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.pass2lead.com/dp-203.html

100% Passing Guarantee
100% Money Back Assurance

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

- **Instant Download** After Purchase
- **100% Money Back** Guarantee
- **365 Days** Free Update
- **800,000+** Satisfied Customers

QUESTION 1

HOTSPOT

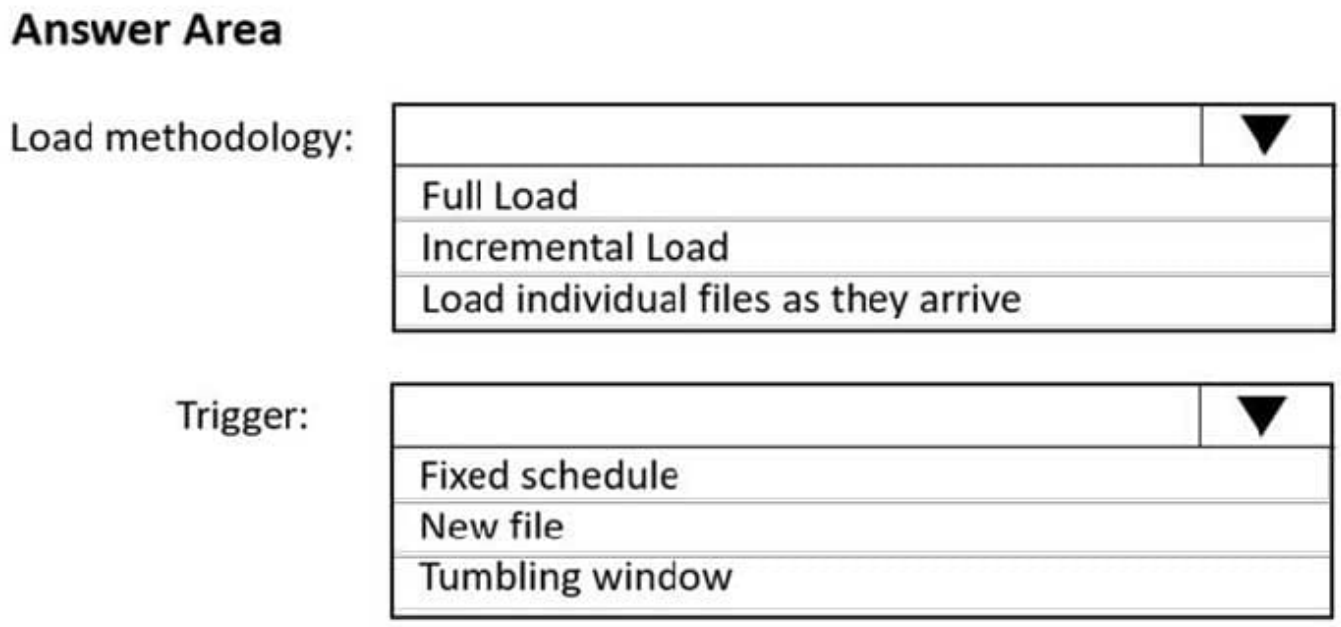
You have an Azure Storage account that generates 200,000 new files daily. The file names have a format of {YYYY}/{MM}/{DD}/{HH}/{CustomerID}.csv.

You need to design an Azure Data Factory solution that will load new data from the storage account to an Azure Data Lake once hourly. The solution must minimize load times and costs.

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

NOTE: Each correct selection is worth one point.

Hot Area:



Correct Answer:

Answer Area

Load methodology:

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

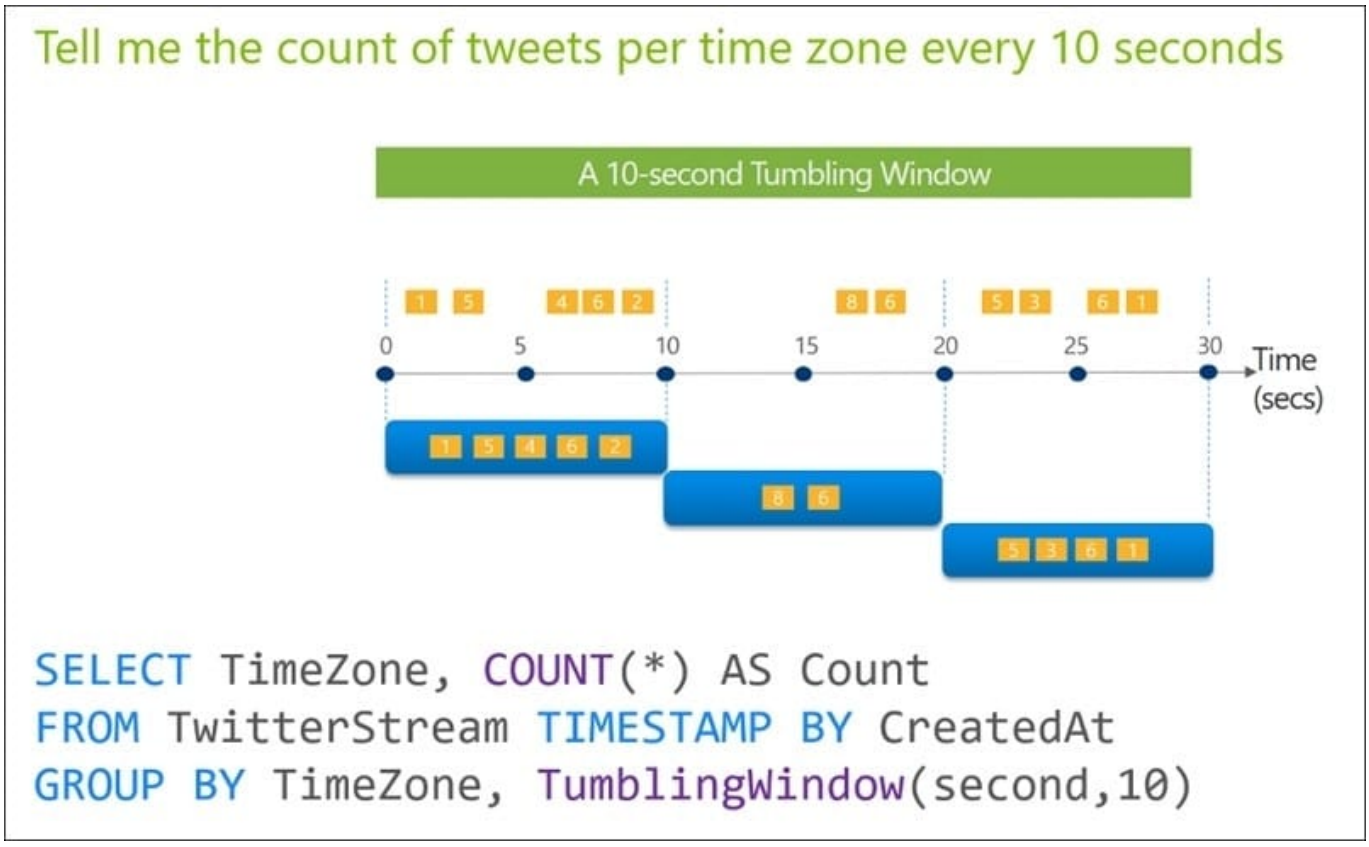
Trigger:

- Fixed schedule
- New file
- **Tumbling window**

Box 1: Incremental load

Box 2: Tumbling window

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



QUESTION 2

You are designing a star schema for a dataset that contains records of online orders. Each record includes an order date, an order due date, and an order ship date. You need to ensure that the design provides the fastest query times of the records when querying for arbitrary date ranges and aggregating by fiscal calendar attributes. Which two actions should you perform? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

A. Create a date dimension table that has a DateTime key.

B. Use built-in SQL functions to extract date attributes.

C. Create a date dimension table that has an integer key in the format of YYYYMMDD.

D. In the fact table, use integer columns for the date fields.

E. Use DateTime columns for the date fields.

Correct Answer: BD

QUESTION 3

You are designing a solution that will use tables in Delta Lake on Azure Databricks. You need to minimize how long it takes to perform the following:

1.

Queries against non-partitioned tables

2.

Joins on non-partitioned columns

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

NOTE: Each correct selection is worth one point.

A. the clone command

B. Z-Ordering

C. Apache Spark caching

D. dynamic file pruning (DFP)

Correct Answer: BD

Best practices: Delta Lake

B: Provide data location hints If you expect a column to be commonly used in query predicates and if that column has high cardinality (that is, a large number of distinct values), then use Z-ORDER BY. Delta Lake automatically lays out the data in the files based on the column values and uses the layout information to skip irrelevant data while querying.

BD: Dynamic file pruning, can significantly improve the performance of many queries on Delta Lake tables. Dynamic file pruning is especially efficient for non-partitioned tables, or for joins on non-partitioned columns. The performance impact

of dynamic file pruning is often correlated to the clustering of data so consider using Z-Ordering to maximize the benefit.

Incorrect:

Not C: Spark caching

Databricks does not recommend that you use Spark caching for the following reasons:

You lose any data skipping that can come from additional filters added on top of the cached DataFrame.

The data that gets cached might not be updated if the table is accessed using a different identifier (for example, you do spark.table(x).cache() but then write to the table using spark.write.save(/some/path).

Reference: https://learn.microsoft.com/en-us/azure/databricks/delta/best-practices#spark-caching
https://learn.microsoft.com/en-us/azure/databricks/optimizations/dynamic-file-pruning

QUESTION 4

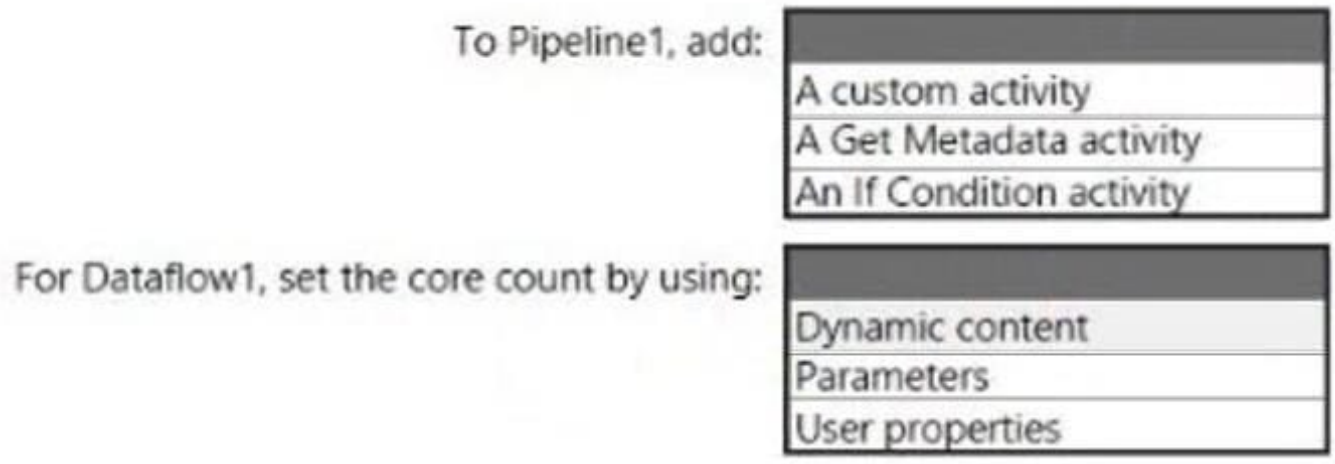
HOTSPOT

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

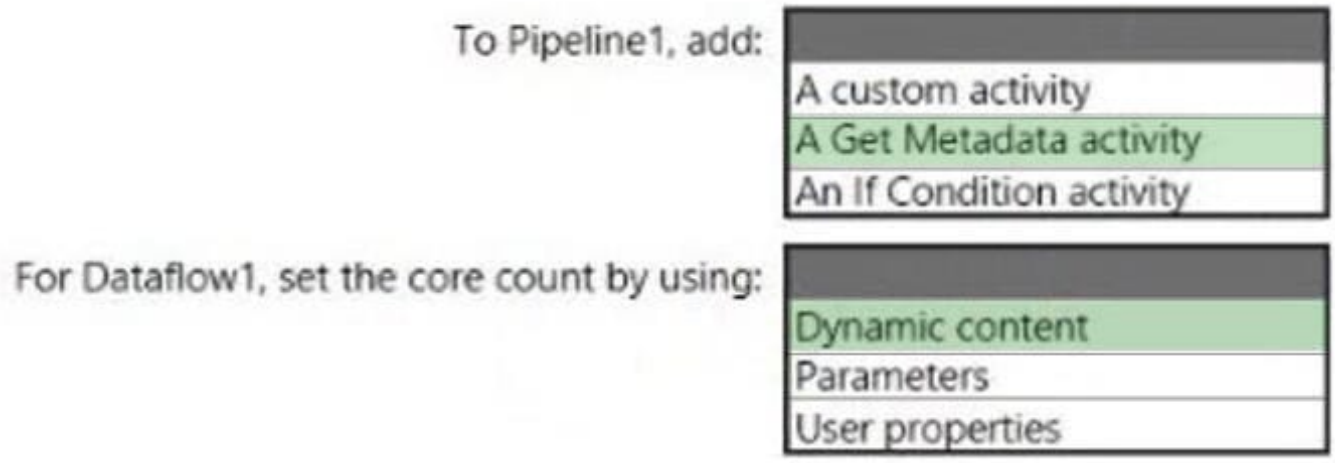
Pipeline1 retrieves files from an Azure Data Lake Storage Gen 2 account named storage1.

Dataflow1 uses the AutoResolveIntegrationRuntime integration runtime configured with a core count of 128. You need to optimize the number of cores used by Dataflow1 to accommodate the size of the files in storage1. What should you configure? To answer, select the appropriate options in the answer area.

Hot Area:



Correct Answer:



QUESTION 5

DRAG DROP

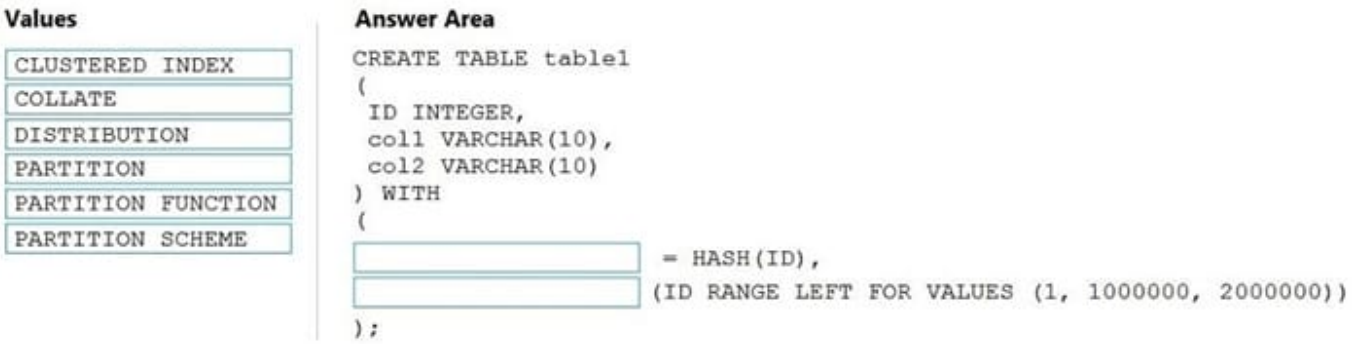
You need to create a partitioned table in an Azure Synapse Analytics dedicated SQL pool.

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

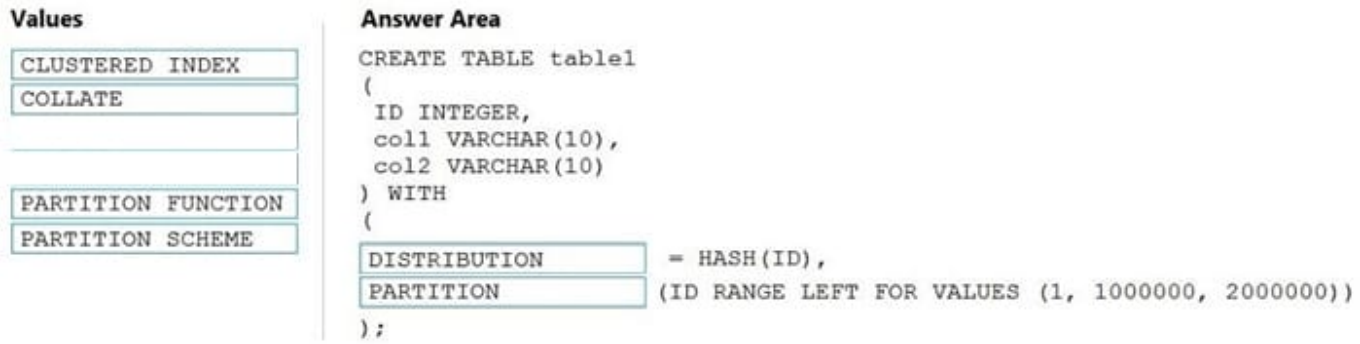
view content.

NOTE: Each correct selection is worth one point.

Select and Place:



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?