DUMPSGATE

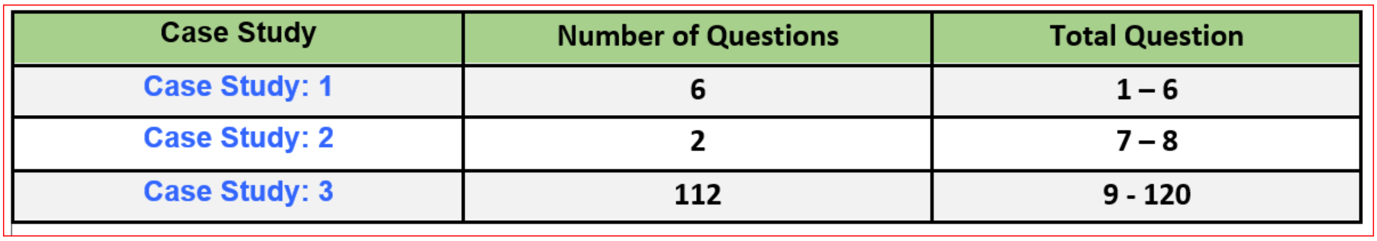
# Microsoft

## DP-203

Data Engineering on Microsoft Azure

## QUESTION & ANSWERS

QUESTION 1



| Case Study | Number of Questions | Total Question |
|---|---|---|
| Case Study: 1 | 6 | 1 – 6 |
| Case Study: 2 | 2 | 7 – 8 |
| Case Study: 3 | 112 | 9 - 120 |

# Case Study: 1

## Contoso

Transactional Date

Streaming Twitter Data

Planned Changes

Sales Transaction Dataset Requirements

### Transactional Date

Contoso has three years of customer, transactional, operation, sourcing, and supplier data comprised of 10 billion records stored across multiple on-premises Microsoft SQL Server servers. The SQL server instances contain data from various operational systems. The data is loaded into the instances by using SQL server integration Services (SSIS) packages.

You estimate that combining all product sales transactions into a company-wide sales transactions dataset will result in a single table that contains 5 billion rows, with one row per transaction.

Most queries targeting the sales transactions data will be used to identify which products were sold in retail stores and which products were sold online during different time period. Sales transaction data that is older than three years will be removed monthly.

You plan to create a retail store table that will contain the address of each retail store. The table will be approximately 2 MB. Queries for retail store sales will include the retail store addresses.

You plan to create a promotional table that will contain a promotion ID. The promotion ID will be associated to a specific product. The product will be identified by a product ID. The table will be approximately 5 GB.

### Streaming Twitter Data

The ecommerce department at Contoso develops and Azure logic app that captures trending Twitter feeds referencing the company's products and pushes the products to Azure Event Hubs.

## Planned Changes

Contoso plans to implement the following changes:

* Load the sales transaction dataset to Azure Synapse Analytics.

* Integrate on-premises data stores with Azure Synapse Analytics by using SSIS packages.

* Use Azure Synapse Analytics to analyze Twitter feeds to assess customer sentiments about products.

## Sales Transaction Dataset Requirements

Contoso identifies the following requirements for the sales transaction dataset:

* Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month. Boundary values must belong: to the partition on the right.

* Ensure that queries joining and filtering sales transaction records based on product ID complete as quickly as possible.

* Implement a surrogate key to account for changes to the retail store addresses.

* Ensure that data storage costs and performance are predictable.

* Minimize how long it takes to remove old records.

## Customer Sentiment Analytics Requirement

Contoso identifies the following requirements for customer sentiment analytics:

* Allow Contoso users to use PolyBase in an A/ure Synapse Analytics dedicated SQL pool to query the content of the data records that host the Twitter feeds. Data must be protected by using row-level security (RLS). The users must be authenticated by using their own A/ureAD credentials.

* Maximize the throughput of ingesting Twitter feeds from Event Hubs to Azure Storage without purchasing additional throughput or capacity units.

* Store Twitter feeds in Azure Storage by using Event Hubs Capture. The feeds will be converted into Parquet files.

* Ensure that the data store supports Azure AD-based access control down to the object level.

* Minimize administrative effort to maintain the Twitter feed data records.

* Purge Twitter feed data records;itftaitJ are older than two years.

## Data Integration Requirements

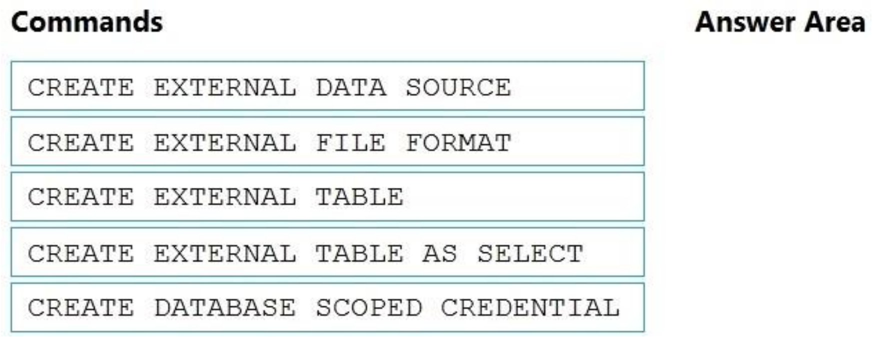
Contoso identifies the following requirements for data integration:

Use an Azure service that leverages the existing SSIS packages to ingest on-premises data into datasets stored in a dedicated SQL pool of Azure Synaps Analytics and transform the data.

Identify a process to ensure that changes to the ingestion and transformation activities can be version controlled and developed independently by multiple data engineers.

## Question: 1

You need to ensure that the Twitter feed data can be analyzed in the dedicated SQL pool. The solution must meet the customer sentiment analytics requirements.
Which three Transaction-SQL DDL commands should you run in sequence? To answer, move the appropriate commands from the list of commands 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.

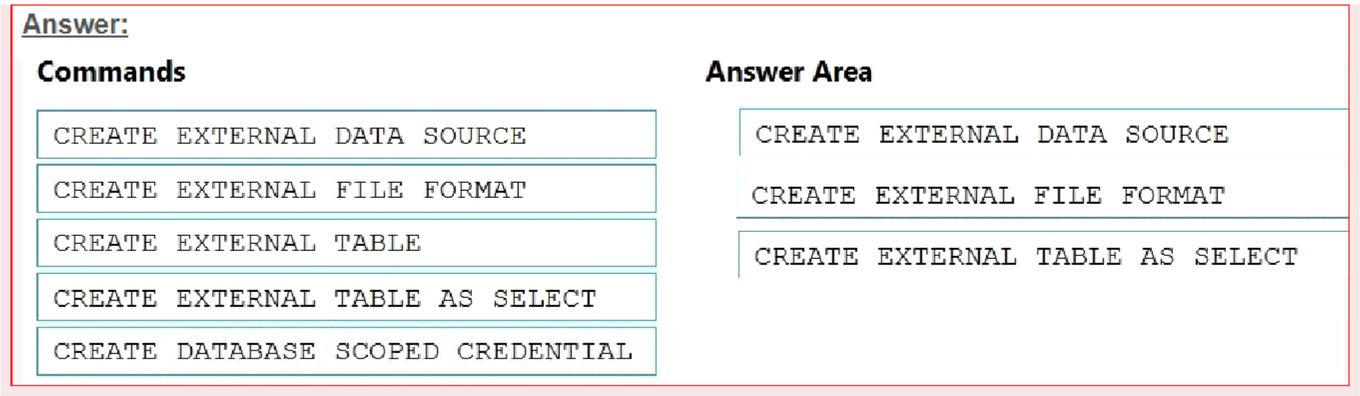


**Commands**

```
CREATE EXTERNAL DATA SOURCE
CREATE EXTERNAL FILE FORMAT
CREATE EXTERNAL TABLE
CREATE EXTERNAL TABLE AS SELECT
CREATE DATABASE SCOPED CREDENTIAL
```

**Answer Area**

Correct Answer:



Answer:

**Commands**

CREATE EXTERNAL DATA SOURCE

CREATE EXTERNAL FILE FORMAT

CREATE EXTERNAL TABLE

CREATE EXTERNAL TABLE AS SELECT

CREATE DATABASE SCOPED CREDENTIAL

**Answer Area**

CREATE EXTERNAL DATA SOURCE

CREATE EXTERNAL FILE FORMAT

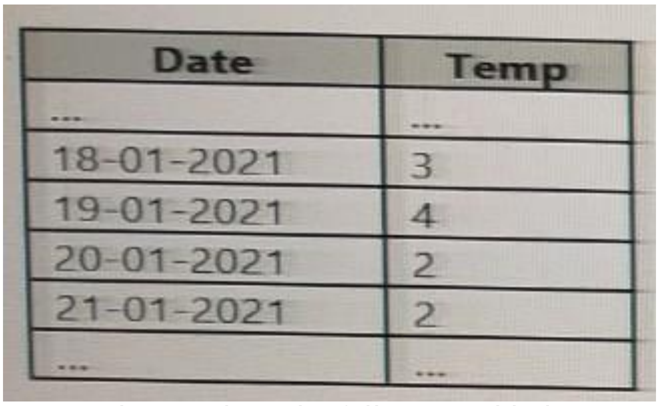
CREATE EXTERNAL TABLE AS SELECT

## Explanation/Reference:

Explanation:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables

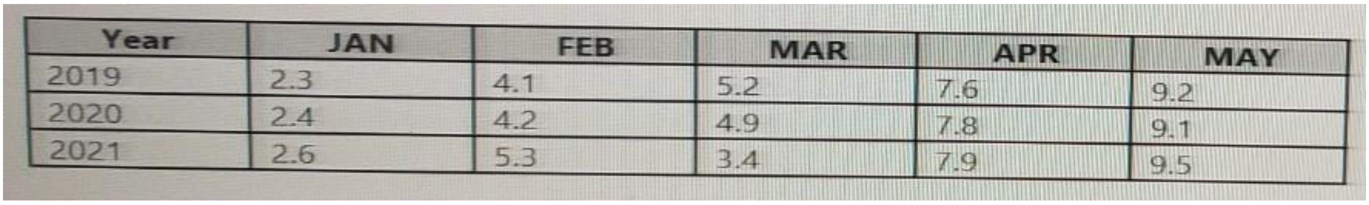
## QUESTION 2

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



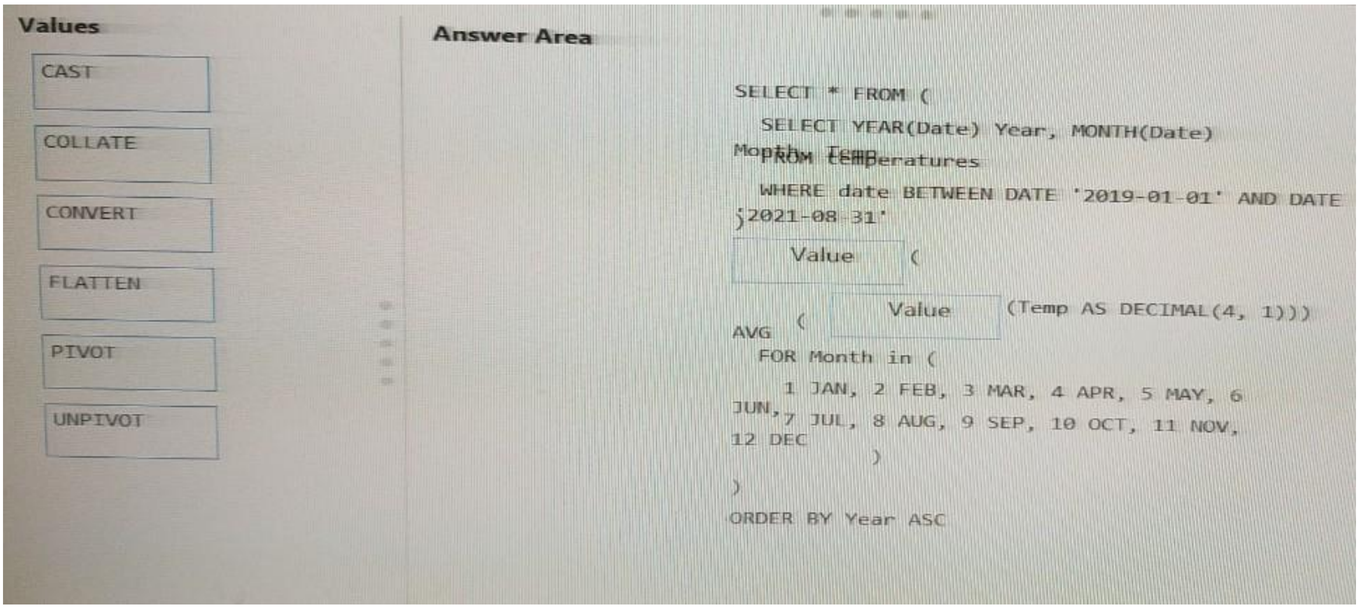
| Date | Temp |
|---|---|
| ... | ... |
| 18-01-2021 | 3 |
| 19-01-2021 | 4 |
| 20-01-2021 | 2 |
| 21-01-2021 | 2 |
| ... | ... |

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.



**Values**

CAST

COLLATE

CONVERT

FLATTEN

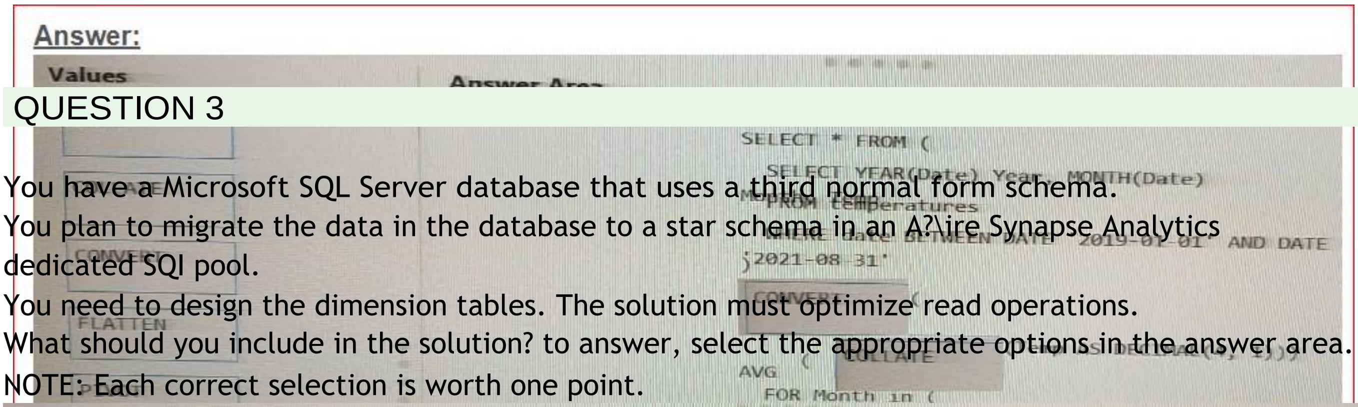
PIVOT

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'
)
     Value      (
               Value      (Temp AS DECIMAL(4, 1)))
AVG   (
  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:



Answer:

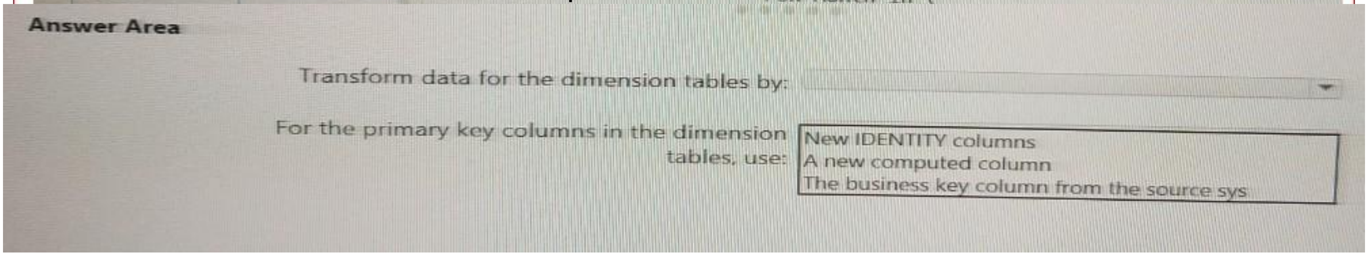
Values

Answer Area

```
SELECT * FROM (
  SELECT YEAR(Date) Year, MONTH(Date)
  FROM temperatures
  WHERE Date BETWEEN DATE '2019-01-01' AND DATE
'2021-08-31'
CONVERT
AVG (   COLLATE
  FOR Month in (
```

QUESTION 3

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 SQl 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.

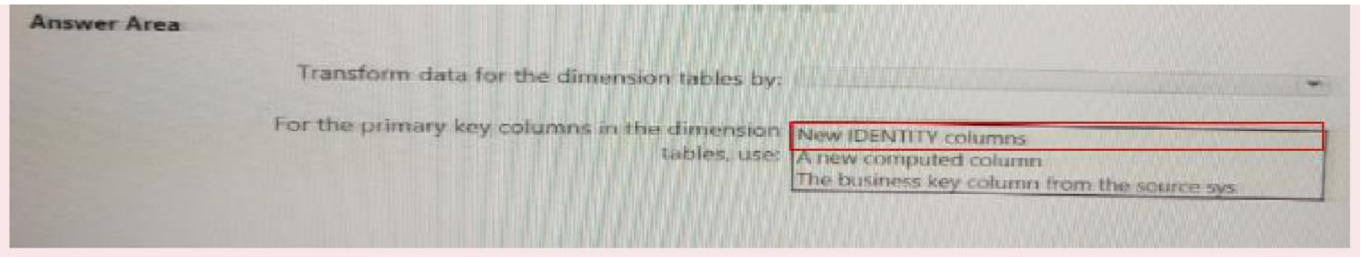


Answer Area

Transform data for the dimension tables by:

For the primary key columns in the dimension tables, use:
- New IDENTITY columns
- A new computed column
- The business key column from the source sys

Correct Answer:



**Answer Area**

Transform data for the dimension tables by:

For the primary key columns in the dimension tables, use:
[✓ New IDENTITY columns]
A new computed column
The business key column from the source sys

QUESTION 4

You have an Azure event hub named retailhub that has 16 partitions. Transactions are posted to retailhub. Each transaction includes the transaction ID, the individual line items, and the payment details. The transaction ID is used as the partition key.

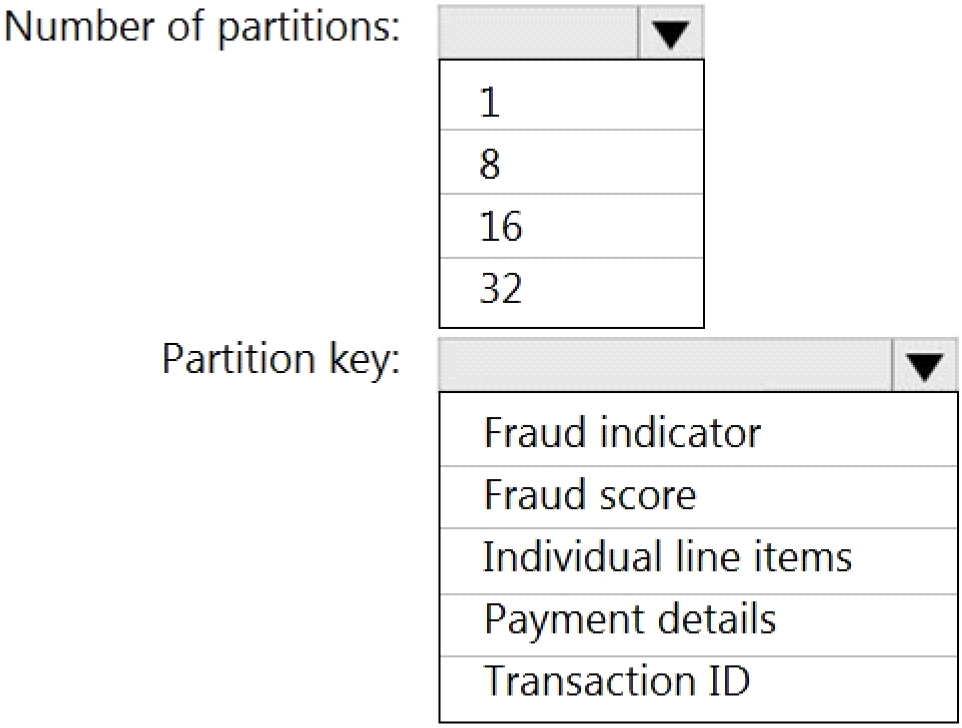
You are designing an Azure Stream Analytics job to identify potentially fraudulent transactions at a retail store. The job will use retailhub as the input. The job will output the transaction ID, the individual line items, the payment details, a fraud score, and a fraud indicator.

You plan to send the output to an Azure event hub named fraudhub.

You need to ensure that the fraud detection solution is highly scalable and processes transactions as quickly as possible.

How should you structure the output of the Stream Analytics job? To answer, select the appropriate options in the answer area.

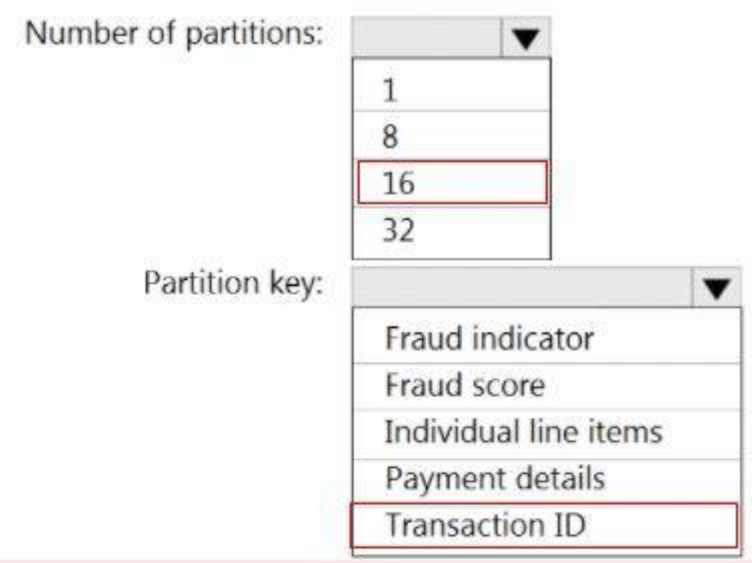
NOTE: Each correct selection is worth one point.



Number of partitions:
1
8
16
32

Partition key:
Fraud indicator
Fraud score
Individual line items
Payment details
Transaction ID

Correct Answer:



Number of partitions:
- 1
- 8
- [✓ 16]
- 32

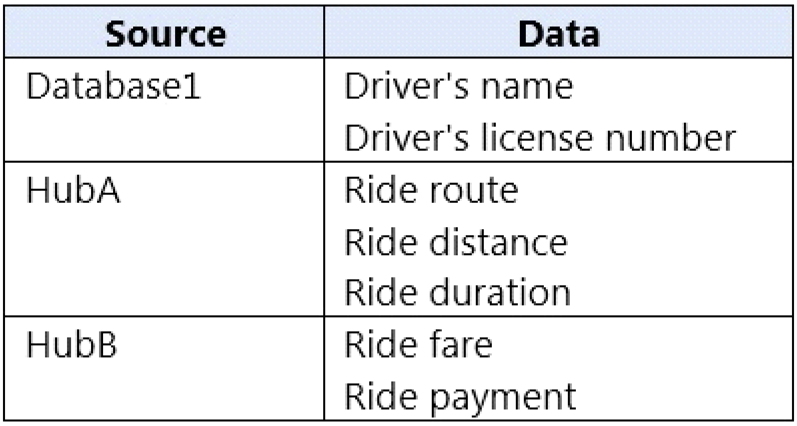
Partition key:
- Fraud indicator
- Fraud score
- Individual line items
- Payment details
- [✓ Transaction ID]

## Explanation/Reference:

Explanation:
https://docs.microsoft.com/en-us/azure/event-hubs/event-hubs-features#partitions

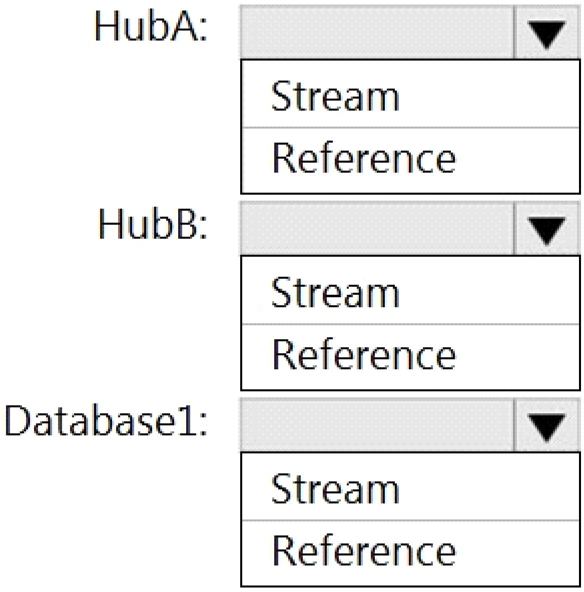
## QUESTION 5

You have an Azure SQL database named Database1 and two Azure event hubs named HubA and HubB. The data consumed from each source is shown in the following table.



| Source | Data |
|---|---|
| Database1 | Driver's name<br>Driver's license number |
| HubA | Ride route<br>Ride distance<br>Ride duration |
| HubB | Ride fare<br>Ride payment |

You need to implement Azure Stream Analytics to calculate the average fare per mile by driver.
How should you configure the Stream Analytics input for each source? To answer, select the appropriate options in the answer are
a.
NOTE: Each correct selection is worth one point.

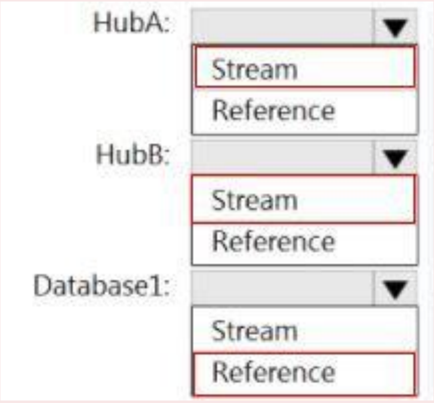


HubA: Stream / Reference

HubB: Stream / Reference

Database1: Stream / Reference

Correct Answer:



HubA: [✓ Stream] Reference

HubB: [✓ Stream] Reference

Database1: Stream [✓ Reference]

## Explanation/Reference:

Explanation:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-use-reference-data

## QUESTION 6

You are designing a solution that will copy Parquet files stored in an Azure Blob storage account to an Azure Data Lake Storage Gen2 account.
The data will be loaded daily to the data lake and will use a folder structure of {Year}/{Month}/{Day}/.
You need to design a daily Azure Data Factory data load to minimize the data transfer between the two
accounts.
Which two configurations should you include in the design? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.

A. Delete the files in the destination before loading new data.
B. Filter by the last modified date of the source files.

C. Delete the source files after they are copied.
D. Specify a file naming pattern for the destination.

Correct Answer: B,C

## Explanation/Reference:

Explanation:
https://docs.microsoft.com/en-us/azure/data-factory/connector-azure-data-lake-storage

## QUESTION 7

You configure monitoring for a Microsoft Azure SQL Data Warehouse implementation. The implementation uses PolyBase to load data from comma-separated value (CSV) files stored in Azure Data Lake Gen 2 using an external table.
Files with an invalid schema cause errors to occur.
You need to monitor for an invalid schema error.
For which error should you monitor?

A. EXTERNAL TABLE access failed due to internal error: 'Java exception raised on call to HdfsBridge_Connect: Error[com.microsoft.polybase.client.KerberosSecureLogin] occurred while accessingexternal files.'
B. EXTERNAL TABLE access failed due to internal error: 'Java exception raised on call to HdfsBridge_Connect: Error [No FileSystem for scheme: wasbs] occurred while accessing external file.'
C. Cannot execute the query 'Remote Query' against OLE DB provider 'SQLNCLI11': for linked server '(null)', Query aborted- the maximum reject threshold (orows) was reached while regarding from an external source: 1 rows rejected out of total 1 rows processed.
D. EXTERNAL TABLE access failed due to internal error: 'Java exception raised on call to HdfsBridge_Connect: Error [Unable to instantiate LoginClass] occurredwhile accessing external files.'

Correct Answer: C

## Explanation/Reference:

Customer Scenario:
SQL Server 2016 or SQL DW connected to Azure blob storage. The CREATE EXTERNAL TABLE DDL points to a directory (and not a specific file) and the directory contains files with different schemas.
SSMS Error:
Select query on the external table gives the following error:
Msg 7320, Level 16, State 110, Line 14
Cannot execute the query 'Remote Query' against OLE DB provider 'SQLNCLI11' for linked server '(null)'. Query aborted-- the maximum reject threshold (0 rows) was reached while reading from an

external source: 1 rows rejected out of total 1 rows processed.
Possible Reason:
The reason this error happens is because each file has different schema. The PolyBase external table DDL when pointed to a directory recursively reads all the files in that directory. When a column or data type mismatch happens, this error could be seen in SSMS.
Possible Solution:
If the data for each table consists of one file, then use the filename in the LOCATION section prepended by the directory of the external files. If there are multiple files per table, put each set of files into different directories in Azure Blob Storage and then you can point LOCATION to the directory instead of a particular file. The latter suggestion is the best practices recommended by SQLCAT even if you have one file per table.
Incorrect Answers:
A: Possible Reason: Kerberos is not enabled in Hadoop Cluster.
References:
https://techcommunity.microsoft.com/t5/DataCAT/PolyBase-Setup-Errors-and-Possible-Solutions/ba-p/305297

QUESTION 8

# Case Study: 2

## Litware, inc.

Overview

Requirements

### Overview

Litware, Inc. owns and operates 300 convenience stores across the US. The company sells a variety of packaged foods and drinks, as well as a variety of prepared foods, such as sandwiches and pizzas.

Litware has a loyalty club whereby members can get daily discounts on specific items by providing their membership number at checkout.

Litware employs business analysts who prefer to analyze data by using Microsoft Power BI, and data scientists who prefer analyzing data in Azure Databricks notebooks.

### Requirements

Litware wants to create a new analytics environment in Azure to meet the following requirements:

See inventory levels across the stores. Data must be updated as close to real time as possible.

Execute ad hoc analytical queries on historical data to identify whether the loyalty club discounts increase sales of the discounted products.

Every four hours, notify store employees about how many prepared food items to produce based on historical demand from the sales data.

## Technical Requirements

Litware identifies the following technical requirements:

Minimize the number of different Azure services needed to achieve the business goals.

Use platform as a service (PaaS) offerings whenever possible and avoid having to provision virtual machines that must be managed by Litware.

Ensure that the analytical data store is accessible only to the company's on-premises network and Azure services.

Use Azure Active Directory (Azure AD) authentication whenever possible.

Use the principle of least privilege when designing security.

Stage Inventory data in Azure Data Lake Storage Gen2 before loading the data into the analytical data store. Litware wants to remove transient data from Data Lake Storage once the data is no longer in use. Files that have a modified date that is older than 14 days must be removed.

Limit the business analysts' access to customer contact information, such as phone numbers, because this type of data is not analytically relevant.

Ensure that you can quickly restore a copy of the analytical data store within one hour in the event of corruption or accidental deletion.

## Planned Environment

Litware plans to implement the following environment:

The application development team will create an Azure event hub to receive real-time sales data, including store number, date, time, product ID, customer loyalty number, price, and discount amount, from the point of sale (POS) system and output the data to data storage in Azure.

Customer data, including name, contact information, and loyalty number, comes from Salesforce, a SaaS application, and can be imported into Azure once every eight hours. Row modified dates are not trusted in the source table.

Product data, including product ID, name, and category, comes from Salesforce and can be imported into Azure once every eight hours. Row modified dates are not trusted in the source table.

Daily inventory data comes from a Microsoft SQL server located on a private network.

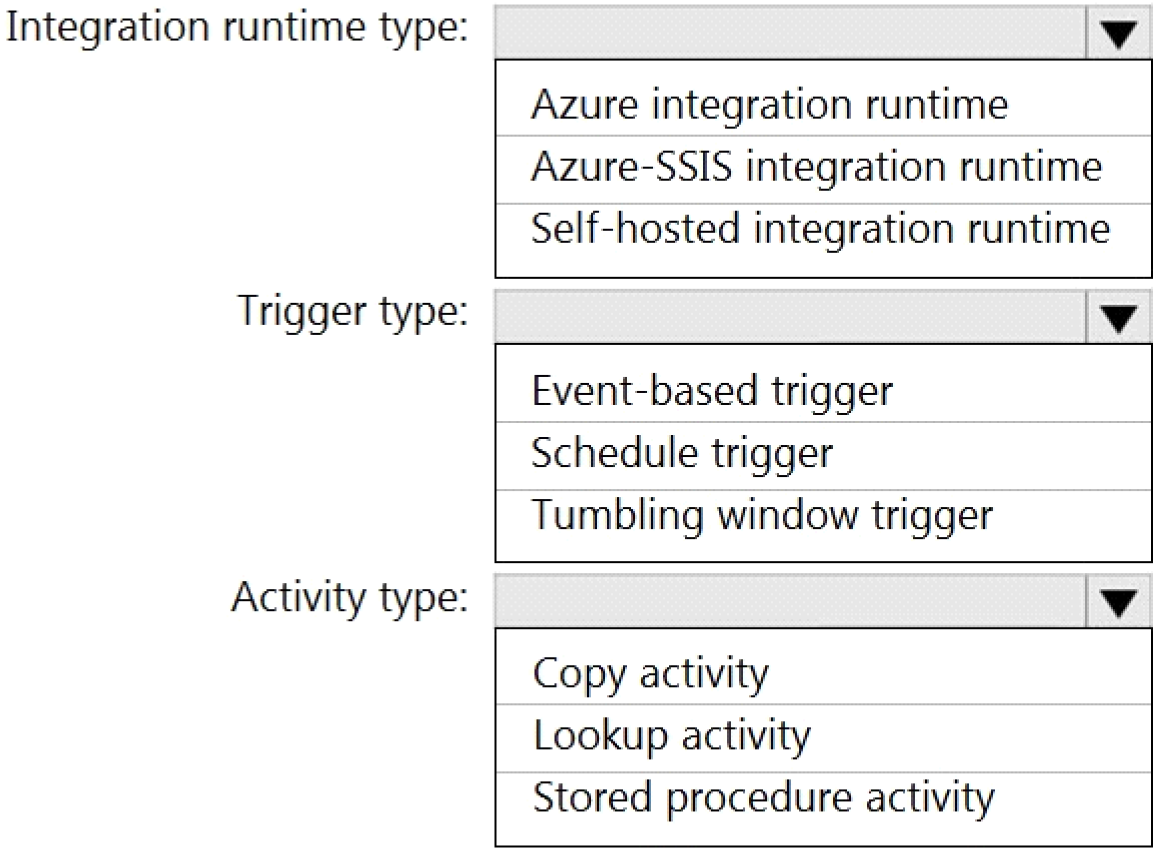
Litware currently has 5 TB of historical sales data and 100 GB of customer data. The company expects approximately 100 GB of new data per month for the next year.

Litware will build a custom application named FoodPrep to provide store employees with the calculation results of how many prepared food items to produce every four hours.

Litware does not plan to implement Azure ExpressRoute or a VPN between the on-premises network and Azure.

Question:

Which Azure Data Factory components should you recommend using together to import the daily inventory data from the SQL server to Azure Data Lake Storage? To answer, select the appropriate options in the answer are
a.
NOTE: Each correct selection is worth one point.



Integration runtime type: ▼

- Azure integration runtime
- Azure-SSIS integration runtime
- Self-hosted integration runtime

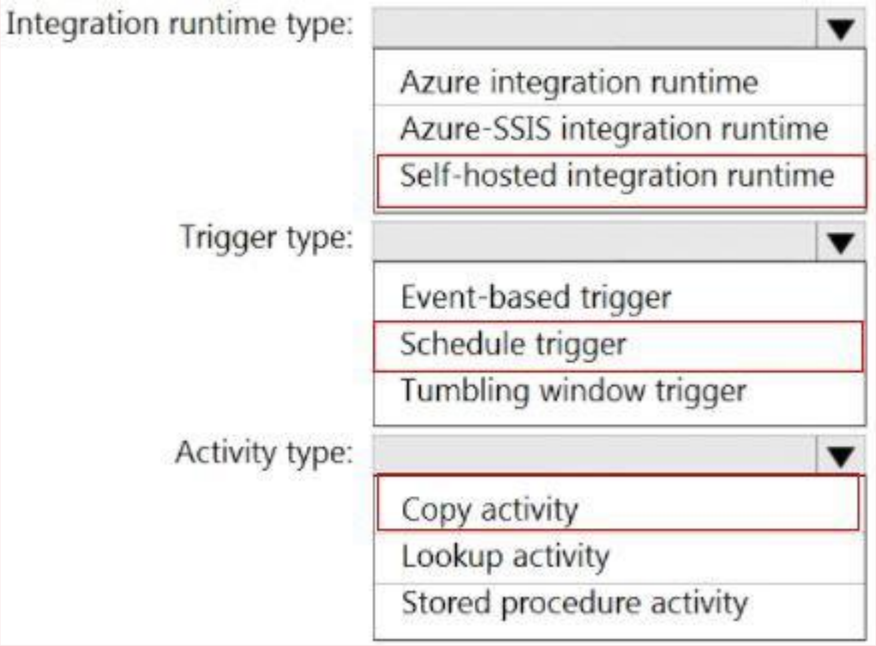
Trigger type: ▼

- Event-based trigger
- Schedule trigger
- Tumbling window trigger

Activity type: ▼

- Copy activity
- Lookup activity
- Stored procedure activity

Correct Answer:



Integration runtime type: ▼

- Azure integration runtime
- Azure-SSIS integration runtime
- [✓ Self-hosted integration runtime]

Trigger type: ▼

- Event-based trigger
- [✓ Schedule trigger]
- Tumbling window trigger

Activity type: ▼

- [✓ Copy activity]
- Lookup activity
- Stored procedure activity

## QUESTION 9

You have an enterprise-wide Azure Data Lake Storage Gen2 account. The data lake is accessible only through an Azure virtual network named VNET1.

You are building a SQL pool in Azure Synapse that will use data from the data lake.

Your company has a sales team. All the members of the sales team are in an Azure Active Directory group named Sales. POSIX controls are used to assign the Sales group access to the files in the data lake.

You plan to load data to the SQL pool every hour.

You need to ensure that the SQL pool can load the sales data from the data lake.

Which three actions should you perform? Each correct answer presents part of the solution.

NOTE: Each area selection is worth one point.

A. Add the managed identity to the Sales group.
B. Use the managed identity as the credentials for the data load process.
C. Create a shared access signature (SAS).
D. Add your Azure Active Directory (Azure AD) account to the Sales group.
E. Use the snared access signature (SAS) as the credentials for the data load process.
F. Create a managed identity.

Correct Answer: A,D,F

## Explanation/Reference:

The managed identity grants permissions to the dedicated SQL pools in the workspace.

Note: Managed identity for Azure resources is a feature of Azure Active Directory. The feature provides Azure
services with an automatically managed identity in Azure AD

https://docs.microsoft.com/en-us/azure/synapse-analytics/security/synapse-workspace-managed-identity

## QUESTION 10

What should you recommend to prevent users outside the Litware on-premises network from accessing the analytical data store?

A. a server-level virtual network rule
B. a database-level virtual network rule
C. a database-level firewall IP rule
D. a server-level firewall IP rule

## Correct Answer: A

## Explanation/Reference:

Virtual network rules are one firewall security feature that controls whether the database server for your single databases and elastic pool in Azure SQL Database or for your databases in SQL Data Warehouse accepts communications that are sent from particular subnets in virtual networks.
Server-level, not database-level: Each virtual network rule applies to your whole Azure SQL Database server, not just to one particular database on the server. In other words, virtual network rule applies at the serverlevel, not at the database-level.
References:
https://docs.microsoft.com/en-us/azure/sql-database/sql-database-vnet-service-endpoint-rule-overview

## QUESTION 11

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

A. Clone the cluster after it is terminated.
B. Terminate the cluster manually when processing completes.
C. Create an Azure runbook that starts the cluster every 90 days.
D. Pin the cluster.

## Correct Answer: D

## Explanation/Reference:

To keep an interactive cluster configuration even after it has been terminated for more than 30 days, an
administrator can pin a cluster to the cluster list.
References:
https://docs.azuredatabricks.net/clusters/clusters-manage.html#automatic-termination