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 arrange them in the correct order.
Select and Place:

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

Hide Solution Discussion 15

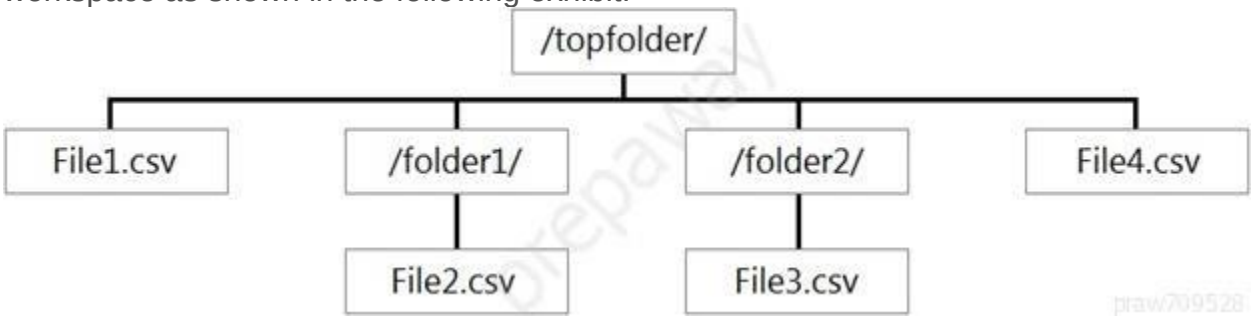
Correct Answer:

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?

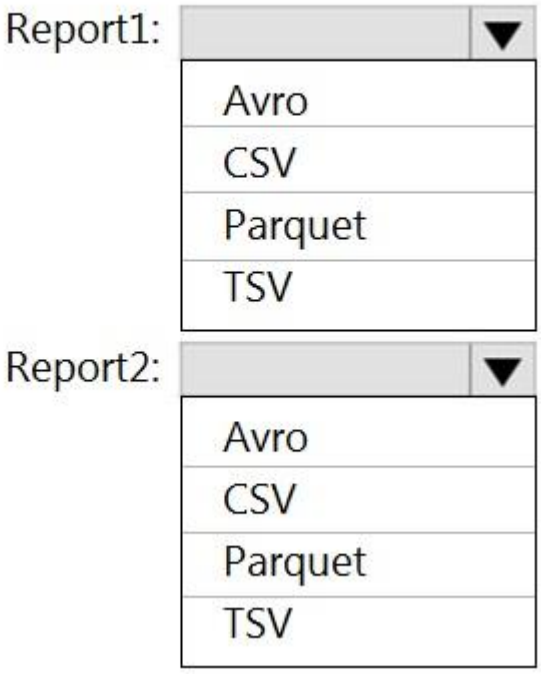
- A. File2.csv and File3.csv only
- B. File1.csv and File4.csv only
- C. File1.csv, File2.csv, File3.csv, and File4.csv
- D. File1.csv only

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.

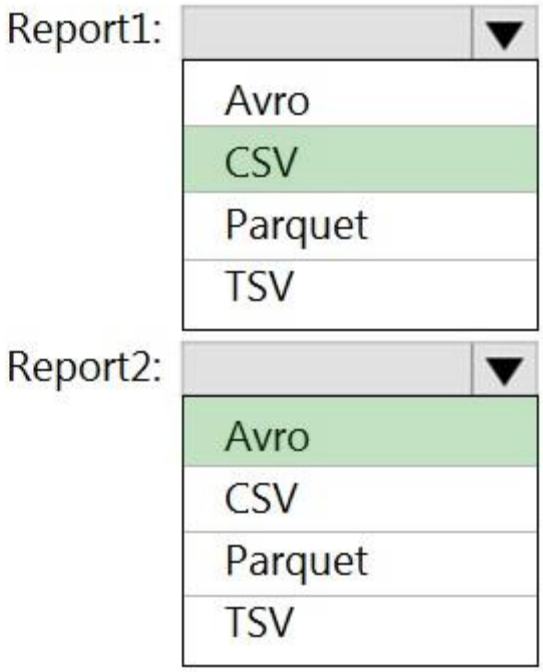
NOTE: Each correct selection is worth one point.
Hot Area:

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}_{DD}.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 -

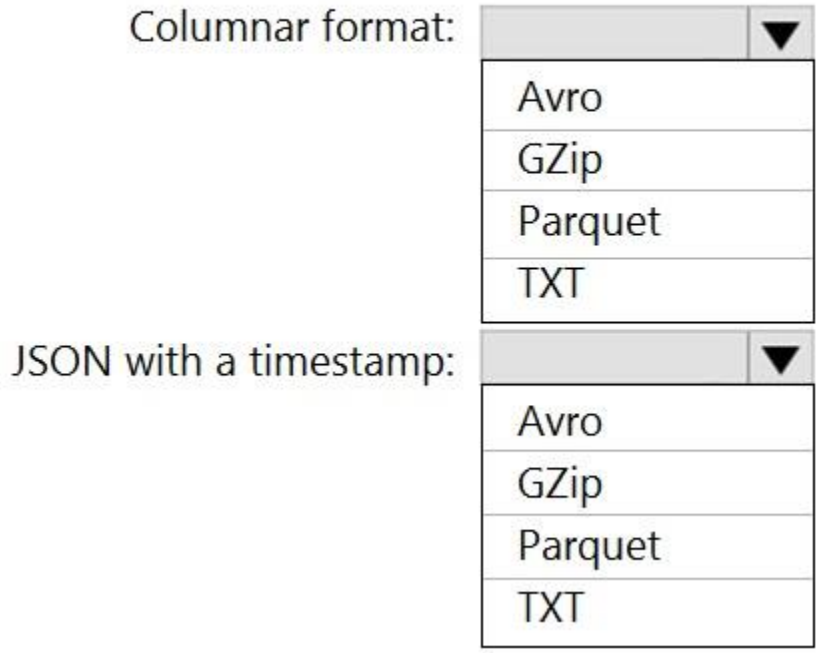
You need to output files from Azure Data Factory.

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.

Hot Area:

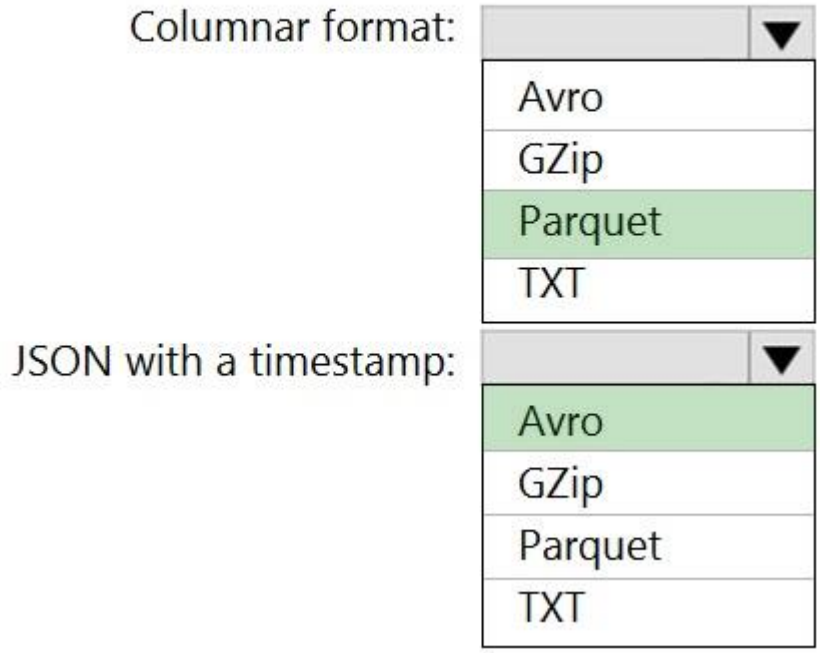
Answer Area



Hide Solution | Discussion 2

Correct

Answer Area



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).

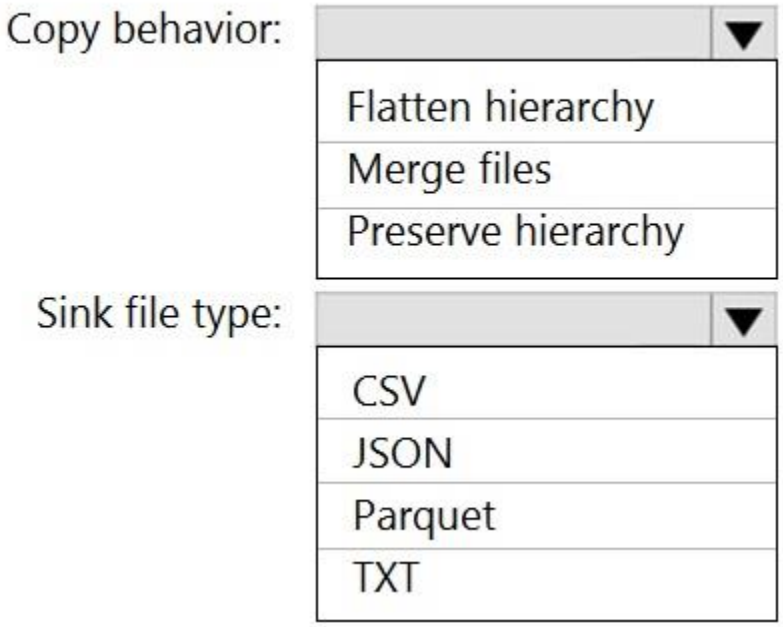
Avro format -

▪
☞ 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:

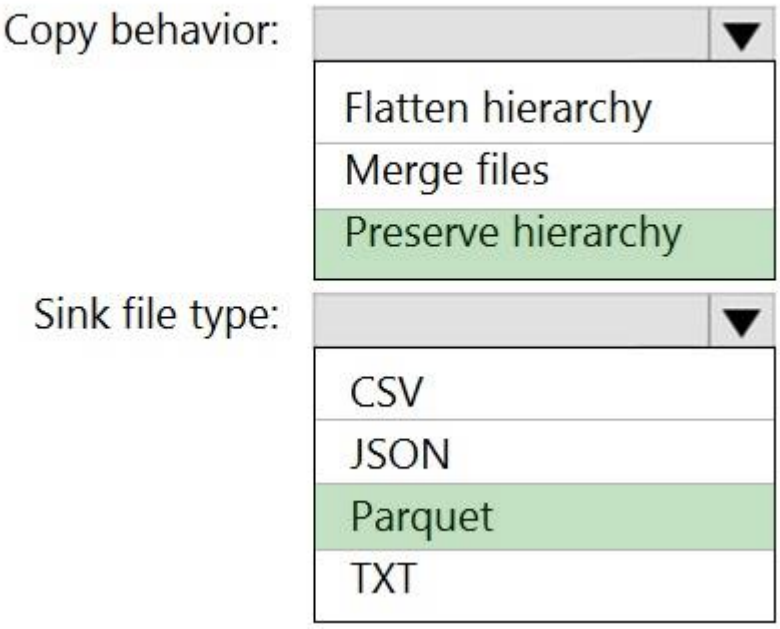
Answer Area



Hide Solution | Discussion 15

Correct

Answer Area



Answer:

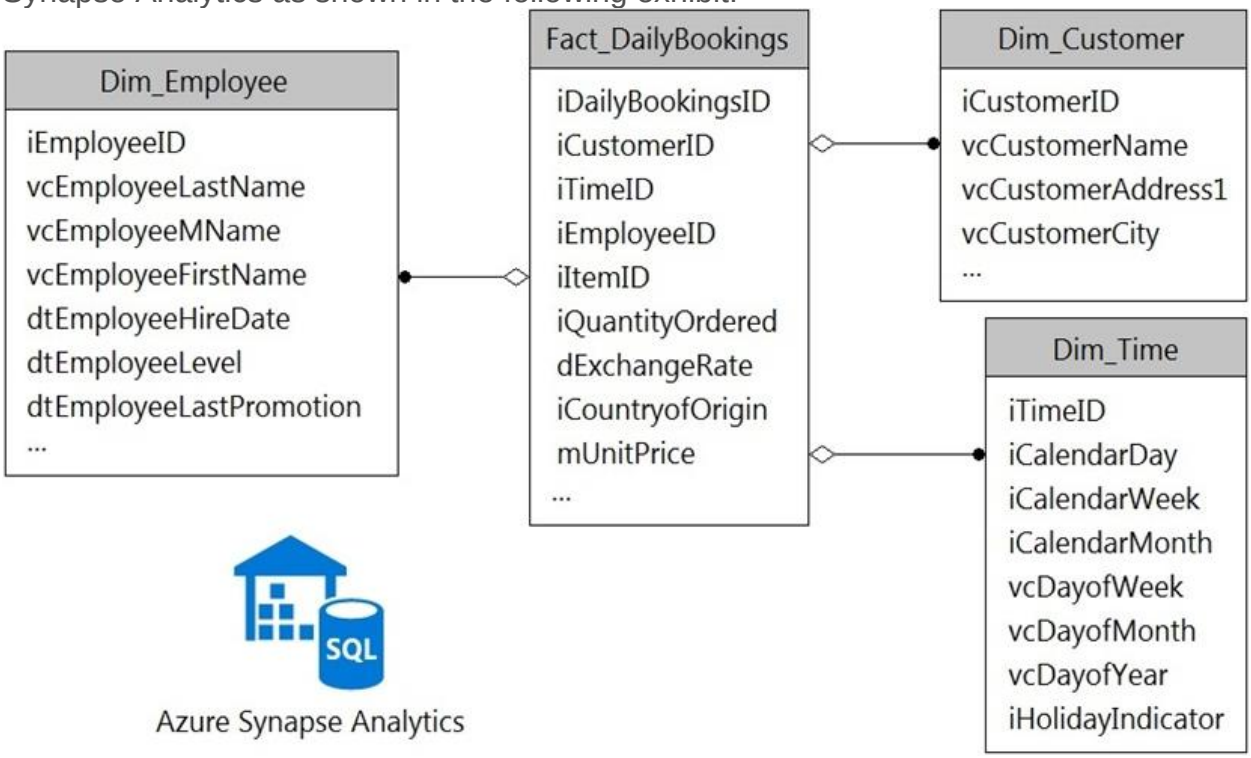
Box 1: Preserver hierarchy -

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

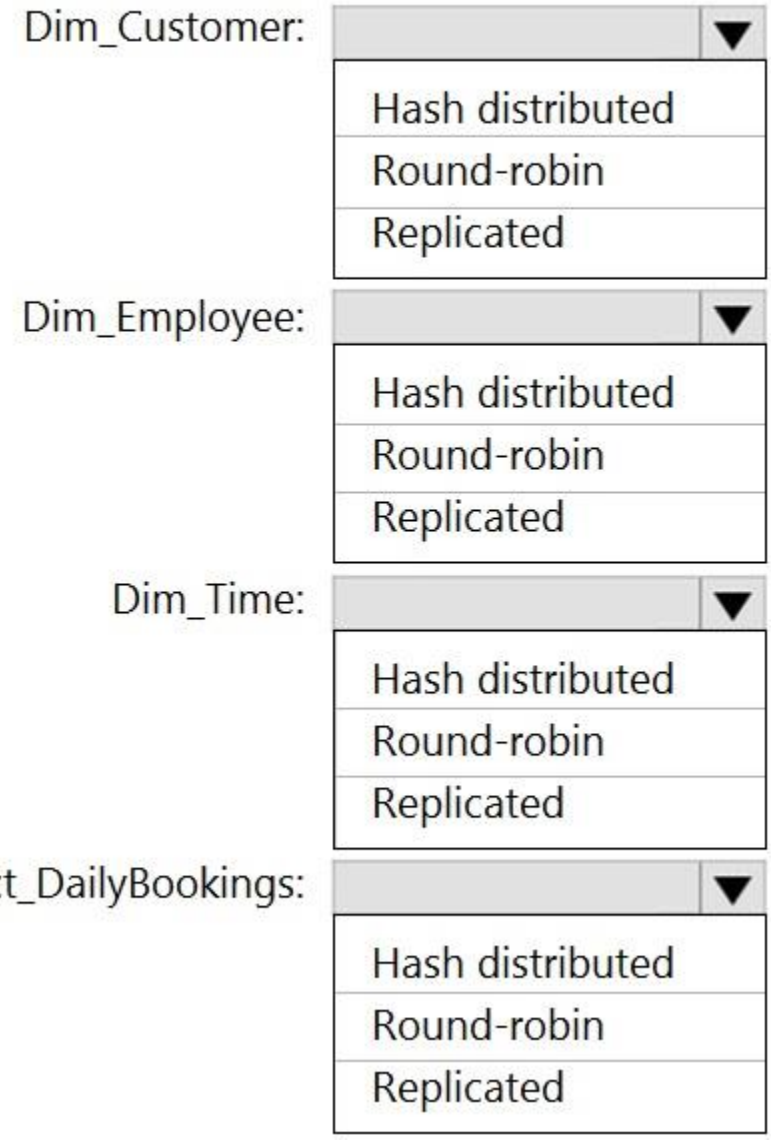
Question #9*Topic 1*
HOTSPOT -
You have a data model that you plan to implement in a data warehouse in Azure Synapse Analytics as shown in the following exhibit.



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:

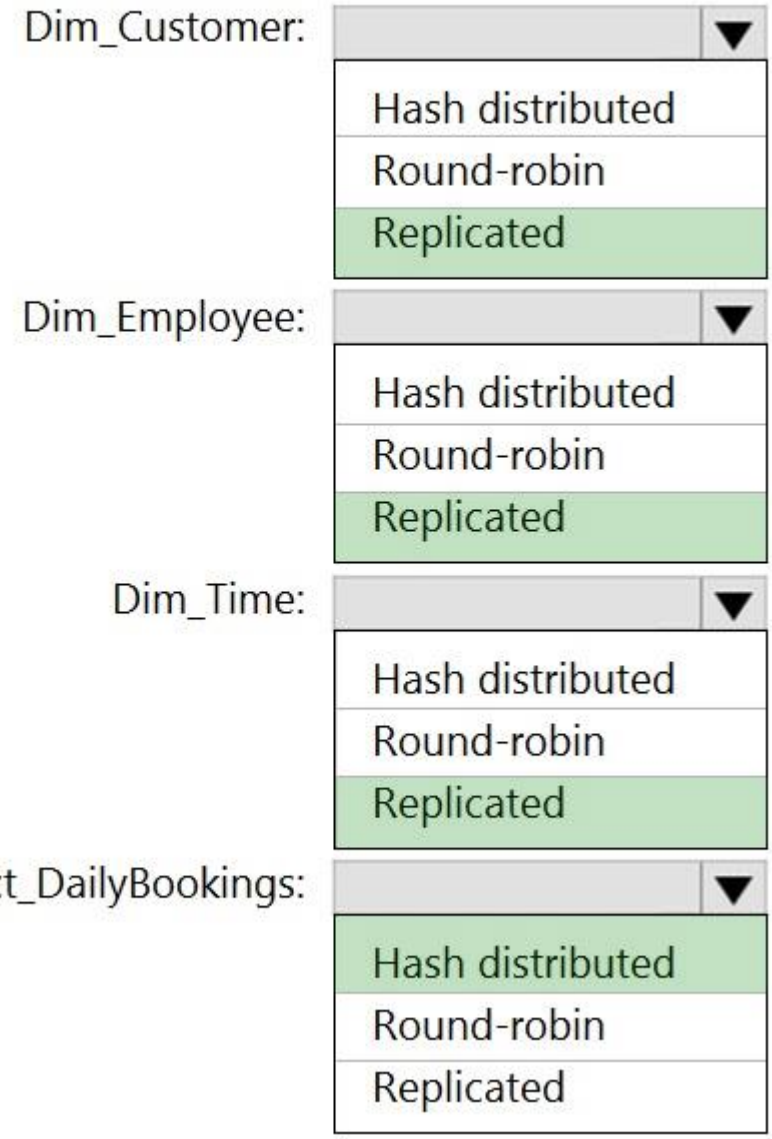
Answer Area



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 -

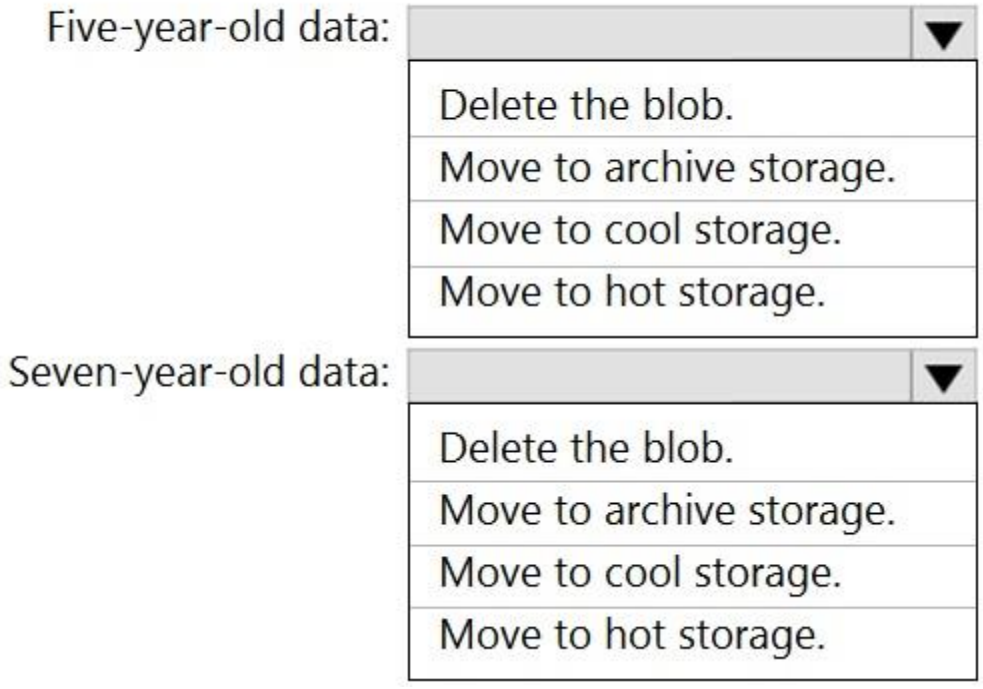
Box 3: Replicated -

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:

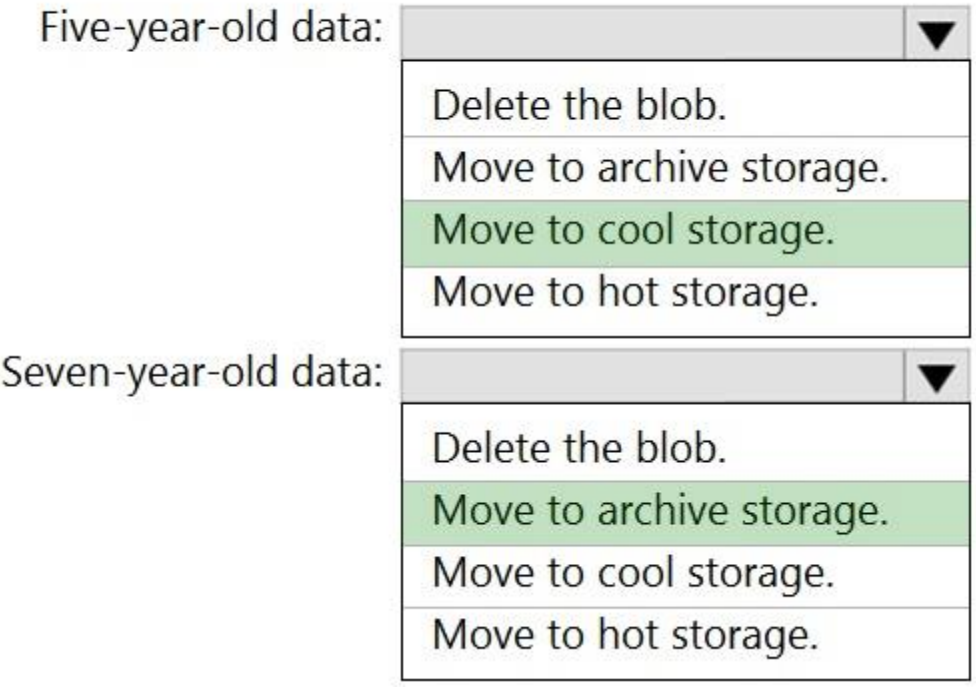
Answer Area



Hide Solution | Discussion 8

Correct Answer:

Answer Area



Box 1: Move to cool storage -

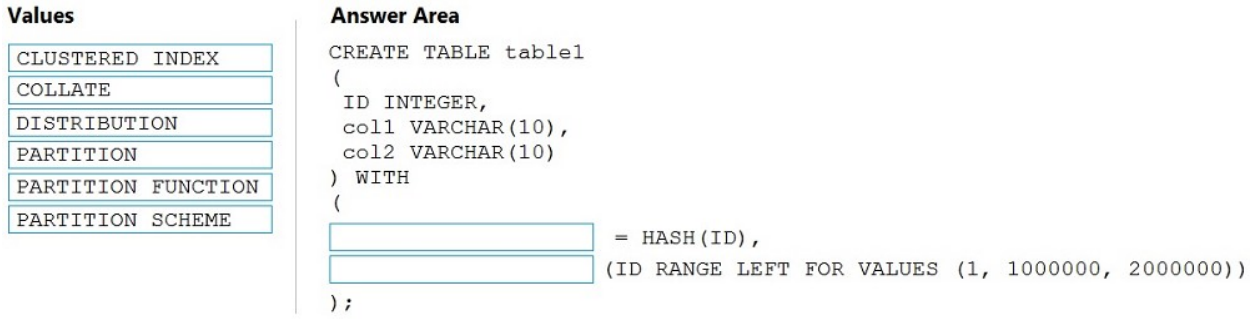
Box 2: Move to archive storage -
Archive - Optimized for storing data that is rarely accessed and stored for at least 180 days with flexible latency requirements, on the order of hours.
The following table shows a comparison of premium performance block blob storage, and the hot, cool, and archive access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

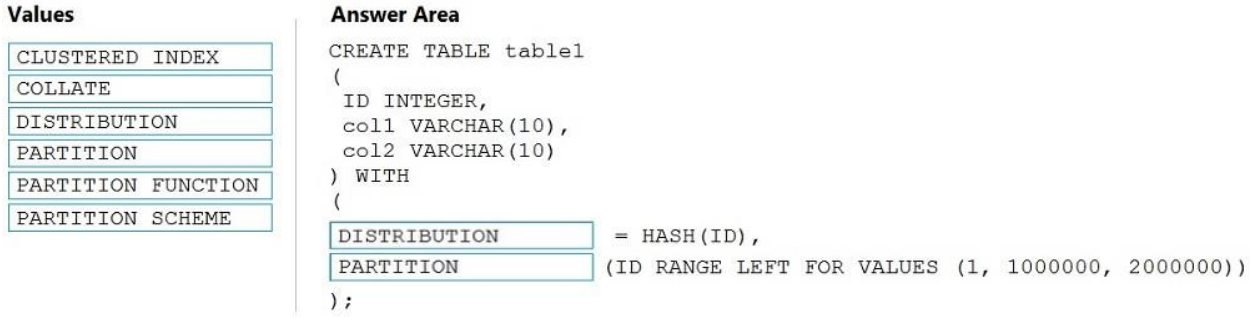
Question #11*Topic 1*
DRAG DROP -
You need to create a partitioned table in an Azure Synapse Analytics dedicated SQL pool.
How should you complete the Transact-SQL statement? To answer, drag the appropriate values to the correct targets. Each value may be used once, more than once, or not at all. You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

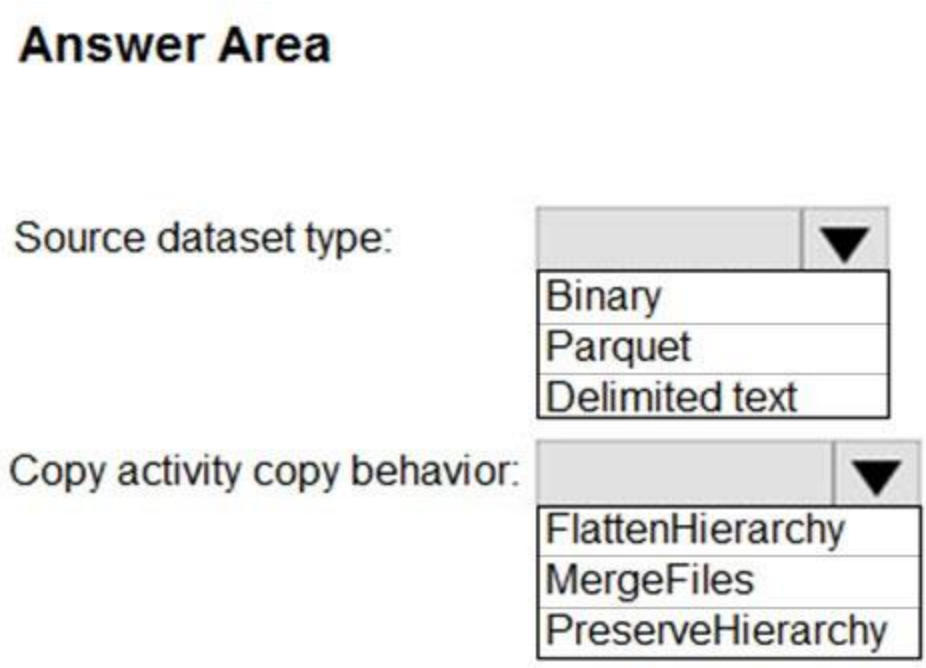
You need to copy folders and files from Storage1 to Storage2 by using a Data Factory copy activity. The solution must meet the following requirements:

- ☞ 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.

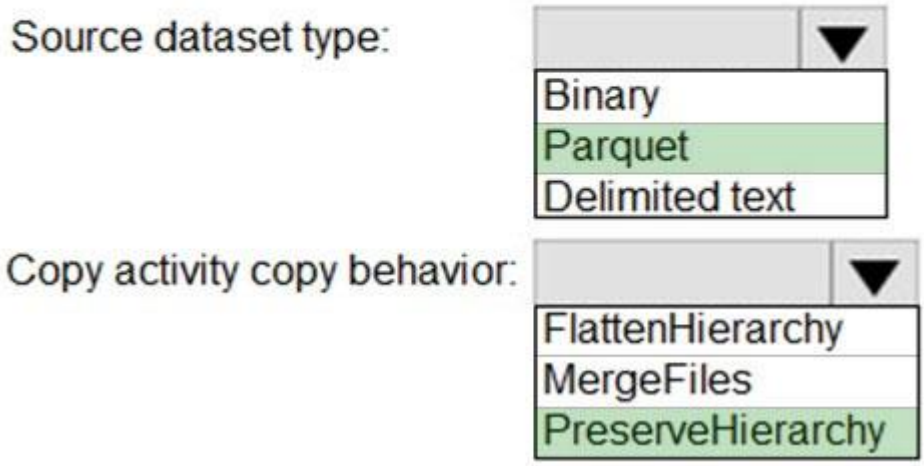
NOTE: Each correct selection is worth one point.

Hot Area:



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*

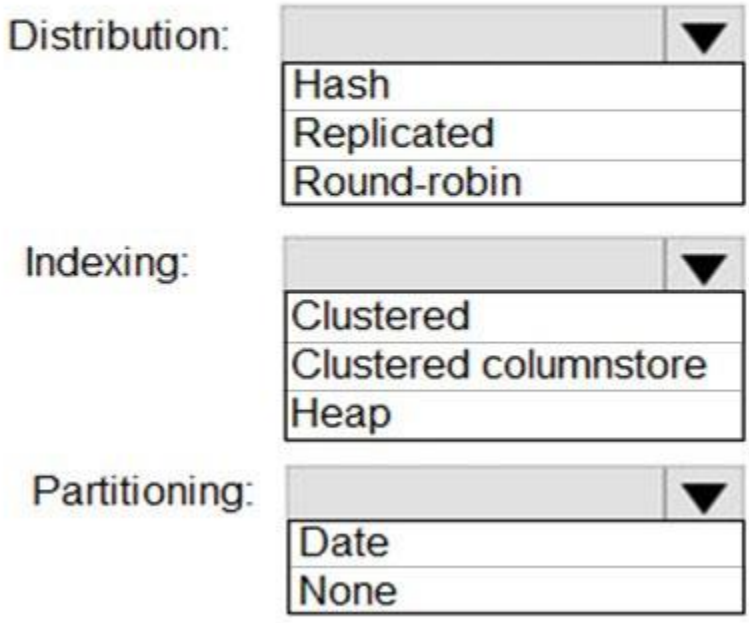
Locally redundant storage (LRS) copies your data synchronously three times within a single physical location in the primary region. LRS is the least expensive replication option
Reference:
https://docs.microsoft.com/en-us/azure/storage/common/storage-redundancy

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

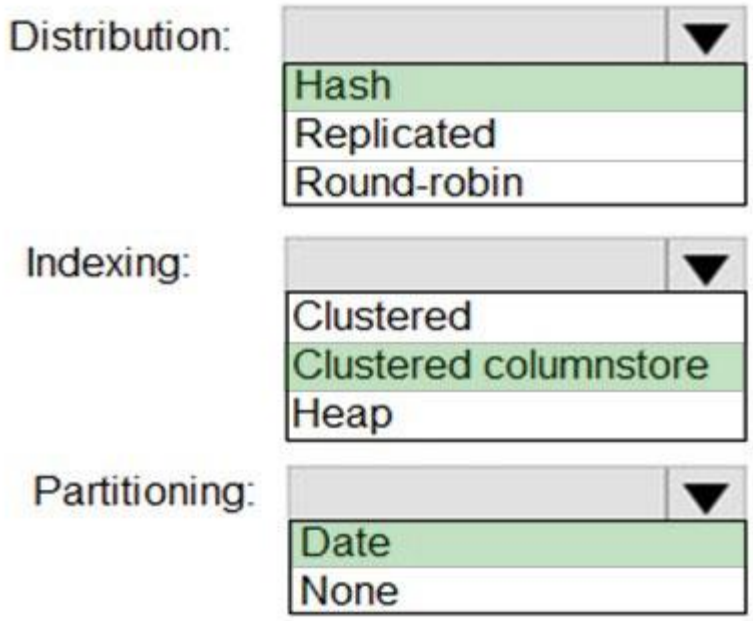
load.
You need to create the staging table. The solution must minimize how long it takes to load the data to the staging table.
How should you configure the table? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
Hot Area:

Answer Area



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

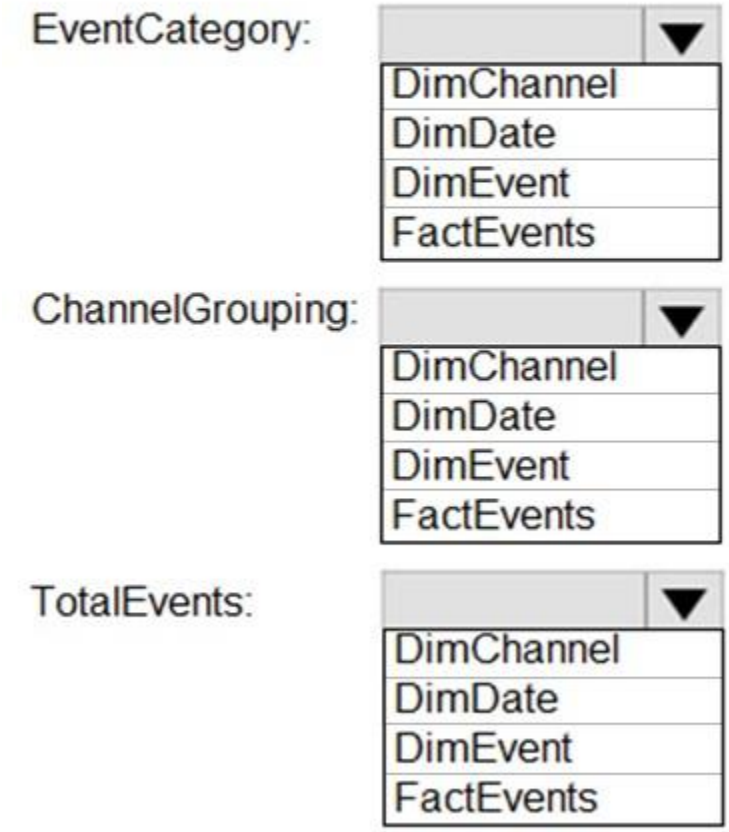
Question #21*Topic 1*
HOTSPOT -
From a website analytics system, you receive data extracts about user interactions such as downloads, link clicks, form submissions, and video plays.
The data contains the following columns.

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.

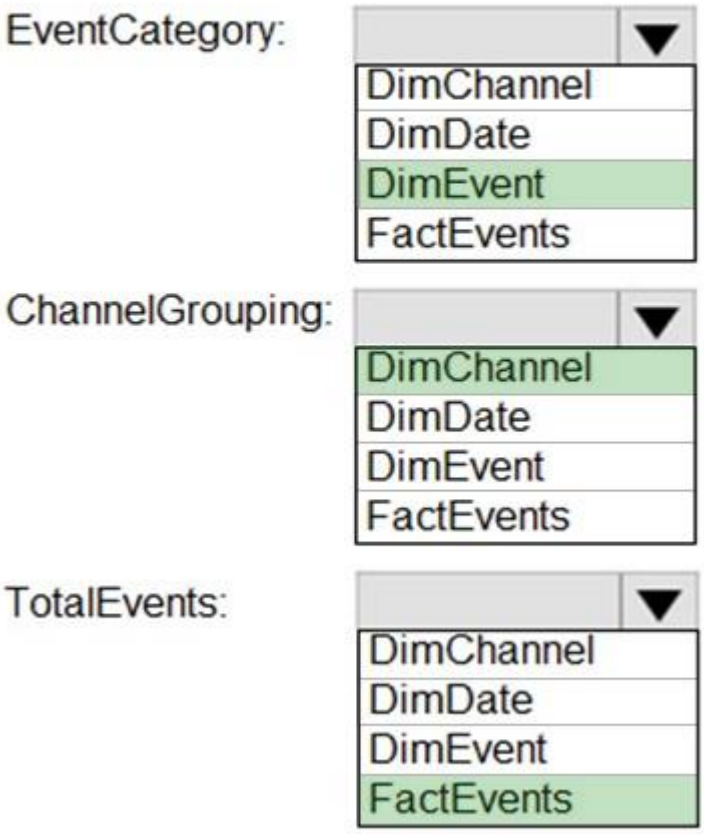
Hot Area:

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

prepaway
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

prepaway
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 Apacheג€™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.managementpolicybaseblob.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:

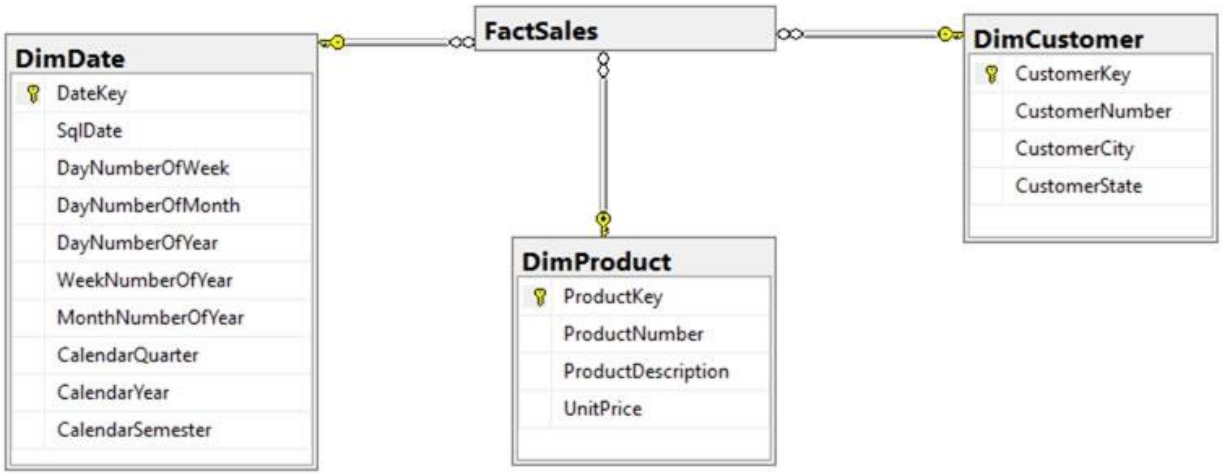
- Maintaining to a third normal form
- Normalizing to a fourth normal form
- Denormalizing to a second normal form

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 ✓

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

- a surrogate key
- a business key ✓
- 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*

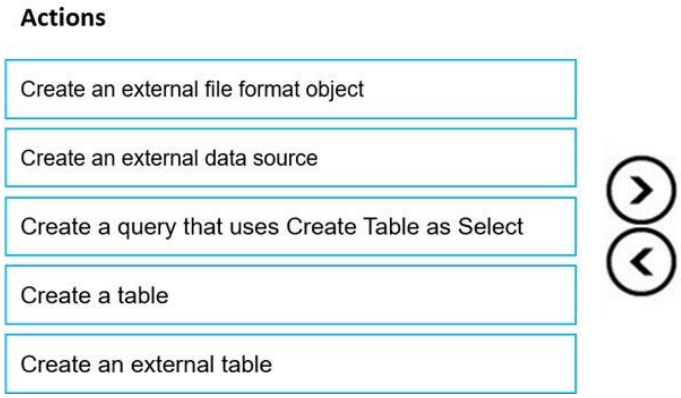
DRAG DROP -

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.

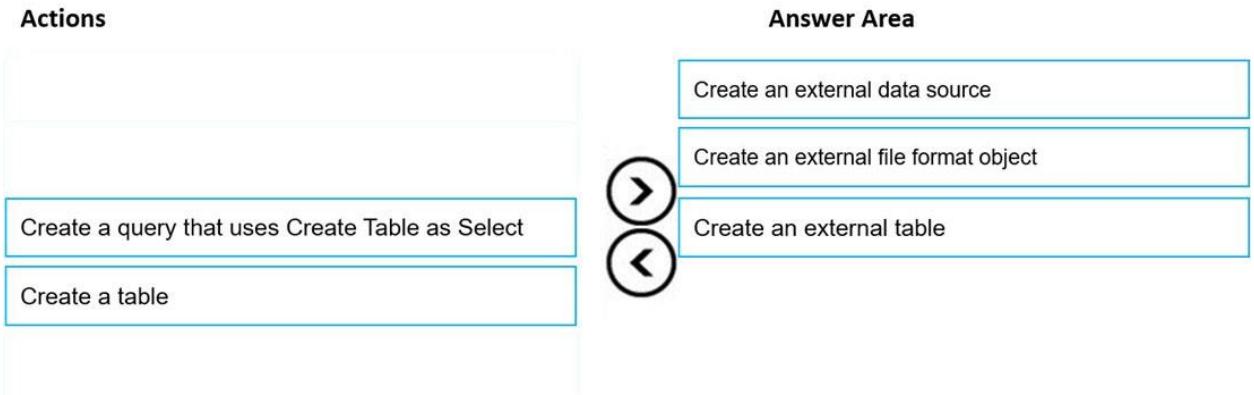
NOTE: More than one order of answer choices is correct. You will receive credit for any of the correct orders you select.

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

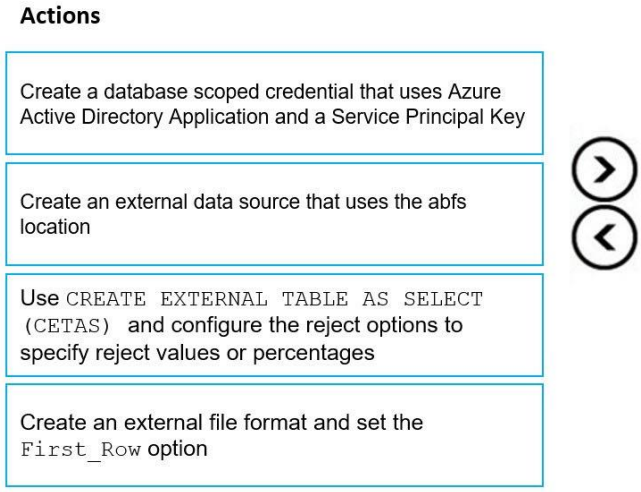
Question #43*Topic 1*
You are designing a dimension table for a data warehouse. The table will track the value of the dimension attributes over time and preserve the history of the data by adding new rows as the data changes.
Which type of slowly changing dimension (SCD) should you use?

- 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

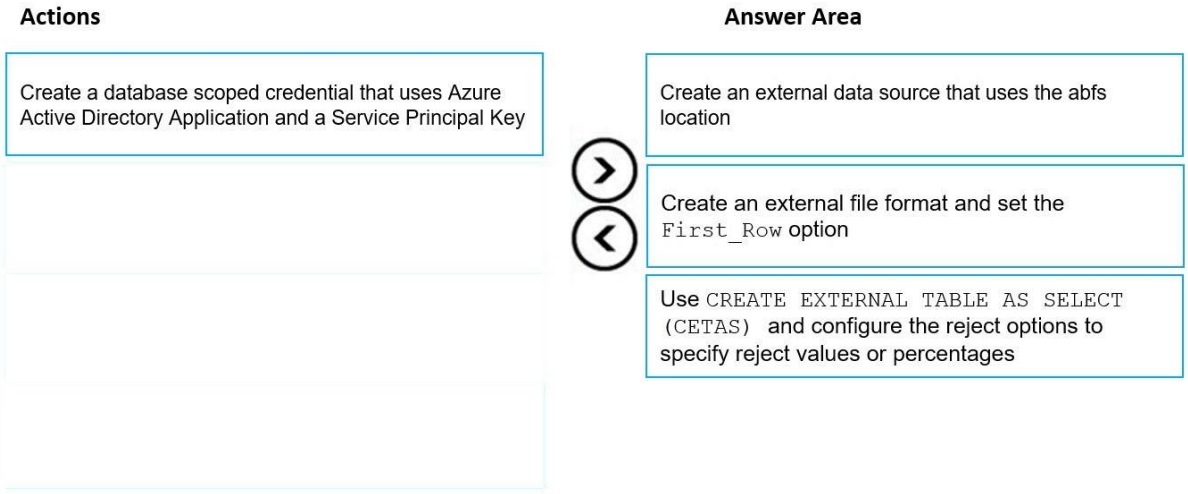
Question #44*Topic 1*
DRAG DROP -
You have data stored in thousands of CSV files in Azure Data Lake Storage Gen2. Each file has a header row followed by a properly formatted carriage return (/ r) and line feed (/n).
You are implementing a pattern that batch loads the files daily into an enterprise data 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

```
(
```

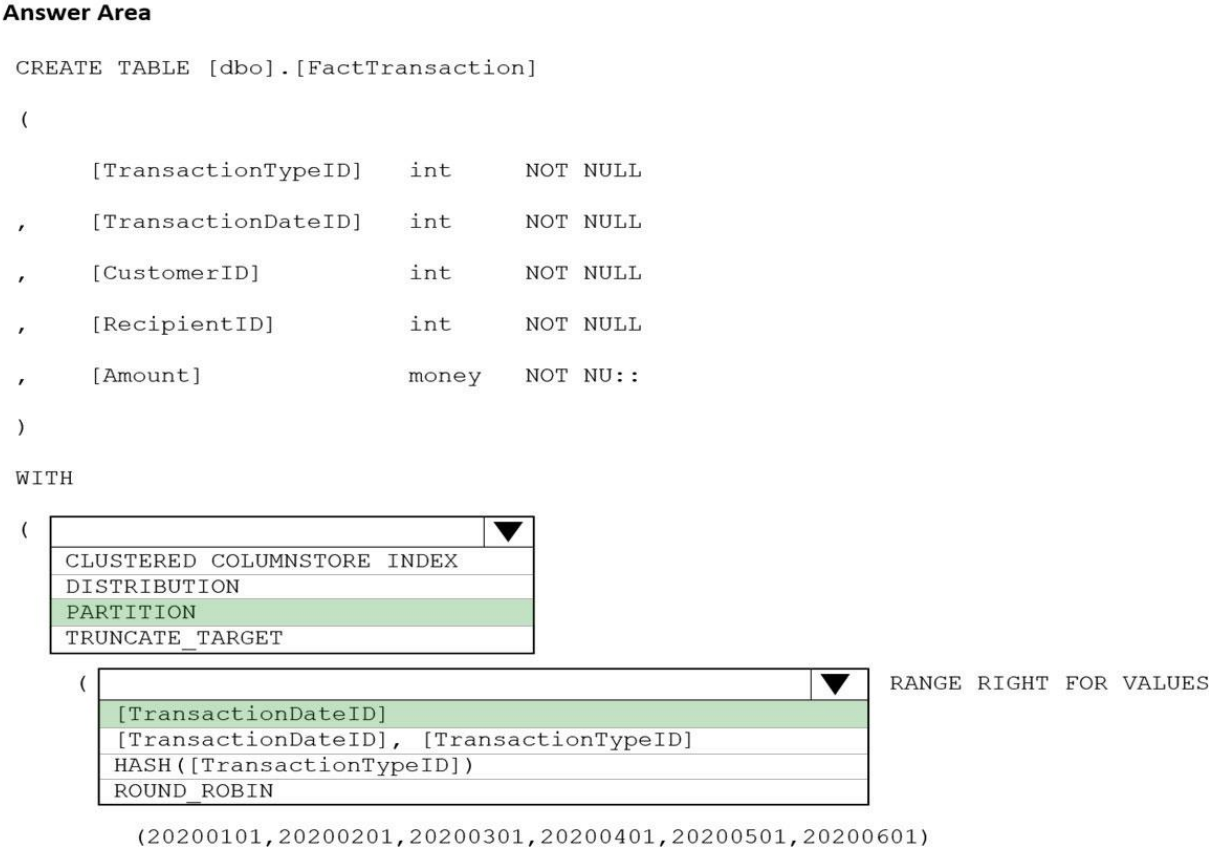
[▼] RANGE RIGHT FOR VALUES
- [TransactionDateID]
- [TransactionDateID], [TransactionTypeID]
- HASH([TransactionTypeID])
- ROUND_ROBIN

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

Hide Solution | Discussion 1

Correct Answer:

Answer Area



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?

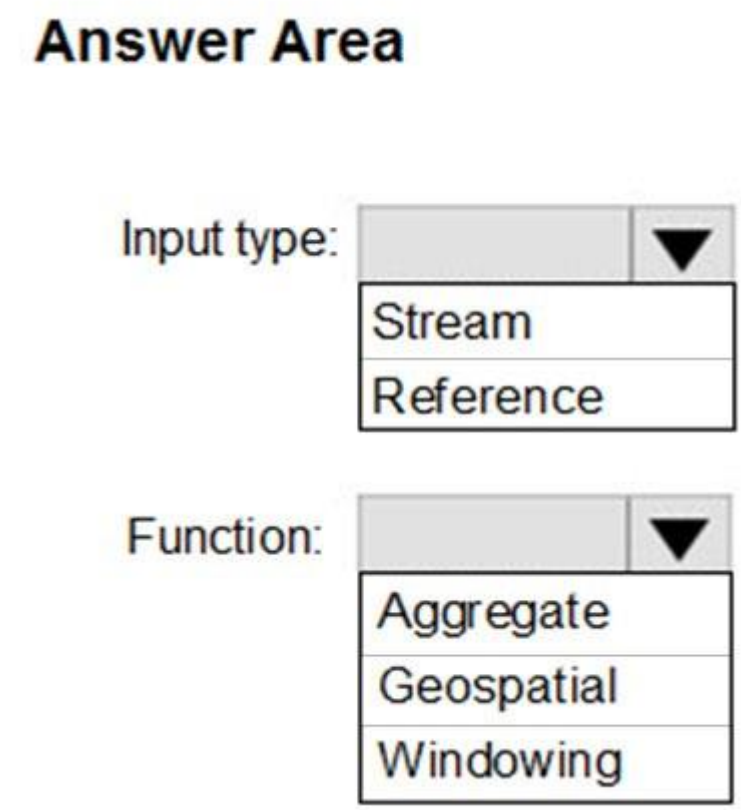
- A. Yes
- B. No

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

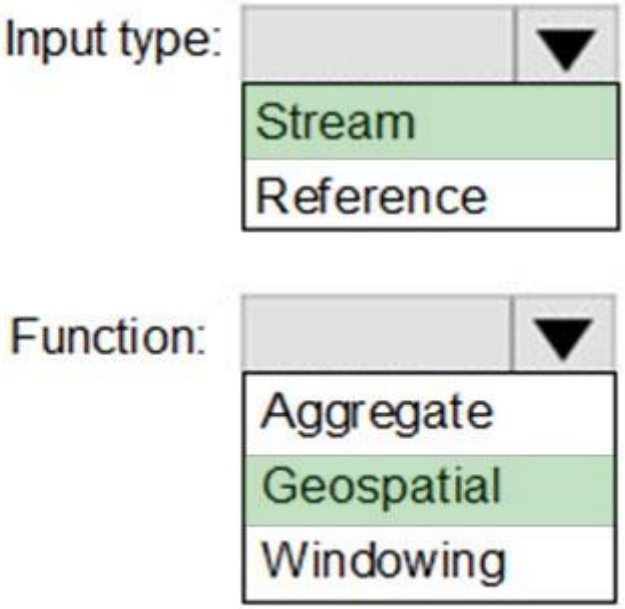
Question #3*Topic 2*
HOTSPOT -
You plan to create a real-time monitoring app that alerts users when a device travels more than 200 meters away from a designated location.
You need to design an Azure Stream Analytics job to process the data for the planned app. The solution must minimize the amount of 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
                [ ▼ ] (Time) AS LastEventTime
                  COUNT
                  MAX
                  MIN
                  TOPONE
        FROM
                Input TIMESTAMP BY Time
        GROUP BY
                [ ▼ ] (minute, 10)
                  HoppingWindow
                  SessionWindow
                  SlidingWindow
                  TumblingWindow
)
SELECT
        Input.License_plate,
        Input.Make,
        Input.Time
FROM
        Input TIMESTAMP BY Time
        INNER JOIN LastInWindow
        ON [ ▼ ] (minute, Input, LastInWindow) BETWEEN 0 AND 10
                  DATEADD
                  DATEDIFF
                  DATENAME
                  DATEPART
AND Input.Time = LastInWindow.LastEventTime
```

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

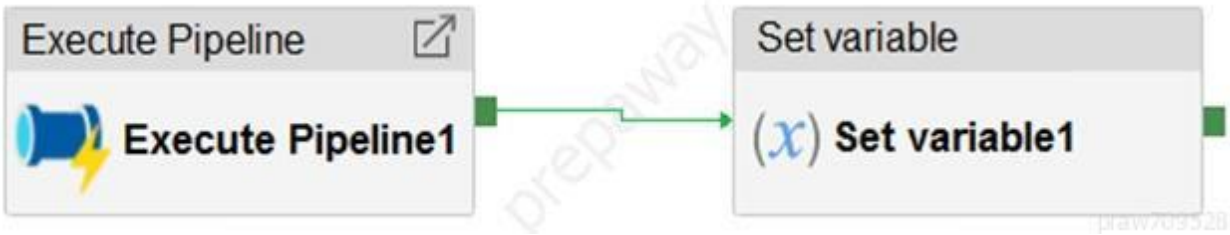
Box 2: TumblingWindow -
Tumbling windows are a series of fixed-sized, non-overlapping and contiguous time intervals.

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.

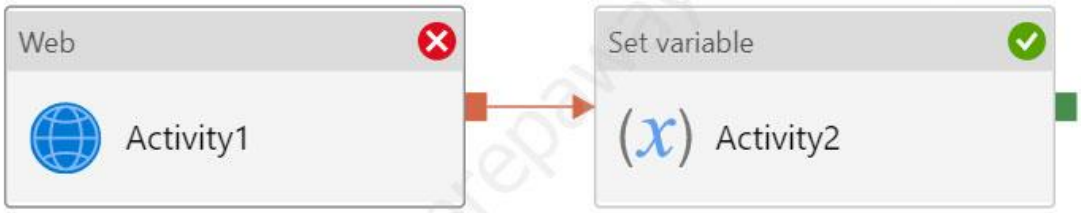
Hide Solution Discussion 9

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:

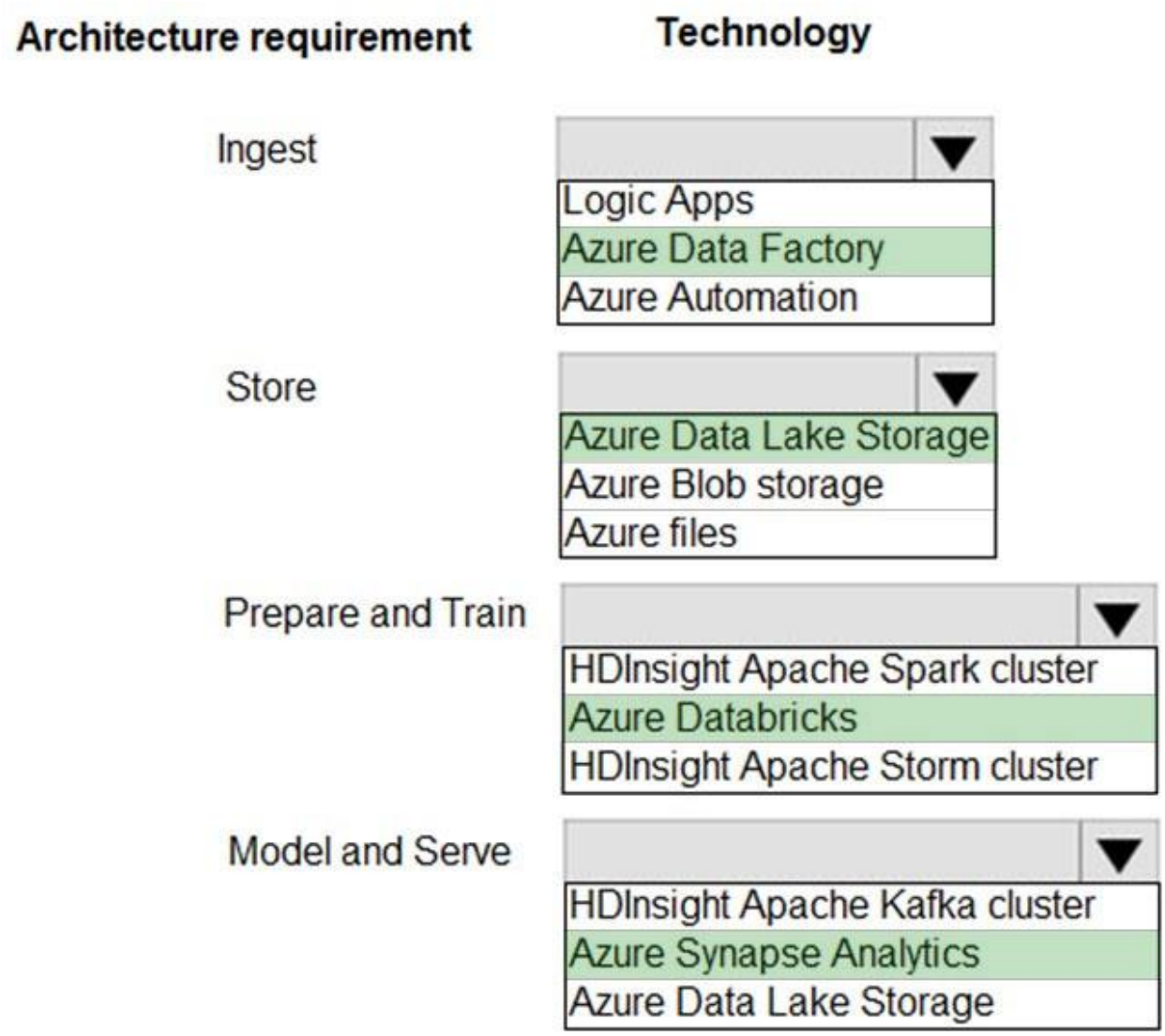
Answer Area

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.

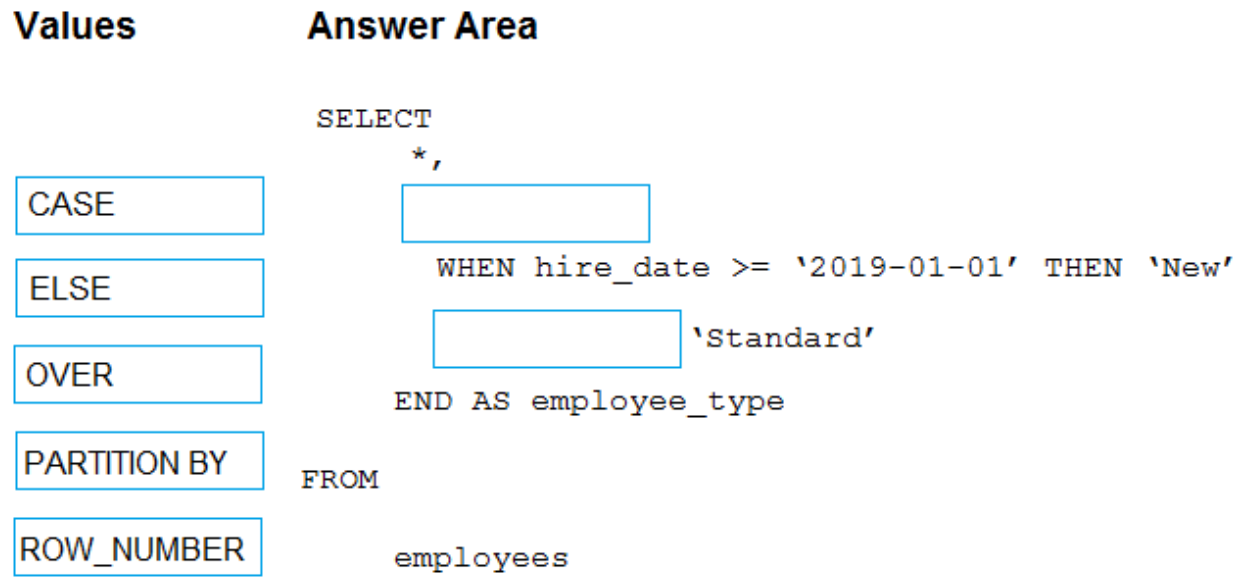
Prepare and Train: Azure Databricks
Azure Databricks provides enterprise-grade Azure security, including Azure Active Directory integration.
With Azure Databricks, you can set up your Apache Spark environment in minutes, autoscale and collaborate on shared projects in an interactive workspace.
Azure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

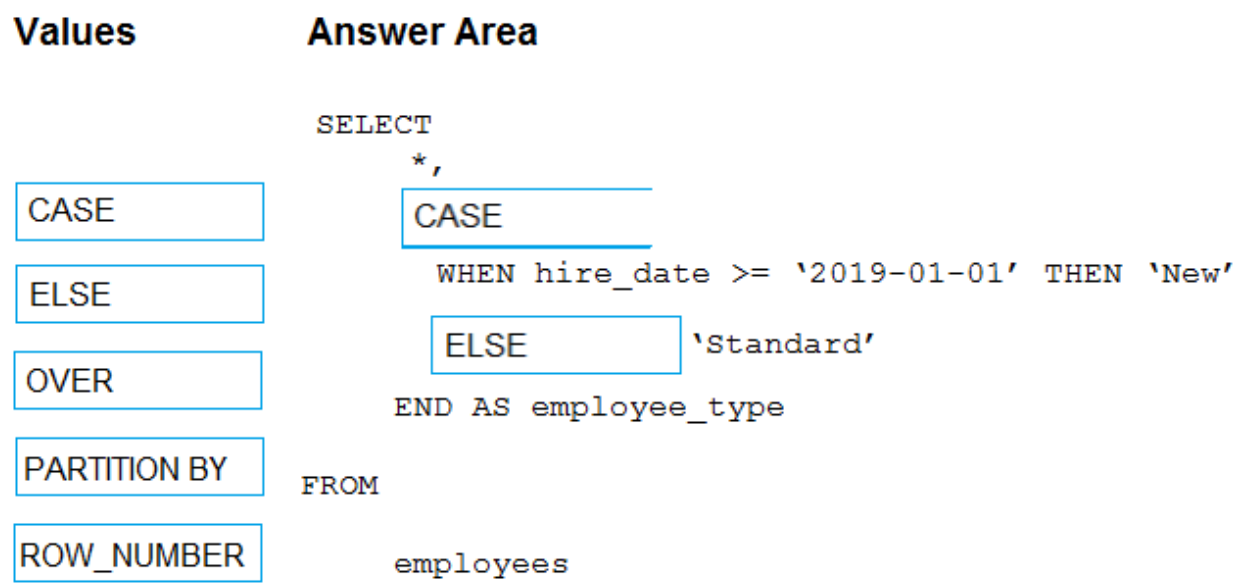
You need to calculate the employee_type value based on the hire_date value.
How should you complete the Transact-SQL statement? To answer, drag the appropriate values to the correct targets. Each value may be used once, more than once, or not at all. You may need to drag the split bar between panes or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