



DP-203

Data Engineering on Microsoft Azure



EXAMKILLER

Help Pass Your Exam At First Try

Microsoft

Exam DP-203

Data Engineering on Microsoft Azure

Version: 17.0

[Total Questions: 289]

Topic break down

Topic	No. of Questions
Topic 1: Contoso Case StudyTransactional Date	11
Topic 2: Litware, inc.	4
Topic 3: Mix Questions	274

Topic 1, Contoso Case Study Transactional Data

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 an 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 Azure 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 AzureAD 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 that 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 No : 1 HOTSPOT - (Topic 1)

You need to design an analytical storage solution for the transactional data. The solution must meet the sales transaction dataset requirements.

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.

Table type to store retail store data:

	▼
Hash	
Replicated	
Round-robin	

Table type to store promotional data:

	▼
Hash	
Replicated	
Round-robin	

Answer:

Table type to store retail store data:

	▼
Hash	
Replicated	
Round-robin	

Table type to store promotional data:

	▼
Hash	
Replicated	
Round-robin	

Explanation:

Table type to store retail store data:

	▼
Hash	
Replicated	
Round-robin	

Table type to store promotional data:

	▼
Hash	
Replicated	
Round-robin	

Graphical user

interface, text, application, table

Description automatically generated

Box 1: Round-robin

Round-robin tables are useful for improving loading speed.

Scenario: Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month.

Box 2: Hash

Hash-distributed tables improve query performance on large fact tables.

Question No : 2 DRAG DROP - (Topic 1)

You need to implement versioned changes to the integration pipelines. The solution must meet the data integration requirements.

In which order should you perform the actions? To answer, move all actions from the list of actions to the answer area and arrange them in the correct order.

Actions	Answer Area
Publish changes.	
Create a feature branch.	
Merge changes.	
Create a repository and a main branch.	
Create a pull request.	

Navigation arrows: > and <

Answer:

Actions	Answer Area
Publish changes.	Create a repository and a main branch.
Create a feature branch.	Create a feature branch.
Merge changes.	Create a pull request.
Create a repository and a main branch.	Merge changes.
Create a pull request.	Publish changes.

Navigation arrows: > and <

Explanation:

Create a repository and a main branch

Create a feature branch

Create a pull request

Merge changes

Publish changes

Graphical user

interface, application

Description automatically generated

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

Step 1: Create a repository and a main branch

You need a Git repository in Azure Pipelines, TFS, or GitHub with your app.

Step 2: Create a feature branch

Step 3: Create a pull request

Step 4: Merge changes

Merge feature branches into the main branch using pull requests.

Step 5: Publish changes

Question No : 3 HOTSPOT - (Topic 1)

You need to design a data ingestion and storage solution for the Twitter feeds. The solution

must meet the customer sentiment analytics requirements.

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

To increase the throughput of ingesting the Twitter feeds:

- Configure Event Hubs partitions.
- Enable Auto-Inflate in Event Hubs.
- Use Event Hubs Dedicated.

To store the Twitter feed data, use:

- An Azure Data Lake Storage Gen2 account
- An Azure Databricks high concurrency cluster
- An Azure General-purpose v2 storage account in the Premium tier

Answer:

Answer Area

To increase the throughput of ingesting the Twitter feeds:

- Configure Event Hubs partitions.
- Enable Auto-Inflate in Event Hubs.
- Use Event Hubs Dedicated.

To store the Twitter feed data, use:

- An Azure Data Lake Storage Gen2 account
- An Azure Databricks high concurrency cluster
- An Azure General-purpose v2 storage account in the Premium tier

Explanation:

To increase the throughput of ingesting the Twitter feeds:

- Configure Event Hubs partitions.
- Enable Auto-Inflate in Event Hubs.
- Use Event Hubs Dedicated.

To store the Twitter feed data, use:

- An Azure Data Lake Storage Gen2 account
- An Azure Databricks high concurrency cluster
- An Azure General-purpose v2 storage account in the Premium tier

Graphical user

interface, text

Description automatically generated

Box 1: Configure Event Hubs partitions

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

Event Hubs is designed to help with processing of large volumes of events. Event Hubs throughput is scaled by using partitions and throughput-unit allocations.

Event Hubs traffic is controlled by TUs (standard tier). Auto-inflate enables you to start small with the minimum required TUs you choose. The feature then scales automatically to the maximum limit of TUs you need, depending on the increase in your traffic.

Box 2: An Azure Data Lake Storage Gen2 account

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

Azure Data Lake Storage Gen2 implements an access control model that supports both Azure role-based access control (Azure RBAC) and POSIX-like access control lists (ACLs).

Question No : 4 DRAG DROP - (Topic 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

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:

```
CREATE EXTERNAL DATA SOURCE
```

```
CREATE EXTERNAL FILE FORMAT
```

```
CREATE EXTERNAL TABLE AS SELECT
```

Scenario: Allow Contoso users to use PolyBase in an Azure 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 Azure AD credentials.

Box 1: CREATE EXTERNAL DATA SOURCE

External data sources are used to connect to storage accounts.

Box 2: CREATE EXTERNAL FILE FORMAT

CREATE EXTERNAL FILE FORMAT creates an external file format object that defines external data stored in Azure Blob Storage or Azure Data Lake Storage. Creating an external file format is a prerequisite for creating an external table.

Box 3: CREATE EXTERNAL TABLE AS SELECT

When used in conjunction with the CREATE TABLE AS SELECT statement, selecting from an external table imports data into a table within the SQL pool. In addition to the COPY statement, external tables are useful for loading data.

Question No : 5 HOTSPOT - (Topic 1)

You need to design a data storage structure for the product sales transactions. The solution must meet the sales transaction dataset requirements.

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

Table type to store the product sales transactions:

When creating the table for sales transactions:

Answer:

Answer Area

Table type to store the product sales transactions:

When creating the table for sales transactions:

Explanation:

Table type to store the product sales transactions:

Hash
Round-robin
Replicated

When creating the table for sales transactions:

Configure a clustered index.
Set the distribution column to product ID.
Set the distribution column to the sales date.

Graphical user

interface, text, application, chat or text message

Description automatically generated

Box 1: Hash

Scenario:

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

A hash distributed table can deliver the highest query performance for joins and aggregations on large tables.

Box 2: Set the distribution column to the sales date.

Scenario: 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.

Question No : 6 - (Topic 1)

You need to integrate the on-premises data sources and Azure Synapse Analytics. The solution must meet the data integration requirements.

Which type of integration runtime should you use?

- A. Azure-SSIS integration runtime
- B. self-hosted integration runtime
- C. Azure integration runtime

Answer: C

Question No : 7 - (Topic 1)

You need to implement the surrogate key for the retail store table. The solution must meet the sales transaction

dataset requirements.

What should you create?

- A. a table that has an IDENTITY property
- B. a system-versioned temporal table
- C. a user-defined SEQUENCE object
- D. a table that has a FOREIGN KEY constraint

Answer: A

Explanation:

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

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.

Reference:

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

Question No : 8 HOTSPOT - (Topic 1)

You need to design the partitions for the product sales transactions. The solution must meet the sales transaction dataset requirements.

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.

Partition product sales transactions data by:

	▼
Sales date	
Product ID	
Promotion ID	

Store product sales transactions data in:

	▼
An Azure Synapse Analytics dedicated SQL pool	
An Azure Synapse Analytics serverless SQL pool	
An Azure Data Lake Storage Gen2 account linked to an Azure Synapse Analytics workspace	

Answer:

Partition product sales transactions data by:

▼
Sales date
Product ID
Promotion ID

Store product sales transactions data in:

▼
An Azure Synapse Analytics dedicated SQL pool
An Azure Synapse Analytics serverless SQL pool
An Azure Data Lake Storage Gen2 account linked to an Azure Synapse Analytics workspace

Explanation:

Partition product sales transactions data by:

▼
Sales date
Product ID
Promotion ID

Store product sales transactions data in:

▼
An Azure Synapse Analytics dedicated SQL pool
An Azure Synapse Analytics serverless SQL pool
An Azure Data Lake Storage Gen2 account linked to an Azure Synapse Analytics workspace

Box 1: Sales date

Scenario: Contoso requirements for data integration include:

- ✍ 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.

Box 2: An Azure Synapse Analytics Dedicated SQL pool

Scenario: Contoso requirements for data integration include:

- ✍ Ensure that data storage costs and performance are predictable.

The size of a dedicated SQL pool (formerly SQL DW) is determined by Data Warehousing Units (DWU).

Dedicated SQL pool (formerly SQL DW) stores data in relational tables with columnar storage. This format significantly reduces the data storage costs, and improves query performance.

Synapse analytics dedicated sql pool

Question No : 9 - (Topic 1)

You need to design a data retention solution for the Twitter feed data records. The solution must meet the customer sentiment analytics requirements.

Which Azure Storage functionality should you include in the solution?

- A. time-based retention
- B. change feed
- C. soft delete
- D. lifecycle management

Answer: D

Question No : 10 HOTSPOT - (Topic 1)

You need to implement an Azure Synapse Analytics database object for storing the sales transactions data. The solution must meet the sales transaction dataset requirements.

What solution must meet the sales transaction dataset requirements.

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

NOTE: Each correct selection is worth one point.

Transact-SQL DDL command to use:

	▼
CREATE EXTERNAL TABLE	
CREATE TABLE	
CREATE VIEW	

Partitioning option to use in the WITH clause of the DDL statement:

	▼
FORMAT_OPTIONS	
FORMAT_TYPE	
RANGE LEFT FOR VALUES	
RANGE RIGHT FOR VALUES	

Answer:

Transact-SQL DDL command to use:

	▼
CREATE EXTERNAL TABLE	
CREATE TABLE	
CREATE VIEW	

Partitioning option to use in the WITH clause of the DDL statement:

	▼
FORMAT_OPTIONS	
FORMAT_TYPE	
RANGE LEFT FOR VALUES	
RANGE RIGHT FOR VALUES	

Explanation:

Transact-SQL DDL command to use:

	▼
CREATE EXTERNAL TABLE	
CREATE TABLE	
CREATE VIEW	

Partitioning option to use in the WITH clause of the DDL statement:

	▼
FORMAT_OPTIONS	
FORMAT_TYPE	
RANGE LEFT FOR VALUES	
RANGE RIGHT FOR VALUES	

Graphical user

interface, text, application, table

Description automatically generated

Box 1: Create table

Scenario: Load the sales transaction dataset to Azure Synapse Analytics

Box 2: RANGE RIGHT FOR VALUES

Scenario: 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.

RANGE RIGHT: Specifies the boundary value belongs to the partition on the right (higher values).

FOR VALUES (boundary_value [,...n]): Specifies the boundary values for the partition.

Scenario: Load the sales transaction dataset to Azure Synapse Analytics.

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.

Question No : 11 - (Topic 1)

You need to design a data retention solution for the Twitter feed data records. The solution must meet the customer sentiment analytics requirements.

Which Azure Storage functionality should you include in the solution?

- A. change feed
- B. soft delete
- C. time-based retention
- D. lifecycle management

Answer: B

Explanation:

Scenario: Purge Twitter feed data records that are older than two years.

Data sets have unique lifecycles. Early in the lifecycle, people access some data often. But the need for access often drops drastically as the data ages. Some data remains idle in the cloud and is rarely accessed once stored. Some data sets expire days or months after creation, while other data sets are actively read and modified throughout their lifetimes. Azure Storage lifecycle management offers a rule-based policy that you can use to transition blob data to the appropriate access tiers or to expire data at the end of the data lifecycle.

Reference:

<https://docs.microsoft.com/en-us/azure/storage/blobs/lifecycle-management-overview>

Topic 2, Litware, inc.

Case study

This is a case study. Case studies are not timed separately. You can use as much exam time as you would like to complete each case. However, there may be additional case studies and sections on this exam. You must manage your time to ensure that you are able to complete all questions included on this exam in the time provided.

To answer the questions included in a case study, you will need to reference information that is provided in the case study. Case studies might contain exhibits and other resources that provide more information about the scenario that is described in the case study. Each question is independent of the other questions in this case study.

At the end of this case study, a review screen will appear. This screen allows you to review your answers and to make changes before you move to the next section of the exam. After you begin a new section, you cannot return to this section.

To start the case study

To display the first question in this case study, click the **Next** button. Use the buttons in the left pane to explore the content of the case study before you answer the questions. Clicking these buttons displays information such as business requirements, existing environment, and problem statements. If the case study has an **All Information** tab, note that the information displayed is identical to the information displayed on the subsequent tabs. When you are ready to answer a question, click the **Question** button to return to the question.

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

Business Goals

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 No : 12 HOTSPOT - (Topic 2)

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 area.


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


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

Explanation:

Integration runtime type:	<div><div></div><div>Azure integration runtime</div><div>Azure-SSIS integration runtime</div><div>Self-hosted integration runtime</div></div>
Trigger type:	<div><div></div><div>Event-based trigger</div><div>Schedule trigger</div><div>Tumbling window trigger</div></div>
Activity type:	<div><div></div><div>Copy activity</div><div>Lookup activity</div><div>Stored procedure activity</div></div>

Box 1: Self-hosted integration runtime

A self-hosted IR is capable of running copy activity between a cloud data stores and a data store in private network.

Box 2: Schedule trigger

Schedule every 8 hours

Box 3: Copy activity

Scenario:

- ✎ Customer data, including name, contact information, and loyalty number, comes from Salesforce 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.

Question No : 13 - (Topic 2)

What should you recommend using to secure sensitive customer contact information?

- A. data labels
- B. column-level security
- C. row-level security
- D. Transparent Data Encryption (TDE)

Answer: B

Explanation:

Scenario: All cloud data must be encrypted at rest and in transit.

Always Encrypted is a feature designed to protect sensitive data stored in specific database columns from access (for example, credit card numbers, national identification numbers, or data on a need to know basis). This includes database administrators or other privileged users who are authorized to access the database to perform management tasks, but have no business need to access the particular data in the encrypted columns. The data is always encrypted, which means the encrypted data is decrypted only for processing by client applications with access to the encryption key.

References:

<https://docs.microsoft.com/en-us/azure/sql-database/sql-database-security-overview>

Question No : 14 - (Topic 2)

What should you do to improve high availability of the real-time data processing solution?

- A. Deploy identical Azure Stream Analytics jobs to paired regions in Azure.
- B. Deploy a High Concurrency Databricks cluster.
- C. Deploy an Azure Stream Analytics job and use an Azure Automation runbook to check the status of the job and to start the job if it stops.
- D. Set Data Lake Storage to use geo-redundant storage (GRS).

Answer: A

Explanation:

Guarantee Stream Analytics job reliability during service updates

Part of being a fully managed service is the capability to introduce new service functionality and improvements at a rapid pace. As a result, Stream Analytics can have a service update deploy on a weekly (or more frequent) basis. No matter how much testing is done there is still a risk that an existing, running job may break due to the introduction of a bug. If you are

running mission critical jobs, these risks need to be avoided. You can reduce this risk by following Azure's paired region model.

Scenario: 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

Reference:

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-job-reliability>

Question No : 15 - (Topic 2)

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

Answer: A

Explanation:

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>

Topic 3, Mix Questions

Question No : 16 - (Topic 3)

You have an Azure Synapse Analytics Apache Spark pool named Pool1.

You plan to load JSON files from an Azure Data Lake Storage Gen2 container into the tables in Pool1. The structure and data types vary by file.

You need to load the files into the tables. The solution must maintain the source data types.

What should you do?

- A. Use a Get Metadata activity in Azure Data Factory.
- B. Use a Conditional Split transformation in an Azure Synapse data flow.
- C. Load the data by using the OPEHROWset Transact-SQL command in an Azure Synapse Analytics serverless SQL pool.
- D. Load the data by using PySpark.

Answer: A

Explanation:

Serverless SQL pool can automatically synchronize metadata from Apache Spark. A serverless SQL pool database will be created for each database existing in serverless Apache Spark pools.

Serverless SQL pool enables you to query data in your data lake. It offers a T-SQL query surface area that accommodates semi-structured and unstructured data queries.

To support a smooth experience for in place querying of data that's located in Azure Storage files, serverless SQL pool uses the OPENROWSET function with additional capabilities.

The easiest way to see to the content of your JSON file is to provide the file URL to the OPENROWSET function, specify csv FORMAT.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/query-json-files>

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/query-data-storage>

Question No : 17 - (Topic 3)

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.

Answer: B,D

Explanation:

Copy data from one place to another. The requirements are : 1- need to minimize transfert and 2- need to adapte data to the destination folder structure. Filter on LastModifiedDate will copy everything that have changed since the latest load while minimizing the data transfert. Specifying the file naming pattern allows to copy data at the right place to the destination Data Lake.

Question No : 18 DRAG DROP - (Topic 3)

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

Data in the table will be retained for five years. Once a year, data that is older than five years will be deleted.

You need to ensure that the data is distributed evenly across partitions. The solution must minimize the amount of time required to delete old data.

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.

Values

- CustomerKey
- HASH
- ROUND_ROBIN
- REPLICATE
- OrderDateKey
- SalesOrderNumber

Answer Area

```
CREATE TABLE [dbo].[FactSales]
(
    [ProductKey]          int          NOT NULL
,   [OrderDateKey]       int          NOT NULL
,   [CustomerKey]        int          NOT NULL
,   [SalesOrderNumber]   nvarchar ( 20 ) NOT NULL
,   [OrderQuantity]      smallint     NOT NULL
,   [UnitPrice]          money        NOT NULL
)
WITH
(   CLUSTERED            COLUMNSTORE      INDEX
,   DISTRIBUTION = [Value] ([ProductKey])
,   PARTITION ( [ [Value] ] RANGE RIGHT FOR VALUES
                (20170101,20180101,20190101,20200101,20210101)
                )
)
```

Answer:

Values

- CustomerKey
- HASH
- ROUND_ROBIN
- REPLICATE
- OrderDateKey
- SalesOrderNumber

Answer Area

```
CREATE TABLE [dbo].[FactSales]
(
    [ProductKey]          int          NOT NULL
,   [OrderDateKey]       int          NOT NULL
,   [CustomerKey]        int          NOT NULL
,   [SalesOrderNumber]   nvarchar ( 20 ) NOT NULL
,   [OrderQuantity]      smallint     NOT NULL
,   [UnitPrice]          money        NOT NULL
)
WITH
(   CLUSTERED            COLUMNSTORE      INDEX
,   DISTRIBUTION = HASH ([ProductKey])
,   PARTITION ( [ OrderDateKey ] RANGE RIGHT FOR VALUES
                (20170101,20180101,20190101,20200101,20210101)
                )
)
```

Explanation:

Box 1: HASH

Box 2: OrderDateKey

In most cases, table partitions are created on a date column.

A way to eliminate rollbacks is to use Metadata Only operations like partition switching for data management. For example, rather than execute a DELETE statement to delete all rows in a table where the order_date was in October of 2001, you could partition your data early. Then you can switch out the partition with data for an empty partition from another table.

Question No : 19 - (Topic 3)

You are performing exploratory analysis of the bus fare data in an Azure Data Lake Storage Gen2 account by using an Azure Synapse Analytics serverless SQL pool.

You execute the Transact-SQL query shown in the following exhibit.

```
SELECT
    payment_type,
    SUM(fare_amount) AS fare_total
FROM OPENROWSET (
    BULK 'csv/busfare/tripdata_2020*.csv',
    DATA_SOURCE = 'BusData',
    FORMAT = 'CSV', PARSER_VERSION = '2.0',
    FIRSTROW = 2
)
WITH (
    payment_type INT 10,
    fare_amount FLOAT 11
) AS nyc
GROUP BY payment_type
ORDER BY payment_type;
```

What do the query results include?

- A. Only CSV files in the tripdata_2020 subfolder.
- B. All files that have file names that beginning with "tripdata_2020".
- C. All CSV files that have file names that contain "tripdata_2020".

D. Only CSV that have file names that beginning with "tripdata_2020".

Answer: D

Question No : 20 - (Topic 3)

You are designing the folder structure for an Azure Data Lake Storage Gen2 container.

Users will query data by using a variety of services including Azure Databricks and Azure Synapse Analytics serverless SQL pools. The data will be secured by subject area. Most queries will include data from the current year or current month.

Which folder structure should you recommend to support fast queries and simplified folder security?

- A. /{SubjectArea}/{DataSource}/{DD}/{MM}/{YYYY}/{FileData}_{YYYY}_{MM}_{DD}.csv
- B. /{DD}/{MM}/{YYYY}/{SubjectArea}/{DataSource}/{FileData}_{YYYY}_{MM}_{DD}.csv
- C. /{YYYY}/{MM}/{DD}/{SubjectArea}/{DataSource}/{FileData}_{YYYY}_{MM}_{DD}.csv
- D. /{SubjectArea}/{DataSource}/{YYYY}/{MM}/{DD}/{FileData}_{YYYY}_{MM}_{DD}.csv

Answer: D

Explanation:

There's an important reason to put the date at the end of the directory structure. If you want to lock down certain regions or subject matters to users/groups, then you can easily do so with the POSIX permissions. Otherwise, if there was a need to restrict a certain security group to viewing just the UK data or certain planes, with the date structure in front a separate permission would be required for numerous directories under every hour directory. Additionally, having the date structure in front would exponentially increase the number of directories as time went on.

Note: In IoT workloads, there can be a great deal of data being landed in the data store that spans across numerous products, devices, organizations, and customers. It's important to pre-plan the directory layout for organization, security, and efficient processing of the data for down-stream consumers. A general template to consider might be the following layout:

{Region}/{SubjectMatter(s)}/{yyyy}/{mm}/{dd}/{hh}/

Question No : 21 - (Topic 3)

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.

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 plan to create an Azure Databricks workspace that has a tiered structure. The workspace will contain the following three workloads:

- ✍ A workload for data engineers who will use Python and SQL.
- ✍ A workload for jobs that will run notebooks that use Python, Scala, and SOL.
- ✍ A workload that data scientists will use to perform ad hoc analysis in Scala and R.

The enterprise architecture team at your company identifies the following standards for Databricks environments:

- ✍ The data engineers must share a cluster.
- ✍ The job cluster will be managed by using a request process whereby data scientists and data engineers provide packaged notebooks for deployment to the cluster.
- ✍ All the data scientists must be assigned their own cluster that terminates automatically after 120 minutes of inactivity. Currently, there are three data scientists.

You need to create the Databricks clusters for the workloads.

Solution: You create a High Concurrency cluster for each data scientist, a High Concurrency cluster for the data engineers, and a Standard cluster for the jobs.

Does this meet the goal?

- A. Yes**
- B. No**