Data Engineering on Microsoft Azure

1. HOTSPOT (Drag and Drop is not supported)

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

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Table type to store the product sales transactions:

- Hash
- Round-robin
- Replicated

When creating the table for sales transactions:

- Configure a clustered index.
- Set the distribution column to produd
- Set the distribution column to the sal

A. See Explanation section for answer.

Answer(s): A

---

2. DRAG DROP (Drag and Drop is not supported)

You need to ensure that the Twitter feed data can be analyzed in the dedicated SQL pool. The solution must meet the customer sentiment analytic requirements.

Which three Transact-SQL DDL commands should you run in sequence? To answer, move the appropriate commands from the list of commands to the answer area and arrange them in the correct order.

NOTE: More than one order of answer choices is correct. You will receive credit for any of the correct orders you select.

Select and Place:

Commands

```
CREATE EXTERNAL DATA SOURCE

CREATE EXTERNAL FILE FORMAT

CREATE EXTERNAL TABLE

CREATE EXTERNAL TABLE AS SELECT

CREATE DATABASE SCOPED CREDENTIAL
```

Answer Area

A. See Explanation section for answer.

Answer(s): A

---

3. HOTSPOT (Drag and Drop is not supported)
You need to design the partitions for the product sales transactions. The solution must meet the sales transaction dataset requirements. What should you include in the solution? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
Hot Area:

Answer Area

Partition product sales transactions data by:

Sales date
Product ID
Promotion ID

Store product sales transactions data in:

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

A. See Explanation section for answer.

Answer(s): A

---

4. You need to implement the surrogate key for the retail store table. The solution must meet the sales transaction dataset requirements. What should you create?

A. a table that has an IDENTITY property

B. a system-versioned temporal table

C. a user-defined SEQUENCE object

D. a table that has a FOREIGN KEY constraint

Answer(s): A

---

5. HOTSPOT (Drag and Drop is not supported)
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 Area

Table type to store retail store data:

▼
Hash
Replicated
Round-robin

Table type to store promotional data:

▼
Hash
Replicated
Round-robin

A. See Explanation section for answer.

Answer(s): A

---

6. HOTSPOT (Drag and Drop is not supported)
You need to implement an Azure Synapse Analytics database object for storing the sales transactions data.
The solution must meet the sales transaction dataset requirements.
What should you do? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
Hot Area:

Transact-SQL DDL command to use:

▼
CREATE EXTERNAL TABLE
CREATE TABLE
CREATE VIEW

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

▼
FORMAT_OPTIONS
FORMAT_TYPE
RANGE LEFT FOR VALUES
RANGE RIGHT FOR VALUES

A. See Explanation section for answer.

Answer(s): A

---

7. You need to design a data retention solution for the Twitter feed data records. The solution must meet the customer sentiment analytics requirements.
Which Azure Storage functionality should you include in the solution?

A. change feed

B. soft delete

C. time-based retention

D. lifecycle management

Answer(s): D

---

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

```
CREATE TABLE [dbo].[DimEmployee](
    [EmployeeKey] [int] IDENTITY(1,1) NOT NULL,
    [EmployeeID] [int] NOT NULL,
    [FirstName] [varchar](100) NOT NULL,
    [LastName] [varchar](100) NOT NULL,
    [JobTitle] [varchar](100) NULL,
    [LastHireDate] [date] NULL,
    [StreetAddress] [varchar](500) NOT NULL,
    [City] [varchar](200) NOT NULL,
    [StateProvince] [varchar](50) NOT NULL,
    [Portalcode] [varchar](10) NOT NULL
    )
```

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?

A. [ManagerEmployeeID] [int] NULL

B. [ManagerEmployeeID] [smallint] NULL

C. [ManagerEmployeeKey] [int] NULL

D. [ManagerName] [varchar](200) NULL

Answer(s): C

---

9. You have an Azure Synapse workspace named MyWorkspace that contains an Apache Spark database named mytestdb. You run the following command in an Azure Synapse Analytics Spark pool in MyWorkspace.

CREATE TABLE mytestdb.myParquetTable(

EmployeeID int,

EmployeeName string,

EmployeeStartDate date)

USING Parquet

You then use Spark to insert a row into mytestdb.myParquetTable. The row contains the following data.

EmployeeName	EmployeeID	EmployeeStartDate
Alice	24	2020-01-25

One minute later, you execute the following query from a serverless SQL pool in MyWorkspace.

SELECT EmployeeID

FROM mytestdb.dbo.myParquetTable

WHERE name = 'Alice';

What will be returned by the query?

A. 24

B. an error

C. a null value

Answer(s): A

---

10. DRAG DROP (Drag and Drop is not supported)
You have a table named SalesFact in an enterprise data warehouse in Azure Synapse Analytics. SalesFact contains sales data from the past 36 months and has the following characteristics:
-Is partitioned by month
-Contains one billion rows
-Has clustered columnstore indexes
At the beginning of each month, you need to remove data from SalesFact that is older than 36 months as quickly as possible.
Which three actions should you perform in sequence in a stored procedure? 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:

| Actions	Answer Area
Switch the partition containing the stale data from SalesFact to SalesFact_Work.	
Truncate the partition containing the stale data.	
Drop the SalesFact_Work table.	
Create an empty table named SalesFact_Work that has the same schema as SalesFact.	
Execute a `DELETE` statement where the value in the Date column is more than 36 months ago.	
Copy the data to a new table by using CREATE TABLE AS SELECT (CTAS).	

A. See Explanation section for answer.

Answer(s): A

---

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

/topfolder/
File1.csv   /folder1/   /folder2/   File4.csv
          File2.csv   File3.csv
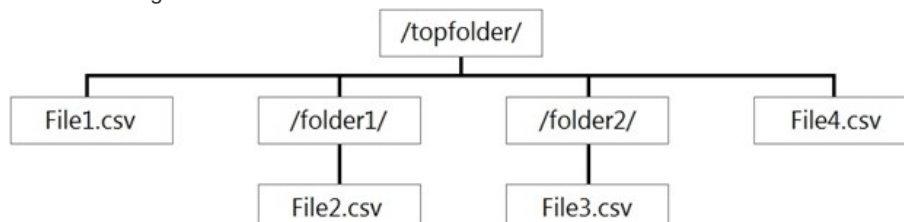
You create an external table named ExtTable that has LOCATION='/topfolder/'.
When you query ExtTable by using an Azure Synapse Analytics serverless SQL pool, which files are returned?

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

Answer(s): B

---

12. HOTSPOT (Drag and Drop is not supported)
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

Report1:
| Avro |
| CSV |
| Parquet |
| TSV |

Report2:
| Avro |
| CSV |
| Parquet |
| TSV |

A. See Explanation section for answer.

Answer(s): A

---

13. You are designing the folder structure for an Azure Data Lake Storage Gen2 container.
Users will query data by using a variety of services including Azure Databricks and Azure Synapse Analytics serverless SQL pools. The data will be secured by subject area. Most queries will include data from the current year or current month.
Which folder structure should you recommend to support fast queries and simplified folder security?

A. /{SubjectArea}/{DataSource}/{DD}/{MM}/{YYYY}/{FileData}_{YYYY}_{MM}_{DD}.csv

B. /{DD}/{MM}/{YYYY}/{SubjectArea}/{DataSource}/{FileData}_{YYYY}_{MM}_{DD}.csv

C. /{YYYY}/{MM}/{DD}/{SubjectArea}/{DataSource}/{FileData}_{YYYY}_{MM}_{DD}.csv

D. /{SubjectArea}/{DataSource}/{YYYY}/{MM}/{DD}/{FileData}_{YYYY}_{MM}_{DD}.csv

Answer(s): D

14. HOTSPOT (Drag and Drop is not supported)

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.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Columnar format:

Avro
GZip
Parquet
TXT

JSON with a timestamp:

Avro
GZip
Parquet
TXT

A. See Explanation section for answer.

Answer(s): A

---

15. HOTSPOT (Drag and Drop is not supported)

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? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Copy behavior:

Flatten hierarchy
Merge files
Preserve hierarchy

Sink file type:

CSV
JSON
Parquet
TXT

A. See Explanation section for answer.

Answer(s): A

___

16. HOTSPOT (Drag and Drop is not supported)
You have a data model that you plan to implement in a data warehouse in Azure Synapse Analytics as shown in the following exhibit.
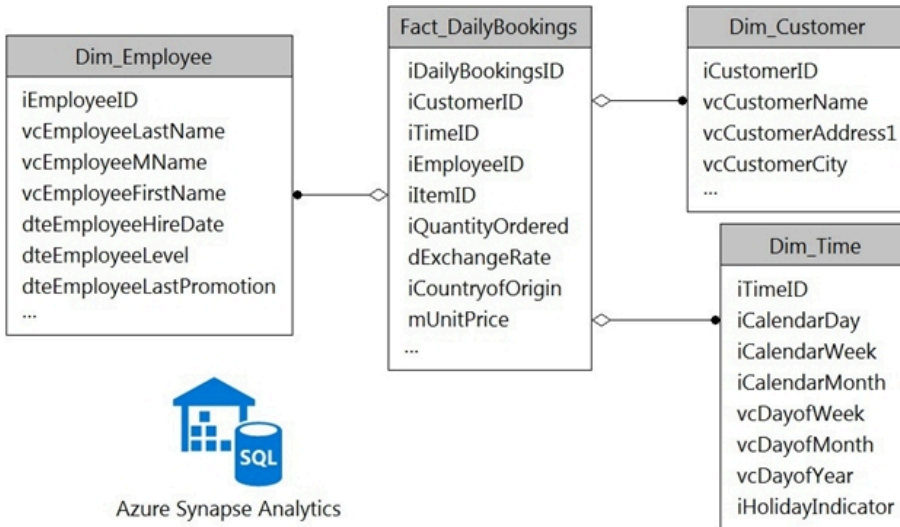
All the dimension tables will be less than 2 GB after compression, and the fact table will be approximately 6 TB. The dimension tables will be relatively static with very few data inserts and updates.
Which type of table should you use for each table? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
Hot Area:

Answer Area

Dim_Customer:
| Hash distributed |
| Round-robin |
| Replicated |

Dim_Employee:
| Hash distributed |
| Round-robin |
| Replicated |

Dim_Time:
| Hash distributed |
| Round-robin |
| Replicated |

Fact_DailyBookings:
| Hash distributed |
| Round-robin |
| Replicated |

A. See Explanation section for answer.

Answer(s): A

17. HOTSPOT (Drag and Drop is not supported)

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? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point

Hot Area:

Answer Area

Five-year-old data:

Delete the blob.
Move to archive storage.
Move to cool storage.
Move to hot storage.

Seven-year-old data:

Delete the blob.
Move to archive storage.
Move to cool storage.
Move to hot storage.

A. See Explanation section for answer.

Answer(s): A

---

18. DRAG DROP (Drag and Drop is not supported)

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

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

CLUSTERED INDEX
COLLATE
DISTRIBUTION
PARTITION
PARTITION FUNCTION
PARTITION SCHEME

Answer Area

```
CREATE TABLE table1
(
  ID INTEGER,
  col1 VARCHAR(10),
  col2 VARCHAR(10)
) WITH
(
  _____  = HASH(ID),
  _____  (ID RANGE LEFT FOR VALUES (1, 1000000
);
```

A. See Explanation section for answer.

Answer(s): A

19. 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

Answer(s): D

---

20. 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.

Answer(s): A B F