

Microsoft DP-203 Exam

Topic 1

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

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

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.

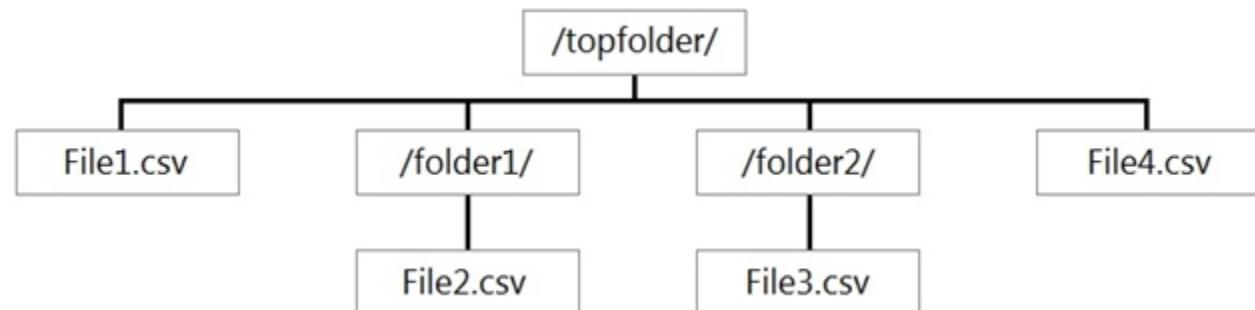
Step 3: Drop the SalesFact_Work table.

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:

| |
|---------|
| Avro |
| CSV |
| Parquet |
| TSV |

Report2:

| |
|---------|
| Avro |
| CSV |
| Parquet |
| TSV |

Correct Answer:

Answer Area

Report1:

| |
|---------|
| Avro |
| CSV |
| Parquet |
| TSV |

Report2:

| |
|---------|
| Avro |
| CSV |
| Parquet |
| TSV |

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 |

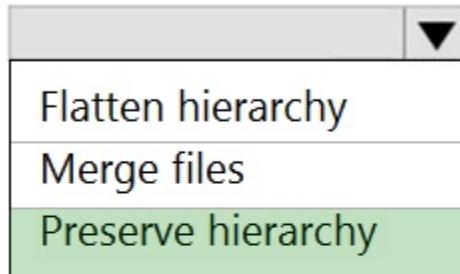
Sink file type:

| |
|---------|
| CSV |
| JSON |
| Parquet |
| TXT |

Correct Answer:

Answer Area

Copy behavior:



Sink file type:



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:

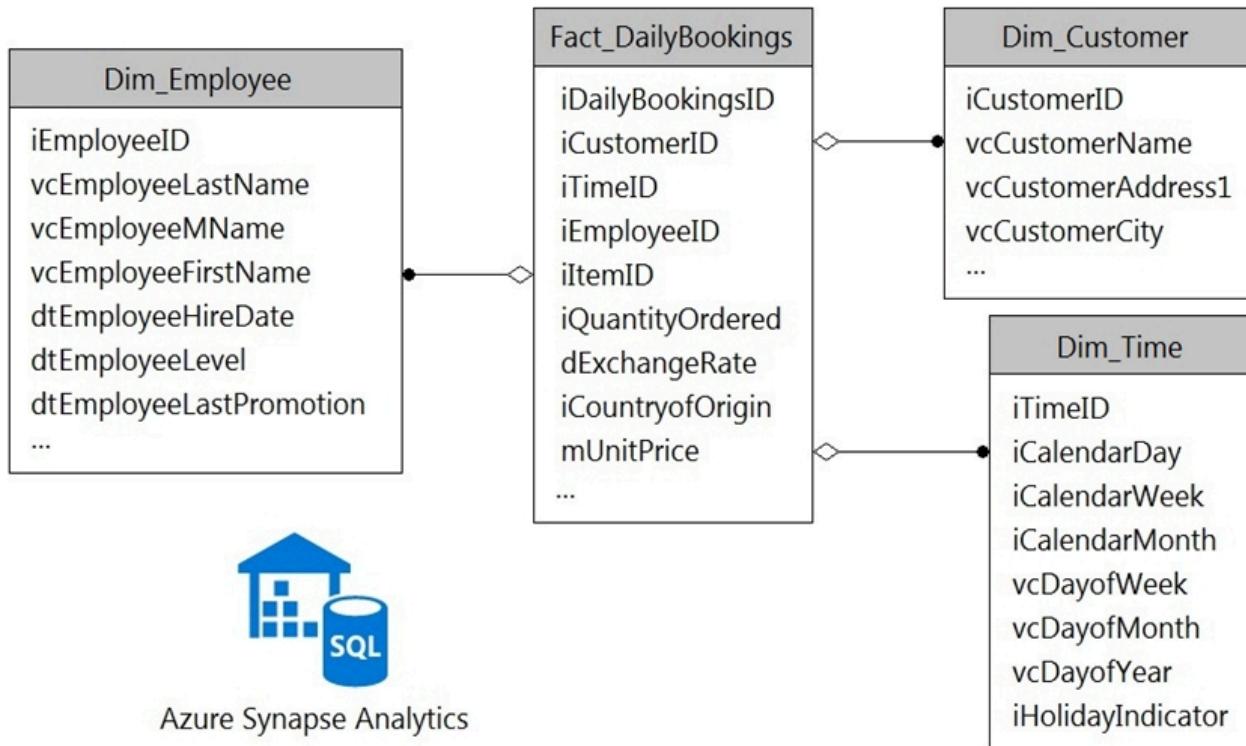
<https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-introduction>

<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 ¹ | 180 days |
| Latency (Time to first byte) | Single-digit milliseconds | milliseconds | milliseconds | hours ² |

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

Answer Area

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

Correct Answer:

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

Answer Area

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

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?

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

Correct Answer:

- D. 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

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

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

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

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

| |
|-----------------------------------|
| a random number |
| the values stored in the database |
| XXXX |
| 0 |

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

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

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:

| |
|----------------|
| Binary |
| Parquet |
| Delimited text |

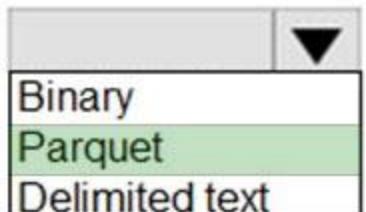
Copy activity copy behavior:

| |
|-------------------|
| FlattenHierarchy |
| MergeFiles |
| PreserveHierarchy |

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 |
| 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 |

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

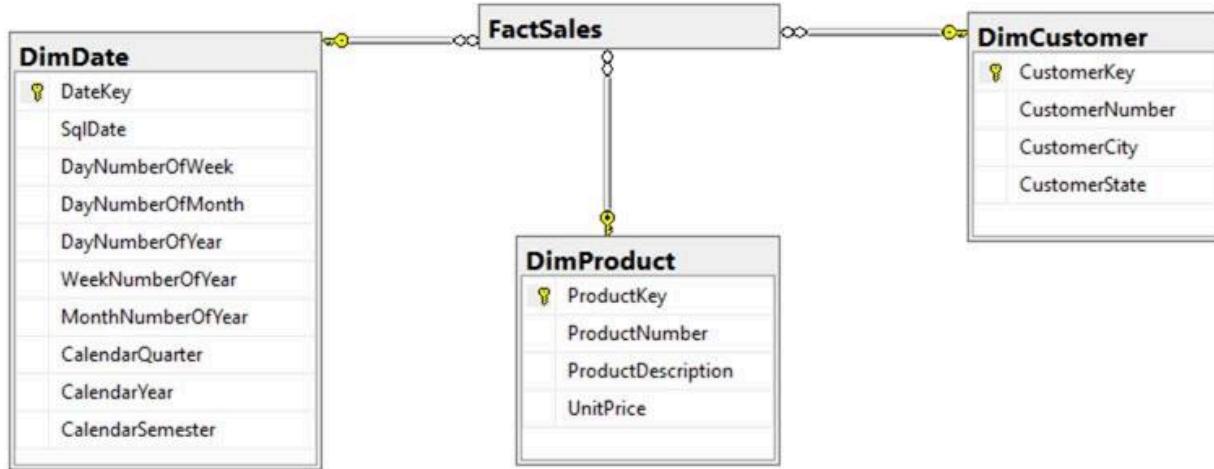
Denormalization is the process of transforming higher normal forms to lower normal forms via storing the join of higher normal form relations as a base relation.

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")

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

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

Correct Answer:**Answer Area**

```
df.write
    .bucketBy
    .partitionBy
    .range
    .sortBy
    .mode("append")
        .csv("/Purchases")
        .json("/Purchases")
        .parquet("/Purchases")
        .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 |
| a surrogate key |
| a business key |
| an audit column |

The ProductKey column is [answer choice].

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

Create an external file format object

Create an external data source

Create a query that uses Create Table as Select

Create a table

Create an external table

Answer Area**Correct Answer:****Actions**

Create a query that uses Create Table as Select

Create a table

Answer Area

Create an external data source

Create an external file format object

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

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

Four large, empty rectangular boxes intended for placing the selected actions from the left side.

Answer Area

Create an external data source that uses the abfs location

Create an external file format and set the First_Row option

Use CREATE EXTERNAL TABLE AS SELECT (CETAS) and configure the reject options to specify reject values or percentages

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]
    [TransactionDateID], [TransactionTypeID]
    HASH([TransactionTypeID])
    ROUND ROBIN
)
RANGE RIGHT FOR VALUES
(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]
    [TransactionDateID], [TransactionTypeID]
    HASH([TransactionTypeID])
    ROUND ROBIN
)
RANGE RIGHT FOR VALUES
(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', PARSE_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.

| owner | age | dog |
|-------|-----|--------|
| 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 spit bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point.

Select and Place:

| Values | Answer Area |
|-----------------|-------------|
| alias | |
| array_union | |
| createDataFrame | |
| explode | |
| 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") |
| array_union | persons = source_df. <input type="text" value="Value"/> <input type="text" value="Value"/> ("persons").alias("persons")) |
| createDataFrame | persons_dogs = persons.select(col("persons.name").alias("owner"), col("persons.age").alias("age"), explode <input type="text" value="Value"/> ("dog")) ("persons-dogs"). display(persons_dogs) |
| | |
| 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 |

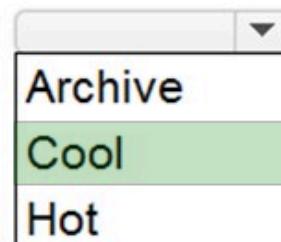
Correct Answer:

Answer Area

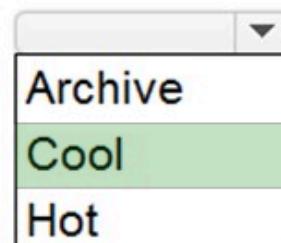
First week:



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 ¹ | 180 days |
| Latency (Time to first byte) | Single-digit milliseconds | milliseconds | milliseconds | hours ² |

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 ⓘ

product.csv

Date format

YYYY/MM/DD

Time format

HH

Event serialization format * ⓘ

CSV

Delimiter ⓘ

comma (,)

Encoding ⓘ

UTF-8

Save ⓘ 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

[..]
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

```
CREATE EXTERNAL TABLE applications  
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
```

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 addresshouseumber, [address_line1] as addressline1
FROM
    (BULK 'https://contoso1.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       | blob  |
| TYPE = BLOB_STORAGE | dfs   |
| TYPE = HADOOP       | table |


)
```

Correct Answer:

Answer Area

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

|                     |       |
|---------------------|-------|
| PUSHDOWN = ON       | blob  |
| TYPE = BLOB_STORAGE | dfs   |
| TYPE = HADOOP       | table |


)
```

Box 1: blob -

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

CREATE EXTERNAL DATA SOURCE YellowTaxi

WITH (LOCATION = 'https://azuredataprocstorage.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 | <pre> 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] mmoney NOT NULL) WITH (CLUSTERED COLUMNSTORE INDEX , DISTRIBUTION = [Value] ([ProductKey]) , PARTITION ([Value]) RANGE RIGHT FOR VALUES (20170101,20180101,20190101,20200101,20210101)) </pre> |

Correct Answer:

| Values | Answer Area |
|------------------|--|
| CustomerKey | <pre>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))</pre> |
| ROUND_ROBIN | |
| REPLICATE | |
| SalesOrderNumber | |

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 ¹ | 180 days |
| Latency (Time to first byte) | Single-digit milliseconds | milliseconds | milliseconds | hours ² |

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"> Contains one row per date for the last 20 years Contains columns named Year, Month, Quarter, and IsWeekend |
| 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 |
|--------------------------|--|
| Hash | Common.Data: <input type="text"/> |
| Replicated | Marketing.Web.Sessions: <input type="text"/> |
| Round-robin | Staging. Web.Sessions: <input type="text"/> |

Correct Answer:

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

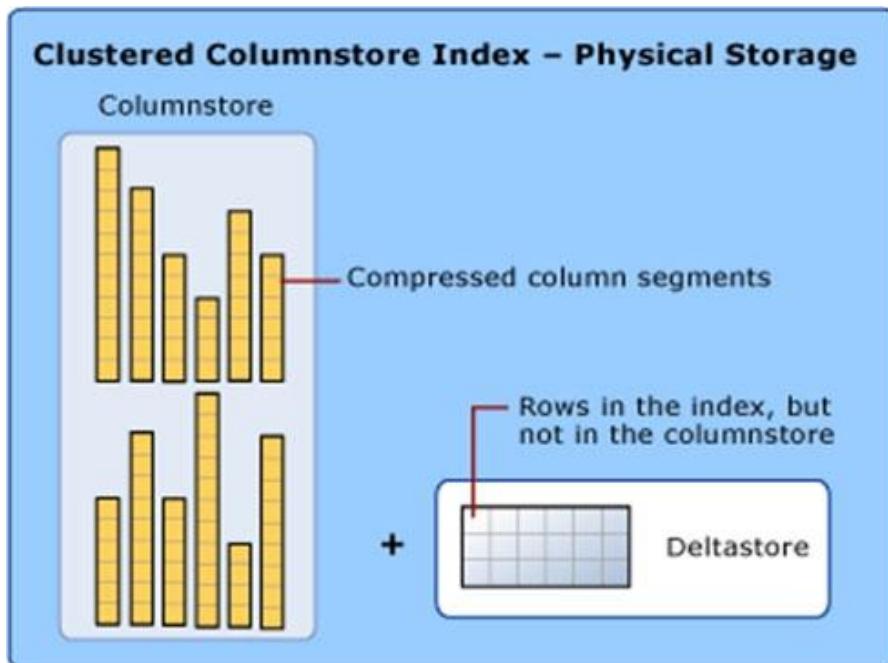
Columnstore indexes are the standard for storing and querying large data warehousing fact tables. This index uses column-based data storage and query processing to achieve gains up to 10 times the query performance in your data warehouse over traditional row-oriented storage.

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:

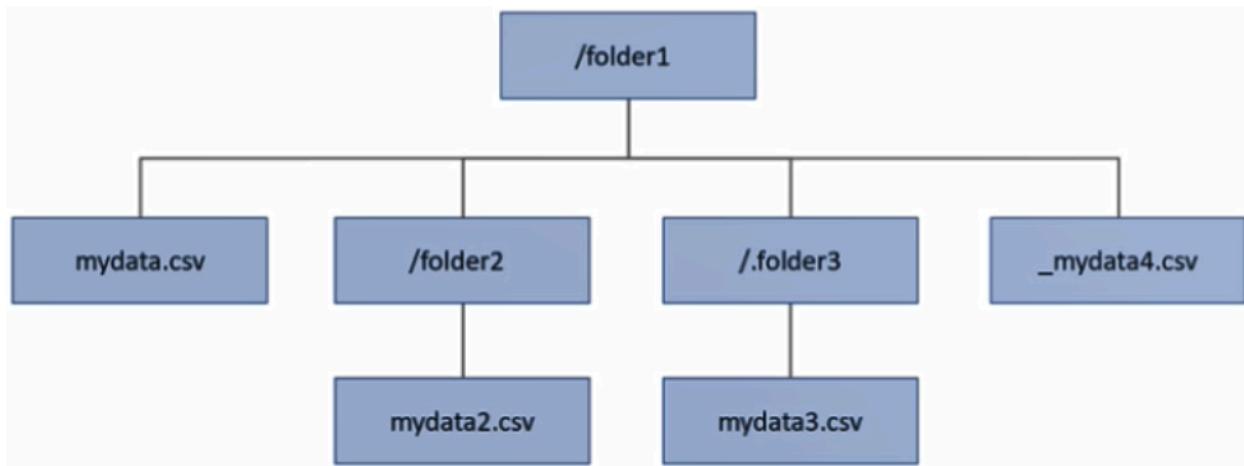
C. an IDENTITY column

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

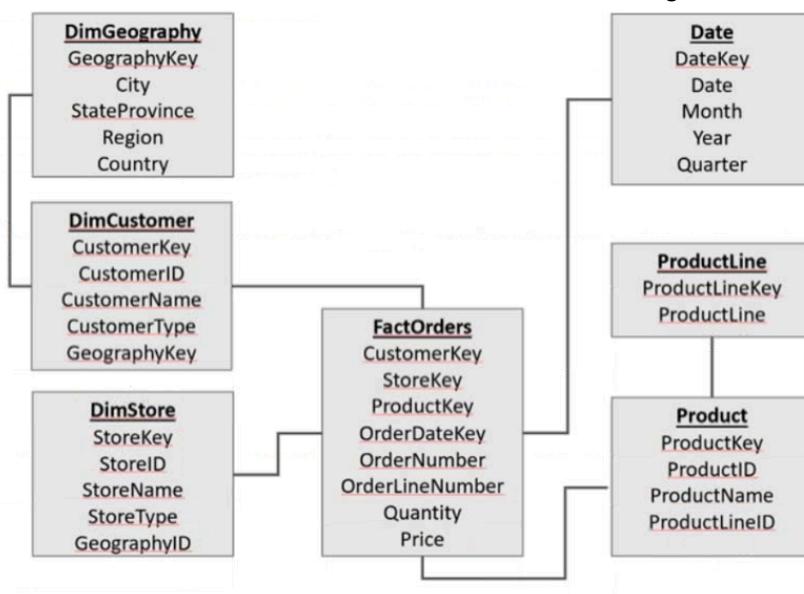
- A. Use a native external table and authenticate by using a shared access signature (SAS).

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 NOT 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:

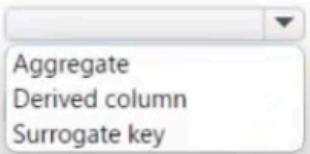
- Detect whether the data of a given customer has changed in the DimCustomer table.
- Perform an upsert to the DimCustomer table.

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

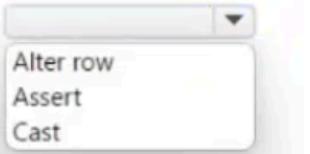
NOTE: Each correct selection is worth one point.

Answer Area

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



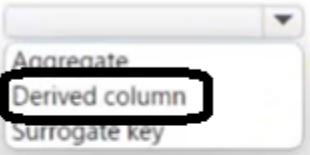
Perform an upsert to the DimCustomer table:



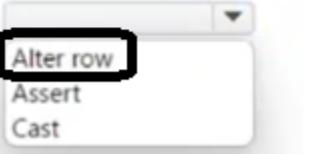
Correct Answer:

Answer Area

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



Perform an upsert to the DimCustomer table:



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="checkbox"/> | <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="checkbox"/> | <input type="radio"/> |
| The users of workspace1 must be assigned the Storage Blob Contributor role for datalake1. | <input type="radio"/> | <input checked="" type="checkbox"/> |

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.

| SCD Type | Answer Area |
|-----------------|-----------------------------------|
| 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: | Type 0 |
| Color: | Type 1 |
| Size: | 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.

Aggregate settings Optimize Inspect Data preview

Output stream name * Learn more [↗](#)

Incoming stream *

[Group by](#) **Aggregates**

Grouped by: ProductID

+ Add [Clone](#) [Delete](#) Open expression builder

| <input type="checkbox"/> Column | Expression |
|--|--|
| <input type="checkbox"/> Each column that matches <code>name != 'ProductID'</code> | <code>creates 1 column(s)</code> ↗ + ↗ - |
| <code>\$\$</code> | <code>abc</code> any + ↗ - |
| | <code>first(\$\$)</code> |

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 checked="" type="radio"/> | <input type="radio"/> |
| There will always be one output row for each unique value of ProductDescr. | <input type="radio"/> | <input checked="" type="radio"/> |
| There will always be one output row for each unique value of ProductID. | <input type="radio"/> | <input checked="" 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="radio"/> |
| There will always be one output row for each unique value of ProductDescr. | <input type="radio"/> | <input checked="" type="radio"/> |
| There will always be one output row for each unique value of ProductID. | <input checked="" type="radio"/> | <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)  
    PARSE_VERSION = '2.0'  
    SINGLE_BLOB
```

Correct Answer:

Answer Area

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

- [EffectiveStartDate] [datetime] NOT NULL
- [EffectiveEndDate] [datetime] NOT NULL
- [OriginalProductDescription] NVARCHAR(2000) NOT NULL
- [IsCurrentRow] [bit] NOT NULL
- [OriginalColor] NVARCHAR(50) NOT NULL
- [OriginalProductName] NVARCHAR(100) NULL

Correct Answer:

- [OriginalProductDescription] NVARCHAR(2000) NOT NULL
- [OriginalColor] NVARCHAR(50) NOT NULL
- [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:

- 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.

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.  
format("csv").  
format("delta").  
format("json").  
format("parquet").  
  
.save(delta_table_path)  
mode("append")  
mode("error")  
mode("ignore")  
mode("overwrite")
```

Correct Answer:

Answer Area

```
new_rows_df.write.  
format("csv").  
format("delta").  
format("json").  
format("parquet").  
  
.save(delta_table_path)  
mode("append")  
mode("error")  
mode("ignore")  
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

Topic 2

Question #1

HOTSPOT -

You plan to create a real-time monitoring app that alerts users when a device travels more than 200 meters away from a designated location.

You need to design an Azure Stream Analytics job to process the data for the planned app. The solution must minimize the amount of code developed and the number of technologies used.

What should you include in the Stream Analytics job? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Input type:

| |
|-----------|
| Stream |
| Reference |

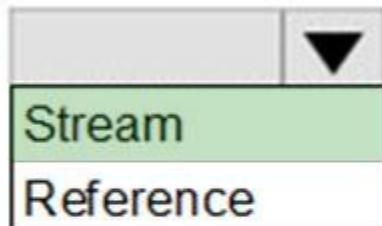
Function:

| |
|------------|
| Aggregate |
| Geospatial |
| Windowing |

Correct Answer

Answer Area

Input type:



Function:



Input type: Stream -

You can process real-time IoT data streams with Azure Stream Analytics.

Function: Geospatial -

With built-in geospatial functions, you can use Azure Stream Analytics to build applications for scenarios such as fleet management, ride sharing, connected cars, and asset tracking.

Note: In a real-world scenario, you could have hundreds of these sensors generating events as a stream. Ideally, a gateway device would run code to push these events to Azure Event Hubs or Azure IoT Hubs.

Reference:

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-get-started-with-azure-stream-analytics-to-process-data-from-iot-devices>

<https://docs.microsoft.com/en-us/azure/stream-analytics/geospatial-scenarios>

Question #2

A company has a real-time data analysis solution that is hosted on Microsoft Azure. The solution uses Azure Event Hub to ingest data and an Azure Stream Analytics cloud job to analyze the data. The cloud job is configured to use 120 Streaming Units (SU).

You need to optimize performance for the Azure Stream Analytics job.

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

NOTE: Each correct selection is worth one point.

- A. Implement event ordering.
- B. Implement Azure Stream Analytics user-defined functions (UDF).
- C. Implement query parallelization by partitioning the data output.D. Scale the SU count for the job up.
- E. Scale the SU count for the job down.
- F. Implement query parallelization by partitioning the data input.

Correct Answer:

- C. Implement query parallelization by partitioning the data output.
- F. Implement query parallelization by partitioning the data input.

Question #3

You need to trigger an Azure Data Factory pipeline when a file arrives in an Azure Data Lake Storage Gen2 container.

Which resource provider should you enable?

- A. Microsoft.Sql
- B. Microsoft.Automation
- C. Microsoft.EventGrid
- D. Microsoft.EventHub

Correct Answer:

- C. Microsoft.EventGrid

Question #4

You plan to perform batch processing in Azure Databricks once daily.

Which type of Databricks cluster should you use?

- A. High Concurrency
- B. automated
- C. interactive

Correct Answer:

- B. automated

Question #5

HOTSPOT -

You are processing streaming data from vehicles that pass through a toll booth.

You need to use Azure Stream Analytics to return the license plate, vehicle make, and hour the last vehicle passed during each 10-minute window.

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

```
WITH LastInWindow AS
(
    SELECT
        

|        |                  |
|--------|------------------|
| (Time) | AS LastEventTime |
| COUNT  |                  |
| MAX    |                  |
| MIN    |                  |
| TOPONE |                  |


    FROM
        Input TIMESTAMP BY Time
    GROUP BY
        

|                |
|----------------|
| (minute, 10)   |
| HoppingWindow  |
| SessionWindow  |
| SlidingWindow  |
| TumblingWindow |


)
SELECT
    Input.License_plate,
    Input.Make,
    Input.Time
FROM
    Input TIMESTAMP BY Time
    INNER JOIN LastInWindow
    ON 

|                               |                  |
|-------------------------------|------------------|
| (minute, Input, LastInWindow) | BETWEEN 0 AND 10 |
| DATEADD                       |                  |
| DATEDIFF                      |                  |
| DATENAME                      |                  |
| DATEPART                      |                  |


    AND Input.Time = LastInWindow.LastEventTime
```

Correct Answer:**Answer Area**

```
WITH LastInWindow AS
(
    SELECT
        (Time) AS LastEventTime
        COUNT
        MAX
        MIN
        TOPONE
    FROM
        Input TIMESTAMP BY Time
    GROUP BY
        (minute, 10)
        HoppingWindow
        SessionWindow
        SlidingWindow
        TumblingWindow
)
SELECT
    Input.License_plate,
    Input.Make,
    Input.Time
FROM
    Input TIMESTAMP BY Time
    INNER JOIN LastInWindow
    ON (minute, Input, LastInWindow) BETWEEN 0 AND 10
        DATEADD
        DATEDIFF
        DATENAME
        DATEPART
    AND Input.Time = LastInWindow.LastEventTime
```

Box 1: MAX -

The first step on the query finds the maximum time stamp in 10-minute windows, that is the time stamp of the last event for that window. The second step joins the results of the first query with the original stream to find the event that match the last time stamps in each window.

Query:

```
WITH LastInWindow AS -
(
    SELECT -
        MAX(Time) AS LastEventTime -
    FROM -
        Input TIMESTAMP BY Time -
    GROUP BY -
        TumblingWindow(minute, 10)
)

SELECT -
    Input.License_plate,
    Input.Make,
    Input.Time -
FROM -
    Input TIMESTAMP BY Time -
INNER JOIN LastInWindow -
    ON DATEDIFF(minute, Input, LastInWindow) BETWEEN 0 AND 10
    AND Input.Time = LastInWindow.LastEventTime

Box 2: TumblingWindow -
Tumbling windows are a series of fixed-sized, non-overlapping and contiguous time intervals.

Box 3: DATEDIFF -
DATEDIFF is a date-specific function that compares and returns the time difference between two DateTime fields, for more information, refer to date functions.
Reference:
https://docs.microsoft.com/en-us/stream-analytics-query/tumbling-window-azure-stream-analytics
```

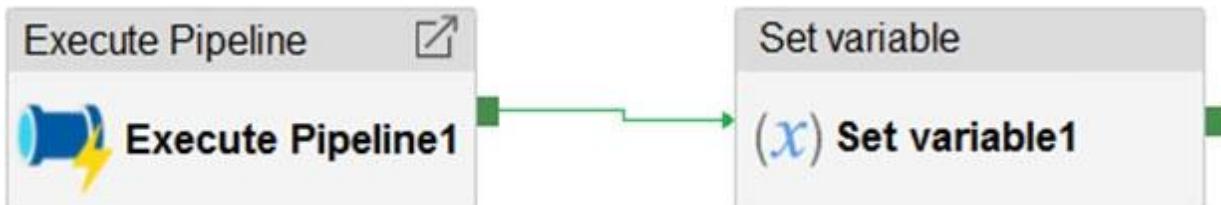
Question #6

You have an Azure Data Factory instance that contains two pipelines named Pipeline1 and Pipeline2.

Pipeline1 has the activities shown in the following exhibit.



Pipeline2 has the activities shown in the following exhibit.



You execute Pipeline2, and Stored procedure1 in Pipeline1 fails.

What is the status of the pipeline runs?

- A. Pipeline1 and Pipeline2 succeeded.
- B. Pipeline1 and Pipeline2 failed.
- C. Pipeline1 succeeded and Pipeline2 failed.
- D. Pipeline1 failed and Pipeline2 succeeded.

Correct Answer:

- A. Pipeline1 and Pipeline2 succeeded

Question #7

HOTSPOT -

A company plans to use Platform-as-a-Service (PaaS) to create the new data pipeline process. The process must meet the following requirements:

Ingest:

- ⇒ Access multiple data sources.
- ⇒ Provide the ability to orchestrate workflow.
- ⇒ Provide the capability to run SQL Server Integration Services packages.

Store:

- ⇒ Optimize storage for big data workloads.
- ⇒ Provide encryption of data at rest.
- ⇒ Operate with no size limits.

Prepare and Train:

- ⇒ Provide a fully-managed and interactive workspace for exploration and visualization.

- Provide the ability to program in R, SQL, Python, Scala, and Java.
- Provide seamless user authentication with Azure Active Directory.

Model & Serve:

- Implement native columnar storage.
- Support for the SQL language
- Provide support for structured streaming.

You need to build the data integration pipeline.

Which technologies should you use? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Architecture requirement

Technology

Ingest

| |
|--------------------|
| Logic Apps |
| Azure Data Factory |
| Azure Automation |

Store

| |
|-------------------------|
| Azure Data Lake Storage |
| Azure Blob storage |
| Azure files |

Prepare and Train

| |
|--------------------------------|
| HDInsight Apache Spark cluster |
| Azure Databricks |
| HDInsight Apache Storm cluster |

Model and Serve

| |
|--------------------------------|
| HDInsight Apache Kafka cluster |
| Azure Synapse Analytics |
| Azure Data Lake Storage |

Correct Answer:

Answer Area

Architecture requirement

Technology

Ingest

- Logic Apps
- Azure Data Factory
- Azure Automation

Store

- Azure Data Lake Storage
- Azure Blob storage
- Azure files

Prepare and Train

- HDInsight Apache Spark cluster
- Azure Databricks
- HDInsight Apache Storm cluster

Model and Serve

- HDInsight Apache Kafka cluster
- Azure Synapse Analytics
- Azure Data Lake Storage

Ingest: Azure Data Factory -

Azure Data Factory pipelines can execute SSIS packages.

In Azure, the following services and tools will meet the core requirements for pipeline orchestration, control flow, and data movement: Azure Data Factory, Oozie on HDInsight, and SQL Server Integration Services (SSIS).

Store: Data Lake Storage -

Data Lake Storage Gen1 provides unlimited storage.

Note: Data at rest includes information that resides in persistent storage on physical media, in any digital format. Microsoft Azure offers a variety of data storage solutions to meet different

needs, including file, disk, blob, and table storage. Microsoft also provides encryption to protect Azure SQL Database, Azure Cosmos DB, and Azure Data Lake.

Prepare and Train: Azure Databricks

Azure Databricks provides enterprise-grade Azure security, including Azure Active Directory integration.

With Azure Databricks, you can set up your Apache Spark environment in minutes, autoscale and collaborate on shared projects in an interactive workspace.

Azure Databricks supports Python, Scala, R, Java and SQL, as well as data science frameworks and libraries including TensorFlow, PyTorch and scikit-learn.

Model and Serve: Azure Synapse Analytics

Azure Synapse Analytics/ SQL Data Warehouse stores data into relational tables with columnar storage.

Azure SQL Data Warehouse connector now offers efficient and scalable structured streaming write support for SQL Data Warehouse. Access SQL Data

Warehouse from Azure Databricks using the SQL Data Warehouse connector.

Note: As of November 2019, Azure SQL Data Warehouse is now Azure Synapse Analytics.

Reference:

<https://docs.microsoft.com/en-us/bs-latn-ba/azure/architecture/data-guide/technology-choices/pipeline-orchestration-data-movement>

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

Question #8

DRAG DROP -

You have the following table named Employees.

| first_name | last_name | hire_date | employee_type |
|------------|-----------|------------|---------------|
| Jane | Doe | 2019-08-23 | new |
| Ben | Smith | 2017-12-15 | Standard |

You need to calculate the employee_type value based on the hire_date value.

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 |
|--------------|---|
| | SELECT |
| | * , |
| CASE | <input type="text"/> |
| ELSE | WHEN hire_date >= '2019-01-01' THEN 'New' |
| OVER | <input type="text"/> 'Standard' |
| PARTITION BY | END AS employee_type |
| ROW_NUMBER | FROM |
| | employees |

Correct Answer:

| Values | Answer Area |
|--------------|---|
| | SELECT |
| | * , |
| CASE | <input type="text"/> CASE |
| ELSE | WHEN hire_date >= '2019-01-01' THEN 'New' |
| OVER | <input type="text"/> ELSE 'Standard' |
| PARTITION BY | END AS employee_type |
| ROW_NUMBER | FROM |
| | employees |

Box 1: CASE -

CASE evaluates a list of conditions and returns one of multiple possible result expressions. CASE can be used in any statement or clause that allows a valid expression. For example, you can use CASE in statements such as SELECT, UPDATE, DELETE and SET, and in clauses such as select_list, IN, WHERE, ORDER BY, and HAVING.

Syntax: Simple CASE expression:

CASE input_expression -
WHEN when_expression THEN result_expression [...n]
[ELSE else_result_expression]

END -

Box 2: ELSE -

Reference:

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

Question #9

DRAG DROP -

You have an Azure Synapse Analytics workspace named WS1.

You have an Azure Data Lake Storage Gen2 container that contains JSON-formatted files in the following format.

```
{
    "id": "66532691-ab20-11ea-8b1d-936b3ec64e54",
    "context": {
        "data": {
            "eventTime": "2020-06-10T13:43:34.553Z",
            "samplingRate": "100.0",
            "isSynthetic": "false"
        },
        "session": {
            "isFirst": "false",
            "id": "38619c14-7a23-4687-8268-95862c5326b1"
        },
        "custom": {
            "dimensions": [
                {
                    "customerInfo": {
                        "ProfileType": "ExpertUser",
                        "RoomName": "",
                        "CustomerName": "diamond",
                        "UserName": "XXXX@yahoo.com"
                    }
                },
                {
                    "customerInfo": {
                        "ProfileType": "Novice",
                        "RoomName": "",
                        "CustomerName": "topaz",
                        "UserName": "XXXX@outlook.com"
                    }
                }
            ]
        }
    }
}
```

You need to use the serverless SQL pool in WS1 to read the files.

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 |
|----------------|---|
| opendatasource | |
| openjson | |
| openquery | |
| openrowset | |
| | select* |
| | FROM |
| | (|
| | BULK 'https://contoso.blob.core.windows.net/contosodw', |
| | FORMAT= 'CSV', |
| | fieldterminator = '0x0b', |
| | fieldquote = '0x0b', |
| | rowterminator = '0x0b' |
| |) |
| | with (id varchar(50), |
| | contextdateeventTime varchar(50) '\$.context.data.eventTime', |
| | contextdatasamplingRate varchar(50) '\$.context.data.samplingRate', |
| | contextdataisSynthetic varchar(50) '\$.context.data.isSynthetic'. |
| | contextsessionisFirst varchar(50) '\$.context.session.isFirst', |
| | contextsession varchar(50) '\$.context.session.id', |
| | contextcustomdimensions varchar(max) '\$.context.custom.dimensions' |
| |) as q |
| | cross apply (contextcustomdimensions) |
| | with (ProfileType varchar(50) '\$.customerInfo.ProfileType', |
| | RoomName varchar(50) '\$.customerInfo.RoomName', |
| | CustomerName varchar(50) '\$.customerInfo.CustomerName', |
| | UserName varchar(50) '\$.customerInfo.UserName' |
| |) |

Correct Answer:

| Values | Answer Area |
|----------------|--|
| | select* |
| | FROM |
| opendatasource | openrowset (|
| | BULK 'https://contoso.blob.core.windows.net/contosodw', FORMAT= 'CSV', fieldterminator = '0x0b', fieldquote = '0x0b', rowterminator = '0x0b' |
| openquery |) with (id varchar(50), contextdateventime varchar(50) '\$.context.data.eventTime', contextdatasamplingRate varchar(50) '\$.context.data.samplingRate', contextdataisSynthetic varchar(50) '\$.context.data.isSynthetic', contextsessionisFirst varchar(50) '\$.context.session isFirst', contextsession varchar(50) '\$.context.session.id', contextcustomdimensions varchar(max) '\$.context.custom.dimensions' |
| |) as q cross apply [openjson (contextcustomdimensions) |
| | with (ProfileType varchar(50) '\$.customerInfo.ProfileType', RoomName varchar(50) '\$.customerInfo.RoomName', CustomerName varchar(50) '\$.customerInfo.CustomerName', UserName varchar(50) '\$.customerInfo.UserName') |

Box 1: openrowset -

The easiest way to see to the content of your CSV file is to provide file URL to OPENROWSET function, specify csv FORMAT.

Example:

```
SELECT *  
FROM OPENROWSET(  
    BULK 'csv/population/population.csv',  
    DATA_SOURCE = 'SqlOnDemandDemo',  
    FORMAT = 'CSV', PARSER_VERSION = '2.0',  
    FIELDTERMINATOR = ',',  
    ROWTERMINATOR = '\n'
```

Box 2: openjson -

You can access your JSON files from the Azure File Storage share by using the mapped drive, as shown in the following example:

```
SELECT book.* FROM -  
OPENROWSET(BULK N't:\books\books.json', SINGLE_CLOB) AS json  
CROSS APPLY OPENJSON(BulkColumn)  
WITH( id nvarchar(100), name nvarchar(100), price float,  
pages_i int, author nvarchar(100)) AS book
```

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/query-single-csv-file>

<https://docs.microsoft.com/en-us/sql/relational-databases/json/import-json-documents-into-sql-server>

Question #10

DRAG DROP -

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

| Date | Temp |
|------------|------|
| ... | ... |
| 18-01-2021 | 3 |
| 19-01-2021 | 4 |
| 20-01-2021 | 2 |
| 21-01-2021 | 2 |
| ... | ... |

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

| Year | JAN | FEB | MAR | APR | MAY |
|------|-----|-----|-----|-----|-----|
| 2019 | 2.3 | 4.1 | 5.2 | 7.6 | 9.2 |
| 2020 | 2.4 | 4.2 | 4.9 | 7.8 | 9.1 |
| 2021 | 2.6 | 5.3 | 3.4 | 7.9 | 9.5 |

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.

Select and Place:

Values **Answer Area**

```
SELECT * FROM (
    SELECT YEAR(Date) Year, MONTH(Date) Month, Temp
    FROM temperatures
    WHERE date BETWEEN DATE '2019-01-01' AND DATE '2021-08-31'
)
(
    AVG ( [ ] (Temp AS DECIMAL(4, 1)))
    FOR Month in (
        1 JAN, 2 FEB, 3 MAR, 4 APR, 5 MAY, 6 JUN,
        7 JUL, 8 AUG, 9 SEP, 10 OCT, 11 NOV, 12 DEC
    )
)
ORDER BY Year ASC
```

CAST
COLLATE
CONVERT
FLATTEN
PIVOT
UNPIVOT

Correct Answer:

Values Answer Area

```
SELECT * FROM (
    SELECT YEAR(Date) Year, MONTH(Date) Month, Temp
    FROM temperatures
    WHERE date BETWEEN DATE '2019-01-01' AND DATE '2021-08-31'
)
PIVOT (
    AVG ( CAST (Temp AS DECIMAL(4, 1)))
    FOR Month in (
        1 JAN, 2 FEB, 3 MAR, 4 APR, 5 MAY, 6 JUN,
        7 JUL, 8 AUG, 9 SEP, 10 OCT, 11 NOV, 12 DEC
    )
)
ORDER BY Year ASC
```

Box 1: PIVOT -

PIVOT rotates a table-valued expression by turning the unique values from one column in the expression into multiple columns in the output. And PIVOT runs aggregations where they're required on any remaining column values that are wanted in the final output.

Incorrect Answers:

UNPIVOT carries out the opposite operation to PIVOT by rotating columns of a table-valued expression into column values.

Box 2: CAST -

If you want to convert an integer value to a DECIMAL data type in SQL Server use the CAST() function.

Example:

```
SELECT -
CAST(12 AS DECIMAL(7,2) ) AS decimal_value;
```

Here is the result:

decimal_value

12.00

Reference:

<https://learnsql.com/cookbook/how-to-convert-an-integer-to-a-decimal-in-sql-server/>

<https://docs.microsoft.com/en-us/sql/t-sql/queries/from-using-pivot-and-unpivot>

Question #11

You have an Azure Data Factory that contains 10 pipelines.

You need to label each pipeline with its main purpose of either ingest, transform, or load. The labels must be available for grouping and filtering when using the monitoring experience in Data Factory.

What should you add to each pipeline?

- A. a resource tag
- B. a correlation ID
- C. a run group ID
- D. an annotation

Correct Answer:

- D. an annotation

Question #12

HOTSPOT -

The following code segment is used to create an Azure Databricks cluster.

```
{  
    "num_workers": null,  
    "autoscale": {  
        "min_workers": 2,  
        "max_workers": 8  
    },  
    "cluster_name": "MyCluster",  
    "spark_version": "latest-stable-scala2.11",  
    "spark_conf": {  
        "spark.databricks.cluster.profile": "serverless",  
        "spark.databricks.repl.allowedLanguages": "sql,python,r"  
    },  
    "node_type_id": "Standard_DS13_v2",  
    "ssh_public_keys": [],  
    "custom_tags": {  
        "ResourceClass": "Serverless"  
    },  
    "spark_env_vars": {  
        "PYSPARK_PYTHON": "/databricks/python3/bin/python3"  
    },  
    "autotermination_minutes": 90,  
    "enable_elastic_disk": true,  
    "init_scripts": []  
}
```

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 Databricks cluster supports multiple concurrent users. | <input type="radio"/> | <input type="radio"/> |
| The Databricks cluster minimizes costs when running scheduled jobs that execute notebooks. | <input type="radio"/> | <input type="radio"/> |
| The Databricks cluster supports the creation of a Delta Lake table. | <input type="radio"/> | <input type="radio"/> |

Correct Answer:

Answer Area

| Statements | Yes | No |
|--|----------------------------------|----------------------------------|
| The Databricks cluster supports multiple concurrent users. | <input checked="" type="radio"/> | <input type="radio"/> |
| The Databricks cluster minimizes costs when running scheduled jobs that execute notebooks. | <input type="radio"/> | <input checked="" type="radio"/> |
| The Databricks cluster supports the creation of a Delta Lake table. | <input checked="" type="radio"/> | <input type="radio"/> |

Box 1: Yes -

A cluster mode of 'High Concurrency' is selected, unlike all the others which are 'Standard'. This results in a worker type of Standard_DS13_v2.

Box 2: No -

When you run a job on a new cluster, the job is treated as a data engineering (job) workload subject to the job workload pricing. When you run a job on an existing cluster, the job is treated as a data analytics (all-purpose) workload subject to all-purpose workload pricing.

Box 3: Yes -

Delta Lake on Databricks allows you to configure Delta Lake based on your workload patterns.

Reference:

<https://adatis.co.uk/databricks-cluster-sizing/>

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

<https://docs.databricks.com/administration-guide/capacity-planning/cmbp.html>

<https://docs.databricks.com/delta/index.html>

Question #13

You are designing a statistical analysis solution that will use custom proprietary Python functions on near real-time data from Azure Event Hubs.

You need to recommend which Azure service to use to perform the statistical analysis. The solution must minimize latency.

What should you recommend?

- A. Azure Synapse Analytics
- B. Azure Databricks
- C. Azure Stream Analytics
- D. Azure SQL Database

Correct Answer:

B. Azure Databricks

Question #14

HOTSPOT -

You have an enterprise data warehouse in Azure Synapse Analytics that contains a table named FactOnlineSales. The table contains data from the start of 2009 to the end of 2012.

You need to improve the performance of queries against FactOnlineSales by using table partitions. The solution must meet the following requirements:

- ☞ Create four partitions based on the order date.
- ☞ Ensure that each partition contains all the orders placed during a given calendar year.

How should you complete the T-SQL command? 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].[FactOnlineSales]
([OnlineSalesKey] [int] NOT NULL,
[OrderDateKey] [datetime] NOT NULL,
[StoreKey] [int] NOT NULL,
[ProductKey] [int] NOT NULL,
[CustomerKey] [int] NOT NULL,
[SalesOrderNumber] [nvarchar](20) NOT NULL,
[SalesQuantity] [int] NOT NULL,
[SalesAmount] [money] NOT NULL,
[UnitPrice] [money] NULL)
WITH (CLUSTERED COLUMNSTORE INDEX)
PARTITION ([OrderDateKey]) RANGE
    FOR VALUES
        RIGHT
        LEFT
(
    20090101,20121231
    20100101,20110101,20120101
    20090101,20100101,20110101,20120101)
```

Correct Answer:

Answer Area

```
CREATE TABLE [dbo].FactOnlineSales
([OnlineSalesKey] [int] NOT NULL,
[OrderDateKey] [datetime] NOT NULL,
[StoreKey] [int] NOT NULL,
[ProductKey] [int] NOT NULL,
[CustomerKey] [int] NOT NULL,
[SalesOrderNumber] [nvarchar] (20) NOT NULL,
[SalesQuantity] [int] NOT NULL,
[SalesAmount] [money] NOT NULL,
[UnitPrice] [money] NULL)
WITH (CLUSTERED COLUMNSTORE INDEX)
PARTITION ([OrderDateKey]) RANGE
    FOR VALUES
        (RIGHT, LEFT)
(
    20090101,20121231
    20100101,20110101,20120101
    20090101,20100101,20110101,20120101
)
```

Range Left or Right, both are creating similar partition but there is difference in comparison

For example: in this scenario, when you use LEFT and 20100101,20110101,20120101

Partition will be, datecol<=20100101, datecol>20100101 and datecol<=20110101,

datecol>20110101 and datecol<=20120101, datecol>20120101

But if you use range RIGHT and 20100101,20110101,20120101

Partition will be, datecol<20100101, datecol>=20100101 and datecol<20110101,

datecol>=20110101 and datecol<20120101, datecol>=20120101

In this example, Range RIGHT will be suitable for calendar comparison Jan 1st to Dec 31st

Reference:

<https://docs.microsoft.com/en-us/sql/t-sql/statements/create-partition-function-transact-sql?view=sql-server-ver15>

Question #15

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 [DBO].[DimProduct] (
[ProductKey] [int] IDENTITY(1,1) NOT NULL,
[ProductSourceID] [int] NOT NULL,
[ProductName] [nvarchar] (100) 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. [EffectiveStartDate] [datetime] NOT NULL,
- B. [CurrentProductCategory] [nvarchar] (100) NOT NULL,
- C. [EffectiveEndDate] [datetime] NULL,
- D. [ProductCategory] [nvarchar] (100) NOT NULL,
- E. [OriginalProductCategory] [nvarchar] (100) NOT NULL,

Correct Answer:

- B. [CurrentProductCategory] [nvarchar] (100) NOT NULL,
- E. [OriginalProductCategory] [nvarchar] (100) NOT NULL

Question #16

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 are designing an Azure Stream Analytics solution that will analyze Twitter data.

You need to count the tweets in each 10-second window. The solution must ensure that each tweet is counted only once.

Solution: You use a hopping window that uses a hop size of 10 seconds and a window size of 10 seconds.

Does this meet the goal?

- A. Yes
- B. No

Correct Answer:

- A. Yes

Question #17

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 are designing an Azure Stream Analytics solution that will analyze Twitter data.

You need to count the tweets in each 10-second window. The solution must ensure that each tweet is counted only once.

Solution: You use a hopping window that uses a hop size of 5 seconds and a window size 10 seconds.

Does this meet the goal?

- A. Yes
- B. No

Correct Answer:

- B. No

Question #18

HOTSPOT -

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

The job receives events based on user actions on the webpage. Each row of data represents an event. Each event has a type of either 'start' or 'end'.

You need to calculate the duration between start and end events.

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

```
SELECT
    [user],
    feature,
    DATEADD(
    DATEDIFF(
    DATEPART(
        second,
        (Time) OVER (PARTITION BY [user], feature LIMIT DURATION(hour, 1) WHEN Event = 'start'),
        ISFIRST
        LAST
        TOPONE
    Time) as duration
FROM input TIMESTAMP BY Time
WHERE
    Event = 'end'
```

Correct Answer:

Answer Area

```
SELECT
    [user],
    feature,
    DATEADD(
    DATEDIFF( LAST
    DATEPART(
        second,
        (Time) OVER (PARTITION BY [user], feature LIMIT DURATION(hour, 1) WHEN Event = 'start'),
        ISFIRST
        LAST
        TOPONE
    Time) as duration
FROM input TIMESTAMP BY Time
WHERE
    Event = 'end'
```

Box 1: DATEDIFF -

DATEDIFF function returns the count (as a signed integer value) of the specified datepart boundaries crossed between the specified startdate and enddate.

Syntax: DATEDIFF (datepart , startdate, enddate)

Box 2: LAST -

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

Example:

```
SELECT -
[user],
feature,
DATEDIFF(
second,
LAST(Time) OVER (PARTITION BY [user], feature LIMIT DURATION(hour,
1) WHEN Event = 'start'),
```

Time) as duration -

FROM input TIMESTAMP BY Time -

WHERE -

Event = 'end'

Reference:

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

Question #19

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

The data flow already contains the following:

- ☞ A source transformation.
- ☞ A Derived Column transformation to set the appropriate types of data.
- ☞ A sink transformation to land the data in the pool.

You need to ensure that the data flow meets the following requirements:

- ☞ All valid rows must be written to the destination table.
- ☞ Truncation errors in the comment column must be avoided proactively.
- ☞ Any rows containing comment values that will cause truncation errors upon insert must be written to a file in blob storage.

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

NOTE: Each correct selection is worth one point.

- A. To the data flow, add a sink transformation to write the rows to a file in blob storage.
- B. To the data flow, add a Conditional Split transformation to separate the rows that will cause truncation errors.
- C. To the data flow, add a filter transformation to filter out rows that will cause truncation errors.
- D. Add a select transformation to select only the rows that will cause truncation errors.

Correct Answer:

- A. To the data flow, add a sink transformation to write the rows to a file in blob storage.
- B. To the data flow, add a Conditional Split transformation to separate the rows that will cause truncation errors.

Question #20

DRAG DROP -

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

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

NOTE: Each correct selection is worth one point.

Select and Place:

| Values | Answer Area |
|--|---|
| all, ecommerce, retail, wholesale | CleanData split(|
| dept=='ecommerce', dept=='retail', dept=='wholesale' | <input type="text"/> |
| dept=='ecommerce', dept== 'wholesale', dept=='retail' | <input type="text"/> |
| disjoint: false |) ~> SplitByDept@(<input type="text"/>) |
| disjoint: true | |
| ecommerce, retail, wholesale, all | |

Correct Answer:

| Values | Answer Area |
|--|---|
| all, ecommerce, retail, wholesale | CleanData split(|
| dept=='ecommerce', dept=='retail', dept=='wholesale' | <input type="text"/> dept=='ecommerce', dept=='retail', dept=='wholesale' |
| dept=='ecommerce', dept== 'wholesale', dept=='retail' | <input type="text"/> disjoint: false |
| disjoint: false |) ~> SplitByDept@(<input type="text"/>) |
| disjoint: true | |
| ecommerce, retail, wholesale, all | ecommerce, retail, wholesale, all |

The conditional split transformation routes data rows to different streams based on matching conditions. The conditional split transformation is similar to a CASE decision structure in a programming language. The transformation evaluates expressions, and based on the results, directs the data row to the specified stream.

Box 1: dept=='ecommerce', dept=='retail', dept=='wholesale'

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

Syntax:

```
<incomingStream>  
split(  
<conditionalExpression1>  
<conditionalExpression2>  
...  
disjoint: {true | false}  
) ~> <splitTx>@(stream1, stream2, ..., <defaultStream>)
```

Box 2: discount : false -

disjoint is false because the data goes to the first matching condition. All remaining rows matching the third condition go to output stream all.

Box 3: ecommerce, retail, wholesale, all

Label the streams -

Reference:

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

Question #21

DRAG DROP -

You have an Azure Data Lake Storage Gen2 account that contains a JSON file for customers. The file contains two attributes named FirstName and LastName.

You need to copy the data from the JSON file to an Azure Synapse Analytics table by using Azure Databricks. A new column must be created that concatenates the FirstName and LastName values.

You create the following components:

- ☞ A destination table in Azure Synapse
- ☞ An Azure Blob storage container
- ☞ A service principal

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

Select and Place:

| Actions | Answer Area |
|--|-------------|
| Mount the Data Lake Storage onto DBFS. | |
| Write the results to a table in Azure Synapse. | |
| Perform transformations on the file. | |
| Specify a temporary folder to stage the data. | |
| Write the results to Data Lake Storage. | |
| Read the file into a data frame. | |
| Drop the data frame. | |
| Perform transformations on the data frame. | |

Correct Answer:

| Actions | Answer Area |
|--|--|
| Mount the Data Lake Storage onto DBFS. | Mount the Data Lake Storage onto DBFS. |
| Write the results to a table in Azure Synapse. | Read the file into a data frame. |
| Perform transformations on the file. | Perform transformations on the data frame. |
| Specify a temporary folder to stage the data. | Specify a temporary folder to stage the data. |
| Write the results to Data Lake Storage. | Write the results to a table in Azure Synapse. |
| Read the file into a data frame. | |
| Drop the data frame. | |
| Perform transformations on the data frame. | |

Step 1: Mount the Data Lake Storage onto DBFS

Begin with creating a file system in the Azure Data Lake Storage Gen2 account.

Step 2: Read the file into a data frame.

You can load the json files as a data frame in Azure Databricks.

Step 3: Perform transformations on the data frame.**Step 4: Specify a temporary folder to stage the data**

Specify a temporary folder to use while moving data between Azure Databricks and Azure Synapse.

Step 5: Write the results to a table in Azure Synapse.

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

Reference:

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

Question #22**HOTSPOT -**

You build an Azure Data Factory pipeline to move data from an Azure Data Lake Storage Gen2 container to a database in an Azure Synapse Analytics dedicated SQL pool.
Data in the container is stored in the following folder structure.

/in/{YYYY}/{MM}/{DD}/{HH}/{mm}

The earliest folder is /in/2021/01/01/00/00. The latest folder is /in/2021/01/15/01/45.

You need to configure a pipeline trigger to meet the following requirements:

- ⇒ Existing data must be loaded.
- ⇒ Data must be loaded every 30 minutes.
- ⇒ Late-arriving data of up to two minutes must be included in the load for the time at which the data should have arrived.

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Type:

| |
|-----------------|
| Event |
| On-demand |
| Schedule |
| Tumbling window |

Additional properties:

| |
|--|
| Prefix: /in/, Event: Blob created |
| Recurrence: 30 minutes, Start time: 2021-01-01T00:00 |
| Recurrence: 30 minutes, Start time: 2021-01-01T00:00, Delay: 2 minutes |
| Recurrence: 32 minutes, Start time: 2021-01-15T01:45 |

Correct Answer:

Answer Area

Type:

| |
|-----------------|
| Event |
| On-demand |
| Schedule |
| Tumbling window |

Additional properties:

| |
|--|
| Prefix: /in/, Event: Blob created |
| Recurrence: 30 minutes, Start time: 2021-01-01T00:00 |
| Recurrence: 30 minutes, Start time: 2021-01-01T00:00, Delay: 2 minutes |
| Recurrence: 32 minutes, Start time: 2021-01-15T01:45 |

Box 1: Tumbling window -

To be able to use the Delay parameter we select Tumbling window.

Box 2:

Recurrence: 30 minutes, not 32 minutes

Delay: 2 minutes.

The amount of time to delay the start of data processing for the window. The pipeline run is started after the expected execution time plus the amount of delay.

The delay defines how long the trigger waits past the due time before triggering a new run. The delay doesn't alter the window startTime.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/how-to-create-tumbling-window-trigger>

Question #23

HOTSPOT -

You are designing a near real-time dashboard solution that will visualize streaming data from remote sensors that connect to the internet. The streaming data must be aggregated to show the average value of each 10-second interval. The data will be discarded after being displayed in the dashboard.

The solution will use Azure Stream Analytics and must meet the following requirements:

- ☞ Minimize latency from an Azure Event hub to the dashboard.
- ☞ Minimize the required storage.
- ☞ Minimize development effort.

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

Azure Stream Analytics input type:

| |
|------------------------|
| Azure Event Hub |
| Azure SQL Database |
| Azure Stream Analytics |
| Microsoft Power BI |

Azure Stream Analytics output type:

| |
|------------------------|
| Azure Event Hub |
| Azure SQL Database |
| Azure Stream Analytics |
| Microsoft Power BI |

Aggregation query location:

| |
|------------------------|
| Azure Event Hub |
| Azure SQL Database |
| Azure Stream Analytics |
| Microsoft Power BI |

Correct Answer:

Answer Area

Azure Stream Analytics input type:

| |
|------------------------|
| Azure Event Hub |
| Azure SQL Database |
| Azure Stream Analytics |
| Microsoft Power BI |

Azure Stream Analytics output type:

| |
|------------------------|
| Azure Event Hub |
| Azure SQL Database |
| Azure Stream Analytics |
| Microsoft Power BI |

Aggregation query location:

| |
|------------------------|
| Azure Event Hub |
| Azure SQL Database |
| Azure Stream Analytics |
| Microsoft Power BI |

Reference:

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-power-bi-dashboard>

Question #24

DRAG DROP -

You have an Azure Stream Analytics job that is a Stream Analytics project solution in Microsoft Visual Studio. The job accepts data generated by IoT devices in the JSON format.

You need to modify the job to accept data generated by the IoT devices in the Protobuf format. Which three actions should you perform from Visual Studio on sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

Select and Place:

| Actions | Answer Area |
|--|-------------|
| Change the Event Serialization Format to Protobuf in the input.json file of the job and reference the DLL. | |
| Add an Azure Stream Analytics Custom Deserializer Project (.NET) project to the solution. | |
| Add .NET deserializer code for Protobuf to the custom deserializer project. | |
| Add .NET deserializer code for Protobuf to the Stream Analytics project. | |
| Add an Azure Stream Analytics Application project to the solution. | |

Correct Answer:

| Actions | Answer Area |
|--|---|
| Change the Event Serialization Format to Protobuf in the input.json file of the job and reference the DLL. | Add an Azure Stream Analytics Custom Deserializer Project (.NET) project to the solution. |
| Add an Azure Stream Analytics Custom Deserializer Project (.NET) project to the solution. | Add .NET deserializer code for Protobuf to the custom deserializer project. |
| Add .NET deserializer code for Protobuf to the custom deserializer project. | Add an Azure Stream Analytics Application project to the solution. |
| Add .NET deserializer code for Protobuf to the Stream Analytics project. | |
| Add an Azure Stream Analytics Application project to the solution. | |

Step 1: Add an Azure Stream Analytics Custom Deserializer Project (.NET) project to the solution.

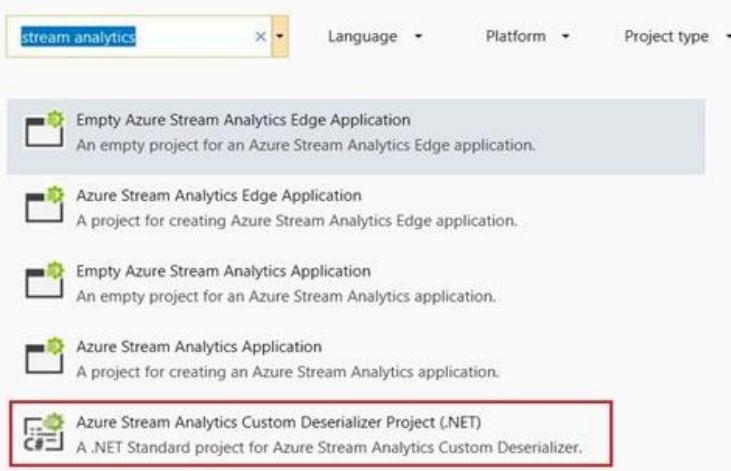
Create a custom deserializer -

1. Open Visual Studio and select File > New > Project. Search for Stream Analytics and select Azure Stream Analytics Custom Deserializer Project (.NET). Give the project a name, like Protobuf Deserializer.

Create a new project

Recent project templates

A list of your recently accessed templates will be displayed here.



2. In Solution Explorer, right-click your Protobuf Deserializer project and select Manage NuGet Packages from the menu. Then install the Microsoft.Azure.StreamAnalytics and Google.Protobuf NuGet packages.
3. Add the MessageBodyProto class and the MessageBodyDeserializer class to your project.
4. Build the Protobuf Deserializer project.

Step 2: Add .NET deserializer code for Protobuf to the custom deserializer project

Azure Stream Analytics has built-in support for three data formats: JSON, CSV, and Avro. With custom .NET deserializers, you can read data from other formats such as Protocol Buffer, Bond and other user defined formats for both cloud and edge jobs.

Step 3: Add an Azure Stream Analytics Application project to the solution

Add an Azure Stream Analytics project

1. In Solution Explorer, right-click the Protobuf Deserializer solution and select Add > New Project. Under Azure Stream Analytics > Stream Analytics, choose Azure Stream Analytics Application. Name it ProtobufCloudDeserializer and select OK.
2. Right-click References under the ProtobufCloudDeserializer Azure Stream Analytics project. Under Projects, add Protobuf Deserializer. It should be automatically populated for you.

Reference:

<https://docs.microsoft.com/en-us/azure/stream-analytics/custom-deserializer>

Question #25

You have an Azure Storage account and a data warehouse in Azure Synapse Analytics in the UK South region.

You need to copy blob data from the storage account to the data warehouse by using Azure Data Factory. The solution must meet the following requirements:

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

Which type of integration runtime should you use?

- A. Azure integration runtime
- B. Azure-SSIS integration runtime
- C. Self-hosted integration runtime

Correct Answer:

- A. Azure integration runtime

Question #26

HOTSPOT -

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

| Source | Data |
|-----------|--|
| Database1 | Driver's name Driver's license number |
| HubA | Ride route Ride distance Ride duration |
| HubB | Ride fare Ride payment |

You need to implement Azure Stream Analytics to calculate the average fare per mile by driver. How should you configure the Stream Analytics input for each source? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

| | | | |
|------------|---|--------|-----------|
| HubA: | <table border="1"><tr><td>Stream</td></tr><tr><td>Reference</td></tr></table> | Stream | Reference |
| Stream | | | |
| Reference | | | |
| HubB: | <table border="1"><tr><td>Stream</td></tr><tr><td>Reference</td></tr></table> | Stream | Reference |
| Stream | | | |
| Reference | | | |
| Database1: | <table border="1"><tr><td>Stream</td></tr><tr><td>Reference</td></tr></table> | Stream | Reference |
| Stream | | | |
| Reference | | | |

Correct Answer:

Answer Area

| | | | |
|------------|---|--------|-----------|
| HubA: | <table border="1"><tr><td>Stream</td></tr><tr><td>Reference</td></tr></table> | Stream | Reference |
| Stream | | | |
| Reference | | | |
| HubB: | <table border="1"><tr><td>Stream</td></tr><tr><td>Reference</td></tr></table> | Stream | Reference |
| Stream | | | |
| Reference | | | |
| Database1: | <table border="1"><tr><td>Stream</td></tr><tr><td>Reference</td></tr></table> | Stream | Reference |
| Stream | | | |
| Reference | | | |

HubA: Stream -

HubB: Stream -

Database1: Reference -

Reference data (also known as a lookup table) is a finite data set that is static or slowly changing in nature, used to perform a lookup or to augment your data streams. For example, in

an IoT scenario, you could store metadata about sensors (which don't change often) in reference data and join it with real time IoT data streams. Azure Stream Analytics loads reference data in memory to achieve low latency stream processing

Reference:

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

Question #27

You have an Azure Stream Analytics job that receives clickstream data from an Azure event hub.

You need to define a query in the Stream Analytics job. The query must meet the following requirements:

- ☞ Count the number of clicks within each 10-second window based on the country of a visitor.
- ☞ Ensure that each click is NOT counted more than once.

How should you define the Query?

- A. SELECT Country, Avg(*) AS Average FROM ClickStream TIMESTAMP BY CreatedAt GROUP BY Country, SlidingWindow(second, 10)
- B. SELECT Country, Count(*) AS Count FROM ClickStream TIMESTAMP BY CreatedAt GROUP BY Country, TumblingWindow(second, 10)
- C. SELECT Country, Avg(*) AS Average FROM ClickStream TIMESTAMP BY CreatedAt GROUP BY Country, HoppingWindow(second, 10, 2)
- D. SELECT Country, Count(*) AS Count FROM ClickStream TIMESTAMP BY CreatedAt GROUP BY Country, SessionWindow(second, 5, 10)

Correct Answer:

- B. SELECT Country, Count(*) AS Count FROM ClickStream TIMESTAMP BY CreatedAt GROUP BY Country, TumblingWindow(second, 10)

Question #28

HOTSPOT -

You are building an Azure Analytics query that will receive input data from Azure IoT Hub and write the results to Azure Blob storage.

You need to calculate the difference in the number of readings per sensor per hour.

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

```
SELECT sensorId,
       growth = reading -
                  (reading) OVER (PARTITION BY sensorId
                                LAG
                                LAST
                                LEAD
                                (hour,1))
                                LIMIT DURATION
                                OFFSET
                                WHEN
FROM input
```

Correct Answer:**Answer Area**

```
SELECT sensorId,
       growth = reading -
                  (reading) OVER (PARTITION BY sensorId
                                LAG
                                (hour,1))
                                LIMIT DURATION
                                OFFSET
                                WHEN
FROM input
```

Box 1: LAG -

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

Box 2: LIMIT DURATION -

Example: Compute the rate of growth, per sensor:

SELECT sensorId,

growth = reading -

LAG(reading) OVER (PARTITION BY sensorId LIMIT DURATION(hour, 1))

FROM input -

Reference:

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

Question #29

You need to schedule an Azure Data Factory pipeline to execute when a new file arrives in an Azure Data Lake Storage Gen2 container.

Which type of trigger should you use?

- A. on-demand
- B. tumbling window
- C. schedule
- D. event

Correct Answer:

- D. event

Question #30

You have two Azure Data Factory instances named ADFdev and ADFprod. ADFdev connects to an Azure DevOps Git repository.

You publish changes from the main branch of the Git repository to ADFdev.

You need to deploy the artifacts from ADFdev to ADFprod.

What should you do first?

- A. From ADFdev, modify the Git configuration.
- B. From ADFdev, create a linked service.
- C. From Azure DevOps, create a release pipeline.
- D. From Azure DevOps, update the main branch.

Correct Answer:

C. From Azure DevOps, create a release pipeline.

Question #31

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 Cosmos DB
- B. Azure Blob storage
- C. Azure IoT Hub
- D. Azure Event Hubs

Correct Answer:

B. Azure Blob storage

Question #32

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

You need to process the events to produce a running average of shopper counts during the previous 15 minutes, calculated at five-minute intervals.

Which type of window should you use?

- A. snapshot
- B. tumbling
- C. hopping
- D. sliding

Correct Answer:

C. hopping

Question #33

HOTSPOT -

You are designing a monitoring solution for a fleet of 500 vehicles. Each vehicle has a GPS tracking device that sends data to an Azure event hub once per minute.

You have a CSV file in an Azure Data Lake Storage Gen2 container. The file maintains the expected geographical area in which each vehicle should be.

You need to ensure that when a GPS position is outside the expected area, a message is added to another event hub for processing within 30 seconds. The solution must minimize cost.

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

Service:

- An Azure Synapse Analytics Apache Spark pool
- An Azure Synapse Analytics serverless SQL pool
- Azure Data Factory
- Azure Stream Analytics

Window:

- Hopping
- No window
- Session
- Tumbling

Analysis type:

- Event pattern matching
- Lagged record comparison
- Point within polygon
- Polygon overlap

Correct Answer:

Answer Area

| | |
|----------------|---|
| Service: | <ul style="list-style-type: none">An Azure Synapse Analytics Apache Spark poolAn Azure Synapse Analytics serverless SQL poolAzure Data FactoryAzure Stream Analytics |
| Window: | <ul style="list-style-type: none">HoppingNo windowSessionTumbling |
| Analysis type: | <ul style="list-style-type: none">Event pattern matchingLagged record comparisonPoint within polygonPolygon overlap |

Box 1: Azure Stream Analytics -

Box 2: Hopping -

Hopping window functions hop forward in time by a fixed period. It may be easy to think of them as Tumbling windows that can overlap and be emitted more often than the window size. Events can belong to more than one Hopping window result set. To make a Hopping window the same as a Tumbling window, specify the hop size to be the same as the window size.

Box 3: Point within polygon -

Reference:

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

Question #34

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

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

What should you include in the solution?

- A. Partition by DateTime fields.
- B. Sink to Azure Queue storage.
- C. Include a watermark column.
- D. Use a JSON format for physical data storage.

Correct Answer:

- B. Sink to Azure Queue storage.

Question #35

HOTSPOT -

You have a self-hosted integration runtime in Azure Data Factory.

The current status of the integration runtime has the following configurations:

- ☞ Status: Running
- ☞ Type: Self-Hosted
- ☞ Version: 4.4.7292.1
- ☞ Running / Registered Node(s): 1/1
- ☞ High Availability Enabled: False
- ☞ Linked Count: 0
- ☞ Queue Length: 0
- ☞ Average Queue Duration. 0.00s

The integration runtime has the following node details:

- ☞ Name: X-M
- ☞ Status: Running
- ☞ Version: 4.4.7292.1
- ☞ Available Memory: 7697MB
- ☞ CPU Utilization: 6%
- ☞ Network (In/Out): 1.21KBps/0.83KBps
- ☞ Concurrent Jobs (Running/Limit): 2/14
- ☞ Role: Dispatcher/Worker
- ☞ Credential Status: In Sync

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

If the X-M node becomes unavailable, all executed pipelines will:

| |
|---------------------------------------|
| fail until the node comes back online |
| switch to another integration runtime |
| exceed the CPU limit |

The number of concurrent jobs and the CPU usage indicate that the Concurrent Jobs (Running/Limit) value should be:

| |
|------------|
| raised |
| lowered |
| left as is |

Correct Answer:

Answer Area

If the X-M node becomes unavailable, all executed pipelines will:

| |
|---------------------------------------|
| fail until the node comes back online |
| switch to another integration runtime |
| exceed the CPU limit |

The number of concurrent jobs and the CPU usage indicate that the Concurrent Jobs (Running/Limit) value should be:

| |
|------------|
| raised |
| lowered |
| left as is |

Box 1: fail until the node comes back online

We see: High Availability Enabled: False

Note: Higher availability of the self-hosted integration runtime so that it's no longer the single point of failure in your big data solution or cloud data integration with Data Factory.

Box 2: lowered -

We see:

Concurrent Jobs (Running/Limit): 2/14

CPU Utilization: 6%

Note: When the processor and available RAM aren't well utilized, but the execution of concurrent jobs reaches a node's limits, scale up by increasing the number of concurrent jobs that a node can run

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/create-self-hosted-integration-runtime>

Question #36

You have an Azure Databricks workspace named workspace1 in the Standard pricing tier. You need to configure workspace1 to support autoscaling all-purpose clusters. The solution must meet the following requirements:

- ☞ Automatically scale down workers when the cluster is underutilized for three minutes.
- ☞ Minimize the time it takes to scale to the maximum number of workers.
- ☞ Minimize costs.

What should you do first?

- A. Enable container services for workspace1.
- B. Upgrade workspace1 to the Premium pricing tier.
- C. Set Cluster Mode to High Concurrency.
- D. Create a cluster policy in workspace1.

Correct Answer:

B. Upgrade workspace1 to the Premium pricing tier.

Question #37

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 are designing an Azure Stream Analytics solution that will analyze Twitter data.

You need to count the tweets in each 10-second window. The solution must ensure that each tweet is counted only once.

Solution: You use a tumbling window, and you set the window size to 10 seconds.

Does this meet the goal?

- A. Yes
- B. No

Correct Answer:

A. Yes

Question #38

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 are designing an Azure Stream Analytics solution that will analyze Twitter data.

You need to count the tweets in each 10-second window. The solution must ensure that each tweet is counted only once.

Solution: You use a session window that uses a timeout size of 10 seconds.

Does this meet the goal?

- A. Yes
- B. No

Correct Answer:

- B. No

Question #39

You use Azure Stream Analytics to receive data from Azure Event Hubs and to output the data to an Azure Blob Storage account.

You need to output the count of records received from the last five minutes every minute.

Which windowing function should you use?

- A. Session
- B. Tumbling
- C. Sliding
- D. Hopping

Correct Answer:

- D. Hopping

Question #40

HOTSPOT -

You configure version control for an Azure Data Factory instance as shown in the following exhibit.

The screenshot shows the 'Git configuration' section of the Azure Data Factory interface. On the left, there's a sidebar with icons for Home, Connections, Linked services, Integration runtimes, Source control (selected), Git configuration (selected), ARM template, Parameterization template, Author (Triggers, Global parameters), Security (Customer managed key, Managed private endpoints), and Publish branch. The main panel is titled 'Git repository' and displays the following configuration:

| Setting | Value |
|----------------------|------------------|
| Repository type | Azure DevOps Git |
| Azure DevOps Account | CONTOSO |
| Project name | Data |
| Repository name | dwh_batchetl |
| Collaboration branch | main |
| Publish branch | adf_publish |
| Root folder | / |

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

Azure Resource Manager (ARM) templates for the pipeline assets are stored in [answer choice]

| |
|---------------------------|
| / |
| adf_publish |
| main |
| Parameterization template |

A Data Factory Azure Resource Manager (ARM) template named contososales can be found in [answer choice]

| |
|--|
| / |
| /contososales |
| /dwh_batchetl/adf_publish/contososales |
| /main |

Correct Answer:

Answer Area

Azure Resource Manager (ARM) templates for the pipeline assets are stored in [answer choice]

| | |
|---------------------------|---|
| / | ▼ |
| adf_publish | |
| main | |
| Parameterization template | |

A Data Factory Azure Resource Manager (ARM) template named contososales can be found in [answer choice]

| | |
|--|---|
| / | ▼ |
| /contososales | |
| /dwh_batchetl/adf_publish/contososales | |
| /main | |

Box 1: adf_publish -

The Publish branch is the branch in your repository where publishing related ARM templates are stored and updated. By default, it's adf_publish.

Box 2: / dwh_batchetl/adf_publish/contososales

Note: RepositoryName (here dwh_batchetl): Your Azure Repos code repository name. Azure Repos projects contain Git repositories to manage your source code as your project grows. You can create a new repository or use an existing repository that's already in your project.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/source-control>

Question #41

HOTSPOT -

You are designing an Azure Stream Analytics solution that receives instant messaging data from an Azure Event Hub.

You need to ensure that the output from the Stream Analytics job counts the number of messages per time zone every 15 seconds.

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Select TimeZone, count (*) AS MessageCount

FROM MessageStream

| |
|--------------------|
| ▼ |
| LAST |
| OVER |
| SYSTEM.TIMESTAMP() |
| TIMESTAMP BY |

CreatedAt

GROUP BY TimeZone,

| |
|----------------|
| ▼ |
| HOPPINGWINDOW |
| SESSIONWINDOW |
| SLIDINGWINDOW |
| TUMBLINGWINDOW |

(second,15)

Correct Answer:

Answer Area

Select TimeZone, count (*) AS MessageCount

FROM MessageStream

| |
|--------------------|
| ▼ |
| LAST |
| OVER |
| SYSTEM.TIMESTAMP() |
| TIMESTAMP BY |

CreatedAt

GROUP BY TimeZone,

| |
|----------------|
| ▼ |
| HOPPINGWINDOW |
| SESSIONWINDOW |
| SLIDINGWINDOW |
| TUMBLINGWINDOW |

(second,15)

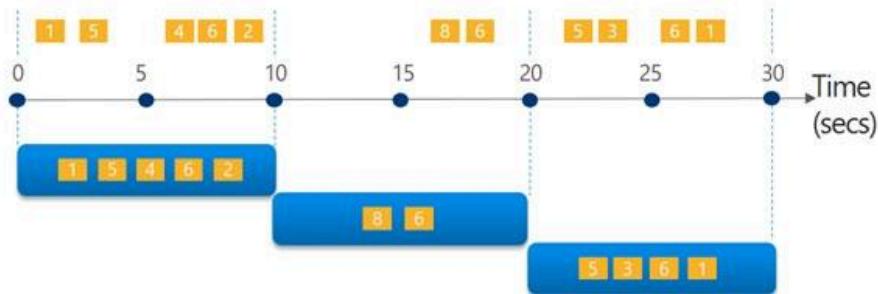
Box 1: timestamp by -

Box 2: TUMBLINGWINDOW -

Tumbling window functions are used to segment a data stream into distinct time segments and perform a function against them, such as the example below. The key differentiators of a Tumbling window are that they repeat, do not overlap, and an event cannot belong to more than one tumbling window.

Tell me the count of Tweets per time zone every 10 seconds

A 10-second Tumbling Window



```
SELECT TimeZone, COUNT(*) AS Count
FROM TwitterStream TIMESTAMP BY CreatedAt
GROUP BY TimeZone, TumblingWindow(second,10)
```

Reference:

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

Question #42

HOTSPOT -

You have an Azure Data Factory instance named ADF1 and two Azure Synapse Analytics workspaces named WS1 and WS2.

ADF1 contains the following pipelines:

- ☞ P1: Uses a copy activity to copy data from a nonpartitioned table in a dedicated SQL pool of WS1 to an Azure Data Lake Storage Gen2 account
- ☞ P2: Uses a copy activity to copy data from text-delimited files in an Azure Data Lake Storage Gen2 account to a nonpartitioned table in a dedicated SQL pool of WS2

You need to configure P1 and P2 to maximize parallelism and performance.

Which dataset settings should you configure for the copy activity if each pipeline? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

P1:

- | | |
|--|---|
| Set the Copy method to Bulk insert | ▼ |
| Set the Copy method to PolyBase | ▼ |
| Set the Isolation level to Repeatable read | ▼ |
| Set the Partition option to Dynamic range | ▼ |

P2:

- | | |
|--|---|
| Set the Copy method to Bulk insert | ▼ |
| Set the Copy method to PolyBase | ▼ |
| Set the Isolation level to Repeatable read | ▼ |
| Set the Partition option to Dynamic range | ▼ |

Correct Answer:

Answer Area

P1:

- | | |
|--|---|
| Set the Copy method to Bulk insert | ▼ |
| Set the Copy method to PolyBase | ▼ |
| Set the Isolation level to Repeatable read | ▼ |
| Set the Partition option to Dynamic range | ▼ |

P2:

- | | |
|--|---|
| Set the Copy method to Bulk insert | ▼ |
| Set the Copy method to PolyBase | ▼ |
| Set the Isolation level to Repeatable read | ▼ |
| Set the Partition option to Dynamic range | ▼ |

Box 1: Set the Copy method to PolyBase

While SQL pool supports many loading methods including non-Polybase options such as BCP and SQL BulkCopy API, the fastest and most scalable way to load data is through PolyBase. PolyBase is a technology that accesses external data stored in Azure Blob storage or Azure Data Lake Store via the T-SQL language.

Box 2: Set the Copy method to Bulk insert

Polybase not possible for text files. Have to use Bulk insert.

Reference:

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

Question #43

HOTSPOT -

You have an Azure Storage account that generates 200,000 new files daily. The file names have a format of {YYYY}/{MM}/{DD}/{HH}/{CustomerID}.csv.

You need to design an Azure Data Factory solution that will load new data from the storage account to an Azure Data Lake once hourly. The solution must minimize load times and costs. How should you configure the solution? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Load methodology:

| | |
|--------------------------------------|--|
| | |
| Full Load | |
| Incremental Load | |
| Load individual files as they arrive | |

Trigger:

| | |
|-----------------|--|
| | |
| Fixed schedule | |
| New file | |
| Tumbling window | |

Correct Answer:

Answer Area

Load methodology:

| | |
|--------------------------------------|---|
| Full Load | ▼ |
| Incremental Load | |
| Load individual files as they arrive | |

Trigger:

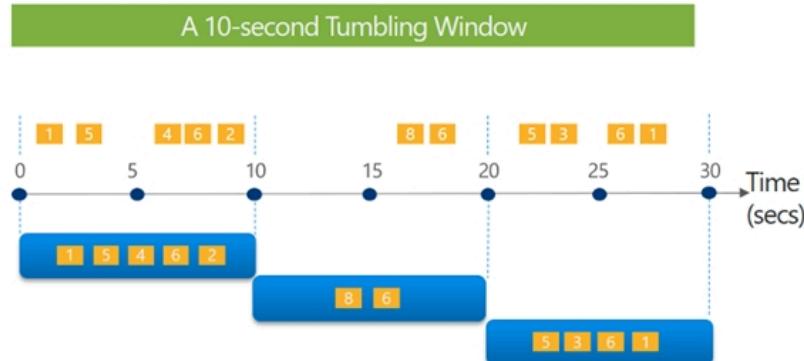
| | |
|-----------------|---|
| Fixed schedule | ▼ |
| New file | |
| Tumbling window | |

Box 1: Incremental load -

Box 2: Tumbling window -

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

Tell me the count of tweets per time zone every 10 seconds



```
SELECT TimeZone, COUNT(*) AS Count
FROM TwitterStream TIMESTAMP BY CreatedAt
GROUP BY TimeZone, TumblingWindow(second,10)
```

Reference:

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

Question #44

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 plan to create an Azure Databricks workspace that has a tiered structure. The workspace will contain the following three workloads:

- ☞ A workload for data engineers who will use Python and SQL.
- ☞ A workload for jobs that will run notebooks that use Python, Scala, and SQL.
- ☞ A workload that data scientists will use to perform ad hoc analysis in Scala and R.

The enterprise architecture team at your company identifies the following standards for Databricks environments:

- ☞ The data engineers must share a cluster.
- ☞ The job cluster will be managed by using a request process whereby data scientists and data engineers provide packaged notebooks for deployment to the cluster.
- ☞ All the data scientists must be assigned their own cluster that terminates automatically after 120 minutes of inactivity. Currently, there are three data scientists.

You need to create the Databricks clusters for the workloads.

Solution: You create a Standard cluster for each data scientist, a Standard cluster for the data engineers, and a High Concurrency cluster for the jobs.

Does this meet the goal?

- A. Yes
- B. No

Correct Answer:

- B. No

Question #45

You have the following Azure Data Factory pipelines:

- ☞ Ingest Data from System1
- ☞ Ingest Data from System2
- ☞ Populate Dimensions
- ☞ Populate Facts

Ingest Data from System1 and Ingest Data from System2 have no dependencies. Populate Dimensions must execute after Ingest Data from System1 and Ingest Data from System2. Populate Facts must execute after Populate Dimensions pipeline. All the pipelines must execute every eight hours.

What should you do to schedule the pipelines for execution?

- A. Add an event trigger to all four pipelines.
- B. Add a schedule trigger to all four pipelines.
- C. Create a patient pipeline that contains the four pipelines and use a schedule trigger.
- D. Create a patient pipeline that contains the four pipelines and use an event trigger.

Correct Answer:

- C. Create a patient pipeline that contains the four pipelines and use a schedule trigger.

Question #46

DRAG DROP -

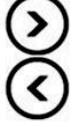
You are responsible for providing access to an Azure Data Lake Storage Gen2 account. Your user account has contributor access to the storage account, and you have the application ID and access key.

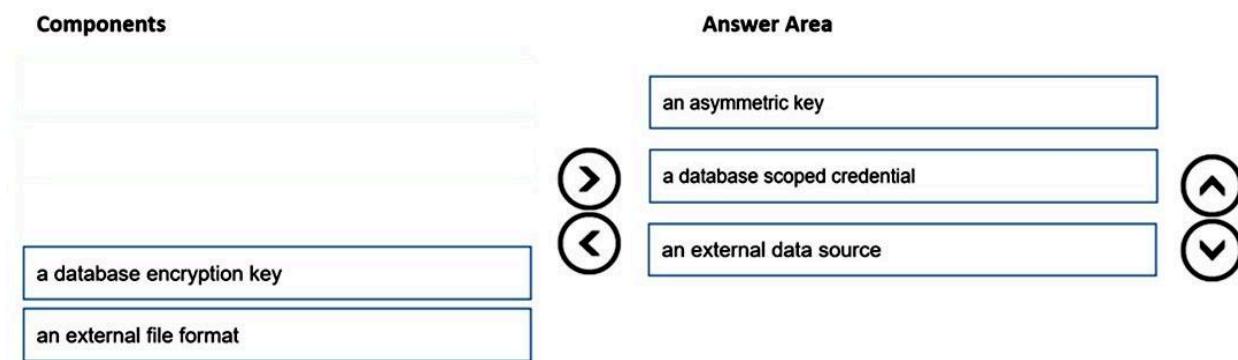
You plan to use PolyBase to load data into an enterprise data warehouse in Azure Synapse Analytics.

You need to configure PolyBase to connect the data warehouse to storage account.

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

Select and Place:

| Components | Answer Area |
|------------------------------|---|
| a database scoped credential | |
| an asymmetric key | |
| an external data source |   |
| a database encryption key | |
| an external file format |   |

Correct Answer:

Step 1: an asymmetric key -

A master key should be created only once in a database. The Database Master Key is a symmetric key used to protect the private keys of certificates and asymmetric keys in the database.

Step 2: a database scoped credential

Create a Database Scoped Credential. A Database Scoped Credential is a record that contains the authentication information required to connect an external resource. The master key needs to be created first before creating the database scoped credential.

Step 3: an external data source -

Create an External Data Source. External data sources are used to establish connectivity for data loading using Polybase.

Reference:

<https://www.sqlservercentral.com/articles/access-external-data-from-azure-synapse-analytics-using-polybase>

Question #47

You are monitoring an Azure Stream Analytics job by using metrics in Azure.

You discover that during the last 12 hours, the average watermark delay is consistently greater than the configured late arrival tolerance.

What is a possible cause of this behavior?

- A. Events whose application timestamp is earlier than their arrival time by more than five minutes arrive as inputs.
- B. There are errors in the input data.
- C. The late arrival policy causes events to be dropped.
- D. The job lacks the resources to process the volume of incoming data.

Correct Answer:

- D. The job lacks the resources to process the volume of incoming data.

Question #48

HOTSPOT -

You are building an Azure Stream Analytics job to retrieve game data.

You need to ensure that the job returns the highest scoring record for each five-minute time interval of each game.

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

SELECT

| |
|--|
| Collect(Score) |
| CollectTop(1) OVER(ORDER BY Score Desc) |
| Game, MAX(Score) |
| TopOne() OVER(PARTITION BY Game ORDER BY Score Desc) |

as HighestScore

FROM input TIMESTAMP BY CreatedAt

GROUP BY

| |
|---|
| Game |
| Hopping(minute,5) |
| Tumbling(minute,5) |
| Windows(TumblingWindow(minute,5),Hopping(minute,5)) |

Correct Answer:

Answer Area

SELECT

| |
|--|
| Collect(Score) |
| CollectTop(1) OVER(ORDER BY Score Desc) |
| Game, MAX(Score) |
| TopOne() OVER(PARTITION BY Game ORDER BY Score Desc) |

as HighestScore

FROM input TIMESTAMP BY CreatedAt

GROUP BY

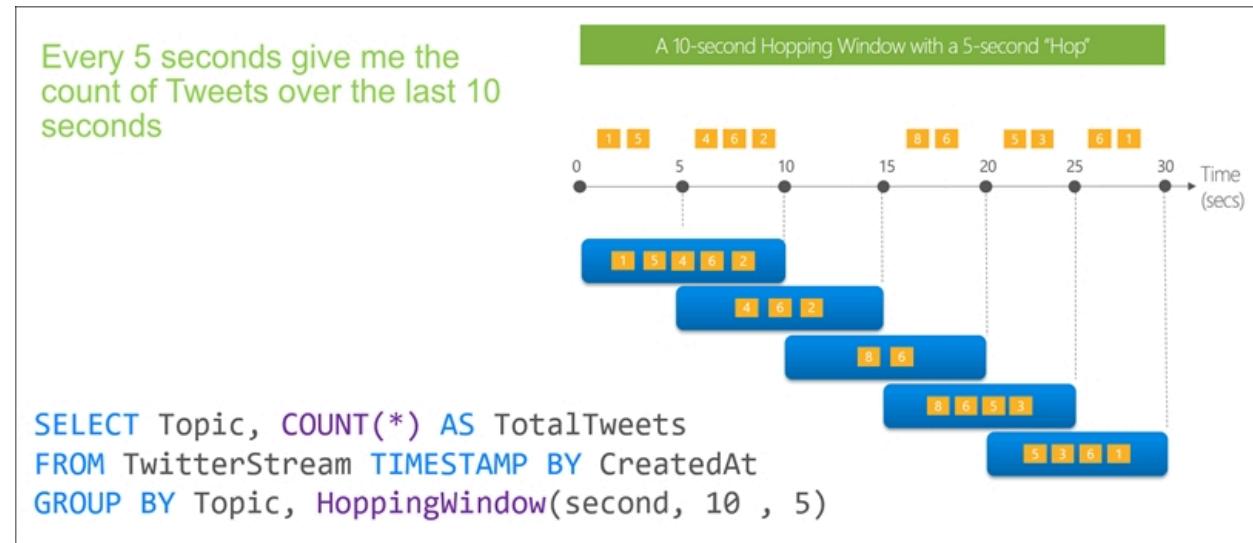
| |
|---|
| Game |
| Hopping(minute,5) |
| Tumbling(minute,5) |
| Windows(TumblingWindow(minute,5),Hopping(minute,5)) |

Box 1: TopOne OVER(PARTITION BY Game ORDER BY Score Desc)

TopOne returns the top-rank record, where rank defines the ranking position of the event in the window according to the specified ordering. Ordering/ranking is based on event columns and can be specified in ORDER BY clause.

Box 2: Hopping(minute,5)

Hopping window functions hop forward in time by a fixed period. It may be easy to think of them as Tumbling windows that can overlap and be emitted more often than the window size. Events can belong to more than one Hopping window result set. To make a Hopping window the same as a Tumbling window, specify the hop size to be the same as the window size.



Reference:

<https://docs.microsoft.com/en-us/stream-analytics-query/topone-azure-stream-analytics>

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

Question #49

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

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

You have an Azure Data Lake Storage account that contains a staging zone.

You need to design a daily process to ingest incremental data from the staging zone, transform the data by executing an R script, and then insert the transformed data into a data warehouse in Azure Synapse Analytics.

Solution: You use an Azure Data Factory schedule trigger to execute a pipeline that copies the data to a staging table in the data warehouse, and then uses a stored procedure to execute the R script.

Does this meet the goal?

- A. Yes
- B. No

Correct Answer:

- B. No

Question #50

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 plan to create an Azure Databricks workspace that has a tiered structure. The workspace will contain the following three workloads:

- ☞ A workload for data engineers who will use Python and SQL.
- ☞ A workload for jobs that will run notebooks that use Python, Scala, and SQL.
- ☞ A workload that data scientists will use to perform ad hoc analysis in Scala and R.

The enterprise architecture team at your company identifies the following standards for Databricks environments:

- ☞ The data engineers must share a cluster.
- ☞ The job cluster will be managed by using a request process whereby data scientists and data engineers provide packaged notebooks for deployment to the cluster.
- ☞ All the data scientists must be assigned their own cluster that terminates automatically after 120 minutes of inactivity. Currently, there are three data scientists.

You need to create the Databricks clusters for the workloads.

Solution: You create a High Concurrency cluster for each data scientist, a High Concurrency cluster for the data engineers, and a Standard cluster for the jobs.

Does this meet the goal?

- A. Yes
- B. No

Correct Answer:

- B. No

Question #51

You are designing an Azure Databricks cluster that runs user-defined local processes.

You need to recommend a cluster configuration that meets the following requirements:

- ☞ Minimize query latency.
- ☞ Maximize the number of users that can run queries on the cluster at the same time.
- ☞ Reduce overall costs without compromising other requirements.

Which cluster type should you recommend?

- A. Standard with Auto Termination
- B. High Concurrency with Autoscaling
- C. High Concurrency with Auto Termination
- D. Standard with Autoscaling

Correct Answer:

B. High Concurrency with Autoscaling

Question #52

HOTSPOT -

You are building an Azure Data Factory solution to process data received from Azure Event Hubs, and then ingested into an Azure Data Lake Storage Gen2 container.

The data will be ingested every five minutes from devices into JSON files. The files have the following naming pattern.

`/{deviceType}/in/{YYYY}/{MM}/{DD}/{HH}/{deviceId}_{YYYY}{MM}{DD}{HH}{mm}.json`

You need to prepare the data for batch data processing so that there is one dataset per hour per deviceType. The solution must minimize read times.

How should you configure the sink for 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

Parameter:

| | |
|---|---|
| <code>@pipeline(),TriggerTime</code> | ▼ |
| <code>@pipeline(),TriggerType</code> | |
| <code>@trigger().outputs.windowStartTime</code> | |
| <code>@trigger().startTime</code> | |

Naming pattern:

| | |
|---|---|
| <code>/{deviceID}/out/{YYYY}/{MM}/{DD}/{HH}.json</code> | ▼ |
| <code>/{YYYY}/{MM}/{DD}/{deviceType}.json</code> | |
| <code>/{YYYY}/{MM}/{DD}/{HH}.json</code> | |
| <code>/{YYYY}/{MM}/{DD}/{HH}_{deviceType}.json</code> | |

Copy behavior:

| | |
|----------------------------------|---|
| <code>Add dynamic content</code> | ▼ |
| <code>Flatten hierarchy</code> | |
| <code>Merge files</code> | |

Correct Answer:

Answer Area

Parameter:

| | |
|---|---|
| <code>@pipeline(),TriggerTime</code> | ▼ |
| <code>@pipeline(),TriggerType</code> | |
| <code>@trigger().outputs.windowStartTime</code> | |
| <code>@trigger().startTime</code> | |

Naming pattern:

| | |
|---|---|
| <code>/{deviceID}/out/{YYYY}/{MM}/{DD}/{HH}.json</code> | ▼ |
| <code>/{YYYY}/{MM}/{DD}/{deviceType}.json</code> | |
| <code>/{YYYY}/{MM}/{DD}/{HH}.json</code> | |
| <code>/{YYYY}/{MM}/{DD}/{HH}_{deviceType}.json</code> | |

Copy behavior:

| | |
|----------------------------------|---|
| <code>Add dynamic content</code> | ▼ |
| <code>Flatten hierarchy</code> | |
| <code>Merge files</code> | |

Box 1: `@trigger().startTime` -

startTime: A date-time value. For basic schedules, the value of the startTime property applies to the first occurrence. For complex schedules, the trigger starts no sooner than the specified startTime value.

Box 2: /{YYYY}/{MM}/{DD}/{HH}_{deviceType}.json

One dataset per hour per deviceType.

Box 3: Flatten hierarchy -

- FlattenHierarchy: All files from the source folder are in the first level of the target folder. The target files have autogenerated names.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipeline-execution-triggers>

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

Question #53

DRAG DROP -

You are designing an Azure Data Lake Storage Gen2 structure for telemetry data from 25 million devices distributed across seven key geographical regions. Each minute, the devices will send a JSON payload of metrics to Azure Event Hubs.

You need to recommend a folder structure for the data. The solution must meet the following requirements:

- ☞ Data engineers from each region must be able to build their own pipelines for the data of their respective region only.
- ☞ The data must be processed at least once every 15 minutes for inclusion in Azure Synapse Analytics serverless SQL pools.

How should you recommend completing the structure? 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 |
|----------------------------|--|
| {deviceID} | / <input type="text" value="Value"/> / <input type="text" value="Value"/> / <input type="text" value="Value"/> .json |
| {mm}/(HH)/(DD)/(MM)/(YYYY) | |
| {regionID}/{deviceID} | |
| {regionID}/raw | |
| {YYYY}/{MM}/{DD}/{HH} | |
| {YYYY}/{MM}/{DD}/{HH}/{mm} | |
| raw/{deviceID} | |
| raw/{regionID} | |

Correct Answer:

| Values | Answer Area |
|----------------------------|---|
| {deviceID} | / raw/{regionID} / {YYYY}/{MM}/{DD}/{HH}/{mm} / {deviceID}.json |
| {mm}/{HH}/{DD}/{MM}/{YYYY} | |
| {regionID}/{deviceID} | |
| {regionID}/raw | |
| {YYYY}/{MM}/{DD}/{HH} | |
| {YYYY}/{MM}/{DD}/{HH}/{mm} | |
| raw/{deviceID} | |
| raw/{regionID} | |

Box 1: {raw/regionID}

Box 2: {YYYY}/{MM}/{DD}/{HH}/{mm}

Box 3: {deviceID}

Reference:

<https://github.com/paolosalvatori/StreamAnalyticsAzureDataLakeStore/blob/master/README.md>

Question #54

HOTSPOT -

You are implementing an Azure Stream Analytics solution to process event data from devices.

The devices output events when there is a fault and emit a repeat of the event every five seconds until the fault is resolved. The devices output a heartbeat event every five seconds after a previous event if there are no faults present.

A sample of the events is shown in the following table.

| DeviceID | EventType | EventTime |
|--------------------------------------|------------------------|-----------------------|
| 78cc5ht9-w357-684r-w4fr-kr16h6p9874e | HeartBeat | 2020-12-01T19:00.000Z |
| 78cc5ht9-w357-684r-w4fr-kr16h6p9874e | HeartBeat | 2020-12-01T19:05.000Z |
| 78cc5ht9-w357-684r-w4fr-kr16h6p9874e | TemperatureSensorFault | 2020-12-01T19:07.000Z |

You need to calculate the uptime between the faults.

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

```
SELECT  
DeviceID,  
MIN(EventTime) as StartTime,  
MAX(EventTime) as EndTime,  
DATEDIFF(second, MIN(EventTime), MAX(EventTime)) AS duration_in_seconds  
FROM input TIMESTAMP BY EventTime
```

| | |
|--|---|
| WHERE EventType='HeartBeat' | ▼ |
| WHERE LAG(EventType, 1) OVER (LIMIT DURATION(second,5)) <> EventType | ▼ |
| WHERE IsFirst(second,5) = 1 | ▼ |

GROUP BY

DeviceID

| | |
|---|---|
| ,SessionWindow(second, 5, 50000) OVER (PARTITION BY DeviceID) | ▼ |
| ,TumblingWindow(second,5) | ▼ |
| HAVING DATEDIFF(second, MIN(EventTime), MAX(EventTime)) > 5 | ▼ |

Correct Answer:

Answer Area

```
SELECT  
DeviceID,  
MIN(EventTime) as StartTime,  
MAX(EventTime) as EndTime,  
DATEDIFF(second, MIN(EventTime), MAX(EventTime)) AS duration_in_seconds  
FROM input TIMESTAMP BY EventTime
```

| | |
|--|---|
| WHERE EventType='HeartBeat' | ▼ |
| WHERE LAG(EventType, 1) OVER (LIMIT DURATION(second,5)) <> EventType | ▼ |
| WHERE IsFirst(second,5) = 1 | ▼ |

GROUP BY

DeviceID

| | |
|---|---|
| ,SessionWindow(second, 5, 50000) OVER (PARTITION BY DeviceID) | ▼ |
| ,TumblingWindow(second,5) | ▼ |
| HAVING DATEDIFF(second, MIN(EventTime), MAX(EventTime)) > 5 | ▼ |

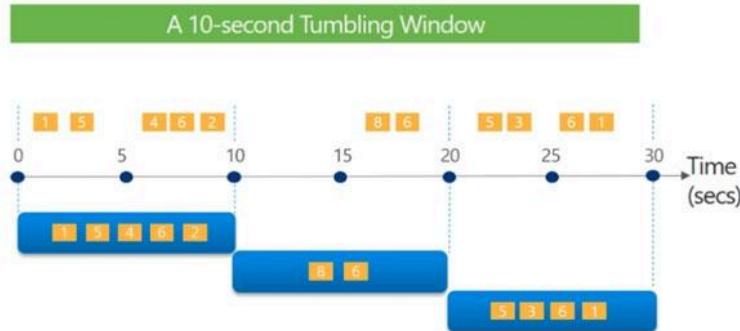
Box 1: WHERE EventType='HeartBeat'

Box 2: ,TumblingWindow(Second, 5)

Tumbling windows are a series of fixed-sized, non-overlapping and contiguous time intervals.

The following diagram illustrates a stream with a series of events and how they are mapped into 10-second tumbling windows.

Tell me the count of tweets per time zone every 10 seconds



```
SELECT TimeZone, COUNT(*) AS Count
FROM TwitterStream TIMESTAMP BY CreatedAt
GROUP BY TimeZone, TumblingWindow(second,10)
```

Incorrect Answers:

,SessionWindow.. : Session windows group events that arrive at similar times, filtering out periods of time where there is no data.

Reference:

<https://docs.microsoft.com/en-us/stream-analytics-query/session-window-azure-stream-analytics>

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

Question #55

You are creating a new notebook in Azure Databricks that will support R as the primary language but will also support Scala and SQL.

Which switch should you use to switch between languages?

- A. %<language>
- B. @<Language >
- C. \\[<language >]
- D. \\(<language >)

Correct Answer:

- A. %<language>

Question #56

You have an Azure Data Factory pipeline that performs an incremental load of source data to an Azure Data Lake Storage Gen2 account.

Data to be loaded is identified by a column named LastUpdatedDate in the source table.

You plan to execute the pipeline every four hours.

You need to ensure that the pipeline execution meets the following requirements:

☞ Automatically retries the execution when the pipeline run fails due to concurrency or throttling limits.

☞ Supports backfilling existing data in the table.

Which type of trigger should you use?

- A. event
- B. on-demand
- C. schedule
- D. tumbling window

Correct Answer:

D. tumbling window

Question #57

You are designing a solution that will copy Parquet files stored in an Azure Blob storage account to an Azure Data Lake Storage Gen2 account.

The data will be loaded daily to the data lake and will use a folder structure of {Year}/{Month}/{Day}/.

You need to design a daily Azure Data Factory data load to minimize the data transfer between the two accounts.

Which two configurations should you include in the design? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point

- A. Specify a file naming pattern for the destination.
- B. Delete the files in the destination before loading the data.
- C. Filter by the last modified date of the source files.
- D. Delete the source files after they are copied.

Correct Answer:

- A. Specify a file naming pattern for the destination.
- C. Filter by the last modified date of the source files.

Question #58

You plan to build a structured streaming solution in Azure Databricks. The solution will count new events in five-minute intervals and report only events that arrive during the interval. The output will be sent to a Delta Lake table.

Which output mode should you use?

- A. update
- B. complete
- C. append

Correct Answer:

C. append

Question #59

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 Synapse Analytics dedicated SQL pool that contains a table named Table1. You have files that are ingested and loaded into an Azure Data Lake Storage Gen2 container named container1.

You plan to insert data from the files in container1 into Table1 and transform the data. Each row of data in the files will produce one row in the serving layer of Table1.

You need to ensure that when the source data files are loaded to container1, the DateTime is stored as an additional column in Table1.

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

Does this meet the goal?

- A. Yes
- B. No

Correct Answer:

A. Yes

Question #60

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 Synapse Analytics dedicated SQL pool that contains a table named Table1. You have files that are ingested and loaded into an Azure Data Lake Storage Gen2 container named container1.

You plan to insert data from the files in container1 into Table1 and transform the data. Each row of data in the files will produce one row in the serving layer of Table1.

You need to ensure that when the source data files are loaded to container1, the DateTime is stored as an additional column in Table1.

Solution: You use a dedicated SQL pool to create an external table that has an additional DateTime column.

Does this meet the goal?

- A. Yes
- B. No

Correct Answer:

- B. No

Question #61

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 Synapse Analytics dedicated SQL pool that contains a table named Table1. You have files that are ingested and loaded into an Azure Data Lake Storage Gen2 container named container1.

You plan to insert data from the files in container1 into Table1 and transform the data. Each row of data in the files will produce one row in the serving layer of Table1.

You need to ensure that when the source data files are loaded to container1, the DateTime is stored as an additional column in Table1.

Solution: You use an Azure Synapse Analytics serverless SQL pool to create an external table that has an additional DateTime column.

Does this meet the goal?

- A. Yes
- B. No

Correct Answer:

- B. No

Question #62

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 Synapse Analytics dedicated SQL pool that contains a table named Table1. You have files that are ingested and loaded into an Azure Data Lake Storage Gen2 container named container1.

You plan to insert data from the files in container1 into Table1 and transform the data. Each row of data in the files will produce one row in the serving layer of Table1.

You need to ensure that when the source data files are loaded to container1, the DateTime is stored as an additional column in Table1.

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

Does this meet the goal?

- A. Yes
- B. No

Correct Answer:

- B. No

Question #63

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

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

You have an Azure Data Lake Storage account that contains a staging zone.

You need to design a daily process to ingest incremental data from the staging zone, transform the data by executing an R script, and then insert the transformed data into a data warehouse in Azure Synapse Analytics.

Solution: You use an Azure Data Factory schedule trigger to execute a pipeline that executes an Azure Databricks notebook, and then inserts the data into the data warehouse.

Does this meet the goal?

- A. Yes
- B. No

Correct Answer:

- A. Yes

Question #64

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

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

You have an Azure Data Lake Storage account that contains a staging zone.

You need to design a daily process to ingest incremental data from the staging zone, transform the data by executing an R script, and then insert the transformed data into a data warehouse in Azure Synapse Analytics.

Solution: You use an Azure Data Factory schedule trigger to execute a pipeline that executes mapping data flow, and then inserts the data into the data warehouse.

Does this meet the goal?

- A. Yes
- B. No

Correct Answer:

- B. No

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 Data Lake Storage account that contains a staging zone.

You need to design a daily process to ingest incremental data from the staging zone, transform the data by executing an R script, and then insert the transformed data into a data warehouse in Azure Synapse Analytics.

Solution: You schedule an Azure Databricks job that executes an R notebook, and then inserts the data into the data warehouse.

Does this meet the goal?

- A. Yes
- B. No

Correct Answer:

- A. Yes

Question #66

You plan to create an Azure Data Factory pipeline that will include a mapping data flow.

You have JSON data containing objects that have nested arrays.

You need to transform the JSON-formatted data into a tabular dataset. The dataset must have one row for each item in the arrays.

Which transformation method should you use in the mapping data flow?

- A. new branch
- B. unpivot
- C. alter row
- D. flatten

Correct Answer:

- D. flatten

Question #67

You use Azure Stream Analytics to receive Twitter data from Azure Event Hubs and to output the data to an Azure Blob storage account.

You need to output the count of tweets during the last five minutes every five minutes. Each tweet must only be counted once.

Which windowing function should you use?

- A. a five-minute Sliding window
- B. a five-minute Session window
- C. a five-minute Hopping window that has a one-minute hop
- D. a five-minute Tumbling window

Correct Answer:

- D. a five-minute Tumbling window

Question #68

You are planning a streaming data solution that will use Azure Databricks. The solution will stream sales transaction data from an online store. The solution has the following specifications: The output data will contain items purchased, quantity, line total sales amount, and line total tax amount.

- ☞ Line total sales amount and line total tax amount will be aggregated in Databricks.
- ☞ Sales transactions will never be updated. Instead, new rows will be added to adjust a sale.

You need to recommend an output mode for the dataset that will be processed by using Structured Streaming. The solution must minimize duplicate data.

What should you recommend?

- A. Update
- B. Complete
- C. Append

Correct Answer:

C. Append

Question #69

You have an enterprise data warehouse in Azure Synapse Analytics named DW1 on a server named Server1.

You need to determine the size of the transaction log file for each distribution of DW1.

What should you do?

- A. On DW1, execute a query against the sys.database_files dynamic management view.
- B. From Azure Monitor in the Azure portal, execute a query against the logs of DW1.
- C. Execute a query against the logs of DW1 by using the Get-AzOperationalInsightsSearchResult PowerShell cmdlet.
- D. On the master database, execute a query against the sys.dm_pdw_nodes_os_performance_counters dynamic management view.

Correct Answer:

- D. On the master database, execute a query against the sys.dm_pdw_nodes_os_performance_counters dynamic management view.

Question #70

You are designing an anomaly detection solution for streaming data from an Azure IoT hub. The solution must meet the following requirements:

- ☞ Send the output to Azure Synapse.
- ☞ Identify spikes and dips in time series data.
- ☞ Minimize development and configuration effort.

Which should you include in the solution?

- A. Azure Databricks
- B. Azure Stream Analytics
- C. Azure SQL Database

Correct Answer:

- B. Azure Stream Analytics

Question #71

A company uses Azure Stream Analytics to monitor devices.

The company plans to double the number of devices that are monitored.

You need to monitor a Stream Analytics job to ensure that there are enough processing resources to handle the additional load.

Which metric should you monitor?

- A. Early Input Events
- B. Late Input Events
- C. Watermark delay
- D. Input Deserialization Errors

Correct Answer:

- C. Watermark delay

Question #72

HOTSPOT -

You are designing an enterprise data warehouse in Azure Synapse Analytics that will store website traffic analytics in a star schema.

You plan to have a fact table for website visits. The table will be approximately 5 GB.

You need to recommend which distribution type and index type to use for the table. The solution must provide the fastest query performance.

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Distribution:

| |
|-------------|
| Hash |
| Round robin |
| Replicated |

Index:

| |
|-----------------------|
| Clustered columnstore |
| Clustered |
| Nonclustered |

Correct Answer:

Answer Area

Distribution:

| |
|-------------|
| Hash |
| Round robin |
| Replicated |

Index:

| |
|-----------------------|
| Clustered columnstore |
| Clustered |
| Nonclustered |

Box 1: Hash -

Consider using a hash-distributed table when:

The table size on disk is more than 2 GB.

The table has frequent insert, update, and delete operations.

Box 2: Clustered columnstore -

Clustered columnstore tables offer both the highest level of data compression and the best overall query performance.

Reference:

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

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

Question #73

You have an Azure Stream Analytics job.

You need to ensure that the job has enough streaming units provisioned.

You configure monitoring of the SU % Utilization metric.

Which two additional metrics should you monitor? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. Backlogged Input Events
- B. Watermark Delay
- C. Function Events
- D. Out of order Events
- E. Late Input Events

Correct Answer:

- A. Backlogged Input Events
- B. Watermark Delay

Question #74

You have an activity in an Azure Data Factory pipeline. The activity calls a stored procedure in a data warehouse in Azure Synapse Analytics and runs daily.

You need to verify the duration of the activity when it ran last.

What should you use?

- A. activity runs in Azure Monitor
- B. Activity log in Azure Synapse Analytics
- C. the sys.dm_pdw_wait_stats data management view in Azure Synapse Analytics
- D. an Azure Resource Manager template

Correct Answer:

- A. activity runs in Azure Monitor

Question #75

You have an Azure Data Factory pipeline that is triggered hourly.

The pipeline has had 100% success for the past seven days.

The pipeline execution fails, and two retries that occur 15 minutes apart also fail. The third failure returns the following error.

ErrorCode=UserErrorFileNotFound,'Type=Microsoft.DataTransfer.Common.Shared.HybridDeliveryException,Message=ADLS Gen2 operation failed for: Operation returned an invalid status code 'NotFound'. Account: 'contosoproduksouth'. Filesystem: wwi. Path: 'BIKES/CARBON/year=2021/month=01/day=10/hour=06'. ErrorCode: 'PathNotFound'.

Message: 'The specified path does not exist.'. RequestId:

'6d269b78-901f-001b-4924-e7a7bc000000'. TimeStamp: 'Sun, 10 Jan 2021 07:45:05'

What is a possible cause of the error?

- A. The parameter used to generate year=2021/month=01/day=10/hour=06 was incorrect.
- B. From 06:00 to 07:00 on January 10, 2021, there was no data in wwi/BIKES/CARBON.
- C. From 06:00 to 07:00 on January 10, 2021, the file format of data in wwi/BIKES/CARBON was incorrect.
- D. The pipeline was triggered too early.

Correct Answer:

B. From 06:00 to 07:00 on January 10, 2021, there was no data in wwi/BIKES/CARBON.

Question #76

You have an Azure Synapse Analytics job that uses Scala.

You need to view the status of the job.

What should you do?

- A. From Synapse Studio, select the workspace. From Monitor, select SQL requests.
- B. From Azure Monitor, run a Kusto query against the AzureDiagnostics table.
- C. From Synapse Studio, select the workspace. From Monitor, select Apache Sparks applications.
- D. From Azure Monitor, run a Kusto query against the SparkLoggingEvent_CL table.

Correct Answer:

C. From Synapse Studio, select the workspace. From Monitor, select Apache Sparks applications.

Question #77

DRAG DROP -

You have an Azure Data Lake Storage Gen2 account that contains a JSON file for customers.

The file contains two attributes named FirstName and LastName.

You need to copy the data from the JSON file to an Azure Synapse Analytics table by using Azure Databricks. A new column must be created that concatenates the FirstName and LastName values.

You create the following components:

- ☞ A destination table in Azure Synapse
- ☞ An Azure Blob storage container
- ☞ A service principal

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

Select and Place:

| Actions | Answer Area |
|--|-------------|
| Mount the Data Lake Storage onto DBFS. | |
| Write the results to a table in Azure Synapse. | |
| Specify a temporary folder to stage the data. | |
| Read the file into a data frame. | |
| Perform transformations on the data frame. | |

Correct Answer:

| Actions | Answer Area |
|---------|--|
| | Mount the Data Lake Storage onto DBFS. |
| | Read the file into a data frame. |
| | Perform transformations on the data frame. |
| | Specify a temporary folder to stage the data. |
| | Write the results to a table in Azure Synapse. |

Step 1: Mount the Data Lake Storage onto DBFS

Begin with creating a file system in the Azure Data Lake Storage Gen2 account.

Step 2: Read the file into a data frame.

You can load the json files as a data frame in Azure Databricks.

Step 3: Perform transformations on the data frame.

Step 4: Specify a temporary folder to stage the data

Specify a temporary folder to use while moving data between Azure Databricks and Azure Synapse.

Step 5: Write the results to a table in Azure Synapse.

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

Reference:

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

Question #78

You have an Azure data factory named ADF1.

You currently publish all pipeline authoring changes directly to ADF1.

You need to implement version control for the changes made to pipeline artifacts. The solution must ensure that you can apply version control to the resources currently defined in the UX Authoring canvas for ADF1.

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

NOTE: Each correct selection is worth one point.

- A. From the UX Authoring canvas, select Set up code repository.
- B. Create a Git repository.
- C. Create a GitHub action.
- D. Create an Azure Data Factory trigger.
- E. From the UX Authoring canvas, select Publish.
- F. From the UX Authoring canvas, run Publish All.

Correct Answer:

- A. From the UX Authoring canvas, select Set up code repository.
- B. Create a Git repository.

Question #79

DRAG DROP -

You have an Azure subscription that contains an Azure Synapse Analytics workspace named workspace1. Workspace1 connects to an Azure DevOps repository named repo1. Repo1 contains a collaboration branch named main and a development branch named branch1.

Branch1 contains an Azure Synapse pipeline named pipeline1.

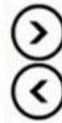
In workspace1, you complete testing of pipeline1.

You need to schedule pipeline1 to run daily at 6 AM.

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

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

Select and Place:

Actions**Answer Area** Create a new branch in Repo1. Merge the changes from branch1 into main. Associate the schedule trigger with pipeline1. Switch to Synapse live mode. Create a schedule trigger. Publish the contents of main.**Correct Answer:****Actions****Answer Area** Create a new branch in Repo1. Create a schedule trigger. Associate the schedule trigger with pipeline1. Merge the changes from branch1 into main. Switch to Synapse live mode. Publish the contents of main.**Question #80****HOTSPOT -**

You have an Azure subscription that contains an Azure Synapse Analytics dedicated SQL pool named Pool1 and an Azure Data Lake Storage account named storage1. Storage1 requires secure transfers.

You need to create an external data source in Pool1 that will be used to read .orc files in storage1.

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

```
CREATE EXTERNAL DATA SOURCE AzureDataLakeStore  
WITH  
( Location1 ` ://data@newyorktaxidataset.dfs.core.windows.net` ,  
    abfs  
    abfss  
    wasb  
    wasbs  
credential = ADLS_credential ,  
TYPE -  
) ;  
    BLOB_STORAGE  
    HADOOP  
    RDBMS  
    SHARP MAP MANAGER
```

Correct Answer:

Answer Area

```
CREATE EXTERNAL DATA SOURCE AzureDataLakeStore  
WITH  
( Location1 ` ://data@newyorktaxidataset.dfs.core.windows.net` ,  
    abfs  
    abfss  
    wasb  
    wasbs  
credential = ADLS_credential ,  
TYPE -  
) ;  
    BLOB_STORAGE  
    HADOOP  
    RDBMS  
    SHARP MAP MANAGER
```

Reference:

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

Question #81

You have an Azure subscription that contains an Azure Synapse Analytics dedicated SQL pool named SQLPool1.

SQLPool1 is currently paused.

You need to restore the current state of SQLPool1 to a new SQL pool.

What should you do first?

- A. Create a workspace.
- B. Create a user-defined restore point.
- C. Resume SQLPool1.
- D. Create a new SQL pool.

Correct Answer:

- C. Resume SQLPool1.

Question #82

You are designing an Azure Synapse Analytics workspace.

You need to recommend a solution to provide double encryption of all the data at rest.

Which two components should you include in the recommendation? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. an X.509 certificate
- B. an RSA key
- C. an Azure virtual network that has a network security group (NSG)
- D. an Azure Policy initiative
- E. an Azure key vault that has purge protection enabled

Correct Answer:

- B. an RSA key

Question #83

You have an Azure Synapse Analytics serverless SQL pool named Pool1 and an Azure Data Lake Storage Gen2 account named storage1. The AllowBlobPublicAccess property is disabled for storage1.

You need to create an external data source that can be used by Azure Active Directory (Azure AD) users to access storage from Pool1.

What should you create first?

- A. an external resource pool
- B. an external library
- C. database scoped credentials
- D. a remote service binding

Correct Answer:

C. database scoped credentials

Question #84

You have an Azure Data Factory pipeline named Pipeline1. Pipeline1 contains a copy activity that sends data to an Azure Data Lake Storage Gen2 account.

Pipeline1 is executed by a schedule trigger.

You change the copy activity sink to a new storage account and merge the changes into the collaboration branch.

After Pipeline1 executes, you discover that data is NOT copied to the new storage account.

You need to ensure that the data is copied to the new storage account.

What should you do?

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

Correct Answer:

A. Publish from the collaboration branch.

Question #85

You have an Azure Data Factory pipeline named pipeline1 that is invoked by a tumbling window trigger named Trigger1. Trigger1 has a recurrence of 60 minutes.

You need to ensure that pipeline1 will execute only if the previous execution completes successfully.

How should you configure the self-dependency for Trigger1?

- A. offset: "-00:01:00" size: "00:01:00"
- B. offset: "01:00:00" size: "-01:00:00"
- C. offset: "01:00:00" size: "01:00:00"
- D. offset: "-01:00:00" size: "01:00:00"

Correct Answer:

D. offset: "-01:00:00" size: "01:00:00"

Question #86

HOTSPOT -

You have an Azure Synapse Analytics pipeline named Pipeline1 that contains a data flow activity named Dataflow1.

Pipeline1 retrieves files from an Azure Data Lake Storage Gen 2 account named storage1.

Dataflow1 uses the AutoResolveIntegrationRuntime integration runtime configured with a core count of 128.

You need to optimize the number of cores used by Dataflow1 to accommodate the size of the files in storage1.

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

Hot Area:

Answer Area

To Pipeline1, add:

- A custom activity
- A Get Metadata activity
- An If Condition activity

For Dataflow1, set the core count by using:

- Dynamic content
- Parameters
- User properties

Correct Answer:

Answer Area

To Pipeline1, add:

- A custom activity
- A Get Metadata activity**
- An If Condition activity

For Dataflow1, set the core count by using:

- Dynamic content**
- Parameters
- User properties

Box 1: A Get Metadata activity -

Dynamically size data flow compute at runtime

The Core Count and Compute Type properties can be set dynamically to adjust to the size of your incoming source data at runtime. Use pipeline activities like

Lookup or Get Metadata in order to find the size of the source dataset data. Then, use Add Dynamic Content in the Data Flow activity properties.

Box 2: Dynamic content -

Reference:

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

Question #87

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 plan to create an Azure Databricks workspace that has a tiered structure. The workspace will contain the following three workloads:

- A workload for data engineers who will use Python and SQL.
 - A workload for jobs that will run notebooks that use Python, Scala, and SQL.
 - A workload that data scientists will use to perform ad hoc analysis in Scala and R.
- The enterprise architecture team at your company identifies the following standards for Databricks environments:
- The data engineers must share a cluster.
 - The job cluster will be managed by using a request process whereby data scientists and data engineers provide packaged notebooks for deployment to the cluster.
 - All the data scientists must be assigned their own cluster that terminates automatically after 120 minutes of inactivity. Currently, there are three data scientists.
- You need to create the Databricks clusters for the workloads.
- Solution: You create a Standard cluster for each data scientist, a High Concurrency cluster for the data engineers, and a Standard cluster for the jobs.
- Does this meet the goal?

- A. Yes
- B. No

Correct Answer:

- A. Yes

Question #88

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 plan to create an Azure Databricks workspace that has a tiered structure. The workspace will contain the following three workloads:

- A workload for data engineers who will use Python and SQL.
 - A workload for jobs that will run notebooks that use Python, Scala, and SQL.
 - A workload that data scientists will use to perform ad hoc analysis in Scala and R.
- The enterprise architecture team at your company identifies the following standards for Databricks environments:
- The data engineers must share a cluster.
 - The job cluster will be managed by using a request process whereby data scientists and data engineers provide packaged notebooks for deployment to the cluster.
 - All the data scientists must be assigned their own cluster that terminates automatically after 120 minutes of inactivity. Currently, there are three data scientists.
- You need to create the Databricks clusters for the workloads.

Solution: You create a Standard cluster for each data scientist, a High Concurrency cluster for the data engineers, and a High Concurrency cluster for the jobs.

Does this meet the goal?

- A. Yes
- B. No

Correct Answer:

- B. No

Question #89

You are designing a folder structure for the files in an Azure Data Lake Storage Gen2 account. The account has one container that contains three years of data.

You need to recommend a folder structure that meets the following requirements:

- ⇒ Supports partition elimination for queries by Azure Synapse Analytics serverless SQL pools
- ⇒ Supports fast data retrieval for data from the current month
- ⇒ Simplifies data security management by department

Which folder structure should you recommend?

- A. \Department\DataSource\YYYY\MM\DataFile_YYYYMMDD.parquet
- B. \DataSource\Department\YYYYMM\DataFile_YYYYMMDD.parquet
- C. \DD\MM\YYYY\Department\DataSource\DataFile_DDMMYY.parquet
- D. \YYYY\MM\DD\Department\DataSource\DataFile_YYYYMMDD.parquet

Correct Answer:

- A. \Department\DataSource\YYYY\MM\DataFile_YYYYMMDD.parquet

Question #90

You have an Azure subscription that contains an Azure Synapse Analytics dedicated SQL pool named Pool1. Pool1 receives new data once every 24 hours.

You have the following function.

```
create function dbo.udfFtoC(F decimal)
return decimal
as
begin
return (F - 32) * 5.0 / 9
end
```

You have the following query.

```
select avg_date, sensorid, avg_f, dbo.udfFtoC(avg_temperature) as avg_c from SensorTemps  
where avg_date = @parameter
```

The query is executed once every 15 minutes and the @parameter value is set to the current date.

You need to minimize the time it takes for the query to return results.

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

NOTE: Each correct selection is worth one point.

- A. Create an index on the avg_f column.
- B. Convert the avg_c column into a calculated column.
- C. Create an index on the sensorid column.
- D. Enable result set caching.
- E. Change the table distribution to replicate.

Correct Answer:

- B. Convert the avg_c column into a calculated column.
- D. Enable result set caching.

Question #91

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

What should you use?

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

Correct Answer:

- B. Structured Streaming in Azure Databricks

Question #92

You are creating an Apache Spark job in Azure Databricks that will ingest JSON-formatted data.

You need to convert a nested JSON string into a DataFrame that will contain multiple rows.

Which Spark SQL function should you use?

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

Correct Answer:

- A. explode

Question #93

DRAG DROP

-

You have an Azure subscription that contains an Azure Databricks workspace. The workspace contains a notebook named Notebook1.

In Notebook1, you create an Apache Spark DataFrame named df_sales that contains the following columns:

- Customer
- SalesPerson
- Region
- Amount

You need to identify the three top performing salespersons by amount for a region named HQ.

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 |
|------------------------------|-------------|
| agg(col('SalesPerson')) | |
| filter(col('SalesPerson')) | |
| groupBy(col('SalesPerson')) | |
| groupBy(col('TotalAmount')) | |
| orderBy(col('TotalAmount')) | |
| orderBy(desc('TotalAmount')) | |

Correct Answer:

Answer Area

```
df_sales.filter(col('Region')=='HQ').groupBy(col('SalesPerson'))
          .agg(sum('Amount').alias('TotalAmount')).orderBy(desc('TotalAmount')) limit(3)
```

Question #94

You need to schedule an Azure Data Factory pipeline to execute when a new file arrives in an Azure Data Lake Storage Gen2 container.

Which type of trigger should you use?

- A. on-demand
- B. tumbling window
- C. schedule
- D. storage event

Correct Answer:

D. storage event

Question #95

DRAG DROP

-

You have a project in Azure DevOps that contains a repository named Repo1. Repo1 contains a branch named main.

You create a new Azure Synapse workspace named Workspace1.

You need to create data processing pipelines in Workspace1. The solution must meet the following requirements:

- Pipeline artifacts must be stored in Repo1

- Source control must be provided for pipeline artifacts.
- All development must be performed in a feature branch.

Which four actions should you perform in sequence in Synapse Studio? 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 pipeline artifacts and save them in the main branch. | |
| Set the main branch as the collaboration branch. | |
| Create a pull request to merge the contents of the main branch into the new branch. | ▶ |
| Create pipeline artifacts and save them in the new branch. | ◀ |
| Create a new branch. | |
| Configure a code repository and select Repo1 . | |



Correct Answer:

| Answer Area |
|---|
| Configure a code repository and select Repo1 . |
| Create a new branch. |
| Create pipeline artifacts and save them in the new branch. |
| Create a pull request to merge the contents of the main branch into the new branch. |

Question #96

You have an Azure subscription that contains an Azure SQL database named DB1 and a storage account named storage1. The storage1 account contains a file named File1.txt. File1.txt contains the names of selected tables in DB1.

You need to use an Azure Synapse pipeline to copy data from the selected tables in DB1 to the files in storage1. The solution must meet the following requirements:

- The Copy activity in the pipeline must be parameterized to use the data in File1.txt to identify the source and destination of the copy.
- Copy activities must occur in parallel as often as possible.

Which two pipeline activities should you include in the pipeline? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

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

Correct Answer:

- B. Lookup
- C. ForEach

Question #97

You have an Azure data factory that connects to a Microsoft Purview account. The data factory is registered in Microsoft Purview.

You update a Data Factory pipeline.

You need to ensure that the updated lineage is available in Microsoft Purview.

What should you do first?

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

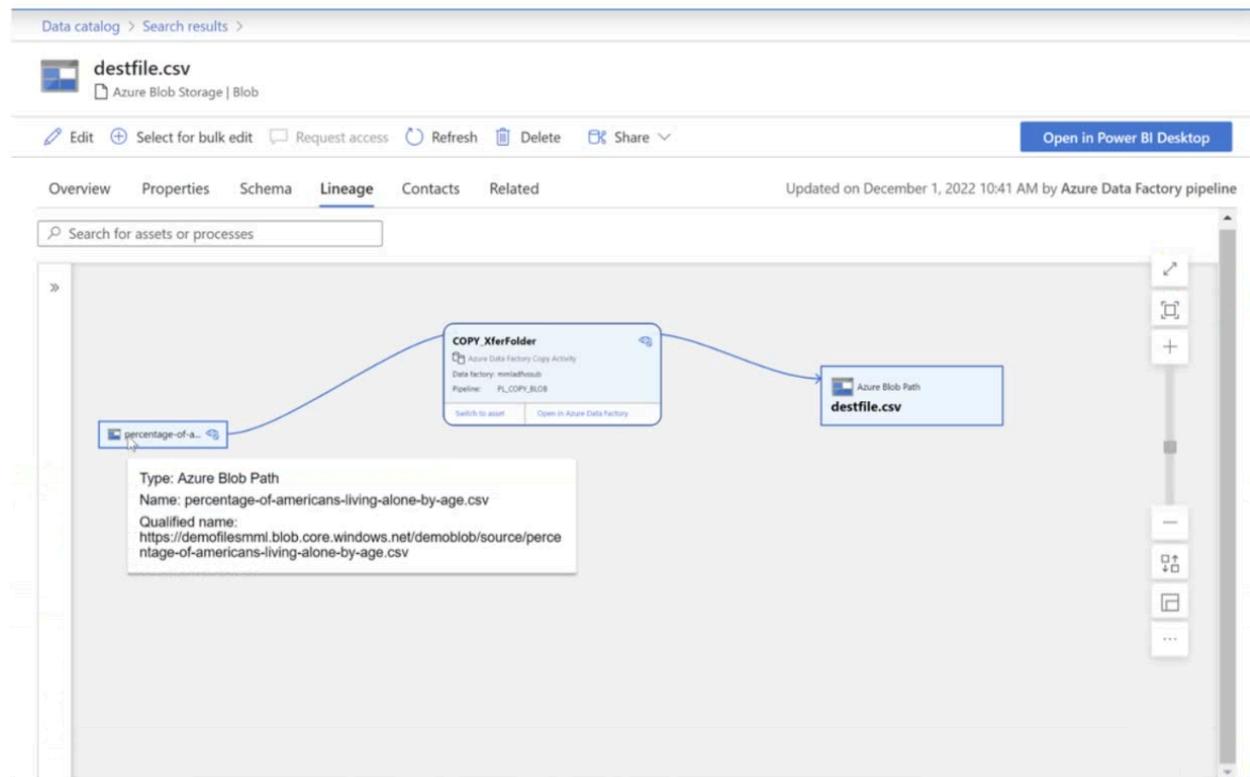
Correct Answer:

- B. Execute the pipeline.

Question #98

You have a Microsoft Purview account.

The Lineage view of a CSV file is shown in the following exhibit.



How is the data for the lineage populated?

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

Correct Answer:

- C. by executing a Data Factory pipeline

Question #99

You have an Azure subscription that contains a Microsoft Purview account named MP1, an Azure data factory named DF1, and a storage account named storage1. MP1 is configured to scan storage1. DF1 is connected to MP1 and contains a dataset named DS1. DS1 references a file in storage1.

In DF1, you plan to create a pipeline that will process data from DS1.

You need to review the schema and lineage information in MP1 for the data referenced by DS1.

Which two features can you use to locate the information? Each correct answer presents a complete solution.

NOTE: Each correct answer is worth one point.

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

Correct Answer:

- A. the search bar in the Microsoft Purview governance portal
- D. the search bar in Azure Data Factory Studio

Question #100

HOTSPOT

-

You have an Azure Blob storage account that contains a folder. The folder contains 120,000 files. Each file contains 62 columns.

Each day, 1,500 new files are added to the folder.

You plan to incrementally load five data columns from each new file into an Azure Synapse Analytics workspace.

You need to minimize how long it takes to perform the incremental loads.

What should you use to store the files and in which format? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

Storage:

- Multiple blob storage accounts
- Multiple containers in the blob storage account
- Timeslice partitioning in the folders

Format:

- Apache Parquet
- CSV
- JSON

Correct Answer:

Answer Area

Storage:

- Multiple blob storage accounts
- Multiple containers in the blob storage account
- Timeslice partitioning in the folders

Format:

- Apache Parquet
- CSV
- JSON

Question #101

DRAG DROP

You are batch loading a table in an Azure Synapse Analytics dedicated SQL pool.

You need to load data from a staging table to the target table. The solution must ensure that if an error occurs while loading the data to the target table, all the inserts in that batch are undone.

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

NOTE: Each correct selection is worth one point.

| Values | Answer Area |
|-------------------------------|-------------|
| BEGIN DISTRIBUTED TRANSACTION | |
| BEGIN TRAN | |
| COMMIT TRAN | |
| ROLLBACK TRAN | |
| SET RESULT_SET_CACHING ON | |
| ... | |
| ... | |
| ... | |
| ... | |
| ... | |

Answer Area

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

Correct Answer:**Answer Area**

```
BEGIN DISTRIBUTED TRANSACTION

BEGIN TRY

    INSERT INTO dbo.Table1 (col1, col2, col3)
    SELECT col1, col2, col3 FROM stage.Table1;

END TRY

BEGIN CATCH

    IF @@TRANCOUNT > 0
        BEGIN
            ROLLBACK TRAN ;
        END
    END CATCH;

    IF @@TRANCOUNT >0
        BEGIN
            COMMIT TRAN;
        END

```

Question #102

HOTSPOT

You have two Azure SQL databases named DB1 and DB2.

DB1 contains a table named Table1. Table1 contains a timestamp column named LastModifiedOn. LastModifiedOn contains the timestamp of the most recent update for each individual row.

DB2 contains a table named Watermark. Watermark contains a single timestamp column named WatermarkValue.

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

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

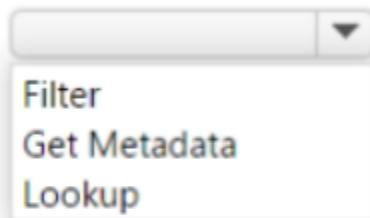
- Minimize the effort to author the pipeline.
- Ensure that the number of data integration units allocated to the upload operation can be controlled.

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

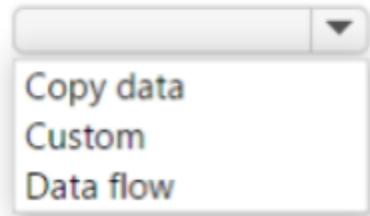
NOTE: Each correct answer is worth one point.

Answer Area

To retrieve the watermark value, use:



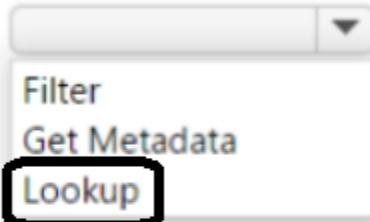
To perform the upload, use:



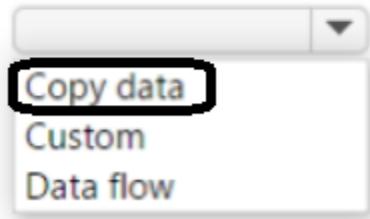
Correct Answer:

Answer Area

To retrieve the watermark value, use:



To perform the upload, use:



Question #103

HOTSPOT

You have an Azure Synapse serverless SQL pool.

You need to read JSON documents from a file by using the OPENROWSET function.

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

NOTE: Each correct selection is worth one point.

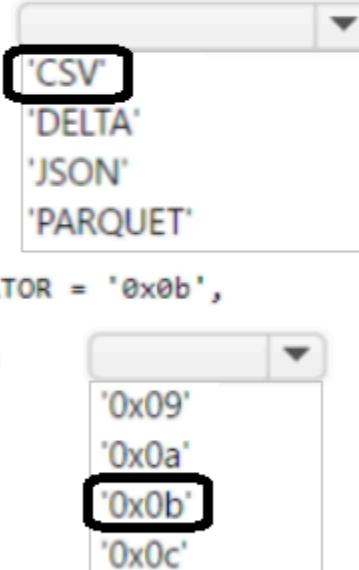
Answer Area

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

Correct Answer:

Answer Area

```
SELECT *
FROM OPENROWSET
(
    BULK
    'https://sourcedatalake.blob.core.windows.net/public/docs.json',
    FORMAT = 'CSV'
    FIELDTERMINATOR = '0x0b',
    FIELDQUOTE = '0x09'
    ROWTERMINATOR = '0x0b'
)
WITH (jsondoc nvarchar(max) AS JsonDocuments
```



Question #104

You use Azure Data Factory to create data pipelines.

You are evaluating whether to integrate Data Factory and GitHub for source and version control.

What are two advantages of the integration? Each correct answer presents a complete solution.

NOTE: Each correct selection is worth one point.

- A. additional triggers
- B. lower pipeline execution times
- C. the ability to save without publishing
- D. the ability to save pipelines that have validation issues

Correct Answer:

- C. the ability to save without publishing
- D. the ability to save pipelines that have validation issues

Question #105

DRAG DROP

-

You have an Azure Synapse Analytics workspace named Workspace1.

You perform the following changes:

- Implement source control for Workspace1.
- Create a branch named Feature based on the collaboration branch.
- Switch to the Feature branch.
- Modify Workspace1.

You need to publish the changes to Azure Synapse.

From which branch should you perform each change? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point

| Branches | Answer Area |
|-----------------|---|
| Collaboration | Create a pull request: <input type="text"/> |
| Publish | Publish the changes: <input type="text"/> |
| Feature | |

Correct Answer:

Answer Area

Create a pull request: Feature

Publish the changes: Collaboration

Question #106

You have two Azure Blob Storage accounts named account1 and account2.

You plan to create an Azure Data Factory pipeline that will use scheduled intervals to replicate newly created or modified blobs from account1 to account2.

You need to recommend a solution to implement the pipeline. The solution must meet the following requirements:

- Ensure that the pipeline only copies blobs that were created or modified since the most recent replication event.
- Minimize the effort to create the pipeline.

What should you recommend?

- A. Run the Copy Data tool and select Metadata-driven copy task.
- B. Create a pipeline that contains a Data Flow activity.
- C. Create a pipeline that contains a flowlet.
- D. Run the Copy Data tool and select Built-in copy task.

Correct Answer:

- D. Run the Copy Data tool and select Built-in copy task.

Question #107

You have an Azure Data Factory pipeline named pipeline1 that contains a data flow activity named activity1.

You need to run pipeline1.

Which runtime will be used to run activity1?

- A. Azure Integration runtime
- B. Self-hosted integration runtime
- C. SSIS integration runtime

Correct Answer:

- A. Azure Integration runtime

Question #108

HOTSPOT

You have an Azure subscription that contains an Azure Synapse Analytics workspace named workspace1. Workspace1 contains a dedicated SQL pool named SQLPool1 and an Apache Spark pool named sparkpool1. Sparkpool1 contains a DataFrame named pyspark_df.

You need to write the contents of pyspark_df to a table in SQLPool1 by using a PySpark notebook.

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

```
pyspark_df.createOrReplaceTempView("pysparkdftemptable")
```

```
%%local  
%%spark  
%%sql
```

val scala_df = spark.sqlContext.sql ("select * from pysparkdftemptable")
scala_df.write.
 jdbc ("sqlpool1.dbo.PySparkTable", Constants.INTERNAL)
 saveAsTable
 synapsesql

Correct Answer:

Answer Area

```
pyspark_df.createOrReplaceTempView("pysparkdftemptable")
```

```
%%local  
%%spark  
%%SQL
```

val scala_df = spark.sqlContext.sql ("select * from pysparkdftemptable")
scala_df.write.
 jdbc ("sqlpool1.dbo.PySparkTable", Constants.INTERNAL)
 saveAsTable
 synapsesql

Question #109

You have an Azure data factory named ADF1 and an Azure Synapse Analytics workspace that contains a pipeline named SynPipeline1. SynPipeline1 includes a Notebook activity.

You create a pipeline in ADF1 named ADPPipeline1.

You need to invoke SynPipeline1 from ADPPipeline1.

Which type of activity should you use?

- A. Web
- B. Spark
- C. Custom
- D. Notebook

Correct Answer:

- A. Web

Question #110

HOTSPOT

-

You have an Azure data factory that contains the linked service shown in the following exhibit.

Edit linked service

 Azure SQL Database [Learn more](#)

i To avoid publishing immediately to Data Factory, please use Azure Key Vault to retrieve secrets securely. Learn more [here](#)

Name *

AzureSqlDatabase1

Description

Connect via integration runtime * ⓘ

AutoResolveIntegrationRuntime

Connection string

Azure Key Vault

Account selection method ⓘ

From Azure subscription Enter manually

Fully qualified domain name *

ssio2022.database.windows.net

Database name *

Contoso

Authentication type *

SQL authentication

User name *

SQLAdmin

Password

Azure Key Vault

Password *

.....

Always encrypted ⓘ



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 answer is worth one point.

Answer Area

When working in a feature branch, changes to the linked service will be published to the live service

upon publishing the changes
upon saving the changes
when the changes are merged into the collaboration branch

A Copy activity that uses the linked service as the source will perform the Copy activity

in the region of the data factory
in the region of the selected external compute
in the region of the source database

Correct Answer:

Answer Area

When working in a feature branch, changes to the linked service will be published to the live service

upon publishing the changes
upon saving the changes
when the changes are merged into the collaboration branch

A Copy activity that uses the linked service as the source will perform the Copy activity

in the region of the data factory
in the region of the selected external compute
in the region of the source database

Question #111

HOTSPOT

-

In Azure Data Factory, you have a schedule trigger that is scheduled in Pacific Time.

Pacific Time observes daylight saving time.

The trigger has the following JSON file.

```
{  
    "name": "Trigger 1",  
    "properties": {  
        "annotations": [],  
        "runtimeState": "Started",  
        "pipelines": [],  
        "type": "ScheduleTrigger",  
        "typeProperties": {  
            "recurrence": {  
                "frequency": "Week",  
                "interval": 1,  
                "startTime": "2022-08-05T04:00:00",  
                "timeZone": "Pacific Standard Time",  
                "schedule": {  
                    "minutes": [  
                        0  
                    ],  
                    "hours": [  
                        3,  
                        21  
                    ],  
                    "weekDays": [  
                        "Sunday",  
                        "Saturday"  
                    ]  
                }  
            }  
        }  
    }  
}
```

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

NOTE: Each correct selection is worth one point.

Answer Area

The trigger will execute [answer choice] on Sunday, March 3, 2024.

| |
|------------|
| one time |
| two times |
| zero times |

The trigger [answer choice] daylight saving time.

| |
|--------------------------------|
| is unaffected by |
| will automatically adjust for |
| will require an adjustment for |

Correct Answer:

Answer Area

The trigger will execute [answer choice] on Sunday, March 3, 2024.

| |
|------------|
| one time |
| two times |
| zero times |

The trigger [answer choice] daylight saving time.

| |
|--------------------------------|
| is unaffected by |
| will automatically adjust for |
| will require an adjustment for |

Question #112

You have an Azure Synapse Analytics dedicated SQL pool.

You need to create a pipeline that will execute a stored procedure in the dedicated SQL pool and use the returned result set as the input for a downstream activity. The solution must minimize development effort.

Which type of activity should you use in the pipeline?

- A. U-SQL
- B. Stored Procedure
- C. Script
- D. Notebook

Correct Answer:

- C. Script

Question #113

You have an Azure SQL database named DB1 and an Azure Data Factory data pipeline named pipeline1.

From Data Factory, you configure a linked service to DB1.

In DB1, you create a stored procedure named SP1. SP1 returns a single row of data that has four columns.

You need to add an activity to pipeline1 to execute SP1. The solution must ensure that the values in the columns are stored as pipeline variables.

Which two types of activities can you use to execute SP1? Each correct answer presents a complete solution.

NOTE: Each correct selection is worth one point.

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

Correct Answer:

- A. Script
- C. Lookup

Question #114

You have an Azure data factory named ADF1.

You currently publish all pipeline authoring changes directly to ADF1.

You need to implement version control for the changes made to pipeline artifacts. The solution must ensure that you can apply version control to the resources currently defined in the Azure Data Factory Studio for ADF1.

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

NOTE: Each correct selection is worth one point.

- A. From the Azure Data Factory Studio, run Publish All.
- B. Create an Azure Data Factory trigger.
- C. Create a Git repository.
- D. Create a GitHub action.

- E. From the Azure Data Factory Studio, select Set up code repository.
- F. From the Azure Data Factory Studio, select Publish.

Correct Answer:

- C. Create a Git repository.
- E. From the Azure Data Factory Studio, select Set up code repository.

Question #115

You have an Azure data factory named ADF1 that contains a pipeline named Pipeline1.

Pipeline1 must execute every 30 minutes with a 15-minute offset.

You need to create a trigger for Pipeline1. The trigger must meet the following requirements:

- Backfill data from the beginning of the day to the current time.
- If Pipeline1 fails, ensure that the pipeline can re-execute within the same 30-minute period.
- Ensure that only one concurrent pipeline execution can occur.
- Minimize development and configuration effort.

Which type of trigger should you create?

- A. schedule
- B. event-based
- C. manual
- D. tumbling window

Correct Answer:

- D. tumbling window

Question #116

You have an Azure Data Lake Storage Gen2 account named account1 and an Azure event hub named Hub1. Data is written to account1 by using Event Hubs Capture.

You plan to query account by using an Apache Spark pool in Azure Synapse Analytics.

You need to create a notebook and ingest the data from account1. The solution must meet the following requirements:

- Retrieve multiple rows of records in their entirety.
- Minimize query execution time.
- Minimize data processing.

Which data format should you use?

- A. Parquet - O. Avro
- C. ORC
- D. JSON

Correct Answer:

- A. Parquet - O. Avro

Question #117

You have an Azure Blob Storage account named blob1 and an Azure Data Factory pipeline named pipeline1.

You need to ensure that pipeline1 runs when a file is deleted from a container in blob1. The solution must minimize development effort.

Which type of trigger should you use?

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

Correct Answer:

- B. storage event

Question #118

HOTSPOT

-

You have Azure Data Factory configured with Azure Repos Git integration. The collaboration branch and the publish branch are set to the default values.

You have a pipeline named pipeline1.

You build a new version of pipeline1 in a branch named feature1.

From the Data Factory Studio, you select Publish.

The source code of which branch will be built, and which branch will contain the output of the Azure Resource Manager (ARM) template? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

Source code:

```
adf_publish  
feature1  
main
```

ARM template output:

```
adf_publish  
feature1  
main
```

Correct Answer:

Answer Area

Source code:

```
adf_publish  
feature1  
main
```

ARM template output:

```
adf publish  
feature1  
main
```

Question #119

DRAG DROP

You have an Azure subscription that contains an Azure data factory.

You are editing an Azure Data Factory activity JSON.

The script needs to copy a file from Azure Blob Storage to multiple destinations. The solution must ensure that the source and destination files have consistent folder paths.

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

NOTE: Each correct selection is worth one point.

| Values | Answer Area |
|-------------------|---|
| FlattenHierarchy | { |
| ForEach | "name": "Pipeline1", "properties": { |
| MergeFiles | "activities": [|
| PreserveHierarchy | { |
| Switch | "name": "Activity1", "type": [REDACTED], |
| Until | "typeProperties": { "isSequential": "true", "items": { "value": "@pipeline ().parameters.mySinkDatasetFolderPath", "type": "Expression"}, "activities" [|
| | { "name": "MyCopyActivity", "type": "Copy", "typeProperties": { "source": { "type": "BlobSource", "recursive": "false" }, "sink": { "type": "BlobSink", "CopyBehavior": [REDACTED] ... } |

Correct Answer:**Answer Area**

```
{  
    "name": "Pipeline1",  
    "properties": {  
        "activities": [  
            {  
                "name": "Activity1",  
                "type": "ForEach",  
                "typeProperties": {  
                    "isSequential": "true",  
                    "items": {  
                        "value": "@pipeline  
().parameters.mySinkDatasetFolderPath",  
                        "type": "Expression",  
                    }  
                },  
                "activities": [  
                    {  
                        "name": "MyCopyActivity",  
                        "type": "Copy",  
                        "typeProperties": {  
                            "source": {  
                                "type": "BlobSource",  
                                "recursive": "false" },  
                            "sink": {  
                                "type": "BlobSink",  
                                "CopyBehavior": "PreserveHierarchy"  
                            }  
                        }  
                    }  
                ]  
            }  
        ]  
    }  
}
```

Question #120

You are building a data flow in Azure Data Factory that upserts data into a table in an Azure Synapse Analytics dedicated SQL pool.

You need to add a transformation to the data flow. The transformation must specify logic indicating when a row from the input data must be upserted into the sink.

Which type of transformation should you add to the data flow?

- A. join
- B. alter row
- C. surrogate key
- D. select

Correct Answer:

- B. alter row

Question #121

You have an on-premises database named db1 and a self-hosted integration runtime.

You have an Azure subscription that contains an Azure Data Lake Storage account named dl1.

You need to develop four data pipeline projects that will use Microsoft Power Query to copy data from db1 to dl1. The solution must meet the following requirements:

- All pipelines must use the self-hosted integration runtime.
- Each project must be stored in a separate Git repository.
- Development effort must be minimized.

What should you use?

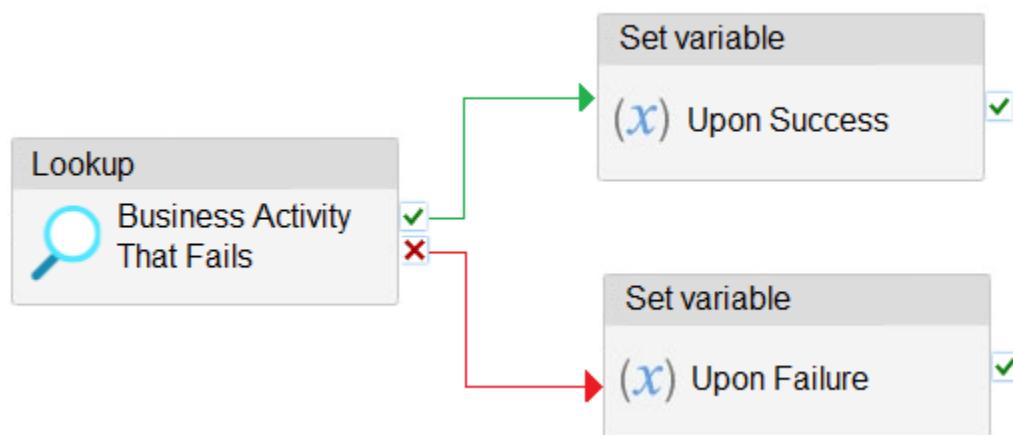
- A. Azure Synapse Analytics
- B. Azure Logic Apps.
- C. Azure Data Factory
- D. Microsoft Power BI

Correct Answer:

C. Azure Data Factory

Question #122

You have the Azure Synapse Analytics pipeline shown in the following exhibit.



You need to add a set variable activity to the pipeline to ensure that after the pipeline's completion, the status of the pipeline is always successful.

What should you configure for the set variable activity?

- A. a skipped dependency on the Upon Failure activity

- B. a skipped dependency on the Upon Success activity
- C. a success dependency on the Business Activity That Fails activity
- D. a failure dependency on the Upon Failure activity

Correct Answer:

- B. a skipped dependency on the Upon Success activity

Question #123

You have an on-premises Linux server that contains a database named DB1.

You have an Azure subscription that contains an Azure data factory named ADF1 and an Azure Data Lake Storage account named ADLS1.

You need to create a pipeline in ADF1 that will copy data from DB1 to ADLS1.

Which type of integration runtime should you use to read the data from DB1?

- A. self-hosted integration runtime
- B. Azure integration runtime
- C. Azure-SQL Server Integration Services (SSIS)

Correct Answer:

- A. self-hosted integration runtime

Question #124

DRAG DROP

-

You have an Azure Data Lake Storage account named account1.

You use an Azure Synapse Analytics serverless SQL pool to access sales data stored in account1.

You need to create a bar chart that displays sales by product. The solution must minimize development effort.

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

Actions

Add a `SELECT` statement that will return the sales by product data.

Switch to the Chart view.

Modify the Chart settings.

Create a SQL script by using Synapse Studio.

Execute the script.

Answer Area



Correct Answer:

Answer Area

Create a SQL script by using Synapse Studio.

Add a `SELECT` statement that will return the sales by product data.

Execute the script.

Switch to the Chart view.

Modify the Chart settings.

Question #125

DRAG DROP

You have an Azure Synapse Analytics dedicated SQL pool.

You need to create a copy of the data warehouse and make the copy available for 28 days. The solution must minimize costs.

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.

Actions

Answer Area

Create a new user-defined restore point.

Restore the latest automatic restore point to a new data warehouse.

Pause the restored data warehouse.

Restore the copy from the latest automatic restore point to the current data warehouse.

Restore the copy from the new user-defined restore point to a new data warehouse.



Correct Answer:

Answer Area

Create a new user-defined restore point.

Restore the copy from the new user-defined restore point to a new data warehouse.

Pause the restored data warehouse.

Question #126

HOTSPOT

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

You need to read data from a CSV file and write the data to a Delta table by using Pool1.

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

NOTE: Each correct selection is worth one point.

Answer Area

```
from delta.tables import *
from pyspark.sql.functions import *
df = spark.read.load
('abfss://container@mydatalake.dfs.core.windows.net/stage/
products.csv', format = 'csv', header = True)
delta_table_path = "/delta/products-delta"
df. ▾ .save(delta_table_path)
cache()
inputFiles()
write.format("delta")
write.parquet
deltaTable = ▾ (spark, delta_table_path)
deltaTable.alias
deltaTable.convertToDelta
deltaTable.forPath
deltaTable.update
```

Correct Answer:

Answer Area

```
from delta.tables import *
from pyspark.sql.functions import *
df = spark.read.load
('abfss://container@mydatalake.dfs.core.windows.net/stage/
products.csv', format = 'csv', header = True)
delta_table_path = "/delta/products-delta"
df. ▾ .save(delta_table_path)

cache()
inputFiles()
write.format("delta")
write.parquet

deltaTable = ▾ (spark, delta_table_path)
deltaTable.alias
deltaTable.convertToDelta
deltaTable.forPath
deltaTable.update
```

Question #127

HOTSPOT

You have an Azure Data Lake Storage account that contains one CSV file per hour for January 1, 2020, through January 31, 2023. The files are partitioned by using the following folder structure.

`csv/system1/{year}/{month}/{filename}.csv`

You need to query the files by using an Azure Synapse Analytics serverless SQL pool. The solution must return the row count of each file created during the last three months of 2022.

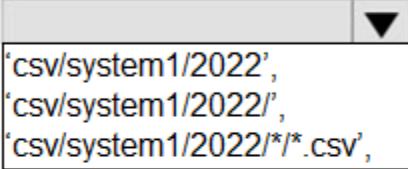
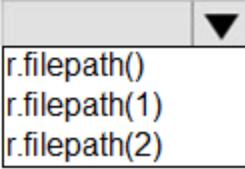
How should you complete the query? To answer, select the appropriate options in the answer area.

Answer Area

```
SELECT
    r.filepath() AS filepath
    ,COUNT_BIG(*) AS [rows]
FROM OPENROWSET(
    BULK
        'csv/system1/2022',
        'csv/system1/2022'
        'csv/system1/2022/*/*.CSV',
    DATA_SOURCE = 'MyDataLake',
    FORMAT = 'CSV',
    PARSER_VERSION = '2.0',
    FIRSTROW = 2 )
WITH (vendor_id INT) AS [r]
WHERE
    r.filepath()
    IN ('10', '11', '12')
    r.filepath(1)
    r.filepath(2)
GROUP BY
```

Correct Answer:

Answer Area

```
SELECT
    r.filepath() AS filepath
    ,COUNT_BIG(*) AS [rows]
FROM OPENROWSET(
    BULK
        
        'csv/system1/2022',
        'csv/system1/2022/',
        'csv/system1/2022/*/*.csv',
    DATA_SOURCE = 'MyDataLake',
    FORMAT = 'CSV',
    PARSER_VERSION = '2.0',
    FIRSTROW = 2 )
WITH (vendor_id INT) AS [r]
WHERE
    
    r.filepath()
    r.filepath(1)
    r.filepath(2)
    IN ('10', '11', '12')
GROUP BY
```

Question #128

HOTSPOT

You have an Azure Synapse Analytics dedicated SQL pool named Pool1 that contains an external table named Sales. Sales contains sales data. Each row in Sales contain data on a single sale, including the name of the salesperson.

You need to implement row-level security (RLS). The solution must ensure that the salespeople can access only their respective sales.

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

NOTE: Each correct selection is worth one point.

Answer Area

Create:

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

Restrict row access by using:

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

Correct Answer:

Answer Area

Create:

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

Restrict row access by using:

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

Question #129

You have an Azure Data Factory pipeline named P1.

You need to schedule P1 to run at 10:15 AM, 12:15 PM, 2:15 PM, and 4:15 PM every day.

Which frequency and interval should you configure for the scheduled trigger?

- A. Frequency: Month - Interval: 1
- B. Frequency: Day - Interval: 1
- C. Frequency: Minute - Interval: 60
- D. Frequency: Hour - Interval: 2

Correct Answer:

B. Frequency: Day - Interval: 1

Question #130

You are creating an Azure Data Factory pipeline.

You need to add an activity to the pipeline. The activity must execute a Transact-SQL stored procedure that has the following characteristics:

- Returns the number of sales invoices for a current date
- Does NOT require input parameters

Which type on activity should you use?

- A. Stored Procedure
- B. Get Metadata
- C. Append Variable
- D. Lookup

Correct Answer:

D. Lookup

Question #131

HOTSPOT

-

You have an Azure Synapse Analytics workspace that contains three pipelines and three triggers named Trigger1, Trigger2, and Trigger3.

Trigger3 has the following definition.

```

...
{
  "name": "Trigger3",
  "properties": {
    "annotations": [],
    "runtimeState": "Stopped",
    "pipeline": {
      "pipelineReference": {
        "referenceName": "Pipeline 3",
        "type": "PipelineReference"
      }
    },
    "type": "TumblingWindowTrigger",
    "typeProperties": {
      "frequency": "Hour",
      "interval": 1,
      "startTime": "2023-11-12T11:00:00Z",
      "delay": "00:00:00",
      "maxConcurrency": 1,
      "retryPolicy": {
        "intervalInSeconds": 30
      },
      "dependsOn": [
        {
          "type": "TumblingWindowTriggerDependencyReference",
          "size": "0.03:00:00",
          "offset": "-0.02:00:00",
          "referenceTrigger": {
            "referenceName": "Trigger1",
            "type": "TriggerReference"
          }
        },
        {
          "type": "TumblingWindowTriggerDependencyReference",
          "size": "0.03:00:00",
          "offset": "-0.02:00:00",
          "referenceTrigger": {
            "referenceName": "Trigger2",
            "type": "TriggerReference"
          }
        },
        {
          "type": "SelfDependencyTumblingWindowTriggerReference",
          "offset": "-0.03:00:00"
        }
      ]
    }
  }
}
...

```

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 |
|--|-----------------------|-----------------------|
| Pipeline3 will execute when Trigger3 fires. | <input type="radio"/> | <input type="radio"/> |
| Up to three instances of Trigger3 can fire simultaneously. | <input type="radio"/> | <input type="radio"/> |
| Trigger3 will fire three hours after Trigger1 has fired three times, and Trigger2 has fired three times. | <input type="radio"/> | <input type="radio"/> |

Correct Answer:

Answer Area

| Statements | Yes | No |
|--|-------------------------------------|-------------------------------------|
| Pipeline3 will execute when Trigger3 fires. | <input checked="" type="checkbox"/> | <input type="radio"/> |
| Up to three instances of Trigger3 can fire simultaneously. | <input type="radio"/> | <input checked="" type="checkbox"/> |
| Trigger3 will fire three hours after Trigger1 has fired three times, and Trigger2 has fired three times. | <input type="radio"/> | <input checked="" type="checkbox"/> |

Question #132

DRAG DROP

You have an Azure Databricks deployment and a local file named /tmp/file1 that contains the following code.

```
[  
  {string:"string1","int":1,"dict": {"key": "'value1"},  
  {  
    "string": "string2",  
    "int": 2,  
    "dict": {  
      "key": "value2",  
      "extra_key": "extra_value2"  
    }  
  }  
]
```

You need to read /tmp/file1 into a data frame by using Scala.

How should you complete the 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.

| Values | Answer Area |
|-----------------|---|
| Source | val df = spark.read.option(" |
| inferSchema | <input type="text"/> ", "true"). <input type="text"/> |
| ignoreExtension | <input type="text"/> ("/tmp/file1) |
| json | |
| Multiline | |
| schema | |
| text | |

Correct Answer:

Answer Area

```
val df = spark.read.option("  ", "true"). json  ("/tmp/file1)
```

Question #133

You have an Azure subscription that contains a Microsoft Purview account.

You need to search the Microsoft Purview Data Catalog to identify assets that have an assetType property of Table or View.

Which query should you run?

- A. assetType IN ('Table', 'View')
- B. assetType:Table OR assetType:view
- C. assetType = (Table OR View)
- D. assetType:(Table OR View)

Correct Answer:

- B. assetType:Table OR assetType:view

Question #134

You have an Azure subscription that contains an Azure Synapse Analytics account. The account is integrated with an Azure Repos repository named Repo1 and contains a pipeline named Pipeline1. Repo1 contains the branches shown in the following table.

| Name | Description |
|-------------------|----------------------|
| featuredev | Working branch |
| main | Collaboration branch |
| pipeline1_publish | Publish branch |

From featuredev, you develop and test changes to Pipeline1.

You need to publish the changes.

What should you do first?

- A. From featuredev, create a pull request.
- B. From main, create a pull request.
- C. Add a Publish_config.json file to the root folder of the collaboration branch.
- D. Switch to live mode.

Correct Answer:

- A. From featuredev, create a pull request.

Question #135

HOTSPOT

-

You have an Azure data factory that has the Git repository settings shown in the following exhibit.

Git repository

Git repository information associated with your data factory. [CI/CD best practices](#)

Edit Overwrite live mode Disconnect Import resources

Repository type Azure DevOps Git

Azure DevOps Account

Project name ADFDeployDemo

Repository name ADFDeployDemo

Collaboration branch main

Publish branch adf_publish

Root folder /

Last published commit 23b144ac4aa7daf16f2fe7c2ab0eb303a8e4ed65

Publish (from ADF Studio) Enabled

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 answer is worth one point.

Answer Area

Changes to pipelines will be saved in Azure DevOps [answer choice].

| |
|--------------------------------|
| every 20 seconds |
| when the pipeline is published |
| when the pipeline is saved |

| |
|--------------------|
| adf_publish branch |
| main branch |
| root folder |

To publish changes by using Azure Data Factory Studio, the changes must first be saved in the [answer choice].

Correct Answer:

Answer Area

Changes to pipelines will be saved in Azure DevOps [answer choice].

every 20 seconds
when the pipeline is published
when the pipeline is saved

To publish changes by using Azure Data Factory Studio, the changes must first be saved in the [answer choice].

adf_publish branch
main branch
root folder

Question #136

HOTSPOT

You have an Azure subscription that contains an Azure Synapse Analytics dedicated SQL pool named Pool1 and a storage account. The storage account contains a blob container. The blob container contains multiple CSV files.

You plan to load the files into Pool1 by using the following code.

```
COPY INTO [staging].[Weather]
FROM <PATH TO CSV FOLDER>
WITH
(
    FILE_TYPE = 'CSV',
    FIELDTERMINATOR = ',',
    ROWTERMINATOR='0X0A',
    COMPRESSION = 'GZIP'
)
OPTION (LABEL = 'COPY : Staging dataset');
```

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 |
|--|-----------------------|-----------------------|
| Identity values will be imported into a table. | <input type="radio"/> | <input type="radio"/> |
| The code expects each line in a CSV file to end with a new line character. | <input type="radio"/> | <input type="radio"/> |
| 'COPY : Staging dataset' will be inserted into each row of a column named LABEL. | <input type="radio"/> | <input type="radio"/> |

Correct Answer:**Answer Area**

| Statements | Yes | No |
|--|----------------------------------|----------------------------------|
| Identity values will be imported into a table. | <input type="radio"/> | <input checked="" type="radio"/> |
| The code expects each line in a CSV file to end with a new line character. | <input checked="" type="radio"/> | <input type="radio"/> |
| 'COPY : Staging dataset' will be inserted into each row of a column named LABEL. | <input checked="" type="radio"/> | <input type="radio"/> |

Question #137

DRAG DROP

You have an Azure subscription that contains an Azure Data Factory account named ADF1 and an Azure Data Lake Storage Gen2 account named storage1. ADF1 contains the objects shown in the following table.

| Name | Description |
|----------------|--|
| DailyIngestion | Data factory pipeline |
| SourceFiles | Parameterized dataset that references the source files from each day |
| TargetFile | Parameterized dataset that references the target file from each day |

You need to configure DailyIngestion to perform the following actions:

- Ingest 2,000 small files into storage1 once every 24 hours.
- Output one large file once every 24 hours.

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.

Answer Area

Add a Data flow activity.

Configure the datasets and parameters for SourceFiles and TargetFile.

Add a Copy data activity.

Correct Answer:**Actions**

Add a Data flow activity.

Add a Copy data activity.

Configure the datasets and parameters for SourceFiles and TargetFile.

Configure the Copy behavior setting.

Enable Auto mapping.

Answer Area**Question #138**

HOTSPOT

-

You have an Azure Synapse Analytics pipeline named pipeline1 that has concurrency set to 1.

To run pipeline1, you create a new trigger as shown in the following exhibit.

Type *

Start date * ⓘ

Time zone ⓘ

Recurrence * ⓘ

Advanced recurrence options

Execute at these times ⓘ

| | | | |
|---------|--|--|--|
| Hours | 10 <input checked="" type="checkbox"/> | 12 <input checked="" type="checkbox"/> | 17 <input checked="" type="checkbox"/> |
| Minutes | 30 <input checked="" type="checkbox"/> | 45 <input checked="" type="checkbox"/> | |

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.

Answer Area

The trigger will run at **[answer choice]**.

| |
|----------|
| 12:30 AM |
| 12:30 PM |
| 7:30 PM |

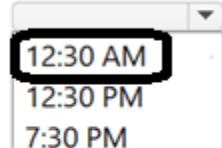
If the previous execution of pipeline1 is still running when the trigger fires next,
the new triggered execution will **[answer choice]**.

| |
|------------|
| be queued |
| be skipped |
| start |

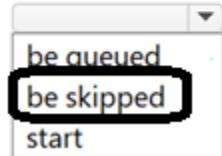
Correct Answer:

Answer Area

The trigger will run at [answer choice].



If the previous execution of pipeline1 is still running when the trigger fires next,
the new triggered execution will [answer choice].



Question #139

HOTSPOT

You have a trigger in Azure Data Factory configured as shown in the following exhibit.

Name *

MyScheduleTrigger

Description

Type *

ScheduleTrigger

Start date * ⓘ

5/11/2023, 8:15:00 PM

Time zone * ⓘ

Brussels, Copenhagen, Madrid, Paris (UTC+1)

ⓘ This time zone observes daylight savings. Trigger will auto-adjust for one hour difference.

Recurrence * ⓘ

Every Minute(s)

Specify an end date

End On * ⓘ

5/12/2025, 5:30:00 PM

Annotations

+ New

Status ⓘ

Started Stopped

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

Answer Area

If the trigger was published on May 12, 2023, at 9:00 AM,
the first execution will occur on [answer choice].

- May 12, 2023, at 9:00 AM
- May 12, 2023, at 9:15 AM
- May 12, 2023, at 8:15 PM

The last expected execution time of the pipeline
will occur on [answer choice].

- May 12, 2025, at 8:15 PM
- May 12, 2025, at 5:30 PM
- May 11, 2025, at 5:15 PM

Correct Answer:

Answer Area

If the trigger was published on May 12, 2023, at 9:00 AM,
the first execution will occur on [answer choice].

- May 12, 2023, at 9:00 AM
- May 12, 2023, at 9:15 AM
- May 12, 2023, at 8:15 PM

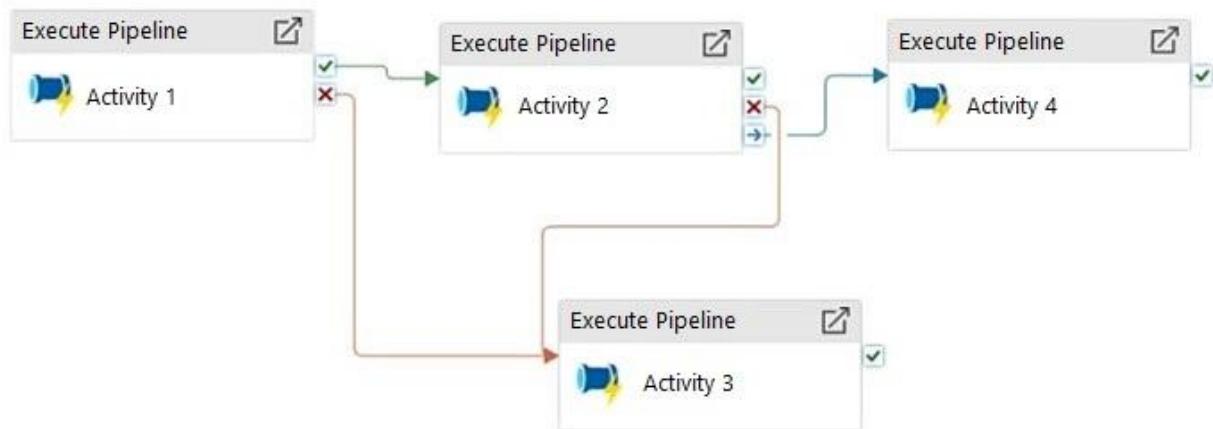
The last expected execution time of the pipeline
will occur on [answer choice].

- May 12, 2025, at 8:15 PM
- May 12, 2025, at 5:30 PM
- May 11, 2025, at 5:15 PM

Question #140

HOTSPOT

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



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 |
|---|-----------------------|-----------------------|
| If Activity 1 fails, Activity 3 will execute. | <input type="radio"/> | <input type="radio"/> |
| If Activity 2 fails, Activity 3 will execute. | <input type="radio"/> | <input type="radio"/> |
| If Activity 2 fails, and Activity 4 succeeds, the pipeline will return a status of Success. | <input type="radio"/> | <input type="radio"/> |

Correct Answer:

Answer Area

| Statements | Yes | No |
|---|-------------------------------------|-------------------------------------|
| If Activity 1 fails, Activity 3 will execute. | <input checked="" type="checkbox"/> | <input type="radio"/> |
| If Activity 2 fails, Activity 3 will execute. | <input checked="" type="checkbox"/> | <input type="radio"/> |
| If Activity 2 fails, and Activity 4 succeeds, the pipeline will return a status of Success. | <input type="radio"/> | <input checked="" type="checkbox"/> |

Question #141

You have an Azure subscription that contains an Azure Synapse Analytics account and a Microsoft Purview account.

You create a pipeline named Pipeline1 for data ingestion to a dedicated SQL pool.

You need to generate data lineage from Pipeline1 to Microsoft Purview.

Which two activities generate data lineage? Each correct answer presents a complete solution.

NOTE: Each correct selection is worth one point.

- A. Web
- B. Copy
- C. WebHook
- D. Data flow
- E. Validation

Correct Answer:

- B. Copy
- D. Data flow

Question #142

DRAG DROP

You have an Azure subscription that contains an Azure Synapse Analytics workspace named Workspace1. Workspace1 contains an Apache Spark pool named spark1 and a pipeline named Pipeline1.

You need to add an activity to Pipeline1 that will run a notebook. The solution must ensure that the activity overrides the value of a variable named inputFile when the notebook runs.

Which five actions should you perform in sequence in Synapse Studio? 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 |
|--|-------------|
| Configure the Base parameters settings. | |
| Add a new notebook that declares and initializes inputFile. | |
| Select Toggle parameter cell. | |
| Publish the notebook. | |
| Drag an activity from the Synapse section to Pipeline1 and link the activity to spark1. | |
| Drag an activity from the Databricks section to Pipeline1 and link the activity to spark1. | |

Correct Answer:**Answer Area**

Drag an activity from the Synapse section to Pipeline1 and link the activity to spark1.

Add a new notebook that declares and initializes inputFile.

Select **Toggle parameter** cell.

Configure the Base parameters settings.

Publish the notebook.

Question #143

You have an Azure subscription that contains an Azure SQL database named SQLDB1 and an Azure Synapse Analytics dedicated SQL pool named Pool1.

You need to replicate data from SQLDB1 to Pool1. The solution must meet the following requirements:

- Minimize performance impact on SQLDB1.
- Support near-real-time (NRT) analytics.
- Minimize administrative effort.

What should you use?

- A. Azure Synapse Link
- B. Azure Private Link
- C. SQL Data Sync for Azure
- D. transactional replication in Microsoft SQL Server

Correct Answer:

A. Azure Synapse Link

Question #144

You have an Azure subscription that contains an Azure SQL database named SQL1 and an Azure data factory named ADF1.

You need to create a pipeline in ADF1 that will perform the following actions:

- Execute a stored procedure to retrieve three configuration values from SQLI.
- For each value, execute the Set Variable activity.

Which activity should you use to retrieve the configuration values?

- A. Get Metadata
- B. Copy
- C. Stored procedure
- D. Lookup

Correct Answer:

- D. Lookup

Topic 3

Question #1

DRAG DROP -

You have an Azure Active Directory (Azure AD) tenant that contains a security group named Group1. You have an Azure Synapse Analytics dedicated SQL pool named dw1 that contains a schema named schema1.

You need to grant Group1 read-only permissions to all the tables and views in schema1. The solution must use the principle of least privilege.

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 a database role named Role1 and grant Role1 SELECT permissions to schema1. | |
| Create a database role named Role1 and grant Role1 SELECT permissions to dw1. | |
| Assign the Azure role-based access control (Azure RBAC) Reader role for dw1 to Group1. | |
| Create a database user in dw1 that represents Group1 and uses the FROM EXTERNAL PROVIDER clause. | |
| Assign Role1 to the Group1 database user. | |

Correct Answer:

| Actions | Answer Area |
|--|--|
| Create a database role named Role1 and grant Role1 SELECT permissions to schema1. | Create a database user in dw1 that represents Group1 and uses the FROM EXTERNAL PROVIDER clause. |
| Create a database role named Role1 and grant Role1 SELECT permissions to dw1. | Create a database role named Role1 and grant Role1 SELECT permissions to schema1. |
| Assign the Azure role-based access control (Azure RBAC) Reader role for dw1 to Group1. | Assign Role1 to the Group1 database user. |
| Create a database user in dw1 that represents Group1 and uses the FROM EXTERNAL PROVIDER clause. | |
| Assign Role1 to the Group1 database user. | |

Step 1: Create a database user named dw1 that represents Group1 and use the FROM EXTERNAL PROVIDER clause.

Step 2: Create a database role named Role1 and grant Role1 SELECT permissions to schema1.

Step 3: Assign Role1 to the Group1 database user.

Reference:

<https://docs.microsoft.com/en-us/azure/data-share/how-to-share-from-sql>

Question #2

HOTSPOT -

You have an Azure subscription that contains a logical Microsoft SQL server named Server1. Server1 hosts an Azure Synapse Analytics SQL dedicated pool named Pool1.

You need to recommend a Transparent Data Encryption (TDE) solution for Server1. The solution must meet the following requirements:

☞ Track the usage of encryption keys.

Maintain the access of client apps to Pool1 in the event of an Azure datacenter outage that affects the availability of the encryption keys.

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 track encryption key usage:

| |
|--------------------------------|
| Always Encrypted |
| TDE with customer-managed keys |
| TDE with platform-managed keys |

To maintain client app access in the event of a datacenter outage:

| |
|--|
| Create and configure Azure key vaults in two Azure regions. |
| Enable Advanced Data Security on Server1. |
| Implement the client apps by using a Microsoft .NET Framework data provider. |

Correct Answer:

Answer Area

To track encryption key usage:

| |
|--------------------------------|
| Always Encrypted |
| TDE with customer-managed keys |
| TDE with platform-managed keys |

To maintain client app access in the event of a datacenter outage:

| |
|--|
| Create and configure Azure key vaults in two Azure regions. |
| Enable Advanced Data Security on Server1. |
| Implement the client apps by using a Microsoft .NET Framework data provider. |

Box 1: TDE with customer-managed keys

Customer-managed keys are stored in the Azure Key Vault. You can monitor how and when your key vaults are accessed, and by whom. You can do this by enabling logging for Azure Key Vault, which saves information in an Azure storage account that you provide.

Box 2: Create and configure Azure key vaults in two Azure regions

The contents of your key vault are replicated within the region and to a secondary region at least 150 miles away, but within the same geography to maintain high durability of your keys and secrets.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/security/workspaces-encryption>

<https://docs.microsoft.com/en-us/azure/key-vault/general/logging>

Question #3

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

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

Which two components should you include in the solution? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. sensitivity-classification labels applied to columns that contain confidential information
- B. resource tags for databases that contain confidential information
- C. audit logs sent to a Log Analytics workspace
- D. dynamic data masking for columns that contain confidential information

Correct Answer:

- A. sensitivity-classification labels applied to columns that contain confidential information
- C. audit logs sent to a Log Analytics workspace

Question #4

You are designing an enterprise data warehouse in Azure Synapse Analytics that will contain a table named Customers. Customers will contain credit card information.

You need to recommend a solution to provide salespeople with the ability to view all the entries in Customers. The solution must prevent all the salespeople from viewing or inferring the credit card information.

What should you include in the recommendation?

- A. data masking
- B. Always Encrypted
- C. column-level security
- D. row-level security

Correct Answer:

- C. column-level security

Question #5

You develop data engineering solutions for a company.

A project requires the deployment of data to Azure Data Lake Storage.

You need to implement role-based access control (RBAC) so that project members can manage the Azure Data Lake Storage resources.

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

NOTE: Each correct selection is worth one point.

- A. Create security groups in Azure Active Directory (Azure AD) and add project members.
- B. Configure end-user authentication for the Azure Data Lake Storage account.
- C. Assign Azure AD security groups to Azure Data Lake Storage.
- D. Configure Service-to-service authentication for the Azure Data Lake Storage account.
- E. Configure access control lists (ACL) for the Azure Data Lake Storage account.

Correct Answer:

- A. Create security groups in Azure Active Directory (Azure AD) and add project members.
- C. Assign Azure AD security groups to Azure Data Lake Storage.

Question #6

You have an Azure Data Factory version 2 (V2) resource named Df1. Df1 contains a linked service.

You have an Azure Key vault named vault1 that contains an encryption key named key1.

You need to encrypt Df1 by using key1.

What should you do first?

- A. Add a private endpoint connection to vault1.
- B. Enable Azure role-based access control on vault1.
- C. Remove the linked service from Df1.
- D. Create a self-hosted integration runtime.

Correct Answer:

C. Remove the linked service from Df1.

Question #7

You are designing an Azure Synapse Analytics dedicated SQL pool.

You need to ensure that you can audit access to Personally Identifiable Information (PII).

What should you include in the solution?

- A. column-level security
- B. dynamic data masking
- C. row-level security (RLS)
- D. sensitivity classifications

Correct Answer:

D. sensitivity classifications

Question #8

HOTSPOT -

You have an Azure subscription that contains an Azure Data Lake Storage account. The storage account contains a data lake named DataLake1.

You plan to use an Azure data factory to ingest data from a folder in DataLake1, transform the data, and land the data in another folder.

You need to ensure that the data factory can read and write data from any folder in the DataLake1 container. The solution must meet the following requirements:

- Minimize the risk of unauthorized user access.
- Use the principle of least privilege.

- Minimize maintenance effort.

How should you configure access to the storage account for the data factory? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

Use

Microsoft Azure Active Directory (Azure AD), part of Microsoft Entra
a shared access signature (SAS)
a shared key

to authenticate by using

a managed identity.
a stored access policy.
an Authorization header.

Correct Answer:

Answer Area

Use

Microsoft Azure Active Directory (Azure AD), part of Microsoft Entra
a shared access signature (SAS)
a shared key

to authenticate by using

a managed identity.
a stored access policy.
an Authorization header.

Question #9

HOTSPOT -

You are designing an Azure Synapse Analytics dedicated SQL pool.

Groups will have access to sensitive data in the pool as shown in the following table.

| Name | Enhanced access |
|------------|--------------------------------------|
| Executives | No access to sensitive data |
| Analysts | Access to in-region sensitive data |
| Engineers | Access to all numeric sensitive data |

You have policies for the sensitive data. The policies vary by region as shown in the following table.

| Region | Data considered sensitive |
|---------|---|
| RegionA | Financial, Personally Identifiable Information (PII) |
| RegionB | Financial, Personally Identifiable Information (PII), medical |
| RegionC | Financial, medical |

You have a table of patients for each region. The tables contain the following potentially sensitive columns.

| Name | Sensitive data | Description |
|--------------|----------------|---|
| CardOnFile | Financial | Debit/credit card number for charges |
| Height | Medical | Patient's height in cm |
| ContactEmail | PII | Email address for secure communications |

You are designing dynamic data masking to maintain compliance.

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 |
|---|-----------------------|-----------------------|
| Analysts in RegionA require dynamic data masking rules for [Patients_RegionA]. | <input type="radio"/> | <input type="radio"/> |
| Engineers in RegionC require a dynamic data masking rule for [Patients_RegionA], [Height] | <input type="radio"/> | <input type="radio"/> |
| Engineers in RegionB require a dynamic data masking rule for [Patients_RegionB], [Height] | <input type="radio"/> | <input type="radio"/> |

Correct Answer:

Answer Area

| Statements | Yes | No |
|---|----------------------------------|----------------------------------|
| Analysts in RegionA require dynamic data masking rules for [Patients_RegionA]. | <input checked="" type="radio"/> | <input type="radio"/> |
| Engineers in RegionC require a dynamic data masking rule for [Patients_RegionA], [Height] | <input type="radio"/> | <input checked="" type="radio"/> |
| Engineers in RegionB require a dynamic data masking rule for [Patients_RegionB], [Height] | <input checked="" type="radio"/> | <input type="radio"/> |

Reference:

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

Question #10

DRAG DROP -

You have an Azure Synapse Analytics SQL pool named Pool1 on a logical Microsoft SQL server named Server1.

You need to implement Transparent Data Encryption (TDE) on Pool1 by using a custom key named key1.

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

Select and Place:

Actions

Enable TDE on Pool1.

Assign a managed identity to Server1.

Configure key1 as the TDE protector for Server1.

Add key1 to the Azure key vault.

Create an Azure key vault and grant the managed identity permissions to the key vault.

Answer Area



Correct Answer:

Actions

Answer Area

Assign a managed identity to Server1.

Create an Azure key vault and grant the managed identity permissions to the key vault.



Add key1 to the Azure key vault.

Configure key1 as the TDE protector for Server1.

Enable TDE on Pool1.

Step 1: Assign a managed identity to Server1

You will need an existing Managed Instance as a prerequisite.

Step 2: Create an Azure key vault and grant the managed identity permissions to the vault

Create Resource and setup Azure Key Vault.

Step 3: Add key1 to the Azure key vault

The recommended way is to import an existing key from a .pfx file or get an existing key from the vault. Alternatively, generate a new key directly in Azure Key Vault.

Step 4: Configure key1 as the TDE protector for Server1

Provide TDE Protector key -

Step 5: Enable TDE on Pool1 -

Reference:

<https://docs.microsoft.com/en-us/azure/azure-sql/managed-instance/scripts/transparent-data-en-cryption-byok-powershell>

Question #11

You have a data warehouse in Azure Synapse Analytics.

You need to ensure that the data in the data warehouse is encrypted at rest.

What should you enable?

- A. Advanced Data Security for this database
- B. Transparent Data Encryption (TDE)
- C. Secure transfer required
- D. Dynamic Data Masking

Correct Answer:

B. Transparent Data Encryption (TDE)

Question #12

You are designing a streaming data solution that will ingest variable volumes of data.

You need to ensure that you can change the partition count after creation.

Which service should you use to ingest the data?

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

Correct Answer:

A. Azure Event Hubs Dedicated

Question #13

You are designing a date dimension table in an Azure Synapse Analytics dedicated SQL pool.

The date dimension table will be used by all the fact tables.

Which distribution type should you recommend to minimize data movement during queries?

- A. HASH
- B. REPLICATE
- C. ROUND_ROBIN

Correct Answer:

B. REPLICATE

Question #14

HOTSPOT -

You develop a dataset named DBTBL1 by using Azure Databricks.

DBTBL1 contains the following columns:

- ☞ SensorTypeID
- ☞ GeographyRegionID
- ☞ Year
- ☞ Month
- ☞ Day
- ☞ Hour
- ☞ Minute
- ☞ Temperature
- ☞ WindSpeed
- ☞ Other

You need to store the data to support daily incremental load pipelines that vary for each GeographyRegionID. 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 | (“*”) |
| .format | (“GeographyRegionID”) |
| .partitionBy | (“GeographyRegionID”, “Year”, “Month”, “Day”) |
| .sortBy | (“Year”, “Month”, “Day”, “GeographyRegionID”) |

.mode (“append”)

| |
|-------------------------|
| .csv(“/DBTBL1”) |
| .json(“/DBTBL1”) |
| .parquet(“/DBTBL1”) |
| .saveAsTable(“/DBTBL1”) |

Correct Answer:

Answer Area

```
df.write
```

| | |
|--------------------------------|---|
| .bucketBy | (“*”) |
| .format | (“GeographyRegionID”) |
| .partitionBy | (“GeographyRegionID”, “Year”, “Month”, “Day”) |
| .sortBy | (“Year”, “Month”, “Day”, “GeographyRegionID”) |
| .mode (“append”) | |
| .csv(“/DBTBL1”) | |
| .json(“/DBTBL1”) | |
| .parquet(“/DBTBL1”) | |
| .saveAsTable(“/DBTBL1”) | |

Box 1: .partitionBy -

Incorrect Answers:

⇒ .format:

Method: format():

Arguments: "parquet", "csv", "txt", "json", "jdbc", "orc", "avro", etc.

⇒ .bucketBy:

Method: bucketBy()

Arguments: (numBuckets, col, col..., coln)

The number of buckets and names of columns to bucket by. Uses Hive's bucketing scheme on a filesystem.

Box 2: ("Year", "Month", "Day", "GeographyRegionID")

Specify the columns on which to do the partition. Use the date columns followed by the GeographyRegionID column.

Box 3: .saveAsTable("/DBTBL1")

Method: saveAsTable()

Argument: "table_name"

The table to save to.

Reference:

<https://www.oreilly.com/library/view/learning-spark-2nd/9781492050032/ch04.html>

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

Question #15

You are designing a security model for an Azure Synapse Analytics dedicated SQL pool that will support multiple companies.

You need to ensure that users from each company can view only the data of their respective company.

Which two objects should you include in the solution? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. a security policy
- B. a custom role-based access control (RBAC) role
- C. a predicate function
- D. a column encryption key
- E. asymmetric keys

Correct Answer:

- A. a security policy
- C. a predicate function

Question #16

You have a SQL pool in Azure Synapse that contains a table named dbo.Customers. The table contains a column name Email.

You need to prevent nonadministrative users from seeing the full email addresses in the Email column. The users must see values in a format of aXXX@XXXX.com instead.

What should you do?

- A. From Microsoft SQL Server Management Studio, set an email mask on the Email column.
- B. From the Azure portal, set a mask on the Email column.
- C. From Microsoft SQL Server Management Studio, grant the SELECT permission to the users for all the columns in the dbo.Customers table except Email.
- D. From the Azure portal, set a sensitivity classification of Confidential for the Email column.

Correct Answer:

- A. From Microsoft SQL Server Management Studio, set an email mask on the Email column.

Question #17

You have an Azure Data Lake Storage Gen2 account named adls2 that is protected by a virtual network.

You are designing a SQL pool in Azure Synapse that will use adls2 as a source.

What should you use to authenticate to adls2?

- A. an Azure Active Directory (Azure AD) user

- B. a shared key
- C. a shared access signature (SAS)
- D. a managed identity

Correct Answer:

- D. a managed identity

Question #18

HOTSPOT -

You have an Azure Synapse Analytics SQL pool named Pool1. In Azure Active Directory (Azure AD), you have a security group named Group1.

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

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

To control access to the columns:

| | |
|-------------------------------|---|
| | ▼ |
| CREATE CRYPTOGRAPHIC PROVIDER | |
| CREATE PARTITION FUNCTION | |
| CREATE SECURITY POLICY | |
| GRANT | |

To control access to the rows:

| | |
|-------------------------------|---|
| | ▼ |
| CREATE CRYPTOGRAPHIC PROVIDER | |
| CREATE PARTITION FUNCTION | |
| CREATE SECURITY POLICY | |
| GRANT | |

Correct Answer:

Answer Area

To control access to the columns:

| | |
|-------------------------------|---|
| | ▼ |
| CREATE CRYPTOGRAPHIC PROVIDER | |
| CREATE PARTITION FUNCTION | |
| CREATE SECURITY POLICY | |
| GRANT | |

To control access to the rows:

| | |
|-------------------------------|---|
| | ▼ |
| CREATE CRYPTOGRAPHIC PROVIDER | |
| CREATE PARTITION FUNCTION | |
| CREATE SECURITY POLICY | |
| GRANT | |

Box 1: GRANT -

You can implement column-level security with the GRANT T-SQL statement. With this mechanism, both SQL and Azure Active Directory (Azure AD) authentication are supported.

Box 2: CREATE SECURITY POLICY -

Implement RLS by using the CREATE SECURITY POLICY Transact-SQL statement, and predicates created as inline table-valued functions.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/column-level-security> <https://docs.microsoft.com/en-us/sql/relational-databases/security/row-level-security>

Question #19

HOTSPOT -

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

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Tier:

| | |
|----------|---|
| Premium | ▼ |
| Standard | |

Advanced option to enable:

| | |
|--|---|
| Azure Data Lake Storage Credential Passthrough | ▼ |
| Table Access Control | |

Correct Answer:

Answer Area

Tier:

| | |
|----------|---|
| Premium | ▼ |
| Standard | |

Advanced option to enable:

| | |
|--|---|
| Azure Data Lake Storage Credential Passthrough | ▼ |
| Table Access Control | |

Box 1: Premium -

Credential passthrough requires an Azure Databricks Premium Plan

Box 2: Azure Data Lake Storage credential passthrough

You can access Azure Data Lake Storage using Azure Active Directory credential passthrough.

When you enable your cluster for Azure Data Lake Storage credential passthrough, commands that you run on that cluster can read and write data in Azure Data

Lake Storage without requiring you to configure service principal credentials for access to storage.

Reference:

<https://docs.microsoft.com/en-us/azure/databricks/security/credential-passthrough/adls-passthrough>

Question #20

You are designing an Azure Synapse solution that will provide a query interface for the data stored in an Azure Storage account. The storage account is only accessible from a virtual network.

You need to recommend an authentication mechanism to ensure that the solution can access the source data.

What should you recommend?

- A. a managed identity
- B. anonymous public read access
- C. a shared key

Correct Answer:

- A. a managed identity

Question #21

You are developing an application that uses Azure Data Lake Storage Gen2.

You need to recommend a solution to grant permissions to a specific application for a limited time period.

What should you include in the recommendation?

- A. role assignments
- B. shared access signatures (SAS)
- C. Azure Active Directory (Azure AD) identities
- D. account keys

Correct Answer:

- B. shared access signatures (SAS)

Question #22

HOTSPOT -

You use Azure Data Lake Storage Gen2 to store data that data scientists and data engineers will query by using Azure Databricks interactive notebooks. Users will have access only to the Data Lake Storage folders that relate to the projects on which they work.

You need to recommend which authentication methods to use for Databricks and Data Lake Storage to provide the users with the appropriate access. The solution must minimize administrative effort and development effort.

Which authentication method should you recommend for each Azure service? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Databricks:

| |
|---|
| Azure Active Directory credential passthrough |
| Azure Key Vault secrets |
| Personal access tokens |

Data Lake Storage:

| |
|---|
| Azure Active Directory credential passthrough |
| Shared access keys |
| Shared access signatures |

Correct Answer:

Answer Area

Databricks:

| |
|---|
| Azure Active Directory credential passthrough |
| Azure Key Vault secrets |
| Personal access tokens |

Data Lake Storage:

| |
|---|
| Azure Active Directory credential passthrough |
| Shared access keys |
| Shared access signatures |

Box 1: Personal access tokens -

You can use storage shared access signatures (SAS) to access an Azure Data Lake Storage Gen2 storage account directly. With SAS, you can restrict access to a storage account using temporary tokens with fine-grained access control.

You can add multiple storage accounts and configure respective SAS token providers in the same Spark session.

Box 2: Azure Active Directory credential passthrough

You can authenticate automatically to Azure Data Lake Storage Gen1 (ADLS Gen1) and Azure Data Lake Storage Gen2 (ADLS Gen2) from Azure Databricks clusters using the same Azure Active Directory (Azure AD) identity that you use to log into Azure Databricks. When you enable your cluster for Azure Data Lake

Storage credential passthrough, commands that you run on that cluster can read and write data in Azure Data Lake Storage without requiring you to configure service principal credentials for access to storage.

After configuring Azure Data Lake Storage credential passthrough and creating storage containers, you can access data directly in Azure Data Lake Storage

Gen1 using an adl:// path and Azure Data Lake Storage Gen2 using an abfss:// path:

Reference:

<https://docs.microsoft.com/en-us/azure/databricks/data/data-sources/azure/adls-gen2/azure-data-lake-gen2-sas-access>

<https://docs.microsoft.com/en-us/azure/databricks/security/credential-passthrough/adls-passthrough>

Question #23

You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Contacts. Contacts contains a column named Phone.

You need to ensure that users in a specific role only see the last four digits of a phone number when querying the Phone column.

What should you include in the solution?

- A. table partitions
- B. a default value
- C. row-level security (RLS)
- D. column encryption
- E. dynamic data masking

Correct Answer:

- E. dynamic data masking

Question #24

You are designing database for an Azure Synapse Analytics dedicated SQL pool to support workloads for detecting ecommerce transaction fraud.

Data will be combined from multiple ecommerce sites and can include sensitive financial information such as credit card numbers.

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

Users must be able to identify potentially fraudulent transactions.

⇒ Users must be able to use credit cards as a potential feature in models.

⇒ Users must NOT be able to access the actual credit card numbers.
What should you include in the recommendation?

- A. Transparent Data Encryption (TDE)
- B. row-level security (RLS)
- C. column-level encryption
- D. Azure Active Directory (Azure AD) pass-through authentication

Correct Answer:

- C. column-level encryption

Question #25

You have an Azure subscription linked to an Azure Active Directory (Azure AD) tenant that contains a service principal named ServicePrincipal1. The subscription contains an Azure Data Lake Storage account named adls1. Adls1 contains a folder named Folder2 that has a URL of <https://adls1.dfs.core.windows.net/container1/Folder1/Folder2/>.
ServicePrincipal1 has the access control list (ACL) permissions shown in the following table.

| Resource | Permission |
|------------|------------------|
| container1 | Access – Execute |
| Folder1 | Access – Execute |
| Folder2 | Access – Read |

You need to ensure that ServicePrincipal1 can perform the following actions:

- ⇒ Traverse child items that are created in Folder2.
- ⇒ Read files that are created in Folder2.

The solution must use the principle of least privilege.

Which two permissions should you grant to ServicePrincipal1 for Folder2? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. Access " Read
- B. Access " Write
- C. Access " Execute
- D. Default " Read
- E. Default " Write
- F. Default " Execute

Correct Answer:

- C. Access " Execute
- D. Default " Read

Question #26

HOTSPOT -

You have an Azure subscription that is linked to a hybrid Azure Active Directory (Azure AD) tenant. The subscription contains an Azure Synapse Analytics SQL pool named Pool1.

You need to recommend an authentication solution for Pool1. The solution must support multi-factor authentication (MFA) and database-level authentication.

Which authentication solution or solutions 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

MFA:

| | |
|-------------------------------------|---|
| Azure AD authentication | ▼ |
| Microsoft SQL Server authentication | ▼ |
| Passwordless authentication | ▼ |
| Windows authentication | ▼ |

Database-level authentication:

| | |
|-----------------------------|---|
| Application roles | ▼ |
| Contained database users | ▼ |
| Database roles | ▼ |
| Microsoft SQL Server logins | ▼ |

Correct Answer:

Answer Area

MFA:

| | |
|-------------------------------------|---|
| Azure AD authentication | ▼ |
| Microsoft SQL Server authentication | |
| Passwordless authentication | |
| Windows authentication | |

Database-level authentication:

| | |
|-----------------------------|---|
| Application roles | ▼ |
| Contained database users | |
| Database roles | |
| Microsoft SQL Server logins | |

Box 1: Azure AD authentication -

Azure AD authentication has the option to include MFA.

Box 2: Contained database users -

Azure AD authentication uses contained database users to authenticate identities at the database level.

Reference:

<https://docs.microsoft.com/en-us/azure/azure-sql/database/authentication-mfa-ssms-overview>
<https://docs.microsoft.com/en-us/azure/azure-sql/database/authentication-aad-overview>

Question #27

DRAG DROP -

You have an Azure data factory.

You need to ensure that pipeline-run data is retained for 120 days. The solution must ensure that you can query the data by using the Kusto query language.

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

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

Select and Place:

Actions**Answer Area**

Select the PipelineRuns category.

Create a Log Analytics workspace that has Data Retention set to 120 days.

Stream to an Azure event hub.

Create an Azure Storage account that has a lifecycle policy.

From the Azure portal, add a diagnostic setting.

Send the data to a Log Analytics workspace.

Select the TriggerRuns category.

Correct Answer:

| Actions | Answer Area |
|---|---|
| Select the PipelineRuns category. | Create an Azure Storage account that has a lifecycle policy. |
| Create a Log Analytics workspace that has Data Retention set to 120 days. | Create a Log Analytics workspace that has Data Retention set to 120 days. |
| Stream to an Azure event hub. | From the Azure portal, add a diagnostic setting. |
| Create an Azure Storage account that has a lifecycle policy. | Send the data to a Log Analytics workspace. |
| From the Azure portal, add a diagnostic setting. | |
| Send the data to a Log Analytics workspace. | |
| Select the TriggerRuns category. | |

Step 1: Create an Azure Storage account that has a lifecycle policy

To automate common data management tasks, Microsoft created a solution based on Azure Data Factory. The service, Data Lifecycle Management, makes frequently accessed data available and archives or purges other data according to retention policies. Teams across the company use the service to reduce storage costs, improve app performance, and comply with data retention policies.

Step 2: Create a Log Analytics workspace that has Data Retention set to 120 days.

Data Factory stores pipeline-run data for only 45 days. Use Azure Monitor if you want to keep that data for a longer time. With Monitor, you can route diagnostic logs for analysis to multiple different targets, such as a Storage Account: Save your diagnostic logs to a storage account for auditing or manual inspection. You can use the diagnostic settings to specify the retention time in days.

Step 3: From Azure Portal, add a diagnostic setting.

Step 4: Send the data to a log Analytics workspace,

Event Hub: A pipeline that transfers events from services to Azure Data Explorer.

Keeping Azure Data Factory metrics and pipeline-run data.

Configure diagnostic settings and workspace.

Create or add diagnostic settings for your data factory.

1. In the portal, go to Monitor. Select Settings > Diagnostic settings.

2. Select the data factory for which you want to set a diagnostic setting.

3. If no settings exist on the selected data factory, you're prompted to create a setting. Select Turn on diagnostics.

4. Give your setting a name, select Send to Log Analytics, and then select a workspace from Log Analytics Workspace.

5. Select Save.

Reference:

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

Question #28

You have an Azure Synapse Analytics dedicated SQL pool.

You need to ensure that data in the pool is encrypted at rest. The solution must NOT require modifying applications that query the data.

What should you do?

- A. Enable encryption at rest for the Azure Data Lake Storage Gen2 account.
- B. Enable Transparent Data Encryption (TDE) for the pool.
- C. Use a customer-managed key to enable double encryption for the Azure Synapse workspace.
- D. Create an Azure key vault in the Azure subscription grant access to the pool.

Correct Answer:

B. Enable Transparent Data Encryption (TDE) for the pool.

Question #29

DRAG DROP -

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

Container1 contains a directory named directory1. Directory1 contains a file named file1.

You have an Azure Active Directory (Azure AD) user named User1 that is assigned the Storage Blob Data Reader role for storage1.

You need to ensure that User1 can append data to file1. The solution must use the principle of least privilege.

Which permissions should you grant? To answer, drag the appropriate permissions to the correct resources. Each permission may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.

Select and Place:

Permissions

Read

Write

Execute

Answer Areacontainer1: directory1: file1: **Correct Answer:****Permissions**

Read

Write

Execute

Answer Areacontainer1: Executedirectory1: Executefile1: Write**Box 1: Execute -**

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

Box 2: Execute -

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

Box 3: Write -

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

Reference:

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

Question #30

HOTSPOT -

You have an Azure subscription that contains an Azure Databricks workspace named databricks1 and an Azure Synapse Analytics workspace named synapse1.

The synapse1 workspace contains an Apache Spark pool named pool1.

You need to share an Apache Hive catalog of pool1 with databricks1.

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

NOTE: Each correct selection is worth one point.

Hot Area:

From synapse1, create a linked service to:

- Azure Cosmos DB
- Azure Data Lake Storage Gen2
- Azure SQL Database

Configure pool1 to use the linked service as:

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

Correct Answer:

From synapse1, create a linked service to:

- Azure Cosmos DB
- Azure Data Lake Storage Gen2
- Azure SQL Database

Configure pool1 to use the linked service as:

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

Box 1: Azure SQL Database -

Use external Hive Metastore for Synapse Spark Pool

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

Set up linked service to Hive Metastore

Follow below steps to set up a linked service to the external Hive Metastore in Synapse workspace.

1. Open Synapse Studio, go to Manage > Linked services at left, click New to create a new linked service.
2. Set up Hive Metastore linked service
3. Choose Azure SQL Database or Azure Database for MySQL based on your database type, click Continue.

4. Provide Name of the linked service. Record the name of the linked service, this info will be used to configure Spark shortly.
5. You can either select Azure SQL Database/Azure Database for MySQL for the external Hive Metastore from Azure subscription list, or enter the info manually.
6. Provide User name and Password to set up the connection.
7. Test connection to verify the username and password.
8. Click Create to create the linked service.

Box 2: A Hive Metastore -

Reference:

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

Question #31

HOTSPOT -

You have an Azure subscription.

You need to deploy an Azure Data Lake Storage Gen2 Premium account. The solution must meet the following requirements:

- * Blobs that are older than 365 days must be deleted.
- * Administrative effort must be minimized.
- * Costs must be minimized.

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

NOTE: Each correct selection is worth one point.

Hot Area:

To minimize costs:

| |
|---------------------------------|
| Locally-redundant storage (LRS) |
| The Archive access tier |
| The Cool access tier |
| Zone-redundant storage (ZRS) |

To delete blobs:

| |
|------------------------------------|
| Azure Automation runbooks |
| Azure Storage lifecycle management |
| Soft delete |

Correct Answer:

To minimize costs:

| |
|---------------------------------|
| Locally-redundant storage (LRS) |
| The Archive access tier |
| The Cool access tier |
| Zone-redundant storage (ZRS) |

To delete blobs:

| |
|------------------------------------|
| Azure Automation runbooks |
| Azure Storage lifecycle management |
| Soft delete |

Box 1: The Archive access tier -

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 Storage lifecycle management

With the lifecycle management policy, you can:

* Delete current versions of a blob, previous versions of a blob, or blob snapshots at the end of their lifecycles.

Transition blobs from cool to hot immediately when they're accessed, to optimize for performance.

Transition current versions of a blob, previous versions of a blob, or blob snapshots to a cooler storage tier if these objects haven't been accessed or modified for a period of time, to optimize for cost. In this scenario, the lifecycle management policy can move objects from hot to cool, from hot to archive, or from cool to archive.

Etc.

Reference:

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

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

Question #32

HOTSPOT -

You are designing an application that will use an Azure Data Lake Storage Gen 2 account to store petabytes of license plate photos from toll booths. The account will use zone-redundant storage (ZRS).

You identify the following usage patterns:

- * The data will be accessed several times a day during the first 30 days after the data is created.
- The data must meet an availability SLA of 99.9%.
- * After 90 days, the data will be accessed infrequently but must be available within 30 seconds.
- * After 365 days, the data will be accessed infrequently but must be available within five minutes.

You need to recommend a data retention solution. The solution must minimize costs.

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

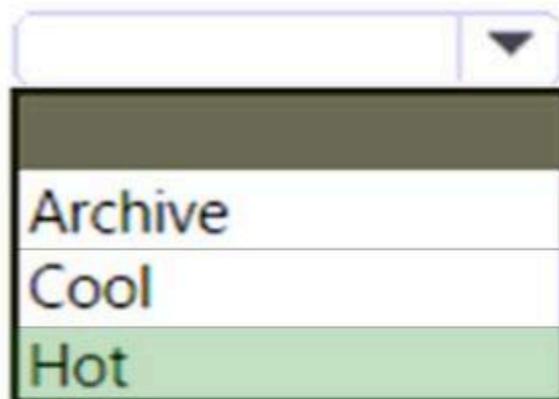
NOTE: Each correct selection is worth one point.

Hot Area:

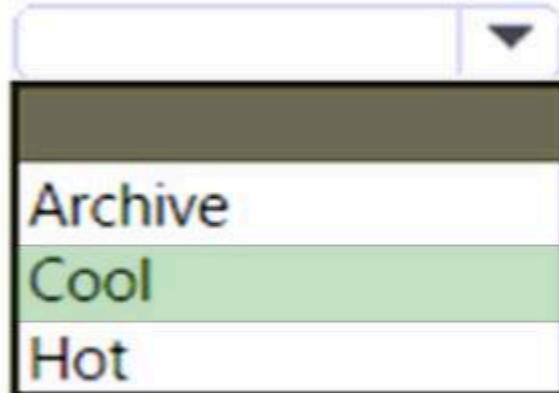
| | | | |
|-----------------|----------------------------------|-------------------------------|---|
| First 30 days: | <input type="checkbox"/> Archive | <input type="checkbox"/> Cool | <input checked="" type="checkbox"/> Hot |
| After 90 days: | <input type="checkbox"/> Archive | <input type="checkbox"/> Cool | <input checked="" type="checkbox"/> Hot |
| After 365 days: | <input type="checkbox"/> Archive | <input type="checkbox"/> Cool | <input checked="" type="checkbox"/> Hot |

Correct Answer:

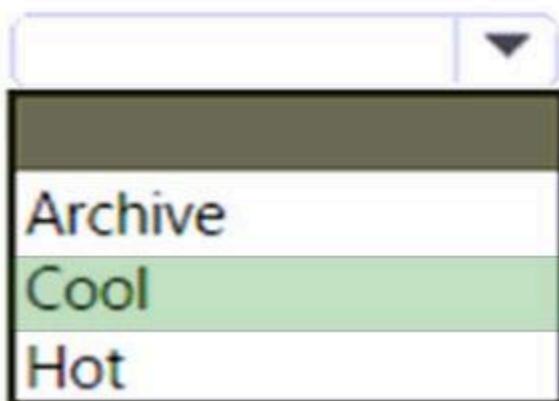
First 30 days:



After 90 days:



After 365 days:



Box 1: Hot -

The data will be accessed several times a day during the first 30 days after the data is created.
The data must meet an availability SLA of 99.9%.

Box 2: Cool -

After 90 days, the data will be accessed infrequently but must be available within 30 seconds.
Data in the Cool tier should be stored for a minimum of 30 days.

When your data is stored in an online access tier (either Hot or Cool), users can access it immediately. The Hot tier is the best choice for data that is in active use, while the Cool tier is ideal for data that is accessed less frequently, but that still must be available for reading and writing.

Box 3: Cool -

After 365 days, the data will be accessed infrequently but must be available within five minutes.

Incorrect:

Not Archive:

While a blob is in the Archive access tier, it's considered to be offline and can't be read or modified. In order to read or modify data in an archived blob, you must first rehydrate the blob to an online tier, either the Hot or Cool tier.

Rehydration priority -

When you rehydrate a blob, you can set the priority for the rehydration operation via the optional `x-ms-rehydrate-priority` header on a Set Blob Tier or Copy Blob operation. Rehydration priority options include:

Standard priority: The rehydration request will be processed in the order it was received and may take up to 15 hours.

High priority: The rehydration request will be prioritized over standard priority requests and may complete in less than one hour for objects under 10 GB in size.

Reference:

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

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

Question #33

DRAG DROP

-

You have an Azure Data Lake Storage Gen 2 account named storage1.

You need to recommend a solution for accessing the content in storage1. The solution must meet the following requirements:

- List and read permissions must be granted at the storage account level.
- Additional permissions can be applied to individual objects in storage1.
- Security principals from Microsoft Azure Active Directory (Azure AD), part of Microsoft Entra, must be used for authentication.

What should you use? To answer, drag the appropriate components to the correct requirements. Each component may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point.

| Components | Answer Area |
|--|--|
| Access control lists (ACLs) | To grant permissions at the storage account level: |
| Role-based access control (RBAC) roles | To grant permissions at the object level: |
| Shared access signatures (SAS) | |
| Shared account keys | |

Correct Answer:

Answer Area

To grant permissions at the storage account level:

To grant permissions at the object level:

Question #34

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

Sales has row-level security (RLS) applied. RLS uses the following predicate filter.

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

A user named SalesUser1 is assigned the db_datareader role for Pool1.

Which rows in the Sales table are returned when SalesUser1 queries the table?

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

Correct Answer:

- D. only the rows for which the value in the SalesRep column is SalesUser1

Question #35

HOTSPOT

You have an Azure Data Lake Storage Gen2 account named account1 that contains the resources shown in the following table.

| Name | Type | Description |
|------------|-----------|---------------------------|
| container1 | Container | A container |
| Directory1 | Directory | A directory in container1 |
| File1 | File | A file in Directory1 |

You need to configure access control lists (ACLs) to allow a user named User1 to delete File1. User1 is NOT assigned any role-based access control (RBAC) roles for account1. The solution must use the principle of least privilege.

Which type of ACL should you configure for each resource? To answer select the appropriate options in the answer area.

Answer Area

container1:

| | |
|---|---|
| | ▼ |
| --- permissions -WX permissions --X permissions | |

Directory1:

| | |
|---|---|
| | ▼ |
| --- permissions -WX permissions --X permissions | |

File1:

| | |
|---|---|
| | ▼ |
| --- permissions -WX permissions --X permissions | |

Correct Answer:

Answer Area

| | |
|-------------|---|
| container1: | <p>--- permissions -WX permissions --X permissions</p> |
| Directory1: | <p>--- permissions -WX permissions --X permissions</p> |
| File1: | <p>--- permissions -WX permissions --X permissions</p> |

Question #36

You have an Azure subscription that is linked to a tenant in Microsoft Azure Active Directory (Azure AD), part of Microsoft Entra. The tenant that contains a security group named Group1. The subscription contains an Azure Data Lake Storage account named myaccount1. The myaccount1 account contains two containers named container1 and container2.

You need to grant Group1 read access to container1. The solution must use the principle of least privilege.

Which role should you assign to Group1?

- A. Storage Table Data Reader for myaccount1
- B. Storage Blob Data Reader for container1
- C. Storage Blob Data Reader for myaccount1
- D. Storage Table Data Reader for container1

Correct Answer:

- B. Storage Blob Data Reader for container1

Question #37

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

You need to prevent a group of users from reading user email addresses from dbo.Users.

What should you use?

- A. column-level security
- B. row-level security (RLS)
- C. Transparent Data Encryption (TDE)
- D. dynamic data masking

Correct Answer:

- A. column-level security

Question #38

HOTSPOT

-

You have an Azure Synapse Analytics dedicated SQL pool that hosts a database named DB1.

You need to ensure that DB1 meets the following security requirements:

- When credit card numbers show in applications, only the last four digits must be visible.
- Tax numbers must be visible only to specific users.

What should you use for each requirement? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

Credit card numbers:

- Column-level security
- Dynamic Data Masking
- Row-level security (RLS)

Tax numbers:

- Column-level security
- Row-level security (RLS)
- Transparent Database Encryption (TDE)

Correct Answer:

Answer Area

Credit card numbers:

- Column-level security
- Dynamic Data Masking
- Row-level security (RLS)

Tax numbers:

- Column-level security
- Row-level security (RLS)
- Transparent Database Encryption (TDE)

Question #39

You have an Azure subscription that contains a storage account named storage1 and an Azure Synapse Analytics dedicated SQL pool. The storage1 account contains a CSV file that requires an account key for access.

You plan to read the contents of the CSV file by using an external table.

You need to create an external data source for the external table.

What should you create first?

- A. a database role
- B. a database scoped credential
- C. a database view
- D. an external file format

Correct Answer:

- B. a database scoped credential

Question #40

You have a tenant in Microsoft Azure Active Directory (Azure AD), part of Microsoft Entra. The tenant contains a group named Group1.

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

| Name | Type | Description |
|-------------|-----------------------------------|---|
| ws1 | Azure Synapse Analytics workspace | None |
| storage1 | Azure Storage account | Contains CSV files |
| credential1 | Database-scoped credential | Stored in the Azure Synapse Analytics serverless SQL pool in ws1 and used to authenticate to storage1 |

You need to ensure that members of Group1 can read CSV files from storage1 by using the OPENROWSET function. The solution must meet the following requirements:

- The members of Group1 must use credential1 to access storage1.
- The principle of least privilege must be followed.

Which permission should you grant to Group1?

- A. EXECUTE
- B. CONTROL
- C. REFERENCES
- D. SELECT

Correct Answer:

- C. REFERENCES

Question #41

You have an Azure subscription that contains an Azure Data Lake Storage account named dl1 and an Azure Analytics Synapse workspace named workspace1.

You need to query the data in dl1 by using an Apache Spark pool named Pool1 in workspace1. The solution must ensure that the data is accessible Pool1.

Which two actions achieve the goal? Each correct answer presents a complete solution.

NOTE: Each correct answer is worth one point.

- A. Implement Azure Synapse Link.
- B. Load the data to the primary storage account of workspace1.
- C. From workspace1, create a linked service for the dl1.
- D. From Microsoft Purview, register dl1 as a data source.

Correct Answer:

- B. Load the data to the primary storage account of workspace1.
- C. From workspace1, create a linked service for the dl1.

Question #42

You have an Azure Synapse Analytics dedicated SQL pool named SQL1 and a user named User1.

You need to ensure that User1 can view requests associated with SQL1 by querying the sys.dm_pdw_exec_requests dynamic management view. The solution must follow the principle of least privilege.

Which permission should you grant to User1?

- A. VIEW DATABASE STATE
- B. SHOWPLAN
- C. CONTROL SERVER
- D. VIEW ANY DATABASE

Correct Answer:

- A. VIEW DATABASE STATE

Question #43

You have a Microsoft Entra tenant.

The tenant contains an Azure Data Lake Storage Gen2 account named storage1 that has two containers named fs1 and fs2.

You have a Microsoft Entra group named DepartmentA.

You need to meet the following requirements:

- DepartmentA must be able to read, write, and list all the files in fs1.
- DepartmentA must be prevented from accessing any files in fs2.
- The solution must use the principle of least privilege.

Which role should you assign to DepartmentA?

- A. Contributor for fs1
- B. Storage Blob Data Owner for fs1
- C. Storage Blob Data Contributor for storage1
- D. Storage Blob Data Contributor for fs1

Correct Answer:

- D. Storage Blob Data Contributor for fs1

Topic 4

Question #1

You implement an enterprise data warehouse in Azure Synapse Analytics.

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

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

| SaleKey | CityKey | CustomerKey | StockItemKey | InvoiceDateKey | Quantity | UnitPrice | TotalExcludingTax |
|---------|---------|-------------|--------------|----------------|----------|-----------|-------------------|
| 49309 | 90858 | 70 | 69 | 10/22/13 | 8 | 16 | 128 |
| 49313 | 55710 | 126 | 69 | 10/22/13 | 2 | 16 | 32 |
| 49343 | 44710 | 234 | 68 | 10/22/13 | 10 | 16 | 160 |
| 49352 | 66109 | 163 | 70 | 10/22/13 | 4 | 16 | 64 |
| 49448 | 65312 | 230 | 70 | 10/22/13 | 8 | 16 | 128 |
| 49646 | 85877 | 271 | 70 | 10/24/13 | 1 | 16 | 16 |
| 49798 | 41238 | 288 | 69 | 10/24/13 | 1 | 16 | 16 |

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

Which technology should you use?

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

Correct Answer:

B. hash distributed table with clustered Columnstore index

Question #2

You have an Azure Synapse Analytics dedicated SQL pool that contains a large fact table. The table contains 50 columns and 5 billion rows and is a heap.

Most queries against the table aggregate values from approximately 100 million rows and return only two columns.

You discover that the queries against the fact table are very slow.

Which type of index should you add to provide the fastest query times?

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

Correct Answer:

B. clustered columnstore

Question #3

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

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

Correct Answer:

B. cluster event logs

Question #4

You have an Azure data factory. You need to examine the pipeline failures from the last 60 days. What should you use?

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

Correct Answer:

D. Azure Monitor

Question #5

You are monitoring an Azure Stream Analytics job. The Backlogged Input Events count has been 20 for the last hour. You need to reduce the Backlogged Input Events count. What should you do?

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

Correct Answer:

C. Increase the streaming units for the job.

Question #6

You are designing an Azure Databricks interactive cluster. The cluster will be used infrequently and will be configured for auto-termination.

You need to ensure that the cluster configuration is retained indefinitely after the cluster is terminated. The solution must minimize costs.

What should you do?

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

Correct Answer:

- A. Pin the cluster.

Question #7

You have an Azure data solution that contains an enterprise data warehouse in Azure Synapse Analytics named DW1.

Several users execute ad hoc queries to DW1 concurrently.

You regularly perform automated data loads to DW1.

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

What should you do?

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

Correct Answer:

- C. Assign a larger resource class to the automated data load queries.

Question #8

You have an Azure Synapse Analytics dedicated SQL pool named Pool1 and a database named DB1. DB1 contains a fact table named Table1.

You need to identify the extent of the data skew in Table1.

What should you do in Synapse Studio?

- A. Connect to the built-in pool and run DBCC PDW_SHOWSPACEUSED.
- B. Connect to the built-in pool and run DBCC CHECKALLOC.
- C. Connect to Pool1 and query sys.dm_pdw_node_status.
- D. Connect to Pool1 and query sys.dm_pdw_nodes_db_partition_stats.

Correct Answer:

- D. Connect to Pool1 and query sys.dm_pdw_nodes_db_partition_stats.

Question #9

HOTSPOT -

You need to collect application metrics, streaming query events, and application log messages for an Azure Databrick cluster.

Which type of library and workspace should you implement? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Library:

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

Workspace:

| |
|------------------------|
| Azure Databricks |
| Azure Log Analytics |
| Azure Machine Learning |

Correct Answer:

Answer Area

Library:

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

Workspace:

| |
|------------------------|
| Azure Databricks |
| Azure Log Analytics |
| Azure Machine Learning |

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

Reference:

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

Question #10

You have a SQL pool in Azure Synapse.

You discover that some queries fail or take a long time to complete.

You need to monitor for transactions that have rolled back.

Which dynamic management view should you query?

- A. sys.dm_pdw_request_steps
- B. sys.dm_pdw_nodes_tran_database_transactions
- C. sys.dm_pdw_waits
- D. sys.dm_pdw_exec_sessions

Correct Answer:

B. sys.dm_pdw_nodes_tran_database_transactions

Question #11

You are monitoring an Azure Stream Analytics job.

You discover that the Backlogged Input Events metric is increasing slowly and is consistently non-zero.

You need to ensure that the job can handle all the events.
What should you do?

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

Correct Answer:

- B. Increase the number of streaming units (SUs).

Question #12

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

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

You identify the following usage patterns:

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

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

On which column should you partition the table?

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

Correct Answer:

- D. WarehouseID

Question #13

You are designing a star schema for a dataset that contains records of online orders. Each record includes an order date, an order due date, and an order ship date.

You need to ensure that the design provides the fastest query times of the records when querying for arbitrary date ranges and aggregating by fiscal calendar attributes.

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

NOTE: Each correct selection is worth one point.

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

Correct Answer:

- C. Create a date dimension table that has an integer key in the format of YYYYMMDD.
- D. In the fact table, use integer columns for the date fields.

Question #14

A company purchases IoT devices to monitor manufacturing machinery. The company uses an Azure IoT Hub to communicate with the IoT devices.

The company must be able to monitor the devices in real-time.

You need to design the solution.

What should you recommend?

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

Correct Answer:

- C. Azure Stream Analytics cloud job using Azure Portal

Question #15

You have a SQL pool in Azure Synapse.

A user reports that queries against the pool take longer than expected to complete. You determine that the issue relates to queried columnstore segments.

You need to add monitoring to the underlying storage to help diagnose the issue.

Which two metrics should you monitor? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

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

Correct Answer:

- B. Cache used percentage
- D. Cache hit percentage

Question #16

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

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

Correct Answer:

- B. Cache hit percentage

Question #17

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

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

Correct Answer:

- A. clusters

Question #18

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

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

Correct Answer:

- D. Last Sync Time

Question #19

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

Files with an invalid schema cause errors to occur.

You need to monitor for an invalid schema error.

For which error should you monitor?

- A. EXTERNAL TABLE access failed due to internal error: 'Java exception raised on call to HdfsBridge_Connect: Error [com.microsoft.polybase.client.KerberosSecureLogin] occurred while accessing external file.'
- B. Cannot execute the query "Remote Query" against OLE DB provider "SQLNCLI11" for linked server "(null)". Query aborted- the maximum reject threshold (0 rows) was reached while reading from an external source: 1 rows rejected out of total 1 rows processed.
- C. EXTERNAL TABLE access failed due to internal error: 'Java exception raised on call to HdfsBridge_Connect: Error [Unable to instantiate LoginClass] occurred while accessing external file.'
- D. EXTERNAL TABLE access failed due to internal error: 'Java exception raised on call to HdfsBridge_Connect: Error [No FileSystem for scheme: wasbs] occurred while accessing external file.'

Correct Answer:

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

Question #20

You have an Azure Synapse Analytics dedicated SQL pool.

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

| ROWS | RESERVED_SPACE | DATA_SPACE | INDEX_SPACE | UNUSED_SPACE | PDW_NODE_ID | DISTRIBUTION_ID |
|------|----------------|------------|-------------|--------------|-------------|-----------------|
| 694 | 2776 | 616 | 48 | 2112 | 1 | 1 |
| 407 | 2704 | 576 | 48 | 2080 | 1 | 2 |
| 53 | 2376 | 512 | 16 | 1848 | 1 | 3 |
| 58 | 2376 | 512 | 16 | 1848 | 1 | 4 |
| 168 | 2632 | 528 | 32 | 2072 | 1 | 5 |
| 195 | 2696 | 536 | 32 | 2128 | 1 | 6 |
| 5995 | 3464 | 1424 | 32 | 2008 | 1 | 7 |
| 0 | 2232 | 496 | 0 | 1736 | 1 | 8 |
| 264 | 2576 | 544 | 40 | 1992 | 1 | 9 |
| 3008 | 3016 | 960 | 32 | 2024 | 1 | 10 |
| ... | ... | ... | ... | ... | ... | ... |
| 1550 | 2832 | 752 | 48 | 2032 | 1 | 50 |
| 1238 | 2832 | 696 | 40 | 2096 | 1 | 51 |
| 192 | 2632 | 528 | 32 | 2072 | 1 | 52 |
| 1127 | 2768 | 680 | 48 | 2040 | 1 | 53 |
| 1244 | 3032 | 704 | 64 | 2264 | 1 | 54 |
| 409 | 2632 | 568 | 32 | 2032 | 1 | 55 |
| 0 | 2232 | 496 | 0 | 1736 | 1 | 56 |
| 1437 | 2832 | 728 | 40 | 2064 | 1 | 57 |
| 0 | 2232 | 496 | 0 | 1736 | 1 | 58 |
| 384 | 2632 | 560 | 32 | 2040 | 1 | 59 |
| 225 | 2768 | 544 | 40 | 2184 | 1 | 60 |

Which statement accurately describes the dbo.FactInternetSales table?

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

Correct Answer:

- D. The table is skewed.

Question #21

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

| Table | Column |
|---------|------------------|
| Flight | ArrivalAirportID |
| | ArrivalDateTime |
| Weather | AirportID |
| | ReportDateTime |

You need to recommend a solution that maximizes query performance.

What should you include in the recommendation?

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

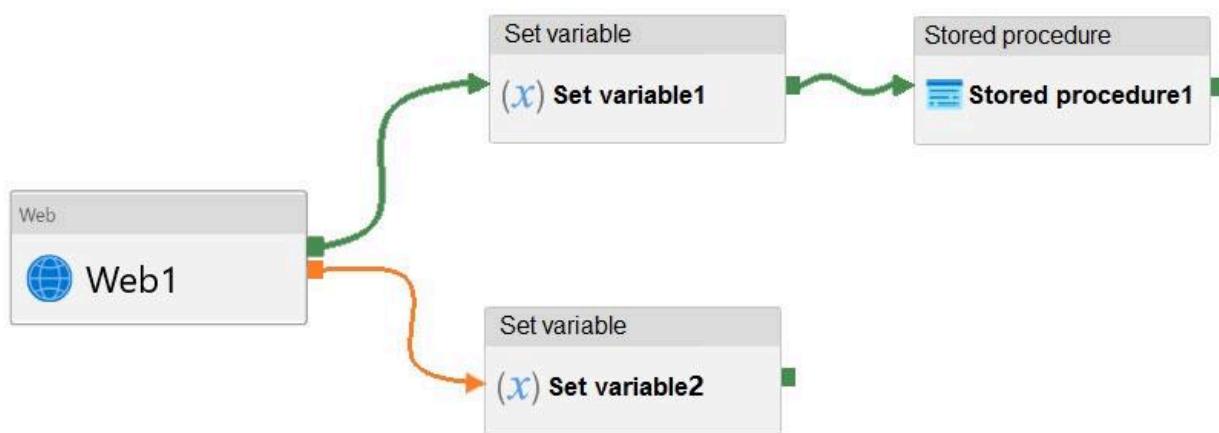
Correct Answer:

- B. In the tables use a hash distribution of ArrivalAirportID and AirportID.

Question #22

HOTSPOT -

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



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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

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

| |
|----------|
| ▼ |
| complete |
| fail |
| succeed |

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

| |
|-----------|
| ▼ |
| Canceled |
| Failed |
| Succeeded |

Correct Answer:

Answer Area

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

| |
|----------|
| ▼ |
| complete |
| fail |
| succeed |

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

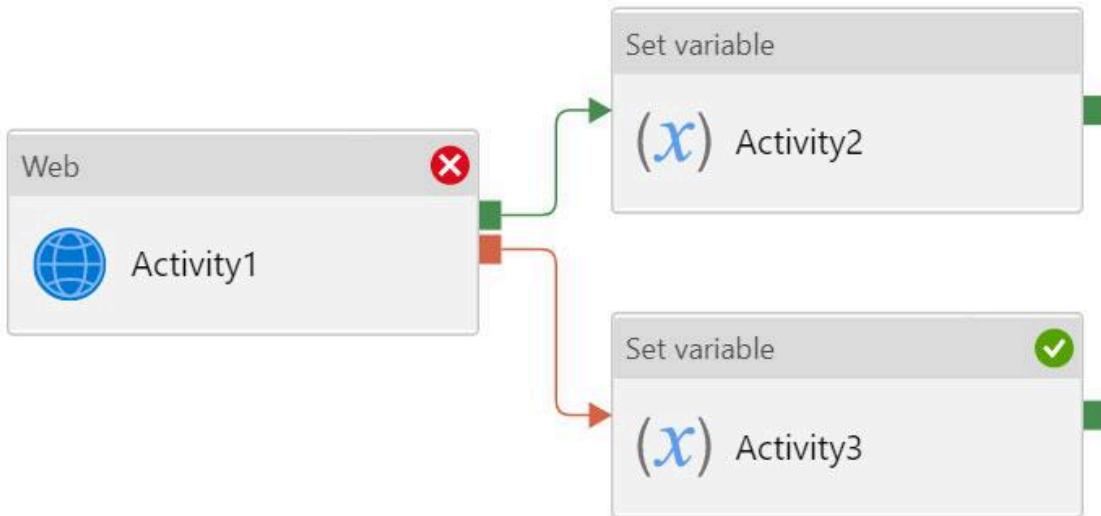
| |
|-----------|
| ▼ |
| Canceled |
| Failed |
| Succeeded |

Box 1: succeed -

Box 2: failed -

Example:

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



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

Reference:

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

Question #23

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

- ☞ Wrangling data flow
- ☞ Notebook
- ☞ Copy
- ☞ Jar

Which two Azure services should you use to debug the activities? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point

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

Correct Answer:

- D. Azure Data Factory
- E. Azure Databricks

Question #24

You have an Azure Synapse Analytics dedicated SQL pool named Pool1 and a database named DB1. DB1 contains a fact table named Table1.

You need to identify the extent of the data skew in Table1.

What should you do in Synapse Studio?

- A. Connect to the built-in pool and run sys.dm_pdw_nodes_db_partition_stats.
- B. Connect to Pool1 and run DBCC CHECKALLOC.
- C. Connect to the built-in pool and run DBCC CHECKALLOC.
- D. Connect to Pool1 and query sys.dm_pdw_nodes_db_partition_stats.

Correct Answer:

- D. Connect to Pool1 and query sys.dm_pdw_nodes_db_partition_stats.

Question #25

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. Local tempdb percentage
- B. Cache used percentage
- C. Data IO percentage
- D. CPU percentage

Correct Answer:

- B. Cache used percentage

Question #26

You have an Azure data factory.

You need to examine the pipeline failures from the last 180 days.

What should you use?

- A. the Activity log blade for the Data Factory resource
- B. Pipeline runs in the Azure Data Factory user experience
- C. the Resource health blade for the Data Factory resource
- D. Azure Data Factory activity runs in Azure Monitor

Correct Answer:

- D. Azure Data Factory activity runs in Azure Monitor

Question #27

A company purchases IoT devices to monitor manufacturing machinery. The company uses an Azure IoT Hub to communicate with the IoT devices.

The company must be able to monitor the devices in real-time.

You need to design the solution.

What should you recommend?

- A. Azure Analysis Services using Azure PowerShell
- B. Azure Stream Analytics Edge application using Microsoft Visual Studio
- C. Azure Analysis Services using Microsoft Visual Studio
- D. Azure Data Factory instance using Azure Portal

Correct Answer:

B. Azure Stream Analytics Edge application using Microsoft Visual Studio

Question #28

You have an Azure Synapse Analytics dedicated SQL pool named SA1 that contains a table named Table1.

You need to identify tables that have a high percentage of deleted rows.

What should you run?

- A. sys.pdw_nodes_column_store_segments
- B. sys.dm_db_column_store_row_group_operational_stats
- C. sys.pdw_nodes_column_store_row_groups
- D. sys.dm_db_column_store_row_group_physical_stats

Correct Answer:

C. sys.pdw_nodes_column_store_row_groups

Question #29

You have an enterprise data warehouse in Azure Synapse Analytics.

You need to monitor the data warehouse to identify whether you must scale up to a higher service level to accommodate the current workloads.

Which is the best metric to monitor?

More than one answer choice may achieve the goal. Select the BEST answer.

- A. DWU used
- B. CPU percentage
- C. DWU percentage
- D. Data IO percentage

Correct Answer:

C. DWU percentage

Question #30

A company purchases IoT devices to monitor manufacturing machinery. The company uses an Azure IoT Hub to communicate with the IoT devices.

The company must be able to monitor the devices in real-time.

You need to design the solution.

What should you recommend?

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

Correct Answer:

C. Azure Stream Analytics cloud job using Azure Portal

Question #31

HOTSPOT -

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

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

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

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

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Number of partitions:

| |
|----|
| 1 |
| 8 |
| 16 |
| 32 |

Partition key:

| |
|-----------------------|
| Fraud indicator |
| Fraud score |
| Individual line items |
| Payment details |
| Transaction ID |

Correct Answer:

Answer Area

Number of partitions:

| |
|----|
| 1 |
| 8 |
| 16 |
| 32 |

Partition key:

| |
|-----------------------|
| Fraud indicator |
| Fraud score |
| Individual line items |
| Payment details |
| Transaction ID |

Box 1: 16 -

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

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

Box 2: Transaction ID -

Reference:

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

Question #32

HOTSPOT -

You have an on-premises data warehouse that includes the following fact tables. Both tables have the following columns: DateKey, ProductKey, RegionKey.

There are 120 unique product keys and 65 unique region keys.

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

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

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

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

NOTE: Each correct selection is worth one point

Hot Area:

Answer Area

| Table | Distribution type | Distribution column |
|-----------|---|---|
| Sales: | <div style="border: 1px solid black; padding: 5px;"><div style="background-color: #f0f0f0; border-bottom: 1px solid black; padding-bottom: 5px;">Hash-distributed</div><div>Round-robin</div></div> | <div style="border: 1px solid black; padding: 5px;"><div style="background-color: #f0f0f0; border-bottom: 1px solid black; padding-bottom: 5px;">DateKey</div><div>ProductKey</div><div>RegionKey</div></div> |
| Invoices: | <div style="border: 1px solid black; padding: 5px;"><div style="background-color: #f0f0f0; border-bottom: 1px solid black; padding-bottom: 5px;">Hash-distributed</div><div>Round-robin</div></div> | <div style="border: 1px solid black; padding: 5px;"><div style="background-color: #f0f0f0; border-bottom: 1px solid black; padding-bottom: 5px;">DateKey</div><div>ProductKey</div><div>RegionKey</div></div> |

Correct Answer:

Answer Area

| Table | Distribution type | Distribution column |
|-----------|---|---|
| Sales: | <div style="border: 1px solid black; padding: 5px;"><div style="background-color: #90EE90; border-bottom: 1px solid black; padding-bottom: 5px;">Hash-distributed</div><div>Round-robin</div></div> | <div style="border: 1px solid black; padding: 5px;"><div style="background-color: #90EE90; border-bottom: 1px solid black; padding-bottom: 5px;">DateKey</div><div style="background-color: #90EE90; border-bottom: 1px solid black; padding-bottom: 5px;">ProductKey</div><div>RegionKey</div></div> |
| Invoices: | <div style="border: 1px solid black; padding: 5px;"><div style="background-color: #90EE90; border-bottom: 1px solid black; padding-bottom: 5px;">Hash-distributed</div><div>Round-robin</div></div> | <div style="border: 1px solid black; padding: 5px;"><div style="background-color: #90EE90; border-bottom: 1px solid black; padding-bottom: 5px;">DateKey</div><div>ProductKey</div><div style="background-color: #90EE90;">RegionKey</div></div> |

Box 1: Hash-distributed -

Box 2: ProductKey -

ProductKey is used extensively in joins.

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

Box 3: Hash-distributed -

Box 4: RegionKey -

Round-robin tables are useful for improving loading speed.

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

- ☞ When getting started as a simple starting point since it is the default
- ☞ If there is no obvious joining key
- ☞ If there is not good candidate column for hash distributing the table
- ☞ If the table does not share a common join key with other tables
- ☞ If the join is less significant than other joins in the query
- ☞ When the table is a temporary staging table

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

Reference:

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

Question #33

You have a partitioned table in an Azure Synapse Analytics dedicated SQL pool.

You need to design queries to maximize the benefits of partition elimination.

What should you include in the Transact-SQL queries?

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

Correct Answer:

B. WHERE

Question #34

You have an Azure Stream Analytics query. The query returns a result set that contains 10,000 distinct values for a column named clusterID.

You monitor the Stream Analytics job and discover high latency.

You need to reduce the latency.

Which two actions should you perform? Each correct answer presents a complete solution.

NOTE: Each correct selection is worth one point.

- A. Add a pass-through query.
- B. Increase the number of streaming units.
- C. Add a temporal analytic function.
- D. Scale out the query by using PARTITION BY.
- E. Convert the query to a reference query.

Correct Answer:

- B. Increase the number of streaming units.
- D. Scale out the query by using PARTITION BY.

Question #35

You have an Azure Synapse Analytics dedicated SQL pool named Pool1 and a database named DB1. DB1 contains a fact table named Table1.

You need to identify the extent of the data skew in Table1.

What should you do in Synapse Studio?

- A. Connect to the built-in pool and query sys.dmv_nodes_db_partition_stats.
- B. Connect to the built-in pool and run DBCC CHECKALLOC.
- C. Connect to Pool1 and query sys.dmv_node_status.
- D. Connect to Pool1 and query sys.dmv_nodes_db_partition_stats.

Correct Answer:

- D. Connect to Pool1 and query sys.dmv_nodes_db_partition_stats.

Question #36

You have an Azure Synapse Analytics dedicated SQL pool named Pool1. Pool1 contains a fact table named Table1.

You need to identify the extent of the data skew in Table1.

What should you do in Synapse Studio?

- A. Connect to Pool1 and DBCC PDW_SHOWSPACEUSED.
- B. Connect to the built-in pool and run DBCC PDW_SHOWSPACEUSED.
- C. Connect to the built-in pool and run DBCC CHECKALLOC.
- D. Connect to the built-in pool and query sys.dmv_sys_info.

Correct Answer:

- A. Connect to Pool1 and DBCC PDW_SHOWSPACEUSED.

Question #37

You use Azure Data Lake Storage Gen2.

You need to ensure that workloads can use filter predicates and column projections to filter data at the time the data is read from disk.

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

NOTE: Each correct selection is worth one point.

- A. Reregister the Azure Storage resource provider.
- B. Create a storage policy that is scoped to a container.
- C. Reregister the Microsoft Data Lake Store resource provider.
- D. Create a storage policy that is scoped to a container prefix filter.
- E. Register the query acceleration feature.

Correct Answer:

- D. Create a storage policy that is scoped to a container prefix filter.
- E. Register the query acceleration feature.

Question #38

You have an Azure Synapse Analytics dedicated SQL pool named Pool1. Pool1 contains a fact table named Table1.

You need to identify the extent of the data skew in Table1.

What should you do in Synapse Studio?

- A. Connect to Pool1 and run DBCC PDW_SHOWSPACEUSED.
- B. Connect to the built-in pool and run DBCC PDW_SHOWSPACEUSED.
- C. Connect to Pool1 and run DBCC CHECKALLOC.
- D. Connect to the built-in pool and query sys.dm_pdw_sys_info.

Correct Answer:

- A. Connect to Pool1 and run DBCC PDW_SHOWSPACEUSED.

Question #39

You have an Azure Data Lake Storage Gen2 account that contains two folders named Folder1 and Folder2.

You use Azure Data Factory to copy multiple files from Folder1 to Folder2.

You receive the following error.

Operation on target Copy_sks failed: Failure happened on 'Sink' side.

```
ErrorCode=DelimitedTextMoreColumnsThanDefined,  
'Type=Microsoft.DataTransfer.Common.Snared.HybridDeliveryException,  
Message=Error found when processing 'Csv/Tsv Format Text' source  
'0_2020_11_09_11_43_32.avro' with row number 53: found more columns than expected  
column count 27., Source=Microsoft.DataTransfer.Common,'
```

What should you do to resolve the error?

- A. Change the Copy activity setting to Binary Copy.
- B. Lower the degree of copy parallelism.
- C. Add an explicit mapping.
- D. Enable fault tolerance to skip incompatible rows.

Correct Answer:

- A. Change the Copy activity setting to Binary Copy.

Question #40

A company plans to use Apache Spark analytics to analyze intrusion detection data.

You need to recommend a solution to analyze network and system activity data for malicious activities and policy violations. The solution must minimize administrative efforts.

What should you recommend?

- A. Azure HDInsight
- B. Azure Data Factory
- C. Azure Data Lake Storage
- D. Azure Databricks

Correct Answer:

- D. Azure Databricks

Question #41

HOTSPOT

-

You have an Azure Synapse Analytics dedicated SQL pool.

You need to monitor the database for long-running queries and identify which queries are waiting on resources.

Which dynamic management view should you use for each requirement? To answer, select the appropriate options in the answer area.

NOTE: Each correct answer is worth one point.

Answer Area

Monitor the database for long-running queries:

| |
|--------------------------|
| ▼ |
| sys.dm_pdw_exec_requests |
| sys.dm_pdw_sql_requests |
| sys.dm_pdw_exec_sessions |

Identify which queries are waiting on resources:

| |
|---------------------------------------|
| ▼ |
| sys.dm_pdw_waits |
| sys.dm_pdw_lock_waits |
| sys.resource_governor_workload_groups |

Correct Answer:

Answer Area

Monitor the database for long-running queries:

| |
|--------------------------|
| ▼ |
| sys.dm_pdw_exec_requests |
| sys.dm_pdw_sql_requests |
| sys.dm_pdw_exec_sessions |

Identify which queries are waiting on resources:

| |
|---------------------------------------|
| ▼ |
| sys.dm_pdw_waits |
| sys.dm_pdw_lock_waits |
| sys.resource_governor_workload_groups |

Question #42

You have an Azure Data Factory pipeline named pipeline1 that includes a Copy activity named Copy1. Copy1 has the following configurations:

- The source of Copy1 is a table in an on-premises Microsoft SQL Server instance that is accessed by using a linked service connected via a self-hosted integration runtime.
- The sink of Copy1 uses a table in an Azure SQL database that is accessed by using a linked service connected via an Azure integration runtime.

You need to maximize the amount of compute resources available to Copy1. The solution must minimize administrative effort.

What should you do?

- A. Scale out the self-hosted integration runtime.
- B. Scale up the data flow runtime of the Azure integration runtime and scale out the self-hosted integration runtime.
- C. Scale up the data flow runtime of the Azure integration runtime.

Correct Answer:

- A. Scale out the self-hosted integration runtime.

Question #43

You are designing a solution that will use tables in Delta Lake on Azure Databricks.

You need to minimize how long it takes to perform the following:

- Queries against non-partitioned tables
- Joins on non-partitioned columns

Which two options should you include in the solution? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. the clone command
- B. Z-Ordering
- C. Apache Spark caching
- D. dynamic file pruning (DFP)

Correct Answer:

- B. Z-Ordering
- D. dynamic file pruning (DFP)

Question #44

You have an Azure Data Lake Storage Gen2 account named account1 that contains a container named container1.

You plan to create lifecycle management policy rules for container1.

You need to ensure that you can create rules that will move blobs between access tiers based on when each blob was accessed last.

What should you do first?

- A. Configure object replication
- B. Create an Azure application
- C. Enable access time tracking
- D. Enable the hierarchical namespace

Correct Answer:

- C. Enable access time tracking

Question #45

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

| Name | Description | Resource group |
|------|--|----------------|
| LA1 | Log Analytics workspace | RG1 |
| DB1 | Azure SQL database | RG2 |
| ADF1 | Azure Data Factory account | RG3 |
| Dw1 | Azure Synapse Analytics dedicated SQL pool | RG4 |

Diagnostic logs from ADF1 are sent to LA1. ADF1 contains a pipeline named Pipeline1 that copies data from DB1 to Dw1.

You need to perform the following actions:

- Create an action group named AG1.
- Configure an alert in ADF1 to use AG1.

In which resource group should you create AG1?

- A. RG1
- B. RG2
- C. RG3
- D. RG4

Correct Answer:

- C. RG3

Question #46

HOTSPOT

-

You have an Azure data factory named DF1 that contains 10 pipelines.

The pipelines are executed hourly by using a schedule trigger. All activities are executed on an Azure integration runtime.

You need to ensure that you can identify trends in queue times across the pipeline executions and activities. The solution must minimize administrative effort.

How should you configure the Diagnostic settings for DF1? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

Collect:

| | |
|----------------------------|---|
| Pipeline activity runs log | ▼ |
| Pipeline runs log | ▼ |
| Trigger runs log | ▼ |

Send to:

| | |
|-------------------------|---|
| Event hub | ▼ |
| Log Analytics workspace | ▼ |
| Storage account | ▼ |

Correct Answer:

Answer Area

Collect:

Pipeline activity runs log

Pipeline runs log

Trigger runs log

Send to:

Event hub

Log Analytics workspace

Storage account

Question #47

You manage an enterprise data warehouse in Azure Synapse Analytics.

Users report slow performance when they run commonly used queries. Users do not report performance changes for infrequently used queries.

You need to monitor resource utilization to determine the source of the performance issues.

Which metric should you monitor?

- A. DWU percentage
- B. Cache hit percentage
- C. Data Warehouse Units (DWU) used
- D. Data IO percentage

Correct Answer:

- B. Cache hit percentage

Question #48

HOTSPOT

-

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

| Name | Type | Description |
|-------|--|---------------------|
| ws1 | Azure Synapse Analytics workspace | None |
| kv1 | Azure Key Vault | None |
| UAMI1 | User-assigned managed identity | Associated with ws1 |
| sp1 | Apache Spark pool in Azure Synapse Analytics | Associated with ws1 |

You need to ensure that you can run Spark notebooks in ws1. The solution must ensure that you can retrieve secrets from kv1 by using UAMI1.

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

NOTE: Each correct selection is worth one point.

Answer Area

In the Azure portal:

- Add a role-based access control (RBAC) role to kv1.
- Add a role-based access control (RBAC) role to ws1.
- Create a linked service to kv1.

In Synapse Studio:

- Add a role-based access control (RBAC) role to kv1.
- Add a role-based access control (RBAC) role to ws1.
- Create a linked service to kv1.

Correct Answer:

Answer Area

In the Azure portal:

- Add a role-based access control (RBAC) role to kv1.
- Add a role-based access control (RBAC) role to ws1.
- Create a linked service to kv1.**

In Synapse Studio:

- Add a role-based access control (RBAC) role to kv1.
- Add a role-based access control (RBAC) role to ws1.**
- Create a linked service to kv1.

Question #49

HOTSPOT

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



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

Activity runs

Pipeline run ID 87f89922-14fa-468f-b13f-2f867606f4ff

All status

Showing 1 - 2 items

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

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

Activity runs

Pipeline run ID a7b5b522-cfaf-4c09-b3a9-fb42986be984

All status

Showing 1 - 3 items

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

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

NOTE: Each correct selection is worth one point.

Answer Area

| Statements | Yes | No |
|---|-----------------------|-----------------------|
| The <code>Retry</code> property of the <code>Web_GetIP</code> activity is set to 1. | <input type="radio"/> | <input type="radio"/> |
| The <code>waitForCompletion</code> property of the <code>Exec_COPY_BLOB</code> activity is set to <code>true</code> . | <input type="radio"/> | <input type="radio"/> |
| The <code>Exec_COPY_BLOB</code> activity was skipped during the second run due to pipeline dependencies. | <input type="radio"/> | <input type="radio"/> |

Correct Answer:**Answer Area**

| Statements | Yes | No |
|---|-------------------------------------|-------------------------------------|
| The <code>Retry</code> property of the <code>Web_GetIP</code> activity is set to 1. | <input checked="" type="checkbox"/> | <input type="radio"/> |
| The <code>waitForCompletion</code> property of the <code>Exec_COPY_BLOB</code> activity is set to <code>true</code> . | <input type="radio"/> | <input checked="" type="checkbox"/> |
| The <code>Exec_COPY_BLOB</code> activity was skipped during the second run due to pipeline dependencies. | <input checked="" type="checkbox"/> | <input type="radio"/> |

Question #50

You have an Azure Synapse Analytics dedicated SQL pool named Pool1. Pool1 contains a fact table named Table1.

You need to identify the extent of the data skew in Table1.

What should you do in Synapse Studio?

- A. Connect to the built-in pool and query `sys.dm_pdw_nodes_db_partition_stats`.
- B. Connect to Pool1 and run `DBCC PDW_SHOWSPACEUSED`.
- C. Connect to Pool1 and query `sys.dm_pdw_node_status`.
- D. Connect to the built-in pool and query `sys.dm_pdw_sys_info`.

Correct Answer:

B. Connect to Pool1 and run `DBCC PDW_SHOWSPACEUSED`.

Question #51

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

- Power Query
- Notebook
- Copy
- Jar

Which two Azure services should you use to debug the activities? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

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

Correct Answer:

- B. Azure Data Factory
- E. Azure Databricks

Question #52

A company purchases IoT devices to monitor manufacturing machinery. The company uses an Azure IoT Hub to communicate with the IoT devices.

The company must be able to monitor the devices in real-time.

You need to design the solution.

What should you recommend?

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

Correct Answer:

- D. Azure Stream Analytics cloud job using Azure Portal

Question #53

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

You need to perform a monthly audit of SQL statements that affect sensitive data. The solution must minimize administrative effort.

What should you include in the solution?

- A. workload management
- B. sensitivity labels
- C. dynamic data masking
- D. Microsoft Defender for SQL

Correct Answer:

- B. sensitivity labels

Question #54

A company purchases IoT devices to monitor manufacturing machinery. The company uses an Azure IoT Hub to communicate with the IoT devices.

The company must be able to monitor the devices in real-time.

You need to design the solution.

What should you recommend?

- A. Azure Analysis Services using Azure Portal
- B. Azure Stream Analytics Edge application using Microsoft Visual Studio
- C. Azure Analysis Services using Azure PowerShell
- D. Azure Analysis Services using Microsoft Visual Studio

Correct Answer:

- B. Azure Stream Analytics Edge application using Microsoft Visual Studio

Question #55

HOTSPOT

-

You have an Azure data factory.

You execute a pipeline that contains an activity named Activity1. Activity1 produces the following output.

```

{
  ...
  "dataRead": 1208,
  "dataWritten": 1208,
  "filesRead": 1,
  "filesWritten": 1,
  "sourcePeakConnections": 3,
  "sinkPeakConnections": 2,
  "copyDuration": 13,
  "throughput": 0.147,
  "effectiveIntegrationRuntime": "AutoResolveIntegrationRuntime (West Central US)",
  "usedDataIntegrationUnits": 4,
  "reportLineageToPurview": {
    "status": "Succeeded",
    "durationInSecond": "4"
  }
}
...
}

```

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 |
|--|---|------------|-----------|
| Activity1 is a Copy activity. | <input type="radio"/> <input type="radio"/> | | |
| Activity1 is executed by using a self-hosted integration runtime. | <input type="radio"/> <input type="radio"/> | | |
| The data factory that executed the pipeline is connected to Microsoft Purview. | <input type="radio"/> <input type="radio"/> | | |

Correct Answer:

| Answer Area | Statements | Yes | No |
|--|--|------------|-----------|
| Activity1 is a Copy activity. | <input checked="" type="checkbox"/> <input type="checkbox"/> | | |
| Activity1 is executed by using a self-hosted integration runtime. | <input type="radio"/> <input checked="" type="checkbox"/> | | |
| The data factory that executed the pipeline is connected to Microsoft Purview. | <input checked="" type="checkbox"/> <input type="radio"/> | | |

Question #56

HOTSPOT -

You have an Azure Data Lake Storage Gen2 account named account1 that contains a container named Container1. Container1 contains two folders named FolderA and FolderB.

You need to configure access control lists (ACLs) to meet the following requirements:

- Group1 must be able to list and read the contents and subfolders of FolderA.
- Group2 must be able to list and read the contents of FolderA and FolderB.
- Group2 must be prevented from reading any other folders at the root of Container1.

How should you configure the ACL permissions for each group? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

Group1:

- Set the access ACLs for FolderA to the Read + Execute permission.
- Set the access ACLs for FolderA to the Read permission.
- Set the default and access ACLs for FolderA to the Read + Execute permission.
- Set the default and access ACLs for the root of Container1 to the Read + Execute permission.

Group2:

- Set the access ACLs for the root of Container1 to the Read + Execute permission.
- Set the default ACLs for the root of Container1 to the Read + Execute permission.
- Set the default and access ACLs for FolderA and FolderB to the Read + Execute permission.

Correct Answer:

Answer Area

Group1:

- Set the access ACLs for FolderA to the Read + Execute permission**
- Set the access ACLs for FolderA to the Read permission.
- Set the default and access ACLs for FolderA to the Read + Execute permission.
- Set the default and access ACLs for the root of Container1 to the Read + Execute permission.

Group2:

- Set the access ACLs for the root of Container1 to the Read + Execute permission.
- Set the default ACLs for the root of Container1 to the Read + Execute permission**
- Set the default and access ACLs for FolderA and FolderB to the Read + Execute permission.**

Question #57

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

You need to ensure that User1 can review the Azure Synapse Analytics database templates from the gallery. The solution must follow the principle of least privilege.

Which role should you assign to User1?

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

Correct Answer:

- D. Synapse User

Question #58

You have a Log Analytics workspace named la1 and an Azure Synapse Analytics dedicated SQL pool named Pool1. Pool1 sends logs to la1.

You need to identify whether a recently executed query on Pool1 used the result set cache.

What are two ways to achieve the goal? Each correct answer presents a complete solution.

NOTE: Each correct selection is worth one point.

- A. Review the sys.dm_pdw_sql_requests dynamic management view in Pool1.
- B. Review the sys.dm_pdw_exec_requests dynamic management view in Pool1.
- C. Use the Monitor hub in Synapse Studio.
- D. Review the AzureDiagnostics table in la1.
- E. Review the sys.dm_pdw_request_steps dynamic management view in Pool1.

Correct Answer:

- B. Review the sys.dm_pdw_exec_requests dynamic management view in Pool1.
- C. Use the Monitor hub in Synapse Studio.

Question #59

HOTSPOT

-

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

You plan to implement row-level security (RLS) for Sales.Orders.

You need to create the security policy that will be used to implement RLS. The solution must ensure that sales representatives only see rows for which the value of the SalesRep column matches their username.

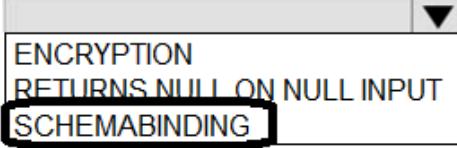
How should you complete the code? To answer, select the appropriate options in the answer area.

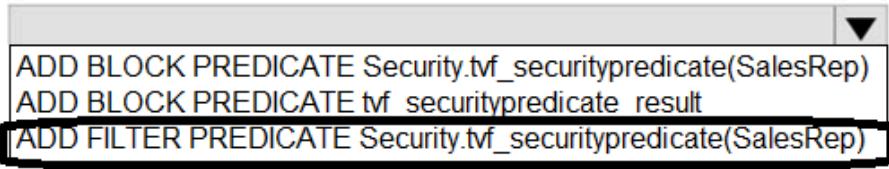
Answer Area

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

Correct Answer:

Answer Area

```
CREATE SCHEMA Security;
GO
CREATE FUNCTION Security.tvf_securitypredicate
(@SalesRep AS nvarchar(50))
RETURNS TABLE
WITH 
AS

    RETURN SELECT 1 AS tvf_securitypredicate_result
WHERE @SalesRep = USER_NAME();
GO
CREATE SECURITY POLICY SalesFilter

ON Sales.Orders
WITH (STATE = ON);
```

Question #60

You have an Azure data factory named DF1. DF1 contains a single pipeline that is executed by using a schedule trigger.

From Diagnostics settings, you configure pipeline runs to be sent to a resource-specific destination table in a Log Analytics workspace.

You need to run KQL queries against the table.

Which table should you query?

- A. ADFPipelineRun
- B. ADFTriggerRun
- C. ADFActivityRun
- D. AzureDiagnostics

Correct Answer:

- A. ADFPipelineRun

Question #61

HOTSPOT

You have an Azure Synapse Analytics dedicated SQL pool named sqlpool1 that contains a table named Sales1.

Each row in the Sales table contains regional sales data and a field that lists the username of a sales analyst.

You need to configure row-level security (RLS) to ensure that the analysts can view only the rows containing their respective data.

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

NOTE: Each correct selection is worth one point.

Answer Area

To configure RLS, create:

- A materialized view in sqlpool1
- A security policy in the Sales table
- Database scoped credentials in sqlpool1

To designate which rows each analyst can access, use:

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

Correct Answer:

Answer Area

To configure RLS, create:

- A materialized view in sqlpool1
- A security policy in the Sales table**
- Database scoped credentials in sqlpool1

To designate which rows each analyst can access, use:

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

Question #62

You have an Azure subscription that contains an Azure Synapse workspace named WS1 and an Azure Monitor action group named Group1. WS1 has a dedicated SQL pool.

You plan to archive monitoring data for integration activity runs.

You need to ensure that you can configure custom alerts based on the archived data that will execute Group1. The solution must minimize administrative effort.

Which diagnostic setting should you select?

- A. Send to Log Analytics workspace
- B. Archive to a storage account
- C. Stream to an event hub
- D. Send to a partner solution

Correct Answer:

A. Send to Log Analytics workspace

Question #63

You have an Azure subscription that contains an Azure Synapse Analytics dedicated SQL pool named Pool1.

You have the queries shown in the following table.

| Name | Users | Result set size |
|--------|-----------------------------------|-----------------|
| Query1 | Deterministic runtime expressions | 25 MB |
| Query2 | Deterministic built-in functions | 1 GB |
| Query3 | User-defined functions (UDFs) | 50 MB |
| Query4 | Row-level security (RLS) | 15 GB |

You are evaluating whether to enable result set caching for Pool1.

Which query results will be cached if result set caching is enabled?

- A. Query1 only
- B. Query2 only
- C. Query1 and Query2 only
- D. Query1 and Query3 only
- E. Query1, Query2, and Query3 only

Correct Answer:

C. Query1 and Query2 only

Question #64

You have an Azure subscription that contains an Azure Synapse Analytics workspace named workspace1, workspace1 contains an Azure Synapse Analytics dedicated SQL pool named Pool1.

You create a mapping data flow in an Azure Synapse pipeline that writes data to Pool1.

You execute the data flow and capture the execution information.

You need to identify how long it takes to write the data to Pool1.

Which metric should you use?

- A. the rows written
- B. the sink processing time
- C. the transformation processing time
- D. the post processing time

Correct Answer:

- B. the sink processing time

Question #65

You have an Azure data factory named DF1. DF1 contains a pipeline that has five activities.

You need to monitor queue times across the activities by using Log Analytics.

What should you do in DF1?

- A. Connect DF1 to a Microsoft Purview account.
- B. Add a diagnostic setting that sends activity runs to a Log Analytics workspace.
- C. Enable auto refresh for the Activity Logs Insights workbook.
- D. Add a diagnostic setting that sends pipeline runs to a Log Analytics workspace.

Correct Answer:

- B. Add a diagnostic setting that sends activity runs to a Log Analytics workspace.

Question #66

You have an Azure subscription that contains an Azure Synapse Analytics dedicated SQL pool named Pool1.

You need to monitor Pool1. The solution must ensure that you capture the start and end times of each query completed in Pool1.

Which diagnostic setting should you use?

- A. Sql Requests
- B. Request Steps
- C. Dms Workers
- D. Exec Requests

Correct Answer:

- D. Exec Requests

Question #67

You have an Azure Stream Analytics job named Job1.

The metrics of Job1 from the last hour are shown in the following table.

| Metric | Time aggregation | Value |
|------------------------------|------------------|-------|
| SU (Memory) % Utilization | Average | 70 |
| CPU % Utilization | Average | 20 |
| Runtime Errors | Total | 0 |
| Watermark Delay | Average | 20 |
| Input Deserialization Errors | Total | 0 |

The late arrival tolerance for Job1 is set to five seconds.

You need to optimize Job1.

Which two actions achieve the goal? Each correct answer presents a complete solution.

NOTE: Each correct answer is worth one point.

- A. Increase the number of SUs.
- B. Parallelize the query.
- C. Resolve errors in output processing.
- D. Resolve errors in input processing.

Correct Answer:

- A. Increase the number of SUs.
- B. Parallelize the query.

Question #68

You have an Azure subscription that contains an Azure data factory named ADF1 and a Log Analytics workspace named Workspace1.

You need to configure ADF1 to send execution information for pipelines to Workspace1.

What should you configure?

- A. diagnostic settings
- B. metrics
- C. logs
- D. alerts

Correct Answer:

- A. diagnostic settings

Question #69

You have an Azure Blob storage account named storage1 and an Azure Synapse Analytics serverless SQL pool named Pool1.

From Pool1, you plan to run ad-hoc queries that target storage1.

You need to ensure that you can use shared access signature (SAS) authorization without defining a data source.

What should you create first?

- A. a stored access policy
- B. a server-level credential
- C. a managed identity
- D. a database scoped credential

Correct Answer:

- D. a database scoped credential

Question #70

You have an Azure subscription that contains an Azure Synapse Analytics workspace named Workspace1, a Log Analytics workspace named Workspace2, and an Azure Data Lake Storage Gen2 container named Container1.

Workspace1 contains an Apache Spark job named Job1 that writes data to Container1.

Workspace1 sends diagnostics to Workspace2.

From Synapse Studio, you submit Job1.

What should you use to review the LogQuery output of the job?

- A. the files in the result subfolder of Container1
- B. the Spark monitoring URL returned after Job1 is submitted
- C. a table in Workspace2
- D. the Apache Spark applications option on the Monitor tab

Correct Answer:

- A. the files in the result subfolder of Container1

Question #71

You have an Azure subscription that contains an Azure Data Lake Storage Gen2 container named Container1 and an Azure Synapse Analytics workspace named Workspace1.

Workspace1 contains multiple Apache Spark jobs that reference a large dataset in Container1.

You need to optimize the run times of the jobs.

What should you do?

- A. For Container1, disable hierarchical namespaces.
- B. Cache the dataset.
- C. Increase the spark.sql.autoBroadcastJoinThreshold value.
- D. Use Resilient Distributed Datasets (RDDs).

Correct Answer:

- B. Cache the dataset.

Question #72

HOTSPOT

You have an Azure subscription that contains an Azure Synapse Analytics workspace named WS1. WS1 has a dedicated SQL pool and a query named Query1.

You execute Query1.

You need to identify whether the results of Query1 were retrieved from the cache or were computed.

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 request_id, command,  
      WHERE request_id = @RequestID
```

FROM

| |
|------------------|
| operation_type |
| result_cache_hit |
| status |

| |
|--------------------------|
| sys.dm_pdw_dms_workers |
| sys.dm_pdw_exec_requests |
| sys.dm_pdw_sql_requests |
| sys.dm_pdw_request_steps |

Correct Answer:

Answer Area

```
SELECT request_id, command,  
      WHERE request_id = @RequestID
```

FROM

| |
|------------------|
| operation_type |
| result_cache_hit |
| status |

| |
|--------------------------|
| sys.dm_pdw_dms_workers |
| sys.dm_pdw_exec_requests |
| sys.dm_pdw_sql_requests |
| sys.dm_pdw_request_steps |

Topic 5

Question #1

Introductory Info

Case study -

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

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

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

To start the case study -

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

Overview -

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

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

Existing Environment -

Transactional Data -

Contoso has three years of customer, transactional, operational, sourcing, and supplier data comprised of 10 billion records stored across multiple on-premises

Microsoft SQL Server servers. The SQL Server instances contain data from various operational systems. The data is loaded into the instances by using SQL Server Integration Services (SSIS) packages.

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

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

You plan to create a retail store table that will contain the address of each retail store. The table will be approximately 2 MB. Queries for retail store sales will include the retail store addresses. You plan to create a promotional table that will contain a promotion ID. The promotion ID will be associated to a specific product. The product will be identified by a product ID. The table will be approximately 200 GB.

Streaming Twitter Data -

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

Planned Changes and Requirements

Planned Changes -

Contoso plans to implement the following changes:

Load the sales transaction dataset to Azure Synapse Analytics.

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

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

Sales Transaction Dataset Requirements

Contoso identifies the following requirements for the sales transaction dataset:

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

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

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

Ensure that data storage costs and performance are predictable.

Minimize how long it takes to remove old records.

Customer Sentiment Analytics Requirements

Contoso identifies the following requirements for customer sentiment analytics:

Allow Contoso users to use PolyBase in an Azure Synapse Analytics dedicated SQL pool to query the content of the data records that host the Twitter feeds.

Data must be protected by using row-level security (RLS). The users must be authenticated by using their own Azure AD credentials.

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

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

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

Minimize administrative effort to maintain the Twitter feed data records.

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

Data Integration Requirements -

Contoso identifies the following requirements for data integration:

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

Analytics and transform the data.

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

Question

HOTSPOT -

You need to design a data storage structure for the product sales transactions. The solution must meet the sales transaction dataset requirements.

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Table type to store the product sales transactions:

| |
|-------------|
| Hash |
| Round-robin |
| Replicated |

When creating the table for sales transactions:

| |
|--|
| Configure a clustered index. |
| Set the distribution column to product ID. |
| Set the distribution column to the sales date. |

Correct Answer:

Answer Area

Table type to store the product sales transactions:

| |
|-------------|
| Hash |
| Round-robin |
| Replicated |

When creating the table for sales transactions:

| |
|--|
| Configure a clustered index. |
| Set the distribution column to product ID. |
| Set the distribution column to the sales date. |

Box 1: Hash -

Scenario:

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

A hash distributed table can deliver the highest query performance for joins and aggregations on large tables.

Box 2: Set the distribution column to the sales date.

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

Reference:

<https://rajanieshkaushikk.com/2020/09/09/how-to-choose-right-data-distribution-strategy-for-azure-synapse/>

Question #2

Introductory Info

Case study -

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

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

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

To start the case study -

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

Overview -

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

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

Existing Environment -

Transactional Data -

Contoso has three years of customer, transactional, operational, sourcing, and supplier data comprised of 10 billion records stored across multiple on-premises

Microsoft SQL Server servers. The SQL Server instances contain data from various operational systems. The data is loaded into the instances by using SQL

Server Integration Services (SSIS) packages.

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

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

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

You plan to create a promotional table that will contain a promotion ID. The promotion ID will be associated to a specific product. The product will be identified by a product ID. The table will be approximately 200 GB.

Streaming Twitter Data -

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

Planned Changes and Requirements

Planned Changes -

Contoso plans to implement the following changes:

Load the sales transaction dataset to Azure Synapse Analytics.

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

Sales Transaction Dataset Requirements

Contoso identifies the following requirements for the sales transaction dataset:

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

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

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

Ensure that data storage costs and performance are predictable.

Minimize how long it takes to remove old records.

Customer Sentiment Analytics Requirements

Contoso identifies the following requirements for customer sentiment analytics:

Allow Contoso users to use PolyBase in an Azure Synapse Analytics dedicated SQL pool to query the content of the data records that host the Twitter feeds.

Data must be protected by using row-level security (RLS). The users must be authenticated by using their own Azure AD credentials.

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

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

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

Minimize administrative effort to maintain the Twitter feed data records.

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

Data Integration Requirements -

Contoso identifies the following requirements for data integration:

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

Analytics and transform the data.

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

Question

DRAG DROP -

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

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

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

Select and Place:

| Commands | Answer Area |
|-----------------------------------|-------------|
| CREATE EXTERNAL DATA SOURCE | |
| CREATE EXTERNAL FILE FORMAT | |
| CREATE EXTERNAL TABLE | |
| CREATE EXTERNAL TABLE AS SELECT | |
| CREATE DATABASE SCOPED CREDENTIAL | |

Correct Answer:

| Commands | Answer Area |
|-----------------------------------|---------------------------------|
| CREATE EXTERNAL DATA SOURCE | CREATE EXTERNAL DATA SOURCE |
| CREATE EXTERNAL FILE FORMAT | CREATE EXTERNAL FILE FORMAT |
| CREATE EXTERNAL TABLE | CREATE EXTERNAL TABLE AS SELECT |
| CREATE EXTERNAL TABLE AS SELECT | |
| CREATE DATABASE SCOPED CREDENTIAL | |

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

Box 1: CREATE EXTERNAL DATA SOURCE

External data sources are used to connect to storage accounts.

Box 2: CREATE EXTERNAL FILE FORMAT

CREATE EXTERNAL FILE FORMAT creates an external file format object that defines external data stored in Azure Blob Storage or Azure Data Lake Storage.

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

Box 3: CREATE EXTERNAL TABLE AS SELECT

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

Incorrect Answers:

CREATE EXTERNAL TABLE -

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

Reference:

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

Question #3

Introductory Info

Case study -

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

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

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

To start the case study -

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

Overview -

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

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

Existing Environment -

Transactional Data -

Contoso has three years of customer, transactional, operational, sourcing, and supplier data comprised of 10 billion records stored across multiple on-premises

Microsoft SQL Server servers. The SQL Server instances contain data from various operational systems. The data is loaded into the instances by using SQL Server Integration Services (SSIS) packages.

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

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

You plan to create a retail store table that will contain the address of each retail store. The table will be approximately 2 MB. Queries for retail store sales will include the retail store addresses. You plan to create a promotional table that will contain a promotion ID. The promotion ID will be associated to a specific product. The product will be identified by a product ID. The table will be approximately 200 GB.

Streaming Twitter Data -

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

Planned Changes and Requirements

Planned Changes -

Contoso plans to implement the following changes:

Load the sales transaction dataset to Azure Synapse Analytics.

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

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

Sales Transaction Dataset Requirements

Contoso identifies the following requirements for the sales transaction dataset:

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

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

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

Ensure that data storage costs and performance are predictable.

Minimize how long it takes to remove old records.

Customer Sentiment Analytics Requirements

Contoso identifies the following requirements for customer sentiment analytics:

Allow Contoso users to use PolyBase in an Azure Synapse Analytics dedicated SQL pool to query the content of the data records that host the Twitter feeds.

Data must be protected by using row-level security (RLS). The users must be authenticated by using their own Azure AD credentials.

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

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

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

Minimize administrative effort to maintain the Twitter feed data records.

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

Data Integration Requirements -

Contoso identifies the following requirements for data integration:

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

Analytics and transform the data.

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

Question

HOTSPOT -

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

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Partition product sales
transactions data by:

| | |
|--------------|---|
| Sales date | ▼ |
| Product ID | ▼ |
| Promotion ID | ▼ |

Store product sales
transactions data in:

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

Correct Answer:

Answer Area

Partition product sales transactions data by:

| |
|--------------|
| Sales date |
| Product ID |
| Promotion ID |

Store product sales transactions data in:

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

Box 1: Sales date -

Scenario: Contoso requirements for data integration include:

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

Box 2: An Azure Synapse Analytics Dedicated SQL pool

Scenario: Contoso requirements for data integration include:

- ☞ Ensure that data storage costs and performance are predictable.

The size of a dedicated SQL pool (formerly SQL DW) is determined by Data Warehousing Units (DWU).

Dedicated SQL pool (formerly SQL DW) stores data in relational tables with columnar storage.

This format significantly reduces the data storage costs, and improves query performance.

Synapse analytics dedicated sql pool

Reference:

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

Question #4

Introductory Info

Case study -

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

To answer the questions included in a case study, you will need to reference information that is provided in the case study. Case studies might contain exhibits and other resources that provide

more information about the scenario that is described in the case study. Each question is independent of the other questions in this case study.

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

To start the case study -

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

Overview -

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

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

Existing Environment -

Transactional Data -

Contoso has three years of customer, transactional, operational, sourcing, and supplier data comprised of 10 billion records stored across multiple on-premises

Microsoft SQL Server servers. The SQL Server instances contain data from various operational systems. The data is loaded into the instances by using SQL Server Integration Services (SSIS) packages.

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

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

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

You plan to create a promotional table that will contain a promotion ID. The promotion ID will be associated to a specific product. The product will be identified by a product ID. The table will be approximately 200 GB.

Streaming Twitter Data -

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

Planned Changes and Requirements

Planned Changes -

Contoso plans to implement the following changes:

Load the sales transaction dataset to Azure Synapse Analytics.

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

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

Sales Transaction Dataset Requirements

Contoso identifies the following requirements for the sales transaction dataset:

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

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

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

Ensure that data storage costs and performance are predictable.

Minimize how long it takes to remove old records.

Customer Sentiment Analytics Requirements

Contoso identifies the following requirements for customer sentiment analytics:

Allow Contoso users to use PolyBase in an Azure Synapse Analytics dedicated SQL pool to query the content of the data records that host the Twitter feeds.

Data must be protected by using row-level security (RLS). The users must be authenticated by using their own Azure AD credentials.

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

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

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

Minimize administrative effort to maintain the Twitter feed data records.

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

Data Integration Requirements -

Contoso identifies the following requirements for data integration:

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

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

Question

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

What should you create?

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

Correct Answer:

- A. a table that has an IDENTITY property

Question #5

Introductory Info

Case study -

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

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

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

To start the case study -

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

Overview -

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

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

Existing Environment -

Transactional Data -

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

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

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

You plan to create a retail store table that will contain the address of each retail store. The table will be approximately 2 MB. Queries for retail store sales will include the retail store addresses. You plan to create a promotional table that will contain a promotion ID. The promotion ID will be associated to a specific product. The product will be identified by a product ID. The table will be approximately 200 GB.

Streaming Twitter Data -

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

Planned Changes and Requirements

Planned Changes -

Contoso plans to implement the following changes:

Load the sales transaction dataset to Azure Synapse Analytics.

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

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

Sales Transaction Dataset Requirements

Contoso identifies the following requirements for the sales transaction dataset:

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

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

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

Ensure that data storage costs and performance are predictable.

Minimize how long it takes to remove old records.

Customer Sentiment Analytics Requirements

Contoso identifies the following requirements for customer sentiment analytics:

Allow Contoso users to use PolyBase in an Azure Synapse Analytics dedicated SQL pool to query the content of the data records that host the Twitter feeds.

Data must be protected by using row-level security (RLS). The users must be authenticated by using their own Azure AD credentials.

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

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

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

Minimize administrative effort to maintain the Twitter feed data records.

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

Data Integration Requirements -

Contoso identifies the following requirements for data integration:

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

Analytics and transform the data.

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

Question

HOTSPOT -

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

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Table type to store retail store data:

| |
|-------------|
| Hash |
| Replicated |
| Round-robin |

Table type to store promotional data:

| |
|-------------|
| Hash |
| Replicated |
| Round-robin |

Correct Answer:

Answer Area

Table type to store retail store data:

| |
|-------------|
| Hash |
| Replicated |
| Round-robin |

Table type to store promotional data:

| |
|-------------|
| Hash |
| Replicated |
| Round-robin |

Box 1: Round-robin -

Round-robin tables are useful for improving loading speed.

Scenario: Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month.

Box 2: Hash -

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

Scenario:

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

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

Reference:

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

Question #6

Introductory Info

Case study -

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

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

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

To start the case study -

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

Overview -

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

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

Existing Environment -

Transactional Data -

Contoso has three years of customer, transactional, operational, sourcing, and supplier data comprised of 10 billion records stored across multiple on-premises

Microsoft SQL Server servers. The SQL Server instances contain data from various operational systems. The data is loaded into the instances by using SQL Server Integration Services (SSIS) packages.

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

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

You plan to create a retail store table that will contain the address of each retail store. The table will be approximately 2 MB. Queries for retail store sales will include the retail store addresses. You plan to create a promotional table that will contain a promotion ID. The promotion ID will be associated to a specific product. The product will be identified by a product ID. The table will be approximately 200 GB.

Streaming Twitter Data -

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

Planned Changes and Requirements

Planned Changes -

Contoso plans to implement the following changes:

Load the sales transaction dataset to Azure Synapse Analytics.

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

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

Sales Transaction Dataset Requirements

Contoso identifies the following requirements for the sales transaction dataset:

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

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

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

Ensure that data storage costs and performance are predictable.

Minimize how long it takes to remove old records.

Customer Sentiment Analytics Requirements

Contoso identifies the following requirements for customer sentiment analytics:

Allow Contoso users to use PolyBase in an Azure Synapse Analytics dedicated SQL pool to query the content of the data records that host the Twitter feeds.

Data must be protected by using row-level security (RLS). The users must be authenticated by using their own Azure AD credentials.

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

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

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

Minimize administrative effort to maintain the Twitter feed data records.

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

Data Integration Requirements -

Contoso identifies the following requirements for data integration:

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

Analytics and transform the data.

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

Question

HOTSPOT -

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

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Transact-SQL DDL command to use:

| |
|-----------------------|
| CREATE EXTERNAL TABLE |
| CREATE TABLE |
| CREATE VIEW |

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

| |
|------------------------|
| FORMAT_OPTIONS |
| FORMAT_TYPE |
| RANGE LEFT FOR VALUES |
| RANGE RIGHT FOR VALUES |

Correct Answer:

Answer Area

Transact-SQL DDL command to use:

| |
|-----------------------|
| CREATE EXTERNAL TABLE |
| CREATE TABLE |
| CREATE VIEW |

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

| |
|------------------------|
| FORMAT_OPTIONS |
| FORMAT_TYPE |
| RANGE LEFT FOR VALUES |
| RANGE RIGHT FOR VALUES |

Box 1: Create table -

Scenario: Load the sales transaction dataset to Azure Synapse Analytics

Box 2: RANGE RIGHT FOR VALUES -

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

RANGE RIGHT: Specifies the boundary value belongs to the partition on the right (higher values).

FOR VALUES (boundary_value [,...n]): Specifies the boundary values for the partition.

Scenario: Load the sales transaction dataset to Azure Synapse Analytics.

Contoso identifies the following requirements for the sales transaction dataset:

- ☞ Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month. Boundary values must belong to the partition on the right.
- ☞ Ensure that queries joining and filtering sales transaction records based on product ID complete as quickly as possible.
- ☞ Implement a surrogate key to account for changes to the retail store addresses.
- ☞ Ensure that data storage costs and performance are predictable.
- ☞ Minimize how long it takes to remove old records.

Reference:

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

Question #7

Introductory Info

Case study -

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

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

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

To start the case study -

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

Overview -

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

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

Existing Environment -

Transactional Data -

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

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

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

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

You plan to create a promotional table that will contain a promotion ID. The promotion ID will be associated to a specific product. The product will be identified by a product ID. The table will be approximately 200 GB.

Streaming Twitter Data -

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

Planned Changes and Requirements

Planned Changes -

Contoso plans to implement the following changes:

Load the sales transaction dataset to Azure Synapse Analytics.

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

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

Sales Transaction Dataset Requirements

Contoso identifies the following requirements for the sales transaction dataset:

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

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

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

Ensure that data storage costs and performance are predictable.

Minimize how long it takes to remove old records.

Customer Sentiment Analytics Requirements

Contoso identifies the following requirements for customer sentiment analytics:

Allow Contoso users to use PolyBase in an Azure Synapse Analytics dedicated SQL pool to query the content of the data records that host the Twitter feeds.

Data must be protected by using row-level security (RLS). The users must be authenticated by using their own Azure AD credentials.

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

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

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

Minimize administrative effort to maintain the Twitter feed data records.

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

Data Integration Requirements -

Contoso identifies the following requirements for data integration:

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

Analytics and transform the data.

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

Question

You need to design a data retention solution for the Twitter feed data records. The solution must meet the customer sentiment analytics requirements.

Which Azure Storage functionality should you include in the solution?

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

Correct Answer:

- D. lifecycle management

Topic 6

Question #1

Introductory Info

Case study -

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

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

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

To start the case study -

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

Overview -

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

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

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

Requirements -

Business Goals -

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

See inventory levels across the stores. Data must be updated as close to real time as possible. Execute ad hoc analytical queries on historical data to identify whether the loyalty club discounts increase sales of the discounted products.

Every four hours, notify store employees about how many prepared food items to produce based on historical demand from the sales data.

Technical Requirements -

Litware identifies the following technical requirements:

Minimize the number of different Azure services needed to achieve the business goals.

Use platform as a service (PaaS) offerings whenever possible and avoid having to provision virtual machines that must be managed by Litware.

Ensure that the analytical data store is accessible only to the company's on-premises network and Azure services.

Use Azure Active Directory (Azure AD) authentication whenever possible.

Use the principle of least privilege when designing security.

Stage Inventory data in Azure Data Lake Storage Gen2 before loading the data into the analytical data store. Litware wants to remove transient data from Data

Lake Storage once the data is no longer in use. Files that have a modified date that is older than 14 days must be removed.

Limit the business analysts' access to customer contact information, such as phone numbers, because this type of data is not analytically relevant.

Ensure that you can quickly restore a copy of the analytical data store within one hour in the event of corruption or accidental deletion.

Planned Environment -

Litware plans to implement the following environment:

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

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

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

Daily inventory data comes from a Microsoft SQL server located on a private network.

Litware currently has 5 TB of historical sales data and 100 GB of customer data. The company expects approximately 100 GB of new data per month for the next year.

Litware will build a custom application named FoodPrep to provide store employees with the calculation results of how many prepared food items to produce every four hours.

Litware does not plan to implement Azure ExpressRoute or a VPN between the on-premises network and Azure.

Question

HOTSPOT -

Which Azure Data Factory components should you recommend using together to import the daily inventory data from the SQL server to Azure Data Lake Storage?

To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Integration runtime type:

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

Trigger type:

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

Activity type:

| |
|---------------------------|
| Copy activity |
| Lookup activity |
| Stored procedure activity |

Correct Answer:

Answer Area

Integration runtime type:

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

Trigger type:

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

Activity type:

| |
|---------------------------|
| Copy activity |
| Lookup activity |
| Stored procedure activity |

Box 1: Self-hosted integration runtime

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

Box 2: Schedule trigger -

Schedule every 8 hours -

Box 3: Copy activity -

Scenario:

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

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

Topic 7

Question #1

Introductory Info

Case study -

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

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

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

To start the case study -

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

Overview -

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

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

Existing Environment -

Transactional Data -

Contoso has three years of customer, transactional, operational, sourcing, and supplier data comprised of 10 billion records stored across multiple on-premises

Microsoft SQL Server servers. The SQL Server instances contain data from various operational systems. The data is loaded into the instances by using SQL Server Integration Services (SSIS) packages.

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

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

You plan to create a retail store table that will contain the address of each retail store. The table will be approximately 2 MB. Queries for retail store sales will include the retail store addresses. You plan to create a promotional table that will contain a promotion ID. The promotion ID will be associated to a specific product. The product will be identified by a product ID. The table will be approximately 200 GB.

Streaming Twitter Data -

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

Planned Changes and Requirements

Planned Changes -

Contoso plans to implement the following changes:

Load the sales transaction dataset to Azure Synapse Analytics.

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

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

Sales Transaction Dataset Requirements

Contoso identifies the following requirements for the sales transaction dataset:

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

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

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

Ensure that data storage costs and performance are predictable.

Minimize how long it takes to remove old records.

Customer Sentiment Analytics Requirements

Contoso identifies the following requirements for customer sentiment analytics:

Allow Contoso users to use PolyBase in an Azure Synapse Analytics dedicated SQL pool to query the content of the data records that host the Twitter feeds.

Data must be protected by using row-level security (RLS). The users must be authenticated by using their own Azure AD credentials.

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

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

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

Minimize administrative effort to maintain the Twitter feed data records.

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

Data Integration Requirements -

Contoso identifies the following requirements for data integration:

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

Analytics and transform the data.

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

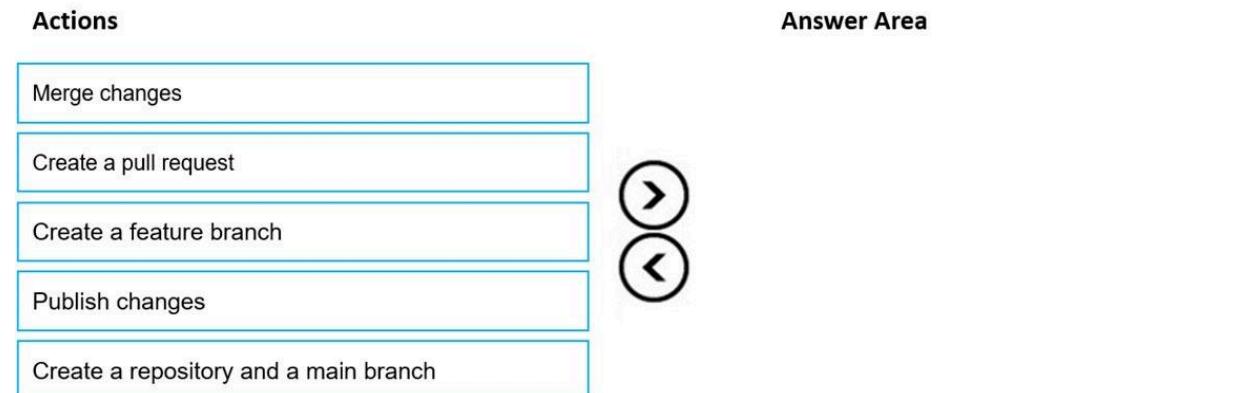
Question

DRAG DROP -

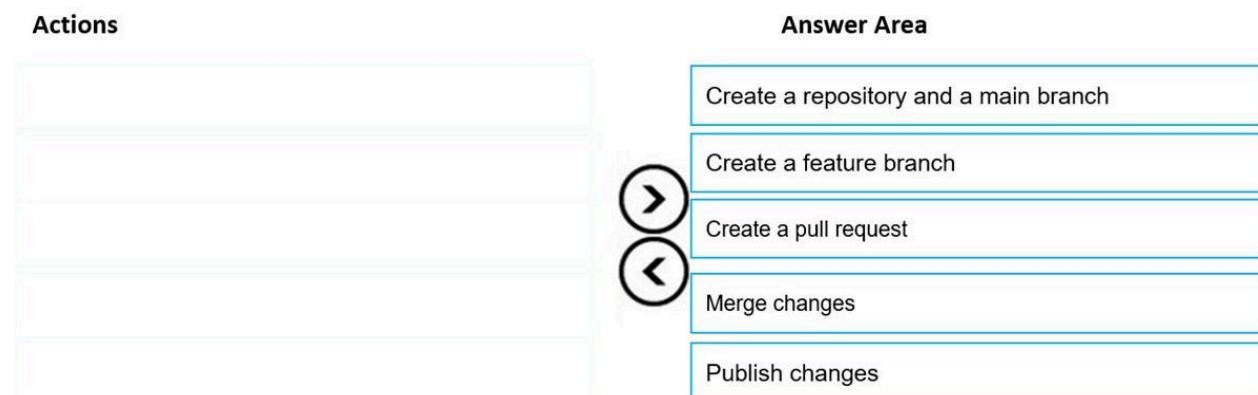
You need to implement versioned changes to the integration pipelines. The solution must meet the data integration requirements.

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

Select and Place:



Correct Answer:



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

Step 1: Create a repository and a main branch

You need a Git repository in Azure Pipelines, TFS, or GitHub with your app.

Step 2: Create a feature branch -

Step 3: Create a pull request -

Step 4: Merge changes -

Merge feature branches into the main branch using pull requests.

Step 5: Publish changes -

Reference:

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

Topic 8

Question #1

Introductory Info

Case study -

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

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

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

To start the case study -

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

Overview -

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

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

Existing Environment -

Transactional Data -

Contoso has three years of customer, transactional, operational, sourcing, and supplier data comprised of 10 billion records stored across multiple on-premises

Microsoft SQL Server servers. The SQL Server instances contain data from various operational systems. The data is loaded into the instances by using SQL Server Integration Services (SSIS) packages.

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

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

You plan to create a retail store table that will contain the address of each retail store. The table will be approximately 2 MB. Queries for retail store sales will include the retail store addresses. You plan to create a promotional table that will contain a promotion ID. The promotion ID will be associated to a specific product. The product will be identified by a product ID. The table will be approximately 200 GB.

Streaming Twitter Data -

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

Planned Changes and Requirements

Planned Changes -

Contoso plans to implement the following changes:

Load the sales transaction dataset to Azure Synapse Analytics.

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

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

Sales Transaction Dataset Requirements

Contoso identifies the following requirements for the sales transaction dataset:

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

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

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

Ensure that data storage costs and performance are predictable.

Minimize how long it takes to remove old records.

Customer Sentiment Analytics Requirements

Contoso identifies the following requirements for customer sentiment analytics:

Allow Contoso users to use PolyBase in an Azure Synapse Analytics dedicated SQL pool to query the content of the data records that host the Twitter feeds.

Data must be protected by using row-level security (RLS). The users must be authenticated by using their own Azure AD credentials.

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

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

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

Minimize administrative effort to maintain the Twitter feed data records.

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

Data Integration Requirements -

Contoso identifies the following requirements for data integration:

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

Analytics and transform the data.

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

Question

HOTSPOT -

You need to design a data ingestion and storage solution for the Twitter feeds. The solution must meet the customer sentiment analytics requirements.

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

To increase the throughput of ingesting the Twitter feeds:

| |
|------------------------------------|
| Configure Event Hubs partitions. |
| Enable Auto-Inflate in Event Hubs. |
| Use Event Hubs Dedicated. |

To store the Twitter feed data, use:

| |
|---|
| An Azure Data Lake Storage Gen2 account |
| An Azure Databricks high concurrency cluster |
| An Azure General-purpose v2 storage account in the Premium tier |

Correct Answer:

Answer Area

To increase the throughput of ingesting the Twitter feeds:

| |
|------------------------------------|
| Configure Event Hubs partitions. |
| Enable Auto-Inflate in Event Hubs. |
| Use Event Hubs Dedicated. |

To store the Twitter feed data, use:

| |
|---|
| An Azure Data Lake Storage Gen2 account |
| An Azure Databricks high concurrency cluster |
| An Azure General-purpose v2 storage account in the Premium tier |

Box 1: Configure Event Hubs partitions

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

Event Hubs is designed to help with processing of large volumes of events. Event Hubs throughput is scaled by using partitions and throughput-unit allocations.

Incorrect Answers:

☞ Event Hubs Dedicated: Event Hubs clusters offer single-tenant deployments for customers with the most demanding streaming needs. This single-tenant offering has a guaranteed 99.99% SLA and is available only on our Dedicated pricing tier.

☞ Auto-Inflate: The Auto-inflate feature of Event Hubs automatically scales up by increasing the number of TUs, to meet usage needs.

Event Hubs traffic is controlled by TUs (standard tier). Auto-inflate enables you to start small with the minimum required TUs you choose. The feature then scales automatically to the maximum limit of TUs you need, depending on the increase in your traffic.

Box 2: An Azure Data Lake Storage Gen2 account

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

Azure Data Lake Storage Gen2 implements an access control model that supports both Azure role-based access control (Azure RBAC) and POSIX-like access control lists (ACLs).

Incorrect Answers:

☞ Azure Databricks: An Azure administrator with the proper permissions can configure Azure Active Directory conditional access to control where and when users are permitted to sign in to Azure Databricks.

☞ Azure Storage supports using Azure Active Directory (Azure AD) to authorize requests to blob data.

You can scope access to Azure blob resources at the following levels, beginning with the narrowest scope:

- An individual container. At this scope, a role assignment applies to all of the blobs in the container, as well as container properties and metadata.
- The storage account. At this scope, a role assignment applies to all containers and their blobs.
- The resource group. At this scope, a role assignment applies to all of the containers in all of the storage accounts in the resource group.
- The subscription. At this scope, a role assignment applies to all of the containers in all of the storage accounts in all of the resource groups in the subscription.
- A management group.

Reference:

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

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

Topic 9

Question #1

Introductory Info

Case study -

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

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

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

To start the case study -

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

Overview -

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

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

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

Requirements -

Business Goals -

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

See inventory levels across the stores. Data must be updated as close to real time as possible. Execute ad hoc analytical queries on historical data to identify whether the loyalty club discounts increase sales of the discounted products.

Every four hours, notify store employees about how many prepared food items to produce based on historical demand from the sales data.

Technical Requirements -

Litware identifies the following technical requirements:

Minimize the number of different Azure services needed to achieve the business goals.

Use platform as a service (PaaS) offerings whenever possible and avoid having to provision virtual machines that must be managed by Litware.

Ensure that the analytical data store is accessible only to the company's on-premises network and Azure services.

Use Azure Active Directory (Azure AD) authentication whenever possible.

Use the principle of least privilege when designing security.

Stage Inventory data in Azure Data Lake Storage Gen2 before loading the data into the analytical data store. Litware wants to remove transient data from Data

Lake Storage once the data is no longer in use. Files that have a modified date that is older than 14 days must be removed.

Limit the business analysts' access to customer contact information, such as phone numbers, because this type of data is not analytically relevant.

Ensure that you can quickly restore a copy of the analytical data store within one hour in the event of corruption or accidental deletion.

Planned Environment -

Litware plans to implement the following environment:

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

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

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

Daily inventory data comes from a Microsoft SQL server located on a private network.

Litware currently has 5 TB of historical sales data and 100 GB of customer data. The company expects approximately 100 GB of new data per month for the next year.

Litware will build a custom application named FoodPrep to provide store employees with the calculation results of how many prepared food items to produce every four hours.

Litware does not plan to implement Azure ExpressRoute or a VPN between the on-premises network and Azure.

Question

What should you recommend to prevent users outside the Litware on-premises network from accessing the analytical data store?

- A. a server-level virtual network rule
- B. a database-level virtual network rule
- C. a server-level firewall IP rule
- D. a database-level firewall IP rule

Correct Answer:

- C. a server-level firewall IP rule

Question #2

Introductory Info

Case study -

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

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

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

To start the case study -

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

Overview -

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

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

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

Requirements -

Business Goals -

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

See inventory levels across the stores. Data must be updated as close to real time as possible. Execute ad hoc analytical queries on historical data to identify whether the loyalty club discounts increase sales of the discounted products.

Every four hours, notify store employees about how many prepared food items to produce based on historical demand from the sales data.

Technical Requirements -

Litware identifies the following technical requirements:

Minimize the number of different Azure services needed to achieve the business goals.

Use platform as a service (PaaS) offerings whenever possible and avoid having to provision virtual machines that must be managed by Litware.

Ensure that the analytical data store is accessible only to the company's on-premises network and Azure services.

Use Azure Active Directory (Azure AD) authentication whenever possible.

Use the principle of least privilege when designing security.

Stage Inventory data in Azure Data Lake Storage Gen2 before loading the data into the analytical data store. Litware wants to remove transient data from Data

Lake Storage once the data is no longer in use. Files that have a modified date that is older than 14 days must be removed.

Limit the business analysts' access to customer contact information, such as phone numbers, because this type of data is not analytically relevant.

Ensure that you can quickly restore a copy of the analytical data store within one hour in the event of corruption or accidental deletion.

Planned Environment -

Litware plans to implement the following environment:

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

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

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

Daily inventory data comes from a Microsoft SQL server located on a private network.

Litware currently has 5 TB of historical sales data and 100 GB of customer data. The company expects approximately 100 GB of new data per month for the next year.

Litware will build a custom application named FoodPrep to provide store employees with the calculation results of how many prepared food items to produce every four hours.

Litware does not plan to implement Azure ExpressRoute or a VPN between the on-premises network and Azure.

Question

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

- A. Transparent Data Encryption (TDE)
- B. row-level security
- C. column-level security
- D. data sensitivity labels

Correct Answer:

- C. column-level security

Topic 10

Question #1

Introductory Info

Case study -

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

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

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

To start the case study -

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

Overview -

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

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

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

Requirements -

Business Goals -

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

See inventory levels across the stores. Data must be updated as close to real time as possible. Execute ad hoc analytical queries on historical data to identify whether the loyalty club discounts increase sales of the discounted products.

Every four hours, notify store employees about how many prepared food items to produce based on historical demand from the sales data.

Technical Requirements -

Litware identifies the following technical requirements:

Minimize the number of different Azure services needed to achieve the business goals.

Use platform as a service (PaaS) offerings whenever possible and avoid having to provision virtual machines that must be managed by Litware.

Ensure that the analytical data store is accessible only to the company's on-premises network and Azure services.

Use Azure Active Directory (Azure AD) authentication whenever possible.

Use the principle of least privilege when designing security.

Stage Inventory data in Azure Data Lake Storage Gen2 before loading the data into the analytical data store. Litware wants to remove transient data from Data

Lake Storage once the data is no longer in use. Files that have a modified date that is older than 14 days must be removed.

Limit the business analysts' access to customer contact information, such as phone numbers, because this type of data is not analytically relevant.

Ensure that you can quickly restore a copy of the analytical data store within one hour in the event of corruption or accidental deletion.

Planned Environment -

Litware plans to implement the following environment:

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

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

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

Daily inventory data comes from a Microsoft SQL server located on a private network.

Litware currently has 5 TB of historical sales data and 100 GB of customer data. The company expects approximately 100 GB of new data per month for the next year.

Litware will build a custom application named FoodPrep to provide store employees with the calculation results of how many prepared food items to produce every four hours.

Litware does not plan to implement Azure ExpressRoute or a VPN between the on-premises network and Azure.

Question

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

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

Correct Answer:

- D. Deploy identical Azure Stream Analytics jobs to paired regions in Azure.