DP-203 Q&As

Data Engineering on Microsoft Azure

Pass Microsoft DP-203 Exam with 100% Guarantee

Free Download Real Questions & Answers **PDF** and **VCE** file from:

https://www.passapply.com/dp-203.html

100% Passing Guarantee
100% Money Back Assurance

Following Questions and Answers are all new published by Microsoft Official Exam Center

QUESTION 1

HOTSPOT

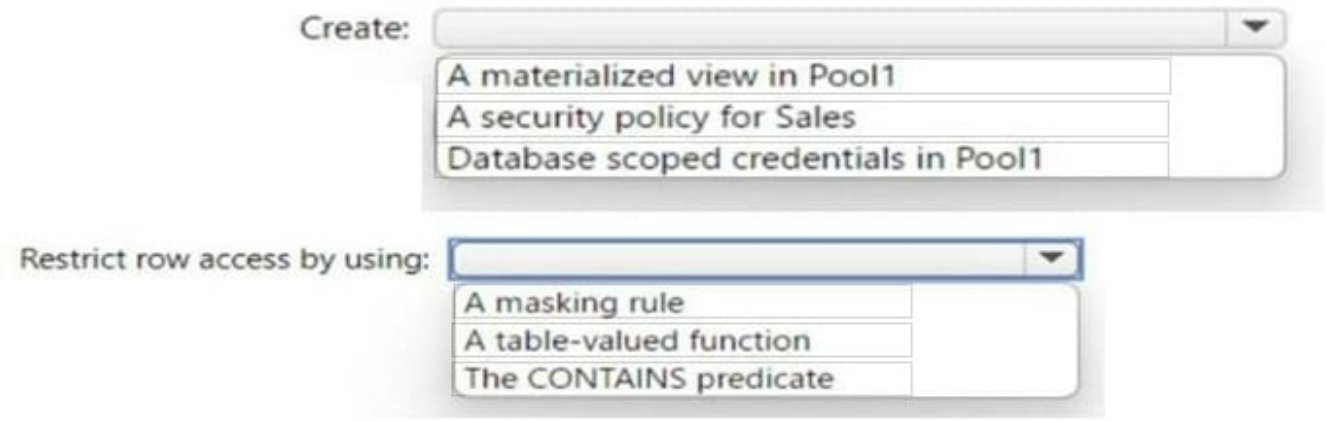
You have an Azure Synapse Analytics dedicated SQL pool named Pool1 that contains an external table named Sales. Sales contains sales data. Each row in Sales contains data on a single sale, including the name of the salesperson.

You need to implement row-level security (RLS). The solution must ensure that the salespeople can access only their respective sales.

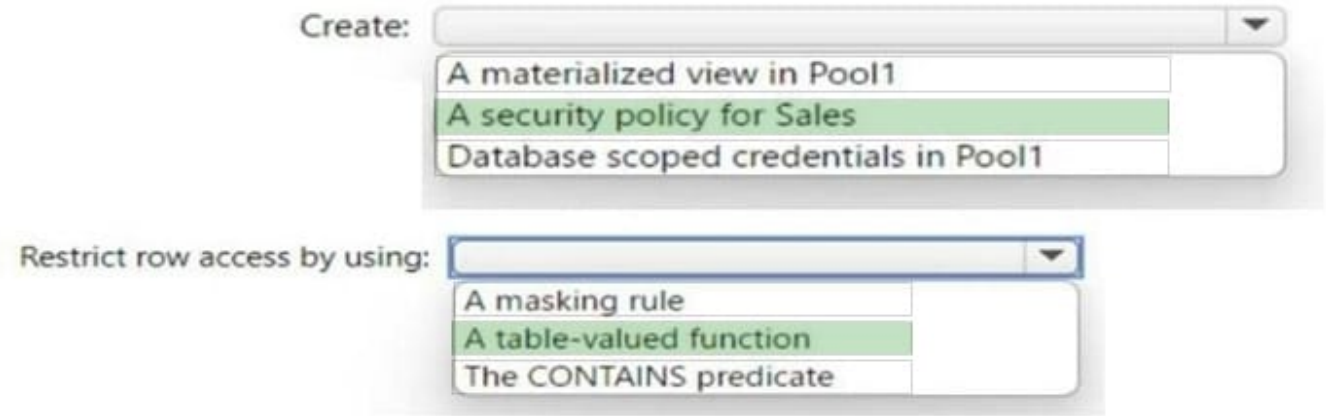
What should you do? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:



Correct Answer:



Box 1: A security policy for sale

Here are the steps to create a security policy for Sales:

Create a user-defined function that returns the name of the current user:

CREATE FUNCTION dbo.GetCurrentUser()

RETURNS NVARCHAR(128)

AS

BEGIN

RETURN SUSER_SNAME();

END;

Create a security predicate function that filters the Sales table based on the

current user:

CREATE FUNCTION dbo.SalesPredicate(@salesperson NVARCHAR(128)) RETURNS TABLE

WITH SCHEMABINDING

AS

RETURN SELECT 1 AS access_result

WHERE @salesperson = SalespersonName;

Create a security policy on the Sales table that uses the SalesPredicate function to filter the data:

CREATE SECURITY POLICY SalesFilter

ADD FILTER PREDICATE dbo.SalesPredicate(dbo.GetCurrentUser()) ON dbo.Sales

WITH (STATE = ON);

By creating a security policy for the Sales table, you ensure that each salesperson can only access their own sales data. The security policy uses a user-defined function to get the name of the current user and a security predicate function to

filter the Sales table based on the current user.

Box 2: table-value function

to restrict row access by using row-level security, you need to create a table-valued function that returns a table of values that represent the rows that a user can access. You then use this function in a security policy that applies a predicate on

the table.

QUESTION 2

You have an Azure Data Lake Storage account that has a virtual network service endpoint configured.

You plan to use Azure Data Factory to extract data from the Data Lake Storage account.

The data will then be loaded to a data warehouse in Azure Synapse Analytics by using PolyBase.

Which authentication method should you use to access Data Lake Storage?

A. shared access key authentication

B. managed identity authentication

C. account key authentication

D. service principal authentication

Correct Answer: B

Reference: https://docs.microsoft.com/en-us/azure/data-factory/connector-azure-sql-data-warehouse#use-polybase-to-load-data-into-azure-sql-data-warehouse

QUESTION 3

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

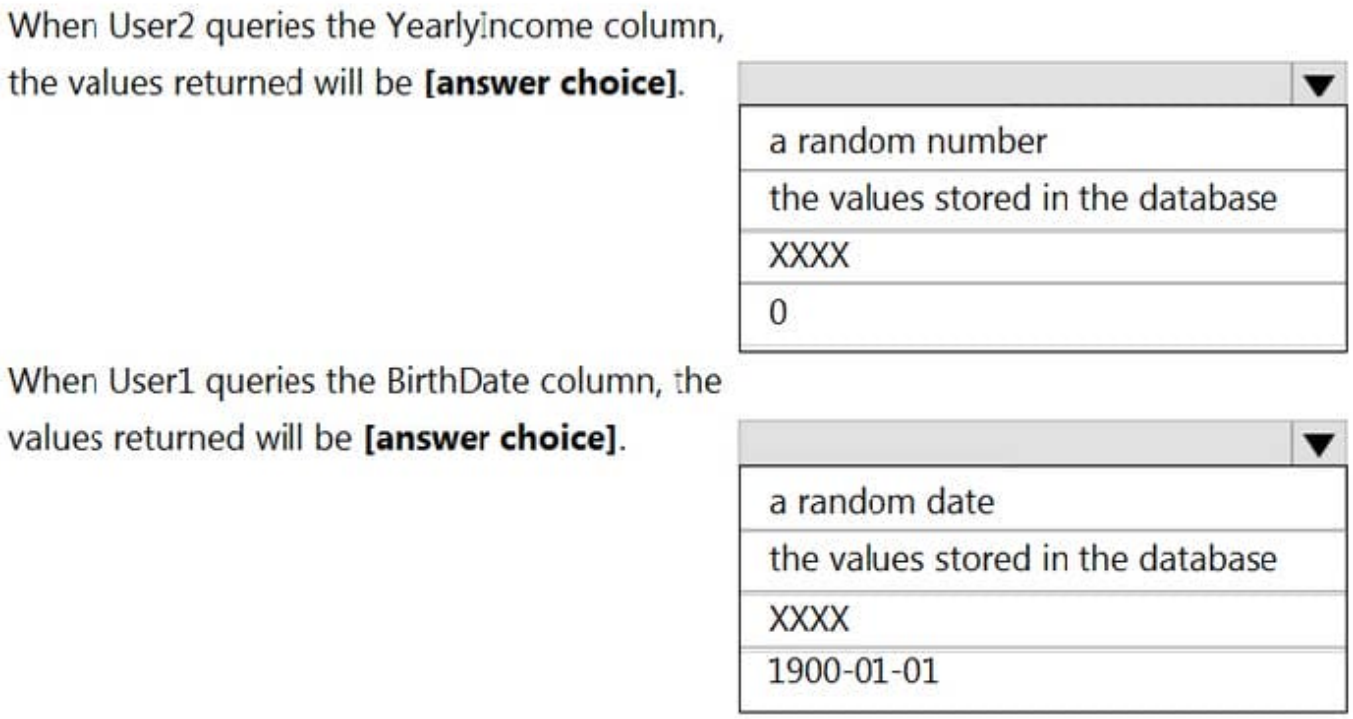
	name	table_name	datatype	is_masked	masking_function
1	BirthDate	DimCustomer	date	1	default()
2	Gender	DimCustomer	nvarchar	1	default()
3	EmailAddress	DimCustomer	nvarchar	1	email()
4	YearlyIncome	DimCustomer	money	1	default()

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

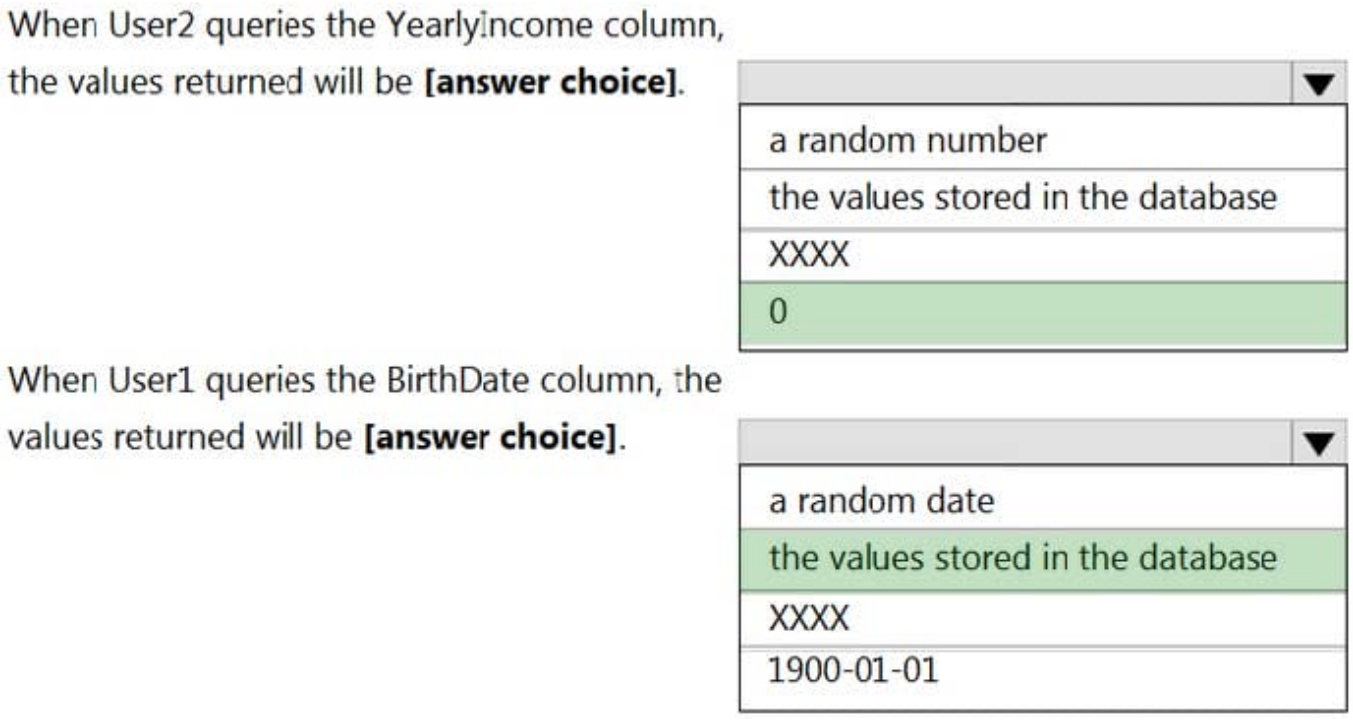
Use the drop-down menus to select the answer choice that completes each statement based on the information presented in the graphic.

NOTE: Each correct selection is worth one point.

Hot Area:



Correct Answer:



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 4

You are designing an inventory updates table in an Azure Synapse Analytics dedicated SQL pool. The table will have a clustered columnstore index and will include the following columns:

Table	Comment
EventDate	One million records are added to the table each day
EventTypeID	The table contains 10 million records for each event type.
WarehouseID	The table contains 100 million records for each warehouse.
ProductCategoryTypeID	The table contains 25 million records for each product category type.

You identify the following usage patterns:

1.

Analysts will most commonly analyze transactions for a warehouse.

2.

Queries will summarize by product category type, date, and/or inventory event type.

You need to recommend a partition strategy for the table to minimize query times.

On which column should you partition the table?

A. EventTypeID

B. ProductCategoryTypeID

C. EventDate

D. WarehouseID

Correct Answer: D

The number of records for each warehouse is big enough for a good partitioning.

Note: Table partitions enable you to divide your data into smaller groups of data. In most cases, table partitions are created on a date column.

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. Before partitions are created, dedicated SQL pool already divides each table into 60 distributed databases.

QUESTION 5

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:

1.

EmployeeID

2.

FirstName

3.

LastName

4.

Recipient

5.

GrossAmount

6.

TransactionID

7.

GovernmentID

8.

NetAmountPaid

9.

TransactionDate

You need to design a star schema data model in an Azure Synapse Analytics dedicated SQL pool for the data mart.

Which two tables should you create? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

A. a dimension table for Transaction

B. a dimension table for EmployeeTransaction

C. a dimension table for Employee

D. a fact table for Employee

E. a fact table for Transaction

Correct Answer: 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 6

You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Table1. Table1 contains the following:

1.

One billion rows

2.

A clustered columnstore index

3.

A hash-distributed column named Product Key

4.

A column named Sales Date that is of the date data type and cannot be null

Thirty million rows will be added to Table1 each month.

You need to partition Table1 based on the Sales Date column. The solution must optimize query performance and data loading. How often should you create a partition?

A. once per month

B. once per year

C. once per day

D. once per week

Correct Answer: B

Need a minimum 1 million rows per distribution. Each table is 60 distributions. 30 millions rows is added each month. Need 2 months to get a minimum of 1 million rows per distribution in a new partition.

Note: 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. Before partitions are created, dedicated SQL pool already divides each table into 60 distributions.

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.

Reference:

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

QUESTION 7

You are deploying a lake database by using an Azure Synapse database template.

You need to add additional tables to the database. The solution must use the same grouping method as the template tables.

Which grouping method should you use?

A. business area

B. size

C. facts and dimensions

D. partition style

Correct Answer: A

Business area: This is how the Azure Synapse database templates group tables by default. Each template consists of one or more enterprise templates that contain tables grouped by business areas. For example, the Retail template has business areas such as Customer, Product, Sales, and Store123. Using the same grouping method as the template tables can help you maintain consistency and compatibility with the industry-specific data model.

https://techcommunity.microsoft.com/t5/azure-synapse-analytics-blog/database-templates- in-azure- synapse-analytics/ba-p/2929112

QUESTION 8

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

You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Table1.

You have files that are ingested and loaded into an Azure Data Lake Storage Gen2 container named container1.

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

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

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

Does this meet the goal?

A. Yes

B. No

Correct Answer: B

Instead use a serverless SQL pool to create an external table with the extra column. Reference: https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/create-use-external-tables

QUESTION 9

You have an Azure Synapse Analytics dedicated SQL pool named Pool1. Pool1 contains a fact table named Table1.

You need to identify the extent of the data skew in Table1.

What should you do in Synapse Studio?

A. Connect to Pool1 and run DBCC PDW_SHOWSPACEUSED.

B. Connect to the built-in pool and run DBCC PDW_SHOWSPACEUSED.

C. Connect to the built-in pool and run DBCC CHECKALLOC.

D. Connect to the built-in pool and query sys.dm_pdw_sys_info.

Correct Answer: A

sys.dm_pdw_sys_info actually provides a set of appliance-level countersthat reflect overall activity on the appliance. DBCC PDW_SHOWSPACEUSEDshould be use instead since it displays the number of rows, disk spacereserved, and disk space used for a specific table, or for all tables ina Azure Synapse Analytics or Analytics Platform System (PDW)

database.

QUESTION 10

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

A. Pin the cluster.

B. Create an Azure runbook that starts the cluster every 90 days.

C. Terminate the cluster manually when processing completes.

D. Clone the cluster after it is terminated.

Correct Answer: A

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

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

QUESTION 11

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

You need to configure workspace1 to support autoscaling all-purpose clusters. The solution must meet the following requirements:

Automatically scale down workers when the cluster is underutilized for three minutes.

Minimize the time it takes to scale to the maximum number of workers.

Minimize costs.

What should you do first?

A. Enable container services for workspace1.

B. Upgrade workspace1 to the Premium pricing tier.

C. Set Cluster Mode to High Concurrency.

D. Create a cluster policy in workspace1.

Correct Answer: B

For clusters running Databricks Runtime 6.4 and above, optimized autoscaling is used by all-purpose clusters in the Premium plan Optimized autoscaling:

Scales up from min to max in 2 steps.

Can scale down even if the cluster is not idle by looking at shuffle file state.

Scales down based on a percentage of current nodes.

On job clusters, scales down if the cluster is underutilized over the last 40 seconds.

On all-purpose clusters, scales down if the cluster is underutilized over the last 150 seconds.

The spark.databricks.aggressiveWindowDownS Spark configuration property specifies in seconds how often a cluster makes down-scaling decisions. Increasing the value causes a cluster to scale down more slowly. The maximum value is

600.

Note: Standard autoscaling

Starts with adding 8 nodes. Thereafter, scales up exponentially, but can take many steps to reach the max. You can customize the first step by setting the spark.databricks.autoscaling.standardFirstStepUp Spark configuration property.

Scales down only when the cluster is completely idle and it has been underutilized for the last 10 minutes.

Scales down exponentially, starting with 1 node.

Reference:

https://docs.databricks.com/clusters/configure.html

QUESTION 12

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

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

Which type of window should you use?

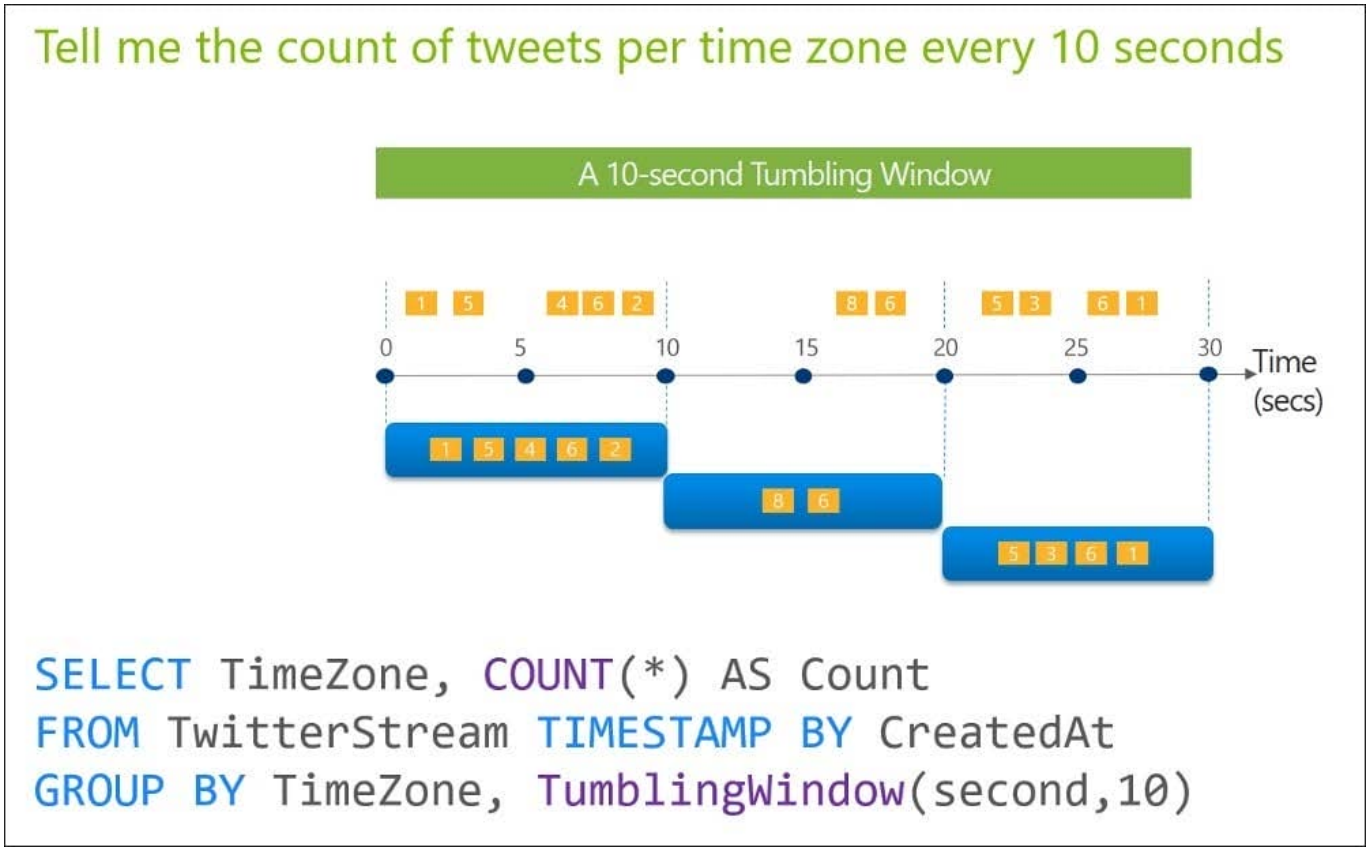
A. snapshot

B. tumbling

C. hopping

D. sliding

Correct Answer: C



Unlike tumbling windows, hopping windows model scheduled overlapping windows. A hopping window specification consist of three parameters: the timeunit, the windowsize (how long each window lasts) and the hopsize (by how much each window moves forward relative to the previous one).

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

QUESTION 13

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 14

HOTSPOT

You develop a dataset named DBTBL1 by using Azure Databricks.

DBTBL1 contains the following columns:

1.

SensorTypeID

2.

GeographyRegionID

3.

Year

4.

Month

5.

Day

6.

Hour

7.

Minute

8.

Temperature

9.

WindSpeed 10.Other

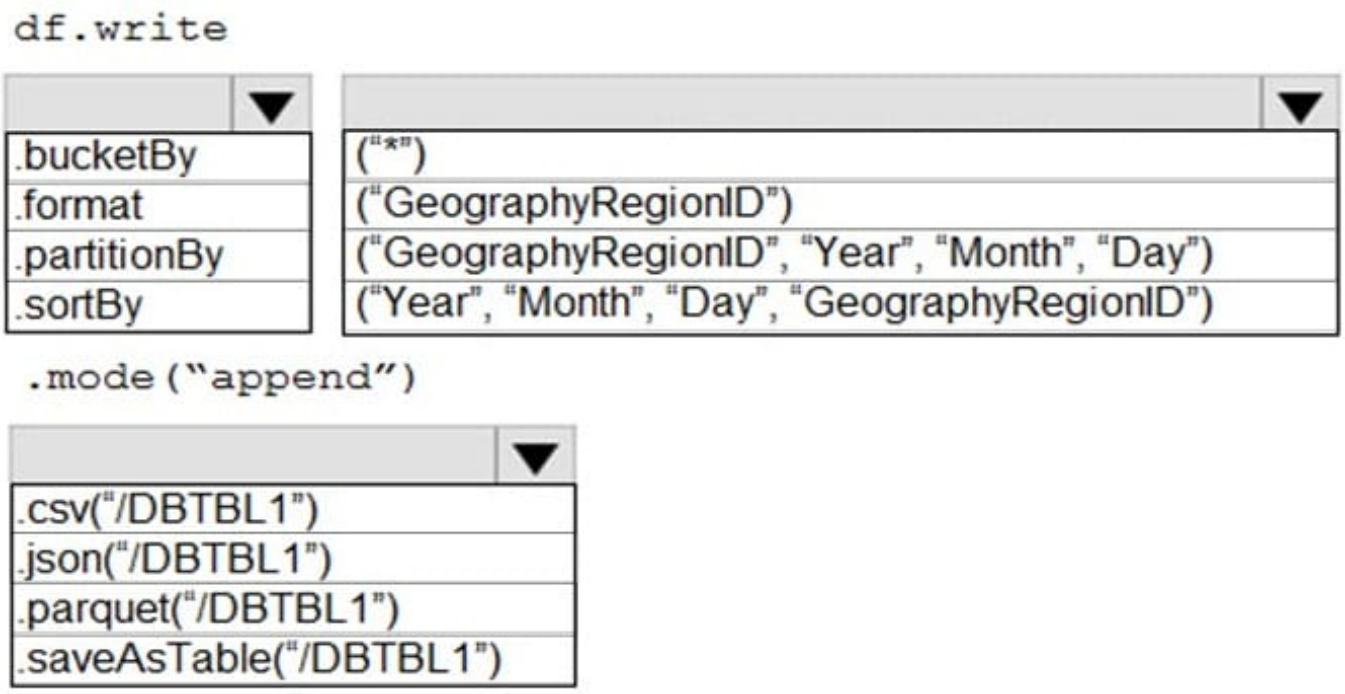
You need to store the data to support daily incremental load pipelines that vary for each GeographyRegionID. The solution must minimize storage costs.

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

NOTE: Each correct selection is worth one point.

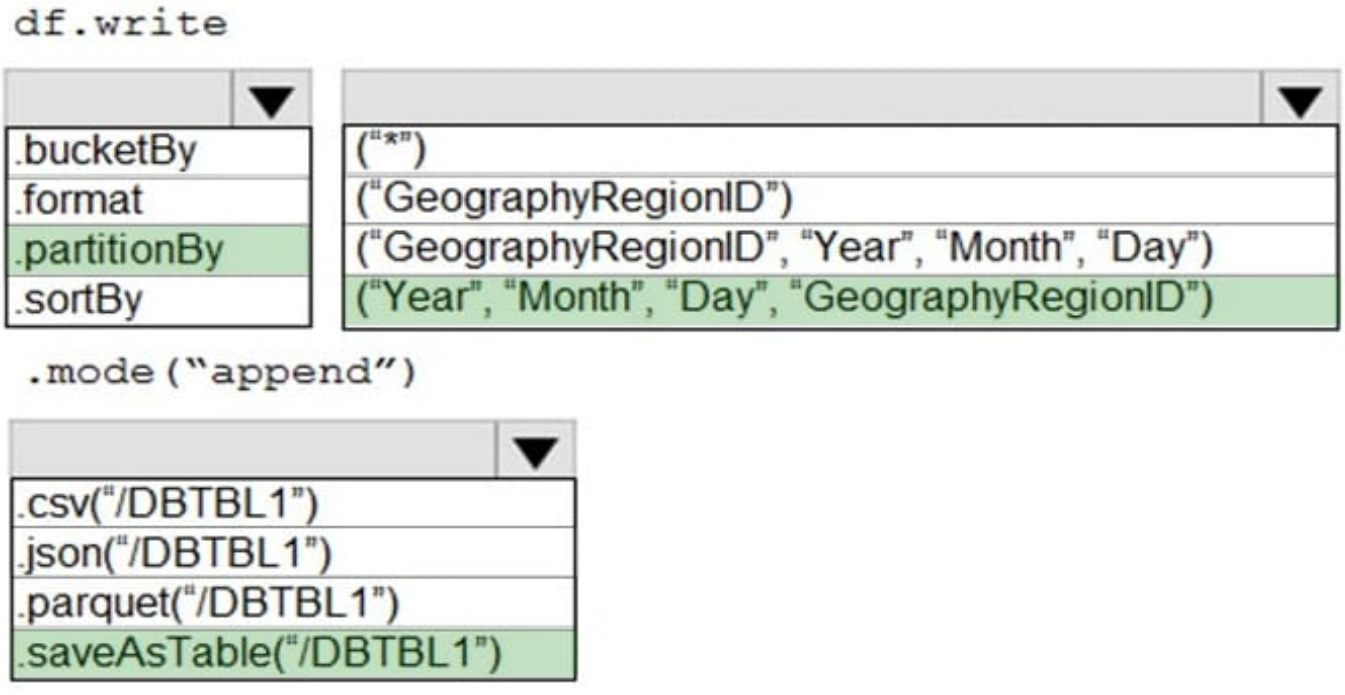
Hot Area:

Answer Area



Correct Answer:

Answer Area



Box 1: .partitionBy

Incorrect Answers:

.format:

Method: format():

Arguments: "parquet", "csv", "txt", "json", "jdbc", "orc", "avro", etc.

.bucketBy:

Method: bucketBy()

Arguments: (numBuckets, col, col..., coln)

The number of buckets and names of columns to bucket by. Uses Hive\\'s bucketing scheme on a filesystem.

Box 2: ("Year", "Month", "Day","GeographyRegionID")

Specify the columns on which to do the partition. Use the date columns followed by the GeographyRegionID column.

Box 3: .saveAsTable("/DBTBL1")

Method: saveAsTable()

Argument: "table_name"

The table to save to.

Reference:

https://www.oreilly.com/library/view/learning-spark-2nd/9781492050032/ch04.html

https://docs.microsoft.com/en-us/azure/databricks/delta/delta-batch

QUESTION 15

You have two fact tables named Flight and Weather. Queries targeting the tables will be based on the join between the following columns.

Table	Column
Flight	ArrivalAirportID ArrivalDateTime
Weather	AirportID ReportDateTime

You need to recommend a solution that maximum query performance. What should you include in the recommendation?

A. In each table, create a column as a composite of the other two columns in the table.

B. In each table, create an IDENTITY column.

C. In the tables, use a hash distribution of ArriveDateTime and ReportDateTime.

D. In the tables, use a hash distribution of ArriveAirPortID and AirportID.

Correct Answer: D