DUMPSARENA

Data Engineering on Microsoft Azure

Microsoft DP-203

Version Demo

Total Demo Questions: 10

Total Premium Questions: 175

Buy Premium PDF

https://dumpsarena.com

sales@dumpsarena.com

Topic Break Down

Topic	No. of Questions
Topic 1, Case Study 1	6
Topic 2, Case Study 2	2
Topic 3, Mixed Questions	167
Total	175

QUESTION NO: 1

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 shared access signature (SAS) as the credentials for the data load process.

F. Create a managed identity.

ANSWER: A D F

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

QUESTION NO: 2

You develop data engineering solutions for a company.

A project requires the deployment of data to Azure Data Lake Storage.

You need to implement role-based access control (RBAC) so that project members can manage the Azure Data Lake Storage resources.

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

NOTE: Each correct selection is worth one point.

A. Create security groups in Azure Active Directory (Azure AD) and add project members.

B. Configure end-user authentication for the Azure Data Lake Storage account.

C. Assign Azure AD security groups to Azure Data Lake Storage.

D. Configure Service-to-service authentication for the Azure Data Lake Storage account.

E. Configure access control lists (ACL) for the Azure Data Lake Storage account.

ANSWER: A C E

Explanation:

AC: Create security groups in Azure Active Directory. Assign users or security groups to Data Lake Storage Gen1 accounts.
E: Assign users or security groups as ACLs to the Data Lake Storage Gen1 file system

Reference: https://docs.microsoft.com/en-us/azure/data-lake-store/data-lake-store-secure-data

QUESTION NO: 3

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?

A. Partition by DateTime fields.

B. Sink to Azure Queue storage.

C. Include a watermark column.

D. Use a JSON format for physical data storage.

ANSWER: B

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 NO: 4 - (DRAG DROP)

DRAG DROP

You have an Azure Data Lake Storage Gen2 account that contains a JSON file for customers. The file contains two attributes named FirstName and LastName.

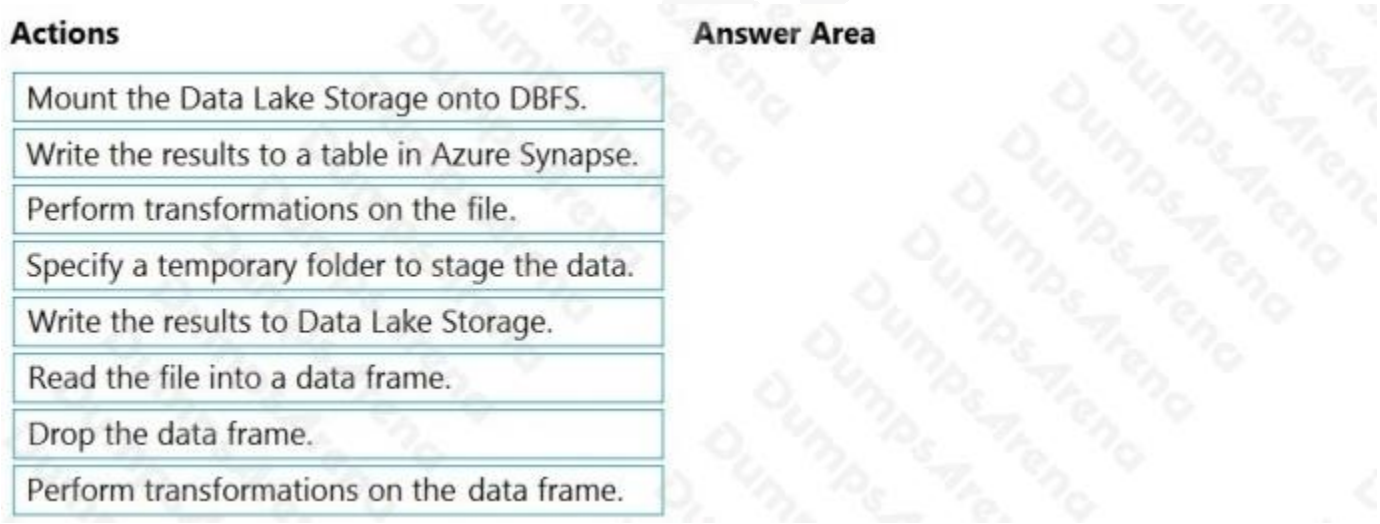
You need to copy the data from the JSON file to an Azure Synapse Analytics table by using Azure Databricks. A new column must be created that concatenates the FirstName and LastName values.

You create the following components:

- A destination table in Azure Synapse
- An Azure Blob storage container
- A service principal

Which five actions should you perform in sequence next in is Databricks notebook? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

Select and Place:



ANSWER:

Actions	Answer Area
Mount the Data Lake Storage onto DBFS.	Read the file into a data frame.
Write the results to a table in Azure Synapse.	Perform transformations on the file.
Perform transformations on the file.	Specify a temporary folder to stage the data.
Specify a temporary folder to stage the data.	Write the results to Data Lake Storage.
Write the results to Data Lake Storage.	Drop the data frame.
Read the file into a data frame.	
Drop the data frame.	
Perform transformations on the data frame.	

Explanation:

Step 1: Read the file into a data frame.

You can load the json files as a data frame in Azure Databricks.

Step 2: Perform transformations on the data frame.

Step 3:Specify a temporary folder to stage the data

Specify a temporary folder to use while moving data between Azure Databricks and Azure Synapse.

Step 4: Write the results to a table in Azure Synapse.

You upload the transformed data frame into Azure Synapse. You use the Azure Synapse connector for Azure Databricks to directly upload a dataframe as a table in a Azure Synapse.

Step 5: Drop the data frame

Clean up resources. You can terminate the cluster. From the Azure Databricks workspace, select Clusters on the left. For the cluster to terminate, under Actions, point to the ellipsis (...) and select the Terminate icon.

Reference: https://docs.microsoft.com/en-us/azure/azure-databricks/databricks-extract-load-sql-data-warehouse

QUESTION NO: 5 - (HOTSPOT)

HOTSPOT

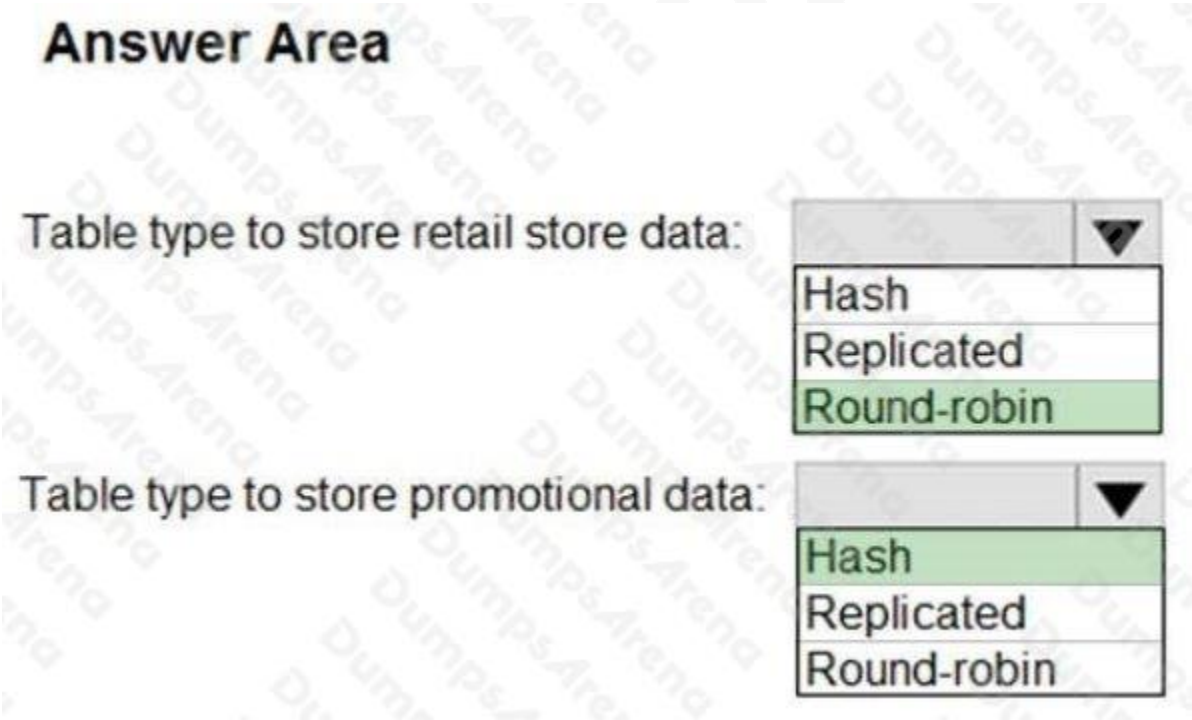
You need to design an analytical storage solution for the transactional data. The solution must meet the sales transaction dataset requirements.

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.

Hot Area:

ANSWER:



Explanation:

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.

Scenario:

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

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

QUESTION NO: 6 - (HOTSPOT)

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.

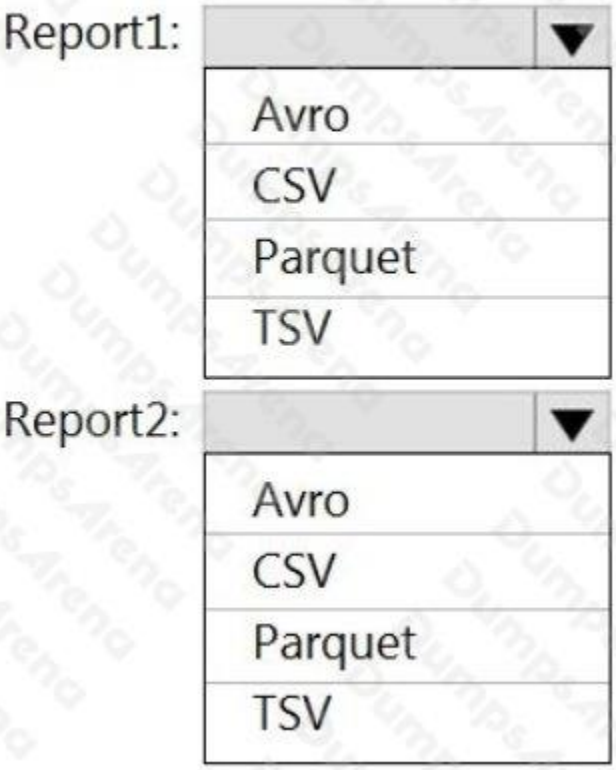
You need to recommend in which format to store the data in the data lake to support the reports. The solution must minimize read times.

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

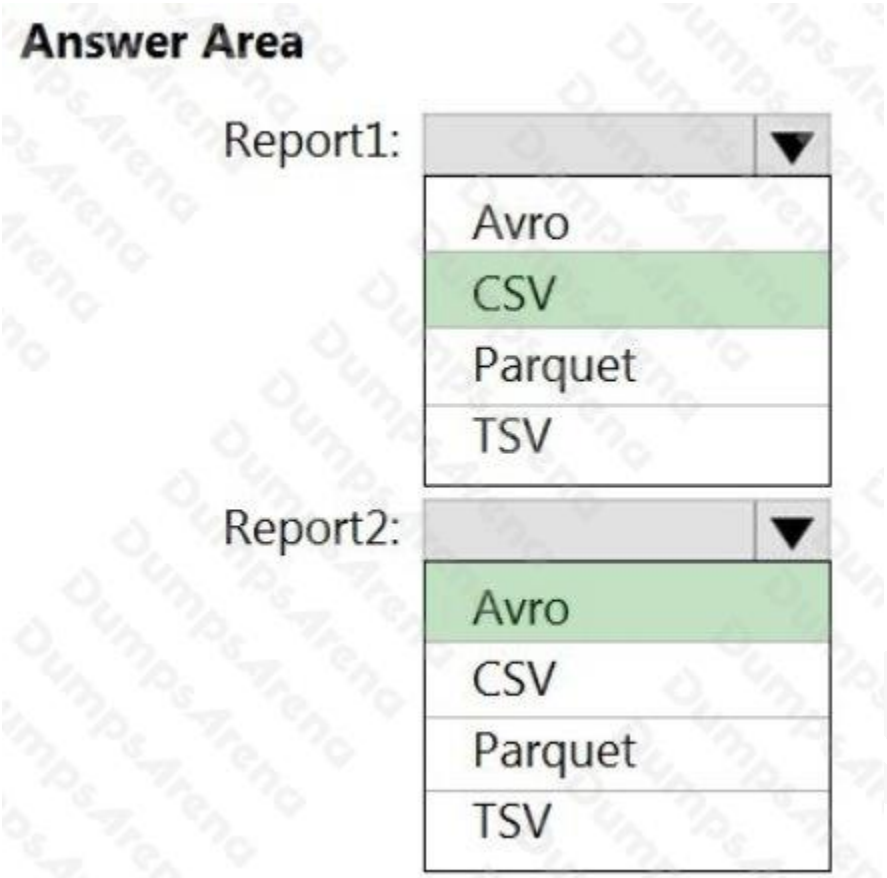
NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area



ANSWER:



Explanation:

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.

Reference: https://streamsets.com/documentation/datacollector/latest/help/datacollector/UserGuide/Destinations/ADLS-G2-D.html

QUESTION NO: 7

You plan to ingest streaming social media data by using Azure Stream Analytics. The data will be stored in files in Azure Data Lake Storage, and then consumed by using Azure Databricks and PolyBase in Azure Synapse Analytics.

You need to recommend a Stream Analytics data output format to ensure that the queries from Databricks and PolyBase against the files encounter the fewest possible errors. The solution must ensure that the files can be queried quickly and that the data type information is retained.

What should you recommend?

A. JSON

B. Parquet

C. CSV

D. Avro

ANSWER: B

Explanation:

Need Parquet to support both Databricks and PolyBase.

Reference:

https://docs.microsoft.com/en-us/sql/t-sql/statements/create-external-file-format-transact-sql

QUESTION NO: 8

You are designing a data mart for the human resources (HR) department at your company. The data mart will contain employee information and employee transactions.

From a source system, you have a flat extract that has the following fields:

- EmployeeID
- FirstName
- LastName
- Recipient
- GrossAmount
- TransactionID
- GovernmentID
- NetAmountPaid
- TransactionDate

You need to design a star schema data model in an Azure Synapse Analytics dedicated SQL pool for the data mart.

Which two tables should you create? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

A. a dimension table for Transaction

B. a dimension table for EmployeeTransaction

C. a dimension table for Employee

D. a fact table for Employee

E. a fact table for Transaction

ANSWER: C E

Explanation:

C: Dimension tables contain attribute data that might change but usually changes infrequently. For example, a customer's name and address are stored in a dimension table and updated only when the customer's profile changes. To minimize the size of a large fact table, the customer's name and address don't need to be in every row of a fact table. Instead, the fact table and the dimension table can share a customer ID. A query can join the two tables to associate a customer's profile and transactions.

E: Fact tables contain quantitative data that are commonly generated in a transactional system, and then loaded into the dedicated SQL pool. For example, a retail business generates sales transactions every day, and then loads the data into a dedicated SQL pool fact table for analysis.

Reference:

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

QUESTION NO: 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 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 SQL.
- 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 duster.
- The job duster 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 Standard cluster for the data engineers, and a High Concurrency cluster for the jobs.

Does this meet the goal?

A. Yes

B. No

ANSWER: B

Explanation:

We need a High Concurrency cluster for the data engineers and 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

QUESTION NO: 10 - (DRAG DROP)

DRAG DROP

You have an Azure Synapse Analytics workspace named WS1.

You have an Azure Data Lake Storage Gen2 container that contains JSON-formatted files in the following format.

```
{
    "id": "66532691-ab20-11ea-8b1d-936b3ec64e54",
    "context": {
        "data": {
            "eventTime": "2020-06-10T13:43:34.553Z",
            "samplingRate": "100.0",
            "isSynthetic": "false"
        },
        "session": {
            "isFirst": "false",
            "id": "38619c14-7a23-4687-8268-95862c5326b1"
        },
        "custom": {
            "dimensions": [
                {
                    "customerInfo": {
                        "ProfileType": "ExpertUser",
                        "RoomName": "",
                        "CustomerName": "diamond",
                        "UserName": "XXXX@yahoo.com"
                    }
                },
                {
                    "customerInfo" {
                        "ProfileType": "Novice",
                        "RoomName": "",
                        "CustomerName": "topaz",
                        "UserName": "XXXX@outlook.com"
                    }
                }
            ]
        }
    }
}
```

You need to use the serverless SQL pool in WS1 to read the files.

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.

Select and Place:

Values

- opendatasource
- openjson
- openquery
- openrowset

Answer Area

```
select*

FROM
[                ] (
    BULK 'https://contoso.blob.core.windows.net/contosodw',
    FORMAT= 'CSV',
    fieldterminator = '0x0b',
    fieldquote = '0x0b',
    rowterminator = '0x0b'
)
with (id varchar(50),
    contextdateventTime varchar(50) '$.context.data.eventTime',
    contextdatasamplingRate varchar(50) '$.context.data.samplingRate',
    contextdataisSynthetic varchar(50) '$.context.data.isSynthetic'.
    contextsessionisFirst varchar(50) '$.context.session.isFirst',
    contextsession varchar(50) '$.context.session.id',
    contextcustomdimensions varchar(max) '$.context.custom.dimensions'

) as q
cross apply [                ] (contextcustomdimensions)

with ( ProfileType varchar(50) '$.customerInfo.ProfileType',
        RoomName varchar(50) '$.customerInfo.RoomName',
        CustomerName varchar(50) '$.customerInfo.CustomerName',
        UserName varchar(50) '$.customerInfo.UserName'
    )
```

ANSWER:

Values

- opendatasource
- openquery

Answer Area

```
select*

FROM
[   openrowset   ] (
    BULK 'https://contoso.blob.core.windows.net/contosodw',
    FORMAT= 'CSV',
    fieldterminator = '0x0b',
    fieldquote = '0x0b',
    rowterminator = '0x0b'
)
with (id varchar(50),
    contextdateventTime varchar(50) '$.context.data.eventTime',
    contextdatasamplingRate varchar(50) '$.context.data.samplingRate',
    contextdataisSynthetic varchar(50) '$.context.data.isSynthetic'.
    contextsessionisFirst varchar(50) '$.context.session.isFirst',
    contextsession varchar(50) '$.context.session.id',
    contextcustomdimensions varchar(max) '$.context.custom.dimensions'

) as q
cross apply [   openjson   ] (contextcustomdimensions)

with ( ProfileType varchar(50) '$.customerInfo.ProfileType',
        RoomName varchar(50) '$.customerInfo.RoomName',
        CustomerName varchar(50) '$.customerInfo.CustomerName',
        UserName varchar(50) '$.customerInfo.UserName'
    )
```

Explanation:

Box 1: openrowset

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

Example:

SELECT *

FROM OPENROWSET(

BULK 'csv/population/population.csv',

DATA_SOURCE = 'SqlOnDemandDemo',

FORMAT = 'CSV', PARSER_VERSION = '2.0',

FIELDTERMINATOR =',',

ROWTERMINATOR = '\n'

Box 2: openjson

You can access your JSON files from the Azure File Storage share by using the mapped drive, as shown in the following example:

SELECT book.* FROM

OPENROWSET(BULK N't:\books\books.json', SINGLE_CLOB) AS json

CROSS APPLY OPENJSON(BulkColumn)

WITH(id nvarchar(100), name nvarchar(100), price float, pages_i int, author nvarchar(100)) AS book

Reference: https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/query-single-csv-file https://docs.microsoft.com/en-us/sql/relational-databases/json/import-json-documents-into-sql-server