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.leads4pass.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

You have an Azure data factory that connects to a Microsoft Purview account. The data factory is registered in Microsoft Purview.

You update a Data Factory pipeline.

You need to ensure that the updated lineage is available in Microsoft Purview.

What should you do first?

A. Locate the related asset in the Microsoft Purview portal.

B. Execute the pipeline.

C. Disconnect the Microsoft Purview account from the data factory.

D. Execute an Azure DevOps build pipeline.

Correct Answer: B

Run pipeline and push lineage data to Microsoft Purview

Step 1: Connect Data Factory to your Microsoft Purview account

Step 2: Run pipeline in Data Factory

You can create pipelines, Copy activities and Dataflow activities in Data Factory. You don\\'t need any additional configuration for lineage data capture. The lineage data will automatically be captured during the activities execution.

Step 3: Monitor lineage reporting status

After you run the pipeline, in the pipeline monitoring view, you can check the lineage reporting status by clicking the following Lineage status button.

Step 4: View lineage information in your Microsoft Purview account

On Microsoft Purview UI, you can browse assets and choose type "Azure Data Factory". You can also search the Data Catalog using keywords.

Reference:

https://learn.microsoft.com/en-us/azure/data-factory/tutorial-push-lineage-to-purview

QUESTION 2

DRAG DROP

You need to create an Azure Data Factory pipeline to process data for the following three departments at your company: Ecommerce, retail, and wholesale. The solution must ensure that data can also be processed for the entire company.

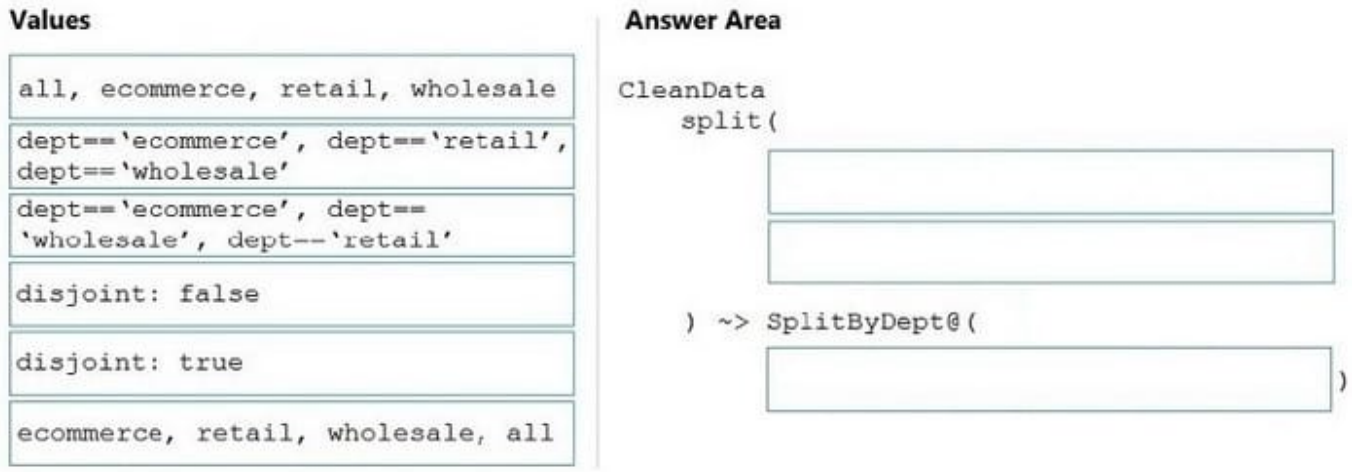
How should you complete the Data Factory data flow script? To answer, drag the appropriate values to the correct

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

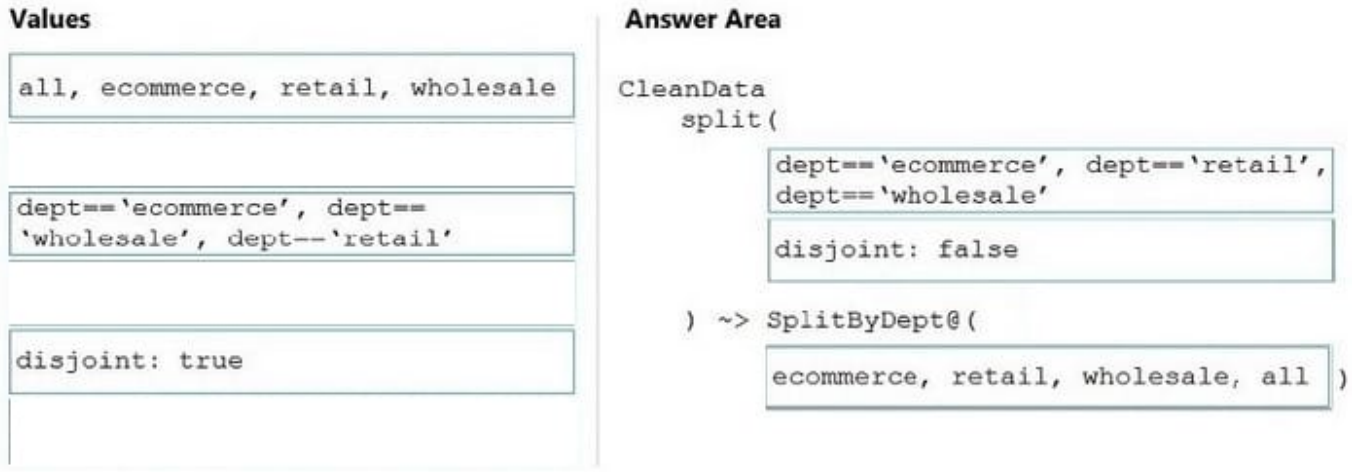
scroll to view content.

NOTE: Each correct selection is worth one point.

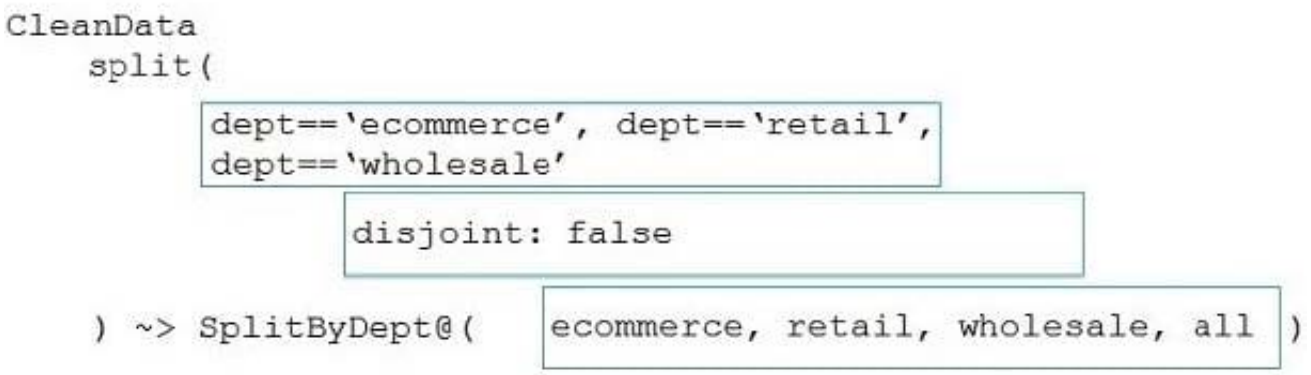
Select and Place:



Correct Answer:



The conditional split transformation routes data rows to different streams based on matching conditions. The conditional split transformation is similar to a CASE decision structure in a programming language. The transformation evaluates expressions, and based on the results, directs the data row to the specified stream.



Box 1: dept==\\'ecommerce\\', dept==\\'retail\\', dept==\\'wholesale\\' First we put the condition. The order must match the stream labeling we define in Box 3.

Syntax:

split(

disjoint: {true | false}

) ~> @(stream1, stream2, ...,)

Box 2: discount : false

disjoint is false because the data goes to the first matching condition. All remaining rows matching the third condition go to output stream all.

Box 3: ecommerce, retail, wholesale, all

Label the streams

QUESTION 3

You are designing a streaming data solution that will ingest variable volumes of data. You need to ensure that you can change the partition count after creation.

Which service should you use to ingest the data?

A. Azure Event Hubs Dedicated

B. Azure Stream Analytics

C. Azure Data Factory

D. Azure Synapse Analytics

Correct Answer: A

You can\\'t change the partition count for an event hub after its creation except for the event hub in a dedicated cluster.

Reference: https://docs.microsoft.com/en-us/azure/event-hubs/event-hubs-features

QUESTION 4

You have an Azure Data Factory that contains 10 pipelines.

You need to label each pipeline with its main purpose of either ingest, transform, or load. The labels must be available for grouping and filtering when using the monitoring experience in Data Factory.

What should you add to each pipeline?

A. a resource tag

B. a correlation ID

C. a run group ID

D. an annotation

Correct Answer: D

Annotations are additional, informative tags that you can add to specific factory resources: pipelines, datasets, linked services, and triggers. By adding annotations, you can easily filter and search for specific factory resources.

Reference: https://www.cathrinewilhelmsen.net/annotations-user-properties-azure-data-factory/

QUESTION 5

HOTSPOT

You build an Azure Data Factory pipeline to move data from an Azure Data Lake Storage Gen2 container to a database in an Azure Synapse Analytics dedicated SQL pool.

Data in the container is stored in the following folder structure.

/in/{YYYY}/{MM}/{DD}/{HH}/{mm}

The earliest folder is /in/2021/01/01/00/00. The latest folder is /in/2021/01/15/01/45.

You need to configure a pipeline trigger to meet the following requirements:

Existing data must be loaded.

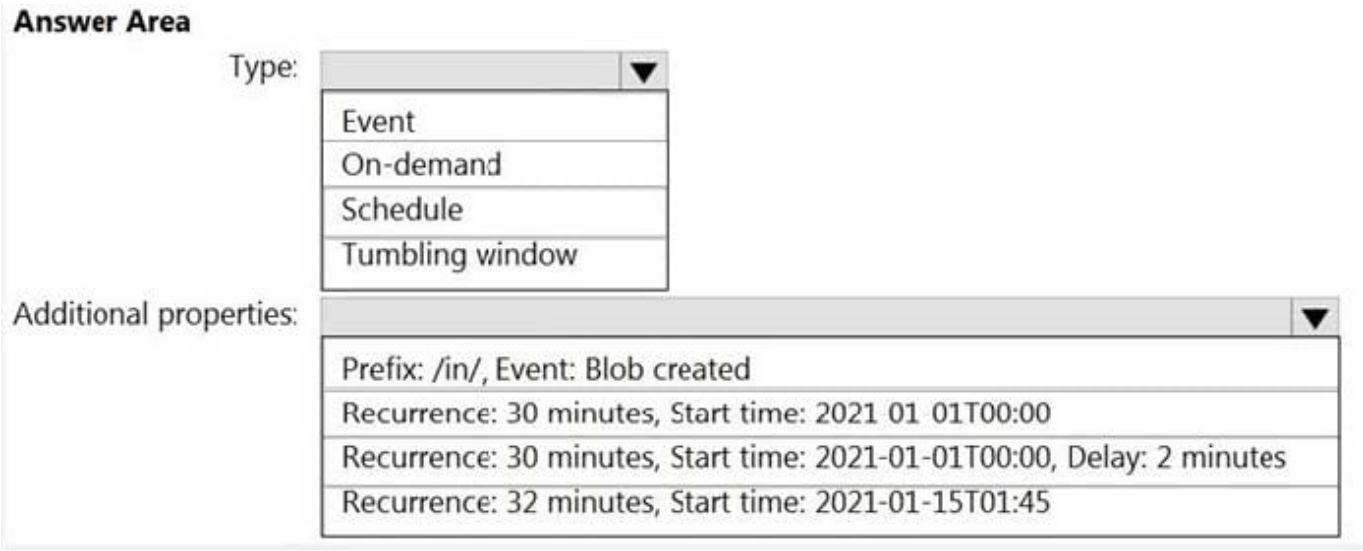
Data must be loaded every 30 minutes.

Late-arriving data of up to two minutes must he included in the load for the time at which the data should have arrived.

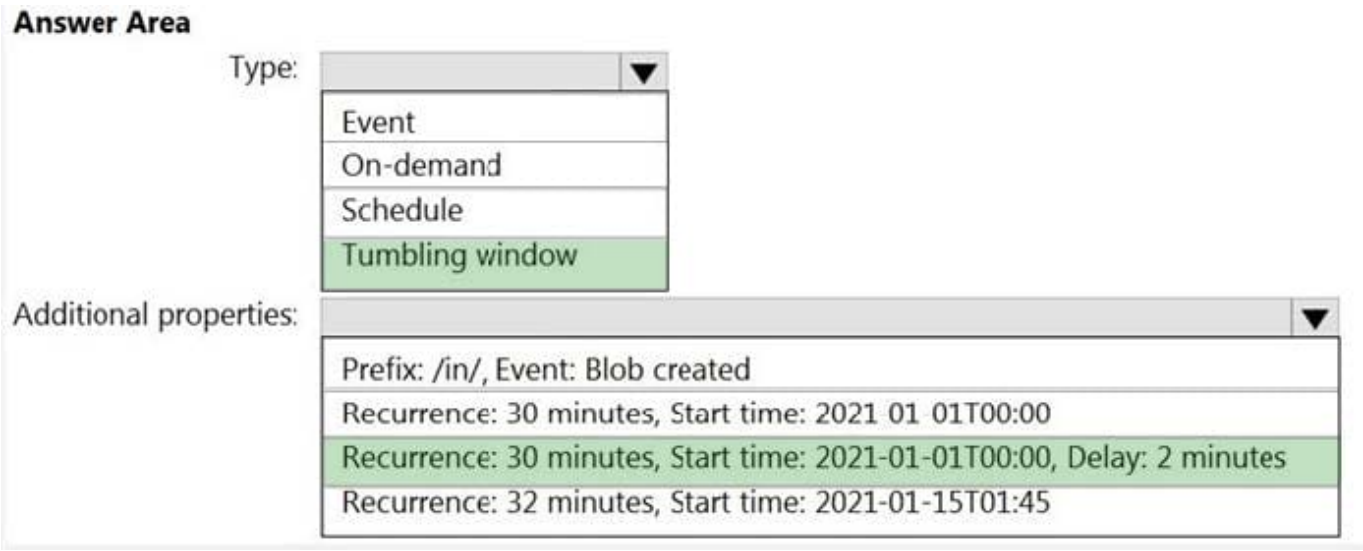
How should you configure the pipeline trigger? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:



Correct Answer:



Box 1: Tumbling window

To be able to use the Delay parameter we select Tumbling window.

Box 2:

Recurrence: 30 minutes, not 32 minutes

Delay: 2 minutes.

The amount of time to delay the start of data processing for the window. The pipeline run is started after the expected execution time plus the amount of delay. The delay defines how long the trigger waits past the due time before triggering a

new run. The delay doesn\\'t alter the window startTime.

Reference:

https://docs.microsoft.com/en-us/azure/data-factory/how-to-create-tumbling-window-trigger

QUESTION 6

HOTSPOT

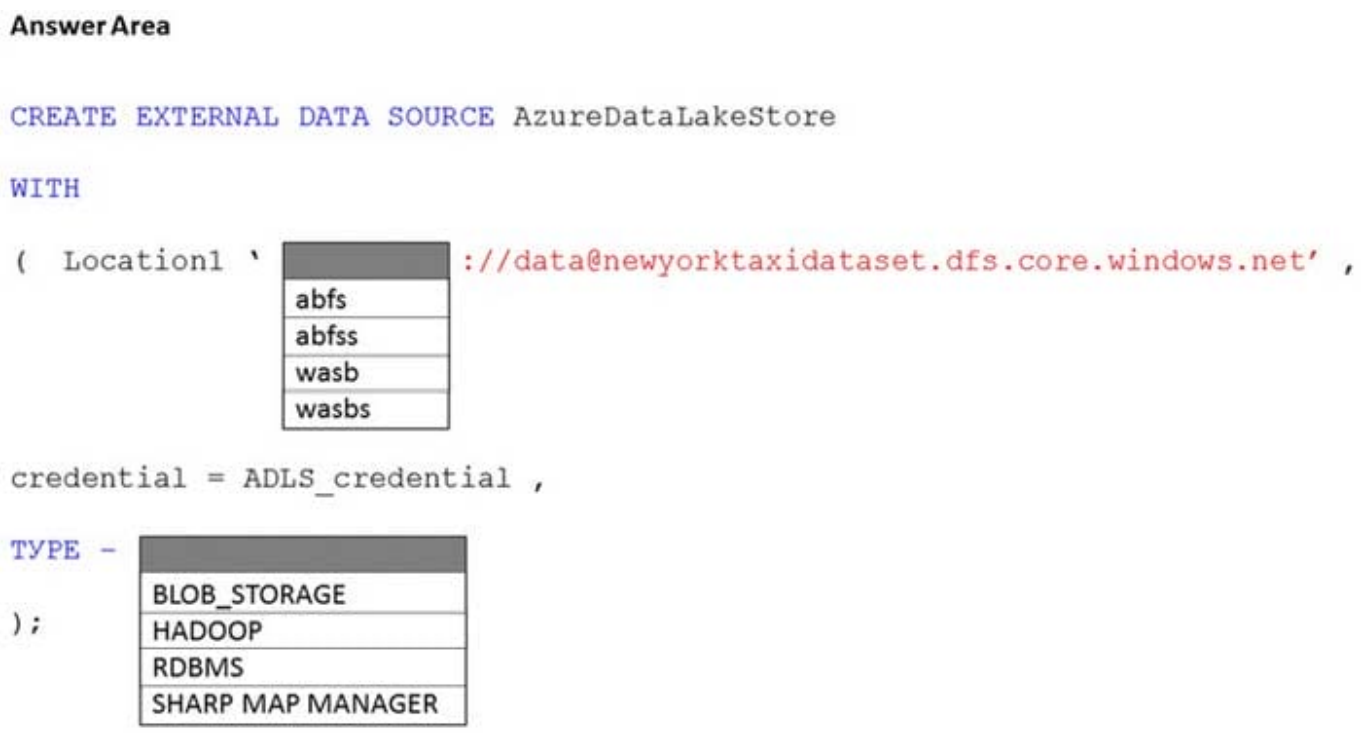
You have an Azure subscription that contains an Azure Synapse Analytics dedicated SQL pool named Pool1 and an Azure Data Lake Storage account named storage1. Storage1 requires secure transfers.

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:



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 7

You manage an enterprise data warehouse in Azure Synapse Analytics.

Users report slow performance when they run commonly used queries. Users do not report performance changes for infrequently used queries.

You need to monitor resource utilization to determine the source of the performance issues.

Which metric should you monitor?

A. DWU percentage

B. Cache hit percentage

C. DWU limit

D. Data IO percentage

Correct Answer: B

You can use Azure Monitor to view cache metrics to troubleshoot query performance.

The key metrics for troubleshooting the cache are Cache hit percentage and Cache used percentage.

Possible scenario: Your current working data set cannot fit into the cache which causes a low cache hit percentage due to physical reads. Consider scaling up your performance level and rerun your workload to populate the cache.

Reference:

https://docs.microsoft.com/da-dk/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-how-to-monitor-cache

QUESTION 8

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?

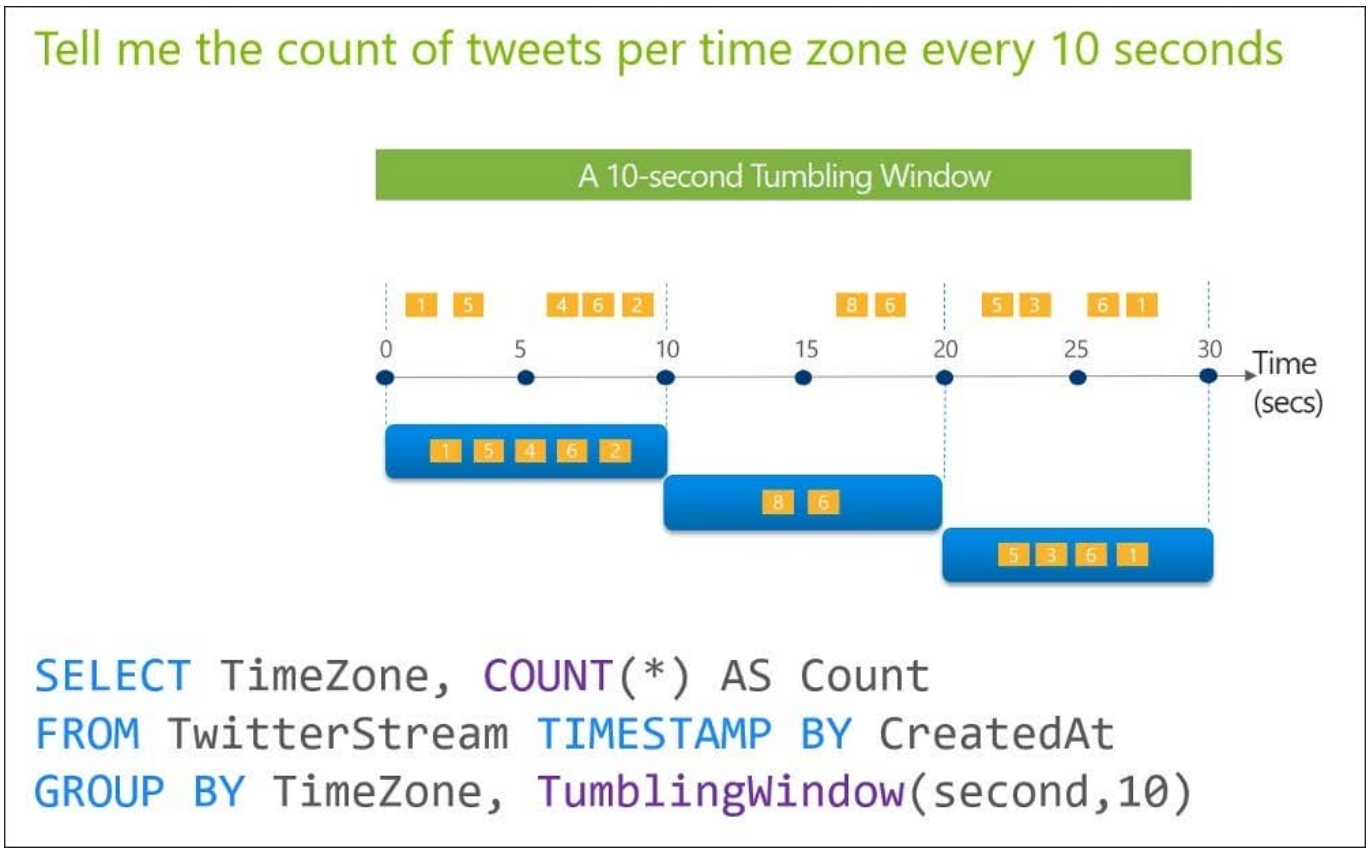
A. snapshot

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 9

HOTSPOT

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

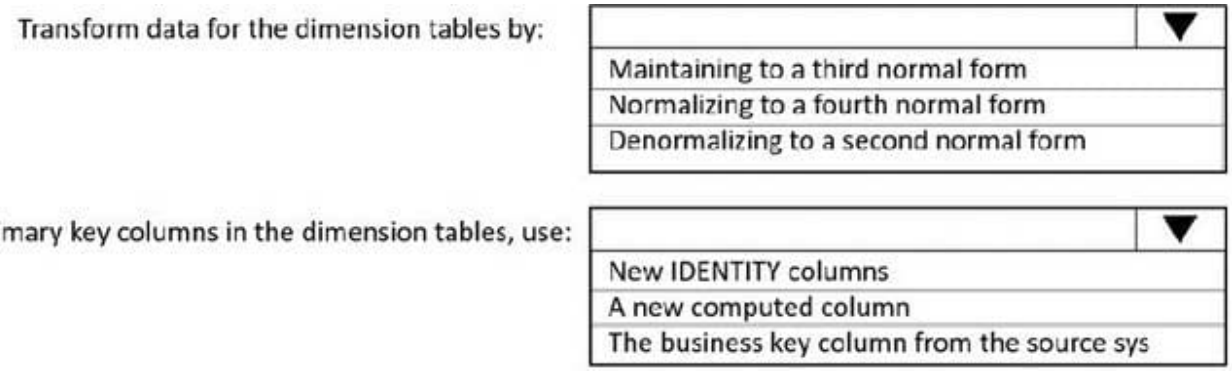
You plan to migrate the data in the database to a star schema in an A?\ire Synapse Analytics dedicated SQI pool.

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

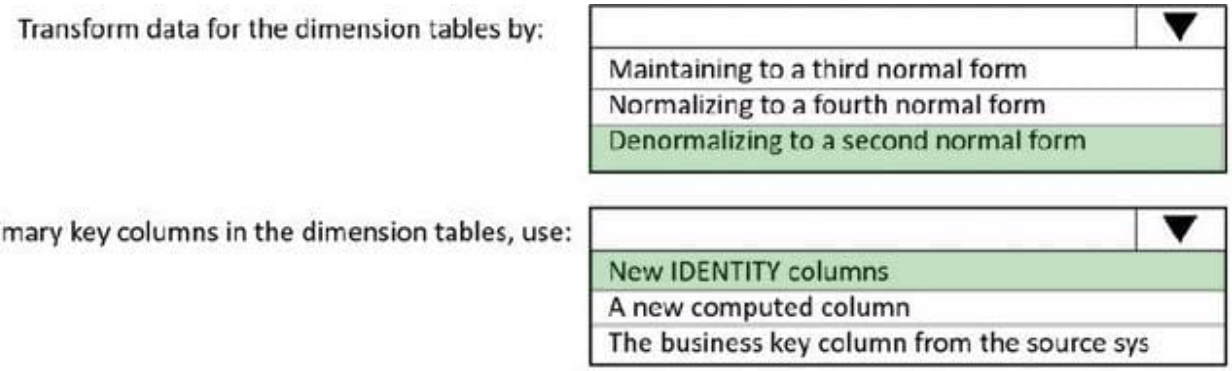
What should you include in the solution? to answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:



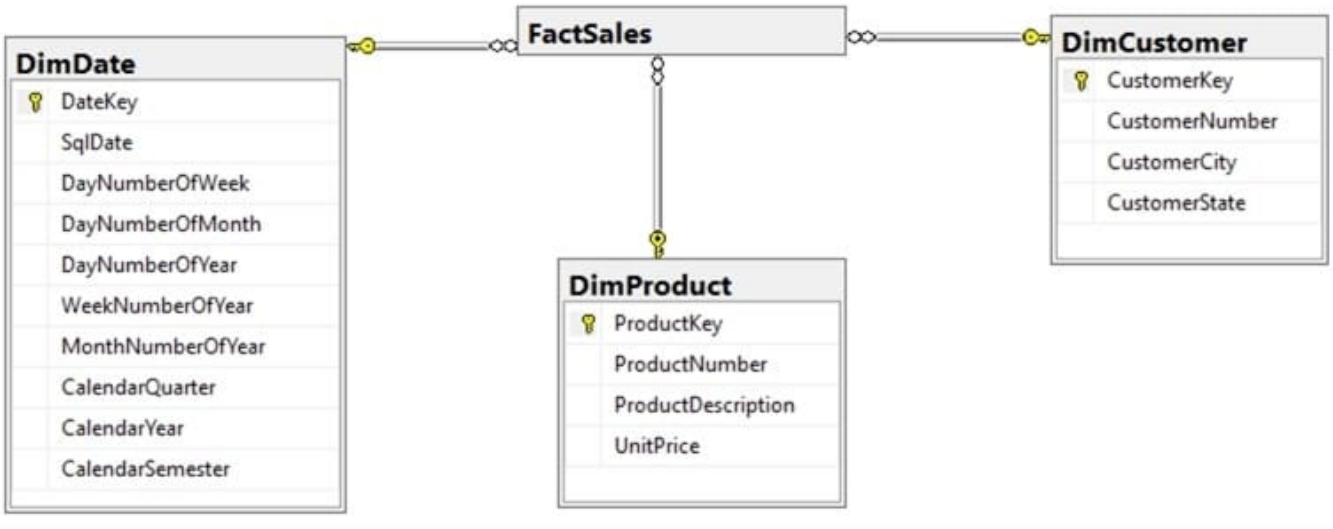
Correct Answer:



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 10

You have two Azure Blob Storage accounts named account1 and account2.

You plan to create an Azure Data Factory pipeline that will use scheduled intervals to replicate newly created or modified blobs from account1 to account2.

You need to recommend a solution to implement the pipeline. The solution must meet the following requirements:

1.

Ensure that the pipeline only copies blobs that were created or modified since the most recent replication event.

2.

Minimize the effort to create the pipeline. What should you recommend?

A. Run the Copy Data tool and select Metadata-driven copy task.

B. Create a pipeline that contains a Data Flow activity.

C. Create a pipeline that contains a flowlet.

D. Run the Copy Data tool and select Built-in copy task.

Correct Answer: A

Build large-scale data copy pipelines with metadata-driven approach in copy data tool

When you want to copy huge amounts of objects (for example, thousands of tables) or load data from large variety of sources, the appropriate approach is to input the name list of the objects with required copy behaviors in a control table,

and then use parameterized pipelines to read the same from the control table and apply them to the jobs accordingly. By doing so, you can maintain (for example, add/remove) the objects list to be copied easily by just updating the object

names in control table instead of redeploying the pipelines. What\\'s more, you will have single place to easily check which objects copied by which pipelines/triggers with defined copy behaviors.

Copy data tool in ADF eases the journey of building such metadata driven data copy pipelines. After you go through an intuitive flow from a wizard-based experience, the tool can generate parameterized pipelines and SQL scripts for you to

create external control tables accordingly. After you run the generated scripts to create the control table in your SQL database, your pipelines will read the metadata from the control table and apply them on the copy jobs automatically.

Incorrect:

Not C: A flowlet is a reusable container of activities that can be created from an existing mapping data flow or started from scratch. By reusing patterns you can prevent logic duplication and apply the same logic across many mapping data

flows.

With flowlets you can create logic to do things such as address cleaning or string trimming. You can then map the input and outputs to columns in the calling data flow for a dynamic code reuse experience.

Reference:

https://learn.microsoft.com/en-us/azure/data-factory/copy-data-tool-metadata-driven

---

QUESTION 11

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:

1.

Provide the fastest possible query times.

2.

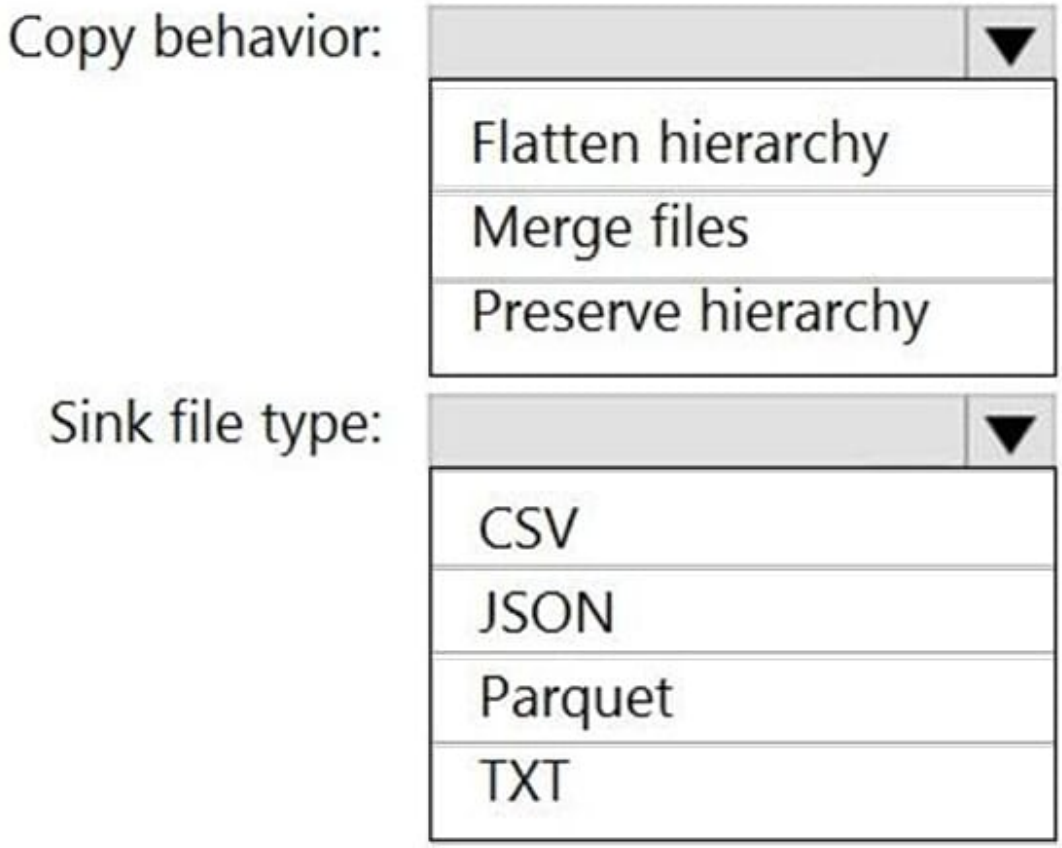
Automatically infer the schema from the underlying files.

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.

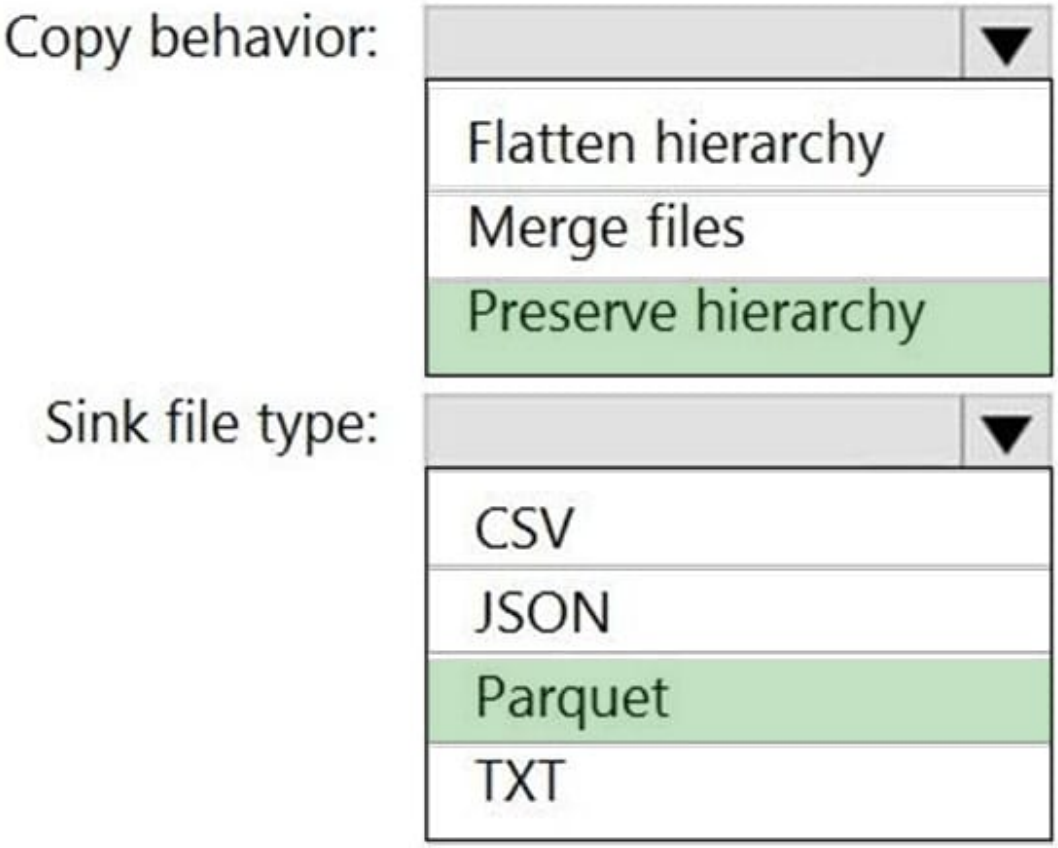
Hot Area:

Answer Area



Correct Answer:

Answer Area



Box 1: Preserver herarchy

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 12

DRAG DROP

You have an Azure subscription that contains an Azure Synapse Analytics workspace named workspace1. Workspace1 connects to an Azure DevOps repository named repo1. Repo1 contains a collaboration branch named main and a

development branch named branch1. Branch1 contains an Azure Synapse pipeline named pipeline1.

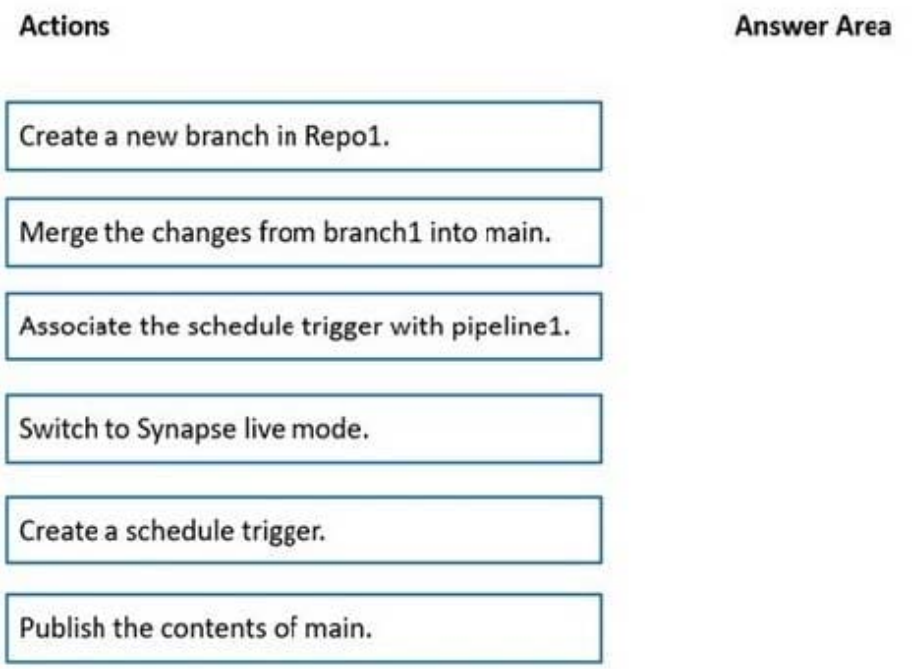
In workspace1, you complete testing of pipeline1.

You need to schedule pipeline1 to run daily at 6 AM.

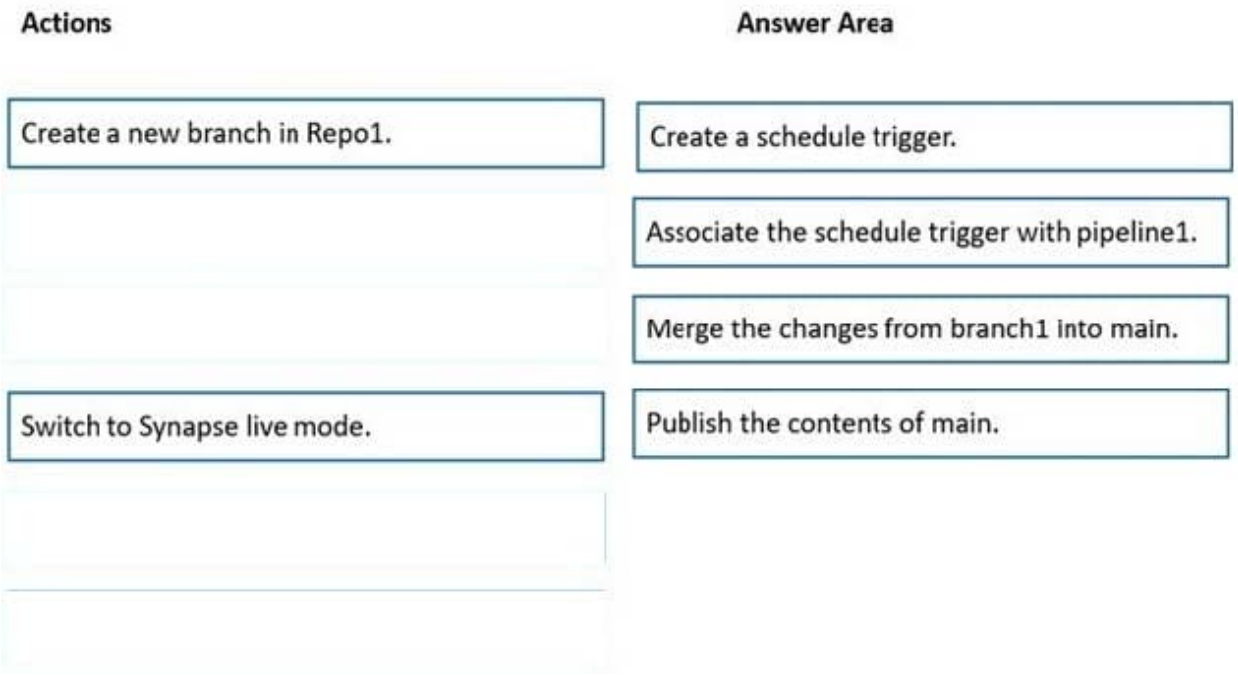
Which four actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

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:



Correct Answer:



QUESTION 13

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:

1.

Access multiple data sources.

2.

Provide the ability to orchestrate workflow.

3.

Provide the capability to run SQL Server Integration Services packages.

Store:

1.

Optimize storage for big data workloads.

2.

Provide encryption of data at rest.

3.

Operate with no size limits.

Prepare and Train:

1.

Provide a fully-managed and interactive workspace for exploration and visualization.

2.

Provide the ability to program in R, SQL, Python, Scala, and Java.

3.

Provide seamless user authentication with Azure Active Directory.

Model and Serve:

1.

Implement native columnar storage.

2.

Support for the SQL language

3.

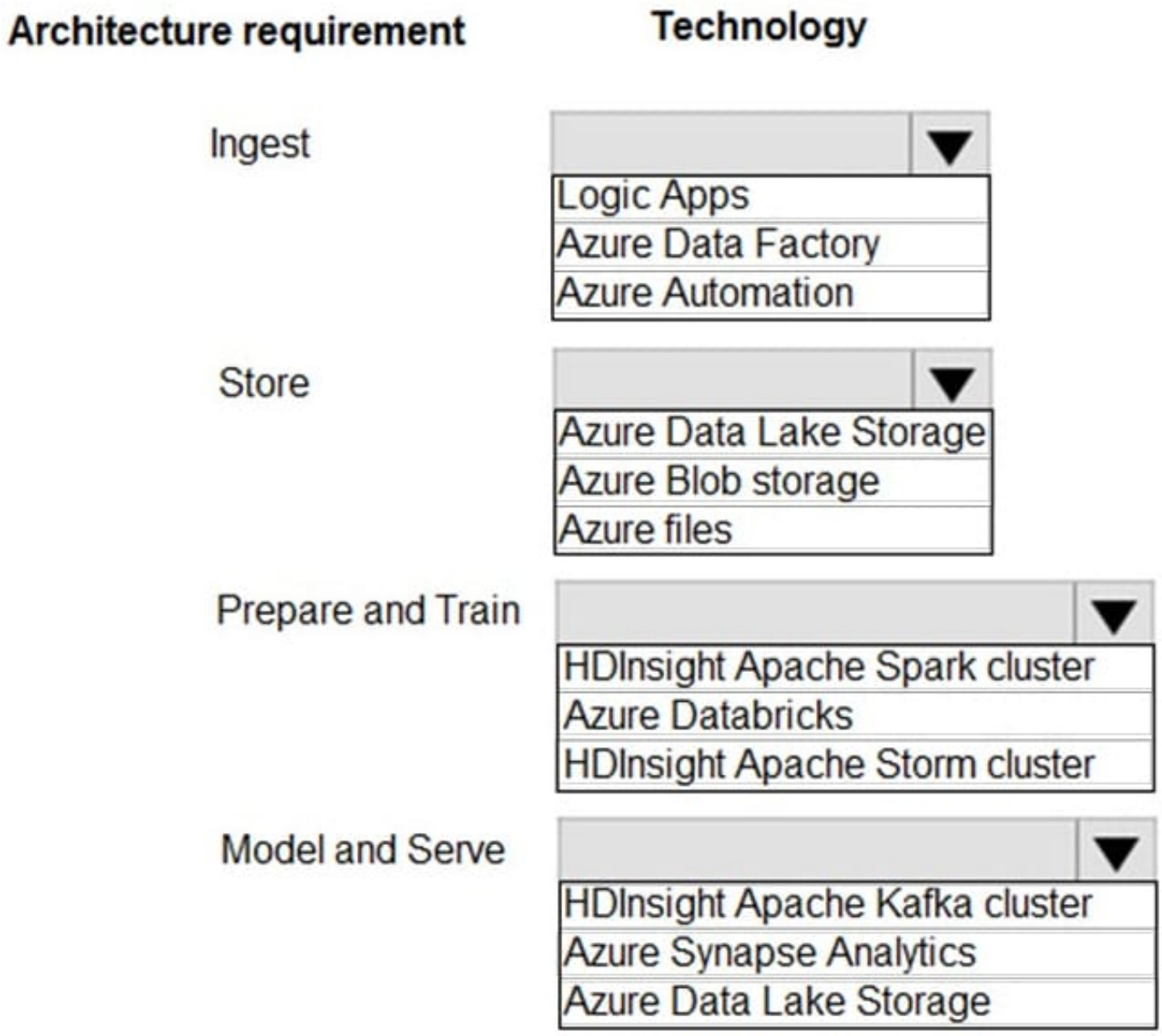
Provide support for structured streaming.

You need to build the data integration pipeline.

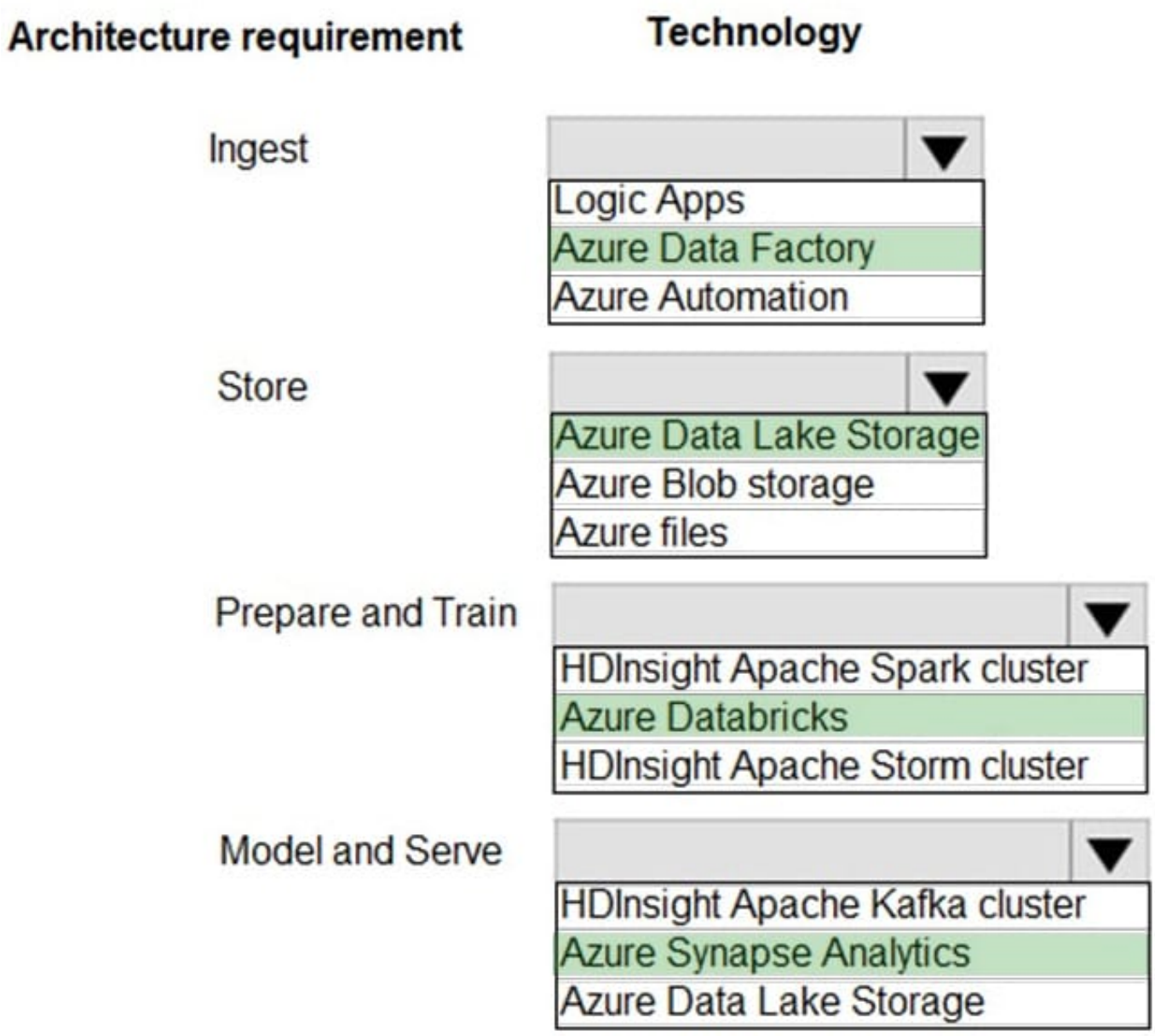
Which technologies should you use? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:



Correct Answer:



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: 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 14

You have an Azure Databricks workspace that contains a Delta Lake dimension table named Table1.

Table1 is a Type 2 slowly changing dimension (SCD) table.

You need to apply updates from a source table to Table1.

Which Apache Spark SQL operation should you use?

A. CREATE

B. UPDATE

C. MERGE

D. ALTER

Correct Answer: C

The Delta provides the ability to infer the schema for data input which further reduces the effort required in managing the schema changes. The Slowly Changing Data(SCD) Type 2 records all the changes made to each key in the dimensional

table. These operations require updating the existing rows to mark the previous values of the keys as old and then inserting new rows as the latest values. Also, Given a source table with the updates and the target table with dimensional data,

SCD Type 2 can be expressed with the merge.

Example:

// Implementing SCD Type 2 operation using merge function customersTable as("customers")

merge(

stagedUpdates.as("staged_updates"),

"customers.customerId = mergeKey")

whenMatched("customers.current = true AND customers.address staged_updates.address") updateExpr(Map(

"current" -> "false",

"endDate" -> "staged_updates.effectiveDate"))

whenNotMatched()

insertExpr(Map(

"customerid" -> "staged_updates.customerId",

"address" -> "staged_updates.address",

"current" -> "true",

"effectiveDate" -> "staged_updates.effectiveDate", "endDate" -> "null")) execute()

}

Reference:

https://www.projectpro.io/recipes/what-is-slowly-changing-data-scd-type-2-operation-delta- table- databricks

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