DP-203

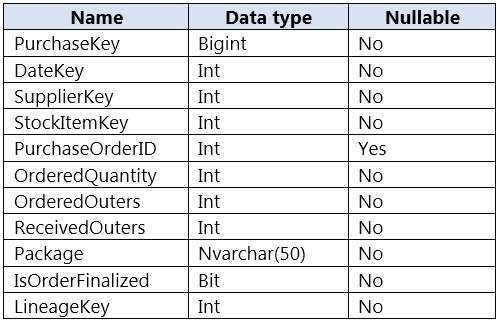
Number: DP-203 Passing Score: 800 Time Limit: 120 min File Version: 1

DP-203

# Design and implement data storage Question Set 1

**QUESTION 1**

You are designing a fact table named FactPurchase in an Azure Synapse Analytics dedicated SQL pool. The table contains purchases from suppliers for a retail store. FactPurchase will contain the following columns.



FactPurchase will have 1 million rows of data added daily and will contain three years of data. Transact-SQL queries similar to the following query will be executed daily.

SELECT

SupplierKey, StockItemKey, COUNT(\*)

FROM FactPurchase

WHERE DateKey >= 20210101 AND DateKey <= 20210131

GROUP By SupplierKey, StockItemKey

Which table distribution will minimize query times?

1. replicated
2. hash-distributed on PurchaseKey
3. round-robin
4. hash-distributed on DateKey

# Correct Answer: B Section: (none) Explanation

**Explanation/Reference:**

Explanation:

Hash-distributed tables improve query performance on large fact tables, and are the focus of this article. Round-robin tables are useful for improving loading speed.

Incorrect:

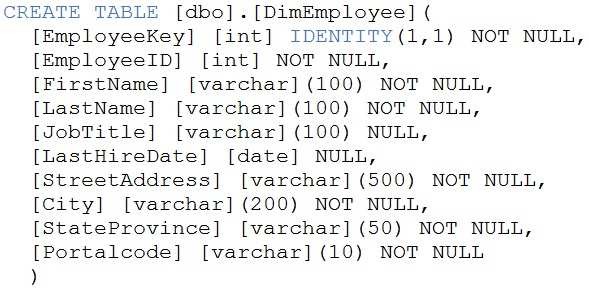
Not D: Do not use a date column. . All data for the same date lands in the same distribution. If several users are all filtering on the same date, then only 1 of the 60 distributions do all the processing work.

Reference:

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

# QUESTION 2

You have a table in an Azure Synapse Analytics dedicated SQL pool. The table was created by using the following Transact-SQL statement.



You need to alter the table to meet the following requirements:

Ensure that users can identify the current manager of employees.

Support creating an employee reporting hierarchy for your entire company. Provide fast lookup of the managers’ attributes such as name and job title.

Which column should you add to the table?

1. [ManagerEmployeeID] [int] NULL
2. [ManagerEmployeeID] [smallint] NULL
3. [ManagerEmployeeKey] [int] NULL
4. [ManagerName] [varchar](200) NULL

# Correct Answer: A Section: (none) Explanation

**Explanation/Reference:**

Explanation:

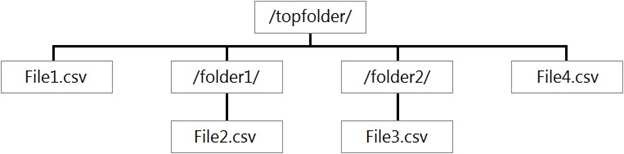
Use the same definition as the EmployeeID column.

Reference:

https://docs.microsoft.com/en-us/analysis-services/tabular-models/hierarchies-ssas-tabular

# QUESTION 3

You have files and folders in Azure Data Lake Storage Gen2 for an Azure Synapse workspace as shown in the following exhibit.



You create an external table named ExtTable that has LOCATION='/topfolder/'.

When you query ExtTable by using an Azure Synapse Analytics serverless SQL pool, which files are returned?

1. File2.csv and File3.csv only
2. File1.csv and File4.csv only
3. File1.csv, File2.csv, File3.csv, and File4.csv
4. File1.csv only

# Correct Answer: C Section: (none) Explanation

**Explanation/Reference:**

Explanation:

To run a T-SQL query over a set of files within a folder or set of folders while treating them as a single entity or rowset, provide a path to a folder or a pattern (using wildcards) over a set of files or folders.

Reference:

https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/query-data-storage#query-multiple-files-or-folders

# QUESTION 4

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?

1. /{SubjectArea}/{DataSource}/{DD}/{MM}/{YYYY}/{FileData}\_{YYYY}\_{MM}\_{DD}.csv
2. /{DD}/{MM}/{YYYY}/{SubjectArea}/{DataSource}/{FileData}\_{YYYY}\_{MM}\_{DD}.csv
3. /{YYYY}/{MM}/{DD}/{SubjectArea}/{DataSource}/{FileData}\_{YYYY}\_{MM}\_{DD}.csv
4. /{SubjectArea}/{DataSource}/{YYYY}/{MM}/{DD}/{FileData}\_{YYYY}\_{MM}\_{DD}.csv

# Correct Answer: D Section: (none) Explanation

**Explanation/Reference:**

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 5

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.

1. Add the managed identity to the Sales group.
2. Use the managed identity as the credentials for the data load process.
3. Create a shared access signature (SAS).
4. Add your Azure Active Directory (Azure AD) account to the Sales group.
5. Use the snared access signature (SAS) as the credentials for the data load process.
6. Create a managed identity.

# Correct Answer: ADF Section: (none) Explanation

**Explanation/Reference:**

Explanation:

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

Reference:

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

# Design and develop data processing Question Set 1

**QUESTION 1**

You are creating an Azure Data Factory data flow that will ingest data from a CSV file, cast columns to specified types of data, and insert the data into a table in an Azure Synapse Analytic dedicated SQL pool. The CSV file contains three columns named username, comment, and date.

The data flow already contains the following: A source transformation.

A Derived Column transformation to set the appropriate types of data. A sink transformation to land the data in the pool.

You need to ensure that the data flow meets the following requirements: All valid rows must be written to the destination table.

Truncation errors in the comment column must be avoided proactively.

Any rows containing comment values that will cause truncation errors upon insert must be written to a file in blob storage.

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

**NOTE:** Each correct selection is worth one point.

1. To the data flow, add a sink transformation to write the rows to a file in blob storage.
2. To the data flow, add a Conditional Split transformation to separate the rows that will cause truncation errors.
3. To the data flow, add a filter transformation to filter out rows that will cause truncation errors.
4. Add a select transformation to select only the rows that will cause truncation errors.

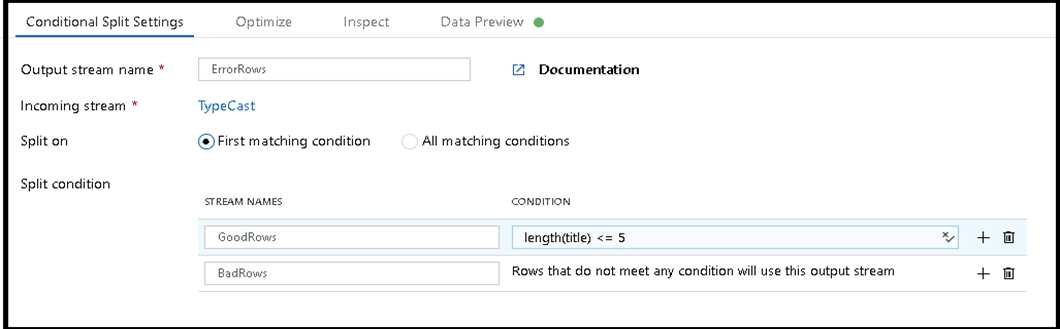
# Correct Answer: AB Section: (none) Explanation

**Explanation/Reference:**

Explanation:

B: Example:

1. This conditional split transformation defines the maximum length of "title" to be five. Any row that is less than or equal to five will go into the GoodRows stream. Any row that is larger than five will go into the BadRows stream.



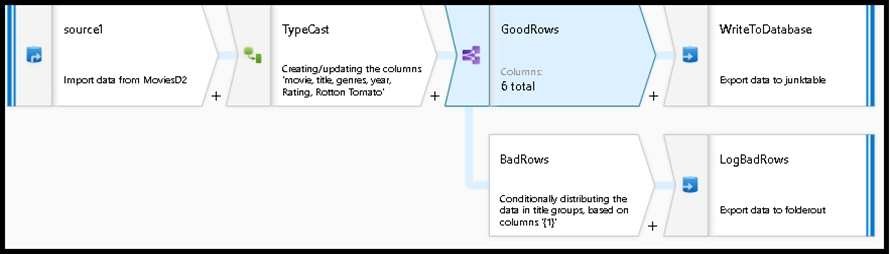
1. This conditional split transformation defines the maximum length of "title" to be five. Any row that is less than or equal to five will go into the GoodRows stream. Any row that is larger than five will go into the BadRows stream.

A:

1. Now we need to log the rows that failed. Add a sink transformation to the BadRows stream for logging. Here, we'll "auto-map" all of the fields so that we have logging of the complete transaction record. This is a text-delimited CSV file output to a single file in Blob Storage. We'll call the log file "badrows.csv".



1. The completed data flow is shown below. We are now able to split off error rows to avoid the SQL truncation errors and put those entries into a log file. Meanwhile, successful rows can continue to write to our target database.



Reference:

https://docs.microsoft.com/en-us/azure/data-factory/how-to-data-flow-error-rows

# QUESTION 2

You have an Azure Storage account and a data warehouse in Azure Synapse Analytics in the UK South region.

You need to copy blob data from the storage account to the data warehouse by using Azure Data Factory. The solution must meet the following requirements: Ensure that the data remains in the UK South region at all times.

Minimize administrative effort.

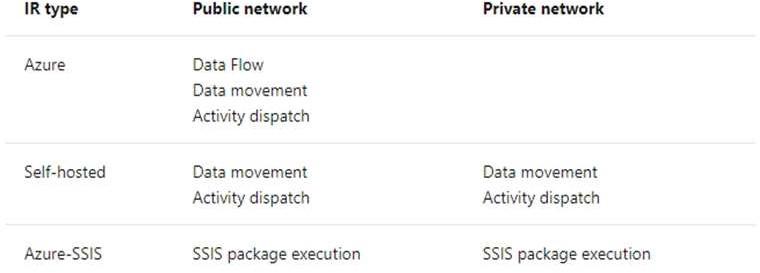
Which type of integration runtime should you use?

1. Azure integration runtime
2. Azure-SSIS integration runtime
3. Self-hosted integration runtime

# Correct Answer: A Section: (none) Explanation

**Explanation/Reference:**

Explanation:



Incorrect Answers:

C: Self-hosted integration runtime is to be used On-premises.

Reference:

https://docs.microsoft.com/en-us/azure/data-factory/concepts-integration-runtime

# QUESTION 3

You have an Azure Stream Analytics job that receives clickstream data from an Azure event hub.

You need to define a query in the Stream Analytics job. The query must meet the following requirements: Count the number of clicks within each 10-second window based on the country of a visitor.

Ensure that each click is **NOT** counted more than once.

How should you define the Query?

1. SELECT Country, Avg(\*) AS Average

FROM ClickStream TIMESTAMP BY CreatedAt GROUP BY Country, SlidingWindow(second, 10)

1. SELECT Country, Count(\*) AS Count

FROM ClickStream TIMESTAMP BY CreatedAt GROUP BY Country, TumblingWindow(second, 10)

1. SELECT Country, Avg(\*) AS Average

FROM ClickStream TIMESTAMP BY CreatedAt

GROUP BY Country, HoppingWindow(second, 10, 2)

1. SELECT Country, Count(\*) AS Count

FROM ClickStream TIMESTAMP BY CreatedAt

GROUP BY Country, SessionWindow(second, 5, 10)

# Correct Answer: B Section: (none) Explanation

**Explanation/Reference:**

Explanation:

Tumbling window functions are used to segment a data stream into distinct time segments and perform a function against them, such as the example below. The key differentiators of a Tumbling window are that they repeat, do not overlap, and an event cannot belong to more than one tumbling window.

Example:

Incorrect Answers:

A: Sliding windows, unlike Tumbling or Hopping windows, output events only for points in time when the content of the window actually changes. In other words, when an event enters or exits the window. Every window has at least one event, like in the case of Hopping windows, events can belong to more than one sliding window.

C: Hopping window functions hop forward in time by a fixed period. It may be easy to think of them as Tumbling windows that can overlap, so events can belong to more than one Hopping window result set. To make a Hopping window the same as a Tumbling window, specify the hop size to be the same as the window size.

D: Session windows group events that arrive at similar times, filtering out periods of time where there is no data.

Reference:

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

# QUESTION 4

You need to schedule an Azure Data Factory pipeline to execute when a new file arrives in an Azure Data Lake Storage Gen2 container. Which type of trigger should you use?

1. on-demand
2. tumbling window
3. schedule
4. event

# Correct Answer: D Section: (none) Explanation

**Explanation/Reference:**

Explanation:

Event-driven architecture (EDA) is a common data integration pattern that involves production, detection, consumption, and reaction to events. Data integration scenarios often require Data Factory 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.

Reference:

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

# QUESTION 5

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?

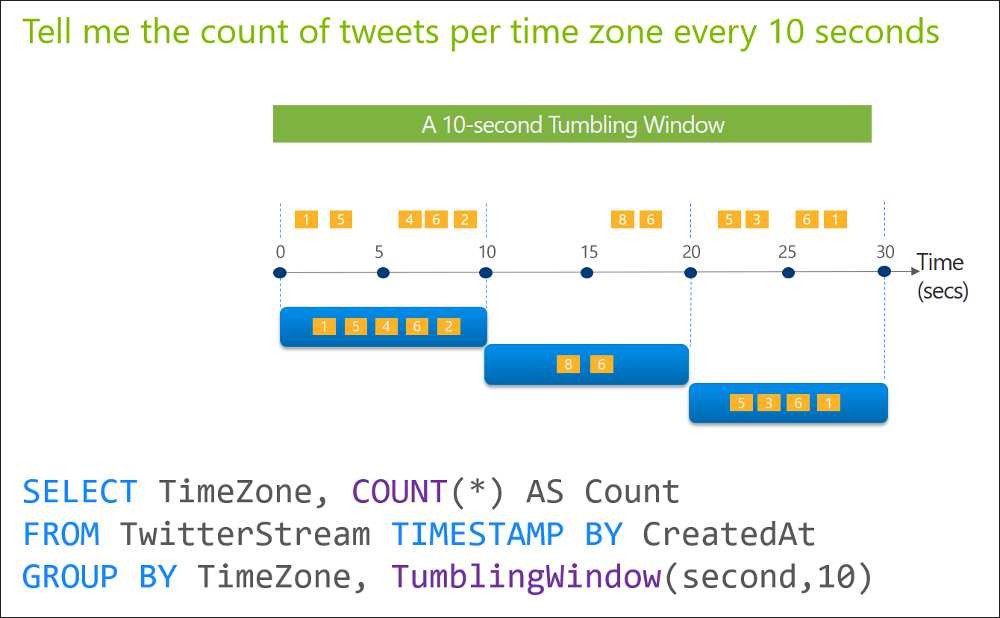
1. snapshot
2. tumbling
3. hopping
4. sliding

# Correct Answer: B Section: (none) Explanation

**Explanation/Reference:**

Explanation:

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



Reference:

https://docs.microsoft.com/en-us/stream-analytics-query/tumbling-window-azure-stream-analytics

# QUESTION 6

You are designing an Azure Databricks table. The table will ingest an average of 20 million streaming events per day.

You need to persist the events in the table for use in incremental load pipeline jobs in Azure Databricks. The solution must minimize storage costs and incremental load times.

What should you include in the solution?

1. Partition by DateTime fields.
2. Sink to Azure Queue storage.
3. Include a watermark column.
4. Use a JSON format for physical data storage.

# Correct Answer: B Section: (none) Explanation

**Explanation/Reference:**

Explanation:

The Databricks ABS-AQS connector uses Azure Queue Storage (AQS) to provide an optimized file source that lets you find new files written to an Azure Blob storage (ABS) container without repeatedly listing all of the files. This provides two major advantages:

Lower latency: no need to list nested directory structures on ABS, which is slow and resource intensive. Lower costs: no more costly LIST API requests made to ABS.

Reference:

https://docs.microsoft.com/en-us/azure/databricks/spark/latest/structured-streaming/aqs

# QUESTION 7

You have an Azure Databricks workspace named workspace1 in the Standard pricing tier.

You need to configure workspace1 to support autoscaling all-purpose clusters. The solution must meet the following requirements: Automatically scale down workers when the cluster is underutilized for three minutes.

Minimize the time it takes to scale to the maximum number of workers. Minimize costs.

What should you do first?

1. Enable container services for workspace1.
2. Upgrade workspace1 to the Premium pricing tier.
3. Set Cluster Mode to High Concurrency.
4. Create a cluster policy in workspace1.

# Correct Answer: B Section: (none) Explanation

**Explanation/Reference:**

Explanation:

For clusters running Databricks Runtime 6.4 and above, optimized autoscaling is used by all-purpose clusters in the Premium plan Optimized autoscaling:

Scales up from min to max in 2 steps.

Can scale down even if the cluster is not idle by looking at shuffle file state. Scales down based on a percentage of current nodes.

On job clusters, scales down if the cluster is underutilized over the last 40 seconds.

On all-purpose clusters, scales down if the cluster is underutilized over the last 150 seconds.

The spark.databricks.aggressiveWindowDownS Spark configuration property specifies in seconds how often a cluster makes down-scaling decisions. Increasing the value causes a cluster to scale down more slowly. The maximum value is 600.

Note: Standard autoscaling

Starts with adding 8 nodes. Thereafter, scales up exponentially, but can take many steps to reach the max. You can customize the first step by setting the spark.databricks.autoscaling.standardFirstStepUp Spark configuration property.

Scales down only when the cluster is completely idle and it has been underutilized for the last 10 minutes. Scales down exponentially, starting with 1 node.

Reference: https://docs.databricks.com/clusters/configure.html

# QUESTION 8

**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 are designing an Azure Stream Analytics solution that will analyze Twitter data.

You need to count the tweets in each 10-second window. The solution must ensure that each tweet is counted only once. Solution: You use a session window that uses a timeout size of 10 seconds.

Does this meet the goal?

1. Yes
2. No

# Correct Answer: B Section: (none) Explanation

**Explanation/Reference:**

Explanation:

Instead use a tumbling window. Tumbling windows are a series of fixed-sized, non-overlapping and contiguous time intervals.

Reference:

https://docs.microsoft.com/en-us/stream-analytics-query/tumbling-window-azure-stream-analytics

# QUESTION 9

**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 are designing an Azure Stream Analytics solution that will analyze Twitter data.

You need to count the tweets in each 10-second window. The solution must ensure that each tweet is counted only once. Solution: You use a hopping window that uses a hop size of 5 seconds and a window size 10 seconds.

Does this meet the goal?

1. Yes
2. No

# Correct Answer: B Section: (none) Explanation

**Explanation/Reference:**

Explanation:

Instead use a tumbling window. Tumbling windows are a series of fixed-sized, non-overlapping and contiguous time intervals.

Reference:

https://docs.microsoft.com/en-us/stream-analytics-query/tumbling-window-azure-stream-analytics

# QUESTION 10

**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?

1. Yes
2. No

# Correct Answer: B Section: (none) Explanation

**Explanation/Reference:**

Explanation:

Need a High Concurrency cluster for the jobs.

Standard clusters are recommended for a single user. Standard can run workloads developed in any language: Python, R, Scala, and SQL.

A high concurrency cluster is a managed cloud resource. The key benefits of high concurrency clusters are that they provide Apache Spark-native fine-grained sharing for maximum resource utilization and minimum query latencies.

Reference: https://docs.azuredatabricks.net/clusters/configure.html

# QUESTION 11

**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 Standard 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?

1. Yes
2. No

# Correct Answer: B Section: (none) Explanation

**Explanation/Reference:**

Explanation:

We would need a High Concurrency cluster for the jobs.

Note:

Standard clusters are recommended for a single user. Standard can run workloads developed in any language: Python, R, Scala, and SQL.

A high concurrency cluster is a managed cloud resource. The key benefits of high concurrency clusters are that they provide Apache Spark-native fine-grained sharing for maximum resource utilization and minimum query latencies.

Reference: https://docs.azuredatabricks.net/clusters/configure.html

# Design and implement data security Testlet 1

**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 1

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

1. a server-level virtual network rule
2. a database-level virtual network rule
3. a server-level firewall IP rule
4. a database-level firewall IP rule

# Correct Answer: A Section: (none) Explanation

**Explanation/Reference:**

Explanation:

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

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 server-level, not at the database-level.

Reference:

https://docs.microsoft.com/en-us/azure/sql-database/sql-database-vnet-service-endpoint-rule-overview

# Design and implement data security Question Set 2

**QUESTION 1**

You have an Azure Synapse Analytics dedicated SQL pool.

You need to ensure that data in the pool is encrypted at rest. The solution must **NOT** require modifying applications that query the data. What should you do?

1. Enable encryption at rest for the Azure Data Lake Storage Gen2 account.
2. Enable Transparent Data Encryption (TDE) for the pool.
3. Use a customer-managed key to enable double encryption for the Azure Synapse workspace.
4. Create an Azure key vault in the Azure subscription grant access to the pool.

# Correct Answer: B Section: (none) Explanation

**Explanation/Reference:**

Explanation:

Transparent Data Encryption (TDE) helps protect against the threat of malicious activity by encrypting and decrypting your data at rest. When you encrypt your database, associated backups and transaction log files are encrypted without requiring any changes to your applications. TDE encrypts the storage of an entire database by using a symmetric key called the database encryption key.

Reference:

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

# QUESTION 2

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

You need to minimize the time it takes to identify queries that return confidential information as defined by the company's data privacy regulations and the users who executed the queues.

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

**NOTE:** Each correct selection is worth one point.

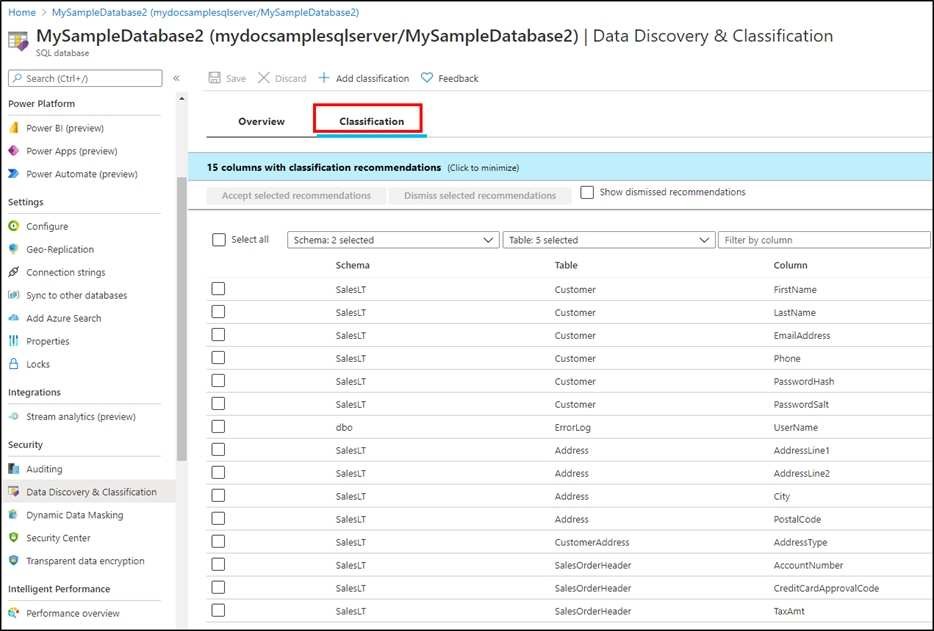
1. sensitivity-classification labels applied to columns that contain confidential information
2. resource tags for databases that contain confidential information
3. audit logs sent to a Log Analytics workspace
4. dynamic data masking for columns that contain confidential information

# Correct Answer: AC Section: (none) Explanation

**Explanation/Reference:**

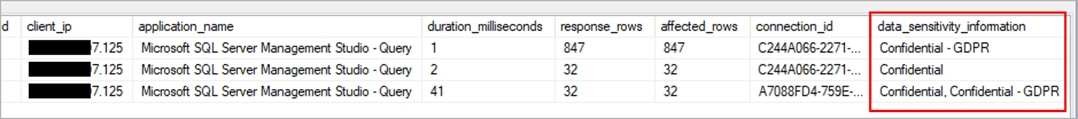
Explanation:

A: You can classify columns manually, as an alternative or in addition to the recommendation-based classification:



* 1. Select Add classification in the top menu of the pane.
  2. In the context window that opens, select the schema, table, and column that you want to classify, and the information type and sensitivity label.
  3. Select Add classification at the bottom of the context window.

C: An important aspect of the information-protection paradigm is the ability to monitor access to sensitive data. Azure SQL Auditing has been enhanced to include a new field in the audit log called data\_sensitivity\_information. This field logs the sensitivity classifications (labels) of the data that was returned by a query. Here's an example:



Reference:

https://docs.microsoft.com/en-us/azure/azure-sql/database/data-discovery-and-classification-overview

# QUESTION 3

You are designing an enterprise data warehouse in Azure Synapse Analytics that will contain a table named Customers. Customers will contain credit card information.

You need to recommend a solution to provide salespeople with the ability to view all the entries in Customers. The solution must prevent all the salespeople from viewing or inferring the credit card information.

What should you include in the recommendation?

* + 1. data masking
    2. Always Encrypted
    3. column-level security
    4. row-level security

# Correct Answer: A Section: (none) Explanation

**Explanation/Reference:**

Explanation:

SQL Database dynamic data masking limits sensitive data exposure by masking it to non-privileged users.

The Credit card masking method exposes the last four digits of the designated fields and adds a constant string as a prefix in the form of a credit card. Example: XXXX-XXXX-XXXX-1234

Reference:

https://docs.microsoft.com/en-us/azure/sql-database/sql-database-dynamic-data-masking-get-started

# Monitor and optimize data storage and data processing Testlet 1

**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 1

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

1. Deploy a High Concurrency Databricks cluster.
2. 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.
3. Set Data Lake Storage to use geo-redundant storage (GRS).
4. Deploy identical Azure Stream Analytics jobs to paired regions in Azure.

# Correct Answer: D Section: (none) Explanation

**Explanation/Reference:**

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:

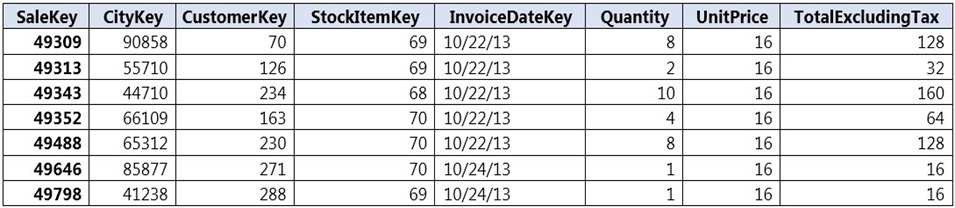
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-job-reliability

# Monitor and optimize data storage and data processing Question Set 2

**QUESTION 1**

You implement an enterprise data warehouse in Azure Synapse Analytics. You have a large fact table that is 10 terabytes (TB) in size.

Incoming queries use the primary key SaleKey column to retrieve data as displayed in the following table:



You need to distribute the large fact table across multiple nodes to optimize performance of the table. Which technology should you use?

1. hash distributed table with clustered index
2. hash distributed table with clustered Columnstore index
3. round robin distributed table with clustered index
4. round robin distributed table with clustered Columnstore index
5. heap table with distribution replicate

# Correct Answer: B Section: (none) Explanation

**Explanation/Reference:**

Explanation:

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

Columnstore indexes can achieve up to 100x better performance on analytics and data warehousing workloads and up to 10x better data compression than traditional rowstore indexes.

Incorrect Answers:

C, D: Round-robin tables are useful for improving loading speed.

Reference:

https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-tables-distribute https://docs.microsoft.com/en-us/sql/relational-databases/indexes/columnstore-indexes-query-performance

# QUESTION 2

You have an Azure Synapse Analytics dedicated SQL pool that contains a large fact table. The table contains 50 columns and 5 billion rows and is a heap. Most queries against the table aggregate values from approximately 100 million rows and return only two columns.

You discover that the queries against the fact table are very slow. Which type of index should you add to provide the fastest query times?

1. nonclustered columnstore
2. clustered columnstore
3. nonclustered
4. clustered

# Correct Answer: B Section: (none) Explanation

**Explanation/Reference:**

Explanation:

Clustered columnstore indexes are one of the most efficient ways you can store your data in dedicated SQL pool. Columnstore tables won't benefit a query unless the table has more than 60 million rows.

Reference:

https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/best-practices-dedicated-sql-pool

# QUESTION 3

You create an Azure Databricks cluster and specify an additional library to install. When you attempt to load the library to a notebook, the library in not found.

You need to identify the cause of the issue. What should you review?

1. notebook logs
2. cluster event logs
3. global init scripts logs
4. workspace logs

# Correct Answer: C Section: (none) Explanation

**Explanation/Reference:**

Explanation:

Cluster-scoped Init Scripts: Init scripts are shell scripts that run during the startup of each cluster node before the Spark driver or worker JVM starts. Databricks customers use init scripts for various purposes such as installing custom libraries, launching background processes, or applying enterprise security policies.

Logs for Cluster-scoped init scripts are now more consistent with Cluster Log Delivery and can be found in the same root folder as driver and executor logs for the cluster.

Reference:

https://databricks.com/blog/2018/08/30/introducing-cluster-scoped-init-scripts.html

# QUESTION 4

You have an Azure data factory.

You need to examine the pipeline failures from the last 60 days. What should you use?

1. the Activity log blade for the Data Factory resource
2. the Monitor & Manage app in Data Factory
3. the Resource health blade for the Data Factory resource
4. Azure Monitor

# Correct Answer: D Section: (none) Explanation

**Explanation/Reference:**

Explanation:

Data Factory stores pipeline-run data for only 45 days. Use Azure Monitor if you want to keep that data for a longer time.

Reference:

https://docs.microsoft.com/en-us/azure/data-factory/monitor-using-azure-monitor