Vendor: Microsoft

Exam Code:DP-203

Exam Name: Data Engineering on Microsoft Azure

Version: Demo

QUESTION 1

You are monitoring an Azure Stream Analytics job.

The Backlogged Input Events count has been 20 for the last hour.

You need to reduce the Backlogged Input Events count.

What should you do?

A. Drop late arriving events from the job.

B. Add an Azure Storage account to the job.

C. Increase the streaming units for the job.

D. Stop the job.

Correct Answer: C

General symptoms of the job hitting system resource limits include:

If the backlog event metric keeps increasing, it\\'s an indicator that the system resource is constrained (either because of output sink throttling, or high CPU).

Note: Backlogged Input Events: Number of input events that are backlogged. A non-zero value for this metric implies that your job isn\\'t able to keep up with the number of incoming events. If this value is slowly increasing or consistently nonzero, you should scale out your job: adjust Streaming Units.

Reference:

https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-scale-jobs

https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-monitoring

---

QUESTION 2

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).

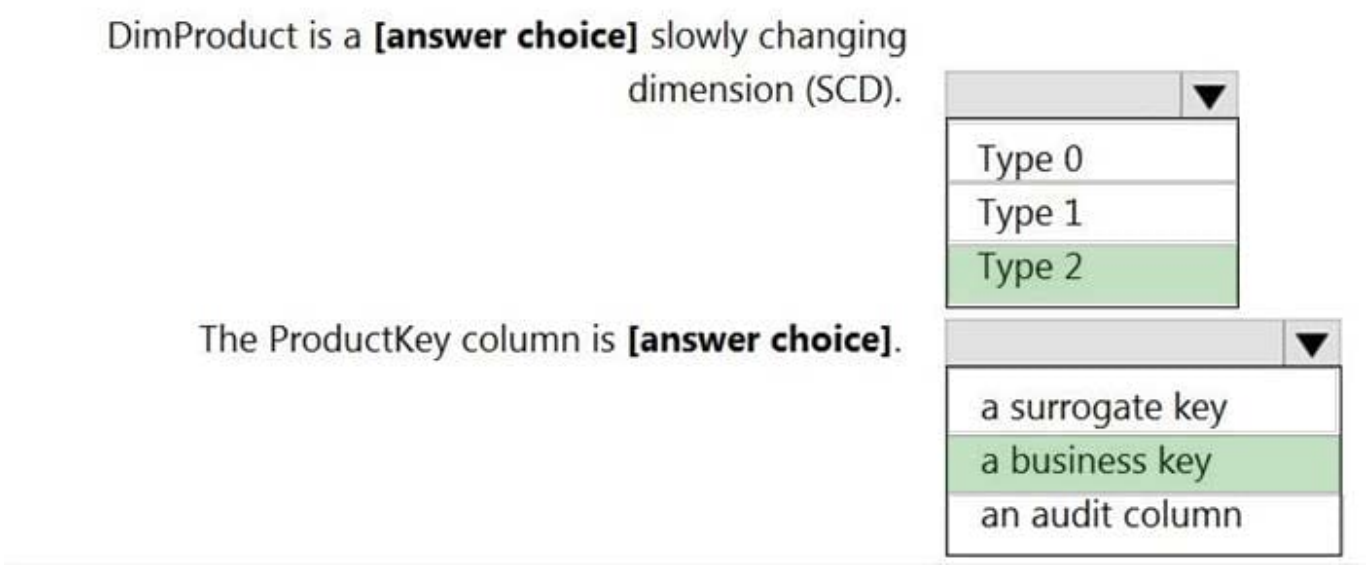
- Type 0
- Type 1
- Type 2

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

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

Correct Answer:

Answer Area



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 3

You have an Azure Data Factory pipeline named pipeline1 that is invoked by a tumbling window trigger named Trigger1. Trigger1 has a recurrence of 60 minutes.

You need to ensure that pipeline1 will execute only if the previous execution completes successfully.

How should you configure the self-dependency for Trigger1?

A. offset: "-00:01:00" size: "00:01:00"

B. offset: "01:00:00" size: "-01:00:00"

C. offset: "01:00:00" size: "01:00:00"

D. offset: "-01:00:00" size: "01:00:00"

Correct Answer: D

Tumbling window self-dependency properties

In scenarios where the trigger shouldn\\'t proceed to the next window until the preceding window is successfully completed, build a self-dependency. A self-dependency trigger that\\'s dependent on the success of earlier runs of itself within the

preceding hour will have the properties indicated in the following code.

Example code:

"name": "DemoSelfDependency",

"properties": {

"runtimeState": "Started",

"pipeline": {

"pipelineReference": {

"referenceName": "Demo",

"type": "PipelineReference"

}

},

"type": "TumblingWindowTrigger",

"typeProperties": {

"frequency": "Hour",

"interval": 1,

"startTime": "2018-10-04T00:00:00Z",

"delay": "00:01:00",

"maxConcurrency": 50,

"retryPolicy": {

"intervalInSeconds": 30

},

"dependsOn": [

{

"type": "SelfDependencyTumblingWindowTriggerReference", "size": "01:00:00", "offset": "-01:00:00"

}

]

}

}

}

Reference: https://docs.microsoft.com/en-us/azure/data-factory/tumbling-window-trigger-dependency

---

QUESTION 4

HOTSPOT

You have an Azure subscription that contains an Azure Synapse Analytics dedicated SQL pool named Pool1 and an Azure Data Lake Storage account named storage1. Storage1 requires secure transfers.

You need to create an external data source in Pool1 that will be used to read .orc files in storage1.

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

```
CREATE EXTERNAL DATA SOURCE AzureDataLakeStore
WITH
(  Location1 ' [          ] ://data@newyorktaxidataset.dfs.core.windows.net' ,
                  abfs
                  abfss
                  wasb
                  wasbs
credential = ADLS_credential ,
TYPE - [                  ]
                  BLOB_STORAGE
);                HADOOP
                  RDBMS
                  SHARP MAP MANAGER
```

Correct Answer:

Answer Area

```
CREATE EXTERNAL DATA SOURCE AzureDataLakeStore
WITH
(  Location1 ' [          ] ://data@newyorktaxidataset.dfs.core.windows.net' ,
                  abfs
                  abfss
                  wasb
                  wasbs
credential = ADLS_credential ,
TYPE - [                  ]
                  BLOB_STORAGE
);                HADOOP
                  RDBMS
                  SHARP MAP MANAGER
```

Reference: https://docs.microsoft.com/en-us/sql/t-sql/statements/create-external-data-source-transact-sql?view=azure-sqldw-latestandpreserve-view=trueandtabs=dedicated

---

QUESTION 5

HOTSPOT

You have a Microsoft SQL Server database that uses a third normal form schema.

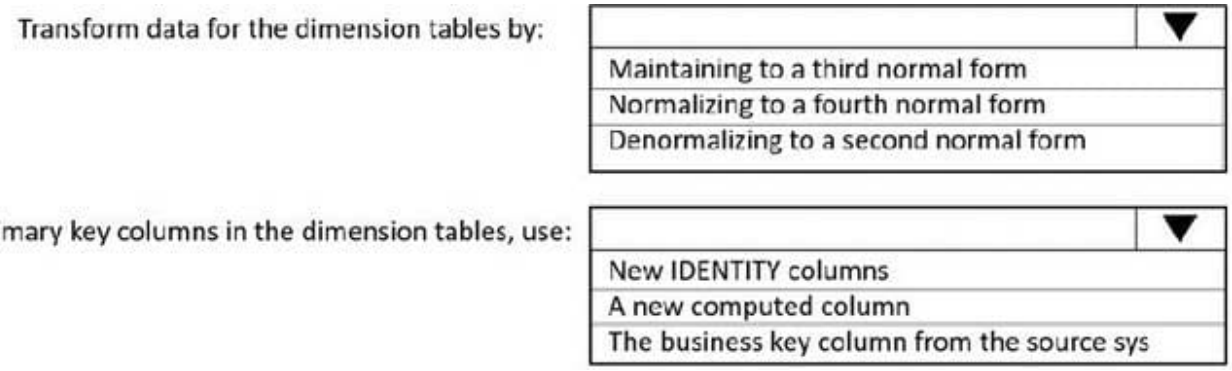
You plan to migrate the data in the database to a star schema in an A?\ire Synapse Analytics dedicated SQI pool.

You need to design the dimension tables. The solution must optimize read operations.

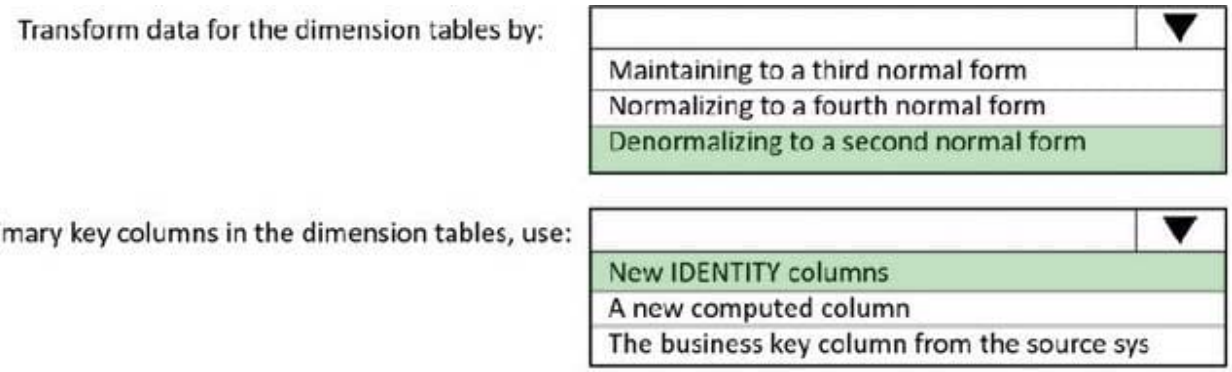
What should you include in the solution? to answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:



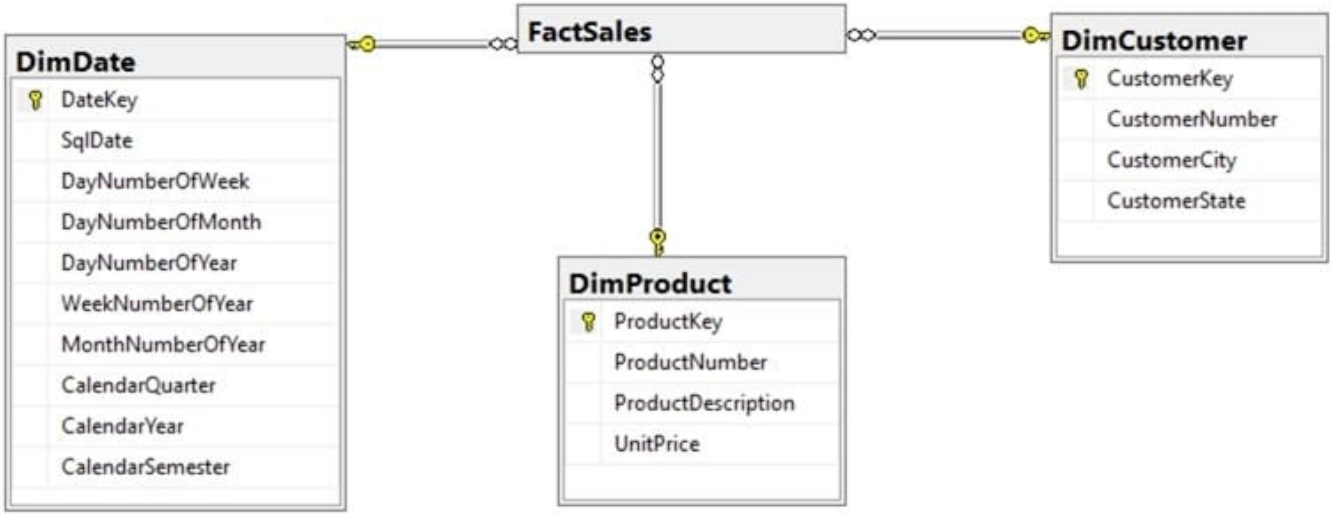
Correct Answer:



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 at 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.

QUESTION 6

You have an Azure Blob Storage account named blob1 and an Azure Data Factory pipeline named pipeline1.

You need to ensure that pipeline1 runs when a file is deleted from a container in blob1. The solution must minimize development effort.

Which type of trigger should you use?

A. schedule

B. storage event

C. tumbling window

D. custom event

Correct Answer: B

Explanation:

You can create a trigger that runs a pipeline in response to a storage event.

Event-driven architecture (EDA) is a common data integration pattern that involves production, detection, consumption, and reaction to events. Data integration scenarios often require 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 and Synapse pipelines natively integrate with Azure Event Grid, which lets you trigger pipelines on such events.

Note

The Storage Event Trigger currently supports only Azure Data Lake Storage Gen2 and General-purpose version 2

storage accounts.

Reference:

https://learn.microsoft.com/en-us/azure/data-factory/how-to-create-event-trigger

---

QUESTION 7

HOTSPOT

The following code segment is used to create an Azure Databricks cluster.

```
{
     "num_workers": null,
     "autoscale": {
          "min_workers": 2,
          "max_workers": 8
     },
     "cluster_name": "MyCluster",
     "spark_version": "latest-stable-scala2.11",
     "spark_conf": {
          "spark.databricks.cluster.profile": "serverless",
          "spark.databricks.repl.allowedLanguages": "sql,python,r"
     },
     "node_type_id": "Standard_DS13_v2",
     "ssh_public_keys": [],
     "custom_tags": {
          "ResourceClass": "Serverless"
     },
     "spark_env_vars": {
          "PYSPARK_PYTHON": "/databricks/python3/bin/python3"
     },
     "autotermination_minutes": 90,
     "enable_elastic_disk": true,
     "init_scripts": []
}
```

For each of the following statements, select Yes if the statement is true. Otherwise, select No.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

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

Correct Answer:

Answer Area

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

Box 1: Yes

A cluster mode of 'High Concurrency' is selected, unlike all the others which are 'Standard'. This results in a worker type of Standard_DS13_v2.

Box 2: No

When you run a job on a new cluster, the job is treated as a data engineering (job) workload subject to the job workload pricing. When you run a job on an existing cluster, the job is treated as a data analytics (all-purpose) workload subject to

all-purpose workload pricing.

Box 3: Yes

Delta Lake on Databricks allows you to configure Delta Lake based on your workload patterns.

Reference:

https://adatis.co.uk/databricks-cluster-sizing/

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

https://docs.databricks.com/administration-guide/capacity-planning/cmbp.html

https://docs.databricks.com/delta/index.html

QUESTION 8

DRAG DROP

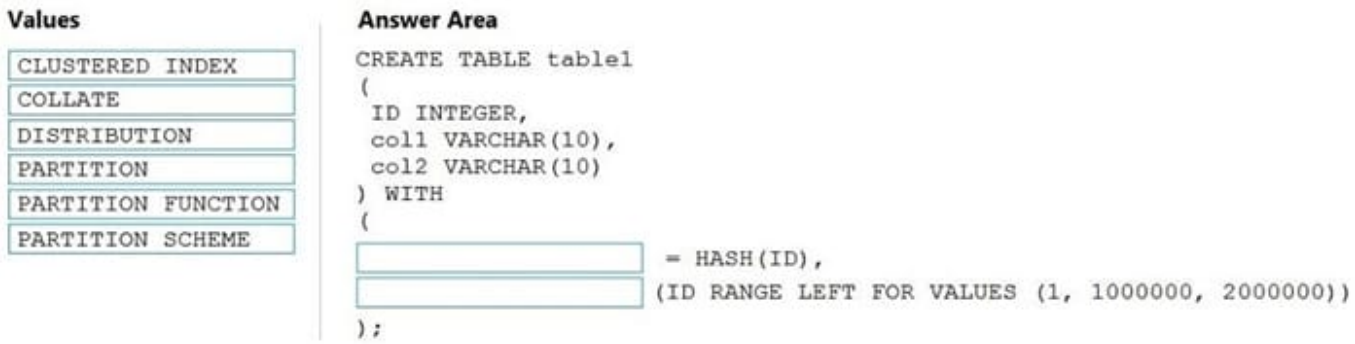
You need to create a partitioned table in an Azure Synapse Analytics dedicated SQL pool.

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

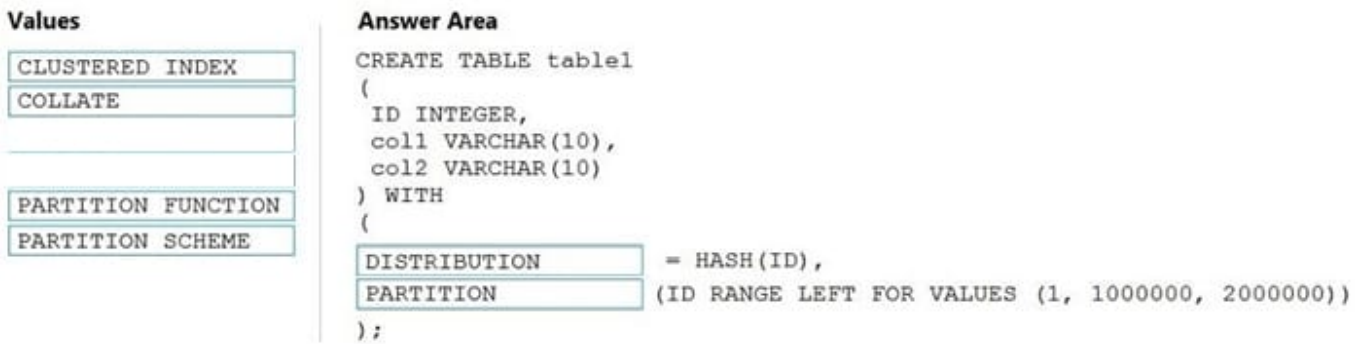
view content.

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?

QUESTION 9

You have an Azure Synapse Analytics dedicated SQL pool named Pool1 and a database named DB1. DB1 contains a fact table named Table1.

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

What should you do in Synapse Studio?

A. Connect to the built-in pool and query sysdm_pdw_sys_info.

B. Connect to Pool1 and run DBCC CHECKALLOC.

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

D. Connect to Pool! and query sys.dm_pdw_nodes_db_partition_stats.

Correct Answer: D

Microsoft recommends use of sys.dm_pdw_nodes_db_partition_stats to analyze any skewness in the data.

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

---

QUESTION 10

You have an Azure SQL database named DB1 and an Azure Data Factory data pipeline named pipeline.

From Data Factory, you configure a linked service to DB1.

In DB1, you create a stored procedure named SP1. SP1 returns a single row of data that has four columns.

You need to add an activity to pipeline to execute SP1. The solution must ensure that the values in the columns are stored as pipeline variables.

Which two types of activities can you use to execute SP1? (Refer to Data Engineering on Microsoft Azure documents or guide for Answers/available at Microsoft.com)

A. Script

B. Copy

C. Lookup

D. Stored Procedure

Correct Answer: AD

A: You use data transformation activities in a Data Factory or Synapse pipeline to transform and process raw data into predictions and insights. The Script activity is one of the transformation activities that pipelines support.

You can use the Script activity to invoke a SQL script in one of the following data stores in your enterprise or on an Azure virtual machine (VM):

Azure SQL Database Azure Synapse Analytics SQL Server Database. Oracle Snowflake

The script may contain either a single SQL statement or multiple SQL statements that run sequentially. You can use the Script task for the following purposes:

Truncate a table in preparation for inserting data.

Create, alter, and drop database objects such as tables and views.

Re-create fact and dimension tables before loading data into them.

*-> Run stored procedures. If the SQL statement invokes a stored procedure that returns results from a temporary table, use the WITH RESULT SETS option to define metadata for the result set.

Save the rowset returned from a query as activity output for downstream consumption.

D: You can transform data by using the SQL Server Stored Procedure activity in Azure Data Factory or Synapse Analytics.

You use data transformation activities in a Data Factory or Synapse pipeline to transform and process raw data into predictions and insights. The Stored Procedure Activity is one of the transformation activities that pipelines support.

You can use the Stored Procedure Activity to invoke a stored procedure in one of the following data stores in your enterprise or on an Azure virtual machine (VM):

Azure SQL Database Azure Synapse Analytics SQL Server Database.

Reference:

https://learn.microsoft.com/en-us/azure/data-factory/transform-data-using-script

https://learn.microsoft.com/en-us/azure/data-factory/transform-data-using-stored-procedure¨

QUESTION 11

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 12

You are designing a date dimension table in an Azure Synapse Analytics dedicated SQL pool. The date dimension table will be used by all the fact tables. Which distribution type should you recommend to minimize data movement?

A. HASH

B. REPLICATE

C. ROUND_ROBIN

Correct Answer: B

A replicated table has a full copy of the table available on every Compute node. Queries run fast on replicated tables since joins on replicated tables don\\'t require data movement. Replication requires extra storage, though, and isn\\'t practical for large tables.

Incorrect Answers:

A: A hash distributed table is designed to achieve high performance for queries on large tables.

C: A round-robin table distributes table rows evenly across all distributions. The rows are distributed randomly. Loading data into a round-robin table is fast. Keep in mind that queries can require more data movement than the other distribution methods.

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