DP-203 Q&As

Data Engineering on Microsoft Azure

Pass Microsoft DP-203 Exam with 100% Guarantee

Free Download Real Questions & Answers **PDF** and **VCE** file from:

https://www.leads4pass.com/dp-203.html

100% Passing Guarantee
100% Money Back Assurance

Following Questions and Answers are all new published by Microsoft Official Exam Center

QUESTION 1

You have an Azure Data Factory pipeline that is triggered hourly.

The pipeline has had 100% success for the past seven days.

The pipeline execution fails, and two retries that occur 15 minutes apart also fail. The third failure returns the following error.

ErrorCode=UserErrorFileNotFound,\\'Type=Microsoft.DataTransfer.Common.Shared.HybridDeliveryException,Message =ADLS Gen2 operation failed for: Operation returned an invalid statuscode \\'NotFound\\'. Account: \\'contosoproduksouth\\'. Filesystem: wwi. Path: \\'BIKES/CARBON/year=2021/month=01/day=10/hour=06\\'. ErrorCode: \\'PathNotFound\\'. Message: \\'The specified pathdoes not exist.\\'. RequestId: \\'6d269b78-901f-001b-4924-e7a7bc000000\\'. TimeStamp: \\'Sun, 10 Jan 2021 07:45:05

What is a possible cause of the error?

A. From 06.00 to 07:00 on January 10.2021 there was no data in w1/bikes/CARBON.

B. The parameter used to generate year.2021/month=0/day=10/hour=06 was incorrect

C. From 06:00 to 07:00 on January 10,2021 the file format of data wi/BiKES/CARBON was incorrect

D. The pipeline was triggered too early.

Correct Answer: B

QUESTION 2

You are designing an Azure Synapse Analytics workspace.

You need to recommend a solution to provide double encryption of all the data at rest.

Which two components should you include in the recommendation? Each coned answer presents part of the solution

NOTE: Each correct selection is worth one point.

A. an X509 certificate

B. an RSA key

C. an Azure key vault that has purge protection enabled

D. an Azure virtual network that has a network security group (NSG)

E. an Azure Policy initiative

Correct Answer: BE

Synapse workspaces encryption uses existing keys or new keys generated in Azure Key Vault. A single key is used to encrypt all the data in a workspace. Synapse workspaces support RSA 2048 and 3072 byte-sized keys, and RSA-HSM

keys.

The Key Vault itself needs to have purge protection enabled.

Reference: https://docs.microsoft.com/en-us/azure/synapse-analytics/security/workspaces-encryption

QUESTION 3

HOTSPOT

You are designing an Azure Data Lake Storage Gen2 container to store data for the human resources (HR) department and the operations department at your company. You have the following data access requirements:

1.

After initial processing, the HR department data will be retained for seven years.

2.

The operations department data will be accessed frequently for the first six months, and then accessed once per month. You need to design a data retention solution to meet the access requirements. The solution must minimize storage costs.

Hot Area:

Correct Answer:

QUESTION 4

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: ADF

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

QUESTION 5

You have an Azure Data Lake Storage Gen2 account named adls2 that is protected by a virtual network.

You are designing a SQL pool in Azure Synapse that will use adls2 as a source.

What should you use to authenticate to adls2?

A. an Azure Active Directory (Azure AD) user

B. a shared key

C. a shared access signature (SAS)

D. a managed identity

Correct Answer: D

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. You can use the Managed Identity capability to authenticate to any service that support Azure AD authentication.

Managed Identity authentication is required when your storage account is attached to a VNet.

Reference: https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/quickstart-bulk-load-copy-tsql-examples

QUESTION 6

You plan to perform batch processing in Azure Databricks once daily. Which type of Databricks cluster should you use?

A. High Concurrency

B. automated

C. interactive

Correct Answer: B

Azure Databricks has two types of clusters: interactive and automated. You use interactive clusters to analyze data collaboratively with interactive notebooks. You use automated clusters to run fast and robust automated jobs.

Example: Scheduled batch workloads (data engineers running ETL jobs)

This scenario involves running batch job JARs and notebooks on a regular cadence through the Databricks platform.

The suggested best practice is to launch a new cluster for each run of critical jobs. This helps avoid any issues (failures, missing SLA, and so on) due to an existing workload (noisy neighbor) on a shared cluster.

Reference:

https://docs.databricks.com/administration-guide/cloud-configurations/aws/cmbp.html#scenario-3-scheduled-batch-workloads-data-engineers-running-etl-jobs

QUESTION 7

You have an Azure Data Factory pipeline named pipeline1 that includes a Copy activity named Copy1. Copy1 has the following configurations:

1.

The source of Copy1 is a table in an on-premises Microsoft SQL Server instance that is accessed by using a linked service connected via a self-hosted integration runtime.

2.

The sink of Copy1 uses a table in an Azure SQL database that is accessed by using a linked service connected via an Azure integration runtime.

You need to maximize the amount of compute resources available to Copy1. The solution must minimize administrative effort.

What should you do?

A. Scale out the self-hosted integration runtime.

B. Scale up the data flow runtime of the Azure integration runtime and scale out the self-hosted integration runtime.

C. Scale up the data flow runtime of the Azure integration runtime.

Correct Answer: C

Scaling the Azure integration runtime is easy with the UI. Scaling the self-hosted integration runtime requires more effort.

Copy activity performance optimization features

Configuring performance features with UI

Note: Data Integration Units

A Data Integration Unit is a measure that represents the power (a combination of CPU, memory, and network resource allocation) of a single unit within the service. Data Integration Unit only applies to Azure integration runtime, but not self-

hosted integration runtime.

The allowed DIUs to empower a copy activity run is between 2 and 256. If not specified or you choose "Auto" on the UI, the service dynamically applies the optimal DIU setting based on your source-sink pair and data pattern.

Incorrect:

Not A, Not B:

Self-hosted integration runtime scalability

If you would like to achieve higher throughput, you can either scale up or scale out the Self-hosted IR:

*

If the CPU and available memory on the Self-hosted IR node are not fully utilized, but the execution of concurrent jobs is reaching the limit, you should scale up by increasing the number of concurrent jobs that can run on a node.

*

If on the other hand, the CPU is high on the Self-hosted IR node or available memory is low, you can add a new node to help scale out the load across the multiple nodes.

Reference: https://learn.microsoft.com/en-us/azure/data-factory/copy-activity-performance-features

---

QUESTION 8

You have an enterprise data warehouse in Azure Synapse Analytics named DW1 on a server named Server1.

You need to determine the size of the transaction log file for each distribution of DW1.

What should you do?

A. On DW1, execute a query against the sys.database_files dynamic management view.

B. From Azure Monitor in the Azure portal, execute a query against the logs of DW1.

C. Execute a query against the logs of DW1 by using the Get-AzOperationalInsightsSearchResult PowerShell cmdlet.

D. On the master database, execute a query against the sys.dm_pdw_nodes_os_performance_counters dynamic management view.

Correct Answer: A

For information about the current log file size, its maximum size, and the autogrow option for the file, you can also use the size, max_size, and growth columns for that log file in sys.database_files.

Reference: https://docs.microsoft.com/en-us/sql/relational-databases/logs/manage-the-size-of-the-transaction-log-file

---

QUESTION 9

HOTSPOT

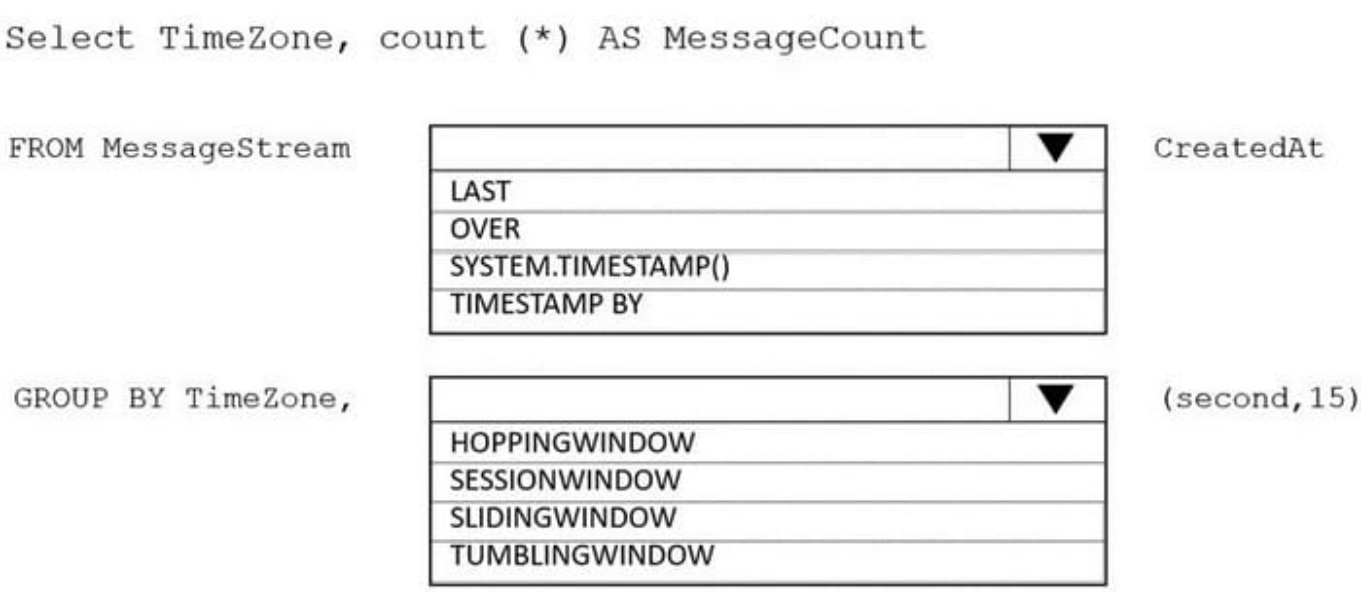
You are designing an Azure Stream Analytics solution that receives instant messaging data from an Azure event hub.

You need to ensure that the output from the Stream Analytics job counts the number of messages per time zone every 15 seconds.

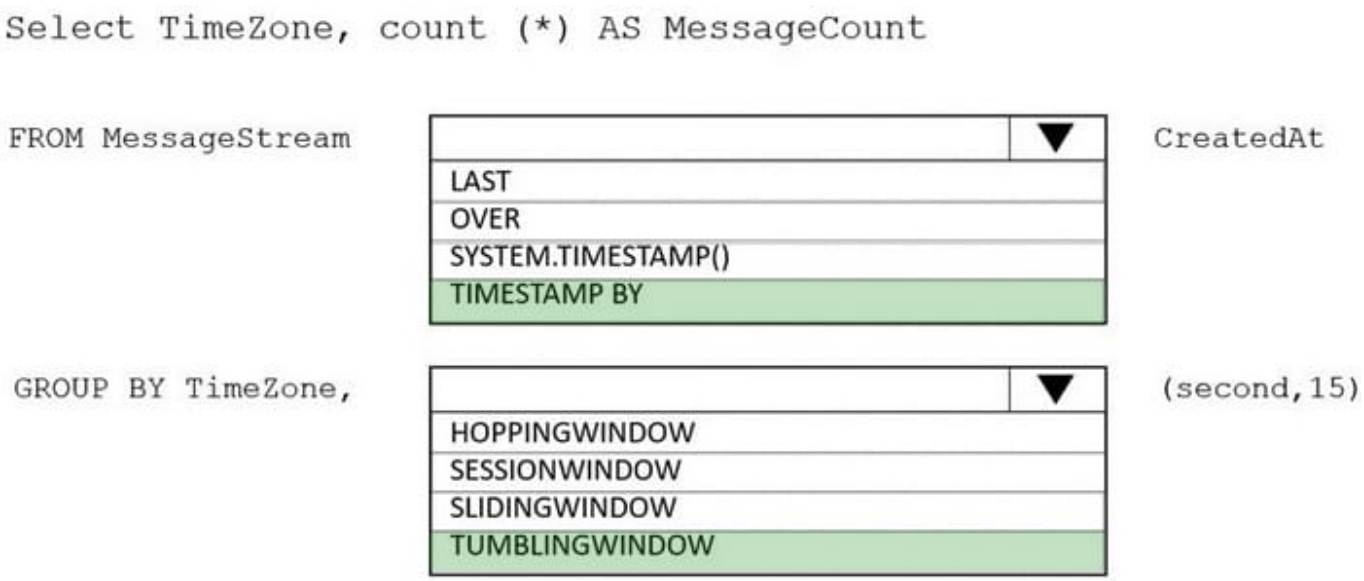
How should you complete the Stream Analytics query? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:



Correct Answer:

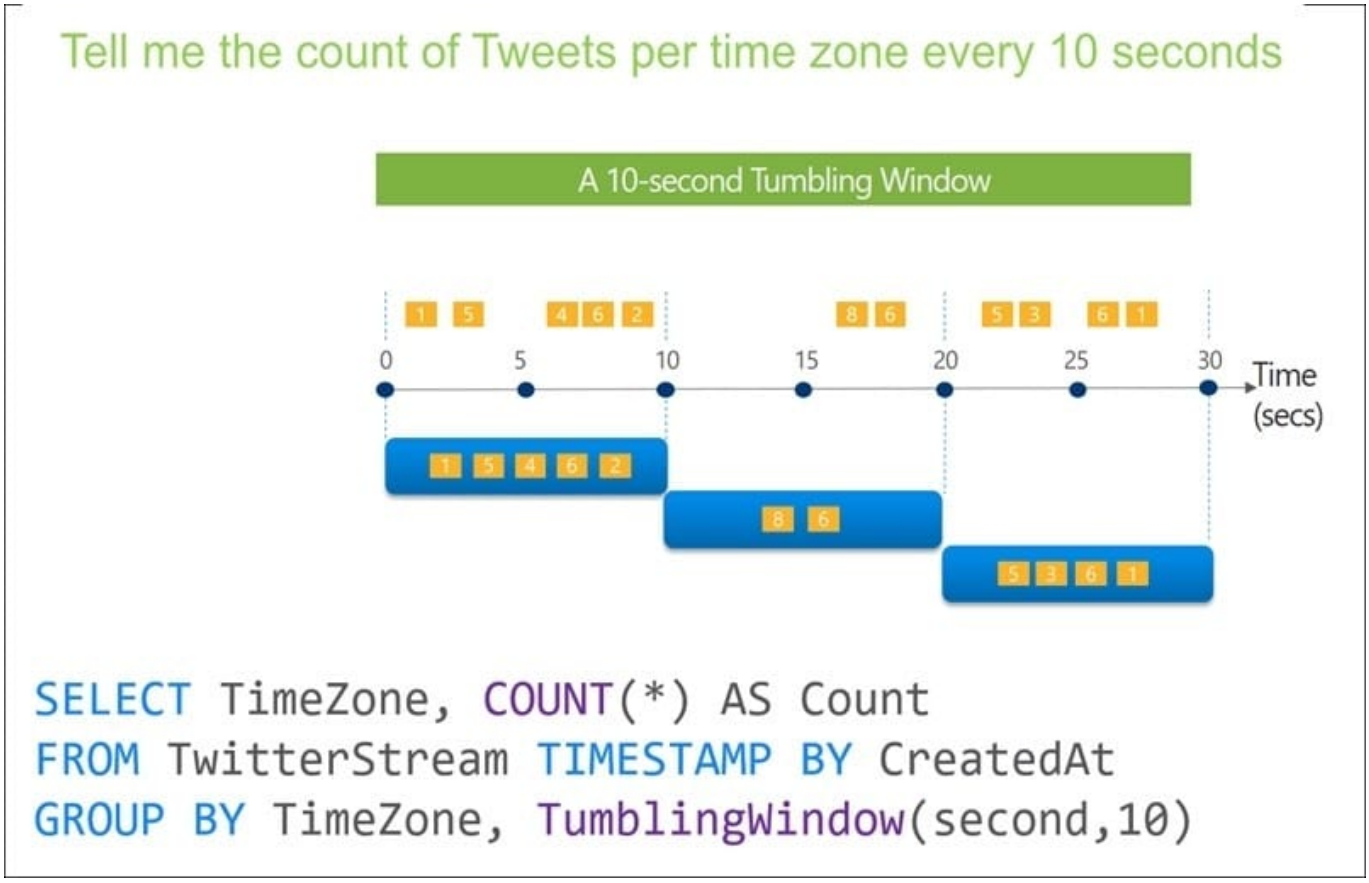


Box 1: timestamp by

Box 2: TUMBLINGWINDOW

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.



QUESTION 10

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

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

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

Which Apache Spark SQL operation should you use?

A. CREATE

B. UPDATE

C. MERGE

D. ALTER

Correct Answer: C

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

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

dimensional data,

SCD Type 2 can be expressed with the merge.

Example:

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

merge(

stagedUpdates.as("staged_updates"),

"customers.customerId = mergeKey")

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

"current" -> "false",

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

whenNotMatched()

insertExpr(Map(

"customerid" -> "staged_updates.customerId",

"address" -> "staged_updates.address",

"current" -> "true",

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

}

Reference:

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

---

QUESTION 11

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:

1.

Ensure that the data remains in the UK South region at all times.

2.

Minimize administrative effort.

Which type of integration runtime should you use?

A. Azure integration runtime

B. Azure-SSIS integration runtime

C. Self-hosted integration runtime

Correct Answer: A

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

QUESTION 12

HOTSPOT

You have an Azure subscription that contains an Azure Data Lake Storage account. The storage account contains a data lake named DataLake1.

You plan to use an Azure data factory to ingest data from a folder in DataLake1, transform the data, and land the data in another folder.

You need to ensure that the data factory can read and write data from any folder in the DataLake1 file system. The solution must meet the following requirements:

1.

Minimize the risk of unauthorized user access.

2.

Use the principle of least privilege.

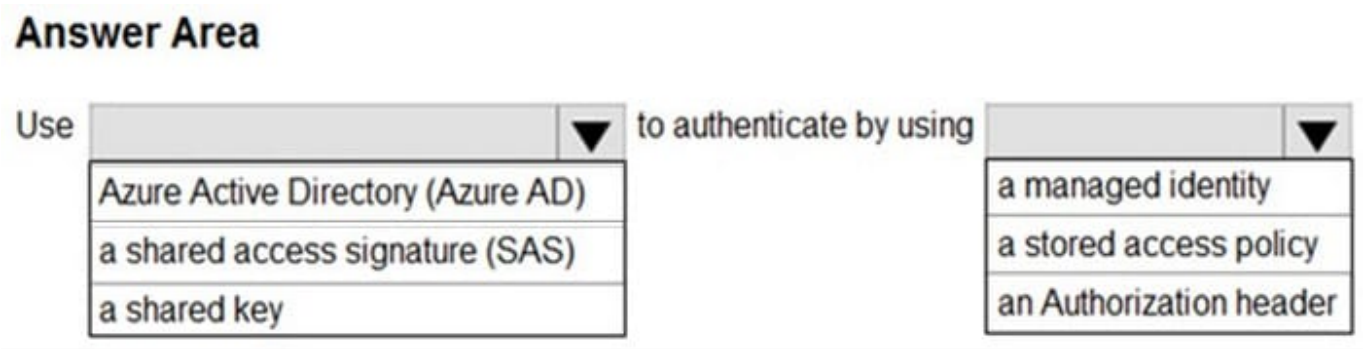
3.

Minimize maintenance effort.

How should you configure access to the storage account for the data factory? To answer, select the appropriate options in the answer area.

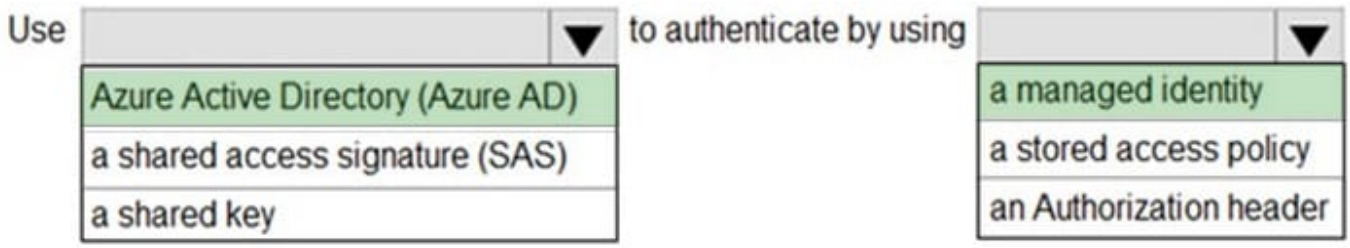
NOTE: Each correct selection is worth one point.

Hot Area:



Correct Answer:

Answer Area



Box 1: Azure Active Directory (Azure AD)

On Azure, managed identities eliminate the need for developers having to manage credentials by providing an identity for the Azure resource in Azure AD and using it to obtain Azure Active Directory (Azure AD) tokens.

Box 2: a managed identity

A data factory can be associated with a managed identity for Azure resources, which represents this specific data factory. You can directly use this managed identity for Data Lake Storage Gen2 authentication, similar to using your own

service principal. It allows this designated factory to access and copy data to or from your Data Lake Storage Gen2.

Note: The Azure Data Lake Storage Gen2 connector supports the following authentication types.

1.

Account key authentication

2.

Service principal authentication

3.

Managed identities for Azure resources authentication

Reference: https://docs.microsoft.com/en-us/azure/active-directory/managed-identities-azure-resources/overview
https://docs.microsoft.com/en-us/azure/data-factory/connector-azure-data-lake-storage

QUESTION 13

You have a data warehouse in Azure Synapse Analytics.

You need to ensure that the data in the data warehouse is encrypted at rest.

What should you enable?

A. Advanced Data Security for this database

B. Transparent Data Encryption (TDE)

C. Secure transfer required

D. Dynamic Data Masking

Correct Answer: B

Azure SQL Database currently supports encryption at rest for Microsoft-managed service side and client-side encryption scenarios.

1.

Support for server encryption is currently provided through the SQL feature called Transparent Data Encryption.

2.

Client-side encryption of Azure SQL Database data is supported through the Always Encrypted feature.

Reference: https://docs.microsoft.com/en-us/azure/security/fundamentals/encryption-atrest

QUESTION 14

You are designing a financial transactions table in an Azure Synapse Analytics dedicated SQL pool. The table will have a clustered columnstore index and will include the following columns:

1.

TransactionType: 40 million rows per transaction type

2.

CustomerSegment: 4 million per customer segment

3.

TransactionMonth: 65 million rows per month

4.

AccountType: 500 million per account type

You have the following query requirements:

1.

Analysts will most commonly analyze transactions for a given month.

2.

Transactions analysis will typically summarize transactions by transaction type, customer segment, and/or account type

You need to recommend a partition strategy for the table to minimize query times.

On which column should you recommend partitioning the table?

A. CustomerSegment

B. AccountType

C. TransactionType

D. TransactionMonth

Correct Answer: D

For optimal compression and performance of clustered columnstore tables, a minimum of 1 million rows per distribution and partition is needed. Before partitions are created, dedicated SQL pool already divides each table into 60 distributed databases.

Example: Any partitioning added to a table is in addition to the distributions created behind the scenes. Using this example, if the sales fact table contained 36 monthly partitions, and given that a dedicated SQL pool has 60 distributions, then the sales fact table should contain 60 million rows per month, or 2.1 billion rows when all months are populated. If a table contains fewer than the recommended minimum number of rows per partition, consider using fewer partitions in order to increase the number of rows per partition.

QUESTION 15

You have an Azure Synapse Analytics job that uses Scala.

You need to view the status of the job.

What should you do?

A. From Azure Monitor, run a Kusto query against the AzureDiagnostics table.

B. From Azure Monitor, run a Kusto query against the SparkLogying1 Event.CL table.

C. From Synapse Studio, select the workspace. From Monitor, select Apache Sparks applications.

D. From Synapse Studio, select the workspace. From Monitor, select SQL requests.

Correct Answer: C

Use Synapse Studio to monitor your Apache Spark applications. To monitor running Apache Spark application Open Monitor, then select Apache Spark applications. To view the details about the Apache Spark applications that are running, select the submitting Apache Spark application and view the details. If the Apache Spark application is still running, you can monitor the progress.

Reference: https://docs.microsoft.com/en-us/azure/synapse-analytics/monitoring/apache-spark-applications