Data Storage]

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?

A. File2.csv and File3.csv only

B. File1.csv and File4.csv only

C. File1.csv, File2.csv, File3.csv, and File4.csv

D. File1.csv only

Correct Answer: C

[Data Storage]

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

Correct Answer: D

[Data Storage]

HOTSPOT

You are planning the deployment of Azure Data Lake Storage Gen2.

You have the following two reports that will access the data lake:

Report1: Reads three columns from a file that contains 50 columns.

Report2: Queries a single record based on a timestamp.

You need to recommend in which format to store the data in the data lake to support the reports. The solution must minimize read times.

What should you recommend for each report? To answer, select the appropriate options in the answer area

Table

Description automatically generated

Correct Answer

Report1: CSV

CSV: The destination writes records as delimited data.

Report2: AVRO

AVRO supports timestamps.

Not Parquet, TSV: Not options for Azure Data Lake Storage Gen2.

[Data Storage]

HOTSPOT

You need to output files from Azure Data Factory.

Which file format should you use for each type of output? To answer, select the appropriate options in the answer area.

Correct Answer

Table

Description automatically generated

Box 1: Parquet

Parquet stores data in columns, while Avro stores data in a row-based format. By their very nature, column-oriented data stores are optimized for read-heavy analytical workloads, while row-based databases are best for write-heavy

transactional workloads.

Box 2: Avro

An Avro schema is created using JSON format.

AVRO supports timestamps.

[Data Storage]

HOTSPOT

You use Azure Data Factory to prepare data to be queried by Azure Synapse Analytics serverless SQL pools.

Files are initially ingested into an Azure Data Lake Storage Gen2 account as 10 small JSON files. Each file contains the same data attributes and data from a subsidiary of your company.

You need to move the files to a different folder and transform the data to meet the following requirements:

Provide the fastest possible query times.

Automatically infer the schema from the underlying files.

How should you configure the Data Factory copy activity?

Table

Description automatically generatedCorrect Answer:

Box 1: Preserve hierarchy

Compared to the flat namespace on Blob storage, the hierarchical namespace greatly improves the performance of directory management operations, which improves overall job performance.

Box 2: Parquet

Azure Data Factory parquet format is supported for Azure Data Lake Storage Gen2.

Parquet supports the schema property.

[Data Storage]

HOTSPOT

You have an Azure Data Lake Storage Gen2 container.

Data is ingested into the container, and then transformed by a data integration application. The data is NOT modified after that. Users can read files in the container but cannot modify the files.

You need to design a data archiving solution that meets the following requirements:

New data is accessed frequently and must be available as quickly as possible.

Data that is older than five years is accessed infrequently but must be available within one second when requested.

Data that is older than seven years is NOT accessed. After seven years, the data must be persisted at the lowest cost possible.

Costs must be minimized while maintaining the required availability.

How should you manage the data?

Table

Description automatically generated

Correct Answer:

Move to cool storage

Move to archive storage

[Data Storage]

HOTSPOT

You have two Azure Storage accounts named Storage1 and Storage2. Each account holds one container and has the hierarchical namespace enabled. The system has files that contain data stored in the Apache Parquet format.

You need to copy folders and files from Storage1 to Storage2 by using a Data Factory copy activity. The solution must meet the following requirements:

- No transformations must be performed.

- The original folder structure must be retained.

- Minimize time required to perform the copy activity.

How should you configure the copy activity?

Graphical user interface, application, table

Description automatically generated

Box 1: Parquet

For Parquet datasets, the type property of the copy activity source must be set to Parquet Source.

Box 2: Preserve Hierarchy

Preserve Hierarchy (default): Preserves the file hierarchy in the target folder. The relative path of the source file to the source folder is identical to the relative path of the target file to the target folder.

[Data Storage]

You have an Azure Data Lake Storage Gen2 container that contains 100 TB of data.

You need to ensure that the data in the container is available for read workloads in a secondary region if an outage occurs in the primary region. The solution must minimize costs.

Which type of data redundancy should you use?

A. geo-redundant storage (GRS)

B. read-access geo-redundant storage (RA-GRS)

C. zone-redundant storage (ZRS)

D. locally-redundant storage (LRS)

Correct Answer: B

[Data Storage]

You plan to implement an Azure Data Lake Gen 2 storage account.

You need to ensure that the data lake will remain available if a data center fails in the primary Azure region. The solution must minimize costs.

Which type of replication should you use for the storage account?

A. geo-redundant storage (GRS)

B. geo-zone-redundant storage (GZRS)

C. locally-redundant storage (LRS)

D. zone-redundant storage (ZRS)

Correct Answer: C

[Data Storage]

You need to design an Azure Synapse Analytics dedicated SQL pool that meets the following requirements:

Can return an employee record from a given point in time.

Maintains the latest employee information.

Minimizes query complexity.

How should you model the employee data?

A. as a temporal table

B. as a SQL graph table

C. as a degenerate dimension table

D. as a Type 2 slowly changing dimension (SCD) table

Correct Answer: D

[Data Storage]

You have a SQL pool in Azure Synapse.

You plan to load data from Azure Blob storage to a staging table. Approximately 1 million rows of data will be loaded daily. The table will be truncated before each daily load.

You need to create the staging table. The solution must minimize how long it takes to load the data to the staging table.

Graphical user interface, application

Description automatically generatedHow should you configure the table?

Correct Answer:

Box 1: Hash

Hash-distributed tables improve query performance on large fact tables. They can have very large numbers of rows and still achieve high performance.

Box 2: Clustered columnstore

When creating partitions on clustered columnstore tables, it is important to consider how many rows belong to each partition. For optimal compression and performance of clustered columnstore tables, a minimum of 1 million rows per

distribution and partition is needed.

Box 3: Date

Table partitions enable you to divide your data into smaller groups of data. In most cases, table partitions are created on a date column.

Partition switching can be used to quickly remove or replace a section of a table.

PART 2

[Data Storage]

HOTSPOT

A company plans to use Platform-as-a-Service (PaaS) to create the new data pipeline process. The process must meet the following requirements:

Ingest:

Access multiple data sources.

Provide the ability to orchestrate workflow.

Provide the capability to run SQL Server Integration Services packages.

Store:

Optimize storage for big data workloads.

Provide encryption of data at rest.

Operate with no size limits.

Prepare and Train:

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

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

Provide seamless user authentication with Azure Active Directory.

Model & Serve:

Implement native columnar storage.

Support for the SQL language

Provide support for structured streaming.

You need to build the data integration pipeline.

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

Table

Description automatically generated

Correct Answer:

Ingest: Azure Data Factory

Azure Data Factory pipelines can execute SSIS packages.

In Azure, the following services and tools will meet the core requirements for pipeline orchestration, control flow, and data movement: Azure Data Factory, Oozie on HDInsight, and SQL Server Integration Services (SSIS).

Store: Data Lake Storage

Data Lake Storage Gen1 provides unlimited storage.

Note: Data at rest includes information that resides in persistent storage on physical media, in any digital format. Microsoft Azure offers a variety of data storage solutions to meet different needs, including file, disk, blob, and table storage.

Microsoft also provides encryption to protect Azure SQL Database, Azure Cosmos DB, and Azure Data Lake.

Prepare and Train: Azure Databricks

Azure Databricks provides enterprise-grade Azure security, including Azure Active Directory integration.

With Azure Databricks, you can set up your Apache Spark environment in minutes, autoscale and collaborate on shared projects in an interactive workspace. Azure Databricks supports Python, Scala, R, Java and SQL, as well as data

science frameworks and libraries including TensorFlow, PyTorch and scikit-learn.

Model and Serve: Azure Synapse Analytics