DP-203$^{\text{Q\&As}}$

Data Engineering on Microsoft Azure

Pass Microsoft DP-203 Exam with 100% Guarantee

QUESTION 1

DRAG DROP

You have an Azure subscription that contains an Azure Databricks workspace. The workspace contains a notebook named Notebook1. In Notebook1, you create an Apache Spark DataFrame named df_sales that contains the following columns:

1.

Customer

2.

Salesperson

3.

Region

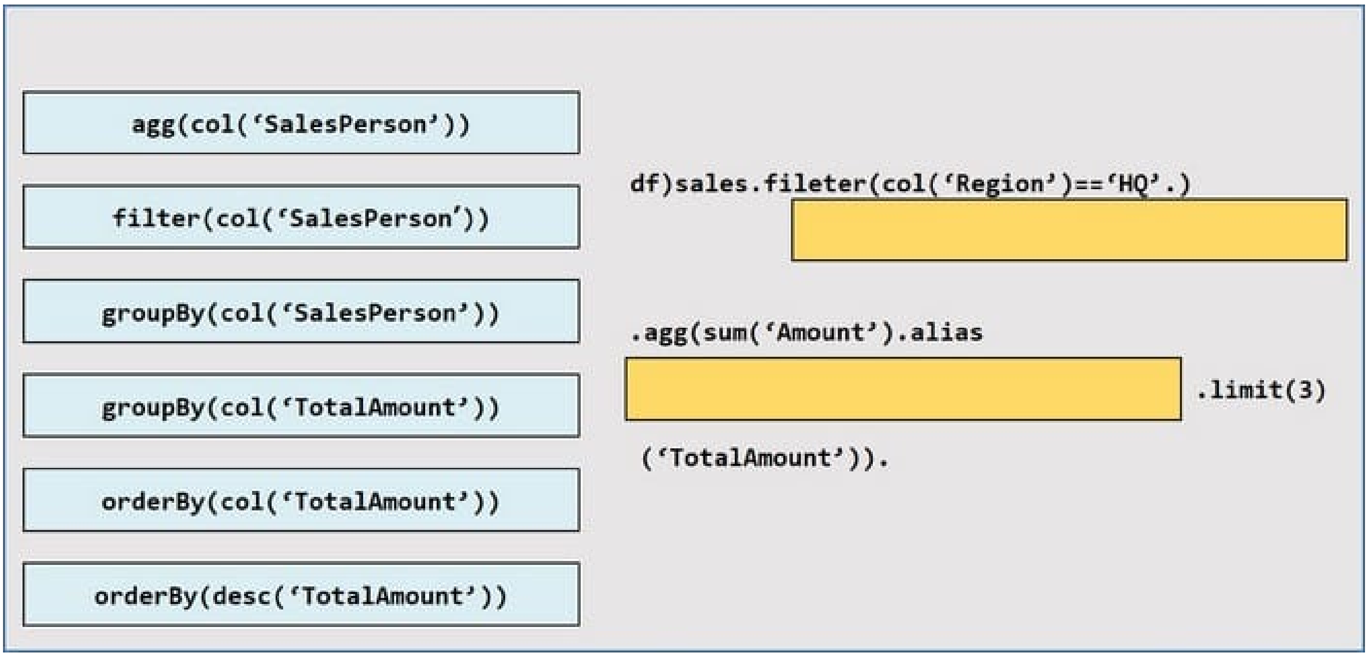
4.

Amount

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

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

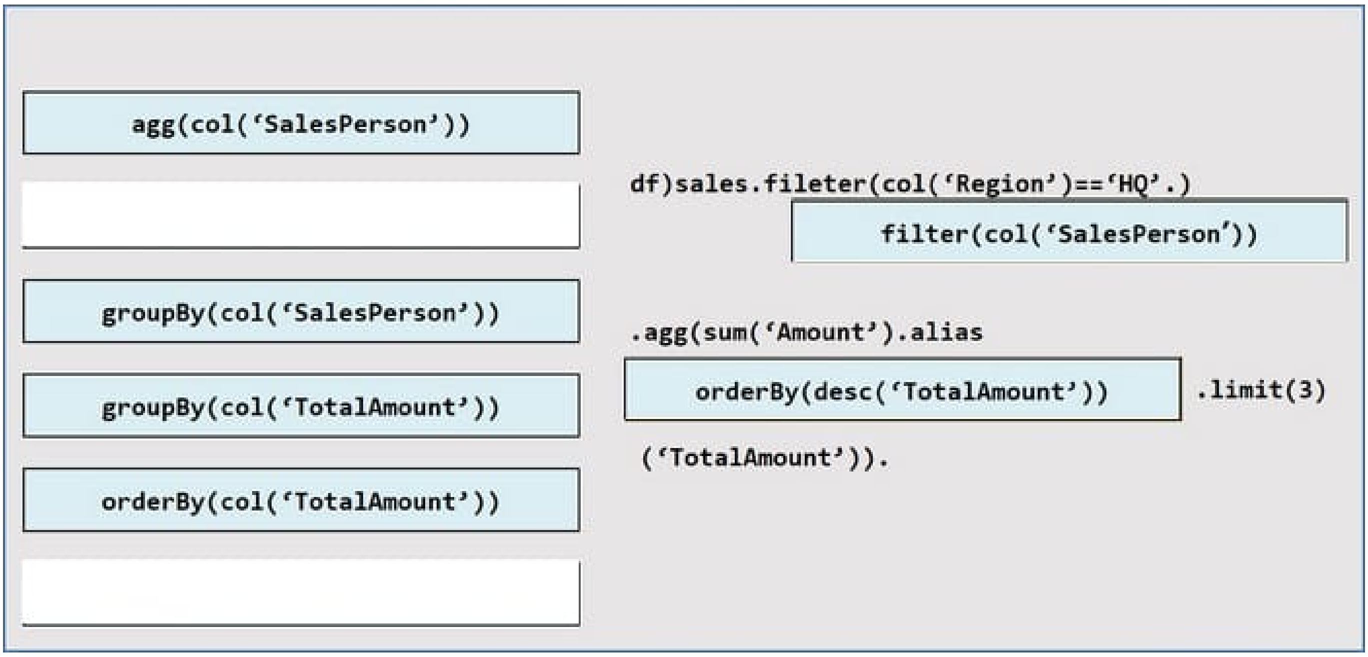
Select and Place:



```
agg(col(‘SalesPerson’))
filter(col(‘SalesPerson’))
groupBy(col(‘SalesPerson’))
groupBy(col(‘TotalAmount’))
orderBy(col(‘TotalAmount’))
orderBy(desc(‘TotalAmount’))
```

```
df)sales.fileter(col(‘Region’)==‘HQ’.)
            ▭
.agg(sum(‘Amount’).alias
▭                              .limit(3)
 (‘TotalAmount’)).
```

Correct Answer:



```
agg(col(‘SalesPerson’))

groupBy(col(‘SalesPerson’))

groupBy(col(‘TotalAmount’))

orderBy(col(‘TotalAmount’))
```

```
df)sales.fileter(col(‘Region’)==‘HQ’.)
            filter(col(‘SalesPerson’))

.agg(sum(‘Amount’).alias
orderBy(desc(‘TotalAmount’))   .limit(3)
(‘TotalAmount’)).
```

QUESTION 2

You have an Azure Factory instance named DF1 that contains a pipeline named PL1.PL1 includes a tumbling window trigger.

You create five clones of PL1. You configure each clone pipeline to use a different data source.

You need to ensure that the execution schedules of the clone pipeline match the execution schedule of PL1.

What should you do?

A. Add a new trigger to each cloned pipeline

B. Associate each cloned pipeline to an existing trigger.

C. Create a tumbling window trigger dependency for the trigger of PL1.

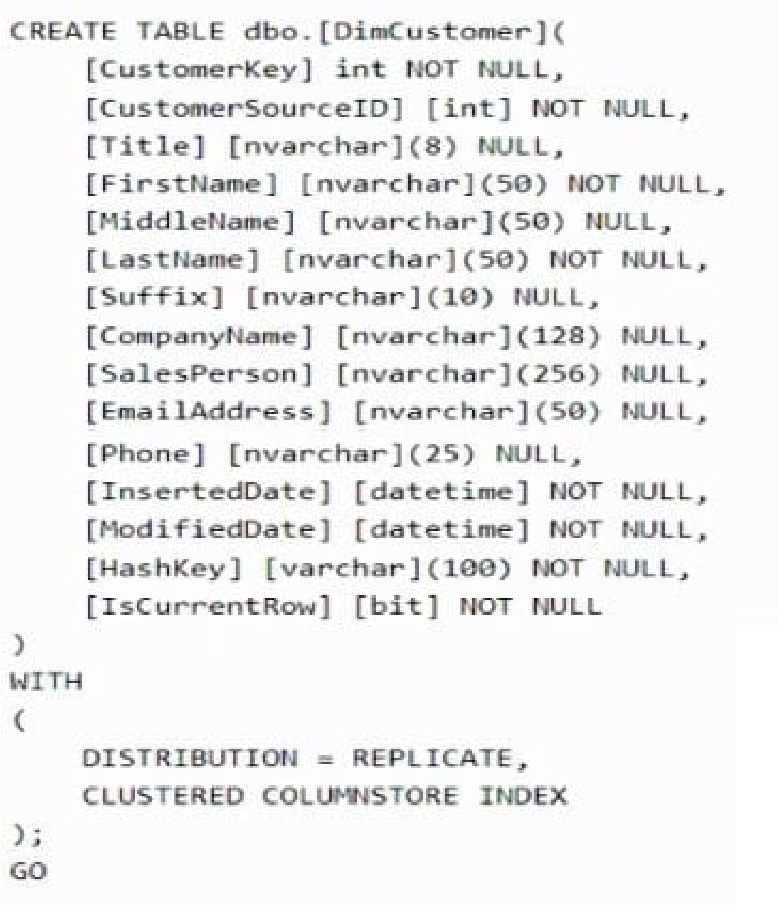
D. Modify the Concurrency setting of each pipeline.

Correct Answer: B

QUESTION 3

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

You plan to implement a star schema in pool1 and create a new table named DimCustomer by using the following code.



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

You need to ensure that DimCustomer has the necessary columns to support a Type 2 slowly changing dimension (SCD).

Which two columns should you add? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

A. [HistoricalSalesPerson] [nvarchar] (256) NOT NULL

B. [EffectiveEndDate] [datetime] NOT NULL

C. [PreviousModifiedDate] [datetime] NOT NULL

D. [RowID] [bigint] NOT NULL

E. [EffectiveStartDate] [datetime] NOT NULL

Correct Answer: BE

"For the SCD Type 2, we need to include three more attributes such as StartDate, EndDate and IsCurrent"

IsCurrentRow is already present! ... ;-)

CustomerKey (in reality is the RowID that many guys wants to add here),

effectiveEndDate will probably set to: 31.12.9999, (to justify the not null).

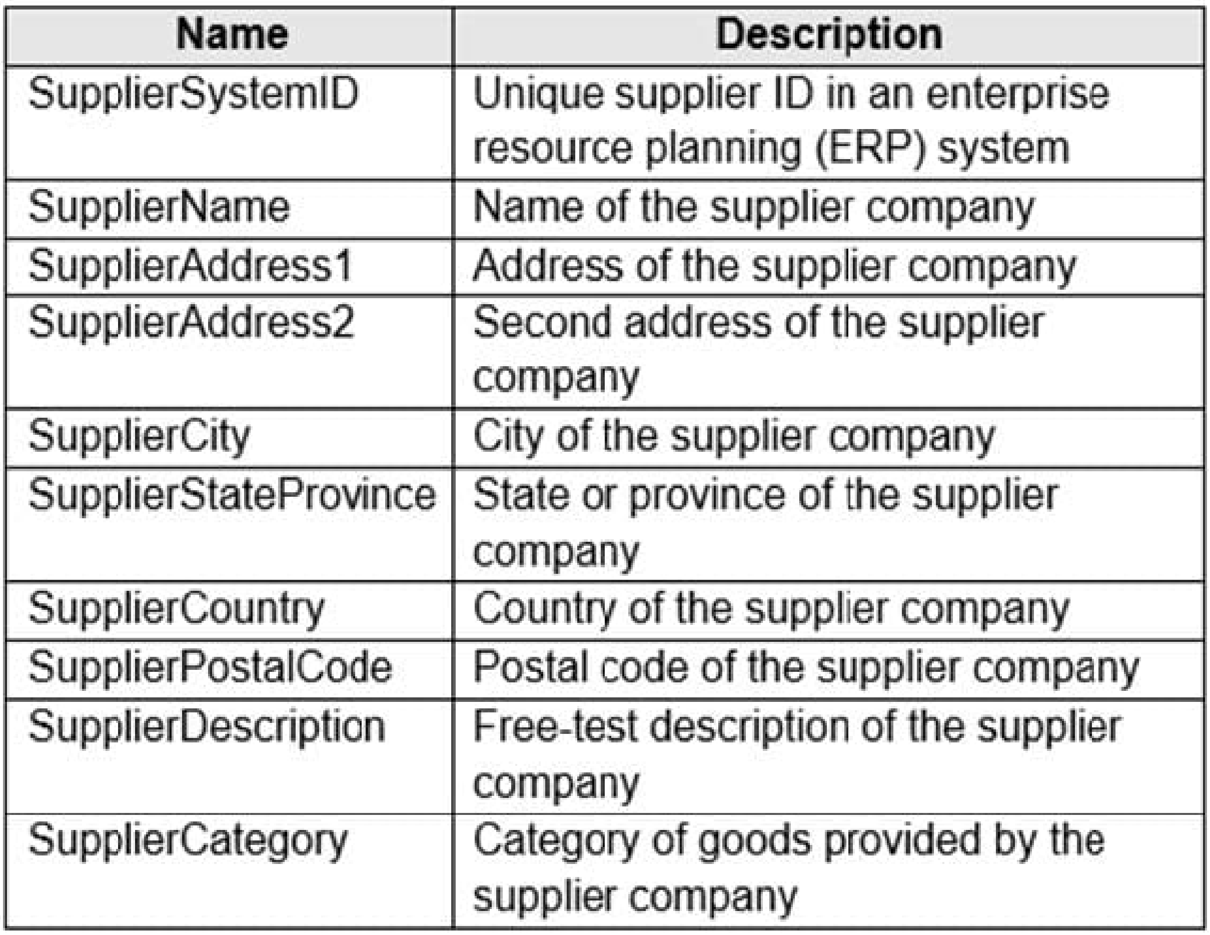
https://www.sqlshack.com/implementing-slowly-changing-dimensions-scds-in-data-warehouses/

QUESTION 4

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

B. effective start date

C. business key

D. last modified date

E. effective end date

F. foreign key

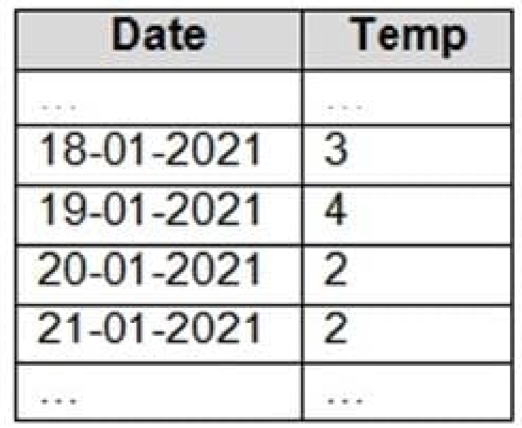
Correct Answer: ABE

Reference: https://docs.microsoft.com/en-us/sql/integration-services/data-flow/transformations/slowly-changing-dimension-transformation

QUESTION 5

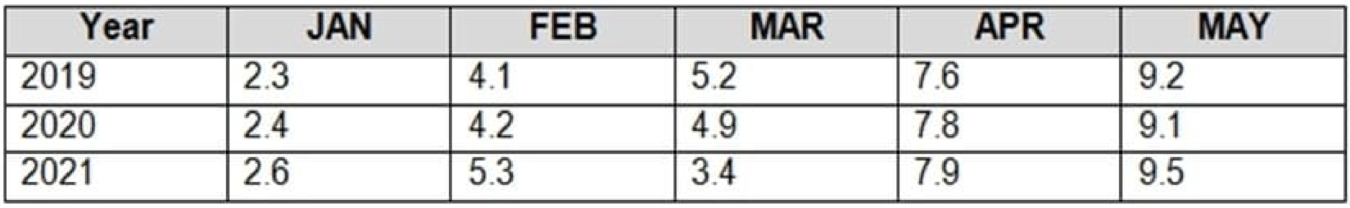
DRAG DROP

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



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

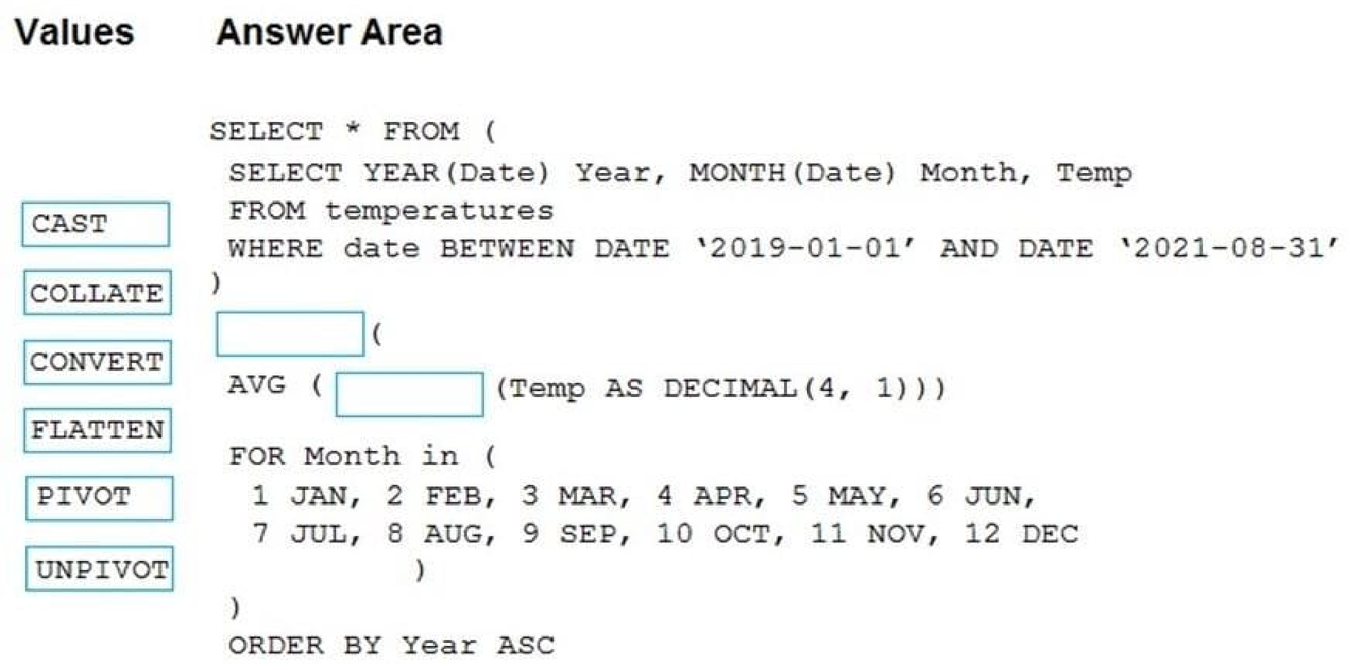
You need to produce the following table by using a Spark SQL query.



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

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

Select and Place:



Values

CAST

COLLATE

CONVERT

FLATTEN

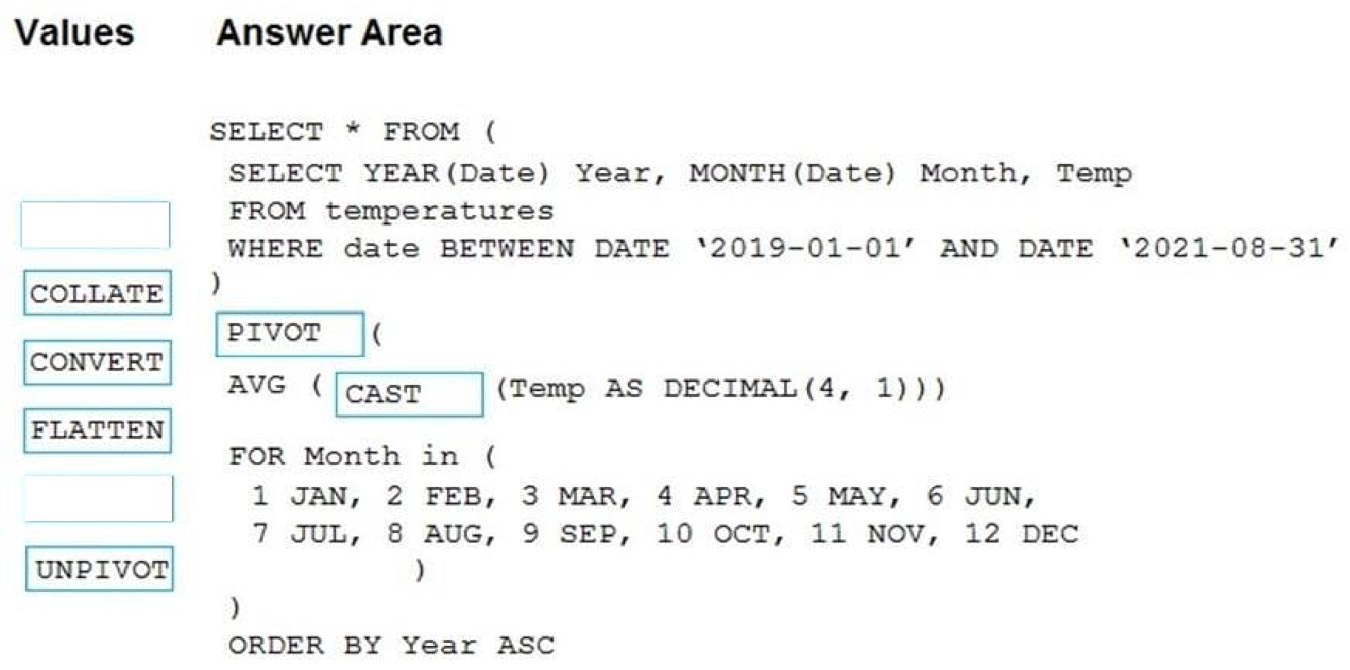
PIVOT

UNPIVOT

Answer Area

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

Correct Answer:



Values

COLLATE

CONVERT

FLATTEN

UNPIVOT

Answer Area

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

QUESTION 6

You have a SQL pool in Azure Synapse.

A user reports that queries against the pool take longer than expected to complete.

You need to add monitoring to the underlying storage to help diagnose the issue.

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

NOTE: Each correct selection is worth one point.

A. Cache used percentage

B. DWU Limit

C. Snapshot Storage Size

D. Active queries

E. Cache hit percentage

Correct Answer: AE

A: Cache used is the sum of all bytes in the local SSD cache across all nodes and cache capacity is the sum of the storage capacity of the local SSD cache across all nodes.

E: Cache hits is the sum of all columnstore segments hits in the local SSD cache and cache miss is the columnstore segments misses in the local SSD cache summed across all nodes

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

QUESTION 7

You have an enterprise data warehouse in Azure Synapse Analytics named DW1 on a server named Server1. You need to verify whether the size of the transaction log file for each distribution of DW1 is smaller than 160 GB. What should you do?

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

B. From Azure Monitor in the Azure portal, execute a query against the logs of DW1.

C. On DW1, execute a query against the sys.database_files dynamic management view.

D. Execute a query against the logs of DW1 by using the Get-AzOperationalInsightSearchResult PowerShell cmdlet.

Correct Answer: A

The following query returns the transaction log size on each distribution. If one of the log files is reaching 160 GB, you should consider scaling up your instance or limiting your transaction size.

-- Transaction log size SELECT instance_name as distribution_db, cntr_value*1.0/1048576 as log_file_size_used_GB,pdw_node_id FROM sys.dm_pdw_nodes_os_performance_counters WHERE instance_name like \\'Distribution_%\\' AND counter_name = \\'Log File(s) Used Size (KB)\\'

References: https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-managemonitor

QUESTION 8

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while

others might not have a correct solution.

After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen. You 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 SOL. A workload that data scientists will use to perform ad hoc analysis in Scala and R.

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

The data engineers must share a cluster.

The job cluster will be managed by using a request process whereby data scientists and data engineers provide packaged notebooks for deployment to the cluster.

All the data scientists must be assigned their own cluster that terminates automatically after 120 minutes of inactivity. Currently, there are three data scientists.

You need to create the Databricks clusters for the workloads.

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

Does this meet the goal?

A. Yes

B. No

Correct Answer: B

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 9

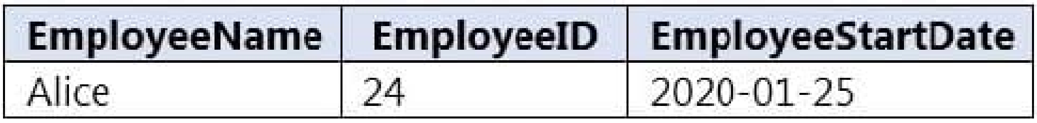
You have an Azure Synapse workspace named MyWorkspace that contains an Apache Spark database named mytestdb.

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 EmployeeIDFROM mytestdb.dbo.myParquetTableWHERE EmployeeName = \\'Alice\\';

What will be returned by the query?

A. 24

B. an error

C. a null value

Correct Answer: B

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 10

HOTSPOT

You have an Azure Synapse Analytics dedicated SQL pool.

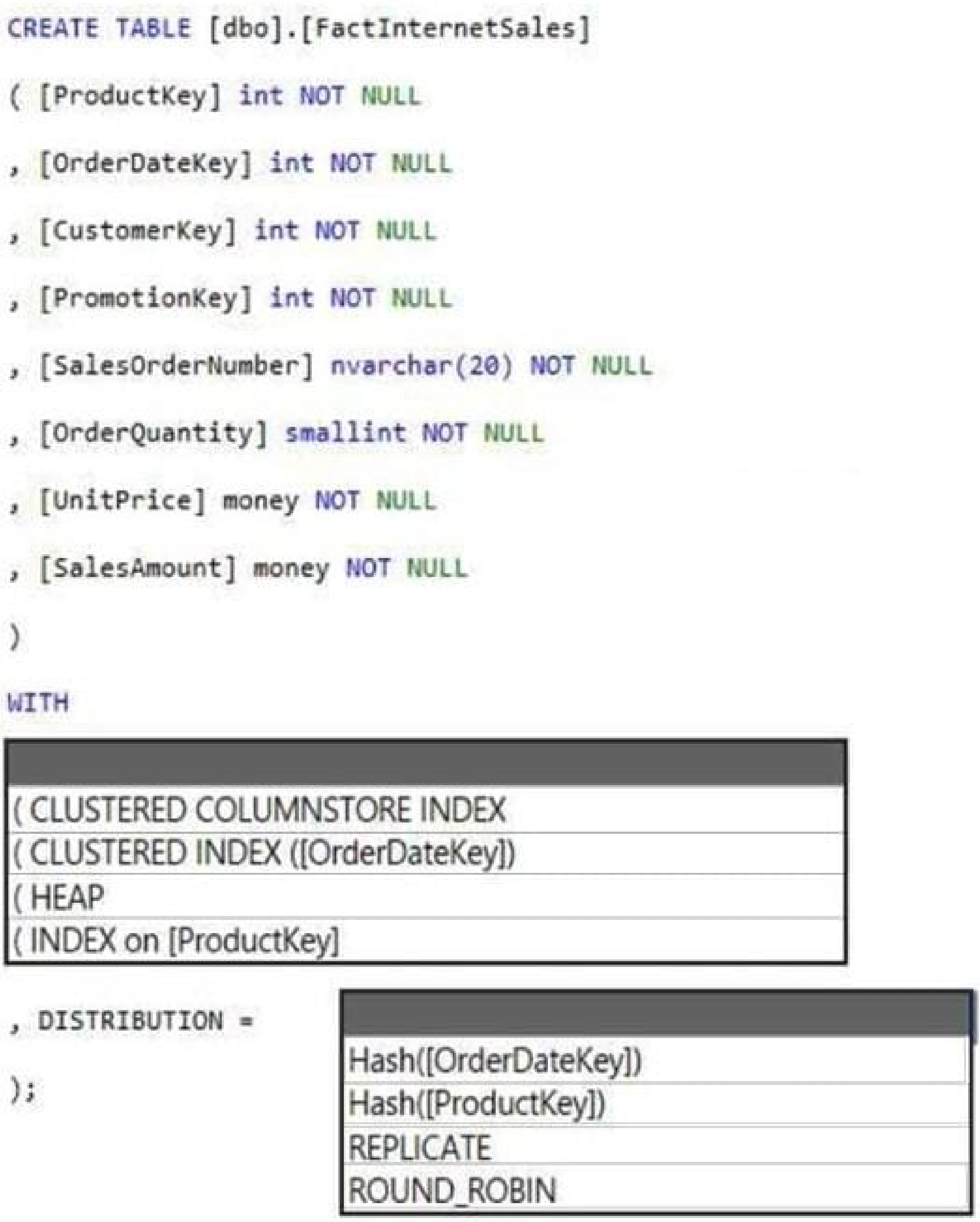
You need to create a table named FactInternetSales that will be a large fact table in a dimensional model. FactInternetSales will contain 100 million rows and two columns named SalesAmount and OrderQuantity. Queries executed on

FactInternetSales will aggregate the values in SalesAmount and OrderQuantity from the last year for a specific product. The solution must minimize the data size and query execution time.

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

NOTE: Each correct selection is worth one point.

Hot Area:



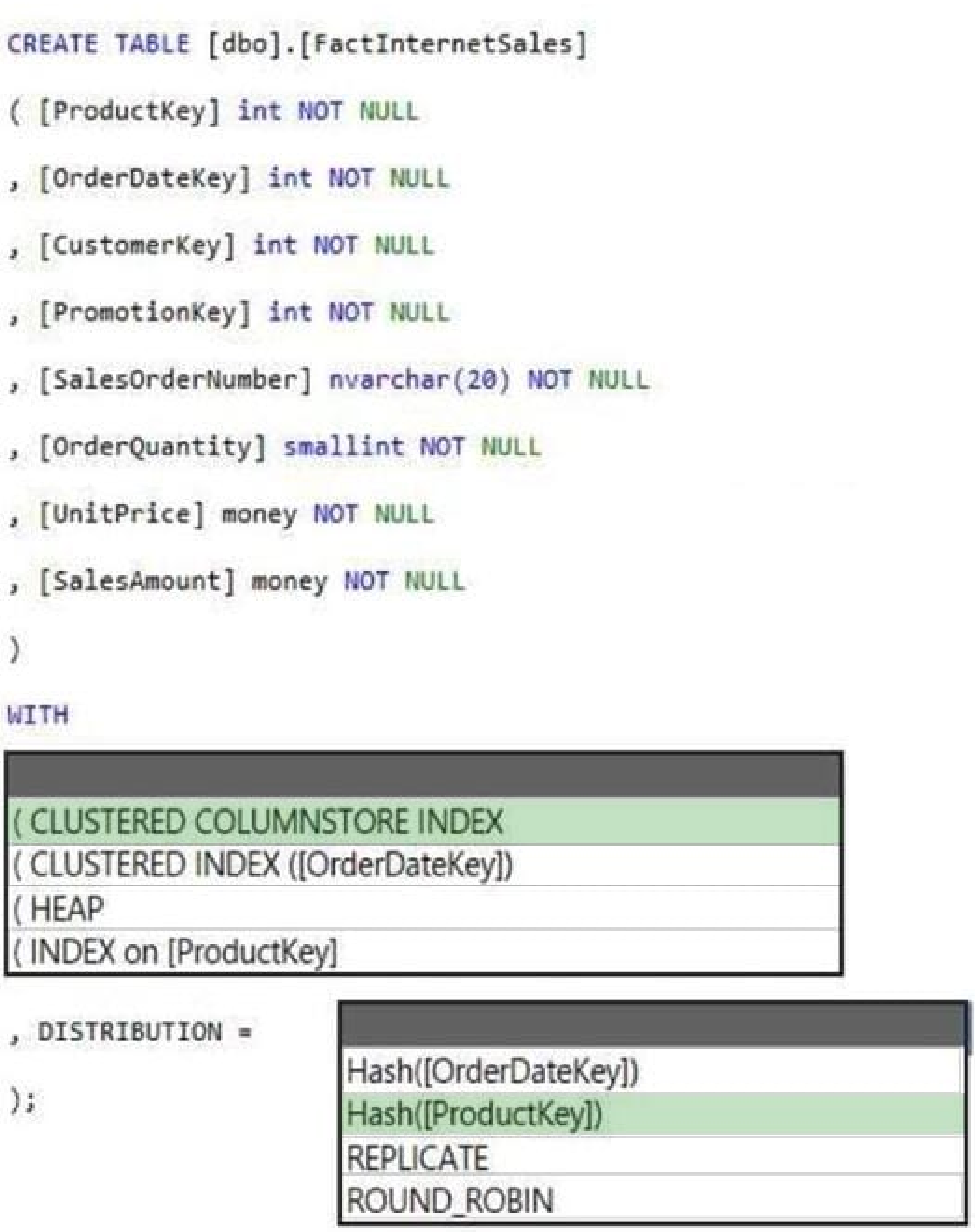
```
CREATE TABLE [dbo].[FactInternetSales]
( [ProductKey] int NOT NULL
, [OrderDateKey] int NOT NULL
, [CustomerKey] int NOT NULL
, [PromotionKey] int NOT NULL
, [SalesOrderNumber] nvarchar(20) NOT NULL
, [OrderQuantity] smallint NOT NULL
, [UnitPrice] money NOT NULL
, [SalesAmount] money NOT NULL
)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

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

```
, DISTRIBUTION =
);
```

Hash([OrderDateKey])
Hash([ProductKey])
REPLICATE
ROUND_ROBIN

Correct Answer:



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

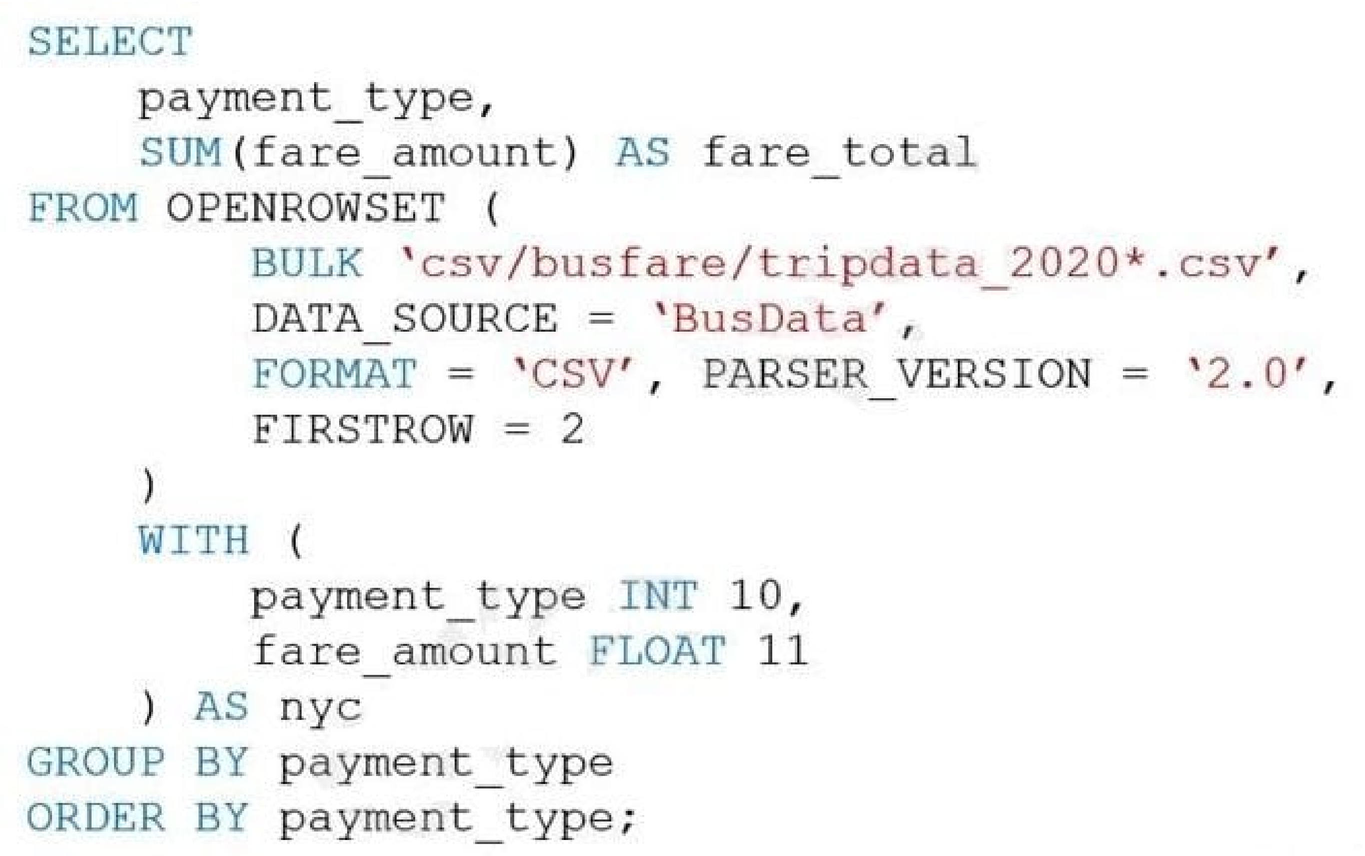
- [✓ (CLUSTERED COLUMNSTORE INDEX]
- (CLUSTERED INDEX ([OrderDateKey])
- (HEAP
- (INDEX on [ProductKey]

```
, DISTRIBUTION =
);
```

- Hash([OrderDateKey])
- [✓ Hash([ProductKey])]
- REPLICATE
- ROUND_ROBIN

QUESTION 11

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



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

What do the query results include?

A. Only CSV files in the tripdata_2020 subfolder.

B. All files that have file names that beginning with "tripdata_2020".

C. All CSV files that have file names that contain "tripdata_2020".

D. Only CSV that have file names that beginning with "tripdata_2020".

Correct Answer: D

QUESTION 12

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:

1.

A workload for data engineers who will use Python and SQL.

2.

A workload for jobs that will run notebooks that use Python, Scala, and SOL.

3.

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:

1.

The data engineers must share a cluster.

2.

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.

3.

All the data scientists must be assigned their own cluster that terminates automatically after 120 minutes of inactivity. Currently, there are three data scientists.

You need to create the Databricks clusters for the workloads.

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

Does this meet the goal?

A. Yes

B. No

Correct Answer: 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 13

HOTSPOT

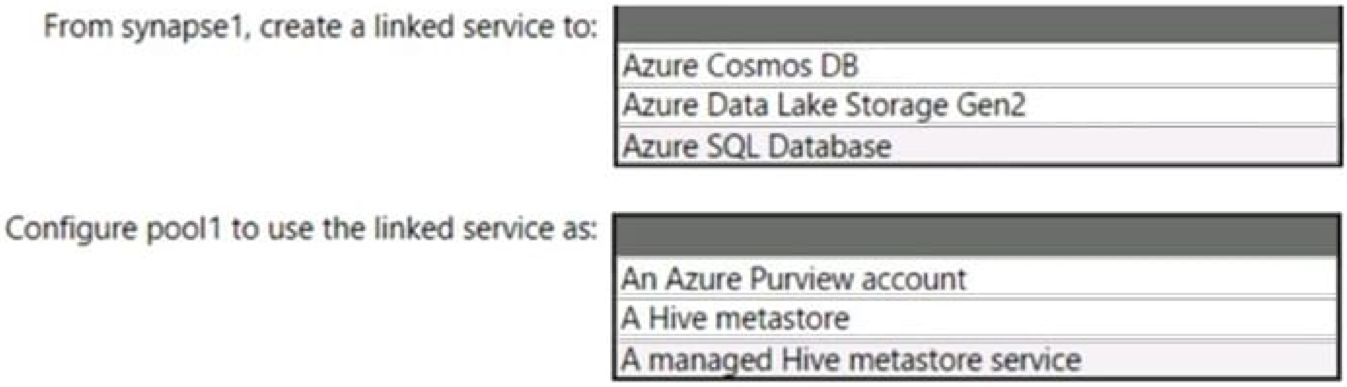
You have an Azure subscription that contains an Azure Databricks workspace named databricks1 and an Azure Synapse Analytics workspace named synapse1. The synapse1 workspace contains an Apache Spark pool named pool1.

You need to share an Apache Hive catalog of pool1 with databricks1.

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

NOTE: Each correct selection is worth one point.

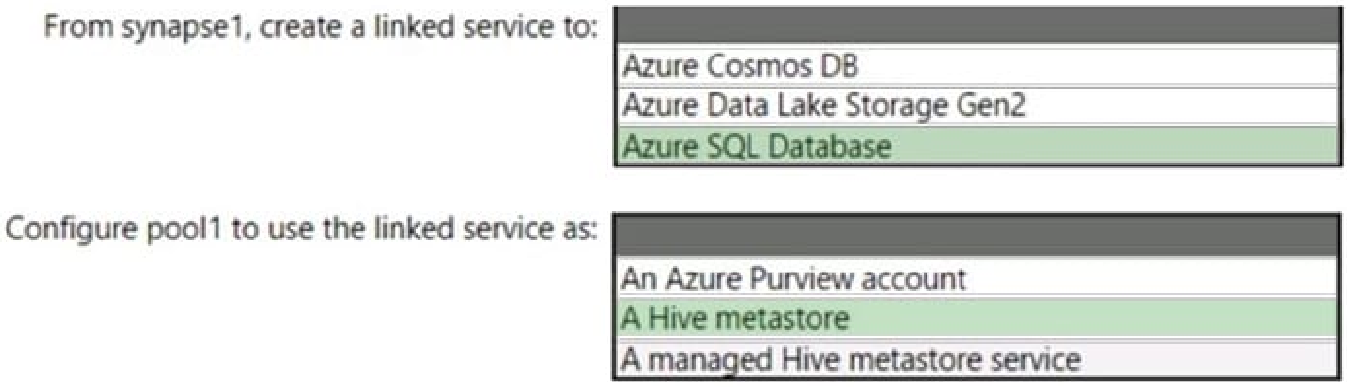
Hot Area:



From synapse1, create a linked service to:
- Azure Cosmos DB
- Azure Data Lake Storage Gen2
- Azure SQL Database

Configure pool1 to use the linked service as:
- An Azure Purview account
- A Hive metastore
- A managed Hive metastore service

Correct Answer:



From synapse1, create a linked service to:
- Azure Cosmos DB
- Azure Data Lake Storage Gen2
- [✓ Azure SQL Database]

Configure pool1 to use the linked service as:
- An Azure Purview account
- [✓ A Hive metastore]
- A managed Hive metastore service

Box 1: Azure SQL Database

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

Set up linked service to Hive Metastore

Follow below steps to set up a linked service to the external Hive Metastore in Synapse workspace.

Open Synapse Studio, go to Manage > Linked services at left, click New to create a new linked service.

Set up Hive Metastore linked service

Choose Azure SQL Database or Azure Database for MySQL based on your database type, click Continue.

Provide Name of the linked service. Record the name of the linked service, this info will be used to configure Spark shortly.

You can either select Azure SQL Database/Azure Database for MySQL for the external Hive Metastore from Azure subscription list, or enter the info manually.

Provide User name and Password to set up the connection.

Test connection to verify the username and password.

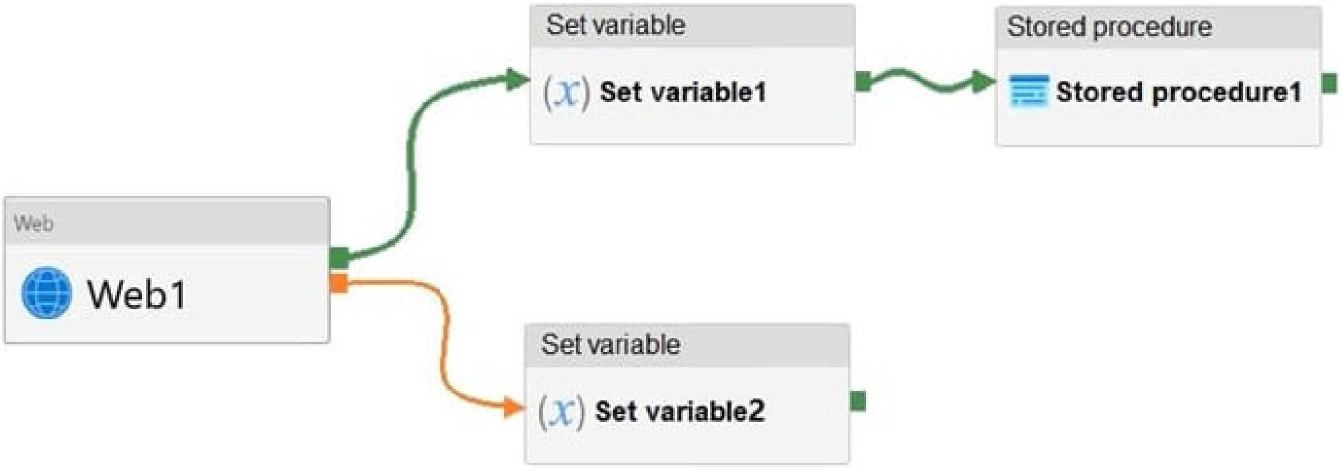
Click Create to create the linked service.

Box 2: A Hive Metastore

QUESTION 14

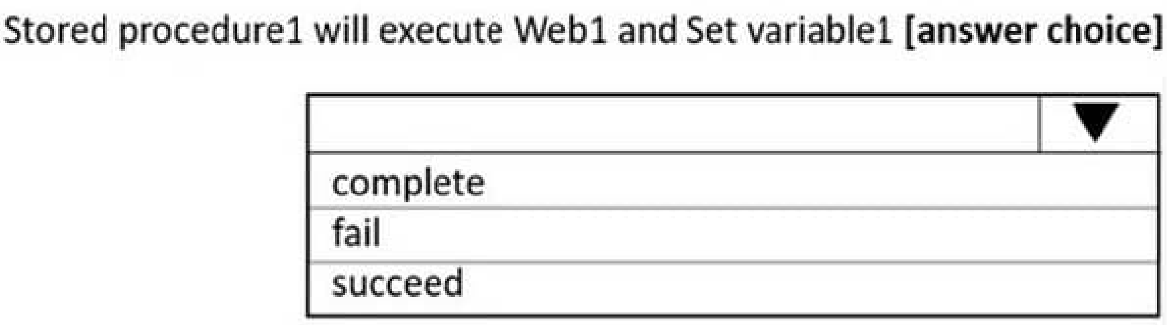
HOTSPOT

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



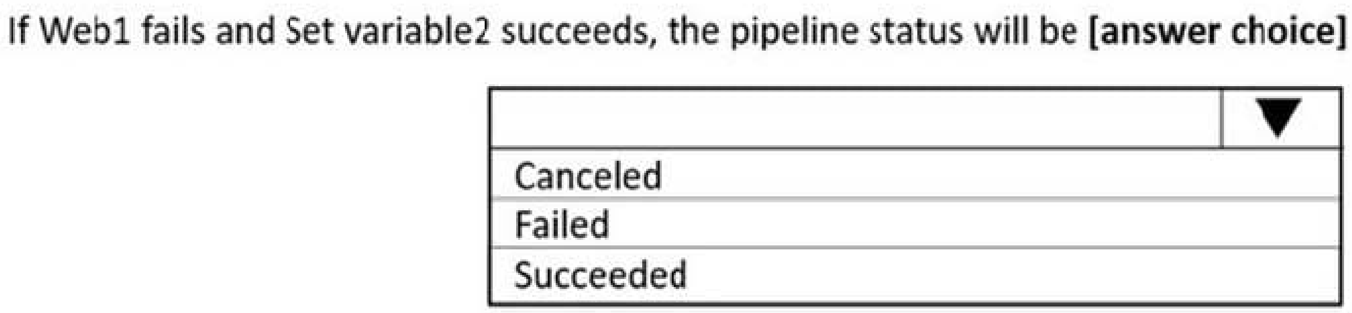
Use the drop-down menus to select the answer choice that completes each statement based on the information presented in the graphic. NOTE: Each correct selection is worth one point.

Hot Area:



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

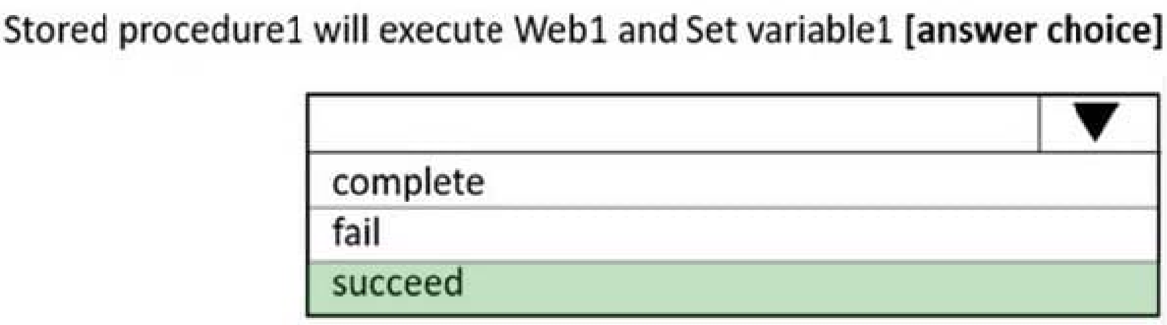
- complete
- fail
- succeed



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

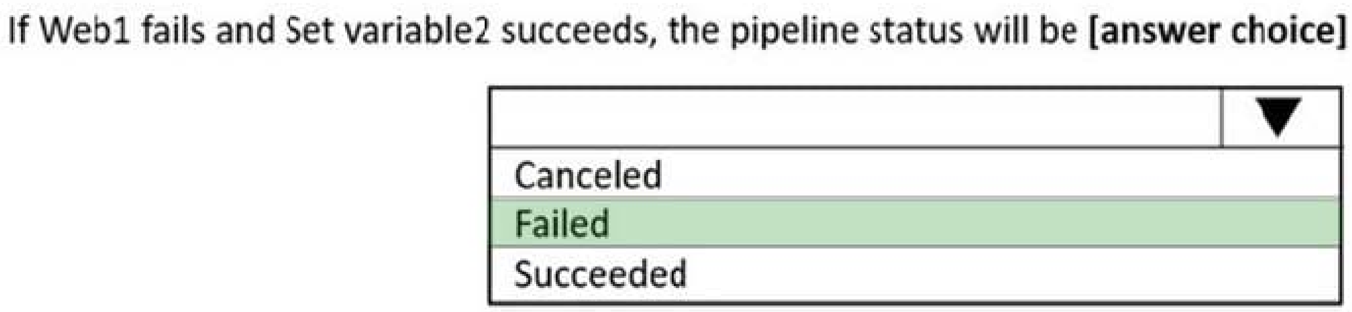
- Canceled
- Failed
- Succeeded

Correct Answer:



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

- complete
- fail
- [✓ succeed]



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

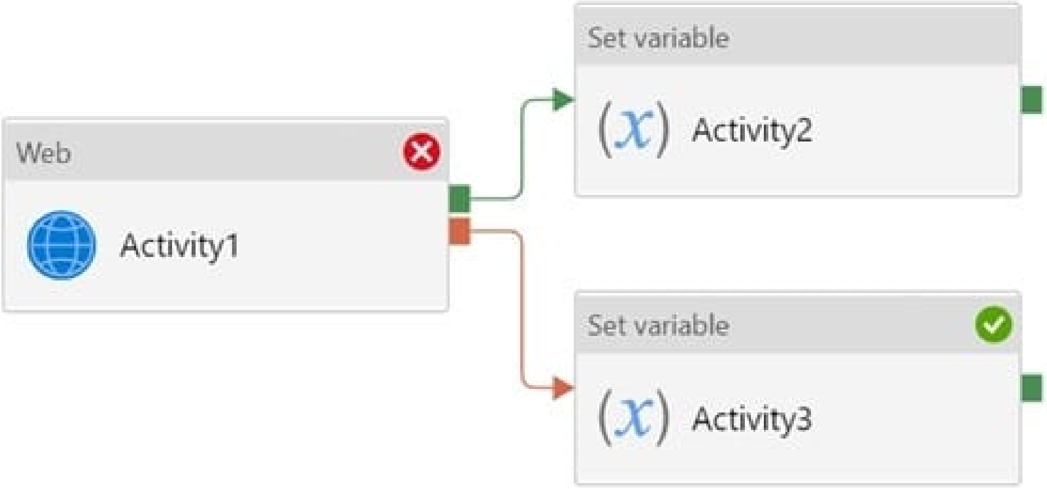
- Canceled
- [✓ Failed]
- Succeeded

Box 1: succeed

Box 2: failed

Example:

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



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

QUESTION 15

DRAG DROP

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

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

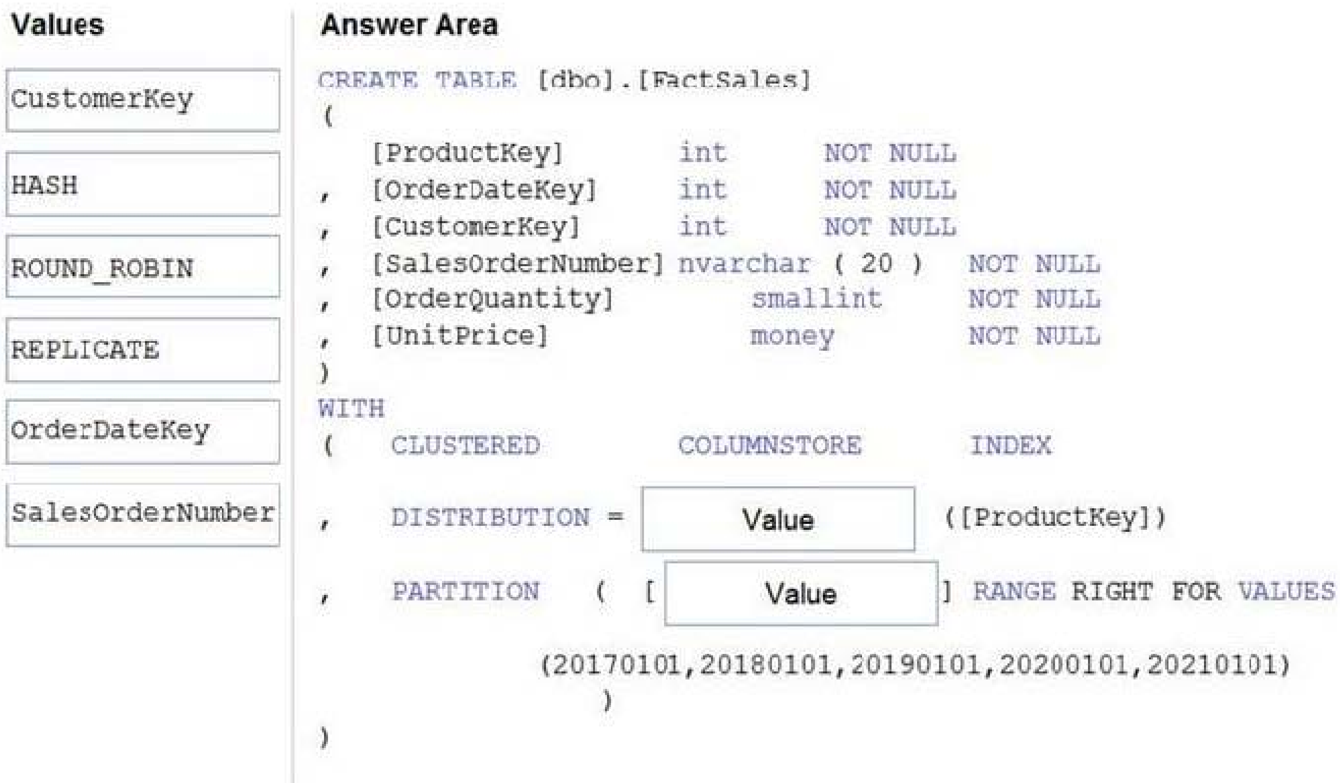
You need to ensure that the data is distributed evenly across partitions. The solution must minimize the amount of time required to delete old data.

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

view content.

NOTE: Each correct selection is worth one point.

Select and Place:



Values

- CustomerKey
- HASH
- ROUND_ROBIN
- REPLICATE
- OrderDateKey
- SalesOrderNumber

Answer Area

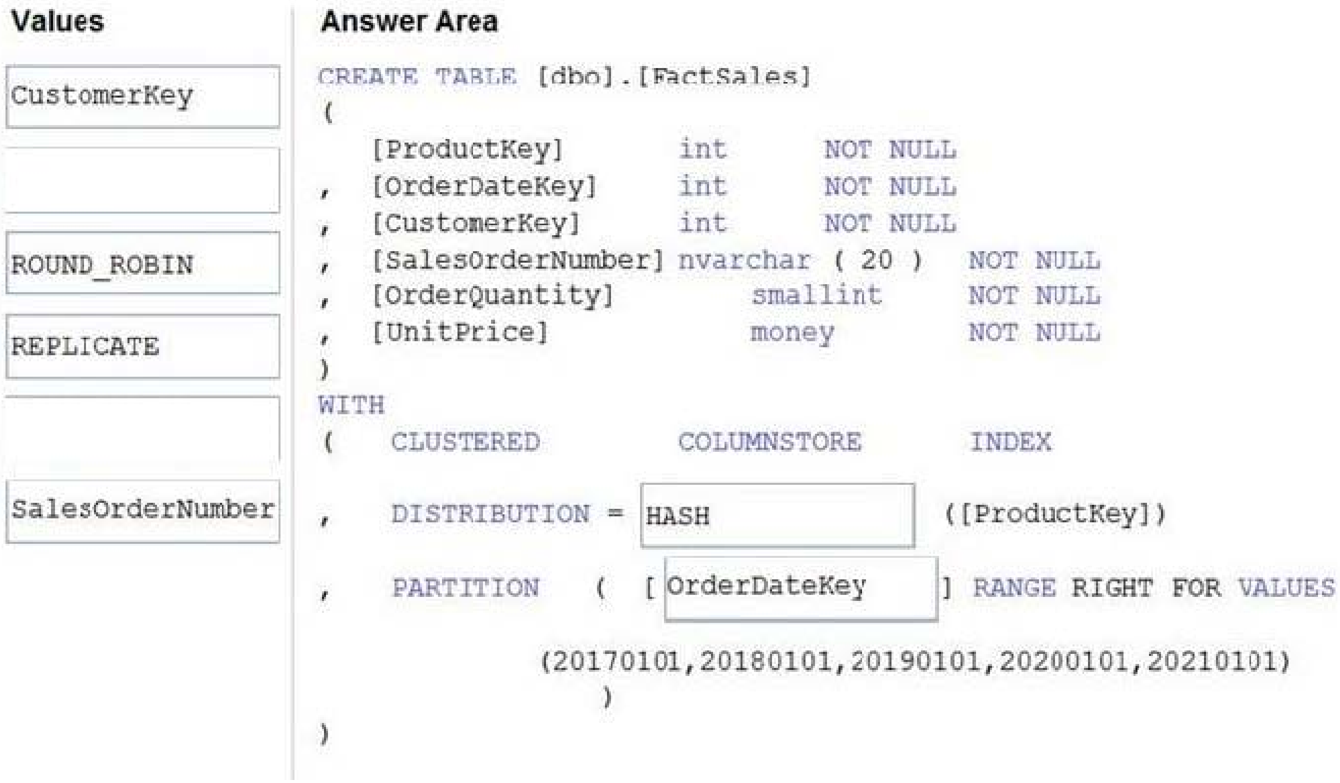
```
CREATE TABLE [dbo].[FactSales]
(
    [ProductKey]         int         NOT NULL
,   [OrderDateKey]       int         NOT NULL
,   [CustomerKey]        int         NOT NULL
,   [SalesOrderNumber] nvarchar ( 20 )   NOT NULL
,   [OrderQuantity]          smallint      NOT NULL
,   [UnitPrice]              money         NOT NULL
)
WITH
(   CLUSTERED          COLUMNSTORE        INDEX

,   DISTRIBUTION =  [ Value ]  ([ProductKey])

,   PARTITION    (  [ [ Value ] ] RANGE RIGHT FOR VALUES

                 (20170101,20180101,20190101,20200101,20210101)
                    )
)
```

Correct Answer:



Values

- CustomerKey
-
- ROUND_ROBIN
- REPLICATE
-
- SalesOrderNumber

Answer Area

```
CREATE TABLE [dbo].[FactSales]
(
    [ProductKey]         int         NOT NULL
,   [OrderDateKey]       int         NOT NULL
,   [CustomerKey]        int         NOT NULL
,   [SalesOrderNumber] nvarchar ( 20 )   NOT NULL
,   [OrderQuantity]          smallint      NOT NULL
,   [UnitPrice]              money         NOT NULL
)
WITH
(   CLUSTERED          COLUMNSTORE        INDEX

,   DISTRIBUTION =  HASH           ([ProductKey])

,   PARTITION    (  [ OrderDateKey  ] RANGE RIGHT FOR VALUES

                 (20170101,20180101,20190101,20200101,20210101)
                    )
)
```

Box 1: HASH

Box 2: OrderDateKey

In most cases, table partitions are created on a date column.

A way to eliminate rollbacks is to use Metadata Only operations like partition switching for data management. For example, rather than execute a DELETE statement to delete all rows in a table where the order_date was in October of 2001,

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