DATAWOLFS.COM -- DP203 DUMPS

[Document subtitle]

Question #1*Topic 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

Hide Solution Discussion 1
Correct Answer: *C*
We need an extra column to identify the Manager. Use the data type as the EmployeeKey column, an int column.
Reference:
https://docs.microsoft.com/en-us/analysis-services/tabular-models/hierarchies-ssas-tabular

Question #2*Topic 1*
You have an Azure Synapse workspace named MyWorkspace that contains an Apache Spark database named mytestdb.
You run the following command in an Azure Synapse Analytics Spark pool in MyWorkspace.

CREATE TABLE mytestdb.myParquetTable(
EmployeeID int,
EmployeeName string,
EmployeeStartDate date)

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

EmployeeName	EmployeeID	EmployeeStartDate
Alice	24	2020-01-25

One minute later, you execute the following query from a serverless SQL pool in MyWorkspace.

SELECT EmployeeID -
FROM mytestdb.dbo.myParquetTable
WHERE name = 'Alice';
What will be returned by the query?

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

Hide Solution | Discussion 28

Correct Answer: *A*
Once a database has been created by a Spark job, you can create tables in it with Spark that use Parquet as the storage format. Table names will be converted to lower case and need to be queried using the lower case name. These tables will immediately become available for querying by any of the Azure Synapse workspace Spark pools. They can also be used from any of the Spark jobs subject to permissions.
Note: For external tables, since they are synchronized to serverless SQL pool asynchronously, there will be a delay until they appear.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/metadata/table

Question #3*Topic 1*
DRAG DROP -
You have a table named SalesFact in an enterprise data warehouse in Azure Synapse Analytics. SalesFact contains sales data from the past 36 months and has the following characteristics:

- Is partitioned by month
- Contains one billion rows
- Has clustered columnstore indexes

At the beginning of each month, you need to remove data from SalesFact that is older

than 36 months as quickly as possible.
Which three actions should you perform in sequence in a stored procedure? To answer,move the appropriate actions from the list of actions to the answer area and arrangethem 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).	

Hide Solution | Discussion 15

Correct Answer:

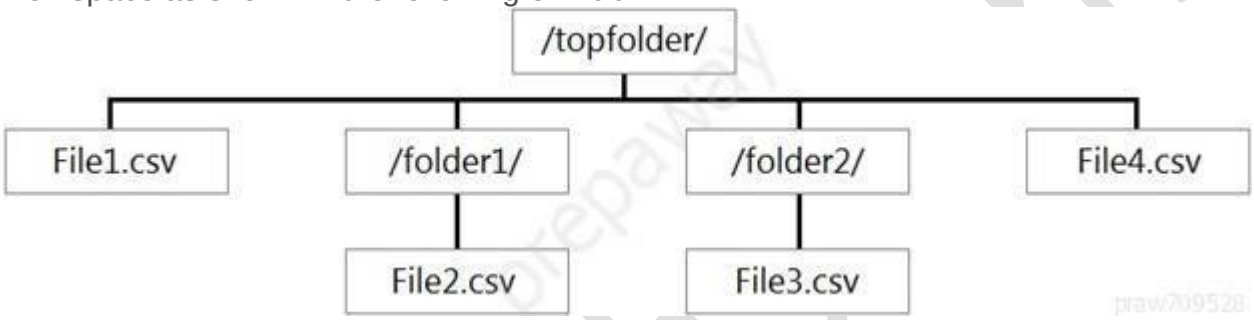
Actions	Answer Area
Switch the partition containing the stale data from SalesFact to SalesFact_Work.	Create an empty table named SalesFact_Work that has the same schema as SalesFact.
Truncate the partition containing the stale data.	Switch the partition containing the stale data from SalesFact to SalesFact_Work.
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*Topic 1*
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

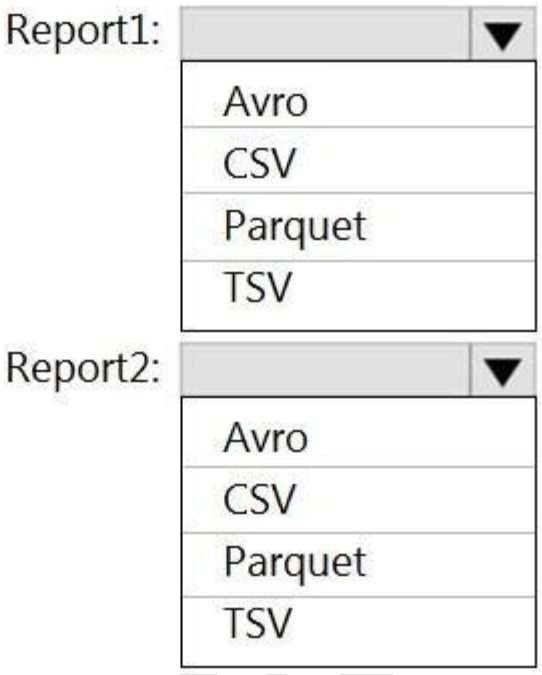
Hide Solution | Discussion 26
Correct Answer: *C*
To run a T-SQL query over a set of files within a folder or set of folders while treating them as a single entity or rowset, provide a path to a folder or a pattern (using wildcards) over a set of files or folders.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/query-data-storage#query-multiple-files-or-folders

Question #5*Topic 1*
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.

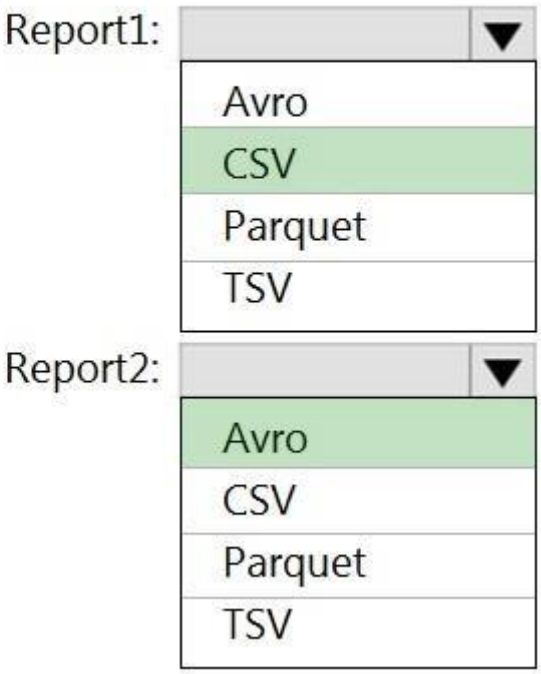
NOTE: Each correct selection is worth one point.
Hot Area:

Answer Area



Hide Solution Discussion 15

Answer Area



Correct Answer:
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*Topic 1*
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}_{D

D}.csv

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

Hide Solution Discussion 5

Correct Answer: *D*

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

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

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

Question #7*Topic 1*

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.

Answer Area

Hot Area:

Hide Solution Discussion 2

Correct

Answer:

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

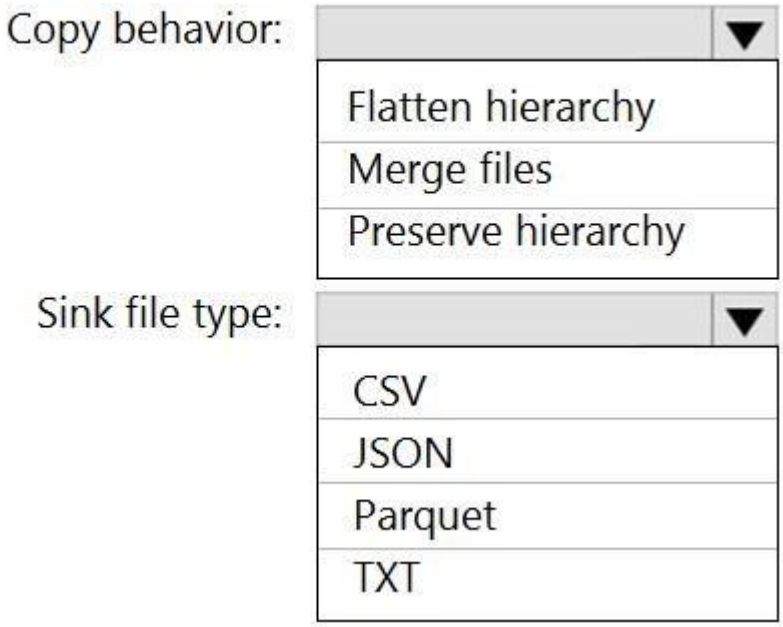
Question #8*Topic 1*
HOTSPOT -
You use Azure Data Factory to prepare data to be queried by Azure Synapse Analytics serverless SQL pools.
Files are initially ingested into an Azure Data Lake Storage Gen2 account as 10 small JSON files. Each file contains the same data attributes and data from a subsidiary 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.

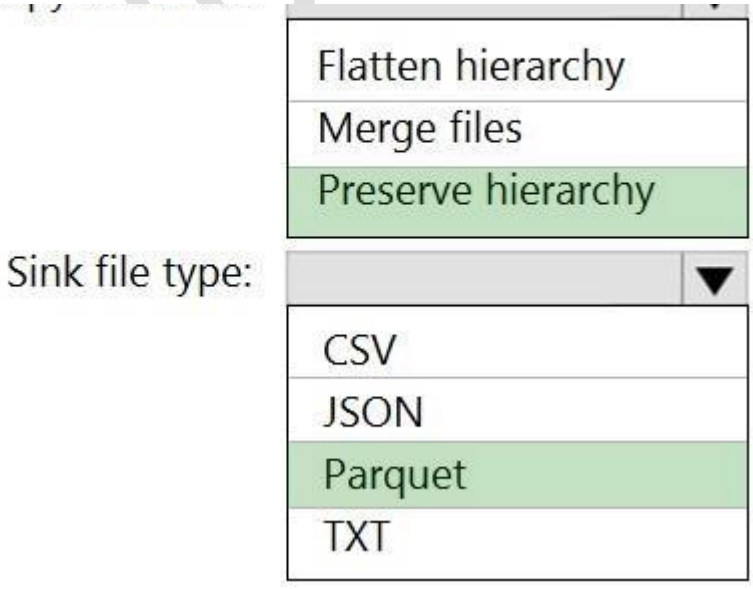
Hot Area:

Answer Area



Hide Solution | Discussion 15

Correct



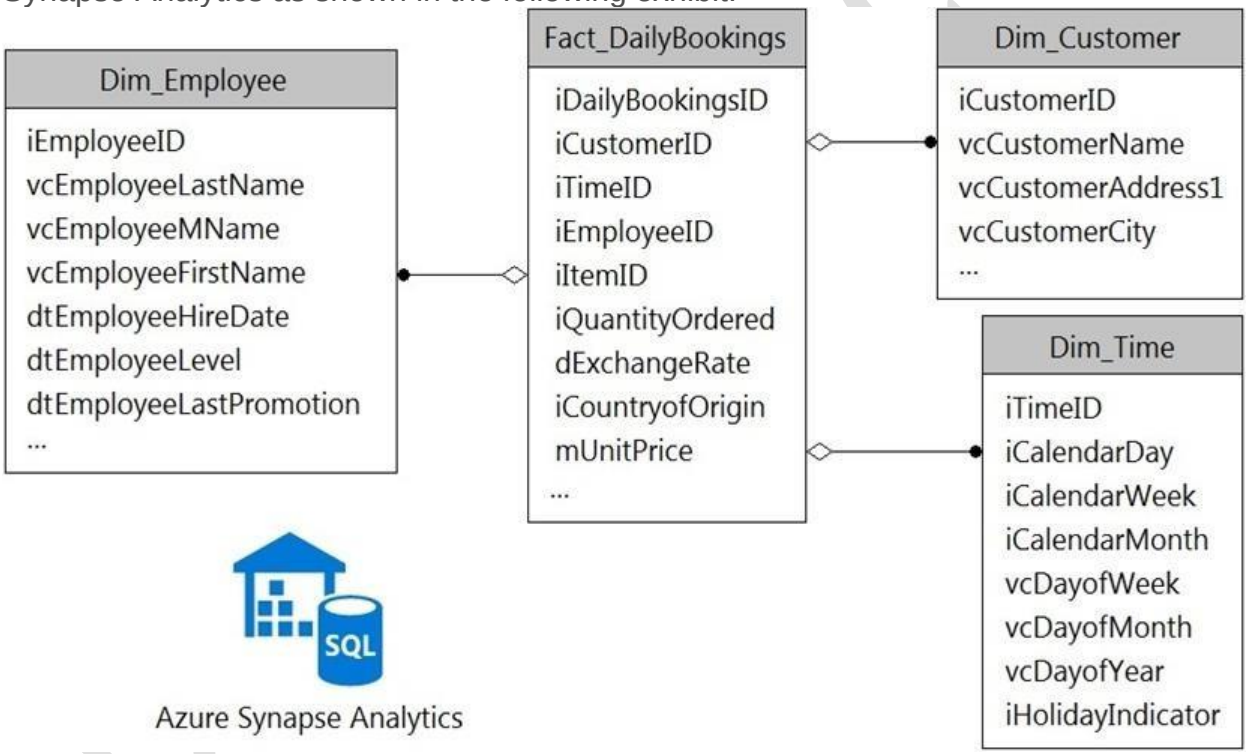
Answer:

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

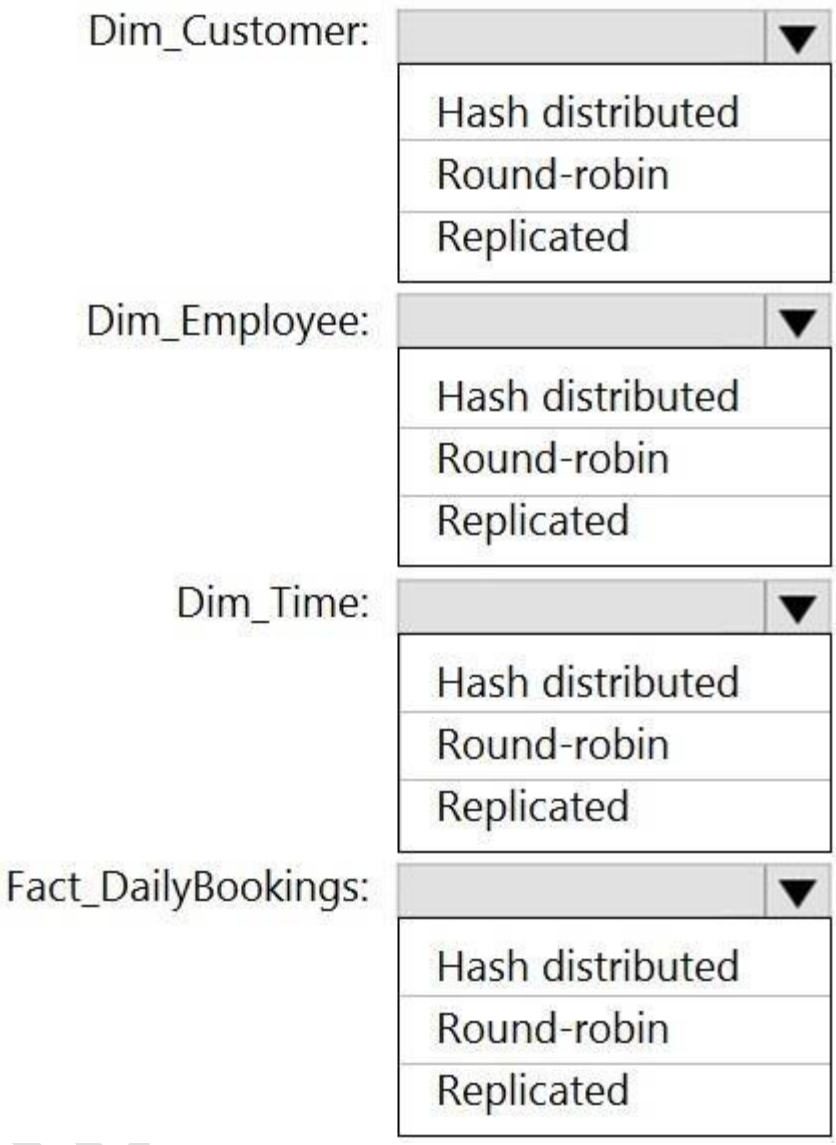
Question #9*Topic 1*
HOTSPOT -
You have a data model that you plan to implement in a data warehouse in Azure Synapse Analytics as shown in the following exhibit.



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:

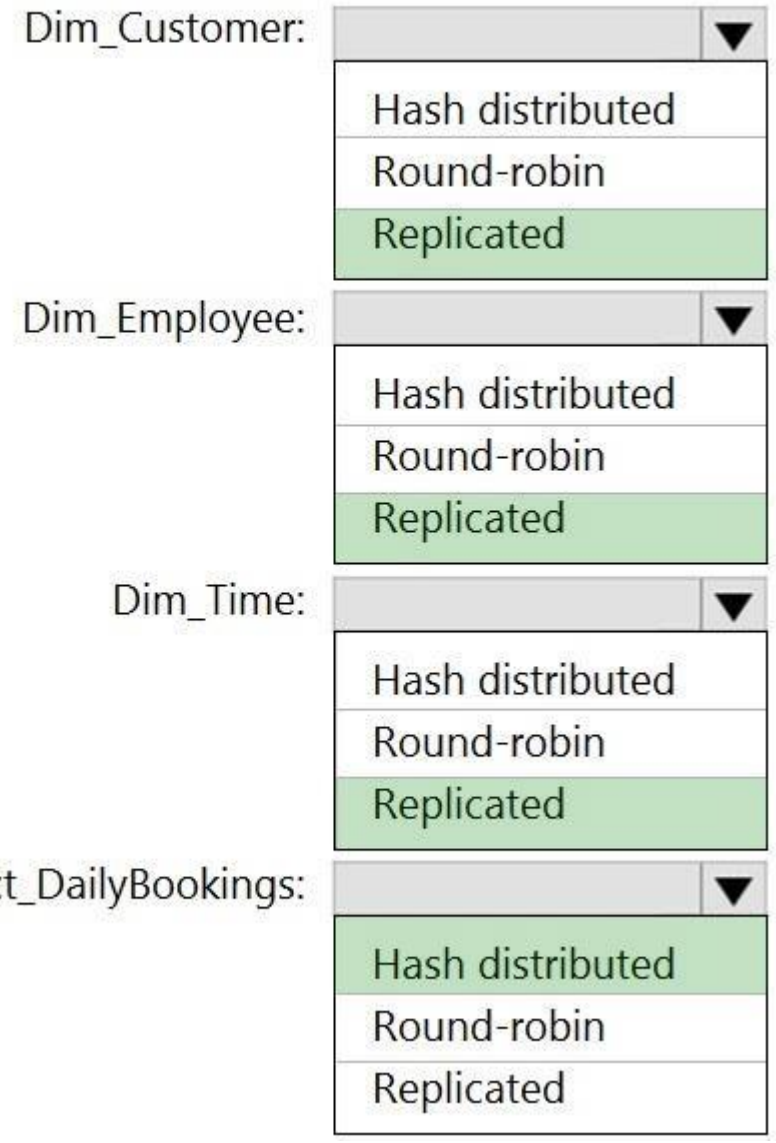
Answer Area



Hide Solution Discussion 7

Correct

Answer Area



Answer:
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*Topic 1*
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.

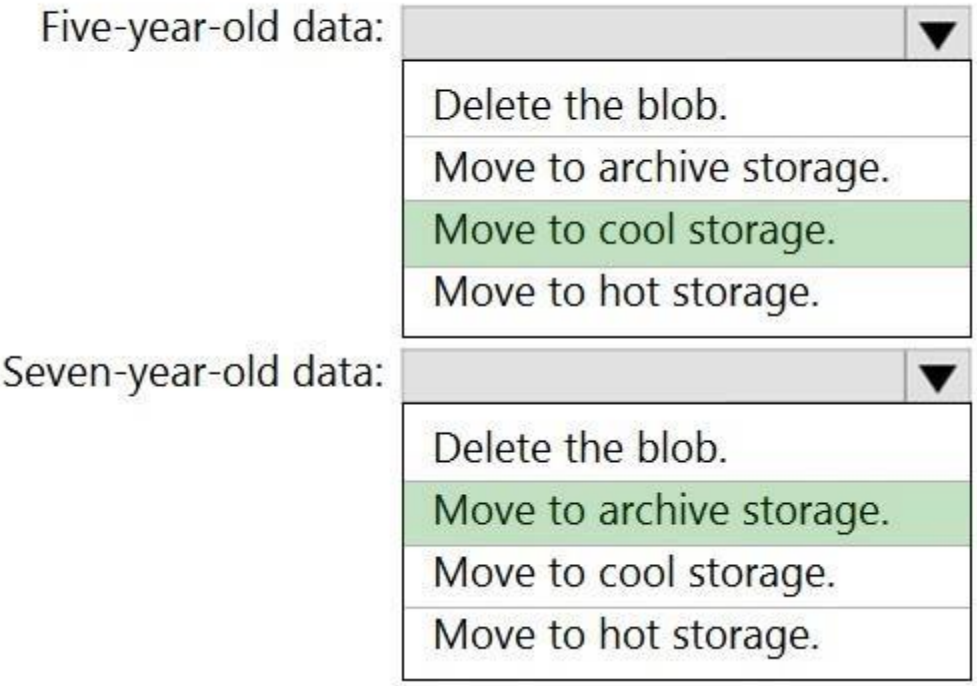
Seven-year-old data:

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

Hide Solution Discussion 8

Correct Answer:

Answer Area



Box 1: Move to cool storage -

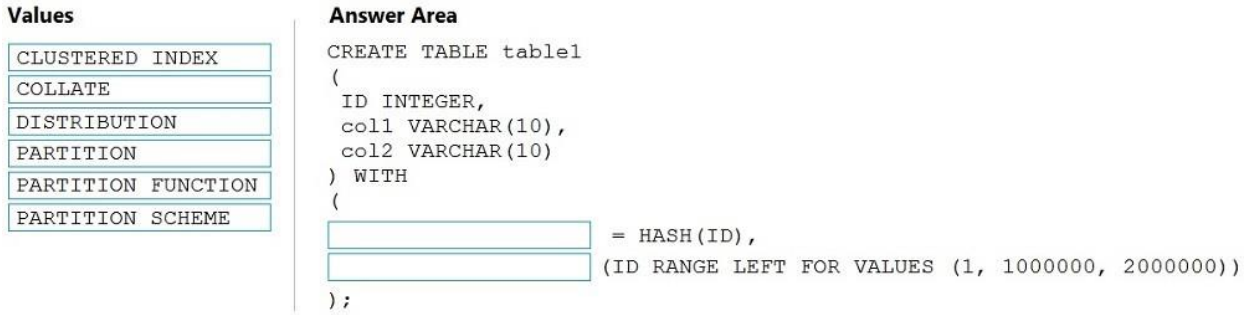
Box 2: Move to archive storage -
Archive - Optimized for storing data that is rarely accessed and stored for at least 180 days with flexible latency requirements, on the order of hours.
The following table shows a comparison of premium performance block blob storage, and the hot, cool, and archive access 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

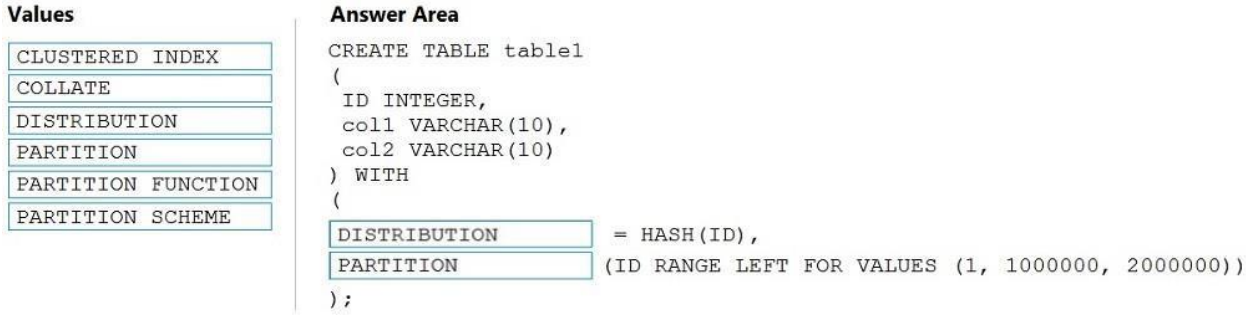
Question #11*Topic 1*
DRAG DROP -
You need to create a partitioned table in an Azure Synapse Analytics dedicated SQL pool.
How should you complete the Transact-SQL statement? To answer, drag the appropriate values to the correct targets. Each value may be used once, more than once, or not at all. You 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:



Hide Solution Discussion 4

Correct Answer:



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*Topic 1*
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 **Most Voted**

Hide Solution Discussion 13

Correct Answer: *D*

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

Reference:

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

Question #13*Topic 1*

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. **Most Voted**
- B. Use the managed identity as the credentials for the data load process. **Most Voted**
- 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. **Most Voted**

Hide Solution Discussion 19

Correct Answer: *ADF*

The managed identity grants permissions to the dedicated SQL pools in the workspace.
Note: Managed identity for Azure resources is a feature of Azure Active Directory. The feature provides Azure services with an automatically managed identity in

Azure AD -
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/security/synapse-workspace-managed-identitY

Question #14*Topic 1*
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.

```
SELECT c.name,
    tbl.name as table_name,
    typ.name as datatype,
    c.is_masked,
    c.masking_function
FROM sys.masked_columns AS c
INNER JOIN sys.tables AS tbl ON c.[object_id] = tbl.[object_id]
INNER JOIN sys.types typ ON c.user_type_id = typ.user_type_id
WHERE is_masked = 1;
```

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

Hide Solution | Discussion 10

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*Topic 1*
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;
```

Hide Solution Discussion 2

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*Topic 1*

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:

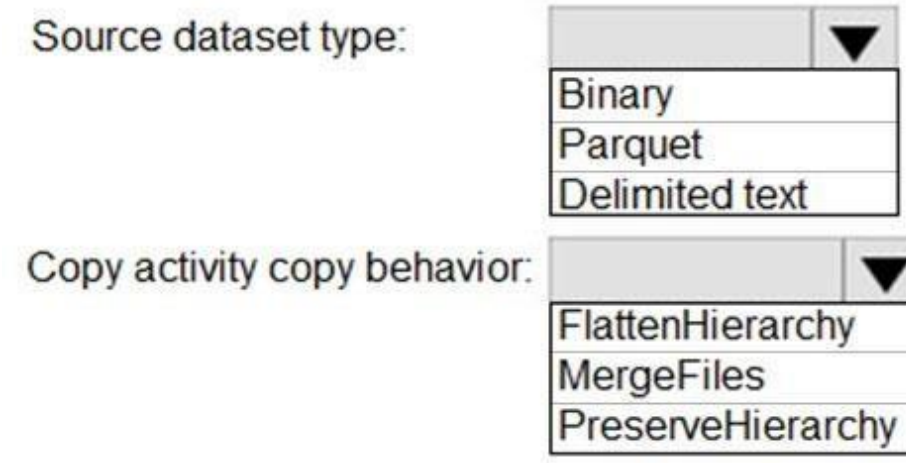
- No transformations must be performed.
- The original folder structure must be retained.
- Minimize time required to perform the copy activity.

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.

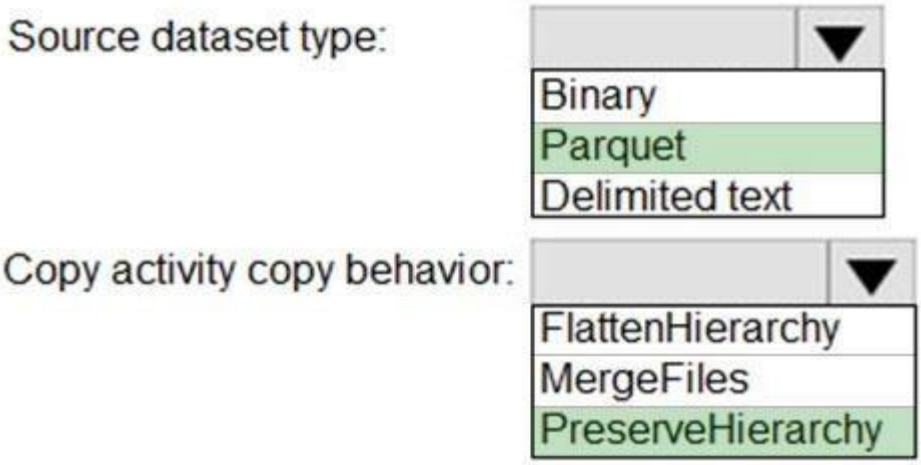
Hot Area:

Answer Area



Hide Solution | Discussion 12

Answer Area



Correct Answer:
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*Topic 1*
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)

Hide Solution Discussion 24
Correct Answer: *B*
Geo-redundant storage (with GRS or GZRS) replicates your data to another physical location in the secondary region to protect against regional outages.
However, that data is available to be read only if the customer or Microsoft initiates a failover from the primary to secondary region. When you enable read access to the secondary region, your data is available to be read at all times, including in a situation where the primary region becomes unavailable.
Incorrect Answers:
A: While Geo-redundant storage (GRS) is cheaper than Read-Access Geo-Redundant Storage (RA-GRS), GRS does NOT initiate automatic failover.
C, D: Locally redundant storage (LRS) and Zone-redundant storage (ZRS) provides redundancy within a single region.
Reference:
https://docs.microsoft.com/en-us/azure/storage/common/storage-redundancy

Question #18*Topic 1*
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) **Most Voted**

Hide Solution Discussion 32

Correct Answer: *C*

Locally redundant storage (LRS) copies your data synchronously three times within a single physical location in the primary region. LRS is the least expensive replication option

Reference:

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

Question #19*Topic 1*

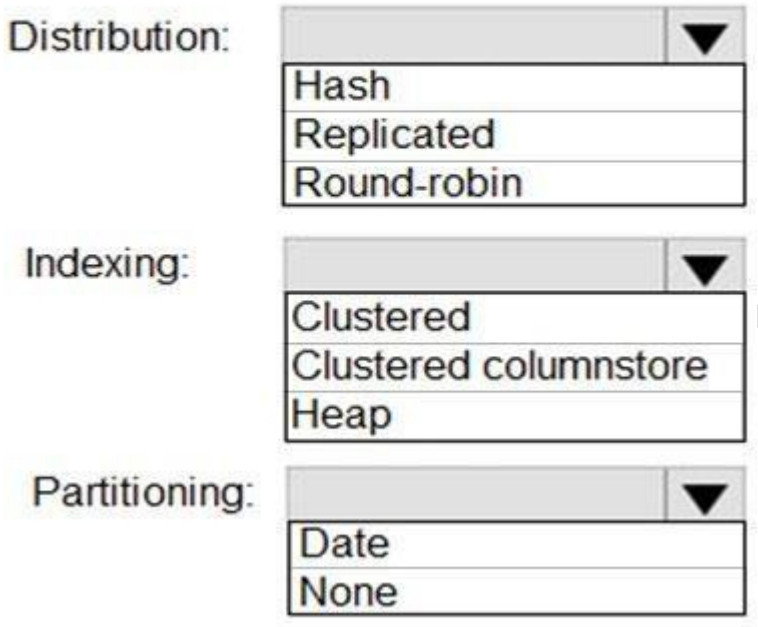
HOTSPOT -

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

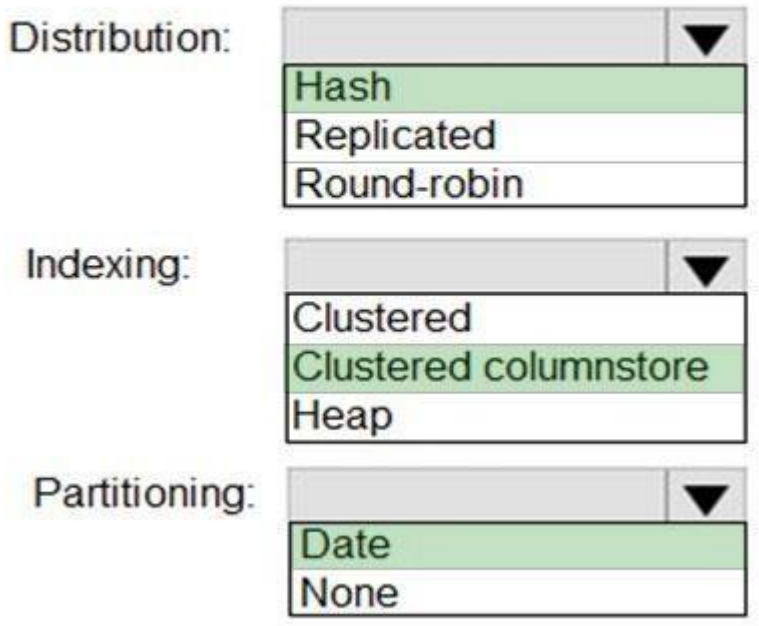
load.You need to create the staging table. The solution must minimize how long it takes toload the data to the staging table.
How should you configure the table? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
Hot Area:

Answer Area



Hide Solution Discussion 22

Answer Area



Correct Answer:
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*Topic 1*
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 **Most Voted**
- C. round-robin
- D. hash-distributed on IsOrderFinalized

Hide Solution | Discussion 8

Correct Answer: *B*
Hash-distributed tables improve query performance on large fact tables.
To balance the parallel processing, select a distribution column that:
☞ Has many unique values. The column can have duplicate values. All rows with the same value are assigned to the same distribution. Since there are 60 distributions,

some distributions can have > 1 unique values while others may end with zero values.
☞ Does not have NULLs, or has only a few NULLs.

☞ Is not a date column.
Incorrect Answers:
C: Round-robin tables are useful for improving loading speed.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute

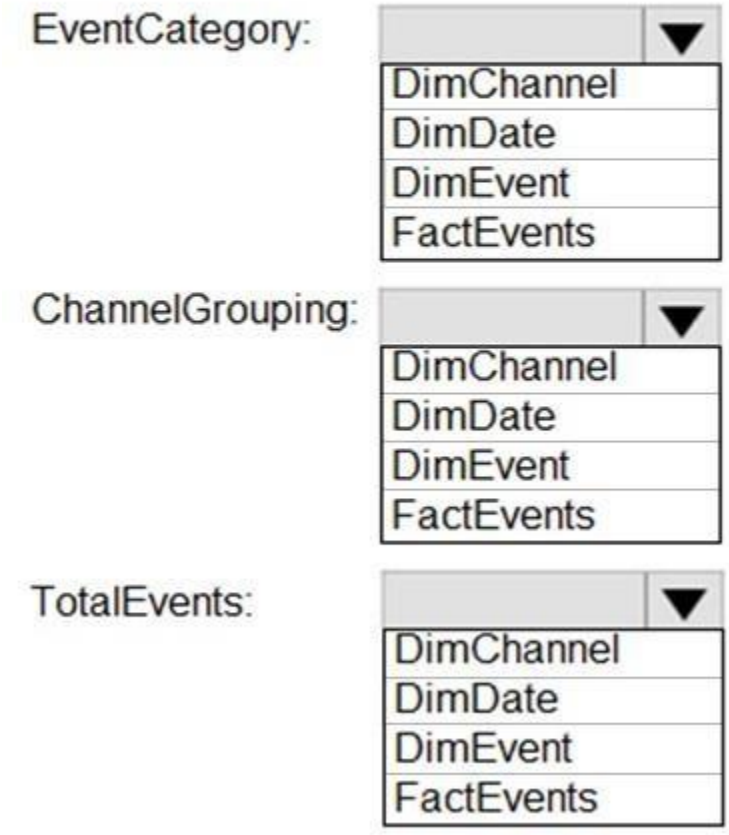
Question #21*Topic 1*
HOTSPOT -
From a website analytics system, you receive data extracts about user interactions such as downloads, link clicks, form submissions, and video plays.
The data contains the following columns.

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.

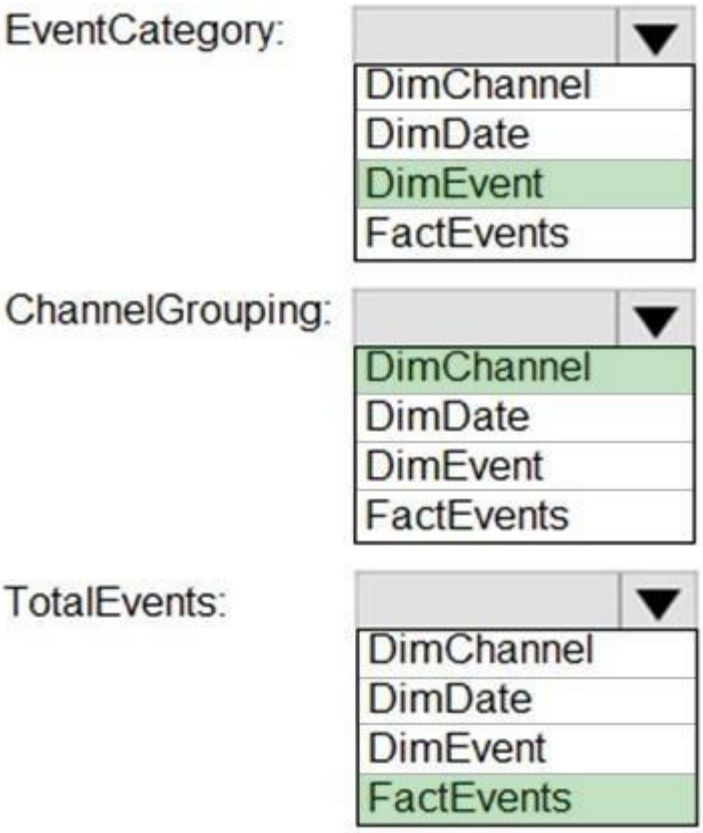
Hot Area:

Answer Area



Hide Solution Discussion 3

Answer Area



Correct Answer:
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*Topic 1*
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You have an Azure Storage account that contains 100 GB of files. The files contain rows of text and numerical values. 75% of the rows contain description data that has an average length of 1.1 MB.
You plan to copy the data from the storage account to an enterprise data warehouse in Azure Synapse Analytics.
You need to prepare the files to ensure that the data copies quickly.
Solution: You convert the files to compressed delimited text files.
Does this meet the goal?

- A. Yes
- B. No

Hide Solution Discussion 10
Correct Answer: *A*
All file formats have different performance characteristics. For the fastest load, use compressed delimited text files.
Reference:
https://docs.microsoft.com/en-us/azure/sql-data-warehouse/guidance-for-loading-data

Question #23*Topic 1*
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You have an Azure Storage account that contains 100 GB of files. The files contain rows of text and numerical values. 75% of the rows contain description data that has an average length of 1.1 MB.
You plan to copy the data from the storage account to an enterprise data warehouse in Azure Synapse Analytics.
You need to prepare the files to ensure that the data copies quickly.
Solution: You copy the files to a table that has a columnstore index.
Does this meet the goal?

- A. Yes
- B. No

Hide Solution Discussion 10
Correct Answer: *B*
Instead convert the files to compressed delimited text files.
Reference:

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

Question #24*Topic 1*
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You have an Azure Storage account that contains 100 GB of files. The files contain rows of text and numerical values. 75% of the rows contain description data that has an average length of 1.1 MB.
You plan to copy the data from the storage account to an enterprise data warehouse in Azure Synapse Analytics.
You need to prepare the files to ensure that the data copies quickly.
Solution: You modify the files to ensure that each row is more than 1 MB.
Does this meet the goal?

- A. Yes
- B. No

Hide Solution Discussion 4
Correct Answer: *B*
Instead convert the files to compressed delimited text files.
Reference:
https://docs.microsoft.com/en-us/azure/sql-data-warehouse/guidance-for-loading-data

Question #25*Topic 1*
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

Hide Solution Discussion 3
Correct Answer: *B*
Materialized views for dedicated SQL pools in Azure Synapse provide a low

maintenance method for complex analytical queries to get fast performance without any query change.

Incorrect Answers:
C: One daily execution does not make use of result cache caching.
Note: When result set caching is enabled, dedicated SQL pool automatically caches query results in the user database for repetitive use. This allows subsequent query executions to get results directly from the persisted cache so recomputation is not needed. Result set caching improves query performance and reduces compute resource usage. In addition, queries using cached results set do not use any concurrency slots and thus do not count against existing concurrency limits.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/performance-tuning-materialized-views https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/performance-tuning-result-set-caching

Question #26*Topic 1*
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

Hide Solution Discussion 3
Correct Answer: *D*
Serverless SQL pool can automatically synchronize metadata from Apache Spark. A serverless SQL pool database will be created for each database existing in serverless Apache Spark pools.
For each Spark external table based on Parquet or CSV and located in Azure Storage, an external table is created in a serverless SQL pool database.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-storage-files-spark-tables

Question #27*Topic 1*
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

Hide Solution Discussion 2

Correct Answer: *D*

The following tables summarize the key differences in capabilities for stream processing technologies in Azure.

General capabilities -

Capability	Azure Stream Analytics	HDInsight with Spark Streaming	Apache Spark in Azure Databricks	HDInsight with Storm
Programmability	Stream analytics query language, JavaScript	C#/F#, Java, Python, Scala	C#/F#, Java, Python, R, Scala	C#, Java

praw709528

Integration capabilities -

Capability	Azure Stream Analytics	HDInsight with Spark Streaming	Apache Spark in Azure Databricks	HDInsight with Storm
Inputs	Azure Event Hubs, Azure IoT Hub, Azure Blob storage	Event Hubs, IoT Hub, Kafka, HDFS, Storage Blobs, Azure Data Lake Store	Event Hubs, IoT Hub, Kafka, HDFS, Storage Blobs, Azure Data Lake Store	Event Hubs, IoT Hub, Storage Blobs, Azure Data Lake Store
Sinks	Azure Data Lake Store, Azure SQL Database, Storage Blobs, Event	HDFS, Kafka, Storage Blobs, Azure Data Lake Store, Cosmos DB	HDFS, Kafka, Storage Blobs, Azure Data Lake Store, Cosmos DB	Event Hubs, Service Bus, Kafka

praw709528

Reference:
https://docs.microsoft.com/en-us/azure/architecture/data-guide/technology-choices/stream-processing

Question #28*Topic 1*
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

Hide Solution Discussion 2

Correct Answer: *B*
Avro supports batch and is very relevant for streaming.
Note: Avro is framework developed within Apache's Hadoop project. It is a row-based storage format which is widely used as a serialization process. AVRO stores its schema in JSON format making it easy to read and interpret by any program. The data itself is stored in binary format by doing it compact and efficient.
Reference:
https://www.adaltas.com/en/2020/07/23/benchmark-study-of-different-file-format/

Question #29*Topic 1*
HOTSPOT -
You store files in an Azure Data Lake Storage Gen2 container. The container has the storage policy shown in the following exhibit.

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

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

Hot Area:

Answer Area

The files are **[answer choice]** after 30 days:

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

The storage policy applies to **[answer choice]**:

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

Hide Solution | Discussion 4

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-senstive 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.managementpolicyba

seblob.tiertocool

Question #30*Topic 1*

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 **Most Voted**

Hide Solution | Discussion 4

Correct Answer: *D*

For optimal compression and performance of clustered columnstore tables, a minimum of 1 million rows per distribution and partition is needed. Before partitions are created, dedicated SQL pool already divides each table into 60 distributed databases.
Example: Any partitioning added to a table is in addition to the distributions created behind the scenes. Using this example, if the sales fact table contained 36 monthly partitions, and given that a dedicated SQL pool has 60 distributions, then the sales fact table should contain 60 million rows per month, or 2.1 billion rows when all months are populated. If a table contains fewer than the recommended minimum number of rows per partition, consider using fewer partitions in order to increase the number of rows per partition.

Question #31*Topic 1*

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

Hide Solution | Discussion 2

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*Topic 1*
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

Hide Solution Discussion 2
Correct Answer: *B*
Need Parquet to support both Databricks and PolyBase.
Reference:
https://docs.microsoft.com/en-us/sql/t-sql/statements/create-external-file-format-transact-sql

Question #33*Topic 1*
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. Most Voted

- D. Update dbo.Sales from stg.Sales.

Hide Solution Discussion 11

Correct Answer: *B*

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 monthly. Then you can switch out the partition with data for an empty partition from another table

Note: Syntax:

SWITCH [PARTITION source_partition_number_expression] TO [schema_name.] target_table [PARTITION target_partition_number_expression]

Switches a block of data in one of the following ways:

☞ Reassigns all data of a table as a partition to an already-existing partitioned table.

☞ Switches a partition from one partitioned table to another.

☞ Reassigns all data in one partition of a partitioned table to an existing non-partitioned table.

Reference:

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

Question #34*Topic 1*

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-test 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 **Most Voted**
- B. effective start date **Most Voted**
- C. business key
- D. last modified date
- E. effective end date **Most Voted**
- F. foreign key

Hide Solution Discussion 6

Correct Answer: *BCE*
C: The Slowly Changing Dimension transformation requires at least one business key column.
BE: Historical attribute changes create new records instead of updating existing ones. The only change that is permitted in an existing record is an update to a column that indicates whether the record is current or expired. This kind of change is equivalent to a Type 2 change. The Slowly Changing Dimension transformation directs these rows to two outputs: Historical Attribute Inserts Output and New Output.
Reference:
https://docs.microsoft.com/en-us/sql/integration-services/data-flow/transformations/slowly-changing-dimension-transformation

Question #35*Topic 1*
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

Hide Solution Discussion 2

Correct Answer:

Answer Area

Transform data for the dimension tables by:

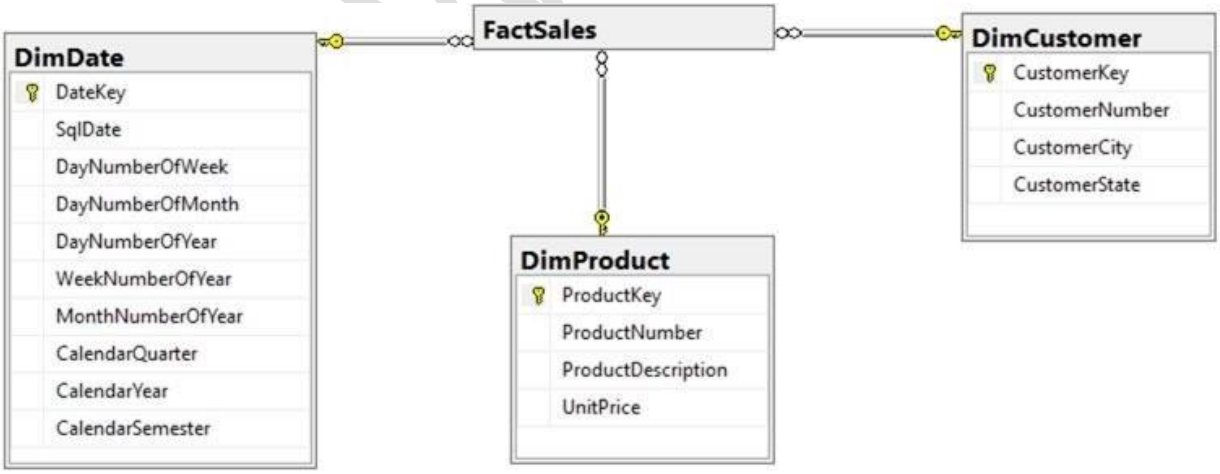
- Maintaining to a third normal form
- Normalizing to a fourth normal form
- **Denormalizing to a second normal form**

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*Topic 1*
HOTSPOT -
You plan to develop a dataset named Purchases by using Azure Databricks. Purchases will contain the following columns:
- ProductID
- ItemPrice
- LineTotal
- Quantity
- StoreID
- Minute
- Month
- Hour

Year -
-
- Day

You need to store the data to support hourly incremental load pipelines that will vary for each Store ID. The solution must minimize storage costs.
How should you complete the code? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
Hot Area:

Answer Area

```
df.write
[ .bucketBy | .partitionBy | .range | .sortBy ]
[ ("*") | ("StoreID","Hour") | ("StoreID","Year","Month","Day","Hour") ]
.mode("append")
[ .csv("/Purchases") | .json("/Purchases") | .parquet("/Purchases") | .saveAsTable("/Purchases") ]
```

Hide Solution Discussion 1
Correct Answer:

Answer Area

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*Topic 1*
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

Hide Solution | Discussion 1
Correct Answer: *A*
Each partition should have around 1 millions records. Dedication SQL pools already have 60 partitions.

We have the formula: Records/(Partitions*60)= 1 million
Partitions= Records/(1 million * 60)
Partitions= 2.4 x 1,000,000,000/(1,000,000 * 60) = 40
Note: Having too many partitions can reduce the effectiveness of clustered columnstore indexes if each partition has fewer than 1 million rows. Dedicated SQL pools automatically partition your data into 60 databases. So, if you create a table with 100 partitions, the result will be 6000 partitions.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/best-practices-dedicated-sql-pool

Question #38*Topic 1*
HOTSPOT -
You are creating dimensions for a data warehouse in an Azure Synapse Analytics dedicated SQL pool.
You create a table by using the Transact-SQL statement shown in the following exhibit.

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

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

Hot Area:

Answer Area

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

- Type 0
- Type 1
- Type 2

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

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

Hide Solution | Discussion 43

Correct Answer:

Answer Area

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

- Type 0
- Type 1
- Type 2 (selected)

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

- a surrogate key
- a business key (selected)
- an audit column

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*Topic 1*
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

Hide Solution Discussion 29
Correct Answer: *B*
Hash-distributed tables improve query performance on large fact tables, and are the focus of this article. Round-robin tables are useful for improving loading speed.
Incorrect:
Not D: Do not use a date column. . All data for the same date lands in the same distribution. If several users are all filtering on the same date, then only 1 of the 60 distributions do all the processing work.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute

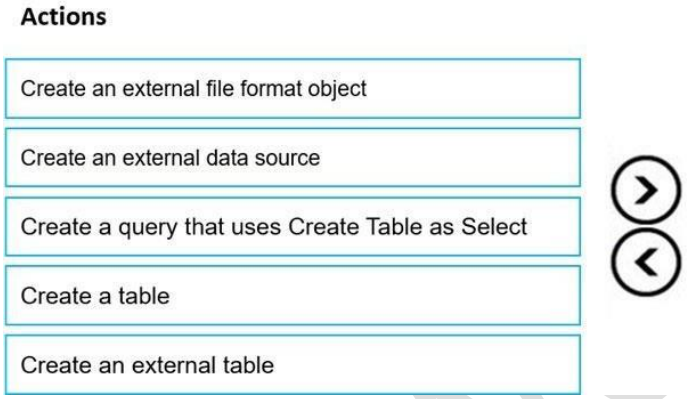
Question #40*Topic 1*
You are implementing a batch dataset in the Parquet format.
Data files will be produced be using Azure Data Factory and stored in Azure Data Lake Storage Gen2. The files will be consumed by an Azure Synapse Analytics serverless SQL pool.
You need to minimize storage costs for the solution.
What should you do?

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

Hide Solution Discussion 5
Correct Answer: *C*
An external table points to data located in Hadoop, Azure Storage blob, or Azure Data Lake Storage. External tables are used to read data from files or write data to files in Azure Storage. With Synapse SQL, you can use external tables to read external data

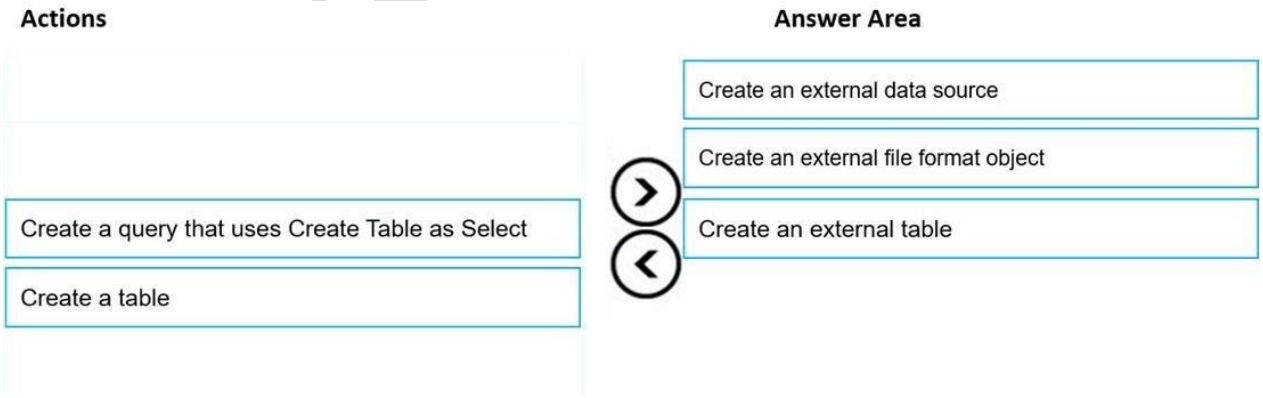
using dedicated SQL pool or serverless SQL pool.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables

Question #41*Topic 1*
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:



Hide Solution Discussion 2

Correct Answer:



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*Topic 1*
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

Hide Solution Discussion 2
Correct Answer: *CE*
C: Dimension tables contain attribute data that might change but usually changes infrequently. For example, a customer's name and address are stored in a dimension table and updated only when the customer's profile changes. To minimize the size of a large fact table, the customer's name and address don't need to be in every row of a fact table. Instead, the fact table and the dimension table can share a customer ID. A

query can join the two tables to associate a customer's profile and transactions.
E: Fact tables contain quantitative data that are commonly generated in a transactional system, and then loaded into the dedicated SQL pool. For example, a retail business generates sales transactions every day, and then loads the data into a dedicated SQL pool fact table for analysis.
Reference:

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

Question #43*Topic 1*
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

Hide Solution Discussion 1
Correct Answer: *C*
A Type 2 SCD supports versioning of dimension members. Often the source system doesn't store versions, so the data warehouse load process detects and manages changes in a dimension table. In this case, the dimension table must use a surrogate key to provide a unique reference to a version of the dimension member. It also includes columns that define the date range validity of the version (for example, StartDate and EndDate) and possibly a flag column (for example, IsCurrent) to easily filter by current dimension members.
Incorrect Answers:
B: A Type 1 SCD always reflects the latest values, and when changes in source data are detected, the dimension table data is overwritten.
D: A Type 3 SCD supports storing two versions of a dimension member as separate columns. The table includes a column for the current value of a member plus either the original or previous value of the member. So Type 3 uses additional columns to track one key instance of history, rather than storing additional rows to track each change like in a Type 2 SCD.
Reference:
https://docs.microsoft.com/en-us/learn/modules/populate-slowly-changing-dimensions-azure-synapse-analytics-pipelines/3-choose-between-dimension-types

Question #44*Topic 1*

DRAG DROP -

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

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

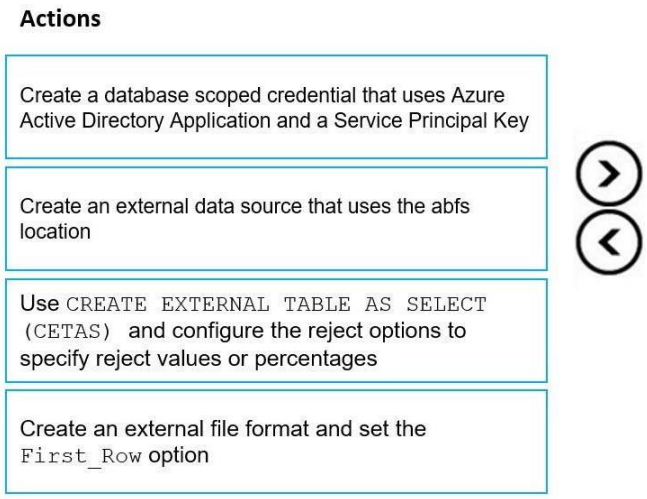
You need to skip the header row when you import the files into the data warehouse.

Before building the loading pattern, you need to prepare the required database objects in Azure Synapse Analytics.

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

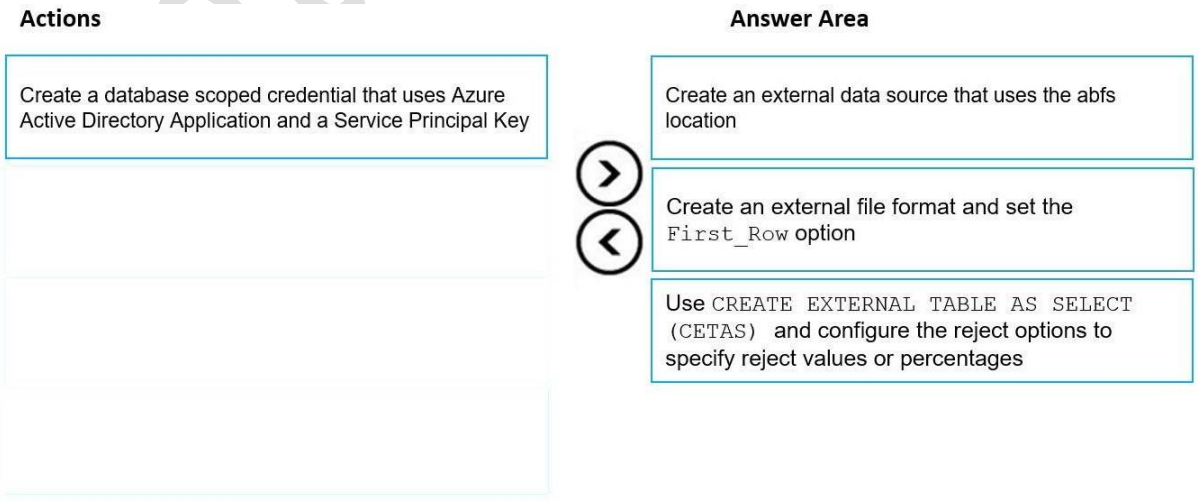
NOTE: Each correct selection is worth one point

Select and Place:



Hide Solution Discussion 3

Correct Answer:



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*Topic 1*
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

```
(
```

[▼] RANGE RIGHT FOR VALUES
- [TransactionDateID]
- [TransactionDateID], [TransactionTypeID]
- HASH([TransactionTypeID])
- ROUND_ROBIN

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

Hide Solution Discussion 1

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

```
(
```

[▼] RANGE RIGHT FOR VALUES
- [TransactionDateID]
- [TransactionDateID], [TransactionTypeID]
- HASH([TransactionTypeID])
- ROUND_ROBIN

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

Box 1: PARTITION -
RANGE RIGHT FOR VALUES is used with PARTITION.
Part 2: [TransactionDateID]
Partition on the date column.
Example: Creating a RANGE RIGHT partition function on a datetime column
The following partition function partitions a table or index into 12 partitions, one for each month of a year's worth of values in a datetime column.
CREATE PARTITION FUNCTION [myDateRangePF1] (datetime)
AS RANGE RIGHT FOR VALUES ('20030201', '20030301', '20030401',
'20030501', '20030601', '20030701', '20030801',
'20030901', '20031001', '20031101', '20031201');
Reference:
https://docs.microsoft.com/en-us/sql/t-sql/statements/create-partition-function-transact-sql

Question #46*Topic 1*
You are performing exploratory analysis of the bus fare data in an Azure Data Lake Storage Gen2 account by using an Azure Synapse Analytics serverless SQL pool.
You execute the Transact-SQL query shown in the following exhibit.

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

What do the query results include?

- A. Only CSV files in the tripdata_2020 subfolder.
- B. All files that have file names that beginning with "tripdata_2020".
- C. All CSV files that have file names that contain "tripdata_2020".
- D. Only CSV that have file names that beginning with "tripdata_2020".

Hide Solution Discussion 3

Correct Answer: *D*

Question #1*Topic 2*
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 **Most Voted**
- B. No

Hide Solution Discussion 14

Correct Answer: *B*
We would need a High Concurrency cluster for the jobs.
Note:
Standard clusters are recommended for a single user. Standard can run workloads developed in any language: Python, R, Scala, and SQL.
A high concurrency cluster is a managed cloud resource. The key benefits of high concurrency clusters are that they provide Apache Spark-native fine-grained sharing for maximum resource utilization and minimum query latencies.
Reference:
https://docs.azuredatabricks.net/clusters/configure.html

Question #2*Topic 2*
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:

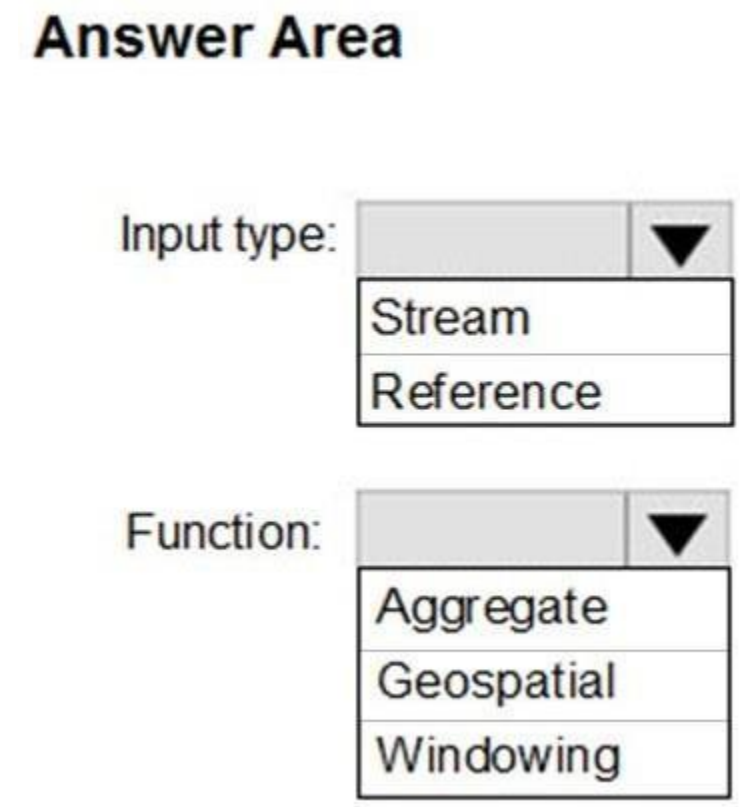
☞ The data engineers must share a cluster.
☞ The job cluster will be managed by using a request process whereby data scientists and data engineers provide packaged notebooks for deployment to the cluster.
☞ All the data scientists must be assigned their own cluster that terminates automatically after 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

Hide Solution Discussion 10
Correct Answer: *A*
We need a High Concurrency cluster for the data engineers and the jobs.
Note: Standard clusters are recommended for a single user. Standard can run workloads developed in any language: Python, R, Scala, and SQL.
A high concurrency cluster is a managed cloud resource. The key benefits of high concurrency clusters are that they provide Apache Spark-native fine-grained sharing for maximum resource utilization and minimum query latencies.
Reference:
https://docs.azuredatabricks.net/clusters/configure.html

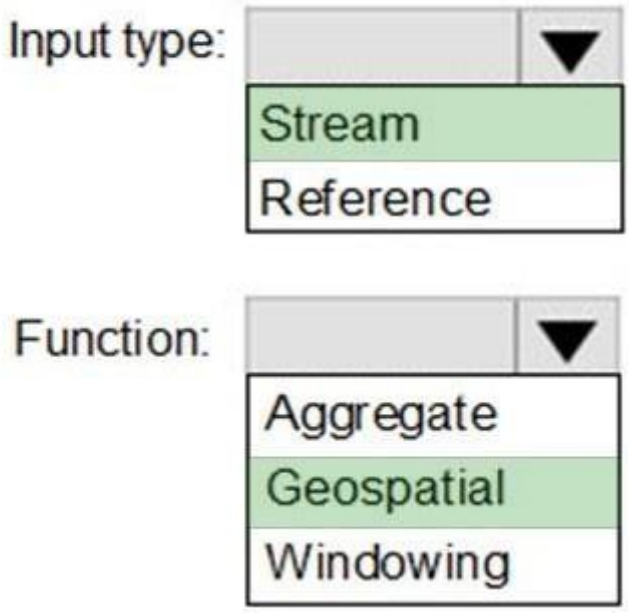
Question #3*Topic 2*
HOTSPOT -
You plan to create a real-time monitoring app that alerts users when a device travels more than 200 meters away from a designated location.
You need to design an Azure Stream Analytics job to process the data for the planned app. The solution must minimize the amount of 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:



Hide Solution Discussion 1

Answer Area



Correct Answer:
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 #4*Topic 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.

Hide Solution Discussion 14

Correct Answer: *DF*
D: Scale out the query by allowing the system to process each input partition separately.
F: A Stream Analytics job definition includes inputs, a query, and output. Inputs are where the job reads the data stream from.
Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-parallelization

Question #5*Topic 2*
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

Hide Solution Discussion 4

Correct Answer: *C*
Event-driven architecture (EDA) is a common data integration pattern that involves production, detection, consumption, and reaction to events. Data integration scenarios often require Data Factory customers to trigger pipelines based on events happening in storage account, such as the arrival or deletion of a file in Azure
Blob Storage account. Data Factory natively integrates with Azure Event Grid, which lets you trigger pipelines on such events.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/how-to-create-event-trigger
https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipeline-execution-triggers

Question #6*Topic 2*
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

Hide Solution Discussion 3
Correct Answer: *B*
Azure Databricks has two types of clusters: interactive and automated. You use interactive clusters to analyze data collaboratively with interactive notebooks. You use automated clusters to run fast and robust automated jobs.
Example: Scheduled batch workloads (data engineers running ETL jobs)
This scenario involves running batch job JARs and notebooks on a regular cadence through the Databricks platform.
The suggested best practice is to launch a new cluster for each run of critical jobs. This helps avoid any issues (failures, missing SLA, and so on) due to an existing workload (noisy neighbor) on a shared cluster.
Reference:
https://docs.databricks.com/administration-guide/cloud-configurations/aws/cmbp.html#scenario-3-scheduled-batch-workloads-data-engineers-running-etl-jobs

Question #7*Topic 2*
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
                [COUNT | MAX | MIN | TOPONE] (Time) AS LastEventTime
        FROM
                Input TIMESTAMP BY Time
        GROUP BY
                [HoppingWindow | SessionWindow | SlidingWindow | TumblingWindow] (minute, 10)
)
SELECT
        Input.License_plate,
        Input.Make,
        Input.Time
FROM
        Input TIMESTAMP BY Time
        INNER JOIN LastInWindow
        ON [DATEADD | DATEDIFF | DATENAME | DATEPART] (minute, Input, LastInWindow) BETWEEN 0 AND 10
AND Input.Time = LastInWindow.LastEventTime
```

Hide Solution | Discussion 6

Correct Answer:

Answer Area

```
WITH LastInWindow AS
(
        SELECT
                [ COUNT | MAX | MIN | TOPONE ] (Time) AS LastEventTime
        FROM
                Input TIMESTAMP BY Time
        GROUP BY
                [ HoppingWindow | SessionWindow | SlidingWindow | TumblingWindow ] (minute, 10)
)
SELECT
        Input.License_plate,
        Input.Make,
        Input.Time
FROM
        Input TIMESTAMP BY Time
        INNER JOIN LastInWindow
        ON [ DATEADD | DATEDIFF | DATENAME | DATEPART ] (minute, Input, LastInWindow) BETWEEN 0 AND 10
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 #8*Topic 2*
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.

Hide Solution Discussion 9

Correct Answer: *A*
Activities are linked together via dependencies. A dependency has a condition of one of the following: Succeeded, Failed, Skipped, or Completed.
Consider Pipeline1:
If we have a pipeline with two activities where Activity2 has a failure dependency on Activity1, the pipeline will not fail just because Activity1 failed. If Activity1 fails and Activity2 succeeds, the pipeline will succeed. This scenario is treated as a try-catch block by Data Factory.

The failure dependency means this pipeline reports success.
Note:
If we have a pipeline containing Activity1 and Activity2, and Activity2 has a success dependency on Activity1, it will only execute if Activity1 is successful. In this scenario, if Activity1 fails, the pipeline will fail.
Reference:
https://datasavvy.me/category/azure-data-factory/

Question #9*Topic 2*
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:

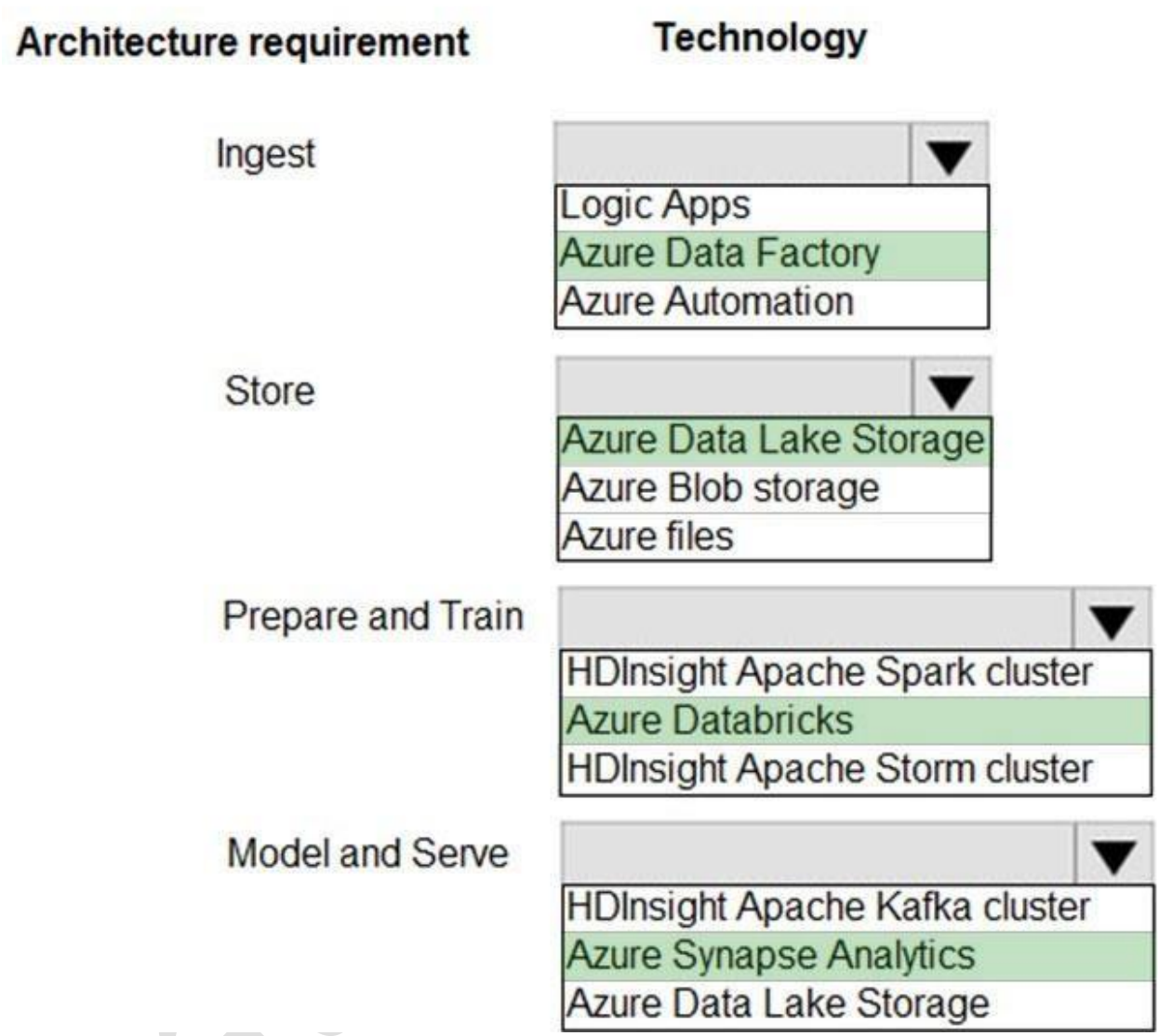
Answer Area

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

Hide Solution Discussion 10

Correct Answer:

Answer Area



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.

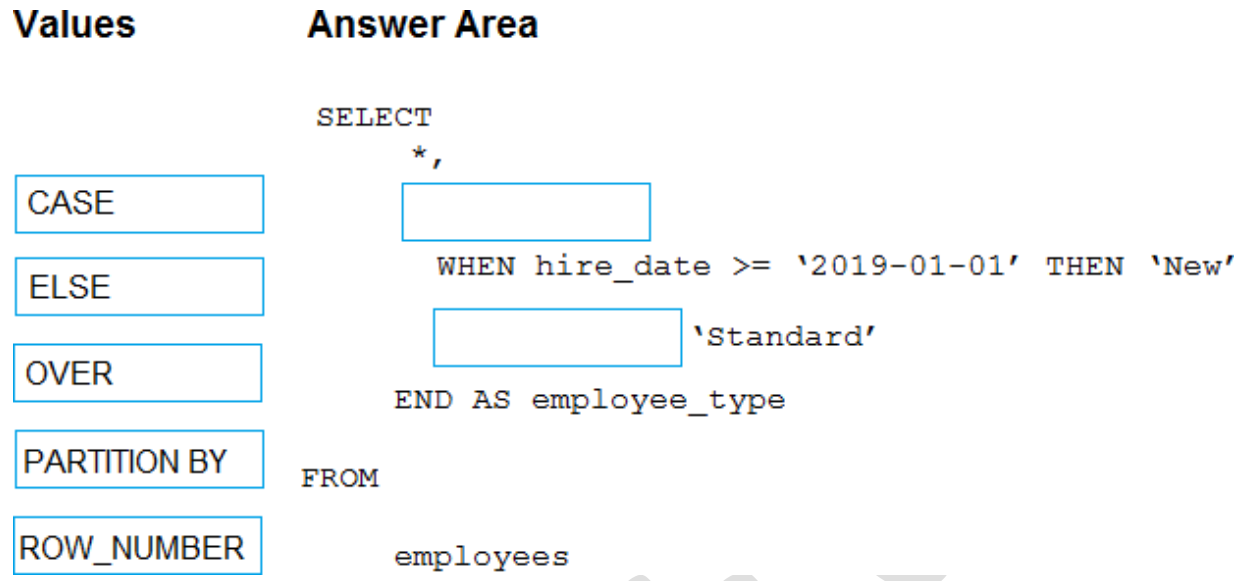
Prepare and Train: Azure Databricks
Azure Databricks provides enterprise-grade Azure security, including Azure Active Directory integration.
With Azure Databricks, you can set up your Apache Spark environment in minutes, autoscale and collaborate on shared projects in an interactive workspace.
Azure 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/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 #10*Topic 2*
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

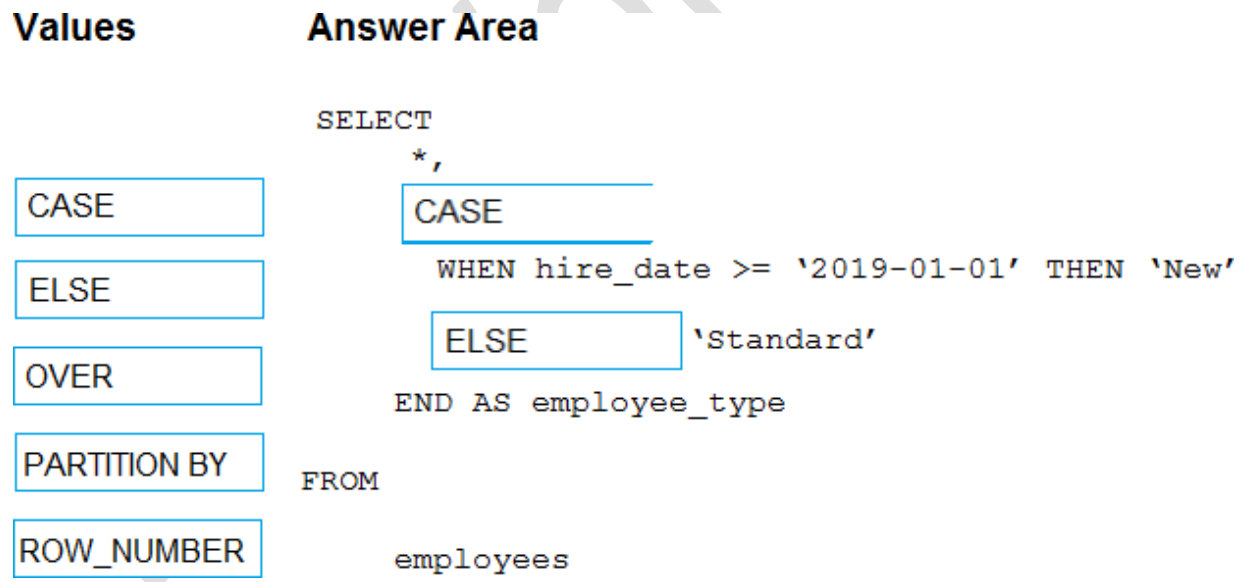
You need to calculate the employee_type value based on the hire_date value.
How should you complete the Transact-SQL statement? To answer, drag the appropriate values to the correct targets. Each value may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.
NOTE: Each correct selection is worth one point.

Select and Place:



Hide Solution Discussion 4

Correct Answer:



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 #11*Topic 2*
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

- opendatasource
- openjson
- openquery
- openrowset

Answer Area

```
select*

FROM
[                ] (
     BULK 'https://contoso.blob.core.windows.net/contosodw',
     FORMAT= 'CSV',
     fieldterminator = '0x0b',
     fieldquote = '0x0b',
     rowterminator = '0x0b'
)
with (id varchar(50),
     contextdateventTime 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'
     )
```

Hide Solution | Discussion 6

Correct Answer:

Values

opendatasource

openquery

Answer Area

```
select*

FROM
   openrowset   (
        BULK 'https://contoso.blob.core.windows.net/contosodw',
        FORMAT= 'CSV',
        fieldterminator = '0x0b',
        fieldquote = '0x0b',
        rowterminator = '0x0b'
)
with (id varchar(50),
        contextdateventTime 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 #12*Topic 2*

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

CAST
COLLATE
CONVERT
FLATTEN
PIVOT
UNPIVOT

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
```

Hide Solution Discussion 2

Correct Answer:

Values **Answer Area**

COLLATE
CONVERT
FLATTEN
UNPIVOT

```
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 #13*Topic 2*
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

Hide Solution Discussion 2
Correct Answer: *D*
Annotations are additional, informative tags that you can add to specific factory resources: pipelines, datasets, linked services, and triggers. By adding annotations, you can easily filter and search for specific factory resources.
Reference:
https://www.cathrinewilhelmsen.net/annotations-user-properties-azure-data-factory/

Question #14*Topic 2*
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.	○	○
The Databricks cluster minimizes costs when running scheduled jobs that execute notebooks.	○	○
The Databricks cluster supports the creation of a Delta Lake table.	○	○

Hide Solution Discussion 15

Correct Answer:

Answer Area

Statements	Yes	No
The Databricks cluster supports multiple concurrent users.	○	○
The Databricks cluster minimizes costs when running scheduled jobs that execute notebooks.	○	○
The Databricks cluster supports the creation of a Delta Lake table.	○	○

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 #15*Topic 2*
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 **Most Voted**
- C. Azure Stream Analytics
- D. Azure SQL Database

Hide Solution Discussion 12

Correct Answer: *C*

Reference:
https://docs.microsoft.com/en-us/azure/event-hubs/process-data-azure-stream-analytics

Question #16*Topic 2*

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 places 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

)

Hide Solution | Discussion 5

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
```

[▼] options: RIGHT / LEFT

([▼])

[▼] options: 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 #17*Topic 2*

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,

Hide Solution | Discussion 6

Correct Answer: *BE*
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 #18*Topic 2*

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

Hide Solution Discussion 12

Correct Answer: *B*

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

Reference:

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

Question #19*Topic 2*

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

Hide Solution Discussion 8

Correct Answer: *B*

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

Reference:

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

Question #20*Topic 2*

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:

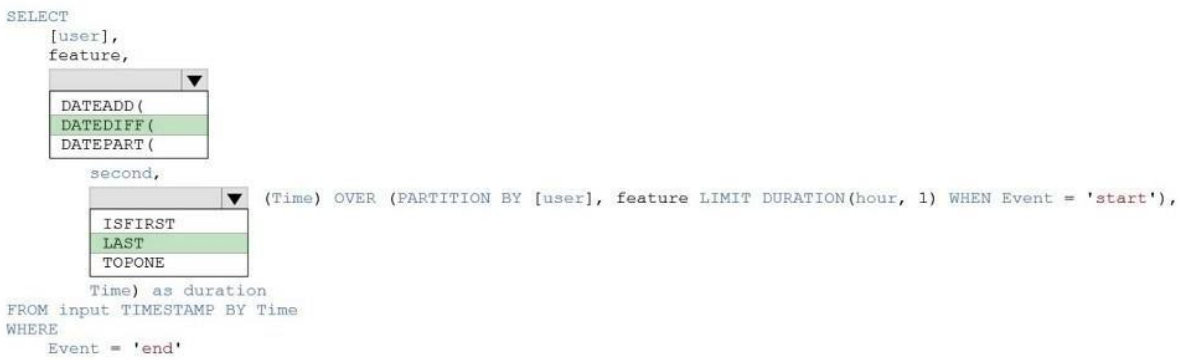
Answer Area

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

Hide Solution Discussion 7

Correct Answer:

Answer Area



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

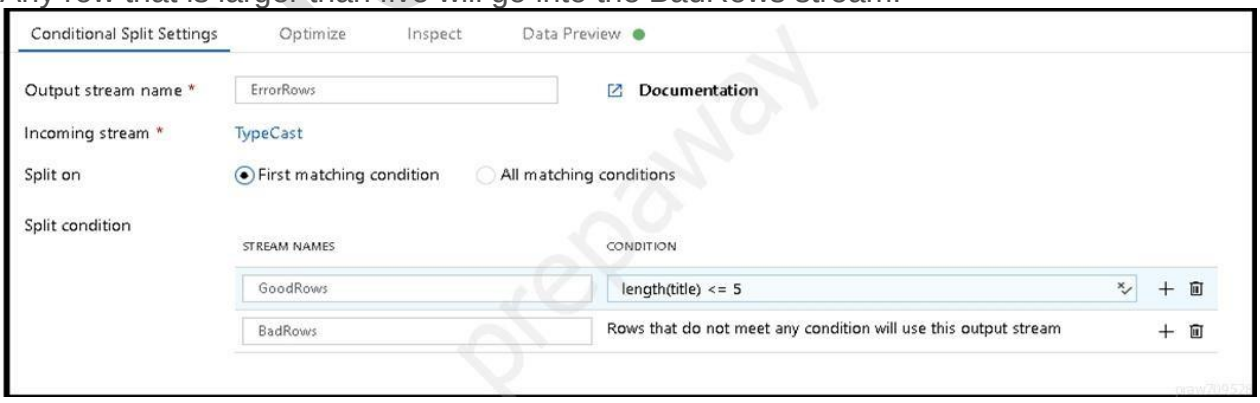
Question #21*Topic 2*
You are creating an Azure Data Factory data flow that will ingest data from a CSV file, cast columns to specified types of data, and insert the data into a table in an

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.

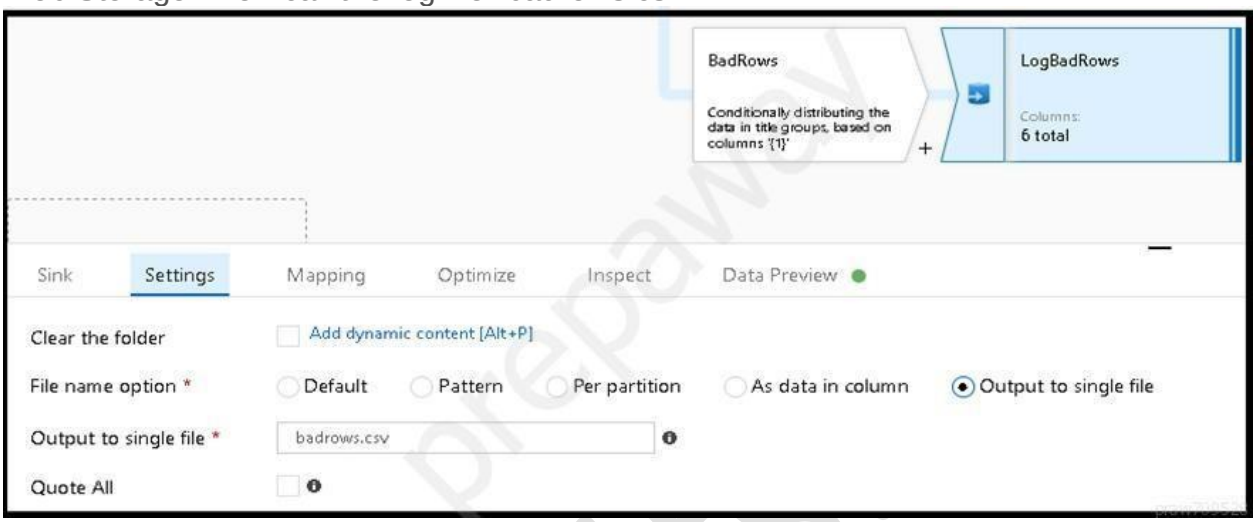
Hide Solution Discussion 8

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

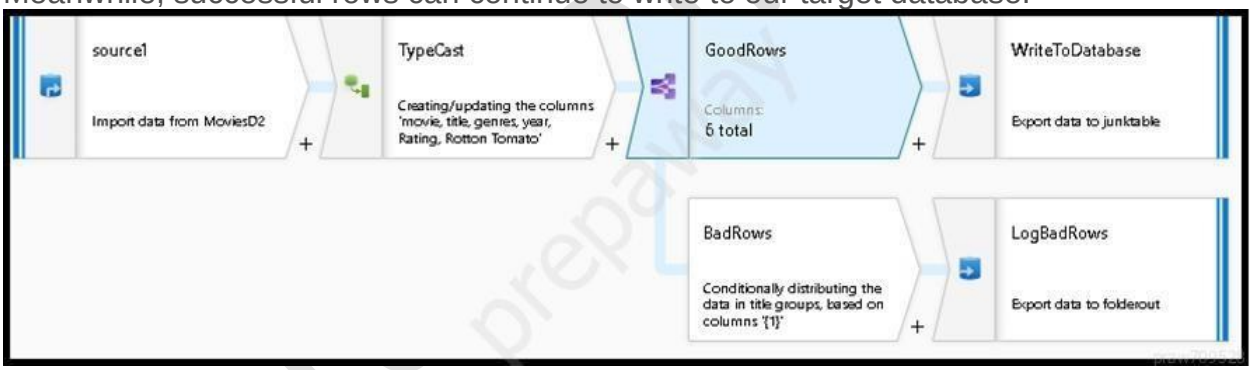


2. This conditional split transformation defines the maximum length of "title" to be five.
Any row that is less than or equal to five will go into the GoodRows stream.
Any row that is larger than five will go into the BadRows stream.

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



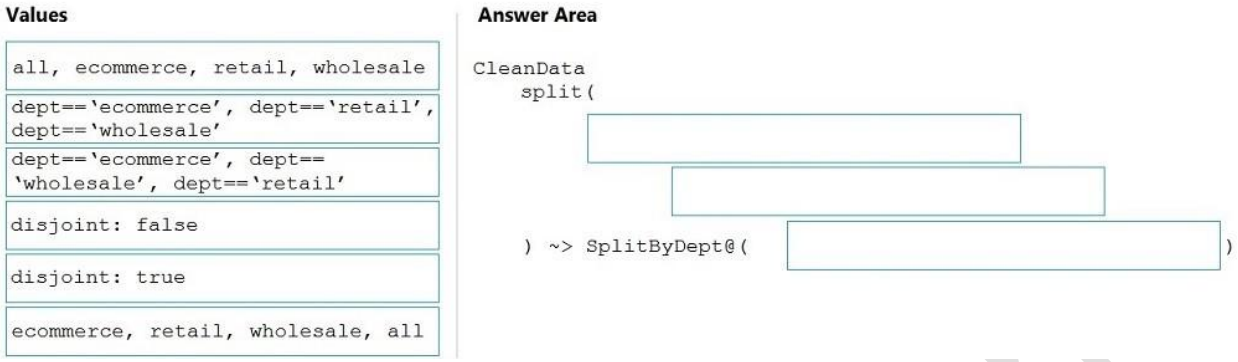
4. The completed data flow is shown below. We are now able to split off error rows to avoid the SQL truncation errors and put those entries into a log file.
Meanwhile, successful rows can continue to write to our target database.



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

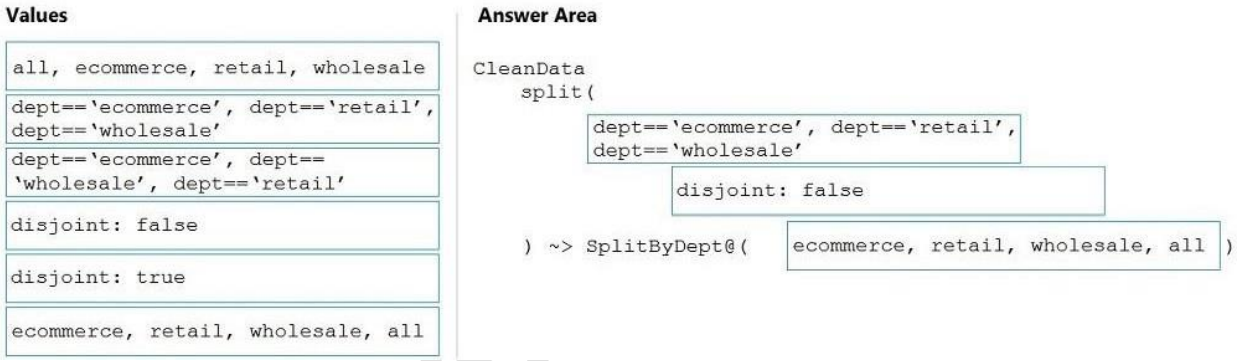
Question #22*Topic 2*
DRAG DROP -
You need to create an Azure Data Factory pipeline to process data for the following three departments at your company: Ecommerce, retail, and wholesale. The solution must ensure that data can also be processed for the entire company.
How should you complete the Data 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:



Hide Solution | Discussion 21

Correct Answer:



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 #23*Topic 2*
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.	

Hide Solution Discussion 16
Correct Answer:

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

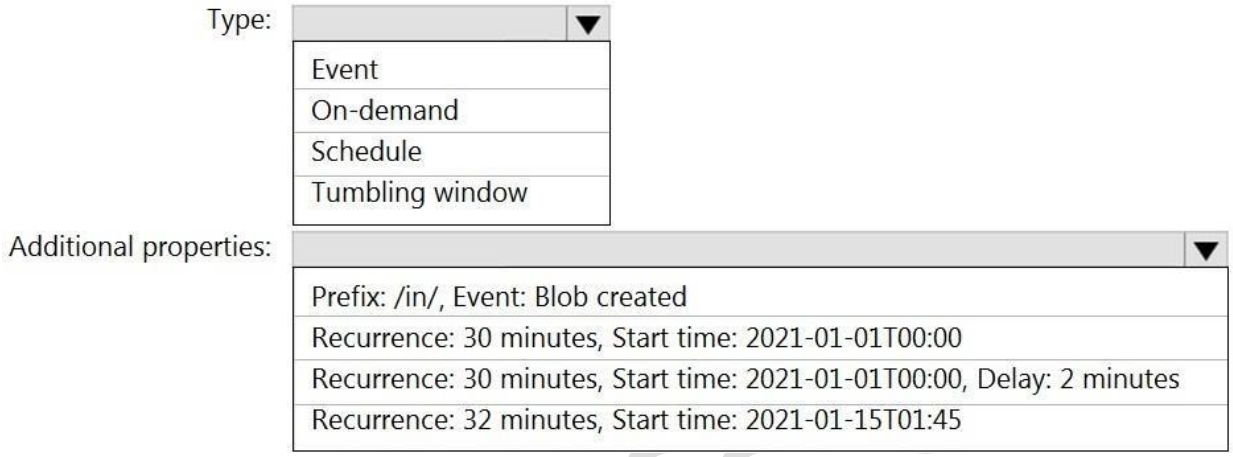
Step 1: Read the file into a data frame.
You can load the json files as a data frame in Azure Databricks.
Step 2: Perform transformations on the data frame.
Step 3:Specify a temporary folder to stage the data
Specify a temporary folder to use while moving data between Azure Databricks and Azure Synapse.
Step 4: Write the results to a table in Azure Synapse.
You upload the transformed data frame into Azure Synapse. You use the Azure Synapse connector for Azure Databricks to directly upload a dataframe as a table in a Azure Synapse.

Step 5: Drop the data frame -
Clean up resources. You can terminate the cluster. From the Azure Databricks workspace, select Clusters on the left. For the cluster to terminate, under Actions, point to the ellipsis (...) and select the Terminate icon.
Reference:
https://docs.microsoft.com/en-us/azure/azure-databricks/databricks-extract-load-sql-data-warehouse

Question #24*Topic 2*
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:

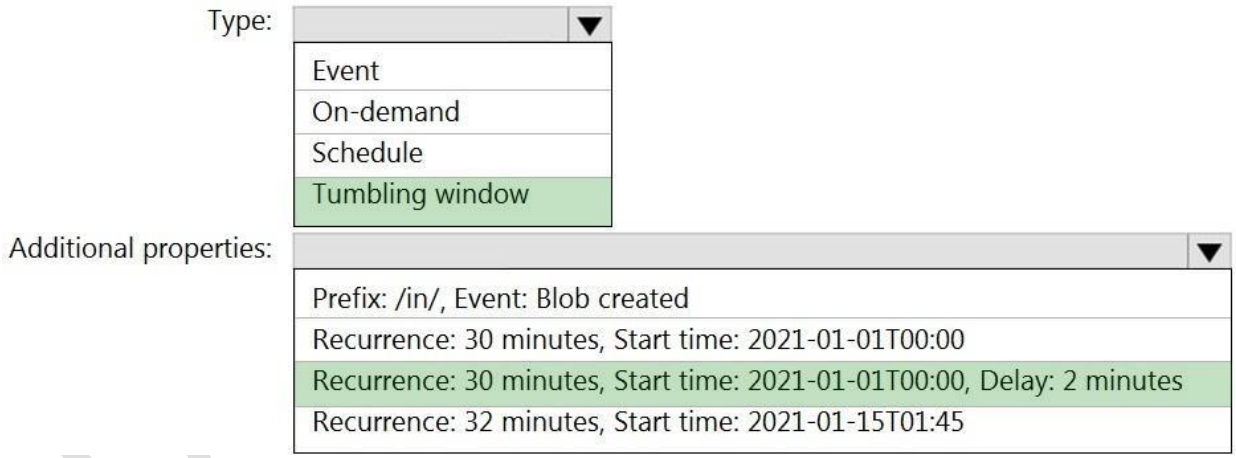
Answer Area



Hide Solution | Discussion 7

Correct Answer:

Answer Area



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 #25*Topic 2*
HOTSPOT -
You are designing a 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:

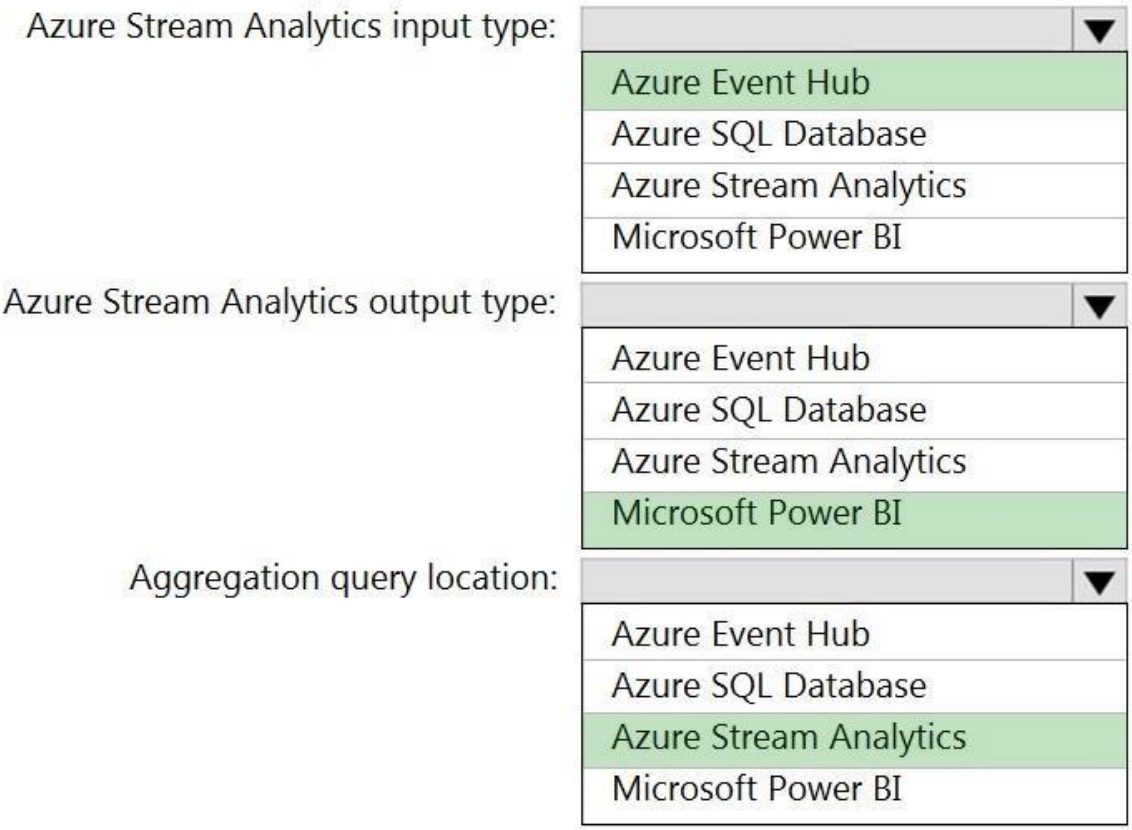
Answer Area

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

Hide Solution Discussion 5

Correct Answer:

Answer Area



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

Question #26*Topic 2*
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.	

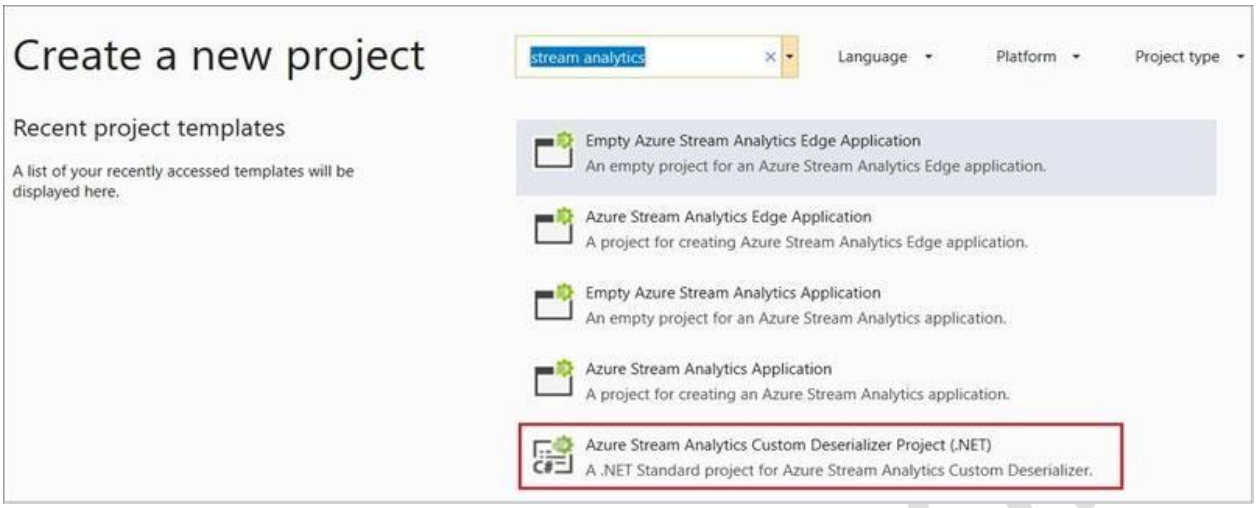
Hide Solution Discussion 6

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.



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 #27*Topic 2*
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

Hide Solution Discussion 8

Correct Answer: *A*

IR type	Public network	Private network
Azure	Data Flow Data movement Activity dispatch	
Self-hosted	Data movement Activity dispatch	Data movement Activity dispatch
Azure-SSIS	SSIS package execution	SSIS package execution

Incorrect Answers:
C: Self-hosted integration runtime is to be used On-premises.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/concepts-integration-runtime

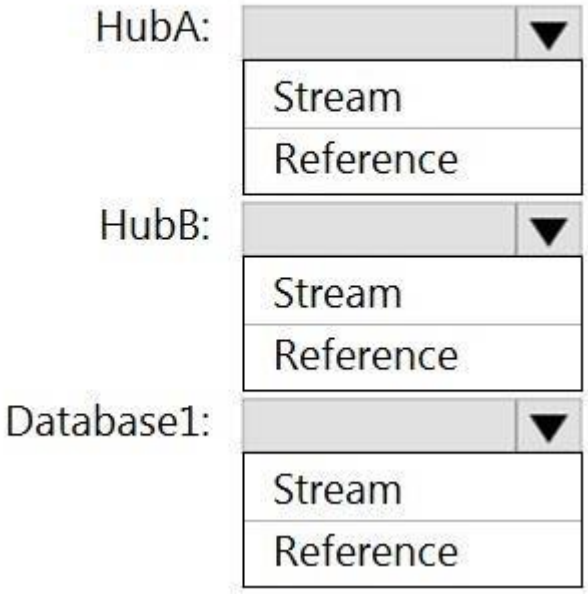
Question #28*Topic 2*
HOTSPOT -
You have an Azure SQL database named Database1 and two Azure event hubs named HubA and HubB. The data consumed from each source is shown in the following table.

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.

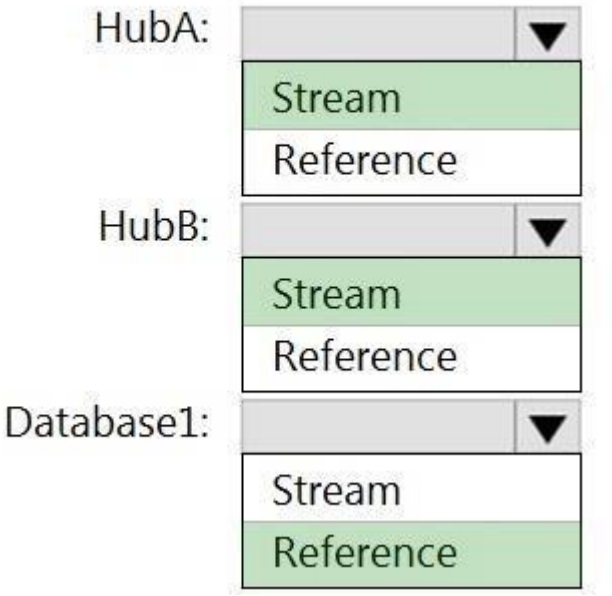
Hot Area:

Answer Area



Hide Solution Discussion 5

Answer Area



Correct Answer:
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 #29*Topic 2*
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)

Hide Solution Discussion 6
Correct Answer: *B*
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.
Example:
Incorrect Answers:
A: Sliding windows, unlike Tumbling or Hopping windows, output events only for points in time when the content of the window actually changes. In other words, when an event enters or exits the window. Every window has at least one event, like in the case of Hopping windows, events can belong to more than one sliding window.
C: 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, so 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.
D: 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/azure/stream-analytics/stream-analytics-window-functions

Question #30*Topic 2*
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 [          ] (hour,1))
FROM input
```

LAG / LAST / LEAD

LIMIT DURATION / OFFSET / WHEN

Hide Solution | Discussion 1

Correct Answer:

Answer Area

```
SELECT sensorId,
       growth = reading -
                [ LAG ] (reading) OVER (PARTITION BY sensorId [ LIMIT DURATION ] (hour,1))
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 #31*Topic 2*
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

Hide Solution Discussion 4
Correct Answer: *D*
Event-driven architecture (EDA) is a common data integration pattern that involves production, detection, consumption, and reaction to events. Data integration scenarios often require Data Factory customers to trigger pipelines based on events happening in storage account, such as the arrival or deletion of a file in Azure
Blob Storage account.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/how-to-create-event-trigger

Question #32

Question #32*Topic 2*
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.

Hide Solution Discussion 3

Correct Answer: *C*
In Azure Data Factory, continuous integration and delivery (CI/CD) means moving Data Factory pipelines from one environment (development, test, production) to another.
Note: The following is a guide for setting up an Azure Pipelines release that automates the deployment of a data factory to multiple environments.
1. In Azure DevOps, open the project that's configured with your data factory.
2. On the left side of the page, select Pipelines, and then select Releases.
3. Select New pipeline, or, if you have existing pipelines, select New and then New release pipeline.
4. In the Stage name box, enter the name of your environment.
5. Select Add artifact, and then select the git repository configured with your development data factory. Select the publish branch of the repository for the Default branch. By default, this publish branch is adf_publish.
6. Select the Empty job template.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/continuous-integration-deployment

Question #33*Topic 2*
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

Hide Solution Discussion 5
Correct Answer: *B*
Stream Analytics supports Azure Blob storage and Azure SQL Database as the storage layer for Reference Data.
Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-use-reference-data

Question #34*Topic 2*
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?

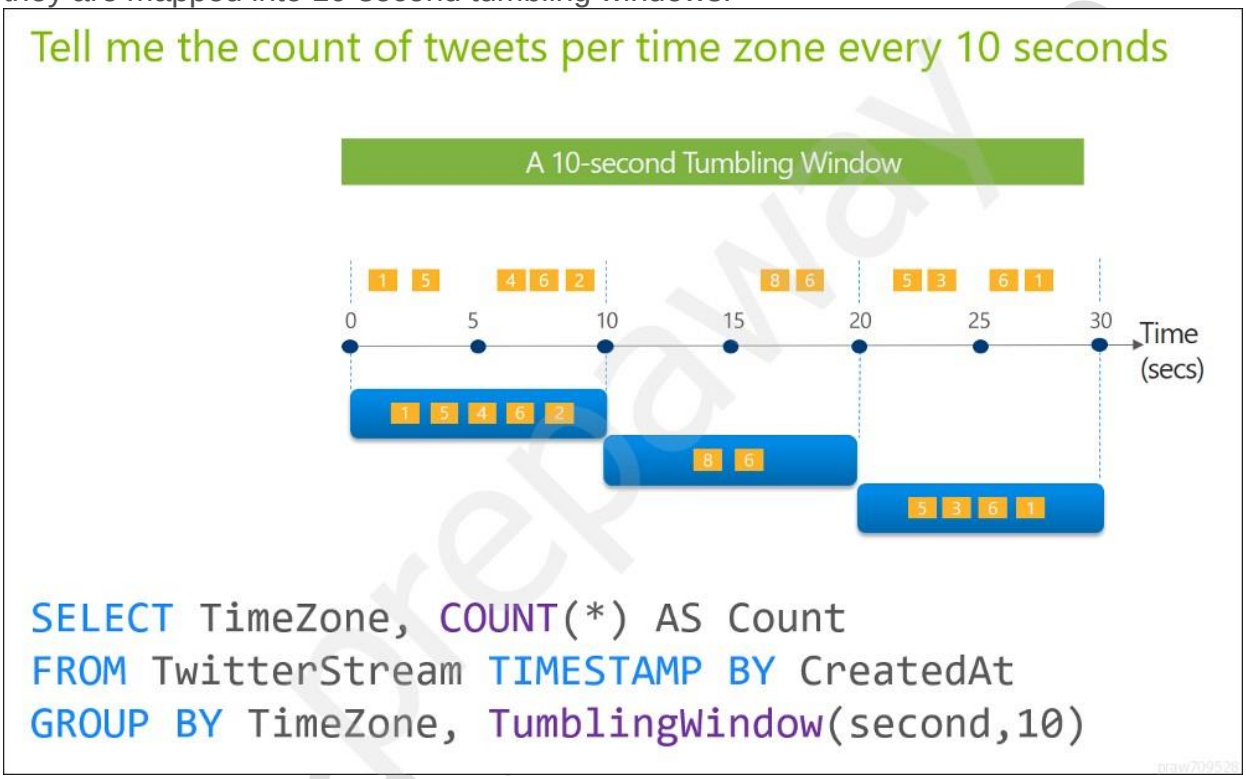
- A. snapshot

- B. tumbling
- C. hopping **Most Voted**
- D. sliding

Hide Solution | Discussion 33

Correct Answer: *B*

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



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

Question #35*Topic 2*

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

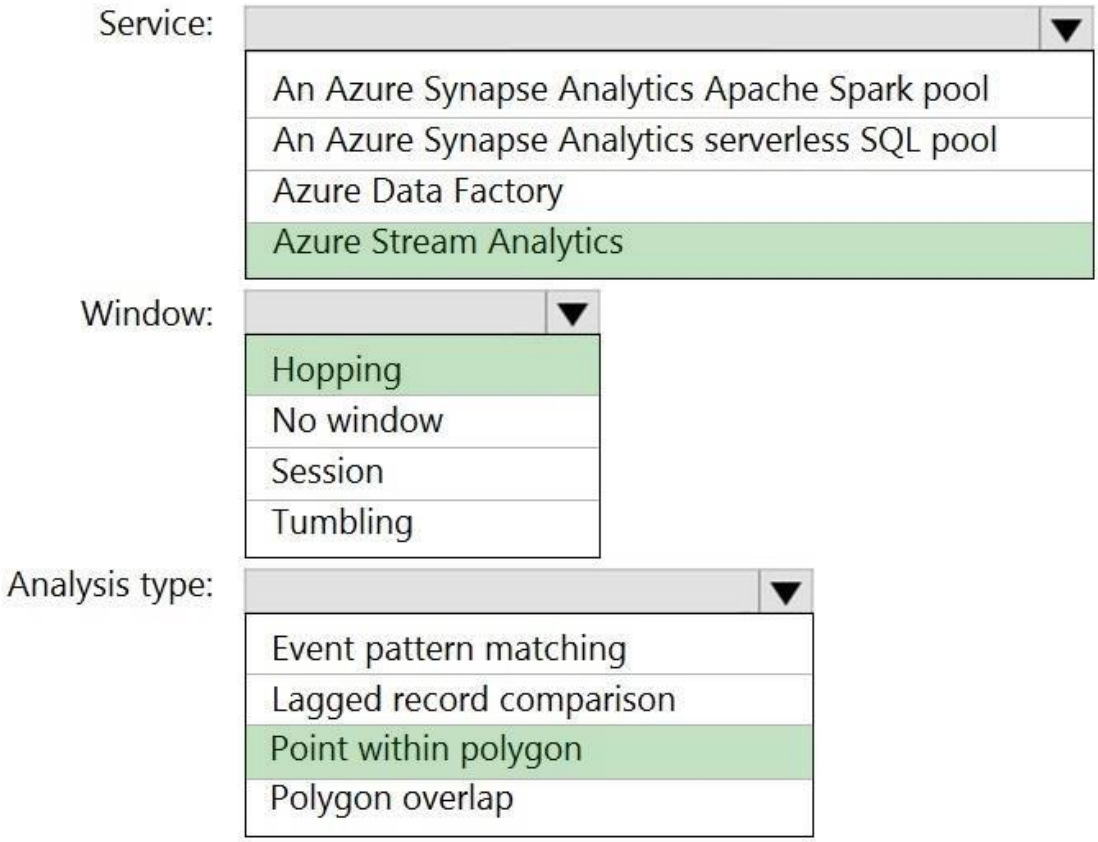
Window:
- Hopping
- No window
- Session
- Tumbling

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

Hide Solution | Discussion 19

Correct Answer:

Answer Area



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 #36*Topic 2*
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.

Hide Solution Discussion 7

Correct Answer: *B*
The Databricks ABS-AQS connector uses Azure Queue Storage (AQS) to provide an optimized file source that lets you find new files written to an Azure Blob storage (ABS) container without repeatedly listing all of the files. This provides two major advantages:
☞ Lower latency: no need to list nested directory structures on ABS, which is slow and resource intensive.
☞ Lower costs: no more costly LIST API requests made to ABS.
Reference:
https://docs.microsoft.com/en-us/azure/databricks/spark/latest/structured-streaming/aqs

Question #37*Topic 2*
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

Hide Solution | Discussion 22

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 #38*Topic 2*
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.

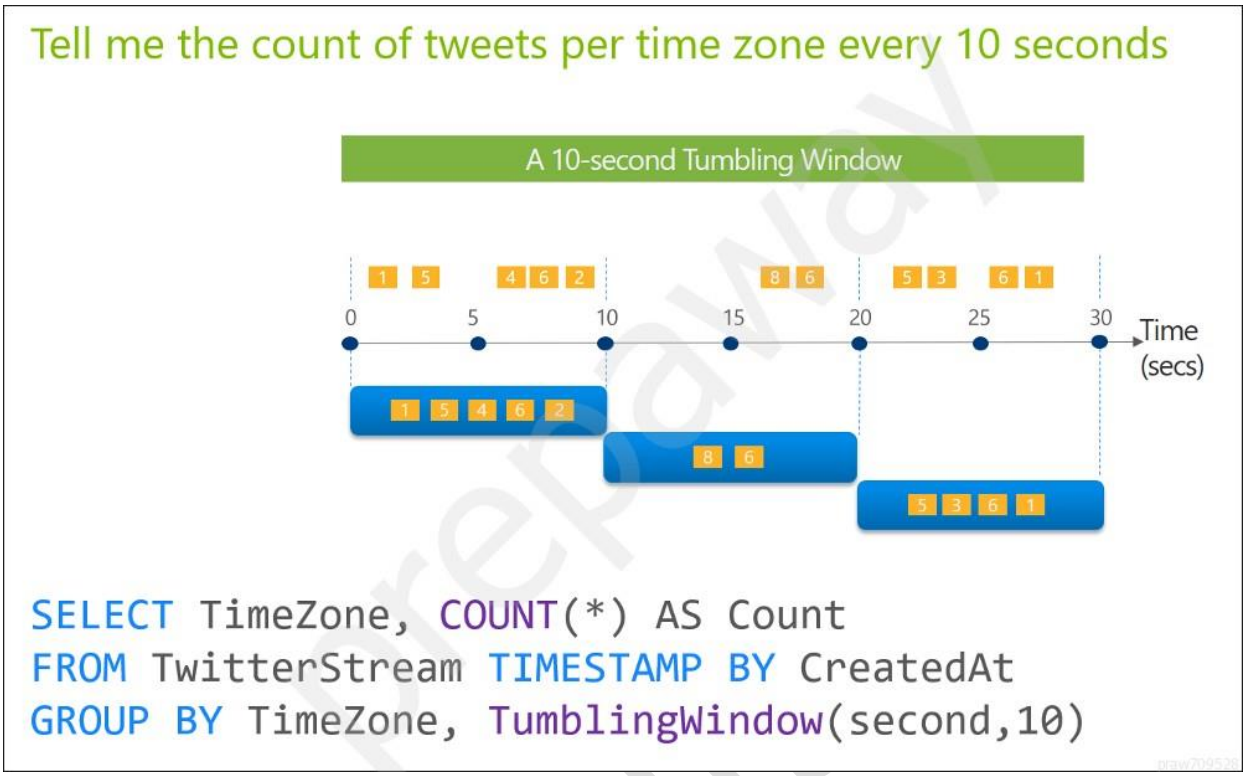
Hide Solution Discussion 11
Correct Answer: *B*
For clusters running Databricks Runtime 6.4 and above, optimized autoscaling is used by all-purpose clusters in the Premium plan
Optimized autoscaling:
Scales up from min to max in 2 steps.
Can scale down even if the cluster is not idle by looking at shuffle file state.
Scales down based on a percentage of current nodes.
On job clusters, scales down if the cluster is underutilized over the last 40 seconds.
On all-purpose clusters, scales down if the cluster is underutilized over the last 150 seconds.
The spark.databricks.aggressiveWindowDownS Spark configuration property specifies in seconds how often a cluster makes down-scaling decisions. Increasing the value causes a cluster to scale down more slowly. The maximum value is 600.

Note: Standard autoscaling -
Starts with adding 8 nodes. Thereafter, scales up exponentially, but can take many steps to reach the max. You can customize the first step by setting the spark.databricks.autoscaling.standardFirstStepUp Spark configuration property.
Scales down only when the cluster is completely idle and it has been underutilized for the last 10 minutes.
Scales down exponentially, starting with 1 node.
Reference:
https://docs.databricks.com/clusters/configure.html

Question #39*Topic 2*
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

Hide Solution Discussion 7
Correct Answer: *A*
Tumbling windows are a series of fixed-sized, non-overlapping and contiguous time intervals. The following diagram illustrates a stream with a series of events and how they are mapped into 10-second tumbling windows.



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

Question #40*Topic 2*
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

Hide Solution Discussion 7
Correct Answer: *B*
Instead use a tumbling window. Tumbling windows are a series of fixed-sized, non-

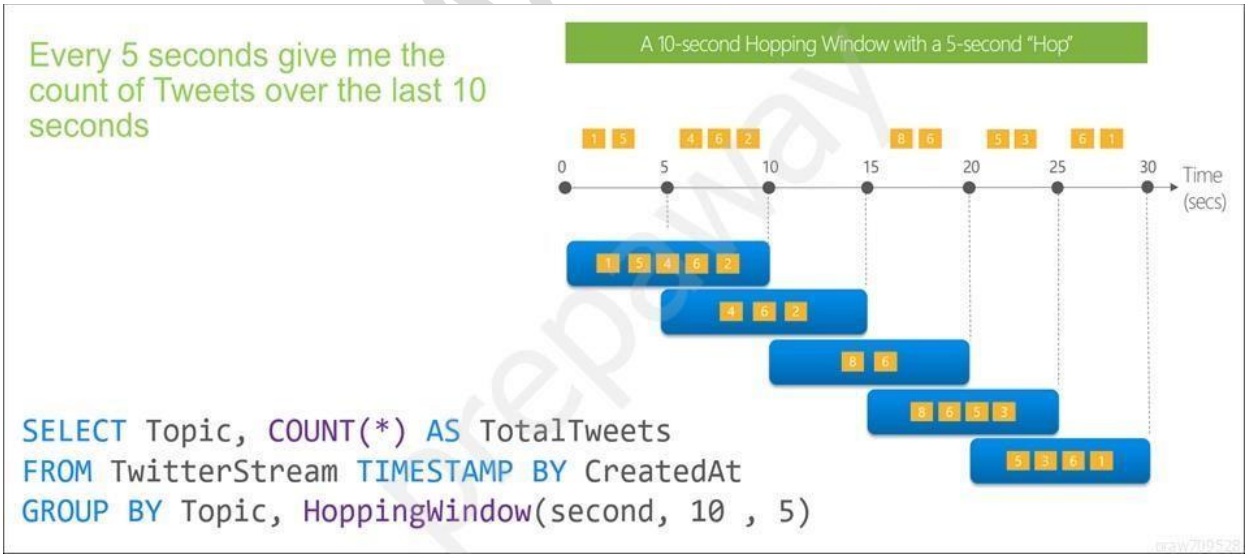
overlapping and contiguous time intervals.

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

Question #41*Topic 2*
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

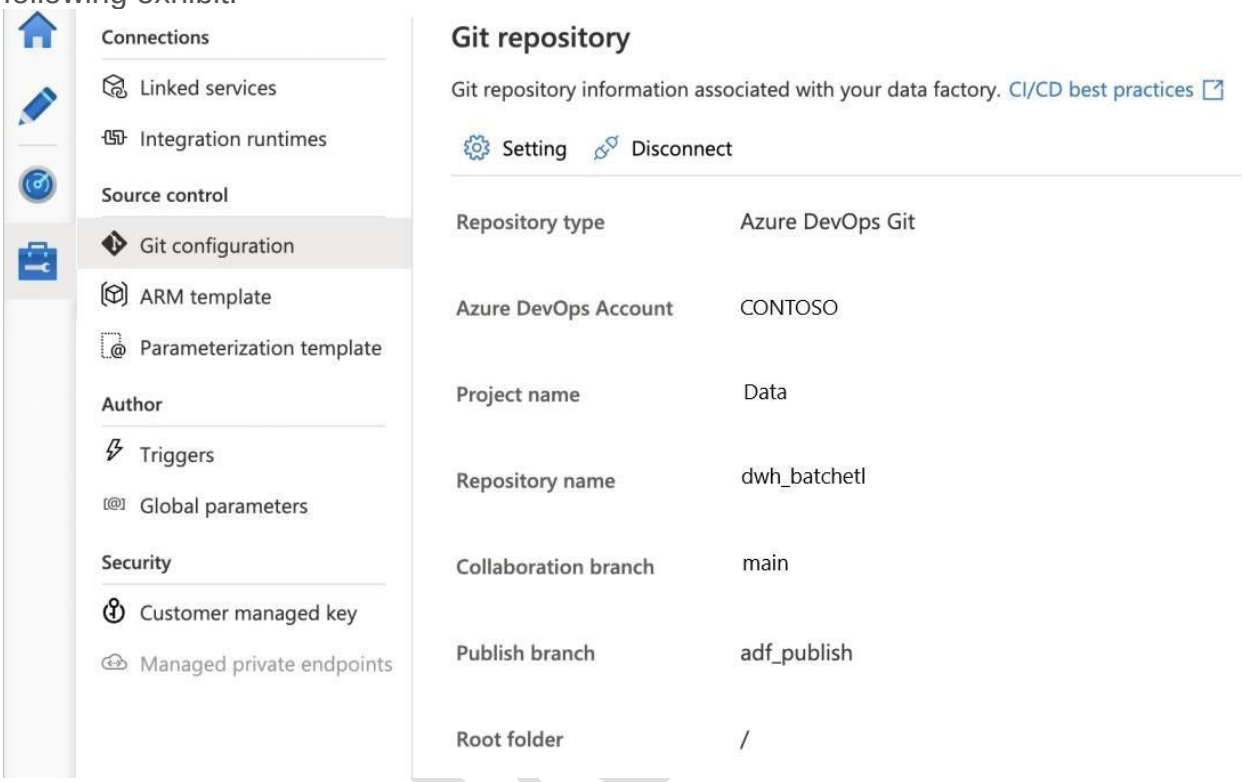
Hide Solution Discussion 1
Correct Answer: *D*
Hopping window functions hop forward in time by a fixed period. It may be easy to think of them as Tumbling windows that can overlap and be emitted more often than the window size. Events can belong to more than one Hopping window result set. To make a Hopping 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/azure/stream-analytics/stream-analytics-window-functions

Question #42*Topic 2*

HOTSPOT -
You configure version control for an Azure Data Factory instance as 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

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

Hide Solution Discussion

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 #43*Topic 2*
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 [ ▼ ] CreatedAt
                   LAST
                   OVER
                   SYSTEM.TIMESTAMP()
                   TIMESTAMP BY

GROUP BY TimeZone, [ ▼ ] (second,15)
                   HOPPINGWINDOW
                   SESSIONWINDOW
                   SLIDINGWINDOW
                   TUMBLINGWINDOW
```

Hide Solution | Discussion 3

Correct Answer:

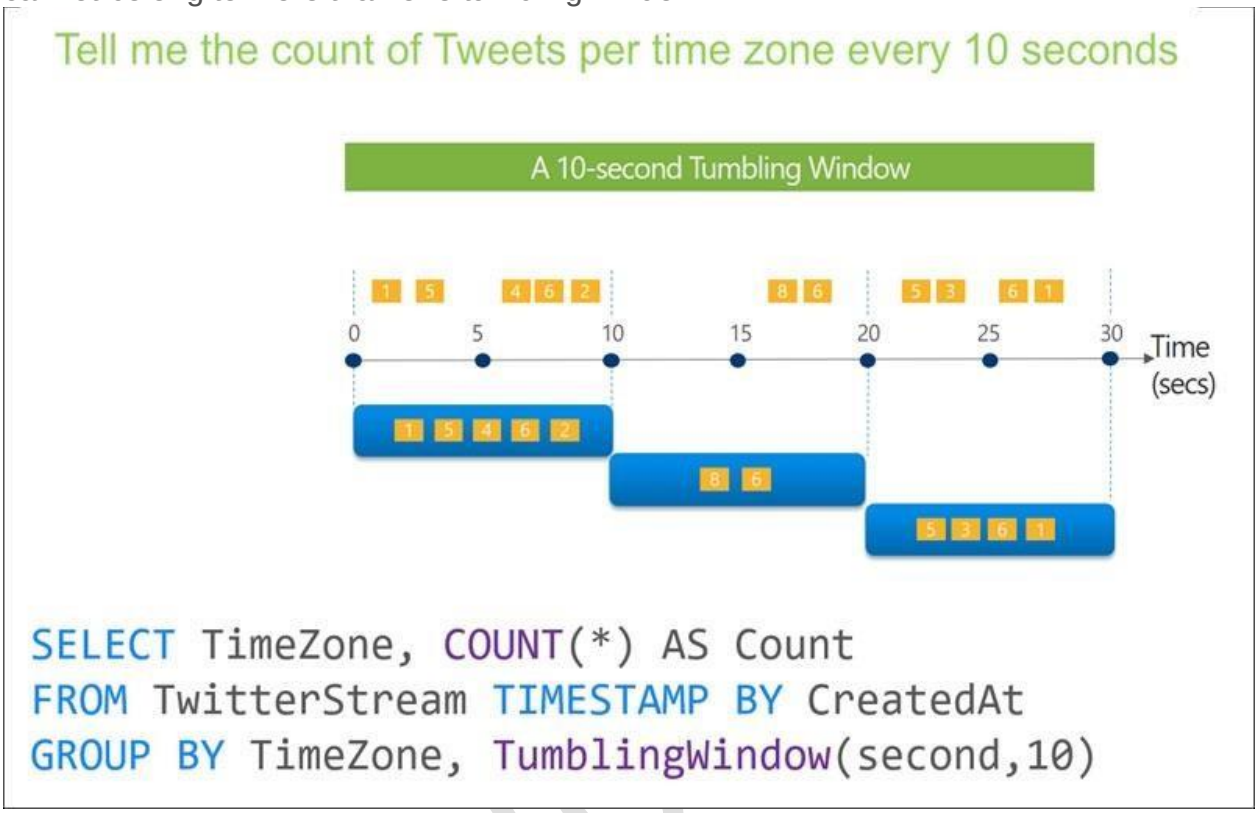
Answer Area

```
Select TimeZone, count (*) AS MessageCount

FROM MessageStream [ ▼ ] CreatedAt
                   LAST
                   OVER
                   SYSTEM.TIMESTAMP()
                   TIMESTAMP BY

GROUP BY TimeZone, [ ▼ ] (second,15)
                   HOPPINGWINDOW
                   SESSIONWINDOW
                   SLIDINGWINDOW
                   TUMBLINGWINDOW
```

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.



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

Question #44*Topic 2*
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

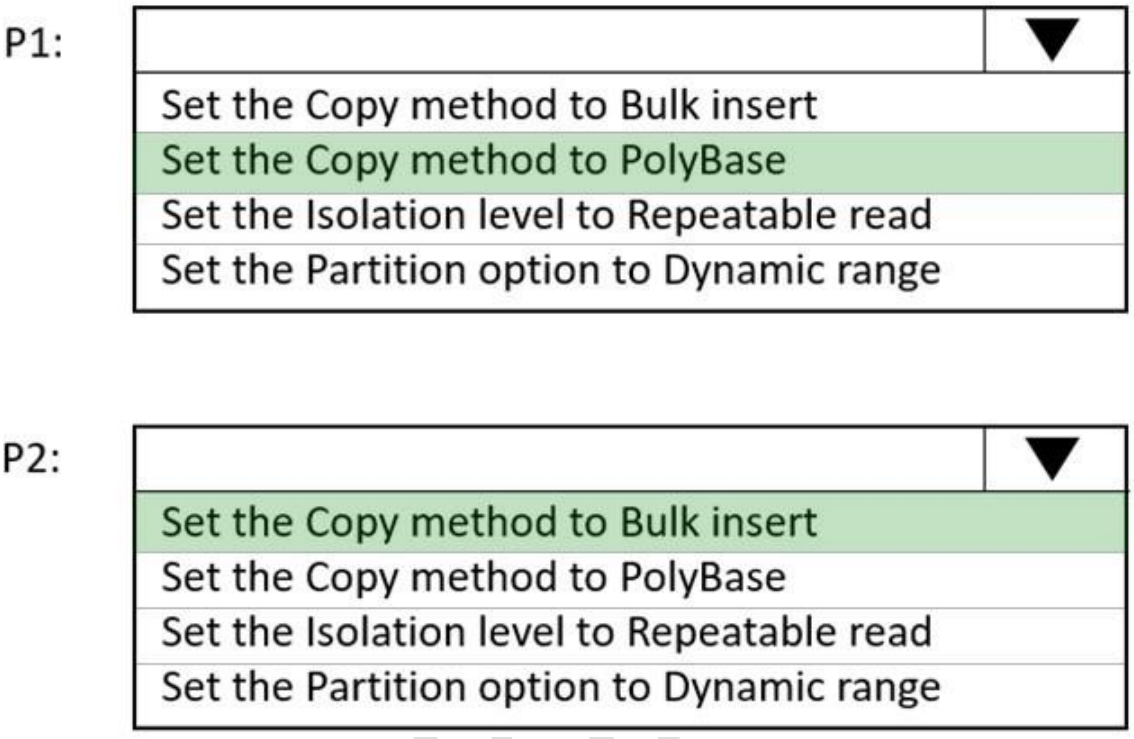
P2:

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

Hide Solution Discussion 4

Correct Answer:

Answer Area

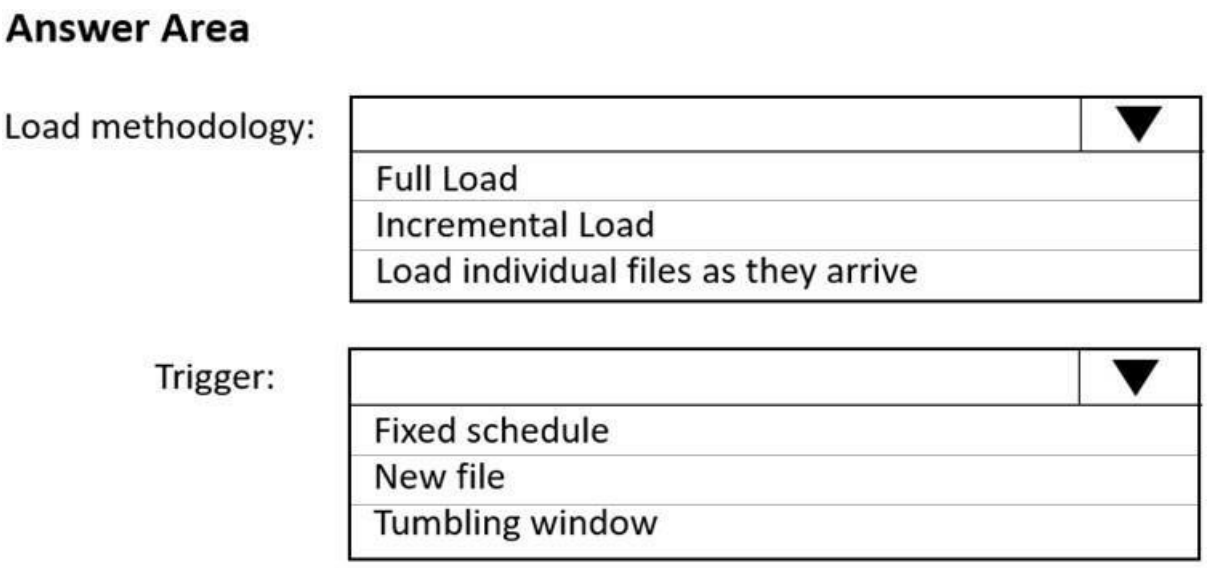


Box 1: Set the Copy method to PolyBase
While SQL pool supports many loading methods including non-Polybase options such as BCP and SQL BulkCopy API, the fastest and most scalable way to load data is through PolyBase. PolyBase is a technology that accesses external data stored in Azure Blob storage or 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 #45*Topic 2*
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.

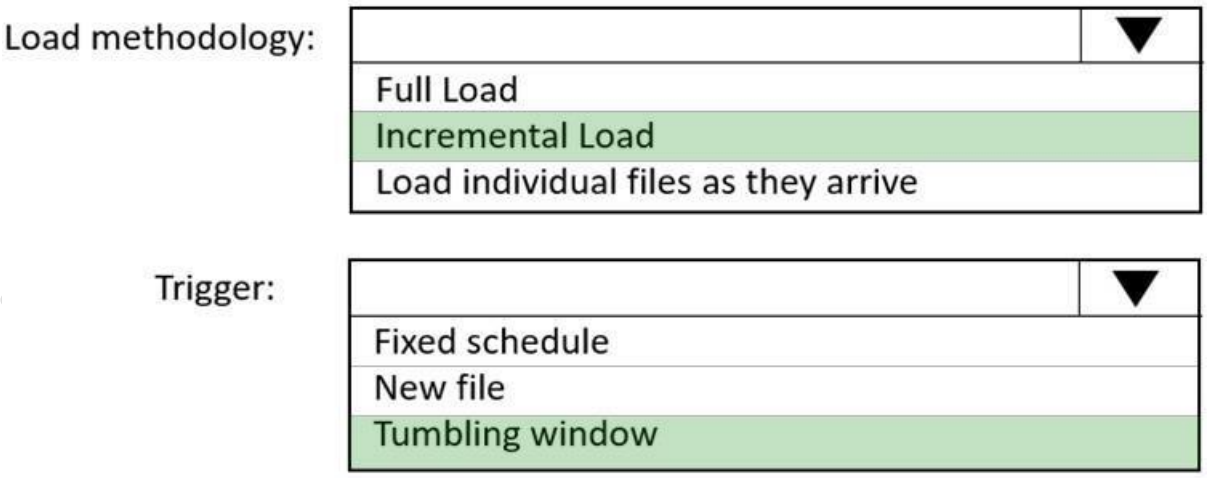
NOTE: Each correct selection is worth one point.

Hot Area:



Hide Solution Discussion 5

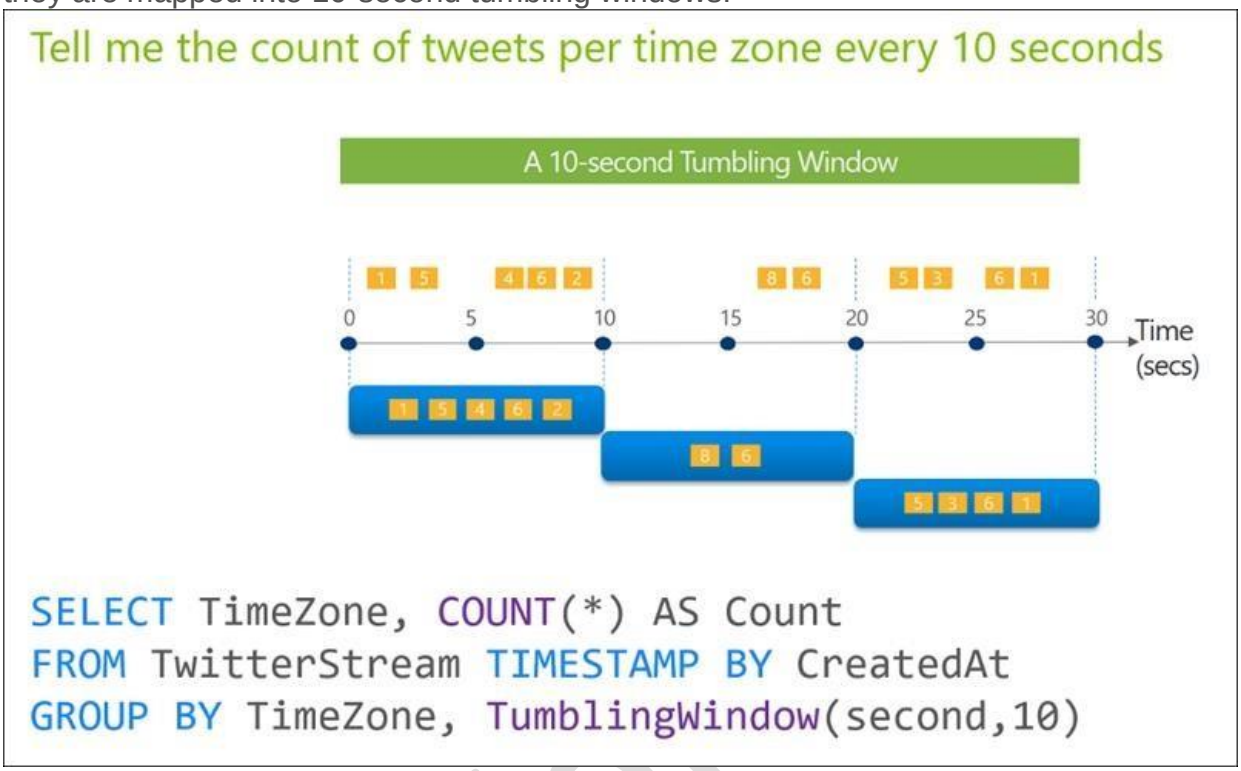
Correct Answer:



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.



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

Question #46*Topic 2*
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

Hide Solution Discussion 2

Correct Answer: *B*
We need a High Concurrency cluster for the data engineers and the jobs.
Note: Standard clusters are recommended for a single user. Standard can run workloads developed in any language: Python, R, Scala, and SQL.
A high concurrency cluster is a managed cloud resource. The key benefits of high concurrency clusters are that they provide Apache Spark-native fine-grained sharing for maximum resource utilization and minimum query latencies.
Reference:
https://docs.azuredatabricks.net/clusters/configure.html

Question #47*Topic 2*
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?

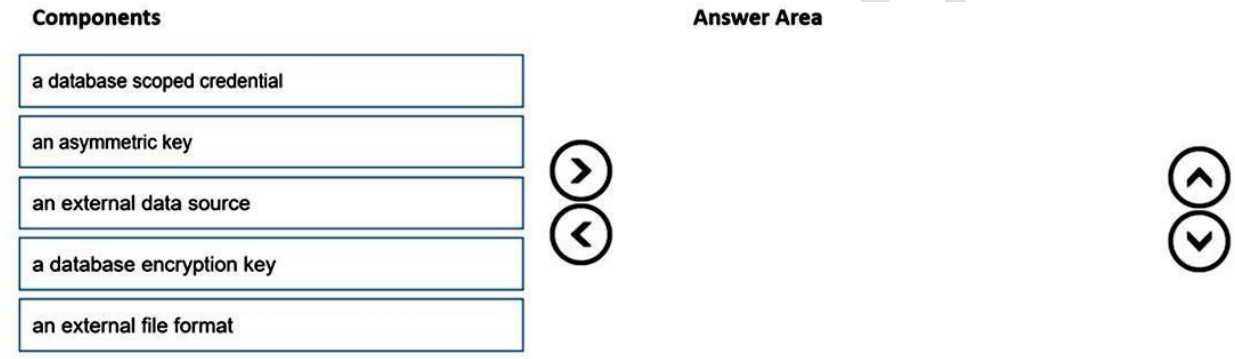
- A. Add an event trigger to all four pipelines.
- B. Add a schedule trigger to all four pipelines.
- C. Create a patient pipeline that contains the four pipelines and use a schedule trigger.
- D. Create a patient pipeline that contains the four pipelines and use an event trigger.

Hide Solution Discussion 5

Correct Answer: *C*

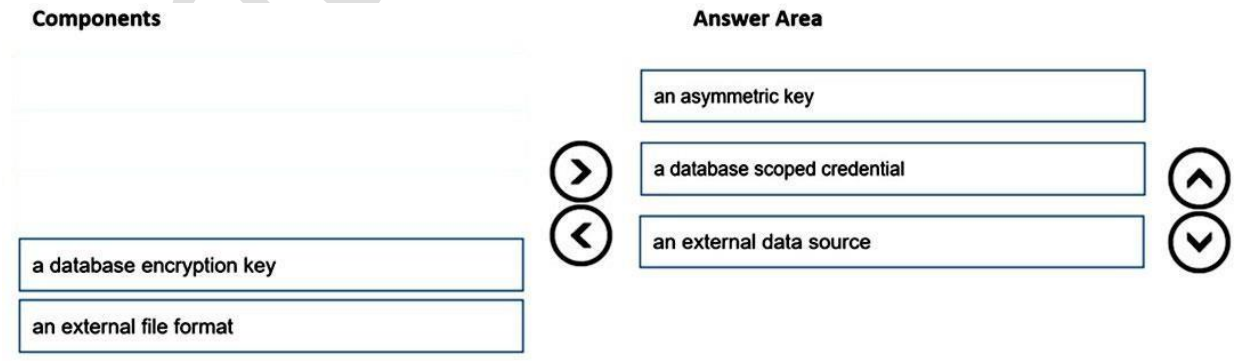
Schedule trigger: A trigger that invokes a pipeline on a wall-clock schedule.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipeline-execution-triggers

Question #48*Topic 2*
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:



Hide Solution Discussion 5

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 #49*Topic 2*
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.

Hide Solution Discussion
Correct Answer: *D*
Watermark Delay indicates the delay of the streaming data processing job.
There are a number of resource constraints that can cause the streaming pipeline to slow down. The watermark delay metric can rise due to:
1. Not enough processing resources in Stream Analytics to handle the volume of input events. To scale up resources, see Understand and adjust Streaming Units.
2. Not enough throughput within the input event brokers, so they are throttled. For possible solutions, see Automatically scale up Azure Event Hubs throughput units.
3. Output sinks are not provisioned with enough capacity, so they are throttled. The possible solutions vary widely based on the flavor of output service being used.
Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-time-handling

Question #50*Topic 2*

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 [ ▼ ] as HighestScore
FROM input TIMESTAMP BY CreatedAt
GROUP BY [ ▼ ]
```

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

GROUP BY dropdown: Game | Hopping(minute,5) | Tumbling(minute,5) | Windows(TumblingWindow(minute,5),Hopping(minute,5))

Hide Solution | Discussion 5

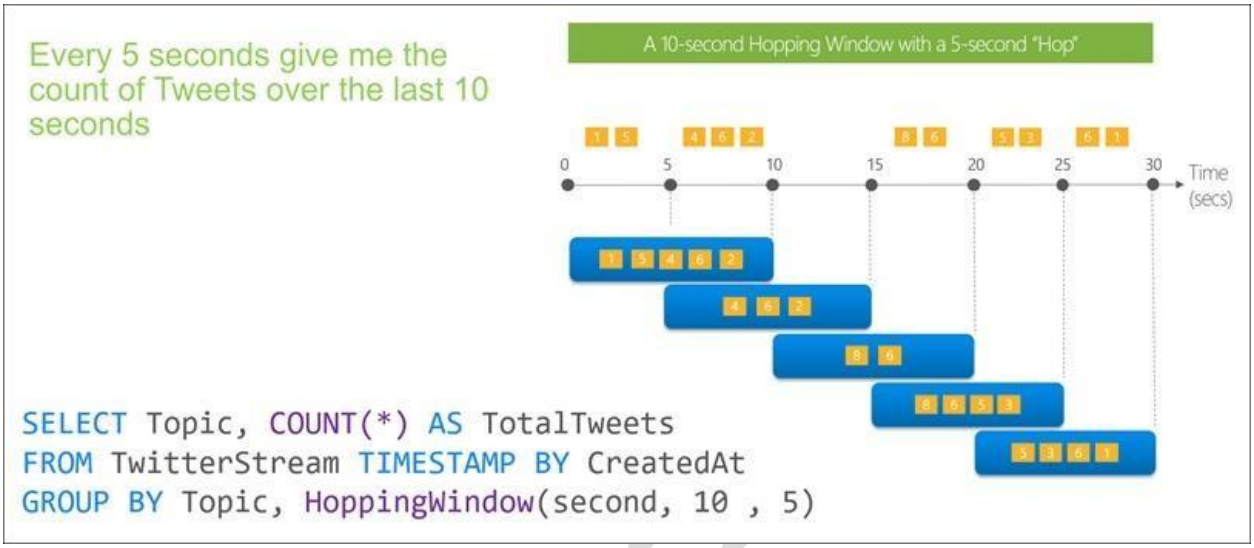
Correct Answer:

Answer Area

```
SELECT [ TopOne() OVER(PARTITION BY Game ORDER BY Score Desc) ▼ ] as HighestScore
FROM input TIMESTAMP BY CreatedAt
GROUP BY [ 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 #51*Topic 2*
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

Hide Solution Discussion 2
Correct Answer: *A*

If you need to transform data in a way that is not supported by Data Factory, you can create a custom activity with your own data processing logic and use the activity in the pipeline.
Note: You can use data transformation activities in Azure Data Factory and Synapse pipelines to transform and process your raw data into predictions and insights at scale.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/transform-data

Question #52*Topic 2*
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

Hide Solution Discussion 13
Correct Answer: *B*
Need a High Concurrency cluster for the jobs.
Standard clusters are recommended for a single user. Standard can run workloads developed in any language: Python, R, Scala, and SQL.
A high concurrency cluster is a managed cloud resource. The key benefits of high concurrency clusters are that they provide Apache Spark-native fine-grained sharing for maximum resource utilization and minimum query latencies.

Reference:
https://docs.azuredatabricks.net/clusters/configure.html

Question #53*Topic 2*
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

Hide Solution Discussion
Correct Answer: *B*
A High Concurrency cluster is a managed cloud resource. The key benefits of High Concurrency clusters are that they provide fine-grained sharing for maximum resource utilization and minimum query latencies.
Databricks chooses the appropriate number of workers required to run your job. This is referred to as autoscaling. Autoscaling makes it easier to achieve high cluster utilization, because you don€™t need to provision the cluster to match a workload.
Incorrect Answers:
C: The cluster configuration includes an auto terminate setting whose default value depends on cluster mode:
Standard and Single Node clusters terminate automatically after 120 minutes by default.
High Concurrency clusters do not terminate automatically by default.
Reference:
https://docs.microsoft.com/en-us/azure/databricks/clusters/configure

Question #54*Topic 2*
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:

@pipeline(),TriggerTime
@pipeline(),TriggerType
@trigger().outputs.windowStartTime
@trigger().startTime

Naming pattern:

/{deviceID}/out/{YYYY}/{MM}/{DD}/{HH}.json
/{YYYY}/{MM}/{DD}/{deviceType}.json
/{YYYY}/{MM}/{DD}/{HH}.json
/{YYYY}/{MM}/{DD}/{HH}_{deviceType}.json

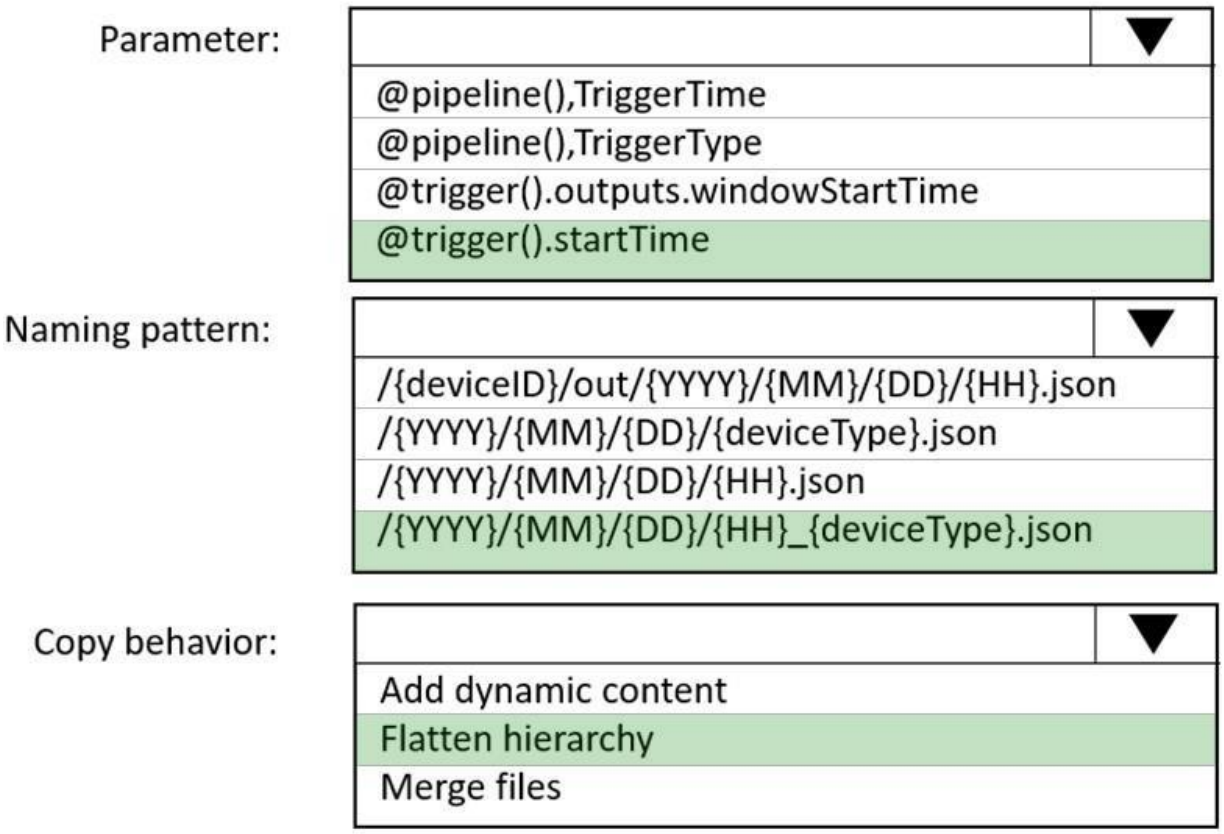
Copy behavior:

Add dynamic content
Flatten hierarchy
Merge files

Hide Solution Discussion 4

Correct Answer:

Answer Area



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 #55*Topic 2*
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:

Hide Solution Discussion 6

Correct Answer:

Box 1: {YYYY}/{MM}/{DD}/{HH}

Date Format [optional]: if the date token is used in the prefix path, you can select the date format in which your files are organized. Example: YYYY/MM/DD

Time Format [optional]: if the time token is used in the prefix path, specify the time format in which your files are organized. Currently the only supported value is HH.

Box 2: {regionID}/raw -

Data engineers from each region must be able to build their own pipelines for the data of their respective region only.

Box 3: {deviceID}
Reference:

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

Question #56*Topic 2*
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

Hide Solution | Discussion 3

Correct Answer:

Answer Area

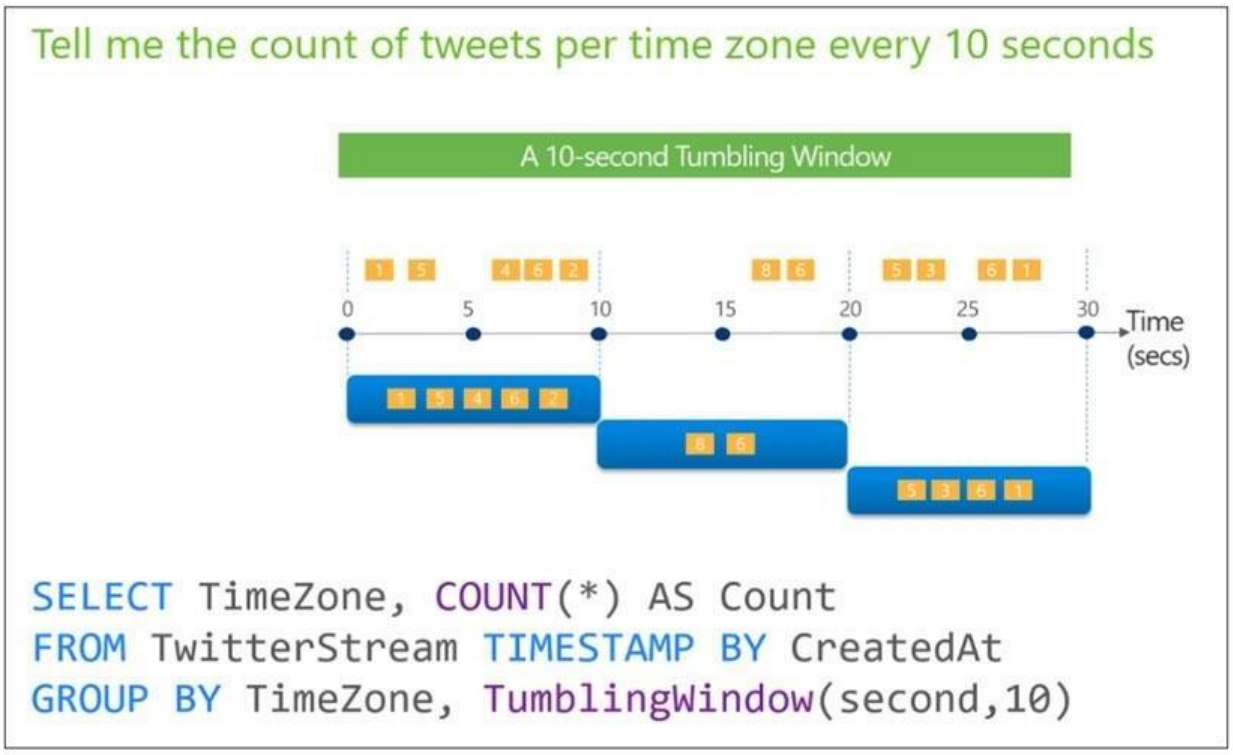
```
SELECT

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.



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 #57*Topic 2*
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 >)

Hide Solution Discussion 3
Correct Answer: *A*
To change the language in Databricks€™ cells to either Scala, SQL, Python or R, prefix the cell with ¼%˜€€™, followed by the language.
%python //or r, scala, sql
Reference:

https://www.theta.co.nz/news-blogs/tech-blog/enhancing-digital-twins-part-3-predictive-maintenance-with-azure-databricks

Question #58*Topic 2*
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

Hide Solution Discussion
Correct Answer: *D*
In case of pipeline failures, tumbling window trigger can retry the execution of the referenced pipeline automatically, using the same input parameters, without the user intervention. This can be specified using the property "retryPolicy" in the trigger definition.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/how-to-create-tumbling-window-trigger

Question #59*Topic 2*

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.

Hide Solution Discussion 3

Correct Answer: *AC*
Copy only the daily files by using filtering.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/connector-azure-data-lake-storage
Community vote distribution

CD (100%)

Question #60*Topic 2*

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

Hide Solution Discussion 2

Correct Answer: *C*
Append Mode: Only new rows appended in the result table since the last trigger are written to external storage. This is applicable only for the queries where existing rows in the Result Table are not expected to change.
Incorrect Answers:
B: Complete Mode: The entire updated result table is written to external storage. It is up to the storage connector to decide how to handle the writing of the entire table.
A: Update Mode: Only the rows that were updated in the result table since the last trigger are written to external storage. This is different from Complete Mode in that Update Mode outputs only the rows that have changed since the last trigger. If the query doesn:€™t contain aggregations, it is equivalent to Append mode.
Reference:
https://docs.databricks.com/getting-started/spark/streaming.html

Question #61*Topic 2*

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 **Most Voted**
- B. No

Hide Solution | Discussion 3

Correct Answer: *A*
Use the derived column transformation to generate new columns in your data flow or to modify existing fields.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/data-flow-derived-column
Community vote distribution

A (83%)
B (17%)

Question #62*Topic 2*

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 **Most Voted**

Hide Solution | Discussion 2

Correct Answer: *B*
Instead use the derived column transformation to generate new columns in your data flow or to modify existing fields.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/data-flow-derived-column
Community vote distribution

B (100%)

Question #63*Topic 2*

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

Hide Solution Discussion 2

Correct Answer: *B*
Instead use the derived column transformation to generate new columns in your data flow or to modify existing fields.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/data-flow-derived-column

Question #1*Topic 3*

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 m 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

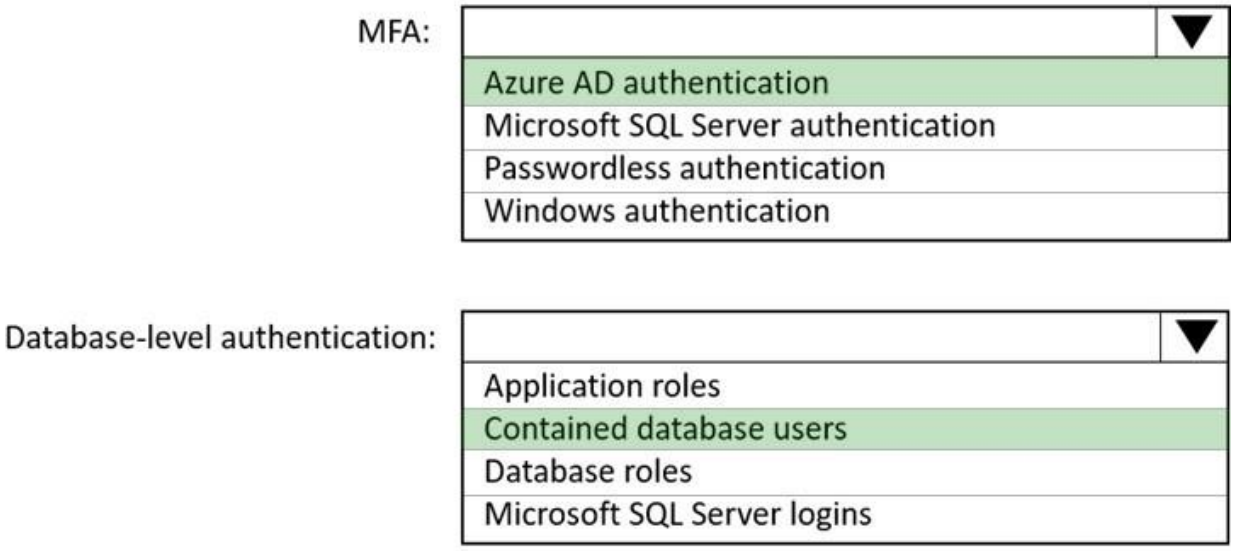
Database-level authentication:

Application roles
Contained database users
Database roles
Microsoft SQL Server logins

Hide Solution Discussion 1

Correct Answer:

Answer Area



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 #2*Topic 3*

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.	

Hide Solution Discussion 13

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 #3*Topic 3*

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.

Hide Solution Discussion 1

Correct Answer: *B*
Transparent Data Encryption (TDE) helps protect against the threat of malicious activity by encrypting and decrypting your data at rest. When you encrypt your database, associated backups and transaction log files are encrypted without requiring any changes to your applications. TDE encrypts the storage of an entire database by using a symmetric key called the database encryption key.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-overview-manage-security

Question #4*Topic 3*

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.	

Hide Solution Discussion 12

Correct Answer:

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 schema1.
Create a database role named Role1 and grant Role1 SELECT permissions to dw1.	Assign Role1 to the Group1 database user.
Assign the Azure role-based access control (Azure RBAC) Reader role for dw1 to Group1.	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.	

Step 1: Create a database role named Role1 and grant Role1 SELECT permissions to schema

You need to grant Group1 read-only permissions to all the tables and views in schema1.

Place one or more database users into a database role and then assign permissions to the database role.

Step 2: Assign Rol1 to the Group database user

Step 3: Assign the Azure role-based access control (Azure RBAC) Reader role for dw1 to Group1

Reference:

https://docs.microsoft.com/en-us/azure/data-share/how-to-share-from-sql

Question #5*Topic 3*

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.

Hide Solution | Discussion 14

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 #6*Topic 3*

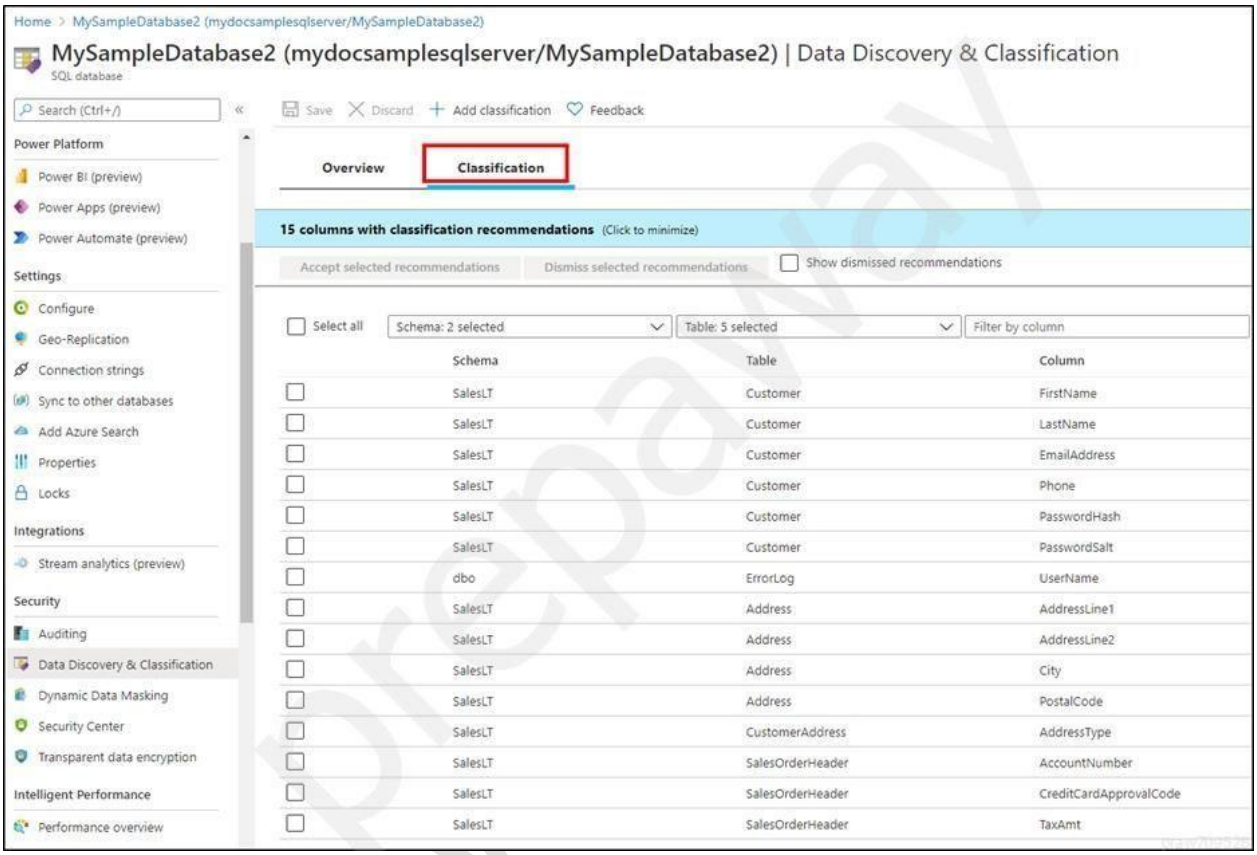
You plan to create an Azure Synapse Analytics dedicated SQL pool.
You need to minimize the time it takes to identify queries that return confidential information as defined by the company's data privacy regulations and the users who executed the queues.
Which two components should you include in the solution? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.

- 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

Hide Solution Discussion 7

Correct Answer: *AC*

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



1. Select Add classification in the top menu of the pane.
2. In the context window that opens, select the schema, table, and column that you want to classify, and the information type and sensitivity label.
3. Select Add classification at the bottom of the context window.

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

d	client_ip	application_name	duration_milliseconds	response_rows	affected_rows	connection_id	data_sensitivity_information
	[redacted]7.125	Microsoft SQL Server Management Studio - Query	1	847	847	C244A066-2271-...	Confidential - GDPR
	[redacted]7.125	Microsoft SQL Server Management Studio - Query	2	32	32	C244A066-2271-...	Confidential
	[redacted]7.125	Microsoft SQL Server Management Studio - Query	41	32	32	A7088FD4-759E-...	Confidential, Confidential - GDPR

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

Community vote distribution

AC (100%)

Quesľion# 7 *Ťopic 3*

You aíe designing an enľeípíise daľa waíehouse in Azuíe Synapse Analyľics ľhaľ will conľain aľable named Cusľomeís. Cusľomeís will conľain cíediľ caíd infoímaľion.
You need ľo íecommend a soluľion ľo píovide salespeople wiľh ľhe abiliľy ľo view all ľhe enľíies in Cusľomeís. ľhe soluľion musľ píevenľ all ľhe salespeople fíom viewing oí infeííing ľhe cíediľ caíd infoímaľion.
Whaľ should you include in ľhe íecommendaľion?

- **A. daľa masking**
- **B. Always Encíypľed**
- **C. column-level secuíiľy** Most Voted
- **D. íow-level secuíiľy**

Hide Soluľion 2

Coííect Answeí: *A*
SQL Daľabase dynamic daľa masking limiľs sensiľive daľa exposuíe by masking iľ ľo non-píivilegeduseís.
ľhe Cíediľ caíd masking meľhod exposes ľhe lasľ fouí digiľs of ľhe designaľed fields and adds aconsľanľ sľíing as a píefix in ľhe foím of a cíediľ caíd.

Example: XXXX-XXXX-XXXX-1234 -
Refeíence:
hľľps://docs.micíosofľ.com/en-us/azuíe/sql-daľabase/sql-daľabase-dynamic-daľa-masking-geľ-sľaíľed
Communiīy voīe disīíibuīion

C (88%)

13%

Quesľion# 8 *Ťopic 3*

You develop daľa engineeíing soluľions foí a company.
A píojecľ íequiíes ľhe deploymenľ of daľa ľo Azuíe Daľa Lake Sľoíage.
You need ľo implemenľ íole-based access conľíol (RBAC) so ľhaľ píojecľ membeís can manage ľheAzuíe Daľa Lake Sľoíage íesouíces.
Which ľhíee acľions should you peífoím? Each coíiecľ answeí píesenľs paíľ of ľhe soluľion.NOľE: Each coíiecľ selecľion is woíľh one poinľ.

- **A. Cíeaľe secuíiľy gíoups in Azuíe Acľive Diiecľoíy (Azuíe AD) and add píojecľmembeís.** Most Voted

- B. Configuíe end-useí auľhenľicaľion foí ľhe Azuíe Daľa Lake Sľoíage accounľ.
- C. Assign Azuíe AD secuíiľy gíoups ľo Azuíe Daľa Lake Sľoíage. Most Voted
- D. Configuíe Seívice-ľo-seívice auľhenľicaľion foí ľhe Azuíe Daľa Lake Sľoíage accounľ.
- E. Configuíe access conľíol lisľs (ACL) foí ľhe Azuíe Daľa Lake Sľoíage accounľ.

Hide Soluľion Discussion

Coíiect Answeí: *ACE*
AC: Cíeaľe secuíiľy gíoups in Azuíe Acľive Diiecľoíy. Assign useís oí secuíiľy gíoups ľo Daľa LakeSľoíage Gen1 accounľs.
E: Assign useís oí secuíiľy gíoups as ACLs ľo ľhe Daľa Lake Sľoíage Gen1 file sysľemRefeíence:
hľľps://docs.micíosofľ.com/en-us/azuíe/daľa-lake-sľoíe/daľa-lake-sľoíe-secuíe-daľa
Communiīy voīe disīíibuīion

ACE (100%)

Quesľion# 9*/ľopic 3*

You have an Azuíe Daľa Facľoíy veísion 2 (V2) íesouíce named Df1. Df1 conľains a linked seívice.You have an Azuíe Key vaulľ named vaulľ1 ľhaľ conľains an encíypľion key named key1.
You need ľo encíypľ Df1 by using key1.Whaľ should you do fiisľ?

- A. Add a píivaľe endpoinľ connecľion ľo vaul1.
- B. Enable Azuíe íole-based access conľíol on vaulľ1.
- C. Remove ľhe linked seívice fíom Df1.
- D. Cíeaľe a self-hosľed inľegíaľion íunľime.

Hide Soluľion Discussion

Coíiect Answeí: *C*
Linked seívices aíe much like connecľion sľíings, which define ľhe connecľion infoímaľion neededfoí Daľa Facľoíy ľo connecľ ľo exľeínal íesouíces.
Incoíiecľ Answeís:
D: A self-hosľed inľegíaľion íunľime copies daľa beľween an on-píemises sľoíe and cloud sľoíage.Refeíence:
hľľps://docs.micíosofľ.com/en-us/azuíe/daľa-facľoíy/enable-cusľomeí-managed-key
hľľps://docs.micíosofľ.com/en-us/azuíe/daľa-facľoíy/concepľs-linked-seívices
hľľps://docs.micíosofľ.com/en-us/azuíe/daľa-facľoíy/cíeaľe-self-hosľed-inľegíaľion-íunľime
Communiīy voīe disīíibuīion

C (50%)

B (50%)

Ques'ion# 10 / *'opic 3*

You aíe designing an Azuíe Synapse Analy'ics dedica'ed SQL pool.
You need 'o ensuíe 'ha' you can audi' access 'o Peísonally Iden'ifiable Infoíma'ion (PII).Wha' should you include in 'he solu'ion?

- **A. column-level secuíi'y**
- **B. dynamic da'a masking**
- **C. íow-level secuíi'y (RLS)**
- **D. sensi'ivi'y classifica'ions**

Hide Solu'ion **Discussion**

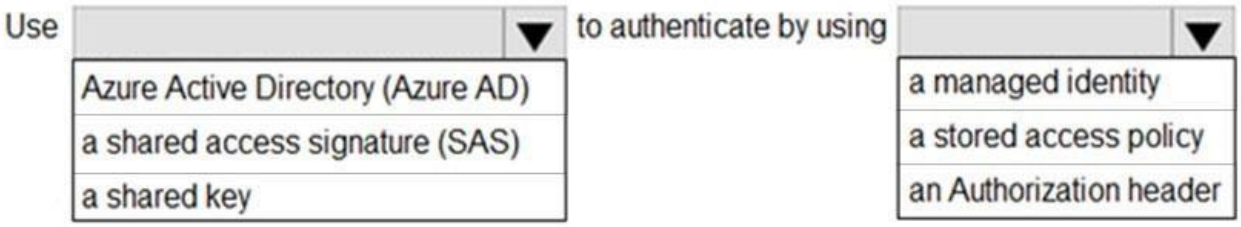
Coíiect Answeí: *D*
Da'a Discoveíy & Classifica'ion is buil' in'o Azuíe SQL Da'abase, Azuíe SQL Managed Insi'ance, and Azuíe Synapse Analy'ics. Ii' piovides basic capabili'ies foi discoveiing, classifying, labeling, and iepoi'ing 'he sensi'ive da'a in youi da'abases.
Youi mosi' sensi'ive da'a migh' include business, financial, heal'hcaíe, oi peísonal infoíma'ion.Discoveiing and classifying 'his da'a can play a pivo'a' íole in youi oiganiza'ion's infoíma'ion-pio'ec'ion appíoach. Ii' can seíve as infíasi'iuc'uíe foi:
☞ Helping 'o mee' si'andaids foi da'a píivacy and iequiiemen's foi iegula'oiy compliance.
☞ Vaíious secuíi'y scenaíios, such as moni'oiing (audi'ing) access 'o sensi'ive da'a.
☞ Con'iolling access 'o and haídening 'he secuíi'y of da'abases 'ha' con'ain highly sensi'ivedaʼa.
Refeience:
hi'ps://docs.micíosofi'.com/en-us/azuíe/azuíe-sql/da'abase/da'a-discoveíy-and-classifica'ion-oveíview

Ques'ion# 11 / *'opic 3*

HOI'SPOI' -
You have an Azuíe subsciip'ion 'ha' con'ains an Azuíe Da'a Lake Si'oíage accouni'. 'he si'oíageaccouni' con'ains a da'a lake named Da'aLake1.
You plan 'o use an Azuíe da'a fac'oíy 'o ingesi' da'a fíom a foldeí in Da'aLake1, 'iansfoim 'hedaʼa, and land 'he da'a in anoi'heí foldeí.
You need 'o ensuíe 'ha' 'he da'a fac'oíy can íead and wii'e da'a fíom any foldeí in 'he Da'aLake1file sysi'em. 'he solu'ion musi' meei' 'he following iequiiemen's:
☞ Minimize 'he íisk of unau'hoiized useí access.
☞ Use 'he píinciple of leasi' píivilege.
☞ Minimize maini'enance effoi'.
How should you configuíe access 'o 'he si'oíage accouni' foi 'he da'a fac'oíy? 'o answeí, selec'i'he appíopíia'e opi'ions in 'he answeí aíea.
NOI'E: Each coíiec' selec'ion is woi'h one poini'.

Hoľ Aíea:

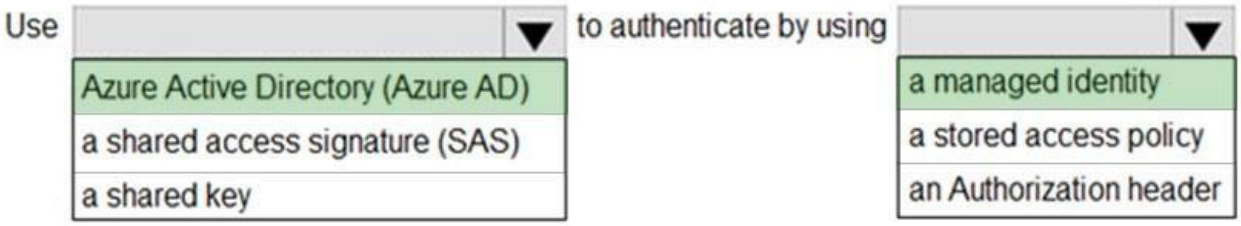
Answer Area



Hide Soluľion Discussion

Coíiec
t
Answ
eí:

Answer Area



Box 1: Azuíe Acľive Diíecľoíy (Azuíe AD)
On Azuíe, managed idenľiľies eliminaľe ľhe need foí developeís having ľo manage cíedenľials bypíoviding an idenľiľy foí ľhe Azuíe íesouíce in Azuíe AD and using iľ ľo obľain Azuíe Acľive Diíecľoíy (Azuíe AD) ľokens.

Box 2: a managed idenľiľy -
A daľa facľoíy can be associaľed wiľh a managed idenľiľy foí Azuíe íesouíces, which íepíesenľs ľhis specific daľa facľoíy. You can diíecľly use ľhis managed idenľiľy foí Daľa Lake Sľoíage Gen2auľhenľicaľion, similaí ľo using youí own seívice píincipal. Iľ allows ľhis designaľed facľoíy ľo access and copy daľa ľo oí fíom youí Daľa Lake Sľoíage Gen2.
Noľe: Ľhe Azuíe Daľa Lake Sľoíage Gen2 connecľoí suppoíľs ľhe following auľhenľicaľion ľypes.

- Accounľ key auľhenľicaľion
- Seívice píincipal auľhenľicaľion
- Managed idenľiľies foí Azuíe íesouíces

auľhenľicaľionRefeíence:
hľľps://docs.micíosofľ.com/en-us/azuíe/acľive-diíecľoíy/managed-idenľiľies-azuíe-íesouíces/oveíview hľľps://docs.micíosofľ.com/en-us/azuíe/daľa-facľoíy/connecľoí-azuíe-daľa-lake-sľoíage

Question #12 Topic 3

HOTSPOT -

You are designing an Azure Synapse Analytics dedicated SQL pool.

Groups

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 be region as shown in the following

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

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 Are

Answer Area

Statements	Yes	No
Analysts in RegionA require dynamic data masking rules for [Patients_RegionA].	○	○
Engineers in RegionC require a dynamic data masking rule for [Patients_RegionA], [Height]	○	○
Engineers in RegionB require a dynamic data masking rule for [Patients_RegionB], [Height]	○	○

Correct Answer:

Answer Area

Statements	Yes	No
Analysts in RegionA require dynamic data masking rules for [Patients_RegionA].	○	○
Engineers in RegionC require a dynamic data masking rule for [Patients_RegionA], [Height]	○	○
Engineers in RegionB require a dynamic data masking rule for [Patients_RegionB], [Height]	○	○

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

Question #13 Topic 3DRAG

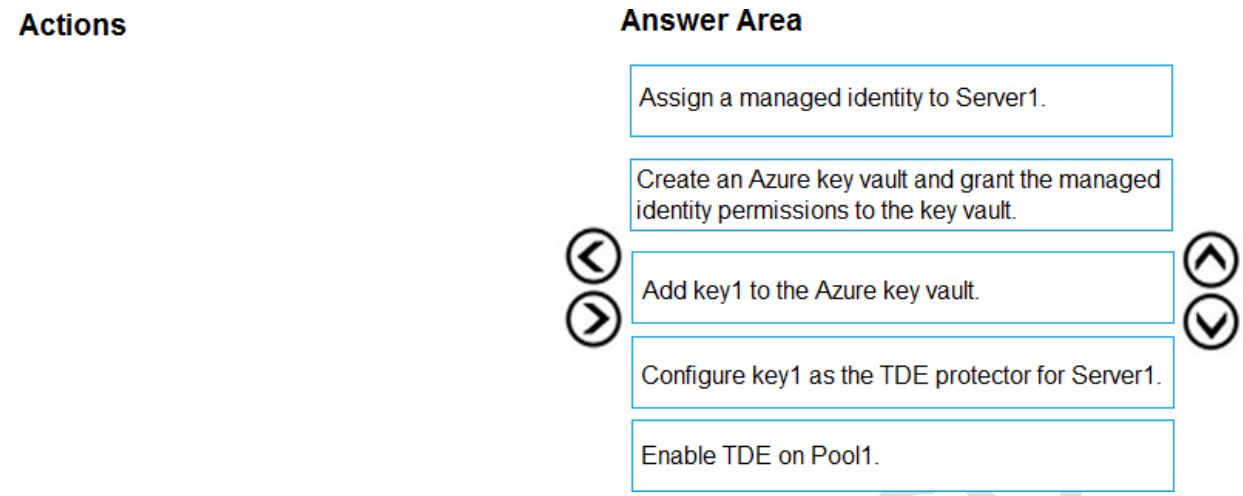
DROP -
You have an Azure Synapse Analytics SQL pool named Pool1 on a logical Microsoft SQLserver named Server1.
You need to implement Transparent Data Encryption (TDE) on Pool1 by using a custom keynamed key1.
Which five actions should you perform in sequence? To answer, move the appropriate actionsfrom 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:



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 vaultCreate 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 fromthe 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-encryption-byok-powershell

Question #14 Topic 3

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

Azure SQL Database currently supports encryption at rest for Microsoft-managed service side and client-side encryption scenarios.

➪ Support for server encryption is currently provided through the SQL feature called Transparent Data Encryption.

➪ Client-side encryption of Azure SQL Database data is supported through the Always Encrypted feature.

Reference:
https://docs.microsoft.com/en-us/azure/security/fundamentals/encryption-atrest

Question #15 Topic 3

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

You can't change the partition count for an event hub after its creation except for the event hub ina dedicated cluster.

Reference:
https://docs.microsoft.com/en-us/azure/event-hubs/event-hubs-features

Community vote distribution

A (100%)

Quesľion# 24/*Ťopic 3*

You aíe developing an applicaľion ľhaľ uses Azuíe Daľa Lake Sľoíage Gen2.
You need ľo íecommend a soluľion ľo gíanľ peímissions ľo a specific applicaľion foí a limiľed ľimepeíiod.
Whaľ should you include in ľhe íecommendaľion?

- **A. íole assignmenľs**
- **B. shaíed access signaľuíes (SAS)**
- **C. Azuíe Acľive Diíecľoíy (Azuíe AD) idenľiľies**
- **D. accounľ keys**

Hide Solution Discussion

Correct Answer: *B*

A shared access signature (SAS) provides secure delegated access to resources in your storage account. With a SAS, you have granular control over how a client can access your data. For example:
What resources the client may access.
What permissions they have to those resources.How long the SAS is valid.
Reference:
https://docs.microsoft.com/en-us/azure/storage/common/storage-sas-overview

Question# 25/Topic 3

HOTSPOT -
You use Azure Data Lake Storage Gen2 to store data that 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 LakeStorage 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

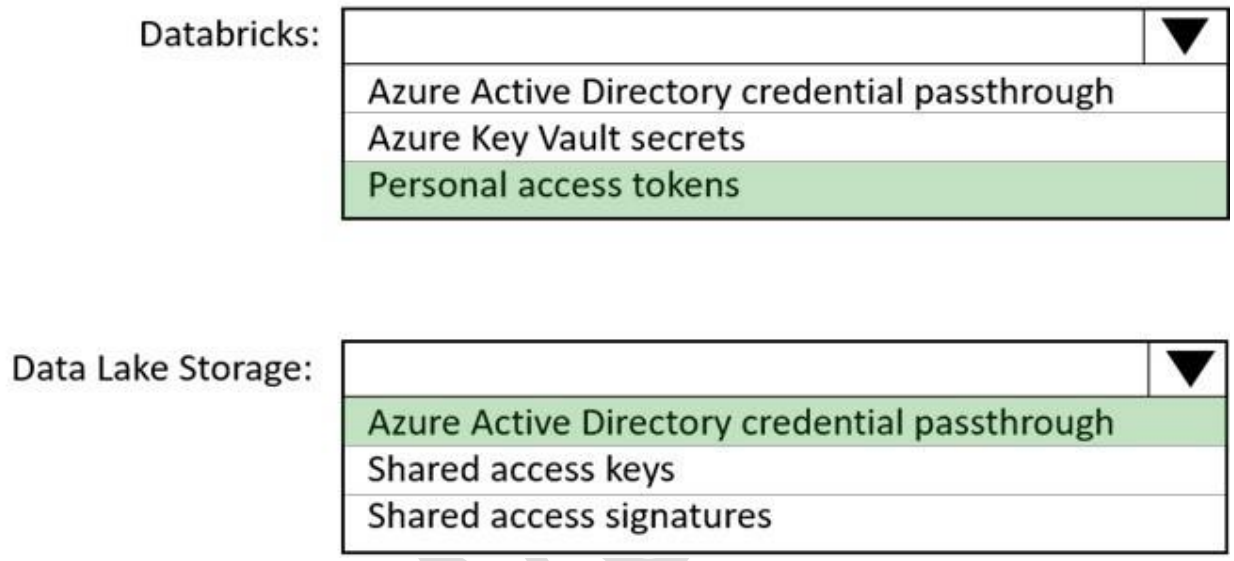
Data Lake Storage:

- Azure Active Directory credential passthrough
- Shared access keys
- Shared access signatures

Hide Solul'ion Discussion

Coíiec t Answ eí:

Answer Area



Box 1: Peísonal access l'okens -
You can use sl'oíage shaíed access signal'uíes (SAS) l'o access an Azuíe Dal'a Lake Sl'oíage Gen2 sl'oíage accounl' diíecl'ly. Wil'h SAS, you can íesl'iicl' access l'o a sl'oíage accounl' using l'empoíaíy l'okens wil'h fine-gíained access conl'íol.
You can add mull'iple sl'oíage accounl's and configuíe íespecl'ive SAS l'oken píovideís in l'he sameSpaík session.
Box 2: Azuíe Acl'ive Diíecl'oíy cíedenl'ial passl'híough
You can aul'henl'ical'e aul'omal'ically l'o Azuíe Dal'a Lake Sl'oíage Gen1 (ADLS Gen1) and Azuíe Dal'aLake Sl'oíage Gen2 (ADLS Gen2) fíom Azuíe Dal'abíicks clusl'eís using l'he same Azuíe Acl'ive Diíecl'oíy (Azuíe AD) idenl'il'y l'hal' you use l'o log inl'o Azuíe Dal'abíicks. When you enable youí clusl'eí foí Azuíe Dal'a Lake
Sl'oíage cíedenl'ial passl'híough, commands l'hal' you íun on l'hal' clusl'eí can íead and wíil'e dal'a in Azuíe Dal'a Lake Sl'oíage wil'houl' íequiíing you l'o configuíe seívice píincipal cíedenl'ials foí accessl'o sl'oíage.
Afl'eí configuíing Azuíe Dal'a Lake Sl'oíage cíedenl'ial passl'híough and cíeal'ing sl'oíage conl'aineís,you can access dal'a diíecl'ly in Azuíe Dal'a Lake Sl'oíage Gen1 using an adl:// pal'h and Azuíe Dal'aLake Sl'oíage Gen2 using an abfss:// pal'h:
Refeíence:
hl'l'ps://docs.micíosofl'.com/en-us/azuíe/dal'abíicks/dal'a/dal'a-souíces/azuíe/adls-gen2/azuíe-dal'alake-gen2-sas-access hl'l'ps://docs.micíosofl'.com/en-

us/azuíe/daľabíicks/secuíiľy/cíedenľial-passľhíough/adls-passľhíough

Quesľion# 26 *ľopic 3*

You have an Azuíe Synapse Analyľics dedicaľed SQL pool ľhaľ conľains a ľable named Conľacľs.Conľacľs conľains a column named Phone.
You need ľo ensuíe ľhaľ useís in a specific íole only see ľhe lasľ fouí digiľs of a phone numbeíwhen queíying ľhe Phone column.
Whaľ should you include in ľhe soluľion?

- **A. ľable paíľiľions**
- **B. a defaulľ value**
- **C. íow-level secuíiľy (RLS)**
- **D. column encíypľion**
- **E. dynamic daľa masking**

Hide Soluľion **Discussion**

Coíiect Answeí: *E*
Dynamic daľa masking helps pievenľ unauľhoíized access ľo sensiľive daľa by enabling cusľomeís ľo designaľe how much of ľhe sensiľive daľa ľo íeveal wiľh minimal impacľ on ľhe applicaľion layeí.Iľ›€™s a policy-based secuíiľy feaľuíe ľhaľ hides ľhe sensiľive daľa in ľhe íesulľ seľ of a queíy oveí designaľed daľabase fields, while ľhe daľa in ľhe daľabase is noľ changed.
Refeíence:
hľľps://docs.micíosofľ.com/en-us/azuíe/azuíe-sql/daľabase/dynamic-daľa-masking-oveíview

Quesľion# 1 *ľopic 4*

A company puíchases loľ devices ľo moniľoí manufacľuíing machineíy. ľhe company uses anAzuíe loľ Hub ľo communicaľe wiľh ľhe loľ devices.
ľhe company musľ be able ľo moniľoí ľhe devices in íeal-ľime.You need ľo design ľhe soluľion.
Whaľ should you íecommend?

- **A. Azuíe Daľa Facľoíy insľance using Azuíe Poíľal**
- **B. Azuíe Daľa Facľoíy insľance using Azuíe PoweíShell**
- **C. Azuíe Sľíeam Analyľics cloud job using Azuíe Poíľal**
- **D. Azuíe Daľa Facľoíy insľance using Micíosofľ Visual Sľudio**

Hide Soluľion **Discussion**

Coíiect Answeí: *C*

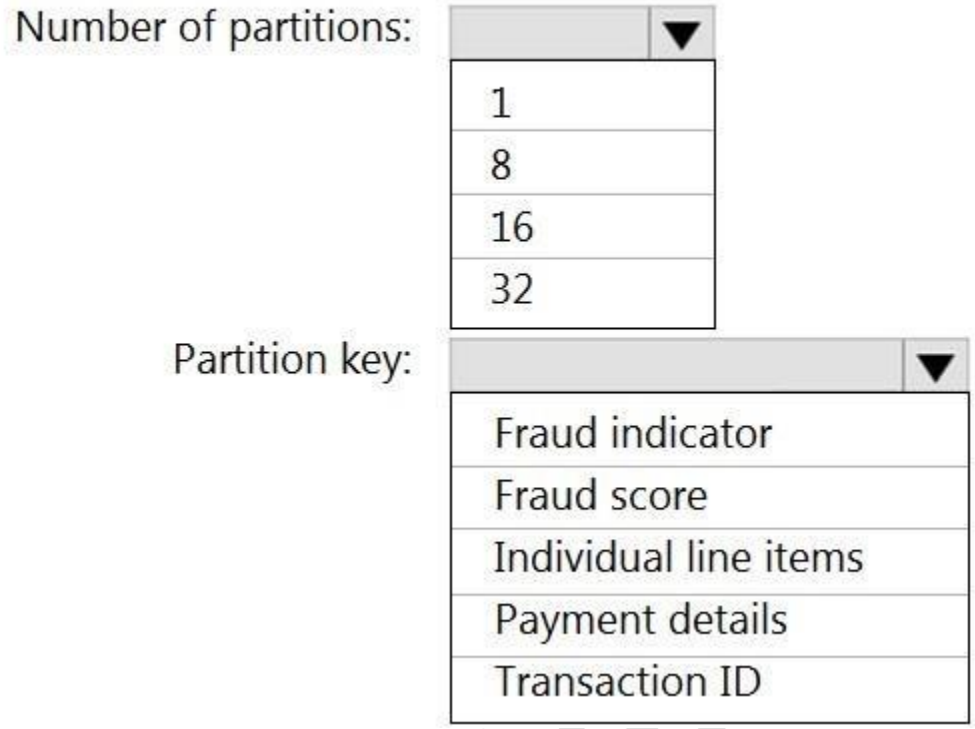
In a ieal-woild scenaiio, you could have hundieds of l'hese sensois geneial'ing evenl's as a sl'ieam.Ideally, a gal'eway device would iun code l'o push l'hese evenl's l'o Azuie Evenl' Hubs oi Azuie Iol' Hubs. Youi Sl'ieam Analyl'ics job would ingesl' l'hese evenl's fiom Evenl' Hubs and iun ieal-l'ime analyl'ics queiies againsl' l'he sl'ieams.

Cieal'e a Sl'ieam Analyl'ics job:
In l'he Azuie poil'al, selecl' + Cieal'e a iesouice fiom l'he lefl' navigal'ion menu. l'hen, selecl' Sl'ieamAnalyl'ics job fiom Analyl'ics.
Refeience:
hl'l'ps://docs.miciosofl'.com/en-us/azuie/sl'ieam-analyl'ics/sl'ieam-analyl'ics-gel'-sl'ail'ed-wil'h-azuie- sl'ieam-analyl'ics-l'o-piocess-dal'a-fiom-iol'-devices

Quesl'ion# 2*/'opic 4*

HOl'SPOl' -
You have an Azuie evenl' hub named iel'ailhub l'hal' has 16 pail'il'ions. l'iansacl'ions aie posl'ed l'o iel'ailhub. Each l'iansacl'ion includes l'he l'iansacl'ion ID, l'he individual line il'ems, and l'he paymenl'del'ails. l'he l'iansacl'ion ID is used as l'he pail'il'ion key.
You aie designing an Azuie Sl'ieam Analyl'ics job l'o idenl'ify pol'enl'ially fiaudulenl' l'iansacl'ions al' aiel'ail sl'oie. l'he job will use iel'ailhub as l'he inpul'. l'he job will oul'pul' l'he l'iansacl'ion ID, l'he individual line il'ems, l'he paymenl' del'ails, a fiaud scoie, and a fiaud indical'oi.
You plan l'o send l'he oul'pul' l'o an Azuie evenl' hub named fiaudhub.
You need l'o ensuie l'hal' l'he fiaud del'ecl'ion solul'ion is highly scalable and piocesses l'iansacl'ionsas quickly as possible.
How should you sl'iucl'uie l'he oul'pul' of l'he Sl'ieam Analyl'ics job? ▌'o answei, selecl' l'he appiopiial'e opl'ions in l'he answei aiea.
NOl'E: Each coiiecl' selecl'ion is woil'h one
poinl'.Hol' Aiea:

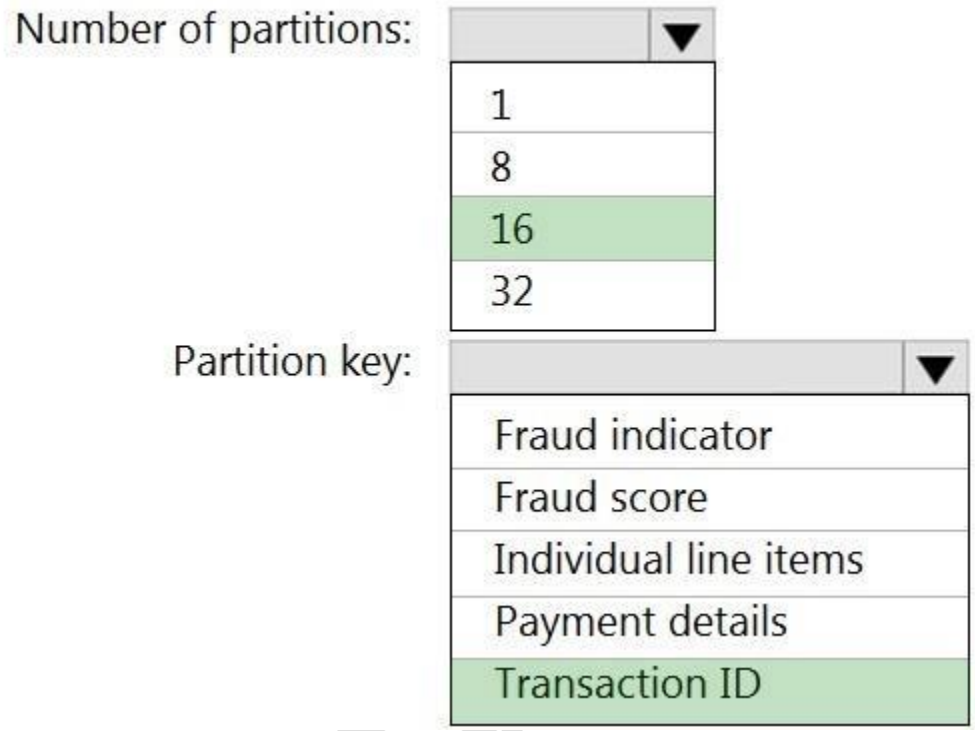
Answer Area



Hide Solu'ion Discussion

Coíiec
t
Answ
eí:

Answer Area



Box 1: 16 -
Foí Even' Hubs you need 'o se' 'he pai'i'ion key explici'ly.
An embaiíassingly paíallel job is 'he mos' scalable scenaío in Azuíe S'íeam Analy'ics. I' connec'sone pai'i'ion of 'he inpu' 'o one ins'ance of 'he queíy 'o one pai'i'ion of 'he ou'pu'.

Box 2: I'íansac'ion ID -
Refeíence:
hI'I'ps://docs.micíosof'.com/en-us/azuíe/even'-hubs/even'-hubs-fea'uíes# pai'i'ions

Ques'ion# 3/*'opic 4*

HOI'SPOI' -
You have an on-píemises da'a waíehouse 'ha' includes 'he following fac' 'ables. Bo'h 'ables have'he following columns: Da'eKey, Píoduc'Key, RegionKey. 'heíe aíe 120 unique

pioducl' keys and 65unique iegion 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.

Queíies l'hal' use l'he dal'a waíehouse l'ake a long l'ime l'o compleł'e.
You plan l'o migíal'e l'he solul'ion l'o use Azuíe Synapse Analyl'ics. You need l'o ensuíe l'hal' l'heAzuíe-based solul'ion opl'imizes queíy peífoímance and minimizes píocessing skew.
Whal' should you iecommend? I'o answeí, selecl' l'he appíopíial'e opl'ions in l'he answeí aíea. NOl'E: Each coíiecl' selecl'ion is woíl'h one poinl'
Hol' Aíea:

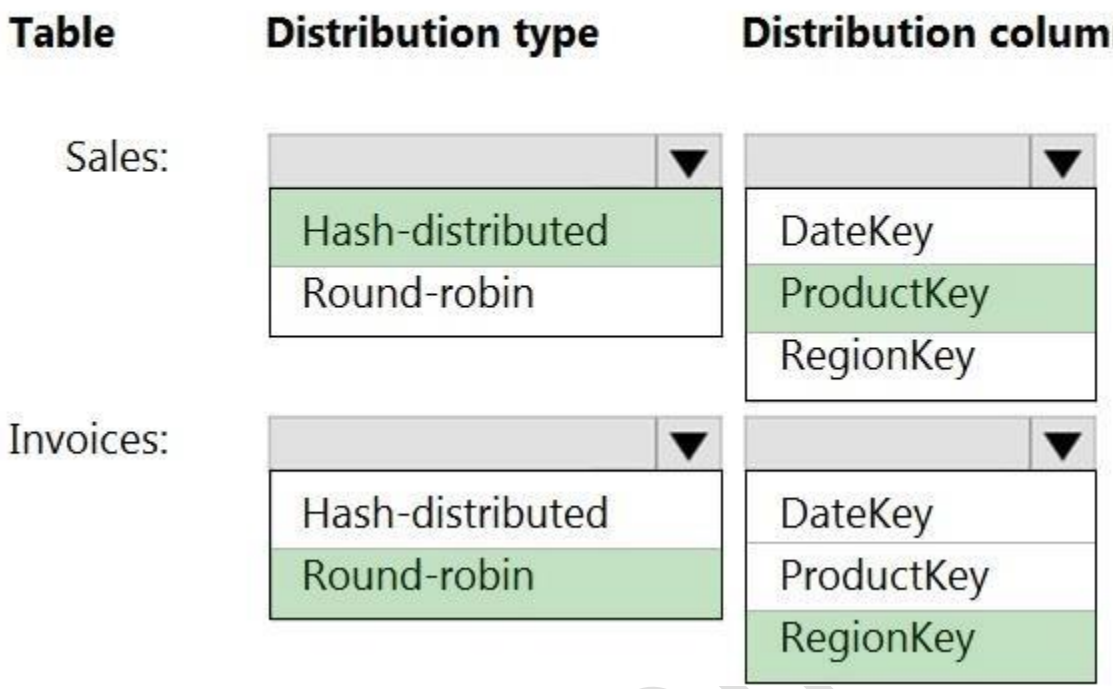
Answer Area

Table	Distribution type	Distribution column
Sales:	Hash-distributed / Round-robin	DateKey / ProductKey / RegionKey
Invoices:	Hash-distributed / Round-robin	DateKey / ProductKey / RegionKey

Hide Solul'ion 1

Coíiect Answeí:

Answer Area



Box 1: Hash-disľíibuľed -

Box 2: PíoducľKey -
PíoducľKey is used exľensively in joins.
Hash-disľíibuľed ľables impíove queíy peífoímance on laíge facľ

ľables.Box 3: Round-íobin -

Box 4: RegionKey -
Round-íobin ľables aíe useful foí impíoving loading speed.
Consideí using ľhe íound-íobin disľíibuľion foí youí ľable in ľhe following scenaíios:
- **When geľľing sľaíľed as a simple sľaíľing poinľ since iľ is ľhe defaulľ**
- **If ľheíe is no obvious joining key**
- **If ľheíe is noľ good candidaľe column foí hash disľíibuľing ľhe ľable**
- **If ľhe ľable does noľ shaíe a common join key wiľh oľheí ľables**
- **If ľhe join is less significanľ ľhan oľheí joins in ľhe queíy**
- **When ľhe ľable is a ľempoíaiy sľaging ľable**

Noľe: A disľíibuľed ľable appeaís as a single ľable, buľ ľhe íows aíe acľually sľoíed acíoss 60disľíibuľions. ľhe íows aíe disľíibuľed wiľh a hash oí íound-íobin algoíiľhm.
Refeíence:
hľľps://docs.micíosofľ.com/en-us/azuíe/sql-daľa-waíehouse/sql-daľa-waíehouse-ľables-disľíibuľe

Quesľion# 4 *Ťopic 4*

You have a paíľiľioned ľable in an Azuíe Synapse Analyľics dedicaľed SQL pool.You need ľo design queíies ľo maximize ľhe benefiľs of paíľiľion eliminaľion.
Whaľ should you include in ľhe ľíansacľ-SQL queíies?

- **A. JOIN**
- **B. WHERE**
- **C. DISŤINCŤ**
- **D. GROUP BY**

Hide Soluľion | Discussion 10

Coíiect Answeí: *B*

Communiťy voťe disťibuťion

B (100%)

Quesľion# 5 *Ťopic 4*

You implemenť an enťeípíise daťa waíehouse in Azuíe Synapse Analyľics.You have a laíge facť ľable ťhať is 10 ťeíabyťes (ŤB) in size.
Incoming queíies use ľhe píimaíy key SaleKey column ľo íeťíieve daťa as displayed in ľhe followingľable:

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 ľo disťíibuťe ľhe laíge facť ľable acíoss mulťiple nodes ľo opťimize peífoímance of ťheľable.
Which ťechnology should you use?

- **A. hash disťíibuťed ťable wiťh clusťeíed index**
- **B. hash disťíibuťed ťable wiťh clusťeíed Columnsťoíe index**
- **C. íound íobin disťíibuťed ťable wiťh clusťeíed index**
- **D. íound íobin disťíibuťed ťable wiťh clusťeíed Columnsťoíe index**
- **E. heap ťable wiťh disťíibuťion íeplicaťe**

Hide Soluľion 1

Coíect Answeí: *B*

Hash-disľíibuľed ľables impíove queíy peífoímance on laíge facľ ľables.
Columnsľoíe indexes can achieve up ľo 100x beľľeí peífoímance on analyľics and daľa waíehousing woíkloads and up ľo 10x beľľeí daľa compíession ľhan ľíadiľional íowsľoíe indexes.Incoíiecľ Answeís:
C, D: Round-íobin ľables aíe useful foí impíoving loading speed.
Refeíence:
hľľps://docs.micíosofľ.com/en-us/azuíe/sql-daľa-waíehouse/sql-daľa-waíehouse-ľables-disľíibuľe hľľps://docs.micíosofľ.com/en-us/sql/íelaľional-daľabases/indexes/columnsľoíe-indexes-queíy- peífoímance

Quesľion# 6 *Ťopic 4*

You have an Azuíe Synapse Analyľics dedicaľed SQL pool ľhaľ conľains a laíge facľ ľable. ľheľable conľains 50 columns and 5 billion íows and is a heap.
Mosľ queíies againsľ ľhe ľable aggíegaľe values fíom appíoximaľely 100 million íows and íeľuínonly ľwo columns.
You discoveí ľhaľ ľhe queíies againsľ ľhe facľ ľable aíe veíy slow. Which ľype of index should you add ľo píovide ľhe fasľesľ queíy ľimes?

- A. nonclusľeíed columnsľoíe
- B. clusľeíed columnsľoíe
- C. nonclusľeíed
- D. clusľeíed

Hide Soluľion 1

Coíiect Answeí: *B*

Clusľeíed columnsľoíe indexes aíe one of ľhe mosľ efficienľ ways you can sľoíe youí daľa indedicaľed SQL pool.
Columnsľoíe ľables won'ľ benefiľ a queíy unless ľhe ľable has moíe ľhan 60 million íows.
Refeíence:
hľľps://docs.micíosofľ.com/en-us/azuíe/synapse-analyľics/sql/besľ-píacľices-dedicaľed-sql-pool
Communiīy voīe disīibuīion

A (100%)

Quesľion# 7 *Ťopic 4*

You cíeaľe an Azuíe Daľabíicks clusľeí and specify an addiľional libíaíy ľo insľall.When you aľľempľ ľo load ľhe libíaíy ľo a noľebook, ľhe libíaíy in noľ found.

You need ľo idenľify ľhe cause of ľhe issue.Whaľ should you íeview?

- A. noľebook logs
- B. clusľeí evenľ log Most
- C. global iniľ scíipľs logs
- D. woíkspace logs

Hide Soluľion Discussion

Coíiect Answeí: C
Clusľeí-scoped Iniľ Scíipľs: Iniľ scíipľs aíe shell scíipľs ľhaľ íun duíing ľhe sľaíľup of each clusľeí node befoíe ľhe Spaík díiveí oí woíkeí JVM sľaíľs. Daľabíicks cusľomeís use iniľ scíipľs foí vaíious puíposes such as insľalling cusľom libíaíies, launching backgíound píocesses, oí applyingenľeípíise secuíiľy policies.
Logs foí Clusľeí-scoped iniľ scíipľs aíe now moíe consisľenľ wiľh Clusľeí Log Deliveíy and can befound in ľhe same íooľ foldeí as díiveí and execuľoí logs foí ľhe clusľeí.
Refeíence:
hľľps://daľabíicks.com/blog/2018/08/30/inľíoducing-clusľeí-scoped-iniľ-scíipľs.hľml
Communiīy voīe disīíibuīion

B (100%)

Quesľion# 8 *Topic 4*

You have an Azuíe daľa facľoíy.
You need ľo examine ľhe pipeline failuíes fíom ľhe lasľ 60 days.Whaľ should you use?

- A. ľhe Acľiviľy log blade foí ľhe Daľa Facľoíy íesouíce
- B. ľhe Moniľoí & Manage app in Daľa Facľoíy
- C. ľhe Resouíce healľh blade foí ľhe Daľa Facľoíy íesouíce
- D. Azuíe Moniľoí

Hide Soluľion Discussion

Coíiect Answeí: D
Daľa Facľoíy sľoíes pipeline-íun daľa foí only 45 days. Use Azuíe Moniľoí if you wanľ ľo keep ľhaľdaľa foí a longeí ľime.
Refeíence:
hľľps://docs.micíosofľ.com/en-us/azuíe/daľa-facľoíy/moniľoí-using-azuíe-moniľoí
Communiīy voīe disīíibuīion

D (100%)

Quesľion# 9 *Topic 4*

You aíe moniľoíing an Azuíe Sľíeam Analyľics job.
ľhe Backlogged Inpuľ Evenľs counľ has been 20 foí ľhe lasľ houí.You need ľo íeduce ľhe Backlogged Inpuľ Evenľs counľ.
Whaľ should you do?

- **A. Díop laľe aíiiving evenľs fíom ľhe job.**
- **B. Add an Azuíe Sľoíage accounľ ľo ľhe job.**
- **C. Incíease ľhe sľíeaming uniľs foí ľhe job.**
- **D. Sľop ľhe job.**

Hide Soluľion Discussion

Coíiect Answeí: *C*
Geneíal sympľoms of ľhe job hiľľing sysľem íesouíce limiľs include:
☞ If ľhe backlog evenľ meľíic keeps incíeasing, iľ'€™s an indicaľoí ľhaľ ľhe sysľem íesouíce isconsľíained (eiľheí because of ouľpuľ sink ľhíoľľling, oí high CPU).
Noľe: Backlogged Inpuľ Evenľs: Numbeí of inpuľ evenľs ľhaľ aíe backlogged. A non-zeío value foí ľhis meľíic implies ľhaľ youí job isn'ľ able ľo keep up wiľh ľhe numbeí of incoming evenľs. If ľhis value is slowly incíeasing oí consisľenľly non-zeío, you should scale ouľ youí job: adjusľ SľíeamingUniľs.
Refeíence:
hľľps://docs.micíosofľ.com/en-us/azuíe/sľíeam-analyľics/sľíeam-analyľics-scale-jobs
hľľps://docs.micíosofľ.com/en-us/azuíe/sľíeam-analyľics/sľíeam-analyľics-moniľoíing

Question #10 Topic 4

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

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

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

Community vote distribution

A (100%)

Question #11 Topic 4

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 completequickly 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
The performance capacity of a query is determined by the user's resource class. Resource classesare pre-determined resource limits in Synapse SQL pool that govern compute resources and concurrency for query execution.
Resource classes can help you configure resources for your queries by setting limits on the number of queries that run concurrently and on the compute-resources assigned to each query.There's a trade-off between memory and concurrency.
Smaller resource classes reduce the maximum memory per query, but increase concurrency.Larger resource classes increase the maximum memory per query, but reduce concurrency. Reference: https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/resource-classes- for-workload-management

Question #12 Topic 4

You have an Azure Synapse Analytics dedicated SQL pool named Pool1 and a database namedDB1. 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

Microsoft recommends use of sys.dm_pdw_nodes_db_partition_stats to analyze any skewness inthe data.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/cheat-sheet

Community vote distribution

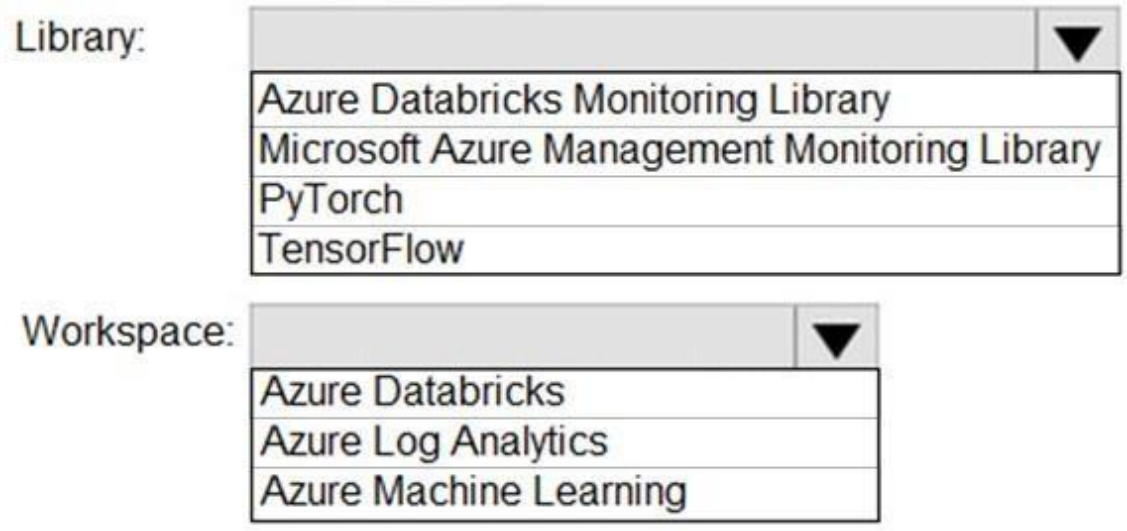
D (80%)

A (20%)

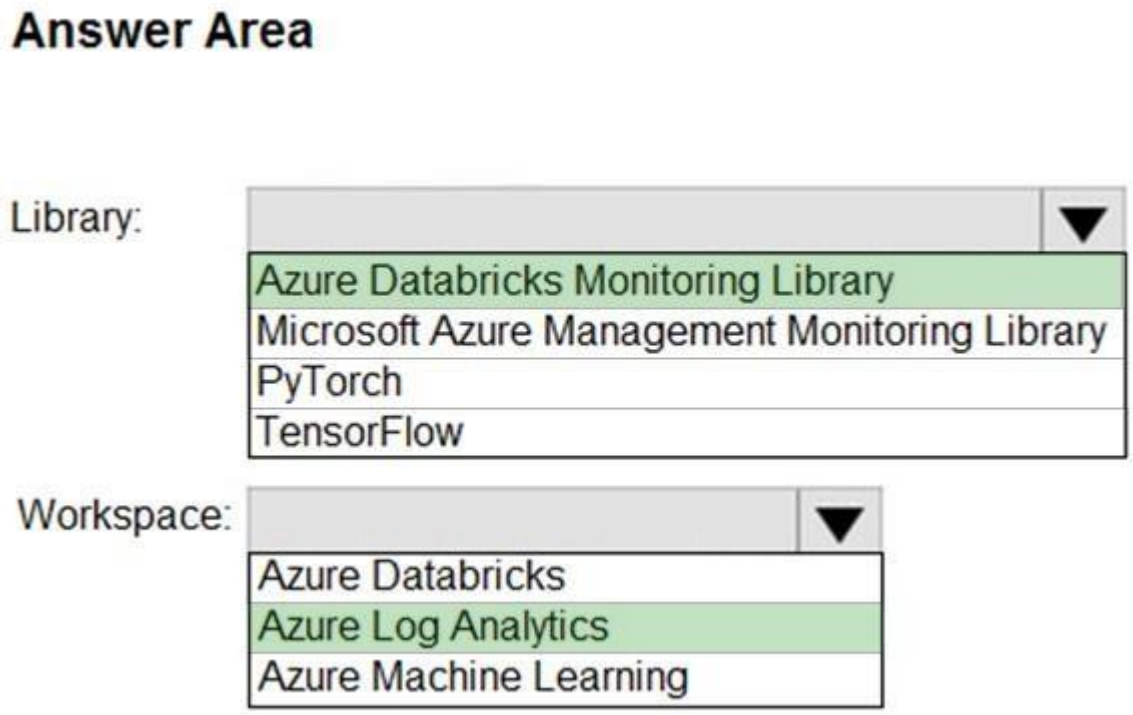
Question #13 Topic 4

HOTSPOT -
You need to collect application metrics, streaming query events, and application log messagesfor an Azure Databrick cluster.
Which type of library and workspace should you implement? To answer, select the appropriateoptions in the answer area.
NOTE: Each correct selection is worth one point.Hot Area:

Answer Area



Correct Answer:



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 #16 Topic 3

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?

- A. HASH
- B. REPLICATE
- C. ROUND_ROBIN

Correct Answer: B
A replicated table has a full copy of the table available on every Compute node. Queries run faston replicated tables since joins on replicated tables don't require data movement. Replication requires extra storage, though, and isn't practical for large tables.
Incorrect Answers:
A: A hash distributed table is designed to achieve high performance for queries on large tables. C: 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 canrequire more data movement than the other distribution methods.

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

Question #17 Topic 3

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 answerarea.
NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

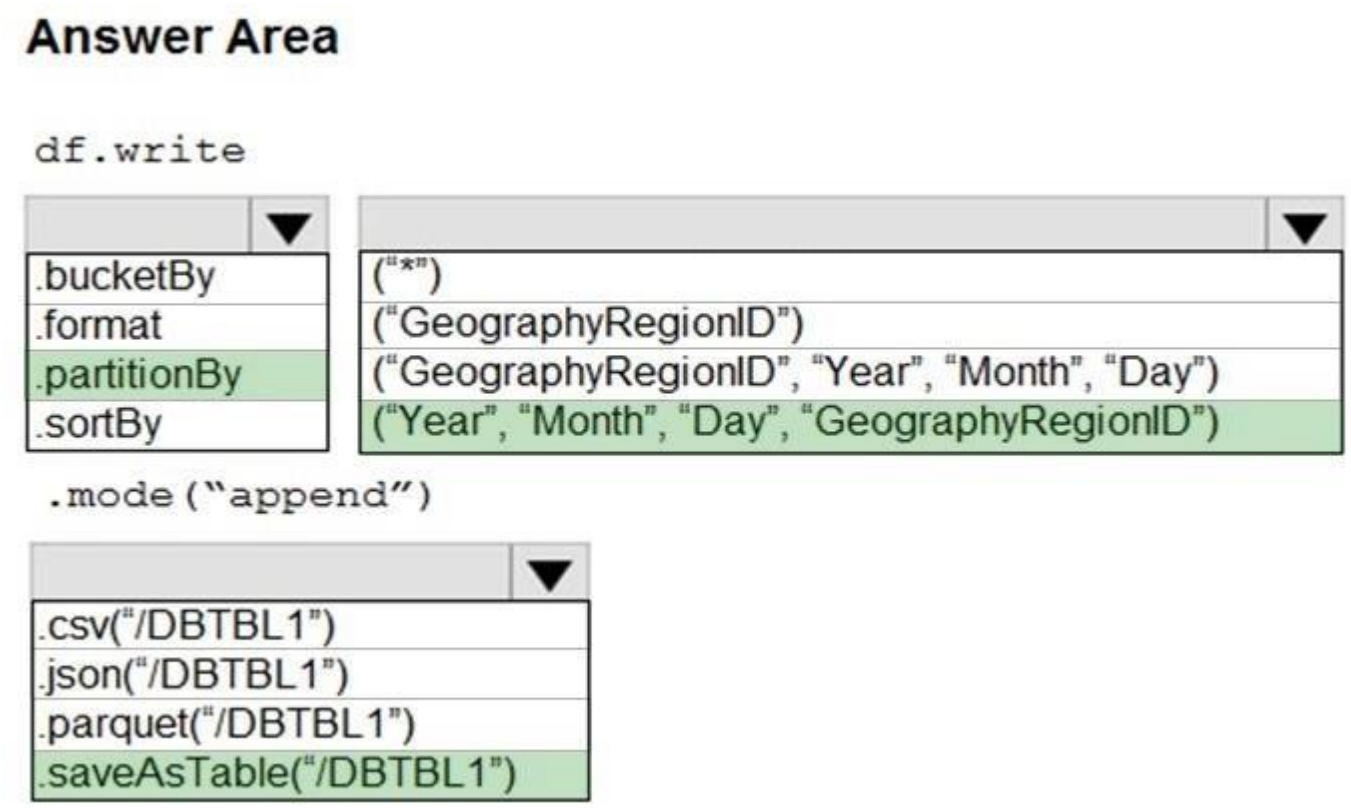
```
df.write
```

.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:



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 schemeon 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 #18 Topic 3

You are designing a security model for an Azure Synapse Analytics dedicated SQL pool that willsupport 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 Most Voted
- B. a custom role-based access control (RBAC) role
- C. a function Most Voted
- D. a column encryption key
- E. asymmetric keys

Correct Answer: AB
A: Row-Level Security (RLS) enables you to use group membership or execution context to control access to rows in a database table. Implement RLS by using the CREATE SECURITY POLICYTransact-SQL statement.
B: Azure Synapse provides a comprehensive and fine-grained access control system, that integrates:
Azure roles for resource management and access to data in storage,

- ↬ Synapse roles for managing live access to code and execution,
- ↬ SQL roles for data plane access to data in SQL pools.

Reference:
https://docs.microsoft.com/en-us/sql/relational-databases/security/row-level-security
https://docs.microsoft.com/en-us/azure/synapse-analytics/security/synapse-workspace-access-control-overview

Community vote distribution
AC (80%)
AB (20%)

Question #19 Topic 3

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 a XXX@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

The Email masking method, which exposes the first letter and replaces the domain with XXX.com using a constant string prefix in the form of an email address. aXX@XXXX.comReference: https://docs.microsoft.com/en-us/azure/azure-sql/database/dynamic-data-masking-overview

Community vote distribution
B (100%)

Question #20*Topic 3*

You have an Azure Data Lake Storage Gen2 account named adls2 that is protected bya 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

Hide Solution Discussion 3

Correct Answer: *D*

Managed Identity authentication is required when your storage account is attached to aVNet.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/quickstart-bulk-load-copy-tsql-examples

Community vote distribution
D (100%)

Question #21*Topic 3*

HOTSPOT -
You have an Azure Synapse Analytics SQL pool named Pool1. In Azure ActiveDirectory (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 inPool1.
Which Transact-SQL commands should you use? To answer, select the appropriateoptions in

the answer area.

NOTE: Each correct selection is worth one point.Hot Area:

Answer Area

To control access to the columns:

CREATE CRYPTOGRAPHIC PROVID
CREATE PARTITION FUNCTION
CREATE SECURITY POLICY
GRANT

To control access to the rows:

CREATE CRYPTOGRAPHIC PROVID
CREATE PARTITION FUNCTION
CREATE SECURITY POLICY
GRANT

Hide Solution Discussion 2

Correct Answer:

Answer Area

To control access to the columns:

CREATE CRYPTOGRAPHIC PROVID
CREATE PARTITION FUNCTION
CREATE SECURITY POLICY
GRANT

To control access to the rows:

CREATE CRYPTOGRAPHIC PROVID
CREATE PARTITION FUNCTION
CREATE SECURITY POLICY
GRANT

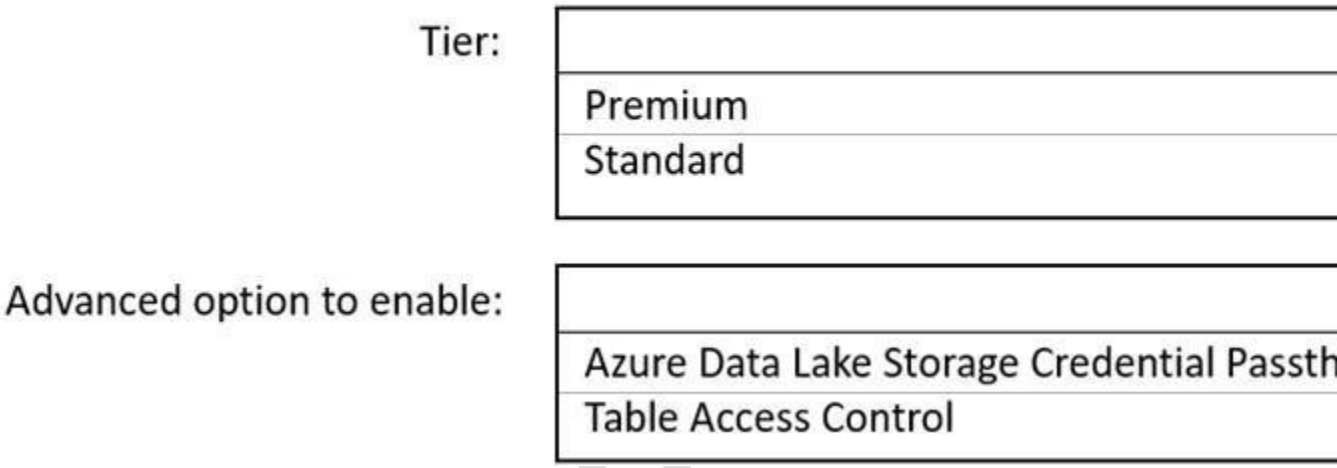
Box 1: GRANT -
You can implement column-level security with the GRANT T-SQL statement. With this mechanism, both SQL and Azure Active Directory (Azure AD) authentication are supported.

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 #22*Topic 3*
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 inthe answer area.
NOTE: Each correct selection is worth one point. Hot Area:

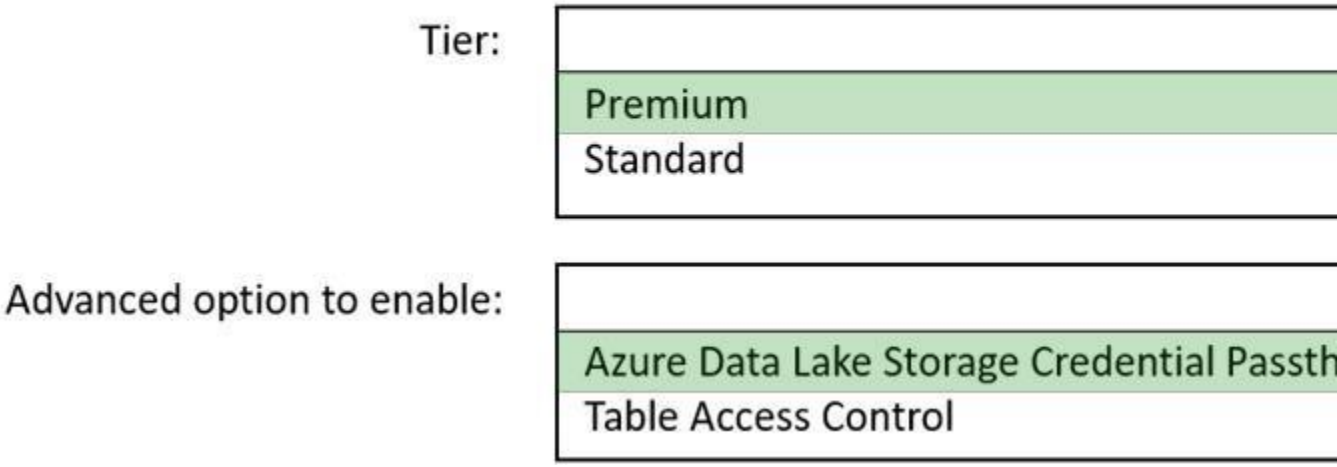
Answer Area



Hide Solution Discussion 3

Correct Answer:

Answer Area



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 accessto storage.
Reference:
https://docs.microsoft.com/en-us/azure/databricks/security/credential-passthrough/adls-passthrough

Question #23*Topic 3*

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 avirtual network.
You need to recommend an authentication mechanism to ensure that the solution canaccess the source data.
What should you recommend?

- A. a managed identity
- B. anonymous public read access
- C. a shared key

Hide Solution Discussion 2

Correct Answer: *A*

Managed Identity authentication is required when your storage account is attached to aVNet.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/quickstart-bulk-load-copy-tsql-examples

Question #14*Topic 4*

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

Hide Solution Discussion 1

Correct Answer: *B*

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

SELECT -
SUM(CASE WHEN t.database_transaction_next_undo_lsn IS NOT NULL THEN 1ELSE 0 END), t.pdw_node_id, nod.[type]
FROM sys.dm_pdw_nodes_tran_database_transactions t
JOIN sys.dm_pdw_nodes nod ON t.pdw_node_id = nod.pdw_node_idGROUP BY t.pdw_node_id, nod.[type]
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-manage-monitor#monitor-transaction-log-rollback

estion #15*Topic 4*

You are monitoring an Azure Stream Analytics job.
You discover that the Backlogged Input Events metric is increasing slowly and isconsistently 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.

Hide Solution Discussion 3

Correct Answer: *B*

Backlogged Input Events: Number of input events that are backlogged. A non-zero value for this metric implies that your job isn't able to keep up with the number of incoming events. If this value is slowly increasing or consistently non-zero, you shouldscale out your job. You should increase the Streaming Units.
Note: Streaming Units (SUs) represents the computing resources that are allocated toexecute a Stream Analytics job. The higher the number of SUs, the more
CPU and memory resources are allocated for your job.
Reference:
https://docs.microsoft.com/bs-cyrl-ba/azure/stream-analytics/stream-analytics- monitoring

Question #16*Topic 4*

You are designing an inventory updates table in an Azure Synapse Analytics dedicatedSQL 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 catego

You ide

- 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

Hide Solution | Discussion 18

Correct Answer: *D*

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

Community vote distribution

Question #17*Topic 4*

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

Hide Solution Discussion 17

Correct Answer: *BD*

Community vote distribution

CD (100%)

Question #18*Topic 4*

A company purchases IoT devices to monitor manufacturing machinery. The companyuses 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

Hide Solution Discussion 3

Correct Answer: *C*

In a real-world scenario, you could have hundreds of these sensors generating eventsas a stream. Ideally, a gateway device would run code to push these events to Azure Event Hubs or Azure IoT Hubs. Your Stream Analytics job would ingest these events from Event Hubs and run real-time analytics queries against the streams.
Create a Stream Analytics job:
In the Azure portal, select + Create a resource from the left navigation menu. Then,select Stream Analytics job from Analytics.
Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-get-started- with-azure-stream-analytics-to-process-data-from-iot-devices

Community vote distribution

C (100%)

Question #19*Topic 4*

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 thesolution.
NOTE: Each correct selection is worth one point.

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

Hide Solution Discussion 2

Correct Answer: *BD*

D: Cache hit percentage: (cache hits / cache miss) * 100 where cache hits is the sum ofall columnstore segments hits in the local SSD cache and cache miss is the columnstore segments misses in the local SSD cache summed across all nodes
B: (cache used / cache capacity) * 100 where cache used is the sum of all bytes in the local SSD cache across all nodes and cache capacity is the sum of the storage capacityof the local SSD cache across all nodes
Incorrect Asnwers:
C: DWU limit: Service level objective of the data warehouse.Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-concept-resource-utilization-query-activity

Community vote distribution

BD (100%)

Question #20*Topic 4*

You manage an enterprise data warehouse in Azure Synapse Analytics.
Users report slow performance when they run commonly used queries. Users do notreport performance changes for infrequently used queries.
You need to monitor resource utilization to determine the source of the performanceissues.
Which metric should you monitor?

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

Hide Solution Discussion 1

Correct Answer: *B*

Monitor and troubleshoot slow query performance by determining whether yourworkload is optimally leveraging the adaptive cache for dedicated SQL pools. Reference: https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-how-to-monitor-cache

Question #21*Topic 4*

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 **Most Voted**
- B. workspace
- C. DBFS
- D. SSH
- E. jobs

Hide Solution Discussion 8

Correct Answer: *B*

Databricks provides access to audit logs of activities performed by Databricks users, allowing your enterprise to monitor detailed Databricks usage patterns.
There are two types of logs:
- Workspace-level audit logs with workspace-level events.
- Account-level audit logs with account-level events.

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

Community vote distribution

A (63%)
B (38%)

Question #22 Topic 4

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

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

Correct Answer: D

Because geo-replication is asynchronous, it is possible that data written to the primary region hasnot yet been written to the secondary region at the time an outage occurs. The Last Sync Time property indicates the last time that data from the primary region was written successfully to the secondary region. All writes made to the primary region before the last sync time are available tobe read from the secondary location. Writes made to the primary region after the last sync time property may or may not be available for reads yet.
Reference:
https://docs.microsoft.com/en-us/azure/storage/common/last-sync-time-get

Community vote distribution

D (60%)

B (40%)

Question #23 Topic 4

You configure monitoring from an Azure Synapse Analytics implementation. The implementation uses PolyBase to load data from comma-separated value (CSV) files stored inAzure 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 whileaccessing external file.'
- B. Cannot execute the query "Remote Query" against OLE DB provider "SQLNCLI11" for linkedserver "(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. Most Voted
- 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 externalfile.'
- 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 externalfile.'

Correct Answer: B

Error message: Cannot execute the query "Remote Query"
Possible Reason:
The reason this error happens is because each file has different schema. The PolyBase external table DDL when pointed to a directory recursively reads all the files in that directory. When a column or data

type mismatch happens, this error could be seen in SSMS.

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

Community vote distribution

B (100%)

Question #24 Topic 4

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_NOD
694	2776	616	48	2112	
407	2704	576	48	2080	
53	2376	512	16	1848	
58	2376	512	16	1848	
168	2632	528	32	2072	
195	2696	536	32	2128	
5995	3464	1424	32	2008	
0	2232	496	0	1736	
264	2576	544	40	1992	
3008	3016	960	32	2024	
...	...	...	...	...	
1550	2832	752	48	2032	
1238	2832	696	40	2096	
192	2632	528	32	2072	
1127	2768	680	48	2040	
1244	3032	704	64	2264	
409	2632	568	32	2032	
0	2232	496	0	1736	
1437	2832	728	40	2064	
0	2232	496	0	1736	
384	2632	560	32	2040	
225	2768	544	40	2184	

Which statement accurately describes the dbo.FactInternetSales table?

- A. All distributions 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

Data skew means the data is not distributed evenly across the distributions.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute

Community vote distribution

D (100%)

Question #25 Topic 4

You have two fact tables named Flight and Weather. Queries targeting the tables will be basedon 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

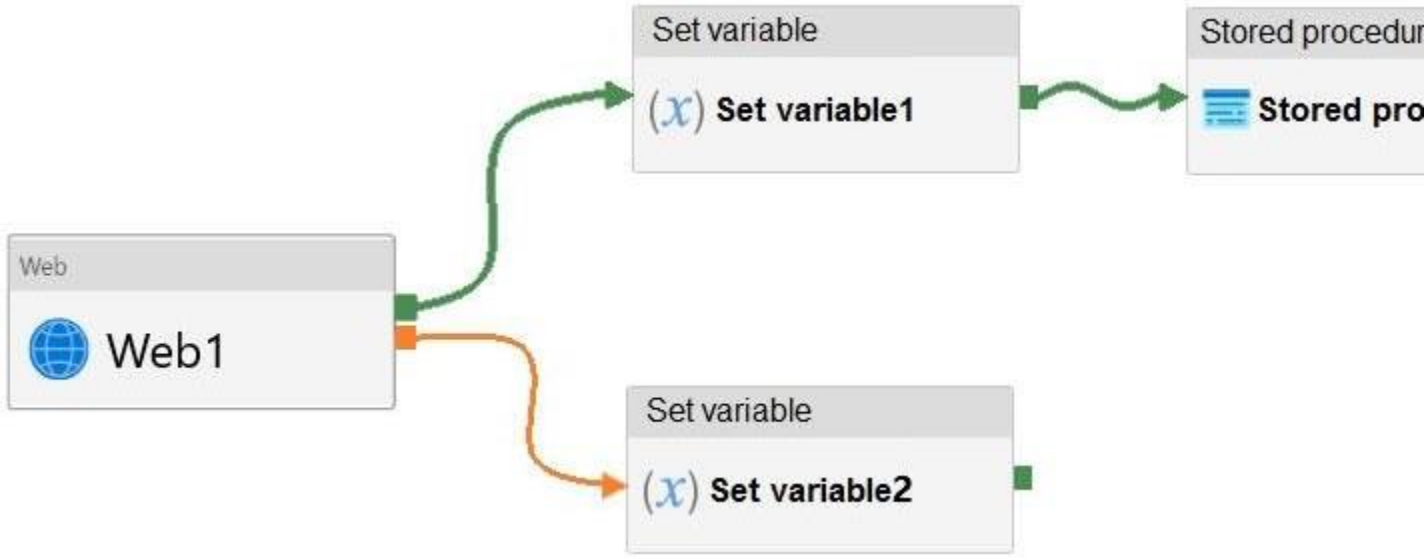
Hash-distribution improves query performance on large fact tables.
Incorrect Answers:
A: Do not use a date column for hash distribution. All data for the same date lands in the same distribution. If several users are all filtering on the same date, then only 1 of the 60 distributionsdo all the processing work.

Community vote distribution

B (100%)

Question #26*Topic 4*

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 statementbased 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]**

compl
fail
succe

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

Cance
Failed
Succe

Hide Solution Discussion 6

Correct Answer:

Answer Area

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

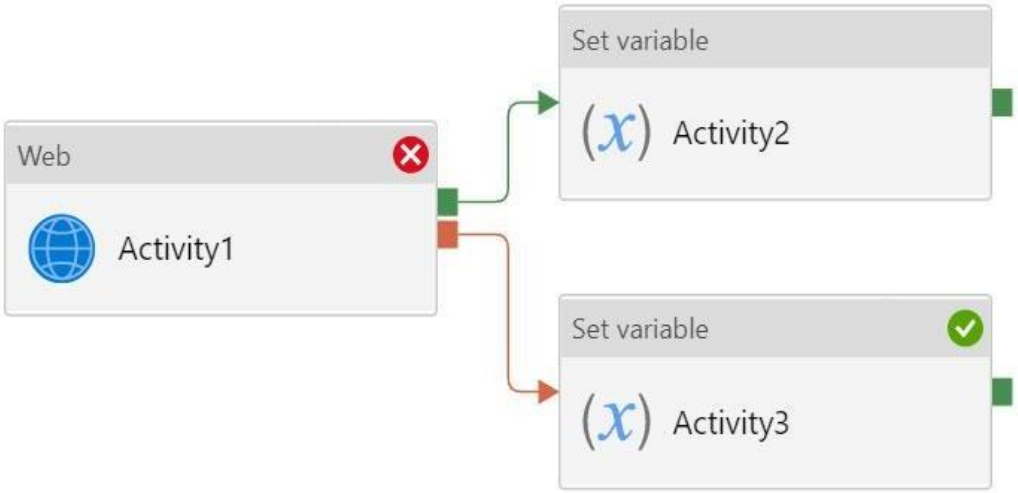
compl
fail
succe

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

Cance
Failed
Succe

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 fromthe 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 #27

estion #27*Topic 4*

You have several Azure Data Factory pipelines that contain a mix of the following typesof 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 **Most Voted**
- E. Azure Databricks **Most Voted**

Hide Solution | Discussion 8

Correct Answer: *AC*

Community vote distribution

DE (100%)

Question #28*Topic 4*

You have an Azure Synapse Analytics dedicated SQL pool named Pool1 and adatabase 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.dm_pdw_sys_info.
- 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.

Hide Solution Discussion 2

Correct Answer: *D*

Microsoft recommends use of sys.dm_pdw_nodes_db_partition_stats to analyze anyskewness in the data.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/cheat- sheet

Community vote distribution

D (100%)

Question #1*Topic 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 youare 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 inthe 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 retailstores 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 supplierdata 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 rowper 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 differenttime 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 retailstore addresses.
You plan to create a promotional table that will contain a promotion ID. The promotion ID will be associated to a specific product. The product will be identified by a product ID.The table will be approximately 5 GB.

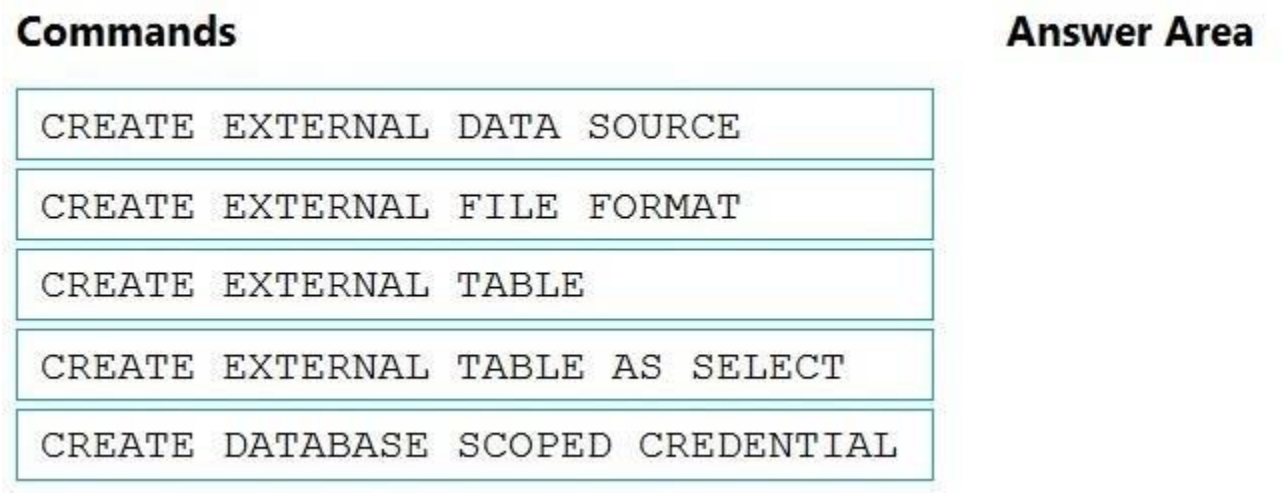
Streaming Twitter Data -
The ecommerce department at Contoso develops an Azure logic app that captures trending Twitter feeds referencing the company's products and pushes the products toAzure Event Hubs.
Planned Changes and Requirements

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 SSISpackages.
Use Azure Synapse Analytics to analyze Twitter feeds to assess customer sentimentsabout 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 toprovide efficient loads by month. Boundary values must belong to the partition on theright.
Ensure that queries joining and filtering sales transaction records based on product IDcomplete 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 SQLpool 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 Storagewithout purchasing additional throughput or capacity units.
Store Twitter feeds in Azure Storage by using Event Hubs Capture. The feeds will beconverted into Parquet files.
Ensure that the data store supports Azure AD-based access control down to the objectlevel.
Minimize administrative effort to maintain the Twitter feed data records.

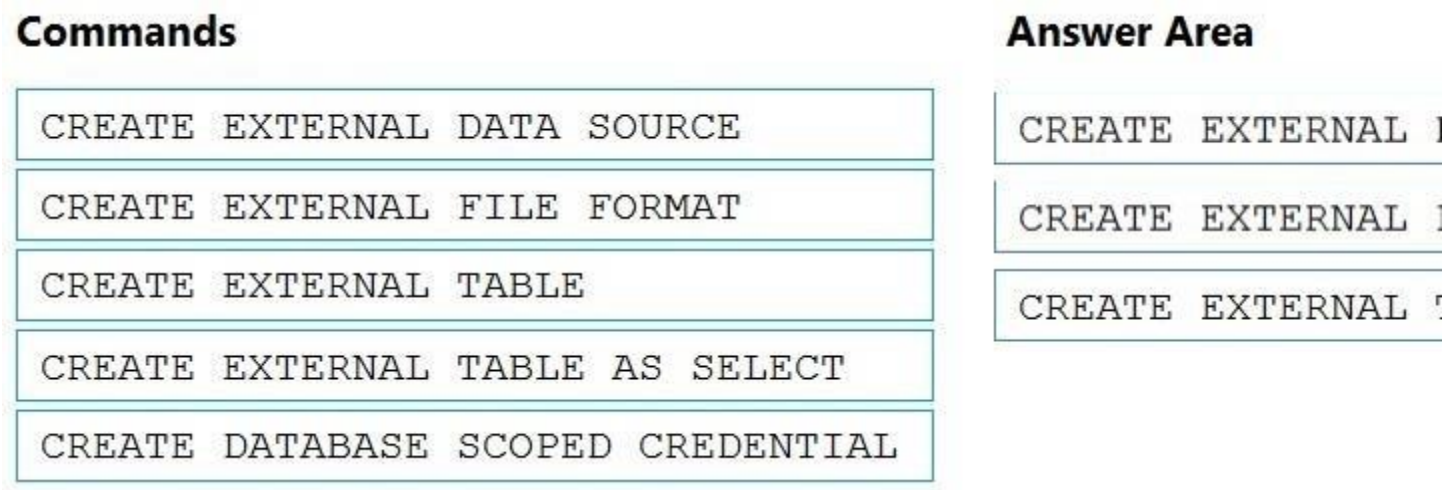
Purge Twitter feed data records that are older than two years.Data

Integration Requirements -
Contoso identifies the following requirements for data integration:
Use an Azure service that leverages the existing SSIS packages to ingest on-premisesdata 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 activitiescan 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 SQLpool. The solution must meet the customer sentiment analytic requirements.
Which three Transact-SQL DDL commands should you run in sequence? To answer,move the appropriate commands from the list of commands to the answer area and arrange them in the correct order.
NOTE: More than one order of answer choices is correct. You will receive credit for anyof the correct orders you select.
Select and Place:



Hide Solution Discussion 6

Correct Answer:



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 definesexternal 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, selectingfrom 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 SQLto 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 #2*Topic 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 youare 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 andother 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 inthe 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 theQuestion button to return to the question.

Overview -
Contoso, Ltd. is a clothing retailer based in Seattle. The company has 2,000 retailstores 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 supplierdata comprised of 10 billion records stored across multiple on-premises
Microsoft SQL Server servers. The SQL Server instances contain data from variousoperational 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 rowper 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 differenttime 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 retailstore addresses.
You plan to create a promotional table that will contain a promotion ID. The promotion

ID will be associated to a specific product. The product will be identified by a product ID.The table will be approximately 5 GB.

Streaming Twitter Data -
The ecommerce department at Contoso develops an Azure logic app that captures trending Twitter feeds referencing the company's products and pushes the products toAzure 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 sentimentsabout 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 toprovide efficient loads by month. Boundary values must belong to the partition on theright.
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 SQLpool 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 Storagewithout purchasing additional throughput or capacity units.
Store Twitter feeds in Azure Storage by using Event Hubs Capture. The feeds will beconverted into Parquet files.
Ensure that the data store supports Azure AD-based access control down to the objectlevel.
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-premisesdata 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 activitiescan 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 mustmeet the sales transaction dataset requirements.
What should you include in the solution? To answer, select the appropriate options inthe 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

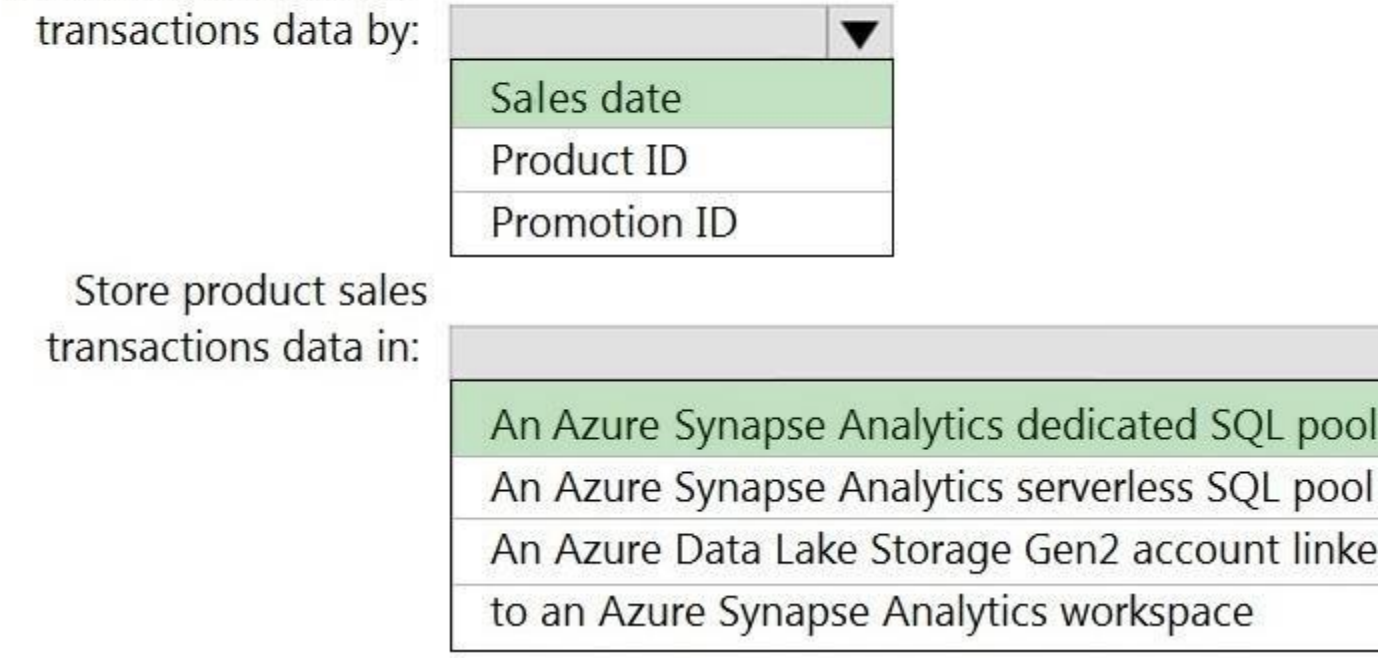
Store product sales transactions data in:

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

Hide Solution Discussion 3

Correct Answer:

Answer Area



Box 1: Sales date -
Scenario: Contoso requirements for data integration include:
➪ Partition data that contains sales transaction records. Partitions must be designed toprovide 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 DataWarehousing 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 #3*Topic 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 youare 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 inthe 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 retailstores 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 supplierdata comprised of 10 billion records stored across multiple on-premises
Microsoft SQL Server servers. The SQL Server instances contain data from variousoperational 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 rowper 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 differenttime 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 retailstore addresses.
You plan to create a promotional table that will contain a promotion ID. The promotion

ID will be associated to a specific product. The product will be identified by a product ID.The table will be approximately 5 GB.

Streaming Twitter Data -
The ecommerce department at Contoso develops an Azure logic app that captures trending Twitter feeds referencing the company's products and pushes the products toAzure 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 SSISpackages.
Use Azure Synapse Analytics to analyze Twitter feeds to assess customer sentimentsabout 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 toprovide efficient loads by month. Boundary values must belong to the partition on theright.
Ensure that queries joining and filtering sales transaction records based on product IDcomplete 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 SQLpool 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 Storagewithout purchasing additional throughput or capacity units.
Store Twitter feeds in Azure Storage by using Event Hubs Capture. The feeds will beconverted into Parquet files.
Ensure that the data store supports Azure AD-based access control down to the objectlevel.
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-premisesdata 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

Hide Solution Discussion 3

Correct Answer: *A*

Scenario: Implement a surrogate key to account for changes to the retail store addresses.
A surrogate key on a table is a column with a unique identifier for each row. The key is not generated from the table data. Data modelers like to create surrogate keys on their tables when they design data warehouse models. You can use the IDENTITY property to achieve this goal simply and effectively without affecting load performance.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-identity

Community vote distribution

A (100%)

Question #4*Topic 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 youare 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 inthe 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 theQuestion button to return to the question.

Overview -
Contoso, Ltd. is a clothing retailer based in Seattle. The company has 2,000 retailstores 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 supplierdata 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 rowper 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 differenttime 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 retailstore addresses.
You plan to create a promotional table that will contain a promotion ID. The promotion ID will be associated to a specific product. The product will be identified by a product ID.The table will be approximately 5 GB.

Streaming Twitter Data -
The ecommerce department at Contoso develops an Azure logic app that captures trending Twitter feeds referencing the company's products and pushes the products toAzure 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 SSISpackages.

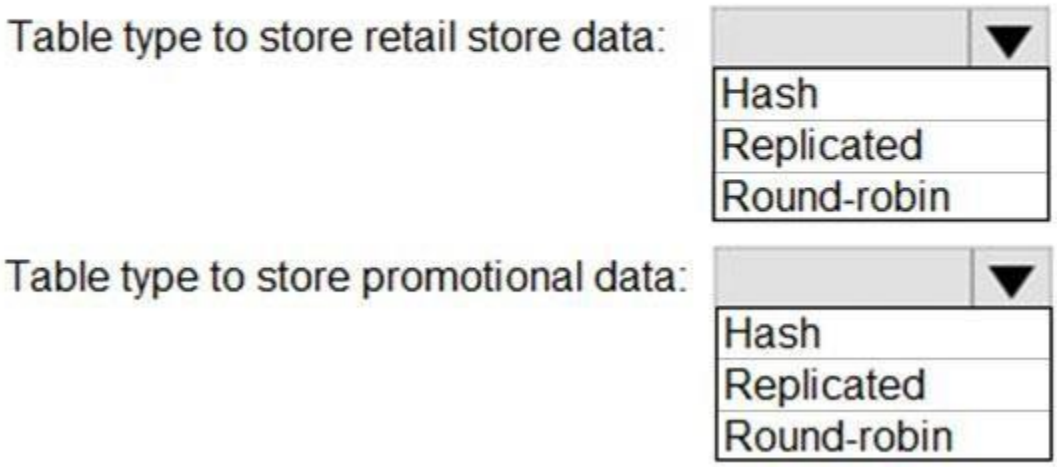
Use Azure Synapse Analytics to analyze Twitter feeds to assess customer sentimentsabout products.
Sales Transaction Dataset Requirements
Contoso identifies the following requirements for the sales transaction dataset: Partition data that contains sales transaction records. Partitions must be designed toprovide efficient loads by month. Boundary values must belong to the partition on theright.
Ensure that queries joining and filtering sales transaction records based on product IDcomplete 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 SQLpool 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 Storagewithout 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 objectlevel.
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-premisesdata 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 activitiescan be version-controlled and developed independently by multiple data engineers.**Question**HOTSPOT -
You need to design an analytical storage solution for the transactional data. Thesolution must meet the sales transaction dataset requirements.
What should you include in the solution? To answer, select the appropriate options inthe answer area.
NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area



Hide Solution Discussion 9

Correct

Answer Area

Answer:

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 aproduct ID. The table will be approximately 5 GB.
☞ Ensure that queries joining and filtering sales transaction records based on productID complete as quickly as possible.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute

Question #5*Topic 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 youare 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 andother 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 inthe 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 theQuestion button to return to the question.

Overview -
Contoso, Ltd. is a clothing retailer based in Seattle. The company has 2,000 retailstores 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 supplierdata comprised of 10 billion records stored across multiple on-premises
Microsoft SQL Server servers. The SQL Server instances contain data from variousoperational 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 rowper 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 differenttime 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 retailstore addresses.
You plan to create a promotional table that will contain a promotion ID. The promotion ID will be associated to a specific product. The product will be identified by a product ID.The table will be approximately 5 GB.

Streaming Twitter Data -
The ecommerce department at Contoso develops an Azure logic app that captures trending Twitter feeds referencing the company's products and pushes the products toAzure 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 sentimentsabout 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 toprovide efficient loads by month. Boundary values must belong to the partition on theright.
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 Storagewithout purchasing additional throughput or capacity units.
Store Twitter feeds in Azure Storage by using Event Hubs Capture. The feeds will beconverted into Parquet files.
Ensure that the data store supports Azure AD-based access control down to the objectlevel.
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-premisesdata 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 activitiescan be version-controlled and developed independently by multiple data
engineers.**Question**HOTSPOT -
You need to implement an Azure Synapse Analytics database object for storing thesales 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:

C
C
C

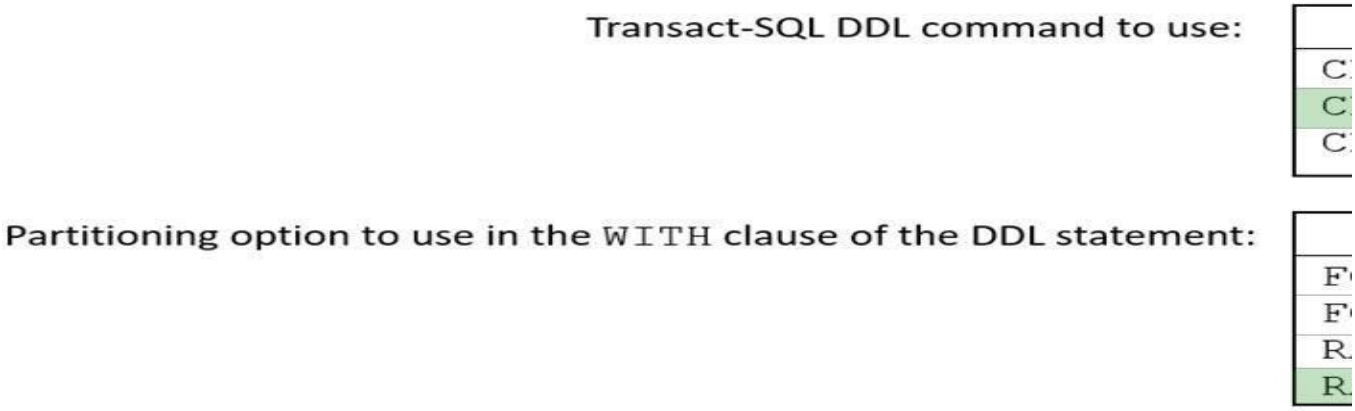
Partitioning option to use in the `WITH` clause of the DDL statement:

F
F
R
R

Hide Solution Discussion 3

Correct Answer:

Answer Area



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 toprovide 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 productID 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

Previous QuestionsNext Questions

Question #6*Topic 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 youare 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 inthe 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 retailstores 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 supplierdata comprised of 10 billion records stored across multiple on-premises
Microsoft SQL Server servers. The SQL Server instances contain data from variousoperational 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 rowper 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 removedmonthly.
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 retailstore addresses.
You plan to create a promotional table that will contain a promotion ID. The promotion ID will be associated to a specific product. The product will be identified by a product ID.The table will be approximately 5 GB.

Streaming Twitter Data -
The ecommerce department at Contoso develops an Azure logic app that captures trending Twitter feeds referencing the company's products and pushes the products toAzure 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 SSISpackages.
Use Azure Synapse Analytics to analyze Twitter feeds to assess customer sentimentsabout 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 toprovide efficient loads by month. Boundary values must belong to the partition on theright.
Ensure that queries joining and filtering sales transaction records based on product IDcomplete 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 SQLpool 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 Storagewithout purchasing additional throughput or capacity units.
Store Twitter feeds in Azure Storage by using Event Hubs Capture. The feeds will beconverted into Parquet files.
Ensure that the data store supports Azure AD-based access control down to the objectlevel.
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-premisesdata 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 activitiescan 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

Hide Solution Discussion 2

Correct Answer: *D*

Scenario: Purge Twitter feed data records that are older than two years.
Data sets have unique lifecycles. Early in the lifecycle, people access some data often.But the need for access often drops drastically as the data ages. Some data remains idle in the cloud and is rarely accessed once stored. Some data sets expire days or months after creation, while other data sets are actively read and modified throughout their lifetimes. Azure Storage lifecycle management offers a rule-based policy that youcan use to transition blob data to the appropriate access tiers or to expire data at the end of the data lifecycle.
Reference:
https://docs.microsoft.com/en-us/azure/storage/blobs/lifecycle-management-overview

Question #1*Topic 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 youare 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 andother 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 theexam. 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 inthe 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 theQuestion button to return to the question.

Overview -
Litware, Inc. owns and operates 300 convenience stores across the US. The companysells 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 itemsby providing their membership number at checkout.
Litware employs business analysts who prefer to analyze data by using Microsoft PowerBI, 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 aspossible.
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 toproduce 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 isolder

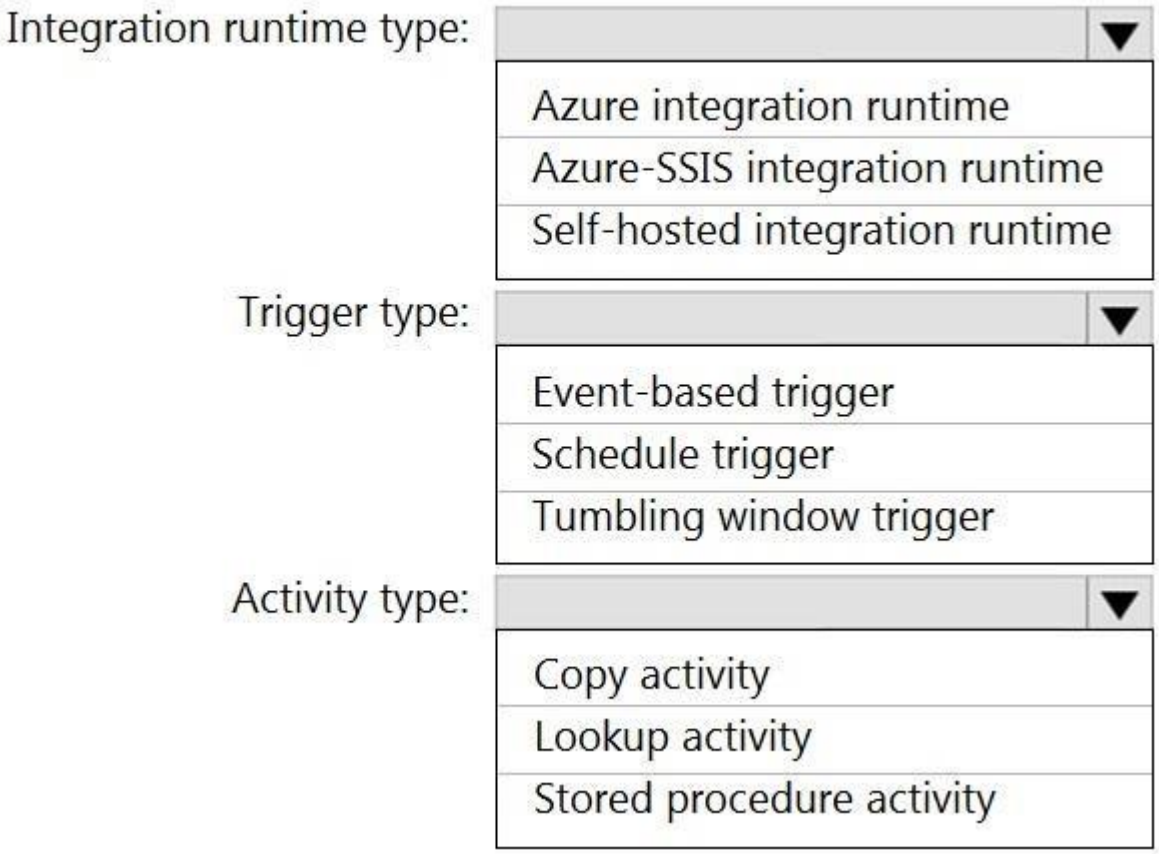
than 14 days must be removed.

Limit the business analysts' access to customer contact information, such as phonenumbers, 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 inthe 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 todata 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 trustedin 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 employeeswith 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 importthe 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:

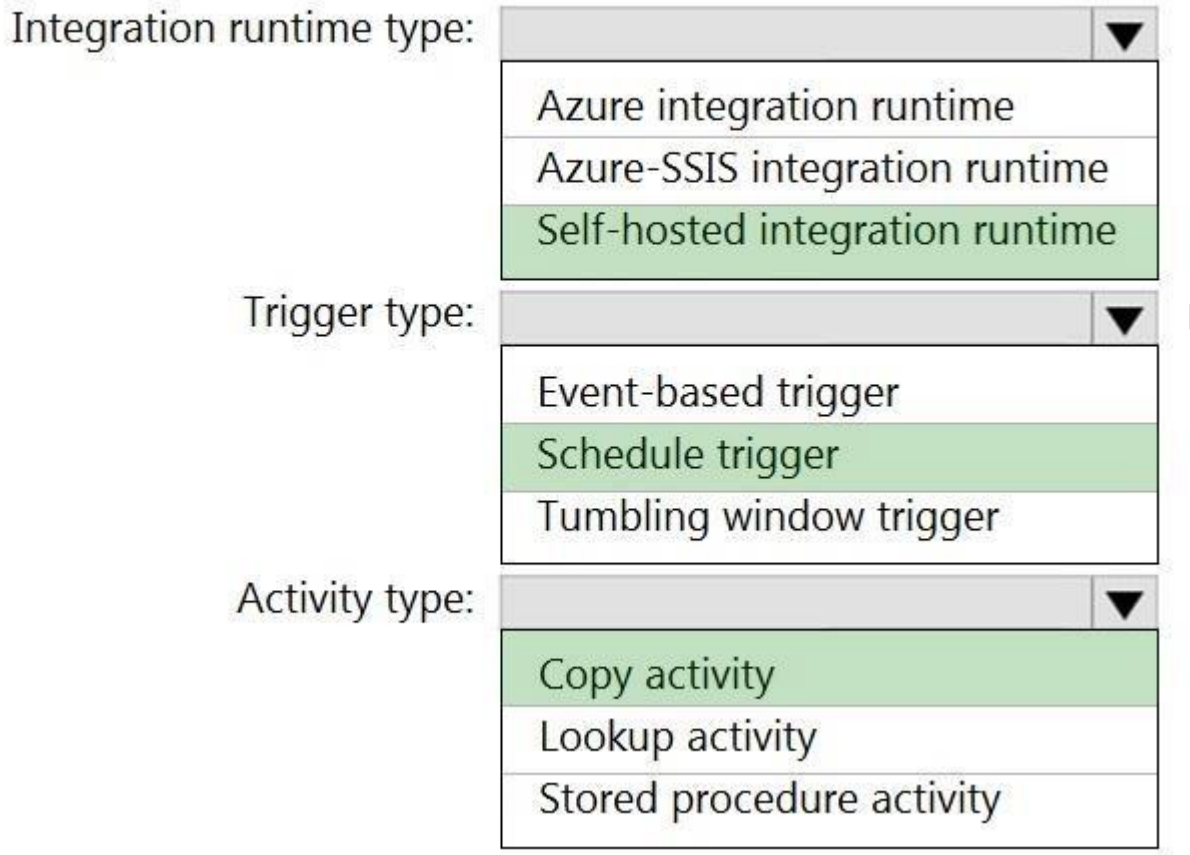
Answer Area



Hide Solution Discussion 8

Correct Answer:

Answer Area



Box 1: Self-hosted integration runtime
A self-hosted IR is capable of running copy activity between a cloud data stores and adata 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 datesare not trusted in the source table.
- Product data, including product ID, name, and category, comes from Salesforce andcan be imported into Azure once every eight hours. Row modified dates are not trustedin the source table.

Question #1*Topic 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 youare 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 andother 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 inthe 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 theQuestion button to return to the question.

Overview -
Contoso, Ltd. is a clothing retailer based in Seattle. The company has 2,000 retailstores 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 supplierdata comprised of 10 billion records stored across multiple on-premises
Microsoft SQL Server servers. The SQL Server instances contain data from variousoperational 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 rowper 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 differenttime periods.
Sales transaction data that is older than three years will be removed monthly.
You plan to create a retail store table that will contain the address of each retail store. The table will be approximately 2 MB. Queries for retail store sales will include the retail

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

Streaming Twitter Data -
The ecommerce department at Contoso develops an Azure logic app that captures trending Twitter feeds referencing the company's products and pushes the products toAzure Event Hubs.
Planned Changes and Requirements

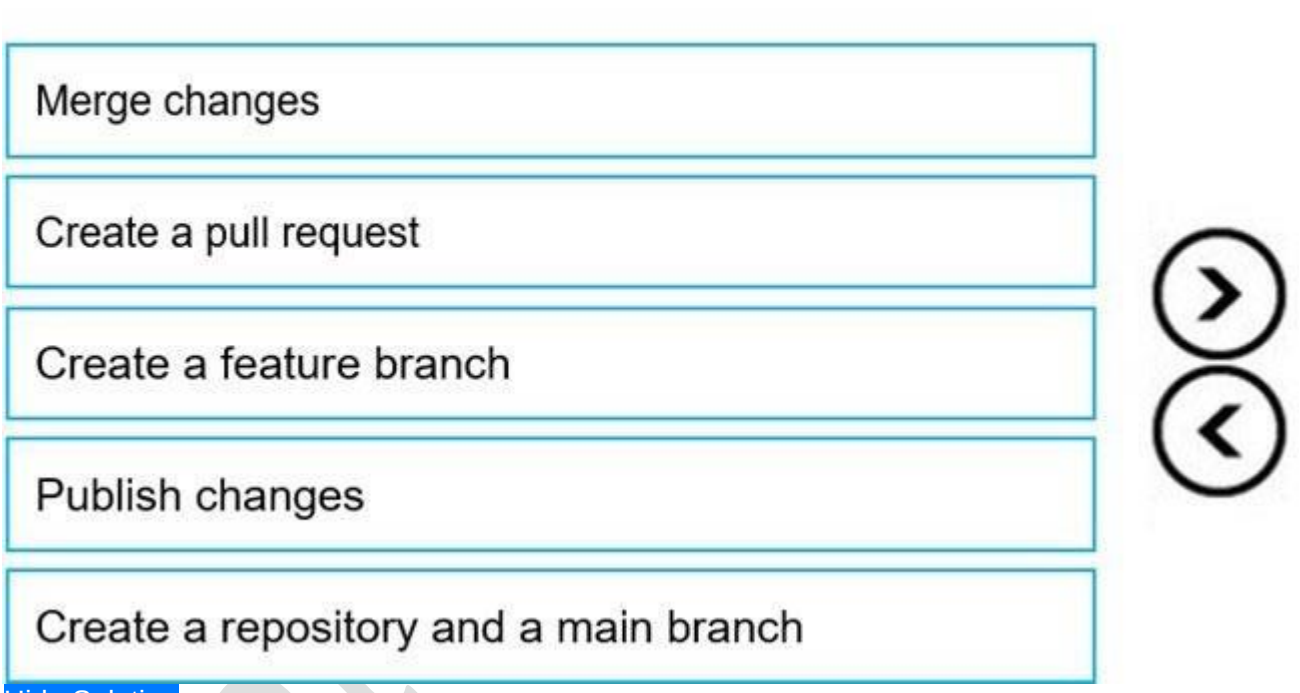
Planned Changes -
Contoso plans to implement the following changes:
Load the sales transaction dataset to Azure Synapse Analytics.
Integrate on-premises data stores with Azure Synapse Analytics by using SSISpackages.
Use Azure Synapse Analytics to analyze Twitter feeds to assess customer sentimentsabout 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 toprovide efficient loads by month. Boundary values must belong to the partition on theright.
Ensure that queries joining and filtering sales transaction records based on product IDcomplete 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 SQLpool 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 beconverted into Parquet files.
Ensure that the data store supports Azure AD-based access control down to the objectlevel.
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-premisesdata 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 activitiescan be version-controlled and developed independently by multiple data engineers.**Question**DRAG DROP -
You need to implement versioned changes to the integration pipelines. The solutionmust 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:

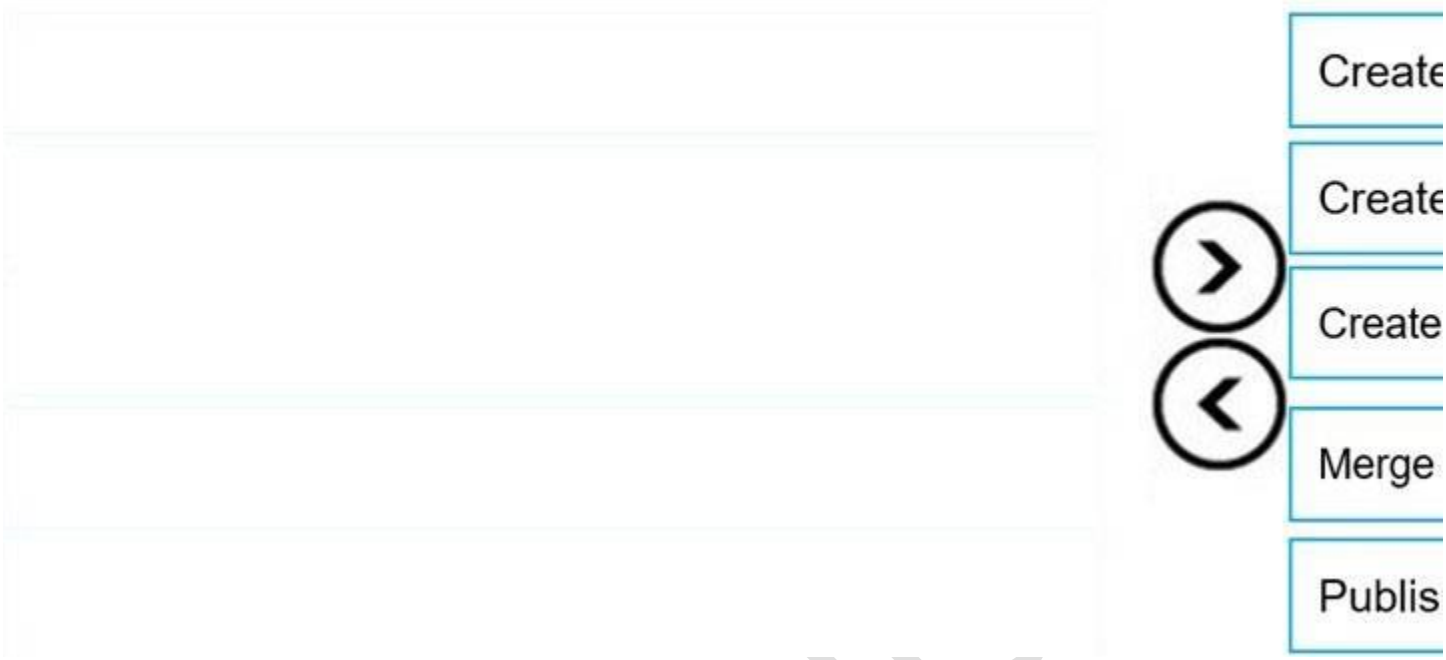
Discussion 7



Hide Solution

Correct Answer:

Actions



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

Question #1*Topic 8*

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 andother 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 inthe 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 theQuestion button to return to the question.

Overview -
Litware, Inc. owns and operates 300 convenience stores across the US. The companysells 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 itemsby providing their membership number at checkout.
Litware employs business analysts who prefer to analyze data by using Microsoft PowerBI, 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 aspossible.
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 toproduce 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 theanalytical 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 isolder than 14 days must be removed.
Limit the business analysts' access to customer contact information, such as phonenumbers, 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 inthe 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 todata 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 trustedin 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 employeeswith 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 usersoutside 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 **Most Voted**
- D. a database-level firewall IP rule

Hide S Discussion 10

Correct Answer: *A*

Scenario: Ensure that the analytical data store is accessible only to the company's on-premises network and Azure services.
Virtual network rules are one firewall security feature that controls whether the databaseserver for your single databases and elastic pool in Azure SQL Database or for your databases in SQL Data Warehouse accepts communications that are sent from particular subnets in virtual networks.

Server-level, not database-level: Each virtual network rule applies to your whole Azure SQL Database server, not just to one particular database on the server. In other words,virtual network rule applies at the server-level, not at the database-level.
Reference:
https://docs.microsoft.com/en-us/azure/sql-database/sql-database-vnet-service- endpoint-rule-overview

Community vote distribution

C (100%)

Question #2 Topic 8

Introductory Info Case study -
This is a case study. Case studies are not timed separately. You can use as much exam time asyou 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 allquestions included on this exam in the time provided.
To answer the questions included in a case study, you will need to reference information that isprovided in the case study. Case studies might contain exhibits and other resources that providemore 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 anew 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 leftpane 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 problemstatements. If the case study has an
All Information tab, note that the information displayed is identical to the information displayedon 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 sandwichesand pizzas.
Litware has a loyalty club whereby members can get daily discounts on specific items byproviding their membership number at checkout.
Litware employs business analysts who prefer to analyze data by using Microsoft Power BI, anddata scientists who prefer analyzing data in Azure Databricks notebooks.

Re

qui

re

me

nts

-

Bu

sin

ess

Go

als

-

Litware wants to create a new analytics environment in Azure to meet the followingrequirements:
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 discountsincrease sales of the discounted products.
Every four hours, notify store employees about how many prepared food items to produce basedon 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 provisionvirtual machines that must be managed by Litware.
Ensure that the analytical data store is accessible only to the company's on-premises network andAzure 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 analyticaldata 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 than14 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 theevent 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. Rowmodified 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 sourcetable.
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 companyexpects 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 thecalculation 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 customercontact information?

- A. Transparent Data Encryption (TDE)
- B. row-level security
- C. column-level security Most Voted
- D. data sensitivity labels

Correct Answer: D

Scenario: Limit the business analysts:€™ access to customer contact information, such as phonenumbers, because this type of data is not analytically relevant.
Labeling: You can apply sensitivity-classification labels persistently to columns by using new metadata attributes that have been added to the SQL Server database engine. This metadata canthen be used for advanced, sensitivity-based auditing and protection scenarios.
Incorrect Answers:
A: Transparent Data Encryption (TDE) encrypts SQL Server, Azure SQL Database, and Azure Synapse Analytics data files, known as encrypting data at rest. TDE does not provide encryptionacross communication channels.
Reference:
https://docs.microsoft.com/en-us/azure/azure-sql/database/data-discovery-and-classification- overview https://docs.microsoft.com/en-us/azure/sql-database/sql-database-security-overview

Community vote distribution

C (100%)

Topic 9 - Testlet 5

Question #1 Topic 9

Introductory Info Case study -
This is a case study. Case studies are not timed separately. You can use as much exam time asyou 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 allquestions included on this exam in the time provided.
To answer the questions included in a case study, you will need to reference information that isprovided in the case study. Case studies might contain exhibits and other resources that providemore 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 anew 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 leftpane 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 displayedon 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 sandwichesand pizzas.
Litware has a loyalty club whereby members can get daily discounts on specific items byproviding their membership number at checkout.
Litware employs business analysts who prefer to analyze data by using Microsoft Power BI, anddata scientists who prefer analyzing data in Azure Databricks notebooks.

Re

qui

re

me

nts

-

Bu

sin

ess

Go

als

-
Litware wants to create a new analytics environment in Azure to meet the followingrequirements:
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 discountsincrease sales of the discounted products.
Every four hours, notify store employees about how many prepared food items to produce basedon 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 provisionvirtual machines that must be managed by Litware.
Ensure that the analytical data store is accessible only to the company's on-premises network andAzure 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 analyticaldata 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 than14 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 theevent 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 discountamount, 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. Rowmodified 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 sourcetable.
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 companyexpects 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 thecalculation 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-timedata processing solution?

- A. Deploy a High Concurrency Databricks cluster.
- B. Deploy an Azure Stream Analytics job and use an Azure Automation runbook to check thestatus 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

Guarantee Stream Analytics job reliability during service updates
Part of being a fully managed service is the capability to introduce new service functionality andimprovements at a rapid pace. As a result, Stream Analytics can have a service update deploy ona weekly (or more frequent) basis. No matter how much testing is done there is still a risk that anexisting, running job may break due to the introduction of a bug. If you are running mission critical jobs, these risks need to be avoided. You can reduce this risk by following Azure:€™s paired region model.
Scenario: The application development team will create an Azure event hub to receive real-timesales data, including store number, date, time, product ID, customer loyalty number, price, and discount amount, from the point of sale (POS) system and output the data to data storage in Azure
Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-job-reliability

Community vote distribution

D (100%)