

## Question #1

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,
    [Portalcode] [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. [ManagerEmployeeKey] [int] NULL

## Question #2

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 EmployeeName = 'Alice';  
What will be returned by the query?
```

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

**Correct Answer:**

A. 24

**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 index

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	Answer Area
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 <code>DELETE</code> statement where the value in the Date column is more than 36 months ago.	
Copy the data to a new table by using <code>CREATE TABLE AS SELECT (CTAS)</code> .	

### Correct Answer:

Actions	Answer Area
Switch the partition containing the stale data from SalesFact to SalesFact_Work.	Create an empty table named SalesFact_Work that has the same schema as SalesFact.
Truncate the partition containing the stale data.	Switch the partition containing the stale data from SalesFact to SalesFact_Work.
Drop the SalesFact_Work table.	Drop the SalesFact_Work table.
Create an empty table named SalesFact_Work that has the same schema as SalesFact.	
Execute a <code>DELETE</code> statement where the value in the Date column is more than 36 months ago.	
Copy the data to a new table by using <code>CREATE TABLE AS SELECT (CTAS)</code> .	

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.

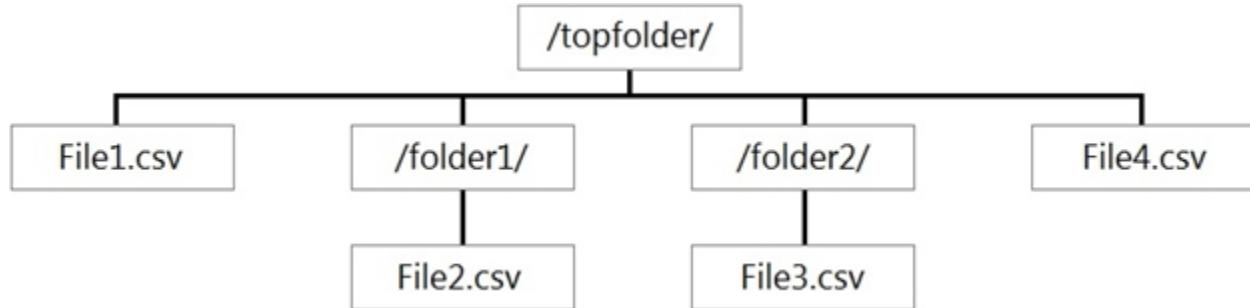
Step 3: Drop the SalesFact\_Work table.

Reference:

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

### Question #4

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. File1.csv and File4.csv only

**Question #5**

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:	

**Correct Answer:**

## **Answer Area**

Report1:	

Report2:	

Report1: CSV -

CSV: The destination writes records as delimited data.

Report2: AVRO -

AVRO supports timestamps.

Not Parquet, TSV: Not options for Azure Data Lake Storage Gen2.

Reference:

<https://streamsets.com/documentation/datacollector/latest/help/datacollector/UserGuide/Destinations/ADLS-G2-D.html>

### Question #6

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. ./{{SubjectArea}}/{{DataSource}}/{{YYYY}}/{{MM}}/{{DD}}/{{FileData}}\_{{YYYY}}\_{{MM}}\_{{DD}}.csv

### Question #7

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

Correct Answer:

## Answer Area

Columnar format:

Avro
GZip
Parquet
TXT

JSON with a timestamp:

Avro
GZip
Parquet
TXT

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 #8

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:

Flatten hierarchy
Merge files
Preserve hierarchy

Sink file type:

CSV
JSON
Parquet
TXT

Correct Answer:

## Answer Area

Copy behavior:

Flatten hierarchy
Merge files
Preserve hierarchy

Sink file type:

CSV
JSON
Parquet
TXT

Box 1: Preserver hierarchy -

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:

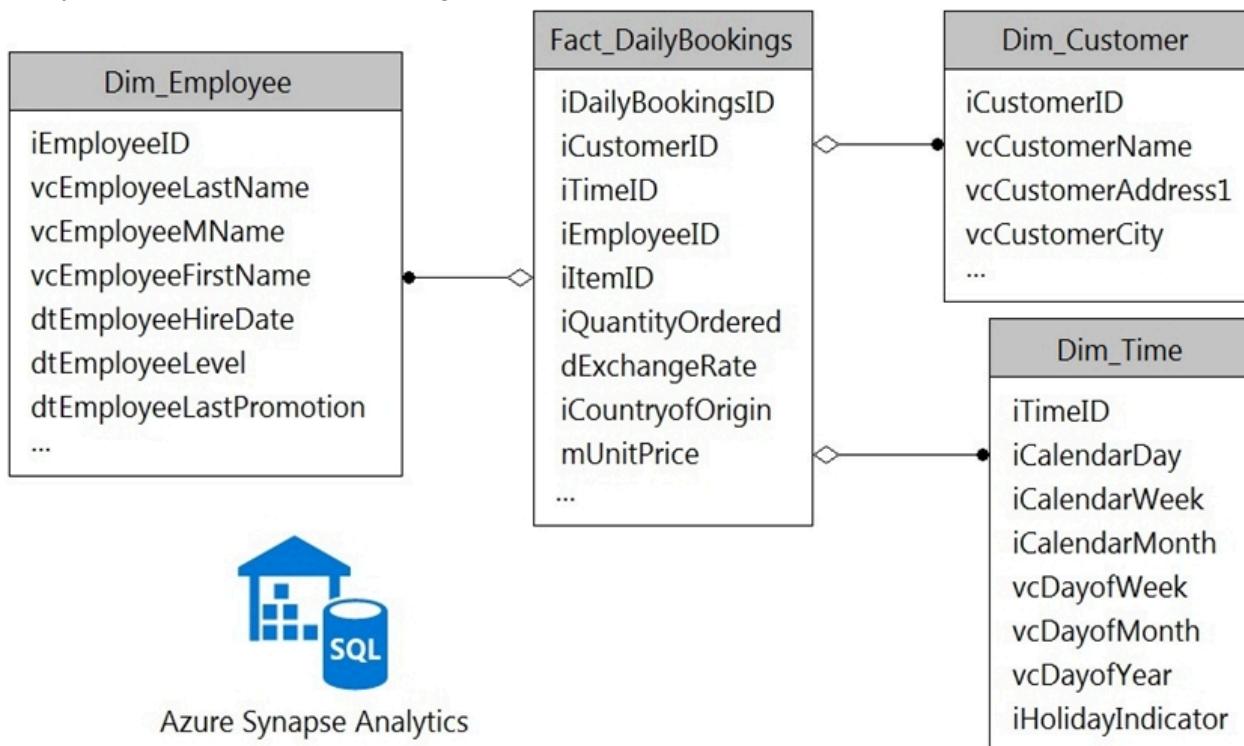
<https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-introduction>

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

## Question #9

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

**Correct Answer:**

## 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

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 #10

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.

**Correct Answer:**

## **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.

Box 1: Move to cool storage -

Box 2: Move to archive storage -

Archive - Optimized for storing data that is rarely accessed and stored for at least 180 days with flexible latency requirements, on the order of hours.

The following table shows a comparison of premium performance block blob storage, and the hot, cool, and archive access tiers.

	Premium performance	Hot tier	Cool tier	Archive tier
Availability	99.9%	99.9%	99%	Offline
Availability (RA-GRS reads)	N/A	99.99%	99.9%	Offline
Usage charges	Higher storage costs, lower access, and transaction cost	Higher storage costs, lower access, and transaction costs	Lower storage costs, higher access, and transaction costs	Lowest storage costs, highest access, and transaction costs
Minimum storage duration	N/A	N/A	30 days <sup>1</sup>	180 days
Latency (Time to first byte)	Single-digit milliseconds	milliseconds	milliseconds	hours <sup>2</sup>

Reference:

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

### Question #11

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	Answer Area
CLUSTERED INDEX	
COLLATE	
DISTRIBUTION	
PARTITION	
PARTITION FUNCTION	
PARTITION SCHEME	

```

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	Answer Area
CLUSTERED INDEX	
COLLATE	
DISTRIBUTION	
PARTITION	
PARTITION FUNCTION	
PARTITION SCHEME	

```

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

```

#### 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 #12

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?

- as a temporal table
- as a SQL graph table
- as a degenerate dimension table
- as a Type 2 slowly changing dimension (SCD) table

**Correct Answer:**

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

**Question #13**

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:**

- A. Add the managed identity to the Sales group.
- B. Use the managed identity as the credentials for the data load process.
- F. Create a managed identity.

**Question #14**

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

	<b>name</b>	<b>table_name</b>	<b>datatype</b>	<b>is_masked</b>	<b>masking_function</b>
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

**Correct Answer:**

## 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

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 #15

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;

```

**Correct Answer:**

C.

**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 #16**

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.

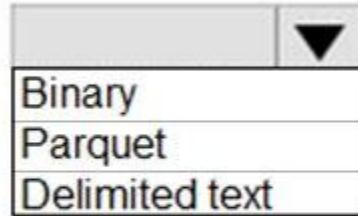
How should you configure the copy activity? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

## Answer Area

Source dataset type:



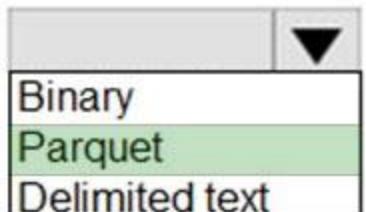
Copy activity copy behavior:



Correct Answer:

## Answer Area

Source dataset type:



Copy activity copy behavior:



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 #17

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:**

- A. geo-redundant storage (GRS)

**Question #18**

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. zone-redundant storage (ZRS)

**Question #19**

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:

Hash
Replicated
Round-robin

Indexing:

Clustered
Clustered columnstore
Heap

Partitioning:

Date
None

Correct Answer:

## Answer Area

Distribution:

Hash
Replicated
Round-robin

Indexing:

Clustered
Clustered columnstore
Heap

Partitioning:

Date
None

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 #20**

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. hash-distributed on PurchaseKey

**Question #21**

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

Correct Answer:

## Answer Area

EventCategory:

DimChannel
DimDate
DimEvent
FactEvents

ChannelGrouping:

DimChannel
DimDate
DimEvent
FactEvents

TotalEvents:

DimChannel
DimDate
DimEvent
FactEvents

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 #22

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. Yes

**Question #23**

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. No

**Question #24**

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. No

**Question #25**

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. a materialized view

**Question #26**

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. Parquet

**Question #27**

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. Azure Databricks

**Question #28**

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:**

D. Merge the files

## Question #29

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": {  
                            "daysAfterCreationGreaterThanOrEqual": 60  
                        }  
                    },  
                    "baseBlob": {  
                        "tierToCool": {  
                            "daysAfterModificationGreaterThanOrEqual":  
                                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	

### Correct Answer:

### Answer Area

The files are [answer choice] after 30 days:

	▼
deleted from the container	
moved to archive storage	
<b>moved to cool storage</b>	
moved to hot storage	

The storage policy applies to [answer choice]:

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

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 #30

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. TransactionMonth

**Question #31**

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

#### Correct Answer:

#### 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

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 #32

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. Parquet

**Question #33**

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. Switch the first partition from stg.Sales to dbo.Sales.

**Question #34**

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 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. surrogate primary key
- B. effective start date
- E. effective end date

**Question #35**

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	

#### Correct Answer:

#### 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	

#### Box 1: Denormalize to a second normal form

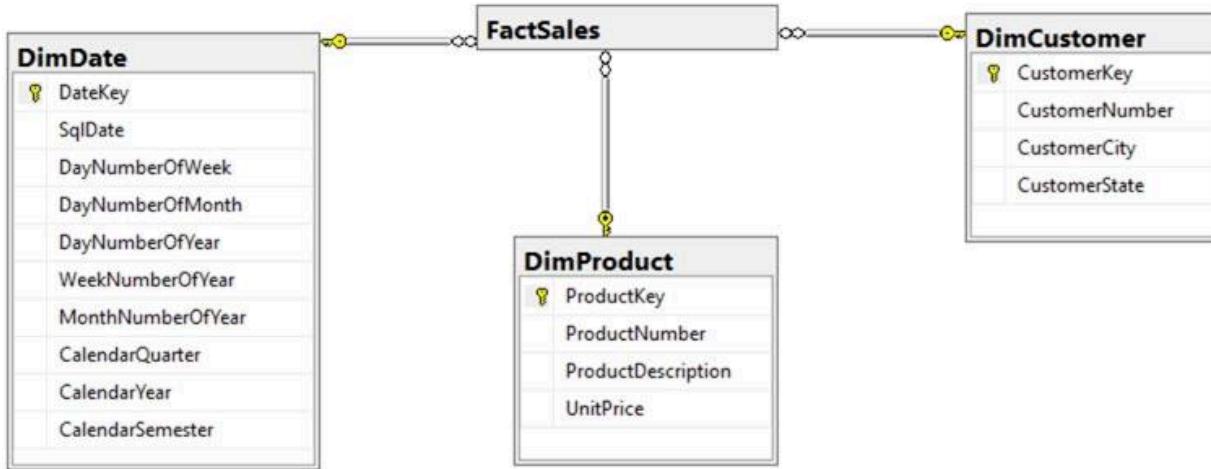
Denormalization is the process of transforming higher normal forms to lower normal forms via storing the join of higher normal form relations as a base relation.

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 flat 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 #36

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  
  
.mode("append")
```

(“*”)	▼
(“StoreID”, “Hour”)	▼
(“StoreID”, “Year”, “Month”, “Day”, “Hour”)	▼

.csv(“/Purchases”)	▼
.json(“/Purchases”)	▼
.parquet(“/Purchases”)	▼
.saveAsTable(“/Purchases”)	▼

**Correct Answer:**

**Answer Area**

```
df.write  
.bucketBy  
.partitionBy  
.range  
.sortBy  
  
.mode("append")
```

(“*”)	▼
(“StoreID”, “Hour”)	▼
(“StoreID”, “Year”, “Month”, “Day”, “Hour”)	▼

.csv(“/Purchases”)	▼
.json(“/Purchases”)	▼
<b>.parquet(“/Purchases”)</b>	▼
.saveAsTable(“/Purchases”)	▼

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 #37**

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. 40

### **Question #38**

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).

Type 0
Type 1
Type 2

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

a surrogate key
a business key
an audit column

Correct Answer:

## Answer Area

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

Type 0
Type 1
Type 2

a surrogate key
a business key
an audit column

The ProductKey column is [answer choice].

Box 1: Type 2 -

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

### Incorrect Answers:

A Type 1 SCD always reflects the latest values, and when changes in source data are detected, the dimension table data is overwritten.

Box 2: a business key -

A business key or natural key is an index which identifies uniqueness of a row based on columns that exist naturally in a table according to business rules. For example business keys are customer code in a customer table, composite of sales order header number and sales order item line number within a sales order details table.

Reference:

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

## Question #39

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. hash-distributed on PurchaseKey

**Question #40**

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 the 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:**

- A. Use Snappy compression for the files.

**Question #41**

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	Answer Area
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	

**Correct Answer:**

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

### Step 1: Create an external data source

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

1. CREATE EXTERNAL DATA SOURCE to reference an external Azure storage and specify the credential that should be used to access the storage.
2. CREATE EXTERNAL FILE FORMAT to describe format of CSV or Parquet files.
3. 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>

## Question #42

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. a dimension table for Employee
- E. a fact table for Transaction

### Question #43

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. Type 2

### Question #44

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 a dedicated SQL pool 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



## Correct Answer:

Actions	Answer Area
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
	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

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 #45

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:

### **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]           RANGE RIGHT FOR VALUES
    [TransactionDateID], [TransactionTypeID]
    HASH([TransactionTypeID])
    ROUND_ROBIN
)
(20200101,20200201,20200301,20200401,20200501,20200601)
```

### **Correct Answer:**

## 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] RANGE RIGHT FOR VALUES
    [TransactionDateID], [TransactionTypeID]
    HASH([TransactionTypeID])
    ROUND_ROBIN
)
(20200101, 20200201, 20200301, 20200401, 20200501, 20200601)
```

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 #46

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. Only CSV that have file names that beginning with "tripdata\_2020".

**Question #47**

DRAG DROP -

You use PySpark in Azure Databricks to parse the following JSON input.

```
{
  "persons": [
    {
      "name": "Keith",
      "age": 30,
      "dogs": ["Fido", "Fluffy"]
    },
    {
      "name": "Donna",
      "age": 46,
      "dogs": ["Spot"]
    }
  ]
}
```

You need to output the data in the following tabular format.

<b>owner</b>	<b>age</b>	<b>dog</b>
Keith	30	Fido
Keith	30	Fluffy
Donna	46	Spot

How should you complete the PySpark 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	Answer Area
alias	dbutils.fs.put("/tmp/source.json", source_json, True) source_df = spark.read.option("multiline", "true").json("/tmp/source.json")
array_union	persons = source_df. <input type="text"/> Value <input type="text"/> Value ("persons").alias("persons")
createDataFrame	persons_dogs = persons.select(col("persons.name").alias("owner"), col("persons.age").alias("age"), explode <input type="text"/> Value ("dog"))
explode	("persons-dogs"). display(persons_dogs)
select	
translate	

**Correct Answer:**

Values	Answer Area
	dbutils.fs.put("/tmp/source.json", source_json, True) source_df = spark.read.option("multiline", "true").json("/tmp/source.json") persons = source_df. select [ ] explode [ ] ("persons").alias("persons")) persons_dogs = persons.select(col("persons.name").alias("owner"), col("persons.age").alias("age"), explode alias [ ] ("dog")) display(persons_dogs)
array_union	
createDataFrame	
translate	

Box 1: select -

Box 2: explode -

Box 3: alias -

pyspark.sql.Column.alias returns this column aliased with a new name or names (in the case of expressions that return more than one column, such as explode).

Reference:

<https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.sql.Column.alias.html>

<https://docs.microsoft.com/en-us/azure/databricks/sql/language-manual/functions/explode>

## Question #48

HOTSPOT -

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

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

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

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

NOTE: Each correct selection is worth one point.

Hot Area:

## **Answer Area**

First week:

Archive
Cool
Hot

After one month:

Archive
Cool
Hot

After one year:

Archive
Cool
Hot

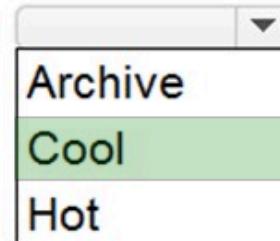
**Correct Answer:**

## Answer Area

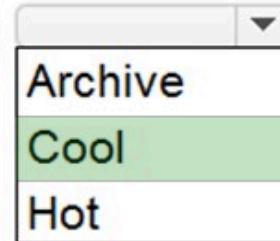
First week:



After one month:



After one year:



Box 1: Hot -

Hot tier - An online tier optimized for storing data that is accessed or modified frequently. The Hot tier has the highest storage costs, but the lowest access costs.

Box 2: Cool -

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.

Box 3: Cool -

Not 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.

	Premium performance	Hot tier	Cool tier	Archive tier
Availability	99.9%	99.9%	99%	Offline
Availability (RA-GRS reads)	N/A	99.99%	99.9%	Offline
Usage charges	Higher storage costs, lower access, and transaction cost	Higher storage costs, lower access, and transaction costs	Lower storage costs, higher access, and transaction costs	Lowest storage costs, highest access, and transaction costs
Minimum object size	N/A	N/A	N/A	N/A
Minimum storage duration	N/A	N/A	30 days <sup>1</sup>	180 days
Latency (Time to first byte)	Single-digit milliseconds	milliseconds	milliseconds	hours <sup>2</sup>

Reference:

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

<https://www.altaro.com/hyper-v/azure-archive-storage/>

### Question #49

You have an Azure Synapse Analytics Apache Spark pool named Pool1.

You plan to load JSON files from an Azure Data Lake Storage Gen2 container into the tables in Pool1. The structure and data types vary by file.

You need to load the files into the tables. The solution must maintain the source data types.

What should you do?

- A. Use a Conditional Split transformation in an Azure Synapse data flow.
- B. Use a Get Metadata activity in Azure Data Factory.
- C. Load the data by using the OPENROWSET Transact-SQL command in an Azure Synapse Analytics serverless SQL pool.
- D. Load the data by using PySpark.

### Correct Answer:

- D. Load the data by using PySpark.

### Question #50

You have an Azure Databricks workspace named workspace1 in the Standard pricing tier.

Workspace1 contains an all-purpose cluster named cluster1.

You need to reduce the time it takes for cluster1 to start and scale up. The solution must minimize costs.

What should you do first?

- A. Configure a global init script for workspace1.
- B. Create a cluster policy in workspace1.
- C. Upgrade workspace1 to the Premium pricing tier.
- D. Create a pool in workspace1.

**Correct Answer:**

- D. Create a pool in workspace1.

### **Question #51**

HOTSPOT -

You are building an Azure Stream Analytics job that queries reference data from a product catalog file. The file is updated daily.

The reference data input details for the file are shown in the Input exhibit. (Click the Input tab.)

## Input Details

X

products

 Test  Delete

Container

Create new  Use existing

refdata

Path pattern ①

product.csv

Date format

YYYY/MM/DD

Time format

HH

Event serialization format \* ①

CSV

Delimiter ①

comma (,)

Encoding ①

UTF-8

**Save**

- ① If the chosen resource and the stream analytics job are located in different regions, you will be billed to move data between regions.

The storage account container view is shown in the Refdata exhibit. (Click the Refdata tab.)

**refdata**  
Container

Search (Ctrl + /) « Upload Add Directory Refresh | Rename Delete

Overview Access Control (IAM)

**Settings**

- Access policy
- Properties
- Metadata

**Authentication method:** Access key ([Switch to Azure AD User Account](#))  
**Location:** refdata / 2020-03-20

Search blobs by prefix (case-sensitive)

Name
<input type="checkbox"/> [..]
<input type="checkbox"/> product.csv

You need to configure the Stream Analytics job to pick up the new reference data.

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

NOTE: Each correct selection is worth one point.

Hot Area:

## Answer Area

Path pattern:

{date}/product.csv
{date}/{time}/product.csv
product.csv
*/product.csv

Date format:

MM/DD/YYYY
YYYY/MM/DD
YYYY-DD-MM
YYYY-MM-DD

**Correct Answer:**

## Answer Area

Path pattern:

{date}/product.csv
{date}/{time}/product.csv
product.csv
*/product.csv

Date format:

MM/DD/YYYY
YYYY/MM/DD
YYYY-DD-MM
YYYY-MM-DD

Box 1: {date}/product.csv -

In the 2nd exhibit we see: Location: refdata / 2020-03-20

Note: Path Pattern: This is a required property that is used to locate your blobs within the specified container. Within the path, you may choose to specify one or more instances of the following 2 variables:

{date}, {time}

Example 1: products/{date}/{time}/product-list.csv

Example 2: products/{date}/product-list.csv

Example 3: product-list.csv -

Box 2: YYYY-MM-DD -

Note: Date Format [optional]: If you have used {date} within the Path Pattern that you specified, then you can select the date format in which your blobs are organized from the drop-down of supported formats.

Example: YYYY/MM/DD, MM/DD/YYYY, etc.

Reference:

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-use-reference-data>

## Question #52

HOTSPOT -

You have the following Azure Stream Analytics query.

WITH

```
step1 AS (SELECT *
           FROM input1
           PARTITION BY StateID
           INTO 10),
step2 AS (SELECT *
           FROM input2
           PARTITION BY StateID
           INTO 10)

SELECT *
INTO output
FROM step1
PARTITION BY StateID
UNION
SELECT * INTO output
           FROM step2
           PARTITION BY StateID
```

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 combines two streams of partitioned data.	<input type="radio"/>	<input type="radio"/>
The stream scheme key and count must match the output scheme.	<input type="radio"/>	<input type="radio"/>
Providing 60 streaming units will optimize the performance of the query.	<input type="radio"/>	<input type="radio"/>

Correct Answer:

## Answer Area

Statements	Yes	No
The query combines two streams of partitioned data.	<input type="radio"/>	<input checked="" type="radio"/>
The stream scheme key and count must match the output scheme.	<input checked="" type="radio"/>	<input type="radio"/>
Providing 60 streaming units will optimize the performance of the query.	<input checked="" type="radio"/>	<input type="radio"/>

Box 1: No -

Note: You can now use a new extension of Azure Stream Analytics SQL to specify the number of partitions of a stream when reshuffling the data.

The outcome is a stream that has the same partition scheme. Please see below for an example:  
WITH step1 AS (SELECT \* FROM [input1] PARTITION BY DeviceID INTO 10), step2 AS  
(SELECT \* FROM [input2] PARTITION BY DeviceID INTO 10)  
SELECT \* INTO [output] FROM step1 PARTITION BY DeviceID UNION step2 PARTITION BY  
DeviceID

Note: The new extension of Azure Stream Analytics SQL includes a keyword INTO that allows you to specify the number of partitions for a stream when performing reshuffling using a PARTITION BY statement.

Box 2: Yes -

When joining two streams of data explicitly repartitioned, these streams must have the same partition key and partition count.

Box 3: Yes -

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.

In general, the best practice is to start with 6 SUs for queries that don't use PARTITION BY. Here there are 10 partitions, so  $6 \times 10 = 60$  SUs is good.

Note: Remember, Streaming Unit (SU) count, which is the unit of scale for Azure Stream Analytics, must be adjusted so the number of physical resources available to the job can fit the partitioned flow. In general, six SUs is a good number to assign to each partition. In case there are insufficient resources assigned to the job, the system will only apply the repartition if it benefits the job.

Reference:

<https://azure.microsoft.com/en-in/blog/maximize-throughput-with-repartitioning-in-azure-stream-analytics/>

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-streaming-unit-consumption>

## Question #53

HOTSPOT -

You are building a database in an Azure Synapse Analytics serverless SQL pool.

You have data stored in Parquet files in an Azure Data Lake Storege Gen2 container.

Records are structured as shown in the following sample.

```
{  
    "id": 123,  
    "address_housenumber": "19c",  
    "address_line": "Memory Lane",  
    "applicant1_name": "Jane",  
    "applicant2_name": "Dev"  
}
```

The records contain two applicants at most.

You need to build a table that includes only the address fields.

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:

### Answer Area

```
applications  
CREATE EXTERNAL TABLE  
CREATE TABLE  
CREATE VIEW  
WITH (  
    LOCATION = 'applications/',  
    DATA_SOURCE = applications_ds,  
    FILE_FORMAT = applications_file_format  
)  
AS  
SELECT id, [address_housenumber] as addresshousenumber, [address_line1] as addressline1  
FROM  
    (BULK 'https://contoso1.dfs.core.windows.net/applications/year=*/*.parquet',  
CROSS APPLY  
OPENJSON  
OPENROWSET  
FORMAT='PARQUET') AS [r]  
GO
```

**Correct Answer:**

## Answer Area

```
applications
CREATE EXTERNAL TABLE
CREATE TABLE
CREATE VIEW
WITH (
    LOCATION = 'applications/',
    DATA_SOURCE = applications_ds,
    FILE_FORMAT = applications_file_format
)
AS
SELECT id, [address_housenumber] as addresshousenumber, [address_line1] as addressline1
FROM
(BULK 'https://contosol1.dfs.core.windows.net/applications/year=/*/*.parquet',
CROSS APPLY
OPENJSON
OPENROWSET
FORMAT='PARQUET') AS [r]
GO
```

### Box 1: CREATE EXTERNAL TABLE -

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.

Syntax:

```
CREATE EXTERNAL TABLE { database_name.schema_name.table_name |
schema_name.table_name | table_name }
( <column_definition> [ ,...n ] )
WITH (
LOCATION = 'folder_or_filepath',
DATA_SOURCE = external_data_source_name,
FILE_FORMAT = external_file_format_name
```

### Box 2. OPENROWSET -

When using serverless SQL pool, CETAS is used to create an external table and export query results to Azure Storage Blob or Azure Data Lake Storage Gen2.

Example:

AS -

```
SELECT decennialTime, stateName, SUM(population) AS population
```

FROM -

```
OPENROWSET(BULK
'https://azureopendatastorage.blob.core.windows.net/censusdatacontainer/release/us_population_county/year=/*/*.parquet',
FORMAT='PARQUET') AS [r]
GROUP BY decennialTime, stateName
```

GO -

Reference:

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

### Question #54

HOTSPOT -

You have an Azure Synapse Analytics dedicated SQL pool named Pool1 and an Azure Data Lake Storage Gen2 account named Account1.

You plan to access the files in Account1 by using an external table.

You need to create a data source in Pool1 that you can reference when you create the external table.

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:

### Answer Area

```
CREATE EXTERNAL DATA SOURCE source1
WITH
    ( LOCATION = 'https://account1.core.windows.net',
      

|                     |
|---------------------|
| PUSHDOWN = ON       |
| TYPE = BLOB_STORAGE |
| TYPE = HADOOP       |


)
```

Correct Answer:

### Answer Area

```
CREATE EXTERNAL DATA SOURCE source1
WITH
    ( LOCATION = 'https://account1.core.windows.net',
      

|                     |
|---------------------|
| PUSHDOWN = ON       |
| TYPE = BLOB_STORAGE |
| TYPE = HADOOP       |


)
```

Box 1: blob -

The following example creates an external data source for Azure Data Lake Gen2

CREATE EXTERNAL DATA SOURCE YellowTaxi

WITH ( LOCATION = 'https://azureopendatastorage.blob.core.windows.net/nyctlc/yellow/' ,

TYPE = HADOOP)

Box 2: HADOOP -

Reference:

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

### Question #55

You have an Azure subscription that contains an Azure Blob Storage account named storage1 and an Azure Synapse Analytics dedicated SQL pool named Pool1.

You need to store data in storage1. The data will be read by Pool1. The solution must meet the following requirements:

Enable Pool1 to skip columns and rows that are unnecessary in a query.

- ☞ Automatically create column statistics.
- ☞ Minimize the size of files.

Which type of file should you use?

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

### Correct Answer:

B. Parquet

### Question #56

DRAG DROP -

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

Data in the table will be retained for five years. Once a year, data that is older than five years will be deleted.

You need to ensure that the data is distributed evenly across partitions. The solution must minimize the amount of time required to delete old data.

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

NOTE: Each correct selection is worth one point.

Select and Place:

Values	Answer Area
CustomerKey	
HASH	
ROUND_ROBIN	
REPLICATE	
OrderDateKey	
SalesOrderNumber	

```

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

```

### Correct Answer:

Values	Answer Area
CustomerKey	
ROUND_ROBIN	
REPLICATE	
SalesOrderNumber	

```

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

```

Box 1: HASH -

Box 2: OrderDateKey -

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

Reference:

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

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

## Question #57

HOTSPOT -

You have an 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 options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

## **Answer Area**

Data over five years old:

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

Data over seven years old:

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

**Correct Answer:**

## **Answer Area**

Data over five years old:

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

Data over seven years old:

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

Box 1: Move to cool storage -

Box 2: Move to archive storage -

Archive - Optimized for storing data that is rarely accessed and stored for at least 180 days with flexible latency requirements, on the order of hours.

The following table shows a comparison of premium performance block blob storage, and the hot, cool, and archive access tiers.

	Premium performance	Hot tier	Cool tier	Archive tier
Availability	99.9%	99.9%	99%	Offline
Availability (RA-GRS reads)	N/A	99.99%	99.9%	Offline
Usage charges	Higher storage costs, lower access, and transaction cost	Higher storage costs, lower access, and transaction costs	Lower storage costs, higher access, and transaction costs	Lowest storage costs, highest access, and transaction costs
Minimum storage duration	N/A	N/A	30 days <sup>1</sup>	180 days
Latency (Time to first byte)	Single-digit milliseconds	milliseconds	milliseconds	hours <sup>2</sup>

Reference:

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

## Question #58

HOTSPOT -

You plan to create an Azure Data Lake Storage Gen2 account.

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

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

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

NOTE: Each correct selection is worth one point.

Hot Area:

**Answer Area**

Replication mechanism:

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

Failover process:

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

**Correct Answer:**

**Answer Area**

Replication mechanism:

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

Failover process:

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

**Reference:**

<https://docs.microsoft.com/en-us/azure/storage/common/storage-disaster-recovery-guidance?toc=/azure/storage/blobs/toc.json>

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

**Question #59**

You need to implement a Type 3 slowly changing dimension (SCD) for product category data in an Azure Synapse Analytics dedicated SQL pool.

You have a table that was created by using the following Transact-SQL statement.

```
CREATE TABLE [DB0].[DimProduct] (
    [ProductKey] [int] IDENTITY(1,1) NOT NULL,
    [ProductSourceID] [int] NOT NULL,
    [ProductName] [nvarchar](100) NOT NULL,
    [Color] [nvarchar](15) NULL,
    [SellStartDate] [date] NOT NULL,
    [SellEndDate] [date] NULL,
    [RowInsertedDateTime] [datetime] NOT NULL,
    [RowUpdatedDateTime] [datetime] NOT NULL,
    [ETLAuditID] [int] NOT NULL
)
```

Which two columns should you add to the table? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

A.

[EffectiveEndDate] [datetime] NULL,

B.

[CurrentProductCategory] [nvarchar](100) NOT NULL,

C.

[ProductCategory] [nvarchar](100) NOT NULL,

D.

[EffectiveStartDate] [datetime] NOT NULL,

E.

[OriginalProductCategory] [nvarchar](100) NOT NULL,

**Correct Answer:**

B,E

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.

This type of tracking may be used for one or two columns in a dimension table. It is not common to use it for many members of the same table. It is often used in combination with Type 1 or Type 2 members.

CustomerID	FirstName	LastName	CurrentEmail	OriginalEmail	CompanyName	InsertedDate	ModifiedDate
2	Keith	Harris	keith0@aw.com	keith0@aw.com	Progressive Sports	2021-03-20	2021-03-20
3	Donna	Carreras	donna0@aw.com	donna0@aw.com	A Bike Store	2021-03-20	2021-03-20

CustomerID	FirstName	LastName	CurrentEmail	OriginalEmail	CompanyName	InsertedDate	ModifiedDate
2	Keith	Harris	keith0@aw.com	keith0@aw.com	Progressive Sports	2021-03-20	2021-03-20
3	Donna	Carreras	dc3@aw.com	donna0@aw.com	A Bike Store	2021-03-20	2021-03-22

Reference:

<https://k21academy.com/microsoft-azure/azure-data-engineer-dp203-q-a-day-2-live-session-review/>

## Question #60

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.

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

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. 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
<input type="checkbox"/> Hash	<input type="text"/> Common.Data:
<input type="checkbox"/> Replicated	<input type="text"/> Marketing.Web.Sessions:
<input type="checkbox"/> Round-robin	<input type="text"/> Staging. Web.Sessions:
⋮	

**Correct Answer:**

Table distribution types	Answer Area
Hash	Common.Data: Replicated
Replicated	Marketing.Web.Sessions: Hash
Round-robin	Staging. Web.Sessions: Round-robin

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-data-warehouse-tables-distribute>

## Question #61

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

**Correct Answer:**

## 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
```

Box 1: (CLUSTERED COLUMNSTORE INDEX

### CLUSTERED COLUMNSTORE INDEX -

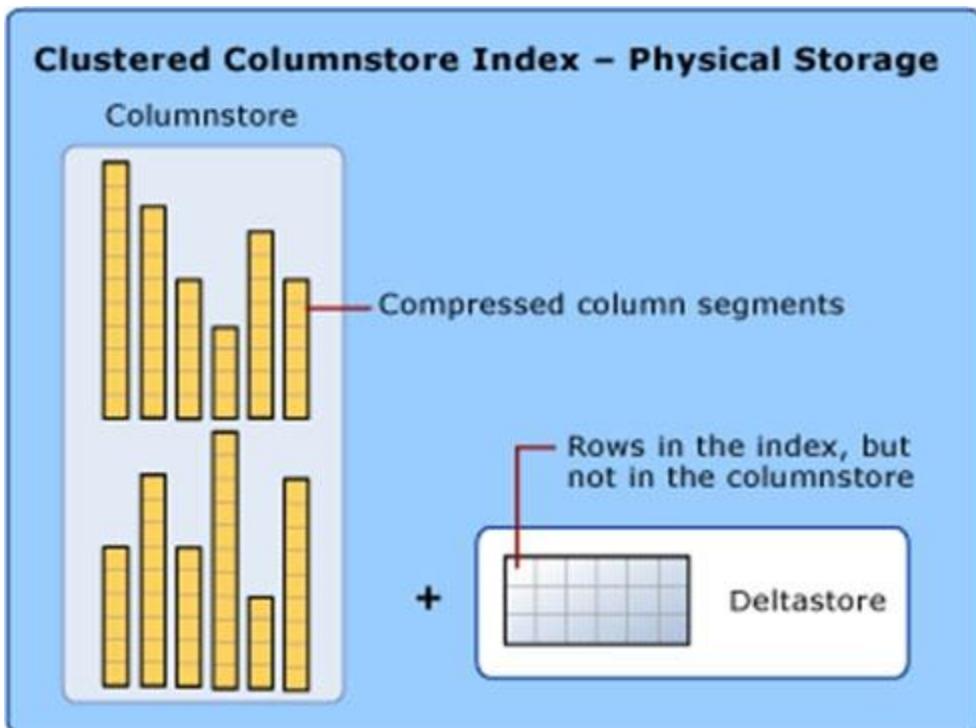
Columnstore indexes are the standard for storing and querying large data warehousing fact tables. This index uses column-based data storage and query processing to achieve gains up to 10 times the query performance in your data warehouse over traditional row-oriented storage.

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

Incorrect:

- \* Not HASH([OrderDateKey]). Is not 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

- \* A replicated table has a full copy of the table available on every Compute node. Queries run fast on replicated tables since joins on replicated tables don't require data movement.

Replication requires extra storage, though, and isn't practical for large tables.

- \* A round-robin table distributes table rows evenly across all distributions. The rows are distributed randomly. Loading data into a round-robin table is fast. Keep in mind that queries can require more data movement than the other distribution methods.

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 #62**

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:**

A. once per month

### **Question #63**

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. ALTER
- D. MERGE

### **Correct Answer:**

D. MERGE

### **Question #64**

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. Apache Parquet

**Question #65**

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 less than 1 MB.

Does this meet the goal?

- A. Yes
- B. No

**Correct Answer:**

A. Yes

**Question #66**

You plan to create a dimension table in Azure Synapse Analytics that will be less than 1 GB.

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

- ⇒ Provide the fastest query time.
- ⇒ Minimize data movement during queries.

Which type of table should you use?

- A. replicated
- B. hash distributed
- C. heap
- D. round-robin

**Correct Answer:**

A. replicated

**Question #67**

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:**

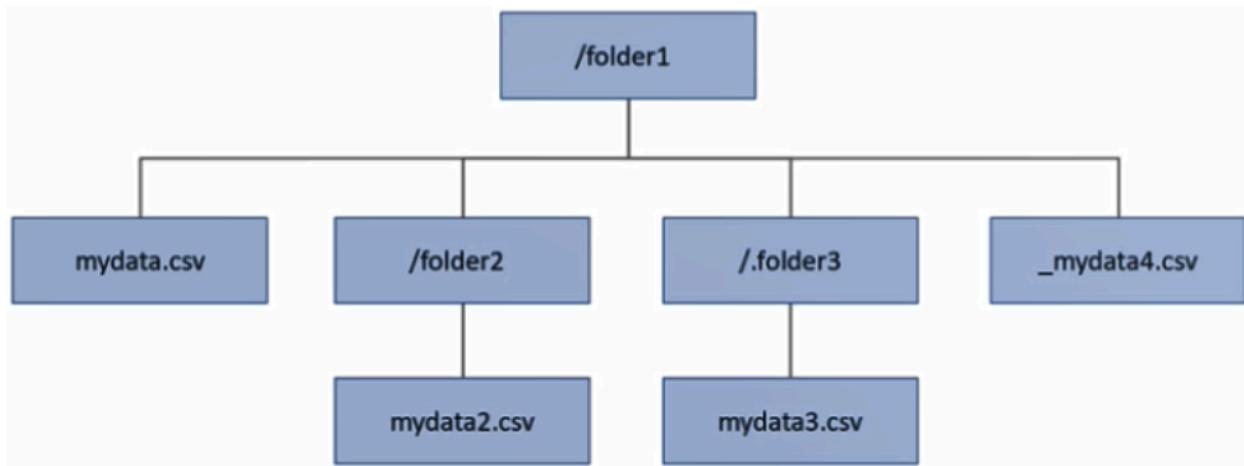
C. an IDENTITY column

**Question #68**

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.

**Answer 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"/>

**Correct Answer:**

**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="checkbox"/>	<input type="checkbox"/>
When selecting all the rows in dbo.Table1, data from the mydata3.csv file will be returned.	<input checked="" type="checkbox"/>	<input type="checkbox"/>
When selecting all the rows in dbo.Table1, data from the _mydata4.csv file will be returned.	<input type="checkbox"/>	<input checked="" type="checkbox"/>

**Question #69**

You have an Azure Synapse Analytics dedicated SQL pool.

You need to create a fact table named Table1 that will store sales data from the last three years. The solution must be optimized for the following query operations:

- Show order counts by week.
- Calculate sales totals by region.
- Calculate sales totals by product.
- Find all the orders from a given month.

Which data should you use to partition Table1?

- A. product
- B. month
- C. week
- D. region

**Correct Answer:**

B. month

**Question #70**

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 pools.
- 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. \DataSource\SubjectArea\YYYY\WW\FileDialog\_YYYY\_MM\_DD.parquet
- B. \DataSource\SubjectArea\YYYY-WW\FileDialog\_YYYY\_MM\_DD.parquet
- C. DataSource\SubjectArea\WW\YYYY\FileDialog\_YYYY\_MM\_DD.parquet
- D. \YYYY\WW\DataSource\SubjectArea\FileDialog\_YYYY\_MM\_DD.parquet
- E. WW\YYYY\SubjectArea\DataSource\FileDialog\_YYYY\_MM\_DD.parquet

**Correct Answer:**

A. \DataSource\SubjectArea\YYYY\WW\FileDialog\_YYYY\_MM\_DD.parquet

**Question #71**

You have an Azure Synapse Analytics dedicated SQL pool named Pool1. Pool1 contains a table named table1.

You load 5 TB of data into table1.

You need to ensure that columnstore compression is maximized for table1.

Which statement should you execute?

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

**Correct Answer:**

D. ALTER INDEX ALL on table1 REBUILD

**Question #72**

You have an Azure Synapse Analytics dedicated SQL pool named pool1.

You plan to implement a star schema in pool 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:**

- B. [EffectiveEndDate] [datetime] NOT NULL
- E. [EffectiveStartDate] [datetime] NOT NULL

**Question #73**

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.

## 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])
        HASH([OrderDate])
        REPLICATE
        ROUND_ROBIN
    CLUSTERED COLUMNSTORE INDEX
)
CREATE TABLE [dbo].[Country]
(
    [CountryId] int NOT NULL
    , [CountryCode] varchar(10) NOT NULL
)
WITH
(
    DISTRIBUTION =
        HASH([CountryCode])
        HASH([CountryId])
        REPLICATE
        ROUND_ROBIN
    CLUSTERED COLUMNSTORE INDEX
)
```

**Correct Answer:**

## 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])
    HASH([OrderDate])
    REPLICATE
    ROUND_ROBIN
    CLUSTERED COLUMNSTORE INDEX
)
CREATE TABLE [dbo].[Country]
(
    [CountryId] int NOT NULL
    , [CountryCode] varchar(10) NOT NULL
)
WITH
(
    DISTRIBUTION = HASH([CountryCode])
    HASH([CountryId])
    REPLICATE
    ROUND_ROBIN
    CLUSTERED COLUMNSTORE INDEX
)
```

**Question #74**

You have an Azure subscription that contains an Azure Data Lake Storage Gen2 account named account1 and an Azure Synapse Analytics workspace named workspace1.

You need to create an external table in a serverless SQL pool in workspace1. The external table will reference CSV files stored in account1. The solution must maximize performance.

How should you configure the external table?

- A. Use a native external table and authenticate by using a shared access signature (SAS).
- B. Use a native external table and authenticate by using a storage account key.
- C. Use an Apache Hadoop external table and authenticate by using a shared access signature (SAS).
- D. Use an Apache Hadoop external table and authenticate by using a service principal in Microsoft Azure Active Directory (Azure AD), part of Microsoft Entra.

**Correct Answer:**

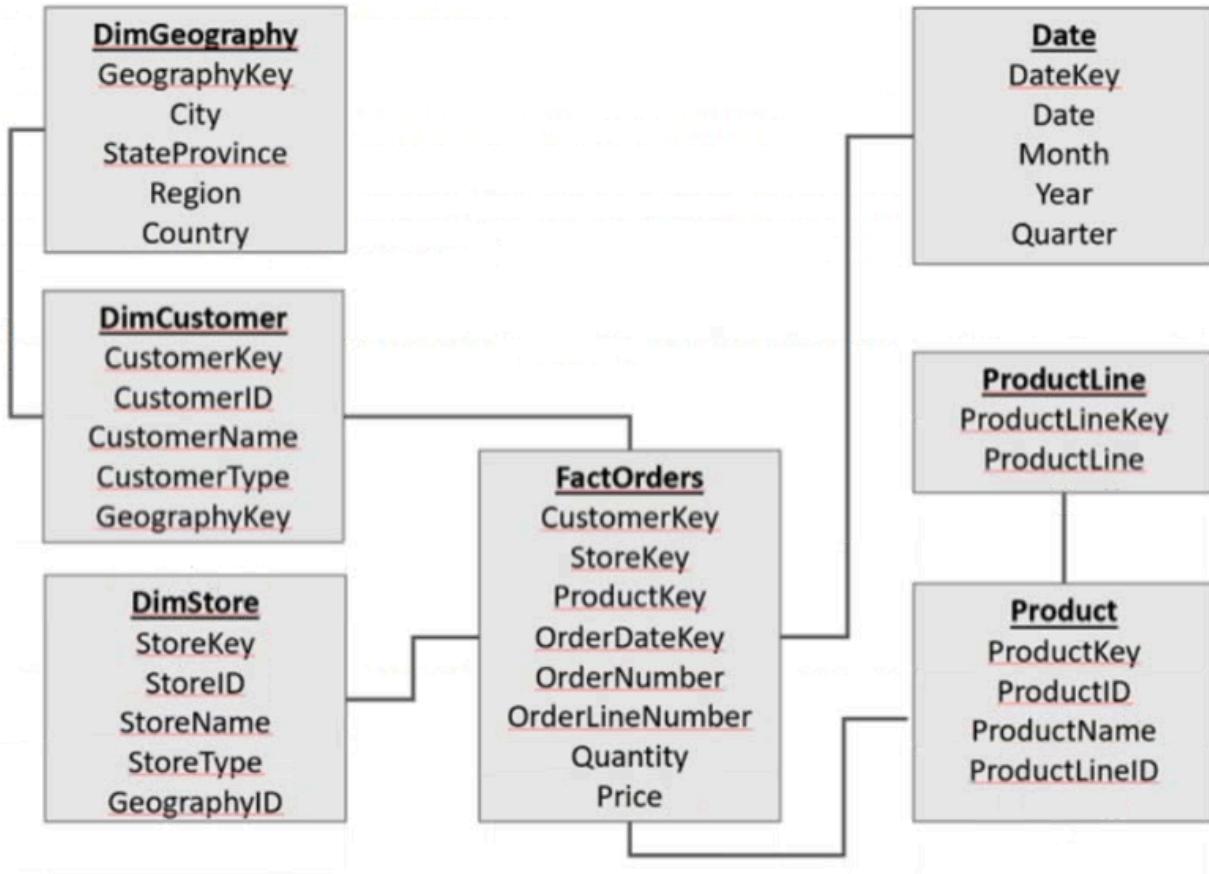
- A. Use a native external table and authenticate by using a shared access signature (SAS).

**Question #75**

HOTSPOT

-

You have an Azure Synapse Analytics serverless SQL pool that contains a database named db1. The data model for db1 is 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 exhibit.

NOTE: Each correct selection is worth one point.

**Answer Area**

To convert the data model to a star schema, [answer choice].

join DimGeography and DimCustomer  
 join DimGeography and FactOrders  
 union DimGeography and DimCustomer  
 union DimGeography and FactOrders

Once the data model is converted into a star schema, there will be [answer choice] tables.

4  
5  
6  
7

**Correct Answer:**

#### Answer Area

To convert the data model to a star schema, [answer choice].

join DimGeography and DimCustomer  
join DimGeography and FactOrders  
union DimGeography and DimCustomer  
union DimGeography and FactOrders

Once the data model is converted into a star schema, there will be [answer choice] tables.

4  
5  
**6**  
7

#### Question #76

You have an Azure Databricks workspace and an Azure Data Lake Storage Gen2 account named storage1.

New files are uploaded daily to storage1.

You need to recommend a solution that configures storage1 as a structured streaming source. The solution must meet the following requirements:

- Incrementally process new files as they are uploaded to storage1.
- Minimize implementation and maintenance effort.
- Minimize the cost of processing millions of files.
- Support schema inference and schema drift.

Which should you include in the recommendation?

- A. COPY INTO
- B. Azure Data Factory
- C. Auto Loader
- D. Apache Spark FileStreamSource

#### Correct Answer:

- C. Auto Loader

#### Question #77

You have an Azure subscription that contains the resources shown in the following table.

Name	Type	Description
storage1	Azure Blob storage account	Contains publicly accessible TSV files that do <b>NOT</b> have a header row
WS1	Azure Synapse Analytics workspace	Contains a serverless SQL pool

You need to read the TSV files by using ad-hoc queries and the OPENROWSET function. The solution must assign a name and override the inferred data type of each column.

What should you include in the OPENROWSET function?

- A. the WITH clause
- B. the ROWSET\_OPTIONS bulk option
- C. the DATAFILETYPE bulk option
- D. the DATA\_SOURCE parameter

**Correct Answer:**

- A. the WITH clause

**Question #78**

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:**

- D. 60 million

**Question #79**

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

- RepSourceID

- SalesRepID
- FirstName
- LastName
- StartDate
- EndDate
- Region

You are developing an Azure Synapse Analytics pipeline that includes a mapping data flow named Dataflow1. Dataflow1 will read sales team data from an external source and use a Type 2 slowly changing dimension (SCD) when loading the data into DimSalesPerson.

You need to update the last name of a salesperson in DimSalesPerson.

Which two actions should you perform? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. Update three columns of an existing row.
- B. Update two columns of an existing row.
- C. Insert an extra row.
- D. Update one column of an existing row.

**Correct Answer:**

- C. Insert an extra row.
- D. Update one column of an existing row.

**Question #80**

HOTSPOT

-

You plan to use an Azure Data Lake Storage Gen2 account to implement a Data Lake development environment that meets the following requirements:

- Read and write access to data must be maintained if an availability zone becomes unavailable.
- Data that was last modified more than two years ago must be deleted automatically.
- Costs must be minimized.

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

NOTE: Each correct selection is worth one point.

### **Answer Area**

For storage redundancy:

- Geo-zone-redundant storage (GZRS)
- Locally-redundant storage (LRS)
- Zone-redundant storage (ZRS)

For data deletion:

- A lifecycle management policy
- Soft delete
- Versioning

### **Correct Answer:**

### **Answer Area**

For storage redundancy:

- Geo-zone-redundant storage (GZRS)
- Locally-redundant storage (LRS)
- Zone-redundant storage (ZRS)

For data deletion:

- A lifecycle management policy
- Soft delete
- Versioning

### **Question #81**

HOTSPOT

-

You are designing an Azure Data Lake Storage Gen2 container to store data for the human resources (HR) department and the operations department at your company.

You have the following data access requirements:

- After initial processing, the HR department data will be retained for seven years and rarely accessed.
- The operations department data will be accessed frequently for the first six months, and then accessed once per month.

You need to design a data retention solution to meet the access requirements. The solution must minimize storage costs.

What should you include in the storage policy for each department? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

**Answer Area**

HR:

- Archive storage after one day and delete storage after 2,555 days.
- Archive storage after 2,555 days.
- Cool storage after one day.
- Cool storage after 180 days.
- Cool storage after 180 days and delete storage after 2,555 days.
- Delete after one day.
- Delete after 180 days.

Operations:

- Archive storage after one day and delete storage after 2,555 days.
- Archive storage after 2,555 days.
- Cool storage after one day.
- Cool storage after 180 days.
- Cool storage after 180 days and delete storage after 2,555 days.
- Delete after one day.
- Delete after 180 days.

**Correct Answer:**

## Answer Area

HR:

Archive storage after one day and delete storage after 2,555 days.

Archive storage after 2,555 days.

Cool storage after one day.

Cool storage after 180 days.

Cool storage after 180 days and delete storage after 2,555 days.

Delete after one day.

Delete after 180 days.

Operations:

Archive storage after one day and delete storage after 2,555 days.

Archive storage after 2,555 days.

Cool storage after one day.

Cool storage after 180 days.

Cool storage after 180 days and delete storage after 2,555 days.

Delete after one day.

Delete after 180 days.

## Question #82

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 DimCustomer 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.

#### Answer Area

Detect whether the data of a given customer has changed in the DimCustomer table:

Aggregate  
Derived column  
Surrogate key

Perform an upsert to the DimCustomer table:

Alter row  
Assert  
Cast

#### Correct Answer:

#### Answer Area

Detect whether the data of a given customer has changed in the DimCustomer table:

Aggregate  
**Derived column**  
Surrogate key

Perform an upsert to the DimCustomer table:

**Alter row**  
Assert  
Cast

#### Question #83

DRAG DROP

You have an Azure Synapse Analytics serverless SQL pool.

You have an Azure Data Lake Storage account named adls1 that contains a public container named container1. The container1 container contains a folder named folder1.

You need to query the top 100 rows of all the CSV files in folder1.

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

NOTE: Each correct selection is worth one point

Values	Answer Area
BULK	SELECT TOP 100 *
DATA_SOURCE	FROM [ ] (
LOCATION	[ ] 'https://adls1.dfs.core.windows.net/container1/folder1/*.csv',
OPENROWSET	FORMAT = 'CSV') AS rows

**Correct Answer:**

**Answer Area**

```
SELECT TOP 100 *
  FROM [OPENROWSET] (
    BULK [ ] 'https://adls1.dfs.core.windows.net/container1/folder1/*.csv',
    FORMAT = 'CSV') AS rows
```

**Question #84**

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. Parquet
- B. ORC
- C. JSON
- D. HIVE

**Correct Answer:**

A. Parquet

**Question #84**

You have an Azure Data Lake Storage Gen2 account named storage1.

You plan to implement query acceleration for storage1.

Which two file types support query acceleration? Each correct answer presents a complete solution.

NOTE: Each correct selection is worth one point.

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

**Correct Answer:**

- A. JSON
- D. CSV

**Question #86**

You have an Azure subscription that contains the resources shown in the following table.

Name	Type	Description
storage1	Azure Blob storage account	Contains publicly accessible JSON files
WS1	Azure Synapse Analytics workspace	Contains a serverless SQL pool

You need to read the files in storage1 by using ad-hoc queries and the OPENROWSET function. The solution must ensure that each rowset contains a single JSON record.

To what should you set the FORMAT option of the OPENROWSET function?

- A. JSON
- B. DELTA
- C. PARQUET
- D. CSV

**Correct Answer:**

- D. CSV

**Question #87**

HOTSPOT

-

You have an Azure subscription that contains the Azure Synapse Analytics workspaces shown in the following table.

Name	Primary storage account
workspace1	datalake1
workspace2	datalake2
workspace3	datalake1

Each workspace must read and write data to datalake1.

Each workspace contains an unused Apache Spark pool.

You plan to configure each Spark pool to share catalog objects that reference datalake1.

For each of the following statements, select Yes if the statement is true. Otherwise, select No.

NOTE: Each correct selection is worth one point.

### Answer Area

Statements	Yes	No
The shared catalog objects can be stored in Azure Database for MySQL.	<input type="radio"/>	<input type="radio"/>
For the Apache Hive Metastore of each workspace, you must configure a linked service that uses user-password authentication.	<input type="radio"/>	<input type="radio"/>
The users of workspace1 must be assigned the Storage Blob Contributor role for datalake1.	<input type="radio"/>	<input type="radio"/>

Correct Answer:

### Answer Area

Statements	Yes	No
The shared catalog objects can be stored in Azure Database for MySQL.	<input checked="" type="radio"/>	<input type="radio"/>
For the Apache Hive Metastore of each workspace, you must configure a linked service that uses user-password authentication.	<input checked="" type="radio"/>	<input type="radio"/>
The users of workspace1 must be assigned the Storage Blob Contributor role for datalake1.	<input type="radio"/>	<input checked="" type="radio"/>

Question #88

## DRAG DROP

---

You have a data warehouse.

You need to implement a slowly changing dimension (SCD) named Product that will include three columns named ProductName, ProductColor, and ProductSize. The solution must meet the following requirements:

- Prevent changes to the values stored in ProductName.
- Retain only the current and the last values in ProductSize.
- Retain all the current and previous values in ProductColor.

Which type of SCD should you implement for each column? To answer, drag the appropriate types to the correct columns. Each 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.

<b>SCD Type</b>	<b>Answer Area</b>
Type 0	ProductName: <input type="text"/>
Type 1	Color: <input type="text"/>
Type 2	Size: <input type="text"/>
Type 3	

**Correct Answer:**

**Answer Area**

ProductName:	<input type="text"/> Type 0
Color:	<input type="text"/> Type 1
Size:	<input type="text"/> Type 2

**Question #89**

## 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.

NOTE: Each correct selection is worth one point.

### Answer Area

Table distribution:

HASH
REPLICATE
ROUND_ROBIN

Table structure:

Clustered index
Columnstore index
Heap

**Correct Answer:**

### Answer Area

Table distribution:

HASH
REPLICATE
ROUND_ROBIN

Table structure:

Clustered index
Columnstore index
Heap

## Question #90

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

You need to ensure that you can query the data in container1 by using the serverless 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. Enable Azure Synapse Link for Cosmos1.
- C. In ws1, create a linked service that references Cosmos1.
- D. Enable the analytical store for container1.

## Question #91

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.

**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"/>

**Correct Answer:**

**Answer Area**

Statements	Yes	No
An external table that uses Ds1 can read the _monthly.csv file.	<input type="radio"/>	<input checked="" type="checkbox"/>
An external table that uses Ds1 can read the Monthly.csv file.	<input checked="" type="checkbox"/>	<input type="radio"/>
An external table that uses Ds2 can read the .testdata.csv file.	<input checked="" type="checkbox"/>	<input type="radio"/>

**Question #92**

## DRAG DROP

You have an Azure subscription that contains an Azure Data Lake Storage Gen2 account named account1 and a user named User1.

In account1, you create a container named container1. In container1, you create a folder named folder1.

You need to ensure that User1 can list and read all the files in folder1. The solution must use the principle of least privilege.

How should you configure the permissions for each folder? To answer, drag the appropriate permissions to the correct folders. 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.

NOTE: Each correct selection is worth one point.

Permissions		Answer Area
Execute	None	container1/: <input type="text"/>
Read	Read and Execute	container1/folder1: <input type="text"/>
Read and Write	Write	

### Correct Answer:

**Answer Area**

container1/: <input type="text"/> Execute
container1/folder1: <input type="text"/> Read and Execute

## Question #93

You have an Azure Data Factory pipeline named pipeline1.

You need to execute pipeline1 at 2 AM every day. The solution must ensure that if the trigger for pipeline1 stops, the next pipeline execution will occur at 2 AM, following a restart of the trigger.

Which type of trigger should you create?

- A. schedule
- B. tumbling
- C. storage event
- D. custom event

**Correct Answer:**

- A. schedule

**Question #94**

HOTSPOT

-

You have an Azure data factory named adf1 that contains a pipeline named ExecProduct. ExecProduct contains a data flow named Product.

The Product data flow contains the following transformations:

1. WeeklyData: A source that points to a CSV file in an Azure Data Lake Storage Gen2 account with 20 columns
2. ProductColumns: A select transformation that selects from WeeklyData six columns named ProductID, ProductDescr, ProductSubCategory, ProductCategory, ProductStatus, and ProductLastUpdated
3. ProductRows: An aggregate transformation
4. ProductList: A sink that outputs data to an Azure Synapse Analytics dedicated SQL pool

The Aggregate settings for ProductRows are configured as shown in the following exhibit.

The screenshot shows the 'Aggregate settings' configuration for the 'ProductRows' transformation. At the top, there are tabs for 'Aggregate settings', 'Optimize', 'Inspect', and 'Data preview'. The 'Aggregate settings' tab is selected. Below the tabs, there are fields for 'Output stream name' (set to 'ProductRows') and 'Incoming stream' (set to 'ProductColumns'). Underneath these, there are two buttons: 'Group by' and 'Aggregates', with 'Aggregates' being the active tab. A note below says 'Grouped by: ProductID'. At the bottom, there are buttons for '+ Add', 'Clone', 'Delete', and 'Open expression builder'. The main area shows a table with two columns: 'Column' and 'Expression'. There is one row in the table, which is a complex expression: 'Each column that matches name != 'ProductID'' followed by 'creates 1 column(s)' and an expression 'first(\$\$)'. There are also '+' and '-' buttons to add or remove items from the list.

For each of the following statements, select Yes if the statement is true. Otherwise, select No.

NOTE: Each correct selection is worth one point.

**Answer Area**

Statements	Yes	No
There will be six columns in the output of ProductRows.	<input type="radio"/>	<input type="radio"/>
There will always be one output row for each unique value of ProductDescr.	<input type="radio"/>	<input type="radio"/>
There will always be one output row for each unique value of ProductID.	<input type="radio"/>	<input type="radio"/>

**Correct Answer:**

**Answer Area**

Statements	Yes	No
There will be six columns in the output of ProductRows.	<input type="radio"/>	<input checked="" type="checkbox"/>
There will always be one output row for each unique value of ProductDescr.	<input type="radio"/>	<input checked="" type="checkbox"/>
There will always be one output row for each unique value of ProductID.	<input checked="" type="checkbox"/>	<input type="radio"/>

**Question #95**

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 limit
- B. Cache hit percentage
- C. Local tempdb percentage
- D. Data IO percentage

**Correct Answer:**

- B. Cache hit percentage

**Question #96**

## HOTSPOT

---

You have an Azure Synapse Analytics serverless SQL pool.

You have an Apache Parquet file that contains 10 columns.

You need to query data from the file. The solution must return only two columns.

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

NOTE: Each correct selection is worth one point.

### Answer Area

```
SELECT * FROM  
OPENROWSET( [ ] N'https://myaccount.dfs.core.windows.net/mycontainer/mysubfolder/data.parquet', FORMAT = 'PARQUET')  
WITH [ ]  
(Col1 int, Col2 varchar(20))  
FILEPATH(2)  
PARSER_VERSION = '2.0'  
SINGLE_BLOB
```

## Correct Answer:

### Answer Area

```
SELECT * FROM  
OPENROWSET( [ ] N'https://myaccount.dfs.core.windows.net/mycontainer/mysubfolder/data.parquet', FORMAT = 'PARQUET')  
WITH [ ]  
[ ]  
(Col1 int, Col2 varchar(20))  
FILEPATH(2)  
PARSER_VERSION = '2.0'  
SINGLE_BLOB
```

## Question #97

You have an Azure Synapse Analytics workspace that contains an Apache Spark pool named SparkPool1. SparkPool1 contains a Delta Lake table named SparkTable1.

You need to recommend a solution that supports Transact-SQL queries against the data referenced by SparkTable1. The solution must ensure that the queries can use partition elimination.

What should you include in the recommendation?

- A. a partitioned table in a dedicated SQL pool
- B. a partitioned view in a dedicated SQL pool
- C. a partitioned index in a dedicated SQL pool
- D. a partitioned view in a serverless SQL pool

**Correct Answer:**

- D. a partitioned view in a serverless SQL pool

**Question #98**

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

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

- A. 1 million
- B. 5 million
- C. 20 million
- D. 60 million

**Correct Answer:**

- A. 1 million

**Question #99**

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. data formats
- C. linked services
- D. table permissions
- E. table definitions

**Correct Answer:**

- A. relationships
- E. table definitions

**Question #100**

You are implementing a star schema in an Azure Synapse Analytics dedicated SQL pool.

You plan to create a table named DimProduct.

DimProduct must be a Type 3 slowly changing dimension (SCD) table that meets the following requirements:

- The values in two columns named ProductKey and ProductSourceID will remain the same.
- The values in three columns named ProductName, ProductDescription, and Color can change.

You need to add additional columns to complete the following table definition.

```
CREATE TABLE [dbo].[dimproduct]
(
    [ProductKey]           INT NOT NULL,
    [ProductSourceID]      INT NOT NULL,
    [ProductName]          NVARCHAR(100) NOT NULL,
    [ProductDescription]   NVARCHAR(2000) NOT NULL,
    [Color]                NVARCHAR(50) NOT NULL
)
WITH
(
    DISTRIBUTION = REPLICATE,
    CLUSTERED COLUMNSTORE INDEX
);
```

Which three columns should you add? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. [EffectiveStartDate] [datetime] NOT NULL
- B. [EffectiveEndDate] [datetime] NOT NULL
- C. [OriginalProductDescription] NVARCHAR(2000) NOT NULL
- D. [IsCurrentRow] [bit] NOT NULL
- E. [OriginalColor] NVARCHAR(50) NOT NULL
- F. [OriginalProductName] NVARCHAR(100) NULL

**Correct Answer:**

- C. [OriginalProductDescription] NVARCHAR(2000) NOT NULL
- E. [OriginalColor] NVARCHAR(50) NOT NULL
- F. [OriginalProductName] NVARCHAR(100) NULL

### Question #101

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.

### 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.

**Correct Answer:**

## Answer Area

Partition the data:	<p>Partition by date with one partition per day.</p> <p>Partition by date with one partition per month. <b>Correct Answer</b></p> <p>Partition by product.</p>
Remove the data:	<p>Delete the old data from Table1 by using a WHERE clause.</p> <p>Delete the old data from Table1 by using a JOIN.</p> <p>Switch the oldest partition to another table named Table2 and drop Table2. <b>Correct Answer</b></p> <p>Truncate the oldest partition.</p>

### Question #102

You have an Azure subscription that contains an Azure Synapse Analytics serverless SQL pool.

You execute the following query.

```
CREATE EXTERNAL TABLE Orders
WITH
(
    LOCATION = 'orders/',
    DATA_SOURCE = sales,
    FILE_FORMAT = SalesOrders
)
AS
SELECT OrderID, CustomerName, OrderTotal
FROM OPENROWSET
(
    BULK 'sales_orders/*.csv',
    DATA_SOURCE = 'sales',
    FORMAT = 'CSV',
    PARSER_VERSION = '2.0',
    HEADER_ROW = TRUE
) AS source_data
WHERE OrderType = 'Customer Order';
```

Where will the rows returned by the query be stored?

- A. in a file in a data lake
- B. in a relational database
- C. in a global temporary table
- D. in a session temporary table

**Correct Answer:**

- A. in a file in a data lake

**Question #103**

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. partition style
- B. business area
- C. size
- D. facts and dimensions

**Correct Answer:**

- B. business area

**Question #104**

You have an Azure data factory connected to a Git repository that contains the following branches:

- main: Collaboration branch
- abc: Feature branch
- xyz: Feature branch

You save changes to a pipeline in the xyz branch.

You need to publish the changes to the live service.

What should you do first?

- A. Publish the data factory.
- B. Create a pull request to merge the changes into the main branch.
- C. Create a pull request to merge the changes into the abc branch.
- D. Push the code to a remote origin.

**Correct Answer:**

- B. Create a pull request to merge the changes into the main branch.

### **Question #105**

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 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.

Does this meet the goal?

- A. Yes
- B. No

#### **Correct Answer:**

- A. Yes

### **Question #106**

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 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 view the JSON code representation of the resource and copy the JSON to a file.

Does this meet the goal?

- A. Yes
- B. No

**Correct Answer:**

- A. Yes

**Question #107**

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 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.

Does this meet the goal?

- A. Yes
- B. No

**Correct Answer:**

- B. No

## Question #108

HOTSPOT

You have an Azure Databricks workspace.

You read data from a CSV file by using a notebook, and then load the data to a DataFrame.

You need to add rows from the DataFrame to an existing Delta table by using Python code.

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

NOTE: Each correct selection is worth one point.

### Answer Area

```
new_rows_df.write.    .save(delta_table_path)
```

format("csv")	mode("append")
format("delta")	mode("error")
format("json")	mode("ignore")
format("parquet")	mode("overwrite")

Correct Answer:

### Answer Area

```
new_rows_df.write.    .save(delta_table_path)
```

format("csv")	mode("append")
format("delta")	mode("error")
format("json")	mode("ignore")
format("parquet")	mode("overwrite")

## Question #109

DRAG DROP

You have an Azure subscription that contains an Azure Cosmos DB for NoSQL account named account1. The account1 account contains a container named Container1 that has the following configurations:

- Analytical store: On
- TTL: 3600

You need to remove analytical store support from Container1. The solution must meet the following requirements:

- Minimize the impact on the apps that reference Container1.
- Minimize storage usage.

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.

### Actions

### Answer Area

Create another container named Container1 and copy the contents of Container2 to Container1.

Set the TTL for Container1 to null.

Create a new container named Container2 and copy the contents of Container1 to Container2.

Set the TTL for Container1 to 0.

Delete Container1.

Delete Container2.



**Correct Answer:**

## Answer Area

Create a new container named Container2 and copy the contents of Container1 to Container2.

Delete Container1.

Create another container named Container1 and copy the contents of Container2 to Container1.

Delete Container2.

## Question #110

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.

## **Actions**

## **Answer Area**

Create a new table named Table1v2 by running CTAS.

Rename Table1v2 as Table1.

Rename Table1 as Table1\_old.

Run DBCC PDW\_SHOWSPACEUSED.

Drop Table1\_old.

Drop the indexes of Table1.



## **Correct Answer:**

## **Answer Area**

Create a new table named Table1v2 by running CTAS.

Rename Table1 as Table1\_old.

Rename Table1v2 as Table1.

Drop Table1\_old.

## **Question #111**

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 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 disable all the triggers for ADF1.

Does this meet the goal?

- A. Yes
- B. No

**Correct Answer:**

- B. No

**Question #112**

You have an Azure subscription that contains the resources shown in the following table.

Name	Type	Description
workspace1	Azure Synapse Analytics workspace	<i>None</i>
sql1	Azure SQL Database server	<i>None</i>
SQLDb1	Azure SQL database	Hosted on sql1

You need to implement Azure Synapse Link for Azure SQL Database.

Which two actions should you perform on sql1? Each correct answer presents a part of the solution.

NOTE: Each correct selection is worth one point.

- A. Update the firewall rules to allow Azure services to access sql1.
- B. Enable the system-assigned managed identity.
- C. From the Access control (IAM) settings, assign the Contributor role to the system-assigned managed identity of workspace1.

D. Disable Transparent Data Encryption (TDE).

**Correct Answer:**

- A. Update the firewall rules to allow Azure services to access sql1.
- B. Enable the system-assigned managed identity.

**Question #113**

You have an Azure subscription that contains an Azure Cosmos DB database. Azure Synapse Link is implemented on the database.

You configure a full fidelity schema for the analytical store.

You perform the following actions:

- Insert {"customerID": 12, "customer": "Tailspin Toys"} as the first document in the container.
- Insert {"customerID": "14", "customer": "Contoso"} as the second document in the container.

How many columns will the analytical store contain?

- A. 1
- B. 2
- C. 3
- D. 4

**Correct Answer:**

C. 3

**Question #114**

HOTSPOT

-

You have an Azure Data Lake Storage account that contains CSV files. The CSV files contain sales order data and are partitioned by using the following format.

/data/salesorders/year=xxxx/month=y

You need to retrieve only the sales orders from January 2023 and February 2023.

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

NOTE: Each correct selection is worth one point.

## Answer Area

```
SELECT product_id, product, list_price
FROM OPENROWSET
(
    BULK 'https://salesdatalake.blob.core.windows.net/data/salesorders/**',
    FORMAT = 'CSV',
    PARSER_VERSION = '2.0'
)
WITH
(
    product_id int,
    product varchar(35) COLLATE Latin1_General_100_BIN2_UTF8,
    list_price decimal(18, 2)
) AS so
WHERE
    so.filepath(0) = 2023
    so.filepath(1) = '2023'
    so.year = 2023
AND
    (so.month = 1 OR so.month = 2)
    so.filepath(1) IN (1,2)
    so.filepath(2) IN ('1', '2')
```

**Correct Answer:**

#### Answer Area

```
SELECT product_id, product, list_price
FROM OPENROWSET
(
    BULK 'https://salesdatalake.blob.core.windows.net/data/salesorders/**',
    FORMAT = 'CSV',
    PARSER_VERSION = '2.0'
)
WITH
(
    product_id int,
    product varchar(35) COLLATE Latin1_General_100_BIN2_UTF8,
    list_price decimal(18, 2)
) AS so
WHERE
    so.filepath(0) = 2023
    so.filepath(1) = '2023'
    so.year = 2023
AND
    (so.month = 1 OR so.month = 2)
    so.filepath(1) IN (1, 2)
    so.filepath(2) IN ('1', '2')
```

#### Question #115

HOTSPOT

You have an Azure subscription that contains an Azure Data Lake Storage Gen2 account named adlsv2. adlsv2 contains a container named logs and an Azure Synapse Analytics dedicated SQL pool named Sqlpool.

You plan to use PolyBase to load external data into Sqlpool.

You need to define the external data source.

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

NOTE: Each correct selection is worth one point.

## Answer Area

```
...  
CREATE EXTERNAL DATA SOURCE logs  
with (  
    TYPE = HADOOP,  
    LOCATION = 

|           |               |                         |
|-----------|---------------|-------------------------|
| 'abfss:// | logs@adls2    | .dfs.core.windows.net', |
| 'adl://   | sqlpool@adls2 |                         |
| 'wasbs:// | dbcred@adls2  |                         |

  
    CREDENTIAL = dbcred  
);  
GO  
...
```

## Correct Answer:

### Answer Area

```
...  
CREATE EXTERNAL DATA SOURCE logs  
with (  
    TYPE = HADOOP,  
    LOCATION = 

|                |               |                         |
|----------------|---------------|-------------------------|
| 'abfss://      | logs@adls2    | .dfs.core.windows.net', |
| <b>'adl://</b> | sqlpool@adls2 |                         |
| 'wasbs://      | dbcred@adls2  |                         |

  
    CREDENTIAL = dbcred  
);  
GO  
...
```

## Question #116

You are developing a solution that will stream to Azure Stream Analytics. The solution will have both streaming data and reference data.

Which input type should you use for the reference data?

- A. Azure Service Bus
- B. Azure Blob storage

- C. Azure IoT Hub
- D. Azure Event Hubs

**Correct Answer:**

- B. Azure Blob storage

**Question #117**

You have an Azure subscription that contains an Azure Synapse Analytics workspace and a user named User1.

You need to ensure that User1 can create a new lake database by using an Azure Synapse database template from Gallery. The solution must follow the principle of least privilege.

Which role should you assign to User1?

- A. Synapse Contributor
- B. Synapse Administrator
- C. Synapse User
- D. Storage Blob Data Contributor

**Correct Answer:**

- A. Synapse Contributor