for-replicated-tables#what-is-a-replicated-table>

**QUESTION 2**

**HOTSPOT**

You need to implement an Azure Synapse Analytics database object for storing the sales transactions data. The solution must meet the sales transaction dataset requirements.

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

NOTE: Each correct selection is worth one point.

**Hot Area:**

**Answer Area**

Transact-SQL DDL command to use:

	▼
CREATE EXTERNAL TABLE	
CREATE TABLE	
CREATE VIEW	

Partitioning option to use in the WITH clause of the DDL statement:

	▼
FORMAT OPTIONS	
FORMAT TYPE	
RANGE LEFT FOR VALUES	
RANGE RIGHT FOR VALUES	

**Answer Area:**

**Answer Area**

Transact-SQL DDL command to use:

	▼
CREATE EXTERNAL TABLE	
CREATE TABLE	
CREATE VIEW	

Partitioning option to use in the WITH clause of the DDL statement:

	▼
FORMAT_OPTIONS	
FORMAT_TYPE	
RANGE LEFT FOR VALUES	
RANGE RIGHT FOR VALUES	

**Section:**

**Explanation:**

Box 1: Create table

Scenario: Load the sales transaction dataset to Azure Synapse Analytics

#### Box 2: RANGE RIGHT FOR VALUES

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

RANGE RIGHT: Specifies the boundary value belongs to the partition on the right (higher values). FOR VALUES ( boundary\_value [,...n] ): Specifies the boundary values for the partition.

Scenario: Load the sales transaction dataset to Azure Synapse Analytics. Contoso identifies the following requirements for the sales transaction dataset:

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. Ensure that queries joining and filtering sales transaction records based on product ID complete as quickly as possible. Implement a surrogate key to account for changes to the retail store addresses. Ensure that data storage costs and performance are predictable.

Minimize how long it takes to remove old records.

Reference:

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

#### QUESTION 3

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

- A. a table that has an IDENTITY property
- B. a system-versioned temporal table
- C. a user-defined SEQUENCE object
- D. a table that has a FOREIGN KEY constraint

**Correct Answer: A**

**Section:**

**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>

#### QUESTION 4

HOTSPOT

You need to design the partitions 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.

**Hot Area:**

**Answer Area**

Partition product sales transactions data by:

	▼
Sales date	
Product ID	
Promotion ID	

Store product sales transactions data in:

	▼
An Azure Synapse Analytics dedicated SQL pool	
An Azure Synapse Analytics serverless SQL pool	
An Azure Data Lake Storage Gen2 account linked to an Azure Synapse Analytics workspace	

Answer Area:

**Answer Area**

Partition product sales transactions data by:

	▼
Sales date	
Product ID	
Promotion ID	

Store product sales transactions data in:

	▼
An Azure Synapse Analytics dedicated SQL pool	
An Azure Synapse Analytics serverless SQL pool	
An Azure Data Lake Storage Gen2 account linked to an Azure Synapse Analytics workspace	



**Section:**

**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-what-is>

#### QUESTION 5

You need to design a data retention solution for the Twitter feed data records. The solution must meet the customer sentiment analytics requirements. Which Azure Storage functionality should you include in the solution?

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

**Correct Answer: B**

**Section:**

**Explanation:**

#### QUESTION 6

DRAG DROP

You need to ensure that the Twitter feed data can be analyzed in the dedicated SQL pool. The solution must meet the customer sentiment analytic requirements.

Which three Transact-SQL DDL commands should you run in sequence? To answer, move the appropriate commands from the list of commands to the answer area and arrange them in the correct order.

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

Select and Place:

**Commands**

- CREATE EXTERNAL DATA SOURCE
- CREATE EXTERNAL FILE FORMAT
- CREATE EXTERNAL TABLE
- CREATE EXTERNAL TABLE AS SELECT
- CREATE DATABASE SCOPED CREDENTIAL

**Answer Area**



**Correct Answer:**

**Commands**

- 
- 
- CREATE EXTERNAL TABLE
- 
- CREATE DATABASE SCOPED CREDENTIAL

**Answer Area**

- CREATE EXTERNAL DATA SOURCE
- CREATE EXTERNAL FILE FORMAT
- CREATE EXTERNAL TABLE AS SELECT

**Section:**

**Explanation:**

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

The users must be authenticated by using their own Azure AD credentials.

Box 1: CREATE EXTERNAL DATA SOURCE

External data sources are used to connect to storage accounts.

Box 2: CREATE EXTERNAL FILE FORMAT

CREATE EXTERNAL FILE FORMAT creates an external file format object that defines external data stored in Azure Blob Storage or Azure Data Lake Storage. Creating an external file format is a prerequisite for creating an external table.

Box 3: CREATE EXTERNAL TABLE AS SELECT

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

Incorrect Answers:

CREATE EXTERNAL TABLE

The CREATE EXTERNAL TABLE command creates an external table for Synapse SQL to access data stored in Azure Blob Storage or Azure Data Lake Storage.

Reference:

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

### QUESTION 7

You have an Azure subscription that contains a Microsoft Purview account named MP1, an Azure data factory named DF1, and a storage account named storage. MP1 is configured to scan storage1. DF1 is connected to MP1 and contains 3 datasets named DS1. DS1 references 2 files in storage. In DF1, you plan to create a pipeline that will process data from DS1. You need to review the schema and lineage information in MP1 for the data referenced by DS1. Which two features can you use to locate the information? Each correct answer presents a complete solution. NOTE: Each correct answer is worth one point.

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

**Correct Answer: C, D**

**Section:**

**Explanation:**

The search bar in the Microsoft Purview governance portal: This is a feature that allows you to search for assets in your data estate using keywords, filters, and facets. You can use the search bar to find the files in storage1 that are referenced by DS1, and then view their schema and lineage information in the asset details page.  
The search bar in Azure Data Factory Studio: This is a feature that allows you to search for datasets, linked services, pipelines, and other resources in your data factory. You can use the search bar to find DS1 in DF1, and then view its schema and lineage information in the dataset details page. You can also click on the Open in Purview button to open the corresponding asset in MP1.  
The two features that can be used to locate the schema and lineage information for the data referenced by DS1 are the search bar in Azure Data Factory Studio and the search bar in the Microsoft Purview governance portal. The search bar in Azure Data Factory Studio allows you to search for the dataset DS1 and view its properties and lineage. This can help you locate information about the source and destination data stores, as well as the transformations that were applied to the data. The search bar in the Microsoft Purview governance portal allows you to search for the storage account and view its metadata, including schema and lineage information. This can help you understand the different data assets that are stored in the storage account and how they are related to each other. The Storage browser of storage1 in the Azure portal may allow you to view the files that are stored in the storage account, but it does not provide lineage or schema information for those files. Similarly, the search bar in the Azure portal may allow you to search for resources in the Azure subscription, but it does not provide detailed information about the data assets themselves.  
Reference: What is Azure Purview? Use Azure Data Factory Studio

### 02-Design and implement data storage

#### QUESTION 1

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

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

- A. Yes
- B. No





**Correct Answer: A**

**Section:**

**Explanation:**

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

Reference:

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

### QUESTION 2

You are performing exploratory analysis of the bus fare data in an Azure Data Lake Storage Gen2 account by using an Azure Synapse Analytics serverless SQL pool.

You execute the Transact-SQL query shown in the following exhibit.

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

What do the query results include?

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

**Correct Answer: D**

**Section:**

### QUESTION 3

DRAG DROP

You have a table named SalesFact in an enterprise data warehouse in Azure Synapse Analytics. SalesFact contains sales data from the past 36 months and has the following characteristics:

Is partitioned by month

Contains one billion rows

Has clustered columnstore indexes

At the beginning of each month, you need to remove data from SalesFact that is older than 36 months as quickly as possible.

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

**Select and Place:**



## Actions

Switch the partition containing the stale data from SalesFact to SalesFact\_Work.

Truncate the partition containing the stale data.

Drop the SalesFact\_Work table.

Create an empty table named SalesFact\_Work that has the same schema as SalesFact.

Execute a DELETE statement where the value in the Date column is more than 36 months ago.

Copy the data to a new table by using CREATE TABLE AS SELECT (CTAS).

## Answer Area

## Answer Area

Create an empty table named SalesFact\_Work that has the same schema as SalesFact.

Switch the partition containing the stale data from SalesFact to SalesFact\_Work.

Drop the SalesFact\_Work table.

Correct Answer:

## Actions

Truncate the partition containing the stale data.

Execute a DELETE statement where the value in the Date column is more than 36 months ago.

Copy the data to a new table by using CREATE TABLE AS SELECT (CTAS).

Section:

Explanation:

Step 1: Create an empty table named SalesFact\_work that has the same schema as SalesFact.

Step 2: Switch the partition containing the stale data from SalesFact to SalesFact\_Work. SQL Data Warehouse supports partition splitting, merging, and switching. To switch partitions between two tables, you must ensure that the partitions align on their respective boundaries and that the table definitions match.

Loading data into partitions with partition switching is a convenient way stage new data in a table that is not visible to users the switch in the new data. Step 3: Drop the SalesFact\_Work table.

Reference:

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

#### QUESTION 4

HOTSPOT

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.

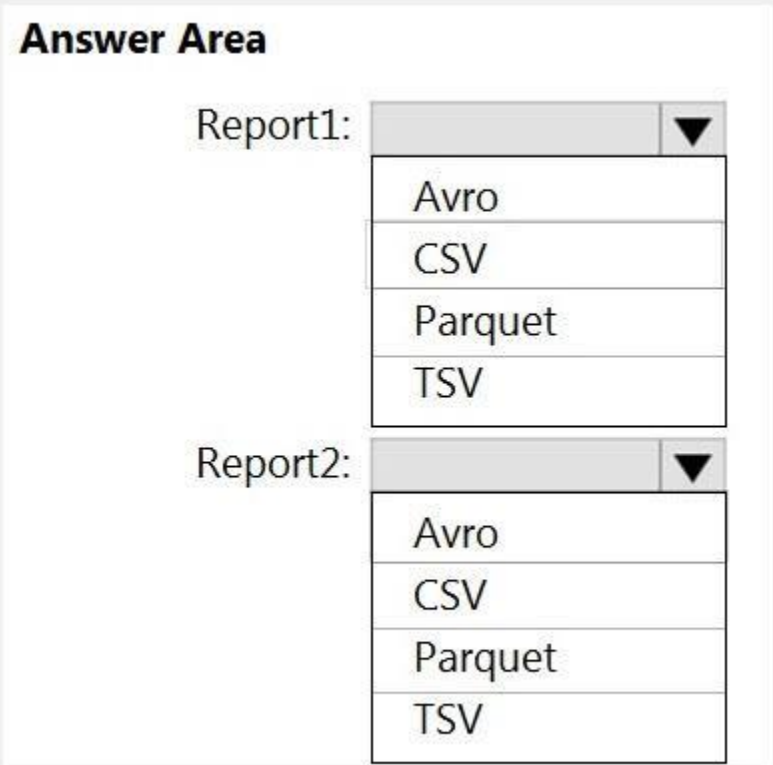
NOTE: Each correct selection is worth one point.

Hot Area:

**Answer Area**

Report1:

Report2:



Answer Area:



**Answer Area**

Report1:  ▼

Avro
CSV
Parquet
TSV

Report2:  ▼

Avro
CSV
Parquet
TSV

**Section:**

**Explanation:**

**QUESTION 5**

HOTSPOT

You need to output files from Azure Data Factory.

Which file format should you use for each type of output? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

**Hot Area:**



**Answer Area**

Columnar format:

	▼
Avro	
GZip	
Parquet	
TXT	

JSON with a timestamp:

	▼
Avro	
GZip	
Parquet	
TXT	

Answer Area:



**Answer Area**

Columnar format:

	▼
Avro	
GZip	
Parquet	
TXT	

JSON with a timestamp:

	▼
Avro	
GZip	
Parquet	
TXT	

Section:

Explanation:

Box 1: Parquet

Parquet stores data in columns, while Avro stores data in a row-based format. By their very nature, column-oriented data stores are optimized for read-heavy analytical workloads, while row-based databases are best for write-heavy transactional workloads.

Box 2: Avro

An Avro schema is created using JSON format.

AVRO supports timestamps.

Note: Azure Data Factory supports the following file formats (not GZip or TXT). Avro format

Binary format

Delimited text format

Excel format

JSON format

ORC format

Parquet format

XML format

Reference:

<https://www.datanami.com/2018/05/16/big-data-file-formats-demystified>

**QUESTION 6**

**HOTSPOT**

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.

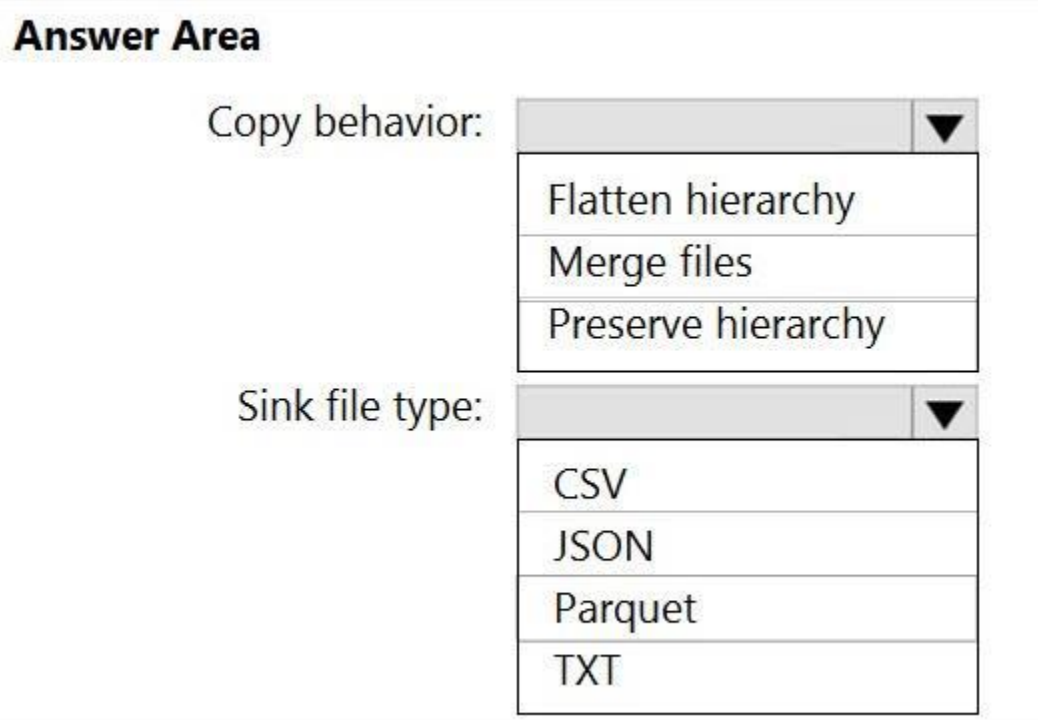
NOTE: Each correct selection is worth one point.

**Hot Area:**

**Answer Area**

Copy behavior:

Sink file type:



**Answer Area:**

**Answer Area**

Copy behavior:

- Flatten hierarchy
- Merge files
- Preserve hierarchy

Sink file type:

- CSV
- JSON
- Parquet
- TXT

**Section:**

**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>

**QUESTION 7**

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

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

- A. Yes
- B. No

**Correct Answer: B**

**Section:**

**Explanation:**

Instead convert the files to compressed delimited text files.

Reference:

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

### QUESTION 8

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

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

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

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

You need to prepare the files to ensure that the data copies quickly.

Solution: You modify the files to ensure that each row is more than 1 MB.

Does this meet the goal?

- A. Yes
- B. No

**Correct Answer: B**

**Section:**

**Explanation:**

### QUESTION 9

You build a data warehouse in an Azure Synapse Analytics dedicated SQL pool. Analysts write a complex SELECT query that contains multiple JOIN and CASE statements to transform data for use in inventory reports. The inventory reports will use the data and additional WHERE parameters depending on the report. The reports will be produced once daily.

You need to implement a solution to make the dataset available for the reports. The solution must minimize query times. What should you implement?

- A. an ordered clustered columnstore index
- B. a materialized view
- C. result set caching
- D. a replicated table



**Correct Answer: B**

**Section:**

**Explanation:**

Materialized views for dedicated SQL pools in Azure Synapse provide a low maintenance method for complex analytical queries to get fast performance without any query change. Incorrect Answers:

C: One daily execution does not make use of result cache caching. Note: When result set caching is enabled, dedicated SQL pool automatically caches query results in the user database for repetitive use. This allows subsequent query executions to get results directly from the persisted cache so recomputation is not needed. Result set caching improves query performance and reduces compute resource usage. In addition, queries using cached results set do not use any concurrency slots and thus do not count against existing concurrency limits.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/performance-tuning-materialized-views> <https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/performance-tuning-result-set-caching>

### QUESTION 10

You have an Azure Synapse Analytics workspace named WS1 that contains an Apache Spark pool named Pool1. You plan to create a database named DB1 in Pool1.

You need to ensure that when tables are created in DB1, the tables are available automatically as external tables to the built-in serverless SQL pool. Which format should you use for the tables in DB1?

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

**Correct Answer: D**



**Section:****Explanation:**

Serverless SQL pool can automatically synchronize metadata from Apache Spark. A serverless SQL pool database will be created for each database existing in serverless Apache Spark pools. For each Spark external table based on Parquet or CSV and located in Azure Storage, an external table is created in a serverless SQL pool database.

Reference:

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

**QUESTION 11**

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

```
CREATE TABLE [dbo].[DimEmployee] (  
    [EmployeeKey] [int] IDENTITY(1,1) NOT NULL,  
    [EmployeeID] [int] NOT NULL,  
    [FirstName] [varchar](100) NOT NULL,  
    [LastName] [varchar](100) NOT NULL,  
    [JobTitle] [varchar](100) NULL,  
    [LastHireDate] [date] NULL,  
    [StreetAddress] [varchar](500) NOT NULL,  
    [City] [varchar](200) NOT NULL,  
    [StateProvince] [varchar](50) NOT NULL,  
    [Postalcode] [varchar](10) NOT NULL  
)
```

You need to alter the table to meet the following requirements:

Ensure that users can identify the current manager of employees. Support creating an employee reporting hierarchy for your entire company. Provide fast lookup of the managers' attributes such as name and job title.

Which column should you add to the table?

- A. [ManagerEmployeeID] [smallint] NULL
- B. [ManagerEmployeeKey] [smallint] NULL
- C. [ManagerEmployeeKey] [int] NULL
- D. [ManagerName] [varchar](200) NULL

**Correct Answer: C**

**Section:****Explanation:**

We need an extra column to identify the Manager. Use the data type as the EmployeeKey column, an int column.

Reference: <https://docs.microsoft.com/en-us/analysis-services/tabular-models/hierarchies-ssas-tabular>

**QUESTION 12**

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

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

FactPurchase will have 1 million rows of data added daily and will contain three years of data.

Transact-SQL queries similar to the following query will be executed daily.

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

Which table distribution will minimize query times?

- A. replicated
- B. hash-distributed on PurchaseKey
- C. round-robin
- D. hash-distributed on IsOrderFinalized

**Correct Answer: B**

**Section:**

**Explanation:**

Hash-distributed tables improve query performance on large fact tables. To balance the parallel processing, select a distribution column that:

Has many unique values. The column can have duplicate values. All rows with the same value are assigned to the same distribution. Since there are 60 distributions, some distributions can have > 1 unique values while others may end with zero values.

Does not have NULLs, or has only a few NULLs. Is not a date column. Incorrect Answers:

C: 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-distribute>

### QUESTION 13

You are planning a solution to aggregate streaming data that originates in Apache Kafka and is output to Azure Data Lake Storage Gen2. The developers who will implement the stream processing solution use Java. Which service should you recommend using to process the streaming data?

- A. Azure Event Hubs
- B. Azure Data Factory
- C. Azure Stream Analytics



D. Azure Databricks

**Correct Answer: D**

**Section:**

**Explanation:**

#### QUESTION 14

You plan to implement an Azure Data Lake Storage Gen2 container that will contain CSV files. The size of the files will vary based on the number of events that occur per hour. File sizes range from 4 KB to 5 GB. You need to ensure that the files stored in the container are optimized for batch processing. What should you do?

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

**Correct Answer: B**

**Section:**

**Explanation:**

Avro supports batch and is very relevant for streaming.

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

Reference:

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



#### QUESTION 15

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

TransactionType: 40 million rows per transaction type

CustomerSegment: 4 million per customer segment

TransactionMonth: 65 million rows per month AccountType: 500 million per account type You have the following query requirements:

Analysts will most commonly analyze transactions for a given month. Transactions analysis will typically summarize transactions by transaction type, customer segment, and/or account type You need to recommend a partition strategy for the table to minimize query times. On which column should you recommend partitioning the table?

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

**Correct Answer: D**

**Section:**

#### QUESTION 16

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.

EmployeeName	EmployeeID	EmployeeStartDate
Alice	24	2020-01-25

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?

- A. 24
- B. an error
- C. a null value

**Correct Answer: A**

**Section:**

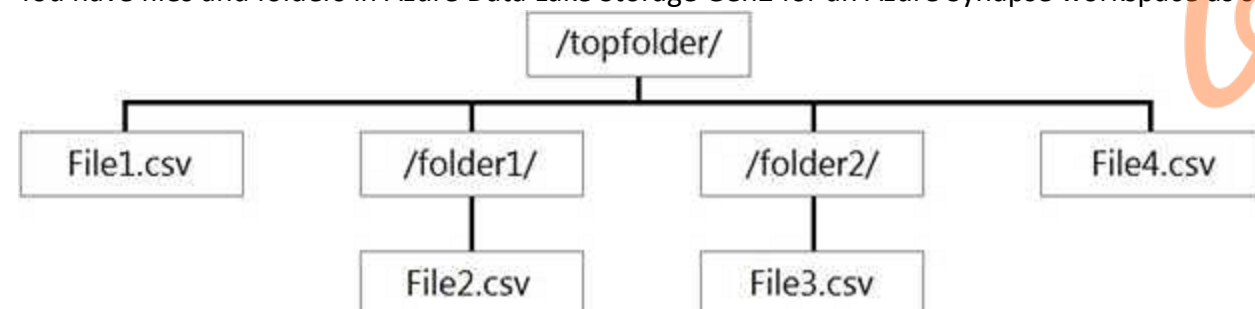
**Explanation:**

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>

#### QUESTION 17

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



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

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

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

**Correct Answer: B**

**Section:**

**Explanation:**

Reference: <https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/query-data-storage#query-multiple-files-or-folders>

#### QUESTION 18

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?

- A. /{SubjectArea}/{DataSource}/{DD}/{MM}/{YYYY}/{FileData}\_{YYYY}\_{MM}\_{DD}.csv
- B. /{DD}/{MM}/{YYYY}/{SubjectArea}/{DataSource}/{FileData}\_{YYYY}\_{MM}\_{DD}.csv
- C. /{YYYY}/{MM}/{DD}/{SubjectArea}/{DataSource}/{FileData}\_{YYYY}\_{MM}\_{DD}.csv
- D. /{SubjectArea}/{DataSource}/{YYYY}/{MM}/{DD}/{FileData}\_{YYYY}\_{MM}\_{DD}.csv

**Correct Answer: D**

**Section:**

**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}/

#### QUESTION 19

You need to design an Azure Synapse Analytics dedicated SQL pool that meets the following requirements:

Can return an employee record from a given point in time.

Maintains the latest employee information. Minimizes query complexity. How should you model the employee data?

- A. as a temporal table
- B. as a SQL graph table
- C. as a degenerate dimension table
- D. as a Type 2 slowly changing dimension (SCD) table

**Correct Answer: D**

**Section:**

**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-pipelines/3-choose-between-dimension-types>

#### QUESTION 20

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.

- A. Add the managed identity to the Sales group.
- B. Use the managed identity as the credentials for the data load process.
- C. Create a shared access signature (SAS).
- D. Add your Azure Active Directory (Azure AD) account to the Sales group.
- E. Use the shared access signature (SAS) as the credentials for the data load process.
- F. Create a managed identity.



**Correct Answer: B, D, F**

**Section:**

**Explanation:**

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

**QUESTION 21**

You have an enterprise data warehouse in Azure Synapse Analytics. Using PolyBase, you create an external table named [Ext].[Items] to query Parquet files stored in Azure Data Lake Storage Gen2 without importing the data to the data warehouse. The external table has three columns.

You discover that the Parquet files have a fourth column named ItemID. Which command should you run to add the ItemID column to the external table?

- A. 

```
ALTER EXTERNAL TABLE [Ext].[Items]
  ADD [ItemID] int;
```
- B. 

```
DROP EXTERNAL FILE FORMAT parquetfile1;
CREATE EXTERNAL FILE FORMAT parquetfile1
WITH (
  FORMAT_TYPE = PARQUET,
  DATA_COMPRESSION = 'org.apache.hadoop.io.compress.SnappyCodec'
);
```
- C. 

```
DROP EXTERNAL TABLE [Ext].[Items]
CREATE EXTERNAL TABLE [Ext].[Items]
([ItemID] [int] NULL,
 [ItemName] nvarchar(50) NULL,
 [ItemType] nvarchar(20) NULL,
 [ItemDescription] nvarchar(250))
WITH
(
  LOCATION= '/Items/',
  DATA_SOURCE = AzureDataLakeStore,
  FILE_FORMAT = PARQUET,
  REJECT_TYPE = VALUE,
  REJECT_VALUE = 0
);
```
- D. 

```
ALTER TABLE [Ext].[Items]
  ADD [ItemID] int;
```

A. Option A

B. Option B

C. Option C

D. Option D

**Correct Answer: C**

**Section:**

**Explanation:**

Incorrect Answers:

A, D: Only these Data Definition Language (DDL) statements are allowed on external tables:



CREATE TABLE and DROP TABLE

CREATE STATISTICS and DROP STATISTICS CREATE VIEW and DROP VIEW

Reference: <https://docs.microsoft.com/en-us/sql/t-sql/statements/create-external-table-transact-sql>

#### QUESTION 22

You have an Azure Data Lake Storage Gen2 container that contains 100 TB of data. You need to ensure that the data in the container is available for read workloads in a secondary region if an outage occurs in the primary region. The solution must minimize costs. Which type of data redundancy should you use?

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

**Correct Answer: B**

**Section:**

**Explanation:**

Geo-redundant storage (with GRS or GZRS) replicates your data to another physical location in the secondary region to protect against regional outages. However, that data is available to be read only if the customer or Microsoft initiates a failover from the primary to secondary region. When you enable read access to the secondary region, your data is available to be read at all times, including in a situation where the primary region becomes unavailable. Incorrect Answers:

A: While Geo-redundant storage (GRS) is cheaper than Read-Access Geo-Redundant Storage (RA-GRS), GRS does NOT initiate automatic failover. C, D: Locally redundant storage (LRS) and Zone-redundant storage (ZRS) provides redundancy within a single region.

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

#### QUESTION 23

You plan to implement an Azure Data Lake Gen 2 storage account. You need to ensure that the data lake will remain available if a data center fails in the primary Azure region. The solution must minimize costs. Which type of replication should you use for the storage account?

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

**Correct Answer: D**

**Section:**

**Explanation:**

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

#### QUESTION 24

You plan to ingest streaming social media data by using Azure Stream Analytics. The data will be stored in files in Azure Data Lake Storage, and then consumed by using Azure Databricks and PolyBase in Azure Synapse Analytics. You need to recommend a Stream Analytics data output format to ensure that the queries from Databricks and PolyBase against the files encounter the fewest possible errors. The solution must ensure that the files can be queried quickly and that the data type information is retained.

What should you recommend?

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

**Correct Answer: B**

**Section:**

**Explanation:**

Need Parquet to support both Databricks and PolyBase.

Reference:

<https://docs.microsoft.com/en-us/sql/t-sql/statements/create-external-file-format-transact-sql>

**QUESTION 25**

**HOTSPOT**

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

The data contains the following columns.

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

You need to design a star schema to support analytical queries of the data. The star schema will contain four tables including a date dimension.

To which table should you add each column? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

**Hot Area:**





### Answer Area

EventCategory:

	▼
DimChannel	
DimDate	
DimEvent	
FactEvents	

ChannelGrouping:

	▼
DimChannel	
DimDate	
DimEvent	
FactEvents	

TotalEvents:

	▼
DimChannel	
DimDate	
DimEvent	
FactEvents	



Answer Area:

**Answer Area**

EventCategory: 

▼
DimChannel
DimDate
DimEvent
FactEvents

ChannelGrouping: 

▼
DimChannel
DimDate
DimEvent
FactEvents

TotalEvents: 

▼
DimChannel
DimDate
DimEvent
FactEvents



**Section:**

**Explanation:**

Box 1: DimEvent

Box 2: DimChannel

Box 3: FactEvents

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

Reference:

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

**QUESTION 26**

HOTSPOT

You store files in an Azure Data Lake Storage Gen2 container. The container has the storage policy shown in the following exhibit.

```
{
  "rules": [
    {
      "enabled": true,
      "name": "contosorule",
      "type": "Lifecycle",
      "definition": {
        "actions": {
          "version": {
            "delete": {
              "daysAfterCreationGreaterThan": 60
            }
          },
          "baseBlob": {
            "tierToCool": {
              "daysAfterModificationGreaterThan":
30
            },
          },
        }
      },
      "filters": {
        "blobTypes": [
          "blockBlob"
        ],
        "prefixMatch": [
          "container1/contoso"
        ]
      }
    }
  ]
}
```



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

NOTE: Each correct selection is worth one point.

**Hot Area:**

## Answer Area

The files are [answer choice] after 30 days:

	▼
deleted from the container	
moved to archive storage	
moved to cool storage	
moved to hot storage	

The storage policy applies to [answer choice]:

	▼
container1/contoso.csv	
container1/docs/contoso.json	
container1/mycontoso/contoso.csv	

Answer Area:

## Answer Area

The files are [answer choice] after 30 days:

	▼
deleted from the container	
moved to archive storage	
moved to cool storage	
moved to hot storage	

The storage policy applies to [answer choice]:

	▼
container1/contoso.csv	
container1/docs/contoso.json	
container1/mycontoso/contoso.csv	

### Section:

### Explanation:

Box 1: moved to cool storage

The ManagementPolicyBaseBlob.TierToCool property gets or sets the function to tier blobs to cool storage. Support blobs currently at Hot tier.

Box 2: container1/contoso.csv

As defined by prefixMatch.

prefixMatch: An array of strings for prefixes to be matched. Each rule can define up to 10 case-sensitive prefixes. A prefix string must start with a container name.

Reference:

<https://docs.microsoft.com/en-us/dotnet/api/microsoft.azure.management.storage.fluent.models.managementpolicybaseblob.tiertocool>

### QUESTION 27

#### HOTSPOT

You have an Azure Data Lake Storage Gen2 account named account1 that stores logs as shown in the following table.

Type	Designated retention period
Application	360 days
Infrastructure	60 days

You do not expect that the logs will be accessed during the retention periods.

You need to recommend a solution for account1 that meets the following requirements:

Automatically deletes the logs at the end of each retention period  
Minimizes storage costs

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

NOTE: Each correct selection is worth one point.

Hot Area:

### Answer Area

To minimize storage costs:

	▼
Store the infrastructure logs and the application logs in the Archive access tier	
Store the infrastructure logs and the application logs in the Cool access tier	
Store the infrastructure logs in the Cool access tier and the application logs in the Archive access tier	

To delete logs automatically:

	▼
Azure Data Factory pipelines	
Azure Blob storage lifecycle management rules	
Immutable Azure Blob storage time-based retention policies	

Answer Area:

## Answer Area

To minimize storage costs:

	▼
Store the infrastructure logs and the application logs in the Archive access tier	
Store the infrastructure logs and the application logs in the Cool access tier	
Store the infrastructure logs in the Cool access tier and the application logs in the Archive access tier	

To delete logs automatically:

	▼
Azure Data Factory pipelines	
Azure Blob storage lifecycle management rules	
Immutable Azure Blob storage time-based retention policies	



### Section:

### Explanation:

Box 1: Store the infrastructure logs in the Cool access tier and the application logs in the Archive access tier For infrastructure logs: Cool tier - An online tier optimized for storing data that is infrequently accessed or modified. Data in the cool tier should be stored for a minimum of 30 days. The cool tier has lower storage costs and higher access costs compared to the hot tier.

For application logs: Archive tier - An offline tier optimized for storing data that is rarely accessed, and that has flexible latency requirements, on the order of hours. Data in the archive tier should be stored for a minimum of 180 days.

Box 2: Azure Blob storage lifecycle management rules

Blob storage lifecycle management offers a rule-based policy that you can use to transition your data to the desired access tier when your specified conditions are met. You can also use lifecycle management to expire data at the end of its life.

Reference:

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

### QUESTION 28

You are implementing a batch dataset in the Parquet format. Data files will be produced by using Azure Data Factory and stored in Azure Data Lake Storage Gen2. The files will be consumed by an Azure Synapse Analytics serverless SQL pool. You need to minimize storage costs for the solution.

What should you do?

- A. Use Snappy compression for files.
- B. Use OPENROWSET to query the Parquet files.
- C. Create an external table that contains a subset of columns from the Parquet files.
- D. Store all data as string in the Parquet files.

**Correct Answer: C**

**Section:**

**Explanation:**

An external table points to data located in Hadoop, Azure Storage blob, or Azure Data Lake Storage. External tables are used to read data from files or write data to files in Azure Storage. With Synapse SQL, you can use external tables to read external data using dedicated SQL pool or serverless SQL pool.

Reference:

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

**QUESTION 29**

DRAG DROP

You have data stored in thousands of CSV files in Azure Data Lake Storage Gen2. Each file has a header row followed by a properly formatted carriage return (/r) and line feed (/n).

You are implementing a pattern that batch loads the files daily into an enterprise data warehouse in Azure Synapse Analytics by using PolyBase.

You need to skip the header row when you import the files into the data warehouse. Before building the loading pattern, you need to prepare the required database objects in Azure Synapse Analytics.

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


NOTE: Each correct selection is worth one point

**Select and Place:**

**Actions**

- Create a database scoped credential that uses Azure Active Directory Application and a Service Principal Key
- Create an external data source that uses the abfs location
- Use CREATE EXTERNAL TABLE AS SELECT (CETAS) and configure the reject options to specify reject values or percentages
- Create an external file format and set the First\_Row option

**Answer Area**



The interface shows a drag-and-drop question. On the left, under 'Actions', there are four boxes with different tasks. On the right, under 'Answer Area', there are two circular arrows (right and left) and a large watermark logo for 'Vdumps'.

**Correct Answer:**

**Actions**

- Create a database scoped credential that uses Azure Active Directory Application and a Service Principal Key
- 
- 
- 

**Answer Area**

- Create an external data source that uses the abfs location
- Create an external file format and set the First\_Row option
- Use CREATE EXTERNAL TABLE AS SELECT (CETAS) and configure the reject options to specify reject values or percentages

The 'Correct Answer' interface shows the same 'Actions' list on the left. In the 'Answer Area' on the right, three boxes are present. The first two boxes are highlighted with a light blue border, indicating they are the correct sequence of actions to be selected and placed in order.

**Section:**

**Explanation:**

Step 1: Create an external data source that uses the abfs location Create External Data Source to reference Azure Data Lake Store Gen 1 or 2

Step 2: Create an external file format and set the First\_Row option. Create External File Format.



Step 3: Use CREATE EXTERNAL TABLE AS SELECT (CETAS) and configure the reject options to specify reject values or percentages To use PolyBase, you must create external tables to reference your external data. Use reject options.

Note: REJECT options don't apply at the time this CREATE EXTERNAL TABLE AS SELECT statement is run. Instead, they're specified here so that the database can use them at a later time when it imports data from the external table. Later, when the CREATE TABLE AS SELECT statement selects data from the external table, the database will use the reject options to determine the number or percentage of rows that can fail to import before it stops the import.

Reference:

<https://docs.microsoft.com/en-us/sql/relational-databases/polybase/polybase-t-sql-objects>

<https://docs.microsoft.com/en-us/sql/t-sql/statements/create-external-table-as-select-transact-sql>

### QUESTION 30

#### HOTSPOT

You have a Microsoft SQL Server database that uses a third normal form schema.

You plan to migrate the data in the database to a star schema in an Azure Synapse Analytics dedicated SQL pool.

You need to design the dimension tables. The solution must optimize read operations.

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.

#### Hot Area:

**Answer Area**

Transform data for the dimension tables by:

	▼
Maintaining to a third normal form	
Normalizing to a fourth normal form	
Denormalizing to a second normal form	

For the primary key columns in the dimension tables, use:

	▼
New IDENTITY columns	
A new computed column	
The business key column from the source sys	

#### Answer Area:

**Answer Area**

Transform data for the dimension tables by:

	▼
Maintaining to a third normal form	
Normalizing to a fourth normal form	
Denormalizing to a second normal form	

For the primary key columns in the dimension tables, use:

	▼
New IDENTITY columns	
A new computed column	
The business key column from the source sys	

#### Section:

#### Explanation:

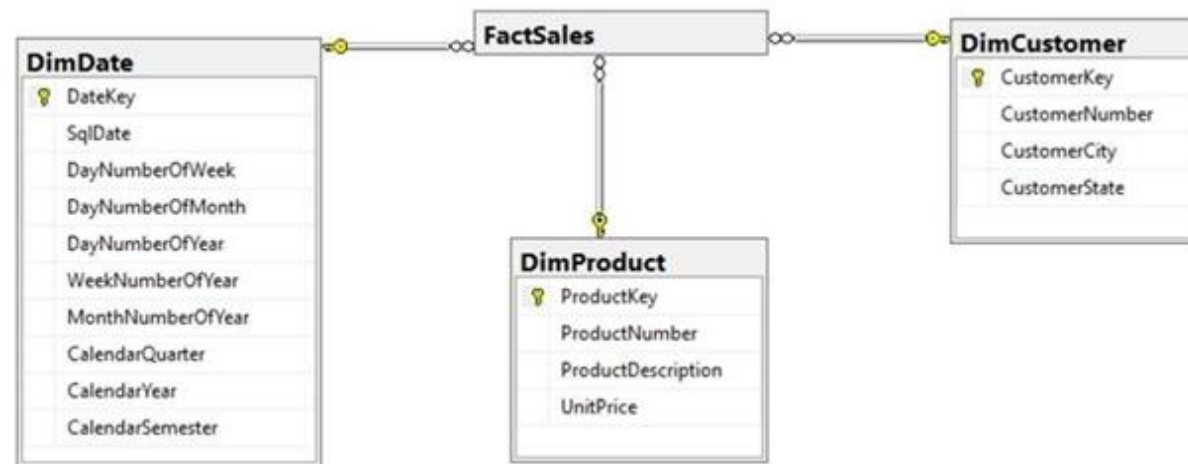
Box 1: Denormalize to a second normal form

Denormalization is the process of transforming higher normal forms to lower normal forms via storing the join of higher normal form relations as a base relation. Denormalization increases the performance in data retrieval at cost of bringing update anomalies to a database.

Box 2: New identity columns

The collapsing relations strategy can be used in this step to collapse classification entities into component entities to obtain fact dimension tables with single-part keys that connect directly to the fact table. The single-part key is a surrogate key generated to ensure it remains unique over time.

Example:



Note: 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://www.mssqltips.com/sqlservertip/5614/explore-the-role-of-normal-forms-in-dimensional-modeling/>

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

### QUESTION 31

#### HOTSPOT

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

ProductID

ItemPrice

LineTotal

Quantity

StoreID

Minute

Month

Hour

Year

Day

You need to store the data to support hourly incremental load pipelines that will vary for each Store ID. The solution must minimize storage costs.

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

NOTE: Each correct selection is worth one point.

**Hot Area:**

### Answer Area

df.write

	▼
.bucketBy	
.partitionBy	
.range	
.sortBy	

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

.mode("append")

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

Answer Area:

### Answer Area

df.write

	▼
.bucketBy	
.partitionBy	
.range	
.sortBy	

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

.mode("append")

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

Section:

Explanation:

Box 1: partitionBy

We should overwrite at the partition level.

Example:

```
df.write.partitionBy("y","m","d")
```

```
.mode(SaveMode.Append)
```

```
.parquet("/data/hive/warehouse/db_name.db/" + tableName)
```

Box 2: ("StoreID", "Year", "Month", "Day", "Hour", "StoreID")

Box 3: parquet("/Purchases")

Reference:

<https://intellipaat.com/community/11744/how-to-partition-and-write-dataframe-in-spark-without-deleting-partitions-with-no-new-data>

### QUESTION 32

You have an Azure Synapse Analytics dedicated SQL pool named Pool1. Pool1 contains a partitioned fact table named dbo.Sales and a staging table named stg.Sales that has the matching table and partition definitions. You need to overwrite the content of the first partition in dbo.Sales with the content of the same partition in stg.Sales. The solution must minimize load times. What should you do?

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

**Correct Answer: C**

**Section:**

**Explanation:**

Reference:

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

### QUESTION 33

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

You plan to keep a record of changes to the available fields.

The supplier data contains the following columns.

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

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

NOTE: Each correct selection is worth one point.

- A. surrogate primary key
- B. effective start date

- C. business key
- D. last modified date
- E. effective end date
- F. foreign key

**Correct Answer: A, B, E**

**Section:**

**Explanation:**

<https://learn.microsoft.com/en-us/training/modules/populate-slowly-changing-dimensions-azure-synapse-analytics-pipelines/3-choose-between-dimension-types>

#### QUESTION 34

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

Contain sales data for 20,000 products.

Use hash distribution on a column named ProductID.

Contain 2.4 billion records for the years 2019 and 2020.

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

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

**Correct Answer: A**

**Section:**

**Explanation:**

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

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

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

Reference:

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

#### QUESTION 35

HOTSPOT

You have two Azure Storage accounts named Storage1 and Storage2. Each account holds one container and has the hierarchical namespace enabled. The system has files that contain data stored in the Apache Parquet format.

You need to copy folders and files from Storage1 to Storage2 by using a Data Factory copy activity. The solution must meet the following requirements:

No transformations must be performed.

The original folder structure must be retained.

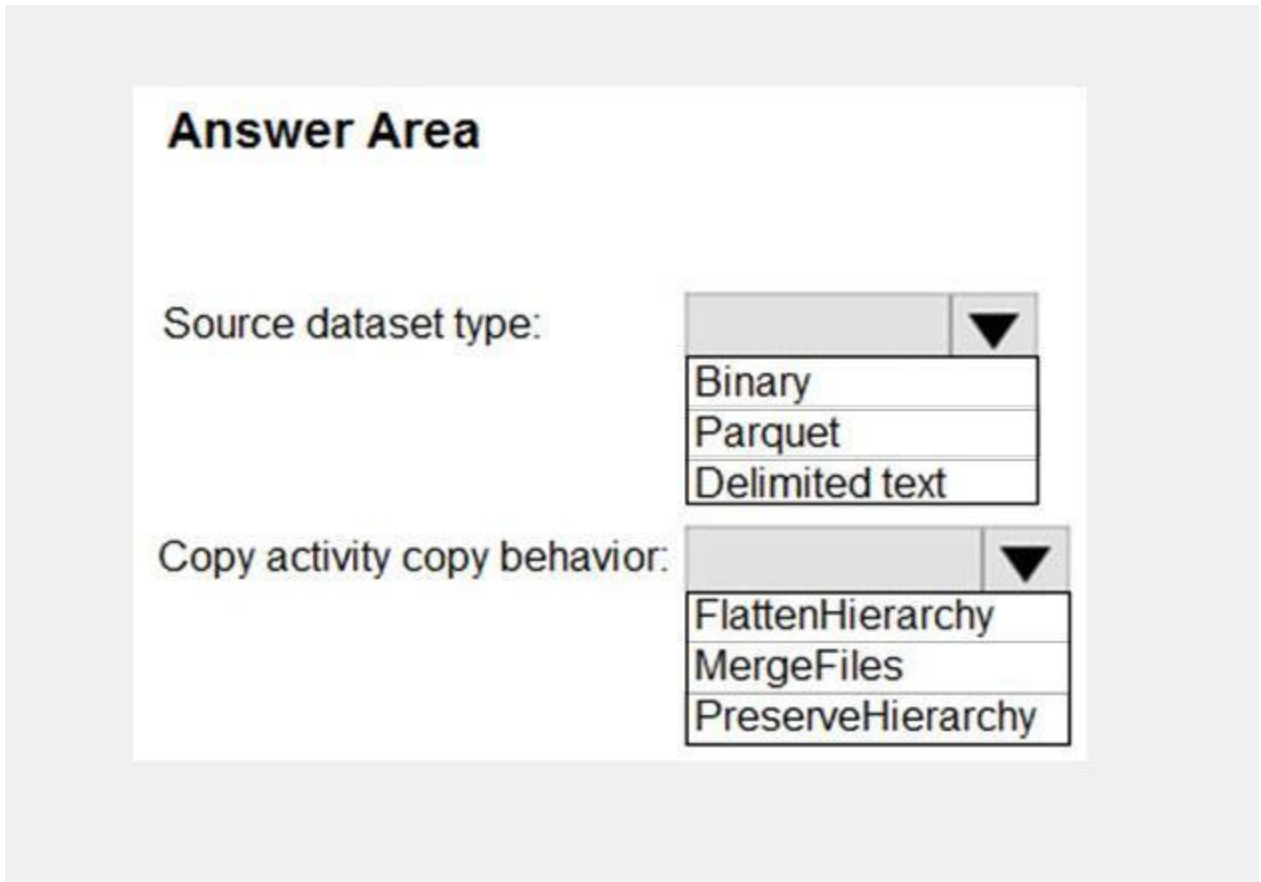
Minimize time required to perform the copy activity.

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

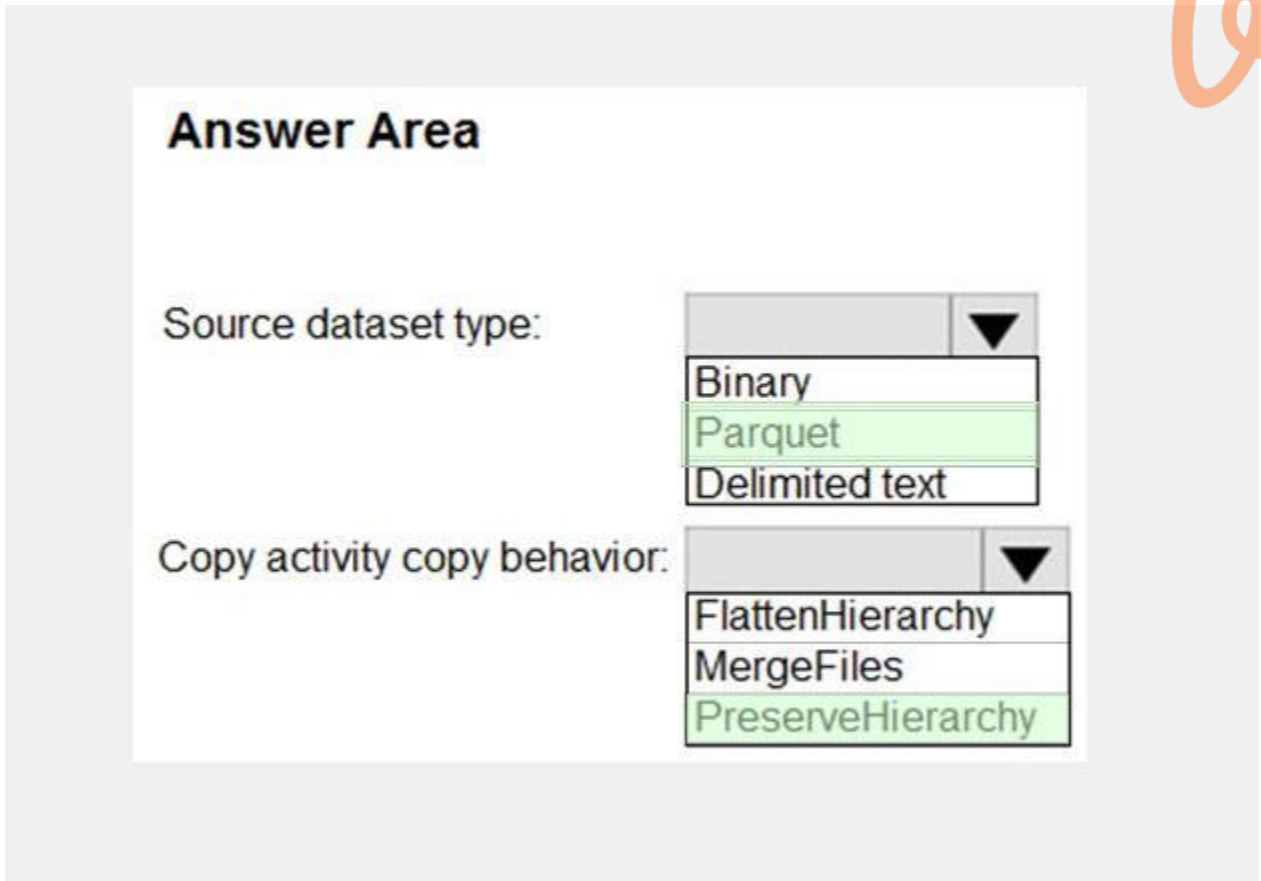
NOTE: Each correct selection is worth one point.

**Hot Area:**





Answer Area:



**Section:**

**Explanation:**

Box 1: Parquet

For Parquet datasets, the type property of the copy activity source must be set to ParquetSource.

Box 2: PreserveHierarchy

PreserveHierarchy (default): Preserves the file hierarchy in the target folder. The relative path of the source file to the source folder is identical to the relative path of the target file to the target folder. Incorrect Answers:

FlattenHierarchy: All files from the source folder are in the first level of the target folder. The target files have autogenerated names. MergeFiles: Merges all files from the source folder to one file. If the file name is specified, the merged file name is the specified name. Otherwise, it's an autogenerated file name.

Reference:

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

<https://docs.microsoft.com/en-us/azure/data-factory/connector-azure-data-lake-storage>

### QUESTION 36

HOTSPOT

You have a SQL pool in Azure Synapse.

You plan to load data from Azure Blob storage to a staging table. Approximately 1 million rows of data will be loaded daily. The table will be truncated before each daily load.

You need to create the staging table. The solution must minimize how long it takes to load the data to the staging table.

How should you configure the table? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

**Answer Area**

Distribution:  ▼

Indexing:  ▼

Partitioning:  ▼

Answer Area:



**Answer Area**

Distribution:

Indexing:

Partitioning:



**Section:**

**Explanation:**

Box 1: Hash

Hash-distributed tables improve query performance on large fact tables. They can have very large numbers of rows and still achieve high performance.

Incorrect Answers:

Round-robin tables are useful for improving loading speed.

Box 2: Clustered columnstore

When creating partitions on clustered columnstore tables, it is important to consider how many rows belong to each partition. For optimal compression and performance of clustered columnstore tables, a minimum of 1 million rows per distribution and partition is needed.

Box 3: Date

Table partitions enable you to divide your data into smaller groups of data. In most cases, table partitions are created on a date column. Partition switching can be used to quickly remove or replace a section of a table.

Reference:

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

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

**QUESTION 37**

**HOTSPOT**

You are creating dimensions for a data warehouse in an Azure Synapse Analytics dedicated SQL pool.

You create a table by using the Transact-SQL statement shown in the following exhibit.



```

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

```


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

Hot Area:

**Answer Area**

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

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



▼

Type 0

Type 1

Type 2

▼

a surrogate key

a business key

an audit column

Answer Area:

### Answer Area

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

▼

- Type 0
- Type 1
- Type 2

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

▼

- a surrogate key
- a business key
- an audit column

Section:

Explanation:



### QUESTION 38

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

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

FactPurchase will have 1 million rows of data added daily and will contain three years of data.

Transact-SQL queries similar to the following query will be executed daily.

```
SELECT  
SupplierKey, StockItemKey, COUNT(*)  
FROM FactPurchase  
WHERE DateKey >= 20210101
```

AND DateKey <= 20210131

GROUP By SupplierKey, StockItemKey

Which table distribution will minimize query times?

- A. replicated
- B. hash-distributed on PurchaseKey
- C. round-robin
- D. hash-distributed on DateKey

**Correct Answer: B**

**Section:**

**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. Incorrect:

Not D: Do not use a date column. . All data for the same date lands in the same distribution. If several users are all filtering on the same date, then only 1 of the 60 distributions do all the processing work.

Reference:

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

### QUESTION 39

HOTSPOT

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

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

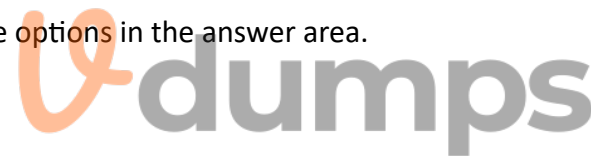
Minimizes the processing time to delete data that is older than 10 years

Minimizes the I/O for queries that use year-to-date values

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

NOTE: Each correct selection is worth one point.

**Hot Area:**



```

CREATE TABLE [dbo].[FactTransaction]
(
    [TransactionTypeID]    int    NOT NULL
,   [TransactionDateID]   int    NOT NULL
,   [CustomerID]          int    NOT NULL
,   [RecipientID]         int    NOT NULL
,   [Amount]              money  NOT NU::
)

```

WITH

	▼
CLUSTERED COLUMNSTORE INDEX	
DISTRIBUTION	
PARTITION	
TRUNCATE_TARGET	

	▼
[TransactionDateID]	
[TransactionDateID], [TransactionTypeID]	
HASH([TransactionTypeID])	
ROUND ROBIN	

RANGE RIGHT FOR VALUES

```

(20200101,20200201,20200301,20200401,20200501,20200601)

```

Answer Area:

```
CREATE TABLE [dbo].[FactTransaction]
(
    [TransactionTypeID]    int    NOT NULL
,   [TransactionDateID]   int    NOT NULL
,   [CustomerID]          int    NOT NULL
,   [RecipientID]        int    NOT NULL
,   [Amount]              money  NOT NU::
)

```

WITH

CLUSTERED COLUMNSTORE INDEX
DISTRIBUTION
PARTITION
TRUNCATE_TARGET

[TransactionDateID]
[TransactionDateID], [TransactionTypeID]
HASH([TransactionTypeID])
ROUND ROBIN

RANGE RIGHT FOR VALUES

```
(20200101, 20200201, 20200301, 20200401, 20200501, 20200601)
```

**Section:**

**Explanation:**

Box 1: PARTITION

RANGE RIGHT FOR VALUES is used with PARTITION.

Part 2: [TransactionDateID]

Partition on the date column.

Example: Creating a RANGE RIGHT partition function on a datetime column The following partition function partitions a table or index into 12 partitions, one for each month of a year's worth of values in a datetime column.

```
CREATE PARTITION FUNCTION [myDateRangePF1] (datetime)
```

```
AS RANGE RIGHT FOR VALUES ('20030201', '20030301', '20030401', '20030501', '20030601', '20030701', '20030801',
'20030901', '20031001', '20031101', '20031201');
```

Reference:

<https://docs.microsoft.com/en-us/sql/t-sql/statements/create-partition-function-transact-sql>

**QUESTION 40**

You are designing a data mart for the human resources (HR) department at your company. The data mart will contain employee information and employee transactions.

From a source system, you have a flat extract that has the following fields:

EmployeeID

FirstName

LastName

Recipient  
GrossAmount  
TransactionID  
GovernmentID  
NetAmountPaid  
TransactionDate

You need to design a star schema data model in an Azure Synapse Analytics dedicated SQL pool for the data mart.

Which two tables should you create? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

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

**Correct Answer: C, E**

**Section:**

**Explanation:**

C: Dimension tables contain attribute data that might change but usually changes infrequently. For example, a customer's name and address are stored in a dimension table and updated only when the customer's profile changes. To minimize the size of a large fact table, the customer's name and address don't need to be in every row of a fact table. Instead, the fact table and the dimension table can share a customer ID. A query can join the two tables to associate a customer's profile and transactions.

E: Fact tables contain quantitative data that are commonly generated in a transactional system, and then loaded into the dedicated SQL pool. For example, a retail business generates sales transactions every day, and then loads the data into a dedicated SQL pool fact table for analysis.

Reference:

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

#### QUESTION 41

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 you use?

- A. Type 0
- B. Type 1
- C. Type 2
- D. Type 3

**Correct Answer: C**

**Section:**

**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.

Incorrect Answers:

B: A Type 1 SCD always reflects the latest values, and when changes in source data are detected, the dimension table data is overwritten. D: A Type 3 SCD supports storing two versions of a dimension member as separate columns. The table includes a column for the current value of a member plus either the original or previous value of the member. So Type 3 uses additional columns to track one key instance of history, rather than storing additional rows to track each change like in a Type 2 SCD.

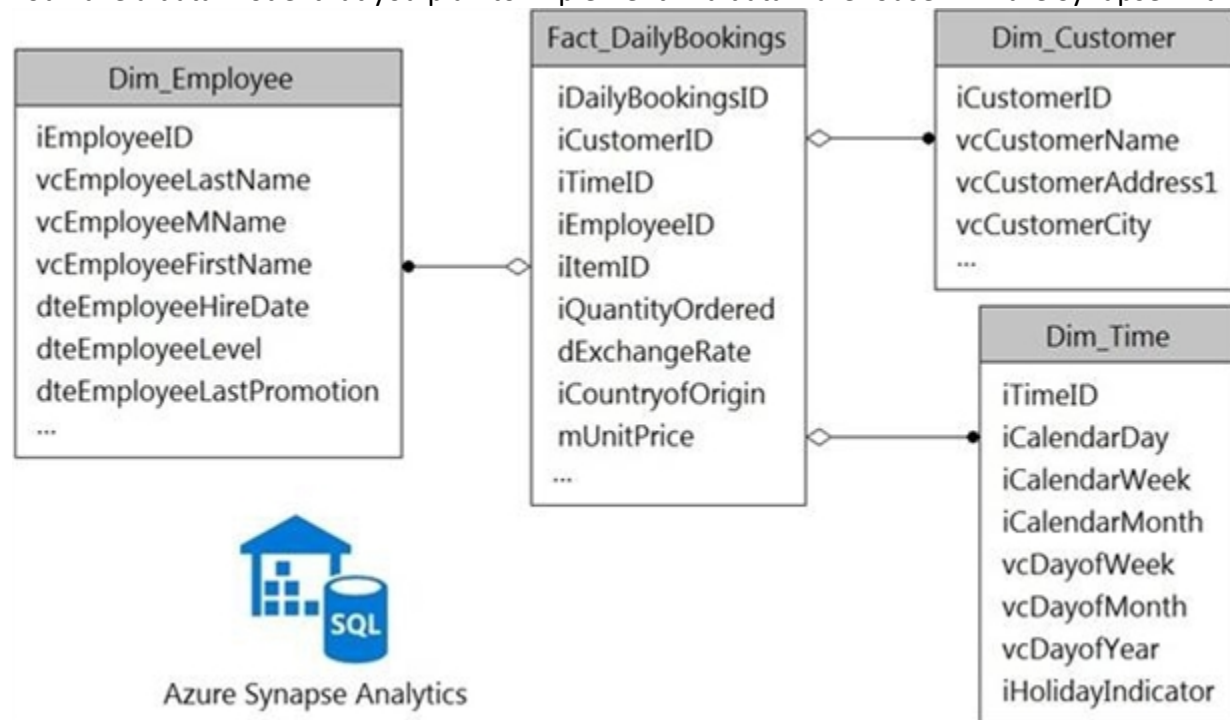
Reference:

<https://docs.microsoft.com/en-us/learn/modules/populate-slowly-changing-dimensions-azure-synapse-analytics-pipelines/3-choose-between-dimension-types>

**QUESTION 42**

**HOTSPOT**

You have a data model that you plan to implement in a data warehouse in Azure Synapse Analytics as shown in the following exhibit.



All the dimension tables will be less than 2 GB after compression, and the fact table will be approximately 6 TB. The dimension tables will be relatively static with very few data inserts and updates.

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

NOTE: Each correct selection is worth one point.

**Hot Area:**



**Answer Area**

Dim\_Customer:  ▼

Hash distributed
Round-robin
Replicated

Dim\_Employee:  ▼

Hash distributed
Round-robin
Replicated

Dim\_Time:  ▼

Hash distributed
Round-robin
Replicated

Fact\_DailyBookings:  ▼

Hash distributed
Round-robin
Replicated



Answer Area:



### Answer Area

Dim\_Customer:

▼
Hash distributed
Round-robin
Replicated

Dim\_Employee:

▼
Hash distributed
Round-robin
Replicated

Dim\_Time:

▼
Hash distributed
Round-robin
Replicated

Fact\_DailyBookings:

▼
Hash distributed
Round-robin
Replicated



**Section:**

**Explanation:**

Box 1: Replicated

Replicated tables are ideal for small star-schema dimension tables, because the fact table is often distributed on a column that is not compatible with the connected dimension tables. If this case applies to your schema, consider changing small dimension tables currently implemented as round-robin to replicated.

Box 2: Replicated

Box 3: Replicated

Box 4: Hash-distributed

For Fact tables use hash-distribution with clustered columnstore index. Performance improves when two hash tables are joined on the same distribution column.

Reference:

<https://azure.microsoft.com/en-us/updates/reduce-data-movement-and-make-your-queries-more-efficient-with-the-general-availability-of-replicated-tables/>

<https://azure.microsoft.com/en-us/blog/replicated-tables-now-generally-available-in-azure-sql-data-warehouse/>

**QUESTION 43**

HOTSPOT

You have an Azure Data Lake Storage Gen2 container.

Data is ingested into the container, and then transformed by a data integration application. The data is NOT modified after that. Users can read files in the container but cannot modify the files.

You need to design a data archiving solution that meets the following requirements:

New data is accessed frequently and must be available as quickly as possible. Data that is older than five years is accessed infrequently but must be available within one second when requested. Data that is older than seven years is NOT accessed. After seven years, the data must be persisted at the lowest cost possible. Costs must be minimized while maintaining the required availability.

How should you manage the data? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point

Hot Area:


**Answer Area**

Five-year-old data:

- Delete the blob.
- Move to archive storage.
- Move to cool storage.
- Move to hot storage.

Seven-year-old data:

- Delete the blob.
- Move to archive storage.
- Move to cool storage.
- Move to hot storage.



Answer Area:

### Answer Area

Five-year-old data:

▼
Delete the blob.
Move to archive storage.
Move to cool storage.
Move to hot storage.

Seven-year-old data:

▼
Delete the blob.
Move to archive storage.
Move to cool storage.
Move to hot storage.

#### Section:

#### Explanation:

HOTSPOT

You have an Azure Data Lake Storage Gen2 container.

Data is ingested into the container, and then transformed by a data integration application. The data is NOT modified after that. Users can read files in the container but cannot modify the files.

You need to design a data archiving solution that meets the following requirements:

New data is accessed frequently and must be available as quickly as possible. Data that is older than five years is accessed infrequently but must be available within one second when requested. Data that is older than seven years is NOT accessed. After seven years, the data must be persisted at the lowest cost possible. Costs must be minimized while maintaining the required availability.

How should you manage the data? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point

#### QUESTION 44

DRAG DROP

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

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

NOTE: Each correct selection is worth one point.

#### Select and Place:



**Values**

- CLUSTERED INDEX
- COLLATE
- DISTRIBUTION
- PARTITION
- PARTITION FUNCTION
- PARTITION SCHEME

**Answer Area**

```
CREATE TABLE table1
(
  ID INTEGER,
  col1 VARCHAR(10),
  col2 VARCHAR(10)
) WITH
(
  [ ] = HASH(ID),
  [ ] (ID RANGE LEFT FOR VALUES (1, 1000000, 2000000))
);
```

**Correct Answer:****Values**

- CLUSTERED INDEX
- COLLATE
- [ ]
- PARTITION FUNCTION
- PARTITION SCHEME

**Answer Area**

```
CREATE TABLE table1
(
  ID INTEGER,
  col1 VARCHAR(10),
  col2 VARCHAR(10)
) WITH
(
  DISTRIBUTION = HASH(ID),
  PARTITION (ID RANGE LEFT FOR VALUES (1, 1000000, 2000000))
);
```

**Section:****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?>**QUESTION 45****HOTSPOT**

You have an Azure Synapse Analytics dedicated SQL pool that contains the users shown in the following table.

Name	Role
User1	Server admin
User2	db_datereader

User1 executes a query on the database, and the query returns the results shown in the following exhibit.

```

1 SELECT c.name,
2     tbl.name as table_name,
3     typ.name as datatype,
4     c.is_masked,
5     c.masking_function
6 FROM sys.masked_columns AS c
7 INNER JOIN sys.tables AS tbl ON c.[object_id] = tbl.[object_id]
8 INNER JOIN sys.types typ ON c.user_type_id = typ.user_type_id
9 WHERE is_masked = 1;
10

```

## Results Messages

	name	table_name	datatype	is_masked	masking_function
1	BirthDate	DimCustomer	date	1	default()
2	Gender	DimCustomer	nvarchar	1	default()
3	EmailAddress	DimCustomer	nvarchar	1	email()
4	YearlyIncome	DimCustomer	money	1	default()

User1 is the only user who has access to the unmasked data.

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

NOTE: Each correct selection is worth one point.

Hot Area:



### Answer Area

When User2 queries the YearlyIncome column, the values returned will be **[answer choice]**.

	▼
a random number	
the values stored in the database	
XXXX	
0	

When User1 queries the BirthDate column, the values returned will be **[answer choice]**.

	▼
a random date	
the values stored in the database	
XXXX	
1900-01-01	

Answer Area:

### Answer Area

When User2 queries the YearlyIncome column, the values returned will be **[answer choice]**.

	▼
a random number	
the values stored in the database	
XXXX	
0	

When User1 queries the BirthDate column, the values returned will be **[answer choice]**.

	▼
a random date	
the values stored in the database	
XXXX	
1900-01-01	

#### Section:

#### Explanation:

Box 1: 0

The YearlyIncome column is of the money data type.

The Default masking function: Full masking according to the data types of the designated fields Use a zero value for numeric data types (bigint, bit, decimal, int, money, numeric, smallint, smallmoney, tinyint, float, real).

Box 2: the values stored in the database

Users with administrator privileges are always excluded from masking, and see the original data without any mask.

Reference:

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

#### QUESTION 46

#### DRAG DROP

You need to build a solution to ensure that users can query specific files in an Azure Data Lake Storage Gen2 account from an Azure Synapse Analytics serverless SQL pool.

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

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

#### Select and Place:

### Actions

- Create an external file format object
- Create an external data source
- Create a query that uses Create Table as Select
- Create a table
- Create an external table

### Answer Area



### Answer Area

- Create an external data source
- Create an external file format object
- Create an external table

Correct Answer:

### Actions

- 
- 
- Create a query that uses Create Table as Select
- Create a table
- 



### Section:

#### Explanation:

Step 1: Create an external data source

You can create external tables in Synapse SQL pools via the following steps:

CREATE EXTERNAL DATA SOURCE to reference an external Azure storage and specify the credential that should be used to access the storage. CREATE EXTERNAL FILE FORMAT to describe format of CSV or Parquet files. CREATE EXTERNAL TABLE on top of the files placed on the data source with the same file format.

Step 2: Create an external file format object

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

Step 3: Create an external table

Reference:

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

### Case 01 - Design and develop data processing

This is a case study. Case studies are not timed separately. You can use as much exam time as you would like to complete each case. However, there may be additional case studies and sections on this exam. You must manage your time to ensure that you are able to complete all questions included on this exam in the time provided.

To answer the questions included in a case study, you will need to reference information that is provided in the case study. Case studies might contain exhibits and other resources that provide more information about the scenario that is described in the case study. Each question is independent of the other questions in this case study.

At the end of this case study, a review screen will appear. This screen allows you to review your answers and to make changes before you move to the next section of the exam. After you begin a new section, you cannot return to this section.

To start the case study

To display the first question in this case study, click the Next button. Use the buttons in the left pane to explore the content of the case study before you answer the questions. Clicking these buttons displays information such as business



requirements, existing environment, and problem statements. If the case study has an All Information tab, note that the information displayed is identical to the information displayed on the subsequent tabs. When you are ready to answer a question, click the Question button to return to the question.

#### Overview

Litware, Inc. owns and operates 300 convenience stores across the US. The company sells a variety of packaged foods and drinks, as well as a variety of prepared foods, such as sandwiches and pizzas.

Litware has a loyalty club whereby members can get daily discounts on specific items by providing their membership number at checkout.

Litware employs business analysts who prefer to analyze data by using Microsoft Power BI, and data scientists who prefer analyzing data in Azure Databricks notebooks.

#### Requirements

##### Business Goals

Litware wants to create a new analytics environment in Azure to meet the following requirements:

See inventory levels across the stores. Data must be updated as close to real time as possible. Execute ad hoc analytical queries on historical data to identify whether the loyalty club discounts increase sales of the discounted products. Every four hours, notify store employees about how many prepared food items to produce based on historical demand from the sales data.

##### Technical Requirements

Litware identifies the following technical requirements:

Minimize the number of different Azure services needed to achieve the business goals. Use platform as a service (PaaS) offerings whenever possible and avoid having to provision virtual machines that must be managed by Litware. Ensure that the analytical data store is accessible only to the company's on-premises network and Azure services. Use Azure Active Directory (Azure AD) authentication whenever possible. Use the principle of least privilege when designing security. Stage Inventory data in Azure Data Lake Storage Gen2 before loading the data into the analytical data store. Litware wants to remove transient data from Data Lake Storage once the data is no longer in use. Files that have a modified date that is older than 14 days must be removed.

Limit the business analysts' access to customer contact information, such as phone numbers, because this type of data is not analytically relevant. Ensure that you can quickly restore a copy of the analytical data store within one hour in the event of corruption or accidental deletion.

##### Planned Environment

Litware plans to implement the following environment:

The application development team will create an Azure event hub to receive real-time sales data, including store number, date, time, product ID, customer loyalty number, price, and discount amount, from the point of sale (POS) system and output the data to data storage in Azure.

Customer data, including name, contact information, and loyalty number, comes from Salesforce, a SaaS application, and can be imported into Azure once every eight hours. Row modified dates are not trusted in the source table. Product data, including product ID, name, and category, comes from Salesforce and can be imported into Azure once every eight hours. Row modified dates are not trusted in the source table. Daily inventory data comes from a Microsoft SQL server located on a private network. Litware currently has 5 TB of historical sales data and 100 GB of customer data. The company expects approximately 100 GB of new data per month for the next year. Litware will build a custom application named FoodPrep to provide store employees with the calculation results of how many prepared food items to produce every four hours. Litware does not plan to implement Azure ExpressRoute or a VPN between the on-premises network and Azure.

#### QUESTION 1

##### HOTSPOT

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.

##### Hot Area:

**Answer Area**

Integration runtime type:  ▼

- Azure integration runtime
- Azure-SSIS integration runtime
- Self-hosted integration runtime

Trigger type:  ▼

- Event-based trigger
- Schedule trigger
- Tumbling window trigger

Activity type:  ▼

- Copy activity
- Lookup activity
- Stored procedure activity

Answer Area:



**Answer Area**

Integration runtime type:  ▼

- Azure integration runtime
- Azure-SSIS integration runtime
- Self-hosted integration runtime

Trigger type:  ▼

- Event-based trigger
- Schedule trigger
- Tumbling window trigger

Activity type:  ▼

- Copy activity
- Lookup activity
- Stored procedure activity

Section:

Explanation:

Explanation:

Box 1: Self-hosted integration runtime

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

Box 2: Schedule trigger

Schedule every 8 hours

Box 3: Copy activity

Scenario:

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

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

## Case 02 - Design and develop data processing

This is a case study. Case studies are not timed separately. You can use as much exam time as you would like to complete each case. However, there may be additional case studies and sections on this exam. You must manage your time to ensure that you are able to complete all questions included on this exam in the time provided.

To answer the questions included in a case study, you will need to reference information that is provided in the case study. Case studies might contain exhibits and other resources that provide more information about the scenario that is described in the case study. Each question is independent of the other questions in this case study.

At the end of this case study, a review screen will appear. This screen allows you to review your answers and to make changes before you move to the next section of the exam. After you begin a new section, you cannot return to this section.

To start the case study

To display the first question in this case study, click the Next button. Use the buttons in the left pane to explore the content of the case study before you answer the questions. Clicking these buttons displays information such as business requirements, existing environment, and problem statements. If the case study has an All Information tab, note that the information displayed is identical to the information displayed on the subsequent tabs. When you are ready to answer a question, click the Question button to return to the question.

Overview

Contoso, Ltd. is a clothing retailer based in Seattle. The company has 2,000 retail stores across the United States and an emerging online presence.

The network contains an Active Directory forest named contoso.com. The forest is integrated with an Azure Active Directory (Azure AD) tenant named contoso.com. Contoso has an Azure subscription associated to the contoso.com Azure AD tenant.

Existing Environment

Transactional Data

Contoso has three years of customer, transactional, operational, sourcing, and supplier data comprised of 10 billion records stored across multiple on-premises Microsoft SQL Server servers. The SQL Server instances contain data from various operational systems. The data is loaded into the instances by using SQL Server Integration Services (SSIS) packages.

You estimate that combining all product sales transactions into a company-wide sales transactions dataset will result in a single table that contains 5 billion rows, with one row per transaction.

Most queries targeting the sales transactions data will be used to identify which products were sold in retail stores and which products were sold online during different time periods. Sales transaction data that is older than three years will be removed monthly.

You plan to create a retail store table that will contain the address of each retail store. The table will be approximately 2 MB. Queries for retail store sales will include the retail store addresses.

You plan to create a promotional table that will contain a promotion ID. The promotion ID will be associated to a specific product. The product will be identified by a product ID. The table will be approximately 5 GB.

Streaming Twitter Data

The ecommerce department at Contoso develops an Azure logic app that captures trending Twitter feeds referencing the company's products and pushes the products to Azure Event Hubs.

Planned Changes and Requirements

Planned Changes

Contoso plans to implement the following changes:

Load the sales transaction dataset to Azure Synapse Analytics. Integrate on-premises data stores with Azure Synapse Analytics by using SSIS packages. Use Azure Synapse Analytics to analyze Twitter feeds to assess customer sentiments about products.

Sales Transaction Dataset Requirements

Contoso identifies the following requirements for the sales transaction dataset:

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. Ensure that queries joining and filtering sales transaction records

based on product ID complete as quickly as possible. Implement a surrogate key to account for changes to the retail store addresses. Ensure that data storage costs and performance are predictable. Minimize how long it takes to remove old records.

#### Customer Sentiment Analytics Requirements

Contoso identifies the following requirements for customer sentiment analytics:

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

Maximize the throughput of ingesting Twitter feeds from Event Hubs to Azure Storage without purchasing additional throughput or capacity units. Store Twitter feeds in Azure Storage by using Event Hubs Capture. The feeds will be converted into Parquet files. Ensure that the data store supports Azure AD-based access control down to the object level. Minimize administrative effort to maintain the Twitter feed data records. Purge Twitter feed data records that are older than two years.

#### Data Integration Requirements

Contoso identifies the following requirements for data integration:

Use an Azure service that leverages the existing SSIS packages to ingest on-premises data into datasets stored in a dedicated SQL pool of Azure Synapse Analytics and transform the data. Identify a process to ensure that changes to the ingestion and transformation activities can be version-controlled and developed independently by multiple data engineers.

### QUESTION 1

DRAG DROP

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

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

Select and Place:

#### Actions

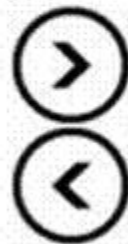
Merge changes

Create a pull request

Create a feature branch

Publish changes

Create a repository and a main branch



Answer Area  
Vdumps

Correct Answer:

## Actions




## Answer Area

Create a repository and a main branch

Create a feature branch

Create a pull request

Merge changes

Publish changes

### Section:

### Explanation:

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

Step 1: Create a repository and a main branch

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

Step 3: Create a pull request

Step 4: Merge changes

Merge feature branches into the main branch using pull requests.

Step 5: Publish changes

Reference:

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



### Case 01 - Monitor and optimize data storage and data processing

This is a case study. Case studies are not timed separately. You can use as much exam time as you would like to complete each case. However, there may be additional case studies and sections on this exam. You must manage your time to ensure that you are able to complete all questions included on this exam in the time provided.

To answer the questions included in a case study, you will need to reference information that is provided in the case study. Case studies might contain exhibits and other resources that provide more information about the scenario that is described in the case study. Each question is independent of the other questions in this case study.

At the end of this case study, a review screen will appear. This screen allows you to review your answers and to make changes before you move to the next section of the exam. After you begin a new section, you cannot return to this section.

To start the case study

To display the first question in this case study, click the Next button. Use the buttons in the left pane to explore the content of the case study before you answer the questions. Clicking these buttons displays information such as business requirements, existing environment, and problem statements. If the case study has an All Information tab, note that the information displayed is identical to the information displayed on the subsequent tabs. When you are ready to answer a question, click the Question button to return to the question.

Overview

Litware, Inc. owns and operates 300 convenience stores across the US. The company sells a variety of packaged foods and drinks, as well as a variety of prepared foods, such as sandwiches and pizzas.

Litware has a loyalty club whereby members can get daily discounts on specific items by providing their membership number at checkout.

Litware employs business analysts who prefer to analyze data by using Microsoft Power BI, and data scientists who prefer analyzing data in Azure Databricks notebooks.

Requirements

## Business Goals

Litware wants to create a new analytics environment in Azure to meet the following requirements:

See inventory levels across the stores. Data must be updated as close to real time as possible. Execute ad hoc analytical queries on historical data to identify whether the loyalty club discounts increase sales of the discounted products. Every four hours, notify store employees about how many prepared food items to produce based on historical demand from the sales data.

## Technical Requirements

Litware identifies the following technical requirements:

Minimize the number of different Azure services needed to achieve the business goals. Use platform as a service (PaaS) offerings whenever possible and avoid having to provision virtual machines that must be managed by Litware. Ensure that the analytical data store is accessible only to the company's on-premises network and Azure services. Use Azure Active Directory (Azure AD) authentication whenever possible. Use the principle of least privilege when designing security. Stage Inventory data in Azure Data Lake Storage Gen2 before loading the data into the analytical data store. Litware wants to remove transient data from Data Lake Storage once the data is no longer in use. Files that have a modified date that is older than 14 days must be removed.

Limit the business analysts' access to customer contact information, such as phone numbers, because this type of data is not analytically relevant. Ensure that you can quickly restore a copy of the analytical data store within one hour in the event of corruption or accidental deletion.

## Planned Environment

Litware plans to implement the following environment:

The application development team will create an Azure event hub to receive real-time sales data, including store number, date, time, product ID, customer loyalty number, price, and discount amount, from the point of sale (POS) system and output the data to data storage in Azure.

Customer data, including name, contact information, and loyalty number, comes from Salesforce, a SaaS application, and can be imported into Azure once every eight hours. Row modified dates are not trusted in the source table. Product data, including product ID, name, and category, comes from Salesforce and can be imported into Azure once every eight hours. Row modified dates are not trusted in the source table. Daily inventory data comes from a Microsoft SQL server located on a private network. Litware currently has 5 TB of historical sales data and 100 GB of customer data. The company expects approximately 100 GB of new data per month for the next year. Litware will build a custom application named FoodPrep to provide store employees with the calculation results of how many prepared food items to produce every four hours. Litware does not plan to implement Azure ExpressRoute or a VPN between the on-premises network and Azure.

### QUESTION 1

What should you do to improve high availability of the real-time data processing solution?



- A. Deploy a High Concurrency Databricks cluster.
- B. Deploy an Azure Stream Analytics job and use an Azure Automation runbook to check the status of the job and to start the job if it stops.
- C. Set Data Lake Storage to use geo-redundant storage (GRS).
- D. Deploy identical Azure Stream Analytics jobs to paired regions in Azure.

**Correct Answer: D**

#### Section:

#### Explanation:

Guarantee Stream Analytics job reliability during service updates

Part of being a fully managed service is the capability to introduce new service functionality and improvements at a rapid pace. As a result, Stream Analytics can have a service update deploy on a weekly (or more frequent) basis. No matter how much testing is done there is still a risk that an existing, running job may break due to the introduction of a bug. If you are running mission critical jobs, these risks need to be avoided. You can reduce this risk by following Azure's paired region model.

Scenario: The application development team will create an Azure event hub to receive real-time sales data, including store number, date, time, product ID, customer loyalty number, price, and discount amount, from the point of sale (POS) system and output the data to data storage in Azure

Reference:

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-job-reliability>

## 02 - Monitor and optimize data storage and data processing

### QUESTION 1

You have a SQL pool in Azure Synapse.

A user reports that queries against the pool take longer than expected to complete. You determine that the issue relates to queried columnstore segments. You need to add monitoring to the underlying storage to help

diagnose the issue. Which two metrics should you monitor? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

- A. Snapshot Storage Size
- B. Cache used percentage
- C. DWU Limit
- D. Cache hit percentage

**Correct Answer: B, D**

**Section:**

**Explanation:**

D: Cache hit percentage:  $(\text{cache hits} / \text{cache miss}) * 100$  where cache hits is the sum of all columnstore segments hits in the local SSD cache and cache miss is the columnstore segments misses in the local SSD cache summed across all nodes

B:  $(\text{cache used} / \text{cache capacity}) * 100$  where cache used is the sum of all bytes in the local SSD cache across all nodes and cache capacity is the sum of the storage capacity of the local SSD cache across all nodes Incorrect Asnwers:

C: DWU limit: Service level objective of the data warehouse.

Reference: <https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-concept-resource-utilization-query-activity>

### QUESTION 2

You manage an enterprise data warehouse in Azure Synapse Analytics. Users report slow performance when they run commonly used queries. Users do not report performance changes for infrequently used queries. You need to monitor resource utilization to determine the source of the performance issues. Which metric should you monitor?

- A. DWU percentage
- B. Cache hit percentage
- C. DWU limit
- D. Data IO percentage



**Correct Answer: B**

**Section:**

**Explanation:**

Monitor and troubleshoot slow query performance by determining whether your workload is optimally leveraging the adaptive cache for dedicated SQL pools.

Reference: <https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-how-to-monitor-cache>

### QUESTION 3

You have an Azure Databricks resource.

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

- A. clusters
- B. workspace
- C. DBFS
- D. SSH
- E. jobs

**Correct Answer: B**

**Section:**

**Explanation:**

Databricks provides access to audit logs of activities performed by Databricks users, allowing your enterprise to monitor detailed Databricks usage patterns. There are two types of logs:

Workspace-level audit logs with workspace-level events. Account-level audit logs with account-level events.

Reference: <https://docs.databricks.com/administration-guide/account-settings/audit-logs.html>

**QUESTION 4**

You are designing a highly available Azure Data Lake Storage solution that will include geo-zone-redundant storage (GZRS). You need to monitor for replication delays that can affect the recovery point objective (RPO). What should you include in the monitoring solution?

- A. 5xx: Server Error errors
- B. Average Success E2E Latency
- C. availability
- D. Last Sync Time

**Correct Answer: D**

**Section:**

**Explanation:**

Because geo-replication is asynchronous, it is possible that data written to the primary region has not yet been written to the secondary region at the time an outage occurs. The Last Sync Time property indicates the last time that data from the primary region was written successfully to the secondary region. All writes made to the primary region before the last sync time are available to be read from the secondary location. Writes made to the primary region after the last sync time property may or may not be available for reads yet.

Reference:

<https://docs.microsoft.com/en-us/azure/storage/common/last-sync-time-get>

**QUESTION 5**

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

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



**Correct Answer: B**

**Section:**

**QUESTION 6**

You implement an enterprise data warehouse in Azure Synapse Analytics.

You have a large fact table that is 10 terabytes (TB) in size.

Incoming queries use the primary key SaleKey column to retrieve data as displayed in the following table:

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

You need to distribute the large fact table across multiple nodes to optimize performance of the table.

Which technology should you use?

- A. hash distributed table with clustered index
- B. hash distributed table with clustered Columnstore index



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

**Correct Answer: B**

**Section:**

**Explanation:**

Hash-distributed tables improve query performance on large fact tables. Columnstore indexes can achieve up to 100x better performance on analytics and data warehousing workloads and up to 10x better data compression than traditional rowstore indexes. Incorrect Answers:

C, D: Round-robin tables are useful for improving loading speed.

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

#### QUESTION 7

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?

- A. nonclustered columnstore
- B. clustered columnstore
- C. nonclustered
- D. clustered

**Correct Answer: B**

**Section:**

**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>

#### QUESTION 8

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

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

**Correct Answer: C**

**Section:**

**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>

#### QUESTION 9

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

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

**Correct Answer: D**

**Section:**

**Explanation:**

Microsoft recommends use of sys.dm\_pdw\_nodes\_db\_partition\_stats to analyze any skewness in the data. Reference:

<https://docs.microsoft.com/en-us/sql/relational-databases/system-dynamic-management-views/sys-dm-db-partition-stats-transact-sql> <https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/cheat-sheet>

#### QUESTION 10

A company purchases IoT devices to monitor manufacturing machinery. The company uses an Azure IoT Hub to communicate with the IoT devices. The company must be able to monitor the devices in real-time. You need to design the solution.

What should you recommend?

- A. Azure Data Factory instance using Azure Portal
- B. Azure Data Factory instance using Azure PowerShell
- C. Azure Stream Analytics cloud job using Azure Portal
- D. Azure Data Factory instance using Microsoft Visual Studio

**Correct Answer: A**

**Section:**

**Explanation:**



#### QUESTION 11

HOTSPOT

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

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

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

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

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

NOTE: Each correct selection is worth one point.

**Hot Area:**

**Answer Area**

Number of partitions:

	▼
1	
8	
16	
32	

Partition key:

	▼
Fraud indicator	
Fraud score	
Individual line items	
Payment details	
Transaction ID	

Answer Area:



### Answer Area

Number of partitions:

	▼
1	
8	
16	
32	

Partition key:

	▼
Fraud indicator	
Fraud score	
Individual line items	
Payment details	
Transaction ID	

#### Section:

#### Explanation:

Box 1: 16

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

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

Box 2: Transaction ID

Reference:

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

#### QUESTION 12

You have an Azure data factory.

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

- A. the Activity log blade for the Data Factory resource
- B. the Monitor & Manage app in Data Factory
- C. the Resource health blade for the Data Factory resource
- D. Azure Monitor

**Correct Answer: D**

#### Section:

#### Explanation:

Data Factory stores pipeline-run data for only 45 days. Use Azure Monitor if you want to keep that data for a longer time.

Reference: <https://docs.microsoft.com/en-us/azure/data-factory/monitor-using-azure-monitor>

### QUESTION 13

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?

- A. Drop late arriving events from the job.
- B. Add an Azure Storage account to the job.
- C. Increase the streaming units for the job.
- D. Stop the job.

**Correct Answer: C**

**Section:**

**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 nonzero, 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>

### QUESTION 14

You are designing an Azure Databricks interactive cluster. The cluster will be used infrequently and will be configured for auto-termination. You need to use that the cluster configuration is retained indefinitely after the cluster is terminated. The solution must minimize costs. What should you do?

- A. Pin the cluster.
- B. Create an Azure runbook that starts the cluster every 90 days.
- C. Terminate the cluster manually when processing completes.
- D. Clone the cluster after it is terminated.



**Correct Answer: A**

**Section:**

**Explanation:**

Azure Databricks retains cluster configuration information for up to 70 all-purpose clusters terminated in the last 30 days and up to 30 job clusters recently terminated by the job scheduler. To keep an all-purpose cluster configuration even after it has been terminated for more than 30 days, an administrator can pin a cluster to the cluster list.

Reference:

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

### QUESTION 15

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

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

**Correct Answer: C**

**Section:****Explanation:**

The performance capacity of a query is determined by the user's resource class. Resource classes are pre-determined resource limits in Synapse SQL pool that govern compute resources and concurrency for query execution. Resource classes can help you configure resources for your queries by setting limits on the number of queries that run concurrently and on the compute-resources assigned to each query. There's a trade-off between memory and concurrency. Smaller resource classes reduce the maximum memory per query, but increase concurrency. Larger resource classes increase the maximum memory per query, but reduce concurrency.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/resource-classes-for-workload-management>

**QUESTION 16**

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

- A. Connect to the built-in pool and run DBCC PDW\_SHOWSPACEUSED.
- B. Connect to the built-in pool and run DBCC CHECKALLOC.
- C. Connect to Pool1 and query sys.dm\_pdw\_node\_status.
- D. Connect to Pool1 and query sys.dm\_pdw\_nodes\_db\_partition\_stats.

**Correct Answer: D**

**Section:****Explanation:**

Microsoft recommends use of sys.dm\_pdw\_nodes\_db\_partition\_stats to analyze any skewness in the data.

Reference:

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

**QUESTION 17**

You have a SQL pool in Azure Synapse.

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

Which dynamic management view should you query?

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

**Correct Answer: B**

**Section:****Explanation:**

You can use Dynamic Management Views (DMVs) to monitor your workload including investigating query execution in SQL pool. If your queries are failing or taking a long time to proceed, you can check and monitor if you have any transactions rolling back. Example:

```
-- Monitor rollback
```

```
SELECT  
SUM(CASE WHEN t.database_transaction_next_undo_lsn IS NOT NULL THEN 1 ELSE 0 END), t.pdw_node_id, nod.[type] FROM sys.dm_pdw_nodes_tran_database_transactions t  
JOIN sys.dm_pdw_nodes nod ON t.pdw_node_id = nod.pdw_node_id GROUP BY t.pdw_node_id, nod.[type]
```

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-manage-monitor#monitor-transaction-log-rollback>

**QUESTION 18**

You are monitoring an Azure Stream Analytics job.

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

What should you do?

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

**Correct Answer: B**

**Section:**

**Explanation:**

Backlogged Input Events: Number of input events that are backlogged. A non-zero value for this metric implies that your job isn't able to keep up with the number of incoming events. If this value is slowly increasing or consistently non-zero, you should scale out your job. You should increase the Streaming Units. Note: Streaming Units (SUs) represents the computing resources that are allocated to execute a Stream Analytics job. The higher the number of SUs, the more CPU and memory resources are allocated for your job.

Reference:

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

#### QUESTION 19

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

Table	Comment
EventDate	One million records are added to the table each day
EventTypeID	The table contains 10 million records for each event type.
WarehouseID	The table contains 100 million records for each warehouse.
ProductCategoryTypeID	The table contains 25 million records for each product category type.

You identify the following usage patterns:

Analysts will most commonly analyze transactions for a warehouse. Queries will summarize by product category type, date, and/or inventory event type.

You need to recommend a partition strategy for the table to minimize query times.

On which column should you partition the table?

- A. EventTypeID
- B. ProductCategoryTypeID
- C. EventDate
- D. WarehouseID

**Correct Answer: D**

**Section:**

**Explanation:**

The number of records for each warehouse is big enough for a good partitioning. Note: Table partitions enable you to divide your data into smaller groups of data. In most cases, table partitions are created on a date column. When creating partitions on clustered columnstore tables, it is important to consider how many rows belong to each partition. For optimal compression and performance of clustered columnstore tables, a minimum of 1 million rows per distribution and partition is needed. Before partitions are created, dedicated SQL pool already divides each table into 60 distributed databases.

#### QUESTION 20

You are designing a star schema for a dataset that contains records of online orders. Each record includes an order date, an order due date, and an order ship date. You need to ensure that the design provides the fastest query times of the records when querying for arbitrary date ranges and aggregating by fiscal calendar attributes. Which two actions should you perform? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

- A. Create a date dimension table that has a DateTime key.
- B. Use built-in SQL functions to extract date attributes.
- C. Create a date dimension table that has an integer key in the format of YYYYMMDD.

- D. In the fact table, use integer columns for the date fields.
- E. Use DateTime columns for the date fields.

**Correct Answer: C, D**

**Section:**

**Explanation:**

#### QUESTION 21

A company purchases IoT devices to monitor manufacturing machinery. The company uses an Azure IoT Hub to communicate with the IoT devices. The company must be able to monitor the devices in real-time. You need to design the solution.

What should you recommend?

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

**Correct Answer: D**

**Section:**

**Explanation:**

Stream Analytics is a cost-effective event processing engine that helps uncover real-time insights from devices, sensors, infrastructure, applications and data quickly and easily. Monitor and manage Stream Analytics resources with Azure PowerShell cmdlets and powershell scripting that execute basic Stream Analytics tasks. <https://cloudblogs.microsoft.com/sqlserver/2014/10/29/microsoft-adds-iot-streaming-analytics-data-production-and-workflow-services-to-azure/>

#### QUESTION 22

You configure monitoring from an Azure Synapse Analytics implementation. The implementation uses PolyBase to load data from comma-separated value (CSV) files stored in Azure Data Lake Storage Gen2 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?

- A. 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 file.'
- B. 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 outof total 1 rows processed.
- C. EXTERNAL TABLE access failed due to internal error: 'Java exception raised on call to HdfsBridge\_Connect: Error [Unable to instantiate LoginClass] occurred while accessing external file.'
- D. 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.'

**Correct Answer: B**

**Section:**

**Explanation:**

Error message: Cannot execute the query "Remote Query"

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.

Reference:

<https://docs.microsoft.com/en-us/sql/relational-databases/polybase/polybase-errors-and-possible-solutions>

#### QUESTION 23

You have an Azure Synapse Analytics dedicated SQL pool.



You run PDW\_SHOWSPACEUSED('dbo.FactInternetSales'); and get the results shown in the following table.

ROWS	RESERVED_SPACE	DATA_SPACE	INDEX_SPACE	UNUSED_SPACE	PDN_MODEL_ID	DISTRIBUTION_ID
694	2776	616	48	2112	1	1
407	2704	576	48	2080	1	2
53	2376	512	16	1848	1	3
58	2376	512	16	1848	1	4
168	2632	528	32	2072	1	5
195	2696	536	32	2128	1	6
5995	3464	1424	32	2008	1	7
0	2232	496	0	1736	1	8
264	2576	544	40	1992	1	9
3008	3016	960	32	2024	1	10
--	--	--	--	--	--	--
1550	2832	752	48	2032	1	50
1238	2832	696	40	2096	1	51
192	2632	528	32	2072	1	52
1127	2768	680	48	2040	1	53
1244	3032	704	64	2264	1	54
409	2632	568	32	2032	1	55
0	2232	496	0	1736	1	56
1437	2832	728	40	2064	1	57
0	2232	496	0	1736	1	58
384	2632	568	32	2040	1	59
225	2768	544	40	2184	1	60

Which statement accurately describes the dbo.FactInternetSales table?

- A. All distributions contain data.
- B. The table contains less than 10,000 rows.
- C. The table uses round-robin distribution.
- D. The table is skewed.



**Correct Answer: D**

**Section:**

**QUESTION 24**

You have two fact tables named Flight and Weather. Queries targeting the tables will be based on the join between the following columns.

Table	Column
Flight	ArrivalAirportID ArrivalDateTime
Weather	AirportID ReportDateTime

You need to recommend a solution that maximizes query performance.

What should you include in the recommendation?

- A. In the tables use a hash distribution of ArrivalDateTime and ReportDateTime.
- B. In the tables use a hash distribution of ArrivalAirportID and AirportID.
- C. In each table, create an IDENTITY column.
- D. In each table, create a column as a composite of the other two columns in the table.

**Correct Answer: B**

**Section:**

**Explanation:**

Hash-distribution improves query performance on large fact tables. Incorrect Answers:

A: Do not use a date column for hash distribution. All data for the same date lands in the same distribution. If several users are all filtering on the same date, then only 1 of the 60 distributions do all the processing work.

**QUESTION 25**

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

- A. Azure Synapse Analytics
- B. Azure HDInsight
- C. Azure Machine Learning
- D. Azure Data Factory
- E. Azure Databricks

**Correct Answer: B, D**

**Section:**

**QUESTION 26**

HOTSPOT

You have an on-premises data warehouse that includes the following fact tables. Both tables have the following columns: DateKey, ProductKey, RegionKey. There are 120 unique product keys and 65 unique region keys.

Table	Comments
Sales	The table is 600 GB in size. DateKey is used extensively in the WHERE clause in queries. ProductKey is used extensively in join operations. RegionKey is used for grouping. Severity-five percent of records relate to one of 40 regions.
Invoice	The table is 6 GB in size. DateKey and ProductKey are used extensively in the WHERE clause in queries. RegionKey is used for grouping.

Vdumps

Queries that use the data warehouse take a long time to complete.

You plan to migrate the solution to use Azure Synapse Analytics. You need to ensure that the Azure-based solution optimizes query performance and minimizes processing skew.

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

NOTE: Each correct selection is worth one point

**Hot Area:**

**Answer Area**

**Table**      **Distribution type**      **Distribution column**

Sales:

	<input type="text" value=""/>	<input type="text" value=""/>
	Hash-distributed	DateKey
	Round-robin	ProductKey
		RegionKey

Invoices:

	<input type="text" value=""/>	<input type="text" value=""/>
	Hash-distributed	DateKey
	Round-robin	ProductKey
		RegionKey



Answer Area:

### Answer Area

Table	Distribution type	Distribution column
-------	-------------------	---------------------

Sales:

<input type="text"/>	<input type="text"/>
Hash-distributed	DateKey
Round-robin	ProductKey
	RegionKey

Invoices:

<input type="text"/>	<input type="text"/>
Hash-distributed	DateKey
Round-robin	ProductKey
	RegionKey

vdumps

#### Section:

#### Explanation:

Box 1: Hash-distributed

Box 2: ProductKey

ProductKey is used extensively in joins.

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

Box 3: Hash-distributed

Box 4: RegionKey

Round-robin tables are useful for improving loading speed.

Consider using the round-robin distribution for your table in the following scenarios:

When getting started as a simple starting point since it is the default if there is no obvious joining key

If there is not good candidate column for hash distributing the table

If the table does not share a common join key with other tables

If the join is less significant than other joins in the query

When the table is a temporary staging table

Note: A distributed table appears as a single table, but the rows are actually stored across 60 distributions. The rows are distributed with a hash or round-robin algorithm.

Reference:

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

#### QUESTION 27

HOTSPOT

You need to collect application metrics, streaming query events, and application log messages for an Azure Databrick cluster. Which type of library and workspace should you implement? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

**Answer Area**

Library:  ▼

- Azure Databricks Monitoring Library
- Microsoft Azure Management Monitoring Library
- PyTorch
- TensorFlow

Workspace:  ▼

- Azure Databricks
- Azure Log Analytics
- Azure Machine Learning

Answer Area:

**Answer Area**

Library:  ▼

- Azure Databricks Monitoring Library
- Microsoft Azure Management Monitoring Library
- PyTorch
- TensorFlow

Workspace:  ▼

- Azure Databricks
- Azure Log Analytics
- Azure Machine Learning



**Section:**

**Explanation:**

You can send application logs and metrics from Azure Databricks to a Log Analytics workspace. It uses the Azure Databricks Monitoring Library, which is available on GitHub.

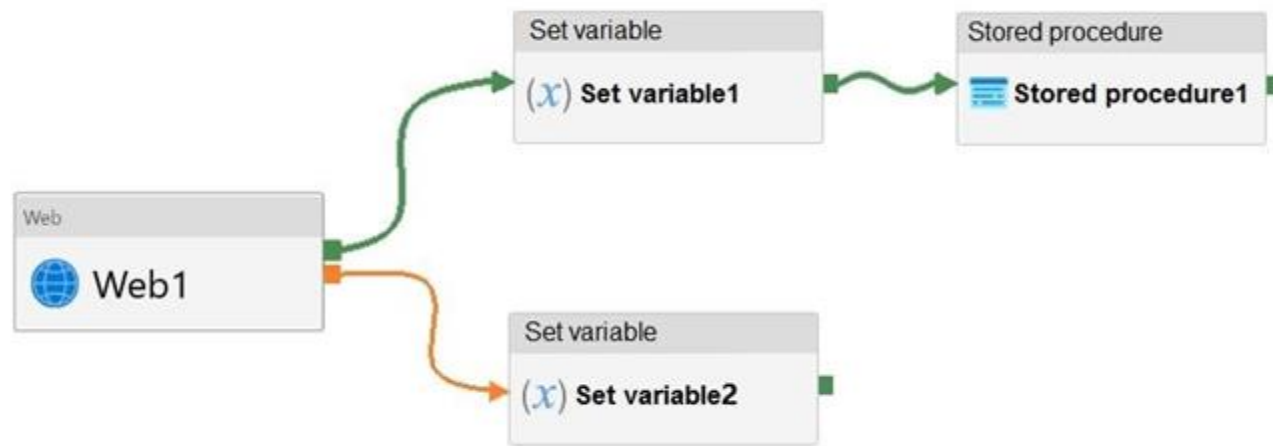
Reference:

<https://docs.microsoft.com/en-us/azure/architecture/databricks-monitoring/application-logs>

**QUESTION 28**

HOTSPOT

You have an Azure Data Factory pipeline that has the activities shown in the following exhibit.



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

Hot Area:

**Answer Area**

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

	▼
complete	
fail	
succeed	

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

	▼
Canceled	
Failed	
Succeeded	

Answer Area:

**Answer Area**

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

	▼
complete	
fail	
succeed	

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

	▼
Canceled	
Failed	
Succeeded	

Section:

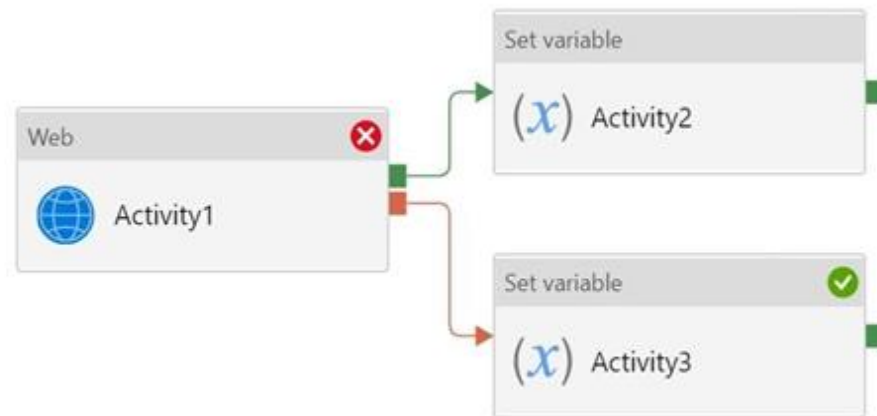
**Explanation:**

Box 1: succeed

Box 2: failed

Example:

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



Activity1 fails, Activity2 is skipped, and Activity3 succeeds. The pipeline reports failure.

Reference:

<https://datasavvy.me/2021/02/18/azure-data-factory-activity-failures-and-pipeline-outcomes/>

**Exam G**



**QUESTION 1**

You have an Azure subscription that contains an Azure SQL database named DB1 and a storage account named storage1. The storage1 account contains a file named File1.txt. File1.txt contains the names of selected tables in DB1. You need to use an Azure Synapse pipeline to copy data from the selected tables in DB1 to the files in storage1. The solution must meet the following requirements:

- The Copy activity in the pipeline must be parameterized to use the data in File1.txt to identify the source and destination of the copy.
- Copy activities must occur in parallel as often as possible. Which two pipeline activities should you include in the pipeline? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

- A. If Condition
- B. ForEach
- C. Lookup
- D. Get Metadata

**Correct Answer: A, D**

**Section:**

**Explanation:**

**QUESTION 2**

You have an Azure data factory that connects to a Microsoft Purview account. The data factory is registered in Microsoft Purview. You update a Data Factory pipeline. You need to ensure that the updated lineage is available in Microsoft Purview. What should you do first?

- A. Locate the related asset in the Microsoft Purview portal.
- B. Execute the pipeline.
- C. Disconnect the Microsoft Purview account from the data factory.

D. Execute an Azure DevOps build pipeline.

**Correct Answer: B**

**Section:**

**QUESTION 3**

HOTSPOT

You have an Azure subscription that contains an Azure Databricks workspace named databricks1 and an Azure Synapse Analytics workspace named synapse1. The synapse1 workspace contains an Apache Spark pool named pool1.

You need to share an Apache Hive catalog of pool1 with databricks1. What should you do? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

**Hot Area:**

From synapse1, create a linked service to:

Azure Cosmos DB
Azure Data Lake Storage Gen2
Azure SQL Database

Configure pool1 to use the linked service as:

An Azure Purview account
A Hive metastore
A managed Hive metastore service

**Answer Area:**

From synapse1, create a linked service to:

Azure Cosmos DB
Azure Data Lake Storage Gen2
Azure SQL Database

Configure pool1 to use the linked service as:

An Azure Purview account
A Hive metastore
A managed Hive metastore service

**Section:**

**Explanation:**

Box 1: Azure SQL Database

Use external Hive Metastore for Synapse Spark Pool

Azure Synapse Analytics allows Apache Spark pools in the same workspace to share a managed HMS (Hive Metastore) compatible metastore as their catalog.

Set up linked service to Hive Metastore

Follow below steps to set up a linked service to the external Hive Metastore in Synapse workspace. Open Synapse Studio, go to Manage > Linked services at left, click New to create a new linked service. Set up Hive Metastore linked service

Choose Azure SQL Database or Azure Database for MySQL based on your database type, click Continue. Provide Name of the linked service. Record the name of the linked service, this info will be used to configure Spark shortly. You can either select Azure SQL Database/Azure Database for MySQL for the external Hive Metastore from Azure subscription list, or enter the info manually. Provide User name and Password to set up the connection.

Test connection to verify the username and password.

Click Create to create the linked service.

Box 2: A Hive Metastore

Reference: <https://docs.microsoft.com/en-us/azure/synapse-analytics/spark/apache-spark-externalmetastore>

**QUESTION 4**

DRAG DROP

You have an Azure subscription that contains an Azure Data Lake Storage Gen2 account named storage1. Storage1 contains a container named container1. Container1 contains a directory named directory1. Directory1



contains a file named file1.

You have an Azure Active Directory (Azure AD) user named User1 that is assigned the Storage Blob Data Reader role for storage1. You need to ensure that User1 can append data to file1. The solution must use the principle of least privilege. Which permissions should you grant? To answer, drag the appropriate permissions to the correct resources. Each permission 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.

Select and Place:

Permissions	Answer Area
Read	container1: Permission
Write	directory1: Permission
Execute	file1: Permission

Correct Answer:

Permissions	Answer Area
Read	container1: Execute
Write	directory1: Execute
Execute	file1: Write



Section:

Explanation:

Box 1: Execute

If you are granting permissions by using only ACLs (no Azure RBAC), then to grant a security principal read or write access to a file, you'll need to give the security principal Execute permissions to the root folder of the container, and to each folder in the hierarchy of folders that lead to the file. Box 2: Execute

On Directory: Execute (X): Required to traverse the child items of a directory Box 3: Write

On file: Write (W): Can write or append to a file.

Reference:

<https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control>

#### QUESTION 5

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

One billion rows

A clustered columnstore index

A hash-distributed column named Product Key

A column named Sales Date that is of the date data type and cannot be null Thirty million rows will be added to Table1 each month. You need to partition Table1 based on the Sales Date column. The solution must optimize query performance and data loading. How often should you create a partition?

- A. once per month
- B. once per year

- C. once per day
- D. once per week

**Correct Answer: B**

**Section:**

**Explanation:**

Need a minimum 1 million rows per distribution. Each table is 60 distributions. 30 millions rows is added each month. Need 2 months to get a minimum of 1 million rows per distribution in a new partition. Note: When creating partitions on clustered columnstore tables, it is important to consider how many rows belong to each partition. For optimal compression and performance of clustered columnstore tables, a minimum of 1 million rows per distribution and partition is needed. Before partitions are created, dedicated SQL pool already divides each table into 60 distributions. Any partitioning added to a table is in addition to the distributions created behind the scenes. Using this example, if the sales fact table contained 36 monthly partitions, and given that a dedicated SQL pool has 60 distributions, then the sales fact table should contain 60 million rows per month, or 2.1 billion rows when all months are populated. If a table contains fewer than the recommended minimum number of rows per partition, consider using fewer partitions in order to increase the number of rows per partition. Reference:

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

#### QUESTION 6

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

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

**Correct Answer: A**

**Section:**

**Explanation:**

Convert nested JSON to a flattened DataFrame

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

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

```
display(DF.select($"id" as "main_id",$"name",$"batters",$"ppu",explode($"topping"))) // Exploding the topping column using explode as it is an array type
```

```
.withColumn("topping_id",$"col.id") // Extracting topping_id from col using DOT form .withColumn("topping_type",$"col.type") // Extracting topping_tytpe from col using DOT form .drop($"col")
```

```
.select($"*", $"batters.*") // Flattened the struct type batters tto array type which is batter .drop($"batters")
```

```
.select($"*",explode($"batter"))
```

```
.drop($"batter")
```

```
.withColumn("batter_id",$"col.id") // Extracting batter_id from col using DOT form .withColumn("battter_type",$"col.type") // Extracting battter_type from col using DOT form .drop($"col")
```

```
)
```

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

#### QUESTION 7

HOTSPOT

You have an Azure Synapse Analytics dedicated SQL pool named Pool1 that contains an external table named Sales. Sales contains sales data. Each row in Sales contains data on a single sale, including the name of the salesperson. You need to implement row-level security (RLS). The solution must ensure that the salespeople can access only their respective sales.

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

**Hot Area:**



Create: 

- A materialized view in Pool1
- A security policy for Sales
- Database scoped credentials in Pool1

Restrict row access by using: 

- A masking rule
- A table-valued function
- The CONTAINS predicate

**Answer Area:**

Create: 

- A materialized view in Pool1
- A security policy for Sales
- Database scoped credentials in Pool1

Restrict row access by using: 

- A masking rule
- A table-valued function
- The CONTAINS predicate



**Section:**

**Explanation:**

Box 1: A security policy for sale

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

Create a user-defined function that returns the name of the current user:

```
CREATE FUNCTION dbo.GetCurrentUser()
RETURNS NVARCHAR(128)
```

AS

BEGIN

RETURN SUSER\_SNAME();

END;

Create a security predicate function that filters the Sales table based on the current user:

```
CREATE FUNCTION dbo.SalesPredicate(@salesperson NVARCHAR(128)) RETURNS TABLE
WITH SCHEMABINDING
```

AS

RETURN SELECT 1 AS access\_result

WHERE @salesperson = SalespersonName;

Create a security policy on the Sales table that uses the SalesPredicate function to filter the data:

```
CREATE SECURITY POLICY SalesFilter
```

```
ADD FILTER PREDICATE dbo.SalesPredicate(dbo.GetCurrentUser()) ON dbo.Sales WITH (STATE = ON);
```

By creating a security policy for the Sales table, you ensure that each salesperson can only access their own sales data. The security policy uses a user-defined function to get the name of the current user and a security predicate function to filter the Sales table based on the current user. Box 2: table-value function

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

**QUESTION 8**

You plan to use an Apache Spark pool in Azure Synapse Analytics to load data to an Azure Data Lake Storage Gen2 account. You need to recommend which file format to use to store the data in the Data Lake Storage account. The solution must meet the following requirements:

- Column names and data types must be defined within the files loaded to the Data Lake Storage account.
- Data must be accessible by using queries from an Azure Synapse Analytics serverless SQL pool.
- Partition elimination must be supported without having to specify a specific partition. What should you recommend?

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

**Correct Answer: D**

**Section:**

**QUESTION 9**

**HOTSPOT**

You have two Azure SQL databases named DB1 and DB2.

DB1 contains a table named Table1. Table1 contains a timestamp column named LastModifiedOn. LastModifiedOn contains the timestamp of the most recent update for each individual row. DB2 contains a table named Watermark. Watermark contains a single timestamp column named WatermarkValue. You plan to create an Azure Data Factory pipeline that will incrementally upload into Azure Blob Storage all the rows in Table1 for which the LastModifiedOn column contains a timestamp newer than the most recent value of the WatermarkValue column in Watermark.

You need to identify which activities to include in the pipeline. The solution must meet the following requirements:

- Minimize the effort to author the pipeline.
- Ensure that the number of data integration units allocated to the upload operation can be controlled. What should you identify? To answer, select the appropriate options in the answer area.

**Hot Area:**

**Answer Area**

To retrieve the watermark value, use:

Lookup  
Filter  
Get Metadata  
Lookup

To perform the upload, use:

Copy data  
Copy data  
Custom  
Data flow

**Answer Area:**

Answer Area

To retrieve the watermark value, use:

Lookup  
Filter  
Get Metadata  
Lookup

To perform the upload, use:

Copy data  
Copy data  
Custom  
Data flow

Section:

Explanation:

QUESTION 10

HOTSPOT

You have an Azure Synapse serverless SQL pool.

You need to read JSON documents from a file by using the OPENROWSET function. How should you complete the query? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

```
SELECT *  
FROM OPENROWSET  
(  
    BULK  
    'https://sourcedatalake.blob.core.windows.net/public/docs.json',  
    FORMAT =  
    FIELDTERMINATOR = '0x0b',  
    FIELDQUOTE =  
    ROWTERMINATOR =  
    WITH (jsondoc nvarchar(1000)) as jsonDocuments
```

FORMAT =  
'JSON'  
'CSV'  
'DELTA'  
'JSON'  
'PARQUET'

FIELDQUOTE =  
'0x0b'  
'0x09'  
'0x0a'  
'0x0b'  
'0x0c'

Answer Area:

Answer Area

```
SELECT *
FROM OPENROWSET
(
    BULK
    'https://sourcedatalake.blob.core.windows.net/public/docs.json',
    FORMAT =
    'JSON',
    FIELDTERMINATOR = '0x0b',
    FIELDQUOTE = '0x0b',
    ROWTERMINATOR = '0x0a'
)
WITH (jsondoc nvarchar(1000))
```

Dropdown menu for FORMAT = with options: 'JSON', 'CSV', 'DELTA', 'JSON', 'PARQUET'. 'JSON' is selected.

Dropdown menu for FIELDQUOTE = with options: '0x0b', '0x09', '0x0a', '0x0b', '0x0c'. '0x0b' is selected.

Section:

Explanation:

QUESTION 11

You have an Azure data factory.

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

- A. the Activity log blade for the Data Factory resource
- B. the Monitor & Manage app in Data Factory
- C. the Resource health blade for the Data Factory resource
- D. Azure Monitor

Correct Answer: D

Section:

Explanation:

Data Factory stores pipeline-run data for only 45 days. Use Azure Monitor if you want to keep that data for a longer time.

Reference: <https://docs.microsoft.com/en-us/azure/data-factory/monitor-using-azure-monitor>

QUESTION 12

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?

- A. Drop late arriving events from the job.
- B. Add an Azure Storage account to the job.
- C. Increase the streaming units for the job.
- D. Stop the job.

**Correct Answer: C**

**Section:**

**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 nonzero, 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>

### QUESTION 13

You are designing an Azure Databricks interactive cluster. The cluster will be used infrequently and will be configured for auto-termination. You need to use that the cluster configuration is retained indefinitely after the cluster is terminated. The solution must minimize costs. What should you do?

- A. Pin the cluster.
- B. Create an Azure runbook that starts the cluster every 90 days.
- C. Terminate the cluster manually when processing completes.
- D. Clone the cluster after it is terminated.

**Correct Answer: A**

**Section:**

**Explanation:**

Azure Databricks retains cluster configuration information for up to 70 all-purpose clusters terminated in the last 30 days and up to 30 job clusters recently terminated by the job scheduler. To keep an all-purpose cluster configuration even after it has been terminated for more than 30 days, an administrator can pin a cluster to the cluster list.

Reference:

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

### QUESTION 14

You have an Azure data solution that contains an enterprise data warehouse in Azure Synapse Analytics named DW1. Several users execute ad hoc queries to DW1 concurrently.

You regularly perform automated data loads to DW1.

You need to ensure that the automated data loads have enough memory available to complete quickly and successfully when the adhoc queries run. What should you do?

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

**Correct Answer: C**

**Section:**

**Explanation:**

The performance capacity of a query is determined by the user's resource class. Resource classes are pre-determined resource limits in Synapse SQL pool that govern compute resources and concurrency for query execution. Resource classes can help you configure resources for your queries by setting limits on the number of queries that run concurrently and on the compute-resources assigned to each query. There's a trade-off between memory and concurrency. Smaller resource classes reduce the maximum memory per query, but increase concurrency. Larger resource classes increase the maximum memory per query, but reduce concurrency.

Reference:



<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/resource-classes-for-workload-management>

#### QUESTION 15

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

- A. Connect to the built-in pool and run DBCC PDW\_SHOWSPACEUSED.
- B. Connect to the built-in pool and run DBCC CHECKALLOC.
- C. Connect to Pool1 and query sys.dm\_pdw\_node\_status.
- D. Connect to Pool1 and query sys.dm\_pdw\_nodes\_db\_partition\_stats.

**Correct Answer: D**

**Section:**

**Explanation:**

Microsoft recommends use of sys.dm\_pdw\_nodes\_db\_partition\_stats to analyze any skewness in the data.

Reference:

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

#### QUESTION 16

You have a SQL pool in Azure Synapse.

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

Which dynamic management view should you query?

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

**Correct Answer: B**

**Section:**

**Explanation:**

You can use Dynamic Management Views (DMVs) to monitor your workload including investigating query execution in SQL pool. If your queries are failing or taking a long time to proceed, you can check and monitor if you have any transactions rolling back. Example:

```
-- Monitor rollback
```

```
SELECT  
SUM(CASE WHEN t.database_transaction_next_undo_lsn IS NOT NULL THEN 1 ELSE 0 END), t.pdw_node_id, nod.[type] FROM sys.dm_pdw_nodes_tran_database_transactions t  
JOIN sys.dm_pdw_nodes nod ON t.pdw_node_id = nod.pdw_node_id GROUP BY t.pdw_node_id, nod.[type]
```

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-manage-monitor#monitor-transaction-log-rollback>

#### QUESTION 17

You are monitoring an Azure Stream Analytics job.

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

What should you do?

- A. Change the compatibility level of the Stream Analytics job.
- B. Increase the number of streaming units (SUs).
- C. Remove any named consumer groups from the connection and use \$default.





D. Create an additional output stream for the existing input stream.

**Correct Answer: B**

**Section:**

**Explanation:**

Backlogged Input Events: Number of input events that are backlogged. A non-zero value for this metric implies that your job isn't able to keep up with the number of incoming events. If this value is slowly increasing or consistently non-zero, you should scale out your job. You should increase the Streaming Units. Note: Streaming Units (SUs) represents the computing resources that are allocated to execute a Stream Analytics job. The higher the number of SUs, the more CPU and memory resources are allocated for your job.

Reference:

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

#### QUESTION 18

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

Table	Comment
EventDate	One million records are added to the table each day
EventTypeID	The table contains 10 million records for each event type.
WarehouseID	The table contains 100 million records for each warehouse.
ProductCategoryTypeID	The table contains 25 million records for each product category type.

You identify the following usage patterns:

Analysts will most commonly analyze transactions for a warehouse. Queries will summarize by product category type, date, and/or inventory event type.

You need to recommend a partition strategy for the table to minimize query times.

On which column should you partition the table?

- A. EventTypeID
- B. ProductCategoryTypeID
- C. EventDate
- D. WarehouseID



**Correct Answer: D**

**Section:**

**Explanation:**

The number of records for each warehouse is big enough for a good partitioning. Note: Table partitions enable you to divide your data into smaller groups of data. In most cases, table partitions are created on a date column. When creating partitions on clustered columnstore tables, it is important to consider how many rows belong to each partition. For optimal compression and performance of clustered columnstore tables, a minimum of 1 million rows per distribution and partition is needed. Before partitions are created, dedicated SQL pool already divides each table into 60 distributed databases.

#### QUESTION 19

You are designing a star schema for a dataset that contains records of online orders. Each record includes an order date, an order due date, and an order ship date. You need to ensure that the design provides the fastest query times of the records when querying for arbitrary date ranges and aggregating by fiscal calendar attributes. Which two actions should you perform? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

- A. Create a date dimension table that has a DateTime key.
- B. Use built-in SQL functions to extract date attributes.
- C. Create a date dimension table that has an integer key in the format of YYYYMMDD.
- D. In the fact table, use integer columns for the date fields.
- E. Use DateTime columns for the date fields.

**Correct Answer: C, D**

**Section:**

**Explanation:**

#### QUESTION 20

A company purchases IoT devices to monitor manufacturing machinery. The company uses an Azure IoT Hub to communicate with the IoT devices. The company must be able to monitor the devices in real-time. You need to design the solution.

What should you recommend?

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

**Correct Answer: D**

**Section:**

**Explanation:**

Stream Analytics is a cost-effective event processing engine that helps uncover real-time insights from devices, sensors, infrastructure, applications and data quickly and easily. Monitor and manage Stream Analytics resources with Azure PowerShell cmdlets and powershell scripting that execute basic Stream Analytics tasks. <https://cloudblogs.microsoft.com/sqlserver/2014/10/29/microsoft-adds-iot-streaming-analytics-data-production-and-workflow-services-to-azure/>

#### QUESTION 21

You configure monitoring from an Azure Synapse Analytics implementation. The implementation uses PolyBase to load data from comma-separated value (CSV) files stored in Azure Data Lake Storage Gen2 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?

- A. 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 file.'
- B. 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 outof total 1 rows processed.
- C. EXTERNAL TABLE access failed due to internal error: 'Java exception raised on call to HdfsBridge\_Connect: Error [Unable to instantiate LoginClass] occurred while accessing external file.'
- D. 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.'

**Correct Answer: B**

**Section:**

**Explanation:**

Error message: Cannot execute the query "Remote Query"

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.

Reference:

<https://docs.microsoft.com/en-us/sql/relational-databases/polybase/polybase-errors-and-possible-solutions>

#### QUESTION 22

You are designing an Azure Synapse Analytics workspace.

You need to recommend a solution to provide double encryption of all the data at rest. Which two components should you include in the recommendation? Each coned answer presents part of the solution NOTE: Each correct selection is worth one point.

- A. an X509 certificate
- B. an RSA key
- C. an Azure key vault that has purge protection enabled

- D. an Azure virtual network that has a network security group (NSG)
- E. an Azure Policy initiative

**Correct Answer: A, D**

**Section:**

**Explanation:**

#### QUESTION 23

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

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

**Correct Answer: C**

**Section:**

**Explanation:**

Apache Spark Structured Streaming is a fast, scalable, and fault-tolerant stream processing API. You can use it to perform analytics on your streaming data in near real-time. With Structured Streaming, you can use SQL queries to process streaming data in the same way that you would process static data.

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

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

Reference:

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

#### QUESTION 24

You have an Azure Synapse Analytics dedicated SQL pool named pool1. You plan to implement a star schema in pool1 and create a new table named DimCustomer by using the following code.

```
CREATE TABLE dbo.[DimCustomer](
    [CustomerKey] int NOT NULL,
    [CustomerSourceID] [int] NOT NULL,
    [Title] [nvarchar](8) NULL,
    [FirstName] [nvarchar](50) NOT NULL,
    [MiddleName] [nvarchar](50) NULL,
    [LastName] [nvarchar](50) NOT NULL,
    [Suffix] [nvarchar](10) NULL,
    [CompanyName] [nvarchar](128) NULL,
    [SalesPerson] [nvarchar](256) NULL,
    [EmailAddress] [nvarchar](50) NULL,
    [Phone] [nvarchar](25) NULL,
    [InsertedDate] [datetime] NOT NULL,
    [ModifiedDate] [datetime] NOT NULL,
    [HashKey] [varchar](100) NOT NULL,
    [IsCurrentRow] [bit] NOT NULL
)
WITH
(
    DISTRIBUTION = REPLICATE,
    CLUSTERED COLUMNSTORE INDEX
);
GO
```

You need to ensure that DimCustomer has the necessary columns to support a Type 2 slowly changing dimension (SCD). Which two columns should you add? Each correct answer presents part of the solution. NOTE: Each correct selection is

worth one point.

- A. [HistoricalSalesPerson] [nvarchar] (256) NOT NULL
- B. [EffectiveEndDate] [datetime] NOT NULL
- C. [PreviousModifiedDate] [datetime] NOT NULL
- D. [RowID] [bigint] NOT NULL
- E. [EffectiveStartDate] [datetime] NOT NULL

**Correct Answer: A, B**

**Section:**

#### QUESTION 25

You have an Azure Data Factory pipeline named Pipeline1!. Pipelinel contains a copy activity that sends data to an Azure Data Lake Storage Gen2 account. Pipeline 1 is executed by a schedule trigger. You change the copy activity sink to a new storage account and merge the changes into the collaboration branch. After Pipelinel executes, you discover that data is NOT copied to the new storage account. You need to ensure that the data is copied to the new storage account. What should you do?

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

**Correct Answer: A**

**Section:**

**Explanation:**

CI/CD lifecycle

A development data factory is created and configured with Azure Repos Git. All developers should have permission to author Data Factory resources like pipelines and datasets. A developer creates a feature branch to make a change. They debug their pipeline runs with their most recent changes After a developer is satisfied with their changes, they create a pull request from their feature branch to the main or collaboration branch to get their changes reviewed by peers.

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

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

#### QUESTION 26

You have an Azure Databricks workspace that contains a Delta Lake dimension table named Table1. Table1 is a Type 2 slowly changing dimension (SCD) table. You need to apply updates from a source table to Table1. Which Apache Spark SQL operation should you use?

- A. CREATE
- B. UPDATE
- C. MERGE
- D. ALTER

**Correct Answer: C**

**Section:**

**Explanation:**

The Delta provides the ability to infer the schema for data input which further reduces the effort required in managing the schema changes. The Slowly Changing Data(SCD) Type 2 records all the changes made to each key in the dimensional table. These operations require updating the existing rows to mark the previous values of the keys as old and then inserting new rows as the latest values. Also, Given a source table with the updates and the target table with dimensional data, SCD Type 2 can be expressed with the merge.

Example:

```
// Implementing SCD Type 2 operation using merge function
```



```

customersTable
.as("customers")
.merge(
stagedUpdates.as("staged_updates"),
"customers.customerId = mergeKey")
.whenMatched("customers.current = true AND customers.address <> staged_updates.address") .updateExpr(Map(
"current" -> "false",
"endDate" -> "staged_updates.effectiveDate"))
.whenNotMatched()
.insertExpr(Map(
"customerid" -> "staged_updates.customerId",
"address" -> "staged_updates.address",
"current" -> "true",
"effectiveDate" -> "staged_updates.effectiveDate",
"endDate" -> "null"))
.execute()
}

```

Reference:

<https://www.projectpro.io/recipes/what-is-slowly-changing-data-scd-type-2-operation-delta-tabledatabricks>

#### QUESTION 27

You have an Azure Synapse Analytics dedicated SQL pool named Pool1 that contains a table named Sales. Sales has row-level security (RLS) applied. RLS uses the following predicate filter.

```

CREATE FUNCTION Security.fn_securitypredicate(@SalesRep AS sysname)
RETURNS TABLE
WITH SCHEMABINDING
AS
RETURN SELECT 1 AS fn_securitypredicate_result
WHERE @SalesRep = USER_NAME() OR USER_NAME() = 'Manager';

```

A user named SalesUser1 is assigned the db\_datareader role for Pool1.

A user named SalesUser1 is assigned the db\_datareader role for Pool1. Which rows in the Sales table are returned when SalesUser1 queries the table?

- A. only the rows for which the value in the User\_Name column is SalesUser1
- B. all the rows
- C. only the rows for which the value in the SalesRep column is Manager
- D. only the rows for which the value in the SalesRep column is SalesUser1

**Correct Answer: A**

**Section:**

#### QUESTION 28

You are designing a dimension table in an Azure Synapse Analytics dedicated SQL pool. You need to create a surrogate key for the table. The solution must provide the fastest query performance. What should you use for the surrogate key?

- A. a GUID column
- B. a sequence object
- C. an IDENTITY column

**Correct Answer: C**

**Section:**

**Explanation:**



Use IDENTITY to create surrogate keys using dedicated SQL pool in Azure Synapse Analytics. Note: 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-datawarehouse-tables-identity>

#### QUESTION 29

You are designing an Azure Data Lake Storage solution that will transform raw JSON files for use in an analytical workload. You need to recommend a format for the transformed files. The solution must meet the following requirements:

Contain information about the data types of each column in the files. Support querying a subset of columns in the files.

Support read-heavy analytical workloads.

Minimize the file size.

What should you recommend?

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

**Correct Answer: D**

**Section:**

**Explanation:**

Parquet, an open-source file format for Hadoop, stores nested data structures in a flat columnar format. Compared to a traditional approach where data is stored in a row-oriented approach, Parquet file format is more efficient in terms of storage and performance. It is especially good for queries that read particular columns from a “wide” (with many columns) table since only needed columns are read, and IO is minimized.

Reference: <https://www.clairvoyant.ai/blog/big-data-file-formats>

#### QUESTION 30

HOTSPOT

You have an Azure subscription that contains an Azure Synapse Analytics dedicated SQL pool. You plan to deploy a solution that will analyze sales data and include the following:

- A table named Country that will contain 195 rows
- A table named Sales that will contain 100 million rows
- A query to identify total sales by country and customer from the past 30 days You need to create the tables. The solution must maximize query performance. How should you complete the script? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

**Hot Area:**



Answer Area

```
CREATE TABLE [dbo].[Sales]
(
    [OrderDate] date NOT NULL
,   [CustomerId] int NOT NULL
,   [CountryId] int NOT NULL
,   [Total] money NOT NULL
)
WITH
(
    DISTRIBUTION = HASH([CustomerId])
    CLUSTERED COLUMNSTORE INDEX
)
CREATE TABLE [dbo].[Country]
```

Answer Area:

Answer Area

```
CREATE TABLE [dbo].[Sales]
(
    [OrderDate] date NOT NULL
,   [CustomerId] int NOT NULL
,   [CountryId] int NOT NULL
,   [Total] money NOT NULL
)
WITH
(
    DISTRIBUTION = HASH([CustomerId])
    CLUSTERED COLUMNSTORE INDEX
)
CREATE TABLE [dbo].[Country]
```



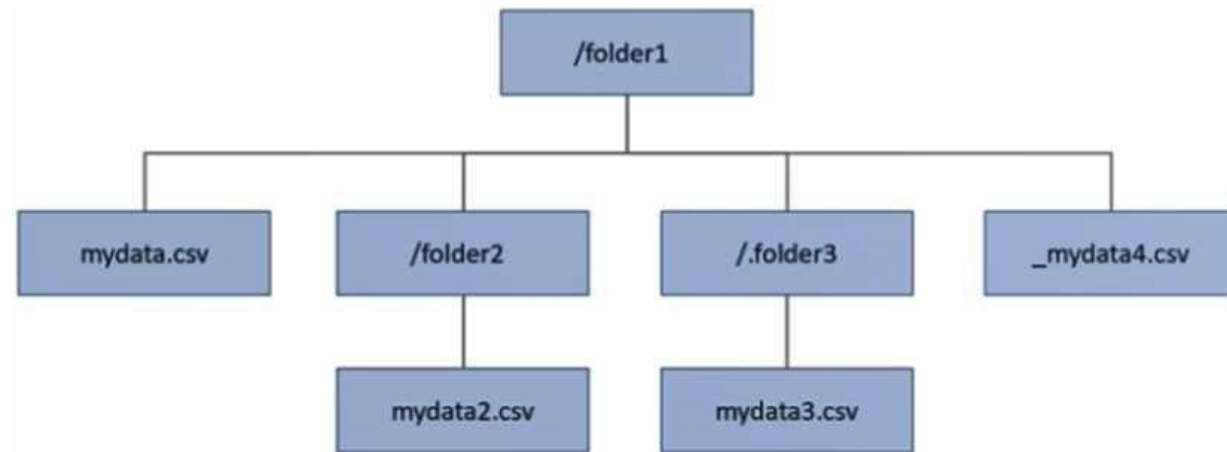
Section:

Explanation:

QUESTION 31

HOTSPOT

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



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

```

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

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

**Hot Area:**

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

**Answer Area:**

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

**Section:**

**Explanation:**

Box 1: Yes

In the serverless SQL pool you can also use recursive wildcards /logs/\*\* to reference Parquet or CSV files in any sub-folder beneath the referenced folder.

Box 2: Yes

Box 3: No

Reference: <https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-externaltables>

**QUESTION 32**

**HOTSPOT**

You have an Azure Synapse Analytics dedicated SQL pool.

You need to create a table named FactInternetSales that will be a large fact table in a dimensional model. FactInternetSales will contain 100 million rows and two columns named SalesAmount and OrderQuantity. Queries executed on FactInternetSales will aggregate the values in SalesAmount and OrderQuantity from the last year for a specific product. The solution must minimize the data size and query execution time. How should you



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

Hot Area:

Answer Area

```
CREATE TABLE [dbo].[FactInternetSales]
(
  [ProductKey] int NOT NULL
  , [OrderDateKey] int NOT NULL
  , [CustomerKey] int NOT NULL
  , [PromotionKey] int NOT NULL
  , [SalesOrderNumber] nvarchar(20) NOT NULL
  , [OrderQuantity] smallint NOT NULL
  , [UnitPrice] money NOT NULL
  , [SalesAmount] money NOT NULL
)
```

WITH

( CLUSTERED COLUMNSTORE INDEX  
( CLUSTERED INDEX ([OrderDateKey])  
( HEAP  
( INDEX on [ProductKey]

, DISTRIBUTION =  
);

Hash([OrderDateKey])  
Hash([ProductKey])  
REPLICATE  
ROUND\_ROBIN

Answer Area:

Answer Area

```
CREATE TABLE [dbo].[FactInternetSales]
(
  [ProductKey] int NOT NULL
, [OrderDateKey] int NOT NULL
, [CustomerKey] int NOT NULL
, [PromotionKey] int NOT NULL
, [SalesOrderNumber] nvarchar(20) NOT NULL
, [OrderQuantity] smallint NOT NULL
, [UnitPrice] money NOT NULL
, [SalesAmount] money NOT NULL
)
WITH
(
  ( CLUSTERED COLUMNSTORE INDEX
  ( CLUSTERED INDEX ([OrderDateKey])
  ( HEAP
  ( INDEX on [ProductKey]
, DISTRIBUTION =
);
```

( CLUSTERED COLUMNSTORE INDEX  
( CLUSTERED INDEX ([OrderDateKey])  
( HEAP  
( INDEX on [ProductKey]

Hash([OrderDateKey])  
Hash([ProductKey])  
REPLICATE  
ROUND\_ROBIN



Section:

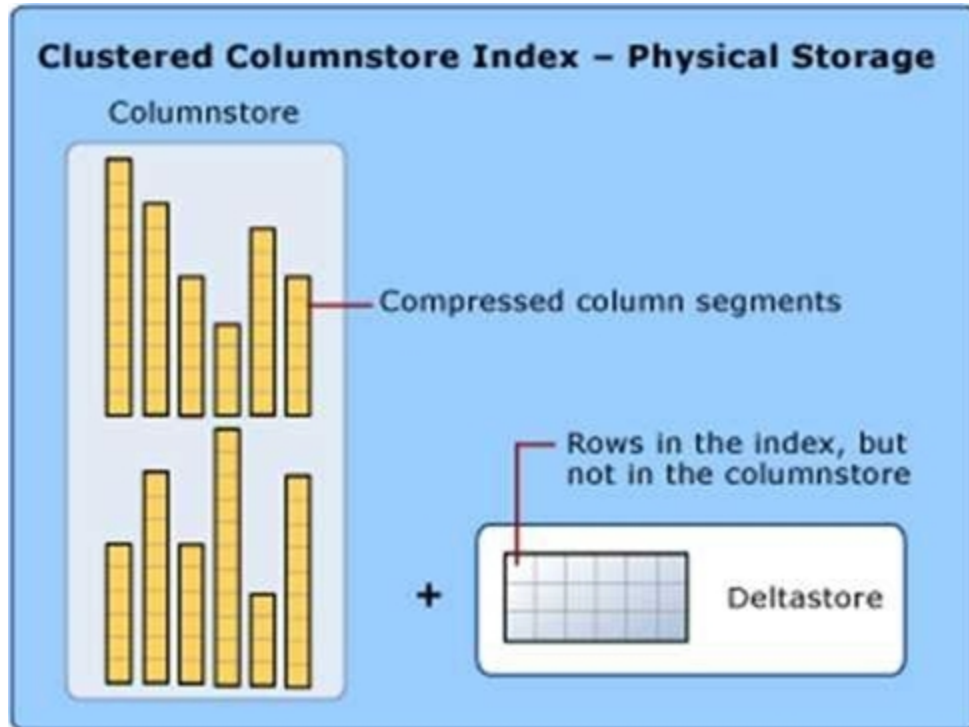
Explanation:

Box 1: (CLUSTERED COLUMNSTORE INDEX  
CLUSTERED COLUMNSTORE INDEX

Columnstore indexes are the standard for storing and querying large data warehousing fact tables. This index uses column-based data storage and query processing to achieve gains up to 10 times the query performance in your data warehouse over traditional row-oriented storage. You can also achieve gains up to 10 times the data compression over the uncompressed data size. Beginning with SQL Server 2016 (13.x) SP1, columnstore indexes enable operational analytics: the ability to run performant real-time analytics on a transactional workload.

Note: Clustered columnstore index

A clustered columnstore index is the physical storage for the entire table.



To reduce fragmentation of the column segments and improve performance, the columnstore index might store some data temporarily into a clustered index called a deltastore and a B-tree list of IDs for deleted rows. The deltastore operations are handled behind the scenes. To return the correct query results, the clustered columnstore index combines query results from both the columnstore and the deltastore.

Box 2: HASH([ProductKey])

A hash distributed table distributes rows based on the value in the distribution column. A hash distributed table is designed to achieve high performance for queries on large tables. Choose a distribution column with data that distributes evenly  
Reference:

<https://docs.microsoft.com/en-us/sql/relational-databases/indexes/columnstore-indexes-overview>

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

### QUESTION 33

#### DRAG DROP

You have an Azure subscription.

You plan to build a data warehouse in an Azure Synapse Analytics dedicated SQL pool named pool1 that will contain staging tables and a dimensional model. Pool1 will contain the following tables. You need to design the table storage for pool1. The solution must meet the following requirements:

Maximize the performance of data loading operations to Staging.WebSessions. Minimize query times for reporting queries against the dimensional model. Which type of table distribution should you use for each table? To answer, drag the appropriate table distribution types to the correct tables.

Name	Number of rows	Update frequency	Description
Common.Date	7,300	New rows inserted yearly	<ul style="list-style-type: none"> <li>Contains one row per date for the last 20 years</li> <li>Contains columns named Year, Month, Quarter, and IsWeekend</li> </ul>
Marketing.WebSessions	1,500,500,000	Hourly inserts and updates	Fact table that contains counts of and updates sessions and page views, including foreign key values for date, channel, device, and medium
Staging.WebSessions	300,000	Hourly truncation and inserts	Staging table for web session data, truncation and including descriptive fields for inserts channel, device, and medium

Each table distribution type 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.

Select and Place:

Table distribution types	Answer Area
Hash	Common.Data: <input type="text"/>
Replicated	Marketing.Web.Sessions: <input type="text"/>
Round-robin	Staging. Web.Sessions: <input type="text"/>

Correct Answer:

Table distribution types	Answer Area
<input type="text"/>	Common.Data: <input type="text" value="Replicated"/>
<input type="text"/>	Marketing.Web.Sessions: <input type="text" value="Hash"/>
<input type="text"/>	Staging. Web.Sessions: <input type="text" value="Round-robin"/>

Section:

Explanation:

Box 1: Replicated

The best table storage option for a small table is to replicate it across all the Compute nodes. Box 2: Hash

Hash-distribution improves query performance on large fact tables. Box 3: Round-robin

Round-robin distribution is useful for improving loading speed.

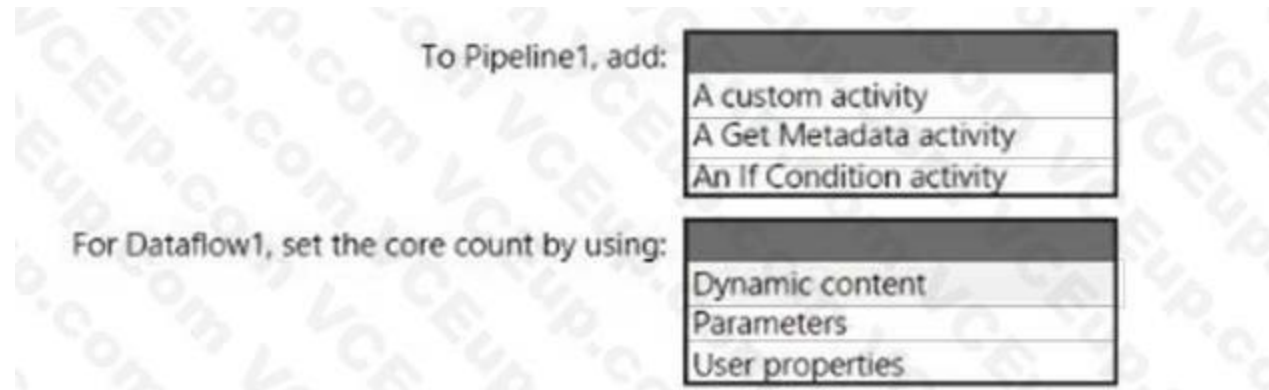
Reference: <https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-datawarehouse-tables-distribute>

QUESTION 34

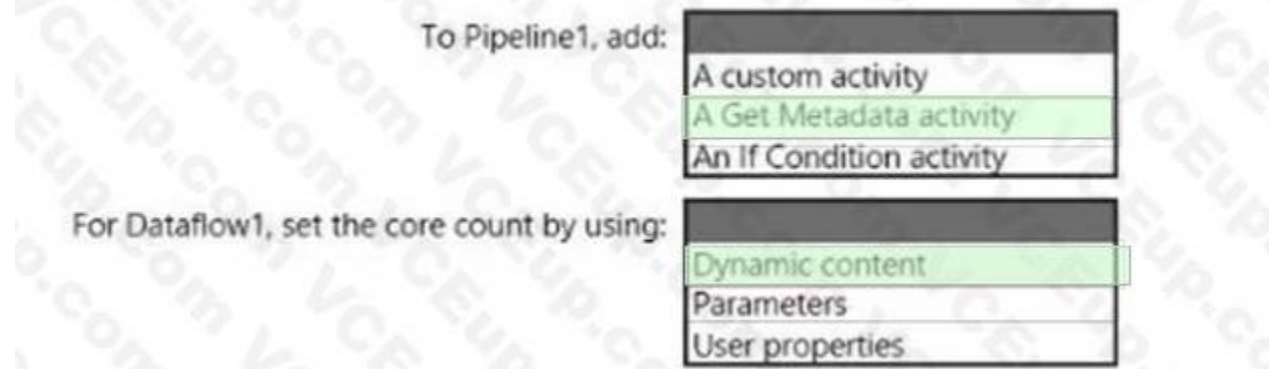
HOTSPOT

You have an Azure Synapse Analytics pipeline named Pipeline1 that contains a data flow activity named Dataflow1. Pipeline1 retrieves files from an Azure Data Lake Storage Gen 2 account named storage1. Dataflow1 uses the AutoResolveIntegrationRuntime integration runtime configured with a core count of 128. You need to optimize the number of cores used by Dataflow1 to accommodate the size of the files in storage1. What should you configure? To answer, select the appropriate options in the answer area.

Hot Area:



**Answer Area:**



**Section:**

**Explanation:**

Box 1: A Get Metadata activity

Dynamically size data flow compute at runtime

The Core Count and Compute Type properties can be set dynamically to adjust to the size of your incoming source data at runtime. Use pipeline activities like Lookup or Get Metadata in order to find the size of the source dataset dat

a. Then, use Add Dynamic Content in the Data Flow activity properties. Box 2: Dynamic content

Reference: <https://docs.microsoft.com/en-us/azure/data-factory/control-flow-execute-data-flowactivity>

**QUESTION 35**

**HOTSPOT**

You have an Azure subscription that contains an Azure Synapse Analytics dedicated SQL pool named Pool1 and an Azure Data Lake Storage account named storage1. Storage1 requires secure transfers. You need to create an external data source in Pool1 that will be used to read .orc files in storage1. How should you complete the code? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

**Hot Area:**



**Answer Area**

```
CREATE EXTERNAL DATA SOURCE AzureDataLakeStore
WITH
( Location1 'abfs://data@newyorktaxidataset.dfs.core.windows.net' ,
credential = ADLS_credential ,
TYPE -
);
```

**Answer Area:**

**Answer Area**

```
CREATE EXTERNAL DATA SOURCE AzureDataLakeStore
WITH
( Location1 'abfs://data@newyorktaxidataset.dfs.core.windows.net' ,
credential = ADLS_credential ,
TYPE -
);
```



**Section:**

**Explanation:**

<https://docs.microsoft.com/en-us/sql/t-sql/statements/create-external-data-source-transactsql?view=azure-sqldw-latest&preserve-view=true&tabs=dedicated>

**QUESTION 36**

DRAG DROP

You have an Azure subscription that contains an Azure Synapse Analytics workspace named workspace1. Workspace1 connects to an Azure DevOps repository named repo1. Repo1 contains a collaboration branch named main and a development branch named branch1. Branch1 contains an Azure Synapse pipeline named pipeline1. In workspace1, you complete testing of pipeline1. You need to schedule pipeline1 to run daily at 6 AM.

Which four actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order. NOTE: More than one order of answer choices is correct. You will receive credit for any of the correct orders you select.

**Select and Place:**

**Actions**

- Create a new branch in Repo1.
- Merge the changes from branch1 into main.
- Associate the schedule trigger with pipeline1.
- Switch to Synapse live mode.
- Create a schedule trigger.
- Publish the contents of main.

**Answer Area**

>

<

**Correct Answer:**

**Actions**

- Create a new branch in Repo1.
- 
- 
- Switch to Synapse live mode.
- 
- 

**Answer Area**

- Create a schedule trigger.
- Associate the schedule trigger with pipeline1.
- Merge the changes from branch1 into main.
- Publish the contents of main.

>

<

**Section:**

**Explanation:**

**QUESTION 37**

DRAG DROP

You have an Azure Data Lake Storage Gen 2 account named storage1. You need to recommend a solution for accessing the content in storage1. The solution must meet the following requirements:

List and read permissions must be granted at the storage account level. Additional permissions can be applied to individual objects in storage1. Security principals from Microsoft Azure Active Directory (Azure AD), part of Microsoft Entra, must be used for authentication. What should you use? To answer, drag the appropriate components to the correct requirements. Each component may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content. NOTE: Each correct selection is worth one point.

**Select and Place:**

Components	Answer Area
Access control lists (ACLs)	To grant permissions at the storage account level: <input type="text"/>
Role-based access control (RBAC) roles	To grant permissions at the object level: <input type="text"/>
Shared access signatures (SAS)	
Shared account keys	

**Correct Answer:**

Components	Answer Area
<input type="text"/>	To grant permissions at the storage account level: <input type="text" value="Role-based access control (RBAC) roles"/>
<input type="text"/>	To grant permissions at the object level: <input type="text" value="Access control lists (ACLs)"/>
Shared access signatures (SAS)	
Shared account keys	

**Section:**

**Explanation:**

Box 1: Role-based access control (RBAC) roles

List and read permissions must be granted at the storage account level. Security principals from Microsoft Azure Active Directory (Azure AD), part of Microsoft Entra, must be used for authentication.

Role-based access control (Azure RBAC)

Azure RBAC uses role assignments to apply sets of permissions to security principals. A security principal is an object that represents a user, group, service principal, or managed identity that is defined in Azure Active Directory (AD). A permission set can give a security principal a "coarse-grain" level of access such as read or write access to all of the data in a storage account or all of the data in a container.

Box 2: Access control lists (ACLs)

Additional permissions can be applied to individual objects in storage1. Access control lists (ACLs)

ACLs give you the ability to apply "finer grain" level of access to directories and files. An ACL is a permission construct that contains a series of ACL entries. Each ACL entry associates security principal with an access level.

Reference: <https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-controlmodel>

### QUESTION 38

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

In DB1, you create a stored procedure named SP1. SP1 returns a single row of data that has four columns. You need to add an activity to pipeline to execute SP1. The solution must ensure that the values in the columns are stored as pipeline variables. Which two types of activities can you use to execute SP1? (Refer to Data Engineering on Microsoft Azure documents or guide for Answers/Explanation available at Microsoft.com)

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

**Correct Answer: A, B**

**Section:**

**Explanation:**

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



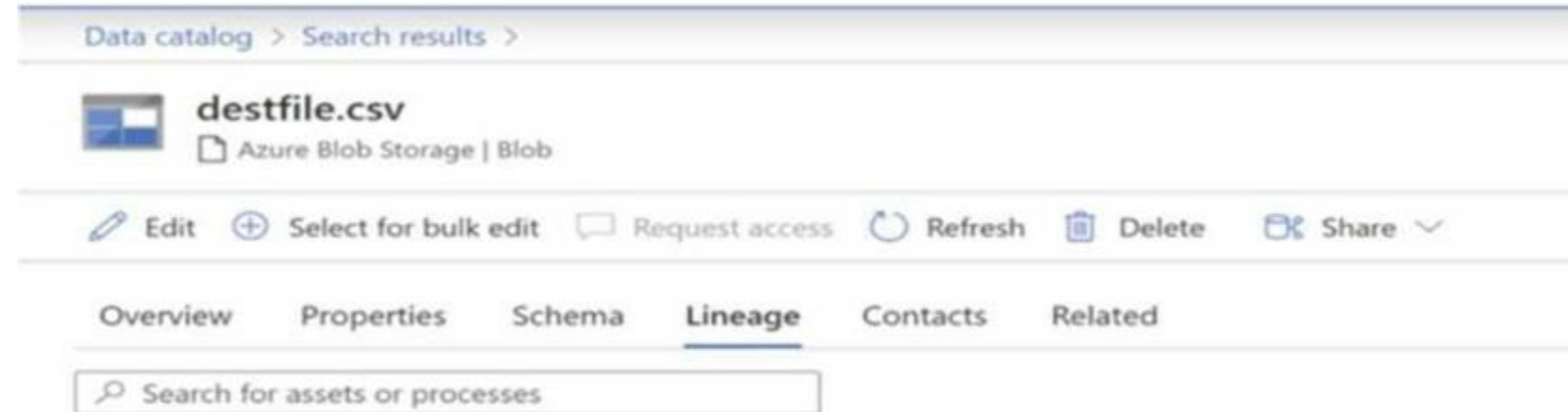
Answer: AB

Explanation:

the two types of activities that you can use to execute SP1 are Stored Procedure and Lookup. A Stored Procedure activity executes a stored procedure on an Azure SQL Database or Azure Synapse Analytics or SQL Server1. You can specify the stored procedure name and parameters in the activity settings1. A Lookup activity retrieves a dataset from any data source that returns a single row of data with four columns2. You can use a query to execute a stored procedure as the source of the Lookup activity2. You can then store the values in the columns as pipeline variables by using expressions2. <https://learn.microsoft.com/en-us/azure/data-factory/transform-data-using-stored-procedure>

### QUESTION 39

You have a Microsoft Purview account. The Lineage view of a CSV file is shown in the following exhibit.



How is the data for the lineage populated?

- A. manually
- B. by scanning data stores
- C. by executing a Data Factory pipeline



**Correct Answer: B**

**Section:**

**Explanation:**

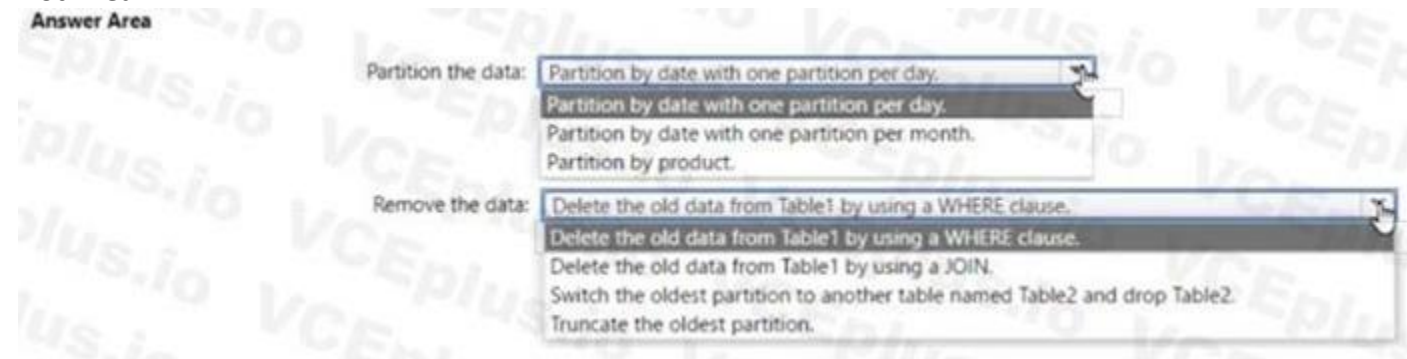
According to Microsoft Purview Data Catalog lineage user guide1, data lineage in Microsoft Purview is a core platform capability that populates the Microsoft Purview Data Map with data movement and transformations across systems2. Lineage is captured as it flows in the enterprise and stitched without gaps irrespective of its source2.

### QUESTION 40

HOTSPOT

You have an Azure Synapse Analytics dedicated SQL pool named Pool1. Pool1 contains a fact table named Table1. Table1 contains sales data. Sixty-five million rows of data are added to Table1 monthly. At the end of each month, you need to remove data that is older than 36 months. The solution must minimize how long it takes to remove the data. How should you partition Table1, and how should you remove the old data? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

**Hot Area:**



**Answer Area:**

**Answer Area**

Partition the data:

- Partition by date with one partition per day.
- Partition by date with one partition per month.
- Partition by product.

Remove the data:

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

**Section:**

**Explanation:**

**QUESTION 41**

**HOTSPOT**

You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Sales.Orders. Sales.Orders contains a column named SalesRep.

You plan to implement row-level security (RLS) for Sales.Orders. You need to create the security policy that will be used to implement RLS. The solution must ensure that sales representatives only see rows for which the value of the SalesRep column matches their username. How should you complete the code? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

**Hot Area:**

**Answer Area**

```
CREATE SCHEMA Security;
GO
CREATE FUNCTION Security.tvf_securitypredicate(@SalesRep AS nvarchar(50))
RETURNS TABLE
WITH
  SCHEMABINDING
  ENCRYPTION
  RETURNS NULL ON NULL INPUT
  SCHEMABINDING
AS
RETURN SELECT 1 AS tvf_securitypredicate_result
WHERE @SalesRep = USER_NAME();
GO
CREATE SECURITY POLICY SalesFilter
  ADD FILTER PREDICATE Security.tvf_securitypredicate(SalesRep)
  ADD BLOCK PREDICATE Security.tvf_securitypredicate(SalesRep)
  ADD BLOCK PREDICATE tvf_securitypredicate_result
  ADD FILTER PREDICATE Security.tvf_securitypredicate(SalesRep)
```

**Answer Area:**

**Answer Area**

```
CREATE SCHEMA Security;  
GO  
CREATE FUNCTION Security.tvf_securitypredicate(@SalesRep AS nvarchar(50))  
RETURNS TABLE  
WITH SCHEMABINDING  
ENCRYPTION  
RETURNS NULL ON NULL INPUT  
AS  
RETURN SELECT 1 AS tvf_securitypredicate_result  
WHERE @SalesRep = USER_NAME();  
GO  
CREATE SECURITY POLICY SalesFilter  
ADD FILTER PREDICATE Security.tvf_securitypredicate(SalesRep)  
ADD BLOCK PREDICATE Security.tvf_securitypredicate(SalesRep)  
ADD BLOCK PREDICATE tvf_securitypredicate_result  
ADD FILTER PREDICATE Security.tvf_securitypredicate(SalesRep)
```

**Section:**

**Explanation:**

**QUESTION 42**

DRAG DROP

You are batch loading a table in an Azure Synapse Analytics dedicated SQL pool. You need to load data from a staging table to the target table. The solution must ensure that if an error occurs while loading the data to the target table, all the inserts in that batch are undone. How should you complete the Transact-SQL code? To answer, drag the appropriate values to the correct targets. Each value may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.

NOTE Each correct selection is worth one point.

**Select and Place:**

Values

BEGIN DISTRIBUTED TRANSACTION

BEGIN TRAN

COMMIT TRAN

ROLLBACK TRAN

SET RESULT\_SET\_CACHING ON

Answer Area

```

BEGIN TRY
  INSERT INTO dbo.Table1 (col1, col2, col3)
  SELECT col1, col2, col3 FROM stage.Table1;
END TRY
BEGIN CATCH
  IF @@TRANCOUNT > 0
  BEGIN
    ;
  END
END CATCH;
IF @@TRANCOUNT >0
BEGIN
  COMMIT TRAN;
END

```

Correct Answer:

Values

BEGIN DISTRIBUTED TRANSACTION

COMMIT TRAN

SET RESULT\_SET\_CACHING ON

Answer Area

BEGIN TRAN

```

BEGIN TRY
  INSERT INTO dbo.Table1 (col1, col2, col3)
  SELECT col1, col2, col3 FROM stage.Table1;
END TRY
BEGIN CATCH
  IF @@TRANCOUNT > 0
  BEGIN
    ROLLBACK TRAN ;
  END
END CATCH;
IF @@TRANCOUNT >0
BEGIN
  COMMIT TRAN;
END

```



Section:

Explanation:

QUESTION 43

**HOTSPOT**

You have an Azure Data Factory pipeline shown the following exhibit.



The execution log for the first pipeline run is shown in the following exhibit.

**Activity runs**  
Pipeline run ID: 87f89922-14fa-468f-b13f-2f86760614ff

All status ▾

Showing 1 - 2 items

Activity name ↑↓	Activity type ↑↓	Run start ↑↓	Duration ↑↓	Status ↑↓
Web_GetIP	Web	Nov 10, 2022, 11:11:36 a	00:00:02	Failed
Exec_COPY_BLOB	Execute Pipeline	Nov 10, 2022, 11:11:25 a	00:00:11	Succeeded

The execution log for the second pipeline run is shown in the following exhibit.

**Activity runs**  
Pipeline run ID: a7b5b522-cfaf-4c09-b3a9-f842986be984

All status ▾

Showing 1 - 3 items

Activity name ↑↓	Activity type ↑↓	Run start ↑↓	Duration ↑↓	Status ↑↓
Set status	Set variable	Nov 10, 2022, 11:13:17 a	00:00:01	Succeeded
Web_GetIP	Web	Nov 10, 2022, 11:12:59 a	00:00:16	Succeeded
Exec_COPY_BLOB	Execute Pipeline	Nov 10, 2022, 11:12:48 a	00:00:11	Skipped



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

**Hot Area:**

**Answer Area**

Statements	Yes	No
The <code>retry</code> property of the <code>Web_GetIP</code> activity is set to 1.	<input type="radio"/>	<input type="radio"/>
The <code>waitOnCompletion</code> property of the <code>Exec_COPY_BLOB</code> activity is set to true.	<input type="radio"/>	<input type="radio"/>
The <code>Exec_COPY_BLOB</code> activity was skipped during the second run due to pipeline dependencies.	<input type="radio"/>	<input type="radio"/>

**Answer Area:**

**Answer Area**

**Statements**

The retry property of the Web\_GetIP activity is set to 1.

The waitOnCompletion property of the Exec\_COPY\_BLOB activity is set to true.

The Exec\_COPY\_BLOB activity was skipped during the second run due to pipeline dependencies.

Yes No

**Section:**

**Explanation:**

**QUESTION 44**

You are designing 2 solution that will use tables in Delta Lake on Azure Databricks. You need to minimize how long it takes to perform the following:

- \*Queries against non-partitioned tables
- \* Joins on non-partitioned columns

Which two options should you include in the solution? Each correct answer presents part of the solution. (Choose Correct Answer and Give Explanation and Reference to Support the answers based from Data Engineering on Microsoft Azure)

- A. Z-Ordering
- B. Apache Spark caching
- C. dynamic file pruning (DFP)
- D. the clone command

**Correct Answer: A, C**

**Section:**

**Explanation:**

- A. Z-Ordering
- B. Apache Spark caching
- C. dynamic file pruning (DFP)
- D. the clone command

Answer: AB

Explanation:

According to the information I found on the web, two options that you should include in the solution to minimize how long it takes to perform queries and joins on non-partitioned tables are:

Z-Ordering: This is a technique to colocate related information in the same set of files. This colocality is automatically used by Delta Lake in data-skipping algorithms. This behavior dramatically reduces the amount of data that Delta Lake on Azure Databricks needs to read. Apache Spark caching: This is a feature that allows you to cache data in memory or on disk for faster access. Caching can improve the performance of repeated queries and joins on the same data. You can cache Delta tables using the CACHE TABLE or CACHE LAZY commands.

To minimize the time it takes to perform queries against non-partitioned tables and joins on nonpartitioned columns in Delta Lake on Azure Databricks, the following options should be included in the solution:

1. Z-Ordering: Z-Ordering improves query performance by co-locating data that share the same column values in the same physical partitions. This reduces the need for shuffling data across nodes during query execution. By using Z-Ordering, you can avoid full table scans and reduce the amount of data processed. 2. Apache Spark caching: Caching data in memory can improve query performance by reducing the amount of data read from disk. This helps to speed up subsequent queries that need to access the same data. When you cache a table, the data is read from the data source and stored in memory.

Subsequent queries can then read the data from memory, which is much faster than reading it from disk.

Reference:

Delta Lake on Databricks: <https://docs.databricks.com/delta/index.html>

Best Practices for Delta Lake on Databricks: <https://databricks.com/blog/2020/05/14/best-practicesfor-delta-lake-on-databricks.html>

**QUESTION 45**

You are deploying a lake database by using an Azure Synapse database template. You need to add additional tables to the database. The solution must use the same grouping method as the template tables. Which grouping method should you use?



- A. business area
- B. size
- C. facts and dimensions
- D. partition style

**Correct Answer: A**

**Section:**

**Explanation:**

Business area: This is how the Azure Synapse database templates group tables by default. Each template consists of one or more enterprise templates that contain tables grouped by business areas. For example, the Retail template has business areas such as Customer, Product, Sales, and Store123. Using the same grouping method as the template tables can help you maintain consistency and compatibility with the industry-specific data model. <https://techcommunity.microsoft.com/t5/azure-synapse-analytics-blog/database-templates-in-azure-synapse-analytics/ba-p/2929112>

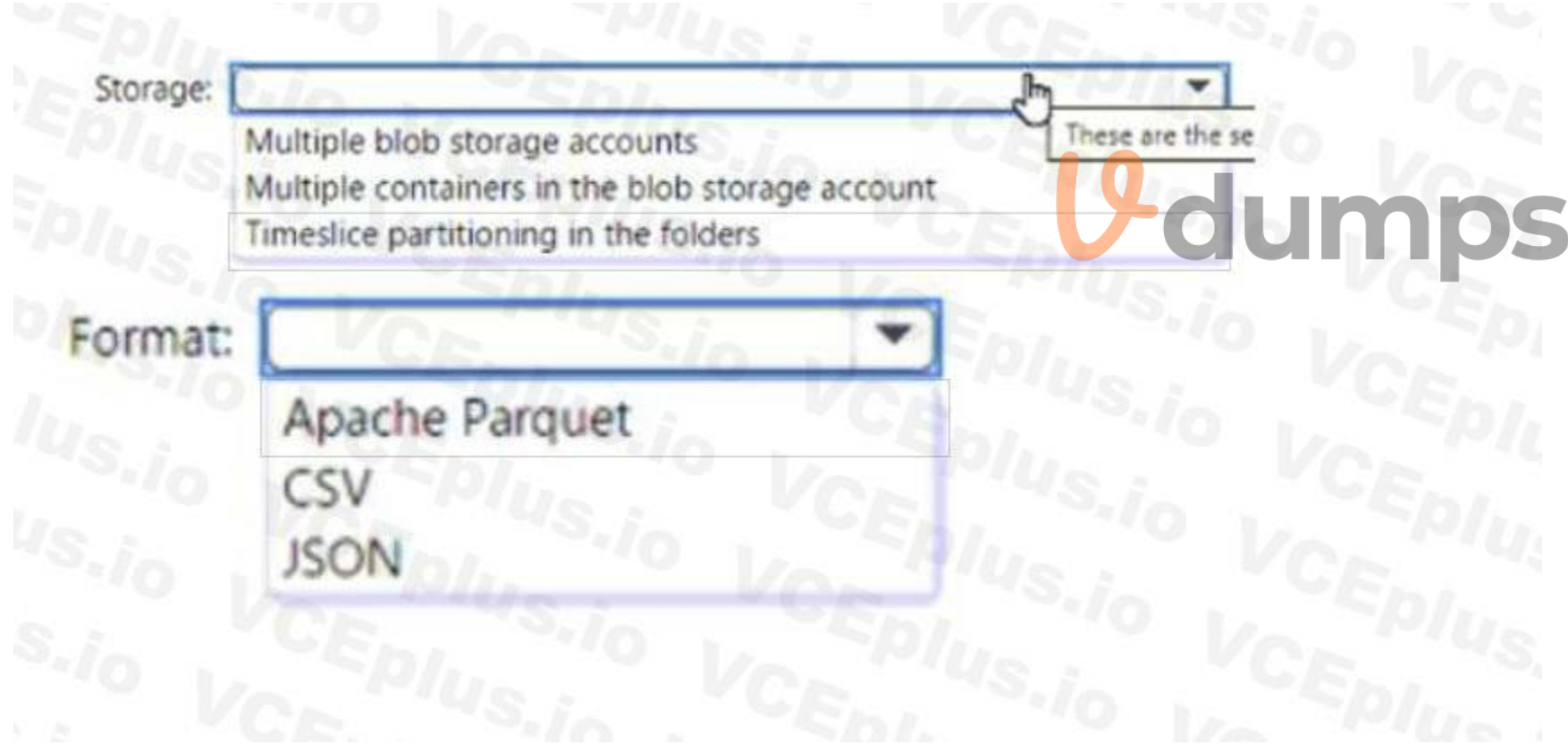
**QUESTION 46**

**HOTSPOT**

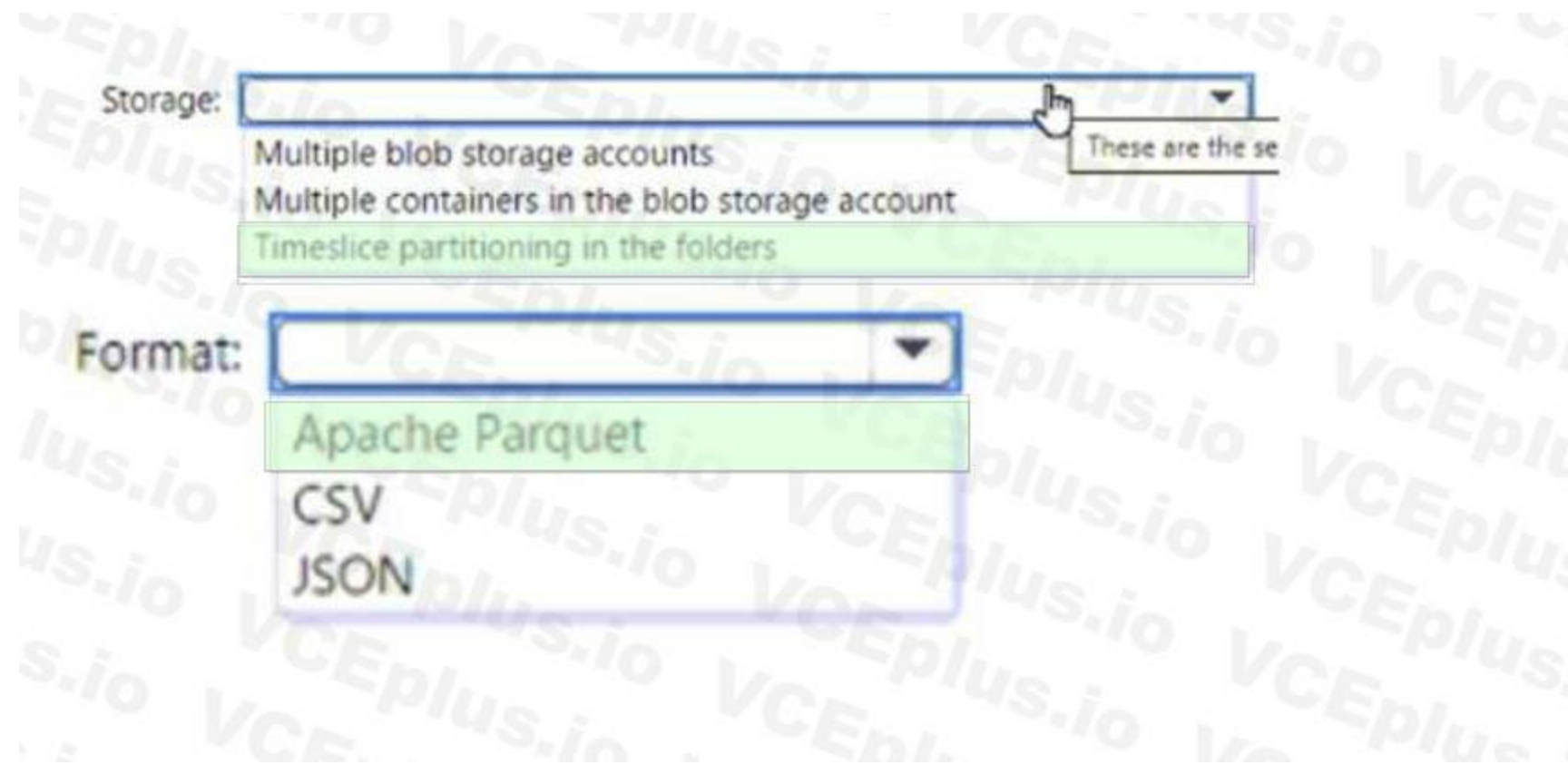
You have an Azure Blob storage account that contains a folder. The folder contains 120,000 files. Each file contains 62 columns. Each day, 1,500 new files are added to the folder.

You plan to incrementally load five data columns from each new file into an Azure Synapse Analytics workspace. You need to minimize how long it takes to perform the incremental loads. What should you use to store the files and format?

**Hot Area:**



**Answer Area:**



**Section:**

**Explanation:**

Box 1 = timeslice partitioning in the folders

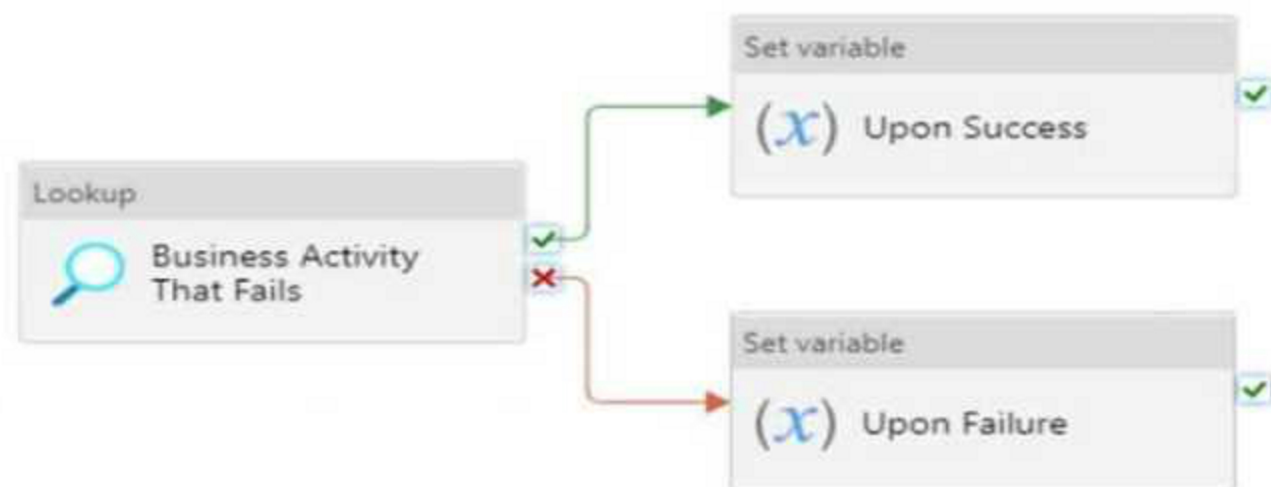
This means that you should organize your files into folders based on a time attribute, such as year, month, day, or hour. For example, you can have a folder structure like /yyyy/mm/dd/file.csv. This way, you can easily identify and load only the new files that are added each day by using a time filter in your Azure Synapse pipeline<sup>12</sup>. Timeslice partitioning can also improve the performance of data loading and querying by reducing the number of files that need to be scanned

Box = 2 Apache Parquet

This is because Parquet is a columnar file format that can efficiently store and compress data with many columns. Parquet files can also be partitioned by a time attribute, which can improve the performance of incremental loading and querying by reducing the number of files that need to be scanned<sup>123</sup>. Parquet files are supported by both dedicated SQL pool and serverless SQL pool in Azure Synapse Analytics<sup>2</sup>.

**QUESTION 47**

You have the Azure Synapse Analytics pipeline shown in the following exhibit.



You need to add a set variable activity to the pipeline to ensure that after the pipeline's completion, the status of the pipeline is always successful. What should you configure for the set variable activity?



- A. a success dependency on the Business Activity That Fails activity
- B. a failure dependency on the Upon Failure activity
- C. a skipped dependency on the Upon Success activity
- D. a skipped dependency on the Upon Failure activity

**Correct Answer: B**

**Section:**

**Explanation:**

A failure dependency means that the activity will run only if the previous activity fails. In this case, setting a failure dependency on the Upon Failure activity will ensure that the set variable activity will run after the pipeline fails and set the status of the pipeline to successful.

<https://www.validexamdumps.com>

#### QUESTION 48

You are building a data flow in Azure Data Factory that upserts data into a table in an Azure Synapse Analytics dedicated SQL pool. You need to add a transformation to the data flow. The transformation must specify logic indicating when a row from the input data must be upserted into the sink.

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

- A. join
- B. select
- C. surrogate key
- D. alter row

**Correct Answer: D**

**Section:**

**Explanation:**

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



#### QUESTION 49

HOTSPOT

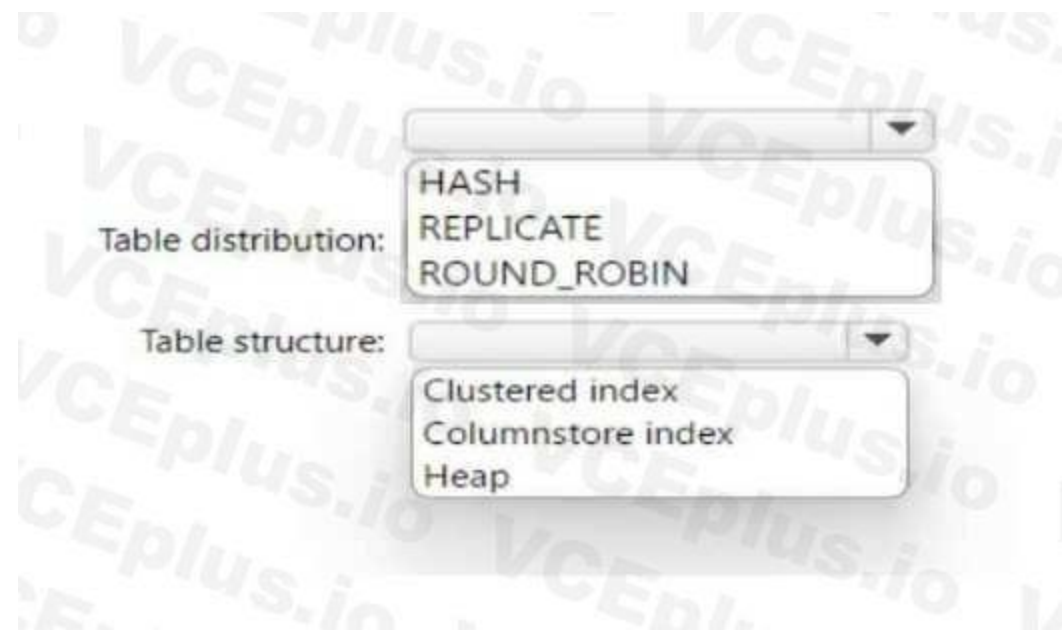
You are incrementally loading data into fact tables in an Azure Synapse Analytics dedicated SQL pool.

Each batch of incoming data is staged before being loaded into the fact tables.

You need to ensure that the incoming data is staged as quickly as possible.

How should you configure the staging tables? To answer, select the appropriate options in the answer area.

**Hot Area:**



**Answer Area:**



**Section:**

**Explanation:**

Round-robin distribution is recommended for staging tables because it distributes data evenly across all the distributions without requiring a hash column. This can improve the speed of data loading and avoid data skew. Heap tables are recommended for staging tables because they do not have any indexes or partitions that can slow down the data loading process. Heap tables are also easier to truncate and reload than clustered index or columnstore index tables.

**QUESTION 50**

DRAG DROP

You have an Azure Synapse Analytics serverless SQ1 pool.

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

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

How should you complete the query? To answer, drag the appropriate values to the correct targets.

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

**Select and Place:**

Values

BULK

DATA\_SOURCE

LOCATION

OPENROWSET

Answer Area

```
SELECT TOP 100 *
FROM [ ] (
[ ] 'https://adls1.dfs.core.windows.net/container1/folder1/*.csv',
FORMAT = 'CSV') AS rows
```

Correct Answer:

Values

DATA\_SOURCE

LOCATION

Answer Area

```
SELECT TOP 100 *
FROM OPENROWSET (
BULK [ ] 'https://adls1.dfs.core.windows.net/container1/folder1/*.csv',
FORMAT = 'CSV') AS rows
```

Section:

Explanation:

### QUESTION 51

HOTSPOT

You have an Azure Synapse Analytics dedicated SQL pool.

You need to monitor the database for long-running queries and identify which queries are waiting on resources.

Which dynamic management view should you use for each requirement? To answer, select the appropriate options in the answer area.

NOTE: Each correct answer is worth one point.

Hot Area:

Answer Area

Monitor the database for long-running queries:

- sys.dm\_pdw\_exec\_requests
- sys.dm\_pdw\_exec\_requests
- sys.dm\_pdw\_sql\_requests
- sys.dm\_pdw\_exec\_sessions

Identify which queries are waiting on resources:

- sys.dm\_pdw\_lock\_waits
- sys.dm\_pdw\_waits
- sys.dm\_pdw\_lock\_waits
- sys.resource\_governor\_workload\_groups

Answer Area:

Answer Area

Monitor the database for long-running queries:

Identify which queries are waiting on resources:

Section:  
Explanation:

QUESTION 52

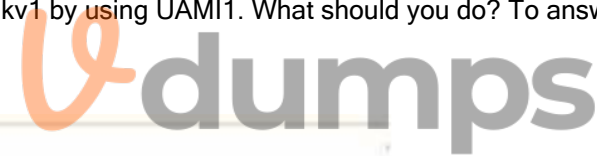
HOTSPOT

You have an Azure subscription that contains the resources shown in the following table.

Name	Type	Description
ws1	Azure Synapse Analytics workspace	None
kv1	Azure Key Vault	None
UAMI1	User-assigned managed identity	Associated with ws1
sp1	Apache Spark pool in Azure Synapse Analytics	Associated with ws1

You need to ensure that you can Spark notebooks in ws1. The solution must ensure secrets from kv1 by using UAMI1. What should you do? To answer, select the appropriate options in the answer area.  
NOTE: Each correct selection is worth one point.

Hot Area:



Answer Area

In the Azure portal:

In Synapse Studio:

Answer Area:

Answer Area

In the Azure portal:

In Synapse Studio:

Section:

**Explanation:**

**QUESTION 53**

You have an Azure Synapse Analytics dedicated SQL pod. You need to create a pipeline that will execute a stored procedure in the dedicated SQL pool and use the returned result set as the input (or a downstream activity). The solution must minimize development effort. Which Type of activity should you use in the pipeline?

- A. Notebook
- B. U-SQL
- C. Script
- D. Stored Procedure

**Correct Answer: D**

**Section:**

**QUESTION 54**

You have an Azure subscription that contains an Azure Synapse Analytics workspace named ws1 and an Azure Cosmos D6 database account named Cosmos1. Cosmos1 contains a container named container 1 and ws1 contains a serverless1 SQL pool.

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

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

**Correct Answer: A, C, D**

**Section:**

**QUESTION 55**

**HOTSPOT**

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

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

- \* Detect whether the data of a given customer has changed in the DimCustomer table.
- Perform an upsert to the DimCustomer table.

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

NOTE; Each correct selection is worth one point.

**Hot Area:**

**Answer Area**

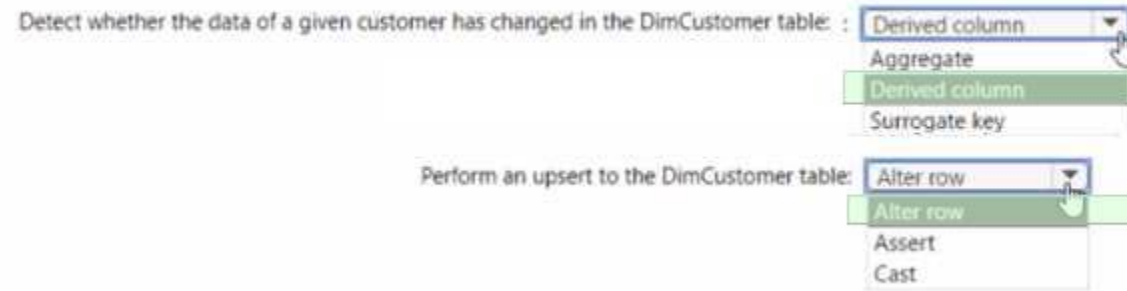
Detect whether the data of a given customer has changed in the DimCustomer table:

Perform an upsert to the DimCustomer table:



**Answer Area:**

Answer Area



**Section:**

**Explanation:**

**QUESTION 56**

**HOTSPOT**

You have an Azure subscription that contains an Azure Cosmos DB analytical store and an Azure Synapse Analytics workspace named WS 1. WS1 has a serverless SQL pool name Pool1. You execute the following query by using Pool1.

```
WITH IDENTITY = 'SHARED /  
SECRET = 'fed4347479872423433563653456345ddfa==';  
  
SELECT clientID AS ClientID,  
       client AS ClientName  
FROM OPENROWSET  
(  
    PROVIDER = 'CosmosDB',  
    CONNECTION = 'Account=account1;Database=database1',  
    OBJECT = 'clients',  
    SERVER_CREDENTIAL = 'AccountCred'  
)  
WITH  
(  
    clientID int,  
    client varchar(50),  
    streetAddress varchar(100)  
) AS c;
```



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

NOTE: Each correct selection is worth one point.

**Hot Area:**

## Answer Area

Statements	Yes	No
The query returns three columns.	<input type="radio"/>	<input type="radio"/>
The container being queried is named <code>clients</code> .	<input type="radio"/>	<input type="radio"/>
Authentication is performed by using an account key.	<input type="radio"/>	<input type="radio"/>

Answer Area:

## Answer Area

Statements	Yes	No
The query returns three columns.	<input type="radio"/>	<input checked="" type="radio"/>
The container being queried is named <code>clients</code> .	<input checked="" type="radio"/>	<input type="radio"/>
Authentication is performed by using an account key.	<input checked="" type="radio"/>	<input type="radio"/>

Section:

Explanation:

### QUESTION 57

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

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

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

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

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

Does this meet the goal?

A. Yes

B. NO

Correct Answer: A

Section:

**QUESTION 58**

You have an Azure Synapse Analytics dedicated SQL pool.

You plan to create a fact table named Table1 that will contain a clustered columnstore index.

You need to optimize data compression and query performance for Table1.

What is the minimum number of rows that Table1 should contain before you create partitions?

- A. 100,000
- B. 600,000
- C. 1 million
- D. 60 million

**Correct Answer: A**

**Section:**

**QUESTION 59**

You have an Azure subscription that contains an Azure Data Factory data pipeline named Pipeline1, a Log Analytics workspace named LA1, and a storage account named account1.

You need to retain pipeline-run data for 90 days. The solution must meet the following requirements:

\* The pipeline-run data must be removed automatically after 90 days.

\* Ongoing costs must be minimized.

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

- A. Configure Pipeline1 to send logs to LA1.
- B. From the Diagnostic settings (classic) settings of account1, set the retention period to 90 days.
- C. Configure Pipeline1 to send logs to account1.
- D. From the Data Retention settings of LA1, set the data retention period to 90 days.



**Correct Answer: A, B**

**Section:**

**QUESTION 60**

HOTSPOT

In Azure Data Factory, you have a schedule trigger that is scheduled in Pacific Time.

Pacific Time observes daylight saving time.

The trigger has the following JSON file.



```
{
  "name": "Trigger 1",
  "properties": {
    "annotations": [],
    "runtimeState": "Started",
    "pipelines": [],
    "type": "ScheduleTrigger",
    "typeProperties": {
      "recurrence": {
        "frequency": "Week",
        "interval": 1,
        "startTime": "2022-08-05T04:00:00",
        "timeZone": "Pacific Standard Time",
        "schedule": {
          "minutes": [
            0
          ],
          "hours": [
            3,
            21
          ],
          "weekDays": [
            "Sunday",
            "Saturday"
          ]
        }
      }
    }
  }
}
```



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

**Hot Area:**

**Answer Area**

The trigger will execute [answer choice] on Sunday, March 3, 2024.

- two times
- one time
- two times
- zero times

The trigger [answer choice] daylight saving time.

- is unaffected by
- is unaffected by
- will automatically adjust for
- will require an adjustment for

Answer Area:

**Answer Area**

The trigger will execute [answer choice] on Sunday, March 3, 2024.

- two times
- one time
- two times
- zero times

The trigger [answer choice] daylight saving time.

- is unaffected by
- is unaffected by
- will automatically adjust for
- will require an adjustment for



Section:

Explanation:

**QUESTION 61**

DRAG DROP

You have an Azure Synapse Analytics dedicated SQL pool named SQL1 that contains a hash-distributed fact table named Table1.

You need to recreate Table1 and add a new distribution column. The solution must maximize the availability of data.

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

Select and Place:

**Actions**

- Drop Table1\_old.
- Run DBCC PDW\_SHOWSPACEUSED.
- Drop the indexes of Table1.
- Create a new table named Table1v2 by running CTAS.
- Rename Table1 as Table1\_old.
- Rename Table1v2 as Table1.

**Answer Area****Correct Answer:****Actions**

- Drop Table1\_old.
- Run DBCC PDW\_SHOWSPACEUSED.
- 
- 
- 
- 

**Answer Area**

- Drop the indexes of Table1.
- Create a new table named Table1v2 by running CTAS.
- Rename Table1 as Table1\_old.
- Rename Table1v2 as Table1.

**Section:****Explanation:**

Drop the indexes of Table1.  
 Create a new table named Table 1v2 by running CTAS  
 Rename Table1 as Table1\_old.  
 Rename Table 1v2 as Table1.

**QUESTION 62**

You have an Azure data factory that connects to a Microsoft Purview account. The data 'factory is registered in Microsoft Purview. You update a Data Factory pipeline. You need to ensure that the updated lineage is available in Microsoft Purview. What should you do first?

- A. Disconnect the Microsoft Purview account from the data factory.
- B. Locate the related asset in the Microsoft Purview portal.
- C. Execute an Azure DevOps build pipeline.
- D. Execute the pipeline.

**Correct Answer: D**

**Section:**

**QUESTION 63**

You have an Azure Synapse Analytics dedicated SQL pool named Pool1.

Pool1 contains two tables named SalesFact\_Staging and SalesFact. Both tables have a matching number of partitions, all of which contain data.

You need to load data from SalesFact\_Staging to SalesFact by switching a partition.

What should you specify when running the alter TABLE statement?

- A. WITH NOCHECK
- B. WITH (TRUNCATE.TASGET = ON)
- C. WITH (TRACK.COLUMNS. UPOATED =ON)
- D. WITH CHECK

**Correct Answer: B**

**Section:**

**QUESTION 64**

You have an Azure Synapse Analytics dedicated SQL pool.

You run PDW\_SHOWSPACEUSED('dbo.FactInternetSales'); and get the results shown in the following table.

ROWS	RESERVED_SPACE	DATA_SPACE	INDEX_SPACE	UNUSED_SPACE	PDW_NODE_ID	DISTRIBUTION_ID
694	2776	616	48	2112	1	1
407	2704	576	48	2080	1	2
53	2376	512	16	1848	1	3
58	2376	512	16	1848	1	4
188	2632	528	32	2072	1	5
195	2696	536	32	2128	1	6
5995	3464	1424	32	2008	1	7
0	2232	496	0	1736	1	8
264	2576	544	40	1992	1	9
3008	3016	960	32	2024	1	10
--	--	--	--	--	--	--
1550	2832	752	48	2032	1	50
1238	2832	696	40	2096	1	51
192	2632	528	32	2072	1	52
1127	2768	680	48	2040	1	53
1244	3032	704	64	2264	1	54
409	2632	568	32	2032	1	55
0	2232	496	0	1736	1	56
1437	2832	728	40	2064	1	57
0	2232	496	0	1736	1	58
384	2632	568	32	2040	1	59
225	2768	544	40	2184	1	60

Which statement accurately describes the dbo.FactInternetSales table?

- A. All distributions contain data.
- B. The table contains less than 10,000 rows.
- C. The table uses round-robin distribution.
- D. The table is skewed.

**Correct Answer: D**

**Section:**

**QUESTION 65**



You have two fact tables named Flight and Weather. Queries targeting the tables will be based on the join between the following columns.

Table	Column
Flight	ArrivalAirportID ArrivalDateTime
Weather	AirportID ReportDateTime

You need to recommend a solution that maximizes query performance. What should you include in the recommendation?

- A. In the tables use a hash distribution of ArrivalDateTime and ReportDateTime.
- B. In the tables use a hash distribution of ArrivalAirportID and AirportID.
- C. In each table, create an IDENTITY column.
- D. In each table, create a column as a composite of the other two columns in the table.

**Correct Answer: B**

**Section:**

**Explanation:**

Hash-distribution improves query performance on large fact tables. Incorrect Answers:

A: Do not use a date column for hash distribution. All data for the same date lands in the same distribution. If several users are all filtering on the same date, then only 1 of the 60 distributions do all the processing work.

#### QUESTION 66

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

- A. Azure Synapse Analytics
- B. Azure HDInsight
- C. Azure Machine Learning
- D. Azure Data Factory
- E. Azure Databricks

**Correct Answer: B, D**

**Section:**

#### QUESTION 67

You are designing the folder structure for an Azure Data Lake Storage Gen2 account. You identify the following usage patterns:

- Users will query data by using Azure Synapse Analytics serverless SQL pools and Azure Synapse Analytics serverless Apache Spark pods.
- Most queries will include a filter on the current year or week.
- Data will be secured by data source.

You need to recommend a folder structure that meets the following requirements:

- Supports the usage patterns
- Simplifies folder security
- Minimizes query times

Which folder structure should you recommend?

A.

`\YYYY\MM\DataSource\SubjectArea\FileData_YYYY_MM_DD.parquet`

B.

`DataSource\SubjectArea\MM\YYYY\FileData_YYYY_MM_DD.parquet`

C.

`\DataSource\SubjectArea\YYYY\MM\FileData_YYYY_MM_DD.parquet`

D.

`\DataSource\SubjectArea\YYYY-MM\FileData_YYYY_MM_DD.parquet`

E.

`MM\YYYY\SubjectArea\DataSource\FileData_YYYY_MM_DD.parquet`

**Correct Answer: C**

**Section:**

**Explanation:**

Data will be secured by data source. -> Use DataSource as top folder. Most queries will include a filter on the current year or week -> Use \YYYY\WW\ as subfolders. Common Use Cases

A common use case is to filter data stored in a date (and possibly time) folder structure such as /YYYY/MM/DD/ or /YYYY/MM/YYYY-MM-DD/. As new data is generated/sent/copied/moved to the storage account, a new folder is created for each specific time period. This strategy organises data into a maintainable folder structure.

Reference: <https://www.serverlesssql.com/optimisation/azurestoragefilteringusingfilepath/>

#### QUESTION 68

You have an Azure subscription that contains an Azure Synapse Analytics workspace and a user named User1.

You need to ensure that User1 can review the Azure Synapse Analytics database templates from the gallery. The solution must follow the principle of least privilege.

Which role should you assign to User1?

- A. Synapse User
- B. Synapse Contributor
- C. Storage blob Data Contributor
- D. Synapse Administrator

**Correct Answer: A**

**Section:**

#### QUESTION 69

HOTSPOT

You have Azure Data Factory configured with Azure Repos Git integration. The collaboration branch and the publish branch are set to the default values.

You have a pipeline named pipeline 1.

You build a new version of pipeline1 in a branch named feature 1.

From the Data Factory Studio, you select Publish

The source code of which branch will be built, and which branch will contain the output of the Azure Resource Manager (ARM) template? To answer, select the appropriate options in the answer area.

**Hot Area:**

Answer Area

Source code:   
adf\_publish  
feature1  
main

ARM template output:   
adf\_publish  
feature1  
main

Answer Area:

Answer Area

Source code:   
adf\_publish  
feature1  
main

ARM template output:   
adf\_publish  
feature1  
main



Section:

Explanation:

QUESTION 70

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

- A. Azure Stream Analytics Edge application using Microsoft Visual Studio.
- B. Azure Analytics Services using Azure portal
- C. Azure Analysis Services using Microsoft visual Studio
- D. Azure Data Factory instance using Microsoft visual Studio

Correct Answer: A

Section:

QUESTION 71

You have an Azure subscription that contains an Azure data factory named ADF1. From Azure Data Factory Studio, you build a complex data pipeline in ADF1. You discover that the Save button is unavailable and there are validation errors that prevent the pipeline from being published. You need to ensure that you can save the logic of the pipeline. Solution: You enable Git integration for ADF1.

- A. Yes

B. No

**Correct Answer: B**

**Section:**

**QUESTION 72**

HOTSPOT

You have an Azure subscription that contains the resources shown in the following table.

Name	Type	Description
Workspace1	Azure Synapse workspace	Contains the Built-in serverless SQL pool
Pool1	Azure Synapse Analytics dedicated SQL pool	Deployed to Workspace1
storage1	Storage account	Hierarchical namespace enabled

The storage1 account contains a container named container1. The container1 container contains the following files.

```
Webdata <root folder>
  Monthly <folder>
    _monthly.csv
    Monthly.csv
  .testdata.csv
  testdata.csv

In Pool1, you run the following script.

CREATE EXTERNAL DATA SOURCE Ds1
WITH
( LOCATION = 'abfss://container1@storage1.dfs.core.windows.net' ,
  CREDENTIAL = credential1,
  TYPE = HADOOP
);
```



In the Built-in serverless SQL pool, you run the following script

```
CREATE EXTERNAL DATA SOURCE Ds2
WITH (
  LOCATION = 'https://storage1.blob.core.windows.net/container1/webdata/',
  CREDENTIAL = credential2
);
```

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

**Hot Area:**

Answer Area	Statements	Yes	No
	An external table that uses Ds1 can read the _monthly.csv file.	<input type="radio"/>	<input type="radio"/>
	An external table that uses Ds1 can read the Monthly.csv file.	<input type="radio"/>	<input type="radio"/>
	An external table that uses Ds2 can read the .testdata.csv file.	<input type="radio"/>	<input type="radio"/>

**Answer Area:**



Answer Area

Statements

An external table that uses Ds1 can read the \_monthly.csv file.

Yes

No

An external table that uses Ds1 can read the Monthly.csv file.

An external table that uses Ds2 can read the .testdata.csv file.

Section:

Explanation:

QUESTION 73

HOTSPOT

You have an Azure Synapse Analytics workspace that contains three pipelines and three triggers named Trigger 1, Trigger2, and Trigger3. Trigger 3 has the following definition.

```
...
{
  "name": "Trigger3",
  "properties": {
    "annotations": [],
    "runtimeState": "Stopped",
    "pipeline": {
      "pipelineReference": {
        "referenceName": "Pipeline 3",
        "type": "PipelineReference"
      }
    },
    "type": "TumblingWindowTrigger",
    "typeProperties": {
      "frequency": "Hour",
      "interval": 1,
    },
    "dependsOn": [
      {
        "type": "TumblingWindowTriggerDependencyReference",
        "size": "0.03:00:00",
        "offset": "-0.02:00:00",
        "referenceTrigger": {
          "referenceName": "Trigger1",
          "type": "TriggerReference"
        }
      },
      {
        "type": "SelfDependencyTumblingWindowTriggerReference",
        "offset": "-0.03:00:00"
      }
    ]
  }
}
...

```

Hot Area:



**Answer Area**

Statements	Yes	No
Pipeline3 will execute when Trigger3 fires.	<input type="radio"/>	<input type="radio"/>
Up to three instances of Trigger3 can fire simultaneously.	<input type="radio"/>	<input type="radio"/>
Trigger3 will fire three hours after Trigger1 has fired three times, and Trigger2 has fired three times.	<input type="radio"/>	<input type="radio"/>

**Answer Area:**  
**Answer Area**

Statements	Yes	No
Pipeline3 will execute when Trigger3 fires.	<input checked="" type="radio"/>	<input type="radio"/>
Up to three instances of Trigger3 can fire simultaneously.	<input type="radio"/>	<input checked="" type="radio"/>
Trigger3 will fire three hours after Trigger1 has fired three times, and Trigger2 has fired three times.	<input checked="" type="radio"/>	<input type="radio"/>

**Section:**  
**Explanation:**

**QUESTION 74**

You have an Azure Stream Analytics job named Job1.  
The metrics of Job1 from the last hour are shown in the following table.



Metric	Time aggregation	Value
SU (Memory) % Utilization	Average	70
CPU % Utilization	Average	20
Runtime Errors	Total	0
Watermark Delay	Average	20
Input Deserialization Errors	Total	0

The late arrival tolerance for Job1 is set to the five seconds.  
You need to optimize Job1.  
Which two actions achieve the goal? Each correct answer presents a complete solution.  
NOTE: Each correct answer is worth one point.

- A. Resolution errors in inputs processing.
- B. Parallelize the query
- C. Resolution errors in output processing
- D. Increase the number of SUs.

**Correct Answer: B, D**  
**Section:**

**QUESTION 75**

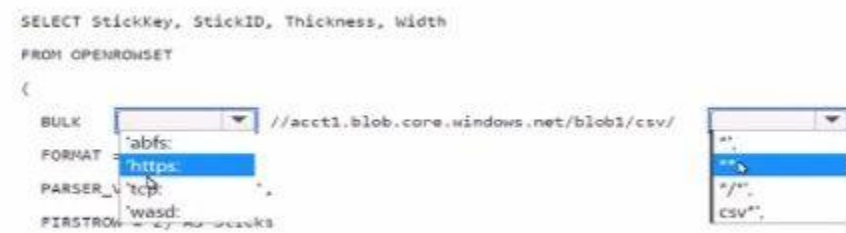
**HOTSPOT**  
You have an Azure subscription that contains a storage account. The account contains a blob container named blob1 and an Azure Synapse Analytic serve-less SQL pool  
You need to Query the CSV files stored in blob1. The solution must ensure that all the files in a (older named csv and all its subfolders are queried

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

**Hot Area:**

Answer Area

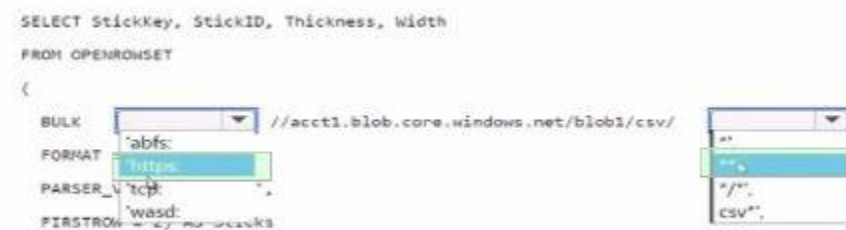
```
SELECT StickKey, StickID, Thickness, Width
FROM OPENROWSET
(
  BULK [abfs://acct1.blob.core.windows.net/blob1/csv/]
  FORMAT 'https:'
  PARSE_NAME('tcp', 'wssd')
  FIRSTROW 1) AS OPENROWSET
```



**Answer Area:**

Answer Area

```
SELECT StickKey, StickID, Thickness, Width
FROM OPENROWSET
(
  BULK [abfs://acct1.blob.core.windows.net/blob1/csv/]
  FORMAT 'https:'
  PARSE_NAME('tcp', 'wssd')
  FIRSTROW 1) AS OPENROWSET
```



**Section:**

**Explanation:**

**QUESTION 76**

You have an Azure subscription that contains a Microsoft Purview account.  
You need to search the Microsoft Purview Data Catalog to identify assets that have an assetType property of Table or View  
Which query should you run?

- A. assetType IN (Table, 'View')
- B. assetType:Table OR assetType:View
- C. assetType - (Table or view)
- D. assetType:(Table OR View)

**Correct Answer: B**

**Section:**

**QUESTION 77**

You have an Azure Synapse Analytics workspace.  
You plan to deploy a lake database by using a database template in Azure Synapse.  
Which two elements are included in the template? Each correct answer presents part of the solution.  
NOTE: Each correct selection is worth one point

- A. relationships
- B. table definitions

- C. table permissions
- D. linked services
- E. data formats

**Correct Answer: A, B**

**Section:**

#### QUESTION 78

You have an Azure subscription that contains an Azure data factory named ADF1.

From Azure Data Factory Studio, you build a complex data pipeline in ADF1.

You discover that the Save button is unavailable and there are validation errors that prevent the pipeline from being published.

You need to ensure that you can save the logic of the pipeline.

Solution: You export ADF1 as an Azure Resource Manager (ARM) template.

- A. Yes
- B. No

**Correct Answer: B**

**Section:**

#### QUESTION 79

You have a Microsoft Entra tenant.

The tenant contains an Azure Data Lake Storage Gen2 account named storage1 that has two containers named fs1 and fs2. You have a Microsoft Entra group named Department

A. You need to meet the following requirements:

\* DepartmentA must be able to read, write, and list all the files in fs1.

\* DepartmentA must be prevented from accessing any files in fs2

\* The solution must use the principle of least privilege.

Which role should you assign to DepartmentA?

- A. Contributor for fs1
- B. Storage Blob Data Owner for fs1
- C. Storage Blob Data Contributor for storage1
- D. Storage Blob Data Contributor for fs1

**Correct Answer: D**

**Section:**

#### QUESTION 80

You have an Azure Stream Analytics job that read data from an Azure event hub.

You need to evaluate whether the job processes data as quickly as the data arrives or cannot keep up.

Which metric should you review?

- A. InputEventLastPunctuationTime
- B. Input Sources Receive
- C. Late input Events
- D. Backlogged input Events

**Correct Answer: B**



**Section:**

**QUESTION 81**

HOTSPOT

A company uses the Azure Data Lake Storage Gen2 service.

You need to design a data archiving solution that meets the following requirements:

Data that is older than five years is accessed infrequently but must be available within one second when requested.

Data that is older than seven years is NOT accessed.

Costs must be minimized while maintaining the required availability.

How should you manage the data? To answer, select the appropriate option in the answers area.

NOTE: Each correct selection is worth one point.

**Hot Area:**

Answer Area

Data over five years old:

Data over seven years old:



**Answer Area:**

Answer Area

Data over five years old:

Data over seven years old:

**Section:**

**Explanation:**

**QUESTION 82**

HOTSPOT

You have an Azure Synapse Analytics dedicated SQL pool that hosts a database named DB1. You need to ensure that DB1 meets the following security requirements:

\* When credit card numbers show in applications, only the last four digits must be visible.

\* Tax numbers must be visible only to specific users.

What should you use for each requirement? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

**Hot Area:**

Answer Area

Credit card numbers:   
Column-level security  
Dynamic Data Masking  
Row-level security (RLS)

Tax numbers:   
Column-level security  
Row-level security (RLS)  
Transparent Database Encryption (TDE)

**Answer Area:**

Answer Area

Credit card numbers:   
Column-level security  
Dynamic Data Masking  
Row-level security (RLS)

Tax numbers:   
Column-level security  
Row-level security (RLS)  
Transparent Database Encryption (TDE)

**Section:**

**Explanation:**

