DP-203 Dumps

Data Engineering on Microsoft Azure

https://www.certleader.com/DP-203-dumps.html

NEW QUESTION 1
- (Exam Topic 3)
You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Table1. You have files that are ingested and loaded into an Azure Data Lake Storage Gen2 container named
container1.
You plan to insert data from the files into Table1 and azure Data Lake Storage Gen2 container named container1.
You plan to insert data from the files into Table1 and transform the data. Each row of data in the files will produce one row in the serving layer of Table1.
You need to ensure that when the source data files are loaded to container1, the DateTime is stored as an additional column in Table1.
Solution: In an Azure Synapse Analytics pipeline, you use a data flow that contains a Derived Column transformation.

A. Yes
B. No

Answer: A

Explanation:
Use the derived column transformation to generate new columns in your data flow or to modify existing fields. Reference:
https://docs.microsoft.com/en-us/azure/data-factory/data-flow-derived-column

NEW QUESTION 2
- (Exam Topic 3)
A company plans to use Apache Spark analytics to analyze intrusion detection data.
You need to recommend a solution to analyze network and system activity data for malicious activities and policy violations. The solution must minimize administrative efforts.
What should you recommend?

A. Azure Data Lake Storage
B. Azure Databricks
C. Azure HDInsight
D. Azure Data Factory

Answer: B

Explanation:
Three common analytics use cases with Microsoft Azure Databricks
Recommendation engines, churn analysis, and intrusion detection are common scenarios that many organizations are solving across multiple industries. They require machine learning, streaming analytics, and utilize massive amounts of data processing that can be difficult to scale without the right tools.
Recommendation engines, churn analysis, and intrusion detection are common scenarios that many organizations are solving across multiple industries. They require machine learning, streaming analytics, and utilize massive amounts of data processing that can be difficult to scale without the right tools.
Note: Recommendation engines, churn analysis, and intrusion detection are common scenarios that many
organizations are solving across multiple industries. They require machine learning, streaming analytics, and utilize massive amounts of data processing that can be difficult to scale without the right tools.
Reference:
https://azure.microsoft.com/es-es/blog/three-critical-analytics-use-cases-with-microsoft-azure-databricks/

NEW QUESTION 3
- (Exam Topic 3)
You have a self-hosted integration runtime in Azure Data Factory.
The current status of the integration runtime has the following configurations:
≫ Status: Running
≫ Type: Self-Hosted
≫ Version: 4.4.7292.1
≫ Running / Registered Node(s): 1/1
≫ High Availability Enabled: False
≫ Linked Count: 0
≫ Queue Length: 0
≫ Average Queue Duration. 0.00s
The integration runtime has the following node details:
≫ Name: X-M
≫ Status: Running
≫ Version: 4.4.7292.1
≫ Available Memory: 7697MB
≫ CPU Utilization: 6%
≫ Network (In/Out): 1.21KBps/0.83KBps
≫ Concurrent Jobs (Running/Limit): 2/14
≫ Role: Dispatcher/Worker
≫ Credential Status: In Sync
Use the drop-down menus to select the answer choice that completes each statement based on the information presented.
NOTE: Each correct selection is worth one point.

If the X-M node becomes unavailable, all executed pipelines will:

| ▼ |
| fail until the node comes back online |
| switch to another integration runtime |
| exceed the CPU limit |

The number of concurrent jobs and the CPU usage indicate that the Concurrent Jobs (Running/Limit) value should be:

| ▼ |
| raised |
| lowered |
| left as is |

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Box 1: fail until the node comes back online
We see: High Availability Enabled: False
Note: Higher availability of the self-hosted integration runtime so that it's no longer the single point of failure in your big data solution or cloud data integration with Data Factory.
Box 2: lowered We see:
Concurrent Jobs (Running/Limit): 2/14 CPU Utilization: 6%
Note: When the processor and available RAM aren't well utilized, but the execution of concurrent jobs reaches a node's limits, scale up by increasing the number of concurrent jobs that a node can run
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/create-self-hosted-integration-runtime

NEW QUESTION 4
- (Exam Topic 3)
You are processing streaming data from vehicles that pass through a toll booth.
You need to use Azure Stream Analytics to return the license plate, vehicle make, and hour the last vehicle passed during each 10-minute window.
How should you complete the query? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

```
WITH LastInWindow AS
(
    SELECT
        [▼] (Time) AS LastEventTime
        COUNT
        MAX
        MIN
        TOPONE
    FROM
        Input TIMESTAMP BY Time
    GROUP BY
        [▼] (minute, 10)
        HoppingWindow
        SessionWindow
        SlidingWindow
        TumblingWindow
)
SELECT
    Input.License_plate,
    Input.Make,
    Input.Time
FROM
    Input TIMESTAMP BY Time
    INNER JOIN LastInWindow
    ON [▼] (minute, Input, LastInWindow) BETWEEN 0 AND 10
        DATEADD
        DATEDIFF
        DATENAME
        DATEPART
    AND Input.Time = LastInWindow.LastEventTime
```

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Graphical user interface, text, application Description automatically generated
Box 1: MAX
The first step on the query finds the maximum time stamp in 10-minute windows, that is the time stamp of the last event for that window. The second step joins the results of the first query with the original stream to find the event that match the last time stamps in each window.
Query:
WITH LastInWindow AS (
SELECT
MAX(Time) AS LastEventTime FROM
Input TIMESTAMP BY Time GROUP BY
TumblingWindow(minute, 10)
) SELECT
Input.License_plate, Input.Make, Input.Time
FROM
Input TIMESTAMP BY Time INNER JOIN LastInWindow
ON DATEDIFF(minute, Input, LastInWindow) BETWEEN 0 AND 10 AND Input.Time = LastInWindow.LastEventTime
Box 2: TumblingWindow
Tumbling windows are a series of fixed-sized, non-overlapping and contiguous time intervals. Box 3: DATEDIFF
DATEDIFF is a date-specific function that compares and returns the time difference between two DateTime fields, for more information, refer to date functions.
Reference:
https://docs.microsoft.com/en-us/stream-analytics-query/tumbling-window-azure-stream-analytics

NEW QUESTION 5
- (Exam Topic 3)
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 have an Azure Synapse Analytics dedicated SQL pool that contains a table named Table1. You have files that are ingested and loaded into an Azure Data Lake Storage Gen2 container named
container1.
You plan to insert data from the files in container1 into Table1 and transform the data. Each row of data in the files will produce one row in the serving layer of Table1.
You need to ensure that when the source data files are loaded to container1, the DateTime is stored as an additional column in Table1.
Solution: You use an Azure Synapse Analytics serverless SQL pool to create an external table that has an additional DateTime column.
Does this meet the goal?

A. Yes
B. No

Answer: B

Explanation:
Instead use the derived column transformation to generate new columns in your data flow or to modify existing fields.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/data-flow-derived-column

NEW QUESTION 6
- (Exam Topic 3)
You have an Azure Databricks workspace and an Azure Data Lake Storage Gen2 account named storage! New files are uploaded daily to storage1.
• Incrementally process new files as they are upkorage1 as a structured streaming source. The solution must meet the following requirements:
• Minimize implementation and maintenance effort.
• Minimize the cost of processing millions of files.
• Support schema inference and schema drift. Which should you include in the recommendation?

A. Auto Loader
B. Apache Spark FileStreamSource
C. COPY INTO
D. Azure Data Factory

Answer: D

NEW QUESTION 7
- (Exam Topic 3)
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?

A. Enable container services for workspace1.
B. Upgrade workspace1 to the Premium pricing tier.
C. Set Cluster Mode to High Concurrency.
D. Create a cluster policy in workspace1.

Answer: B

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

NEW QUESTION 8
- (Exam Topic 3)
You have an enterprise data warehouse in Azure Synapse Analytics that contains a table named FactOnlineSales. The table contains data from the start of 2009 to the end of 2012.
You need to improve the performance of queries against FactOnlineSales by using table partitions. The solution must meet the following requirements:

≫ Create four partitions based on the order date.

≫ Ensure that each partition contains all the orders places during a given calendar year.
How should you complete the T-SQL command? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

```
CREATE TABLE [dbo].FactOnlineSales
([OnlineSalesKey] [int] NOT NULL,
[OrderDateKey] [datetime]    NOT NULL,
[StoreKey] [int]        NOT NULL,
[ProductKey] [int]       NOT NULL,
[CustomerKey] [int]      NOT NULL,
[SalesOrderNumber] [varchar](20) NOT NULL,
[SalesQuantity] [int]    NOT NULL,
[SalesAmount] [money]    NOT NULL,
[UnitPrice]    [money]   NULL)
WITH (CLUSTERED COLUMNSTORE INDEX)
PARTITION ([OrderDateKey] RANGE [▼] FOR VALUES
    RIGHT
    LEFT

(  [▼]  )
    20090101,20121231
    20100101,20110101,20120101
    20090101,20100101,20110101,20120101
```

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Text Description automatically generated
Range Left or Right, both are creating similar partition but there is difference in comparison For example: in this scenario, when you use LEFT and 20100101,20110101,20120101
Partition will be, datecol<=20100101, datecol>20100101 and datecol<=20110101, datecol>20110101 and datecol<=20120101, datecol>20120101
But if you use range RIGHT and 20100101,20110101,20120101
Partition will be, datecol<20100101, datecol>=20100101 and datecol<20110101, datecol>=20110101 and datecol<20120101, datecol>=20120101
In this example, Range RIGHT will be suitable for calendar comparison Jan 1st to Dec 31st Reference:
https://docs.microsoft.com/en-us/sql/t-sql/statements/create-partition-function-transact-sql?view=sql-server-ver1

NEW QUESTION 9
- (Exam Topic 3)
You are designing an Azure Synapse Analytics dedicated SQL pool.
You need to ensure that you can audit access to Personally Identifiable information (PII). What should you include in the solution?

A. dynamic data masking
B. row-level security (RLS)
C. sensitivity classifications
D. column-level security

Answer: C

Explanation:
Data Discovery & Classification is built into Azure SQL Database, Azure SQL Managed Instance, and Azure Synapse Analytics. It provides basic capabilities for discovering, classifying, labeling, and reporting the sensitive data in your databases.
Your most sensitive data might include business, financial, healthcare, or personal information. Discovering and classifying this data can play a pivotal role in your organization's information-protection approach. It can serve as infrastructure for:

> Helping to meet standards for data privacy and requirements for regulatory compliance.

> Various security scenarios, such as monitoring (auditing) access to sensitive data.

> Controlling access to and hardening the security of databases that contain highly sensitive data.
Reference:
https://docs.microsoft.com/en-us/azure/azure-sql/database/data-discovery-and-classification-overview

NEW QUESTION 10
- (Exam Topic 3)
You have an Azure Data Factory pipeline named pipeline1 that is invoked by a tumbling window trigger named Trigger1. Trigger1 has a recurrence of 60 minutes.
You need to ensure that pipeline1 will execute only if the previous execution completes successfully. How should you configure the self-dependency for Trigger1?

A. offset: "-00:01:00" size: "00:01:00"
B. offset: "01:00:00" size: "-01:00:00"
C. offset: "01:00:00" size: "01:00:00"
D. offset: "-01:00:00" size: "01:00:00"

Answer: D

Explanation:
Tumbling window self-dependency properties
In scenarios where the trigger shouldn't proceed to the next window until the preceding window is successfully completed, build a self-dependency. A self-dependency trigger that's dependent on the success of earlier runs of itself within the preceding hour will have the properties indicated in the following code.
Example code:
"name": "DemoSelfDependency", "properties": {
"runtimeState": "Started", "pipeline": { "pipelineReference": { "referenceName": "Demo", "type": "PipelineReference"
}
},
"type": "TumblingWindowTrigger", "typeProperties": {
"frequency": "Hour", "interval": 1,
"startTime": "2018-10-04T00:00:00Z", "delay": "00:01:00",
"maxConcurrency": 50, "retryPolicy": { "intervalInSeconds": 30
},
"dependsOn": [
{
"type": "SelfDependencyTumblingWindowTriggerReference",
"size": "01:00:00",
"offset": "-01:00:00"
}
]
}
}
}
Reference: https://docs.microsoft.com/en-us/azure/data-factory/tumbling-window-trigger-dependency

NEW QUESTION 10
- (Exam Topic 3)
You have an Azure Synapse Analytics dedicated SQL pool named Pool1 and a database named DB1. DB1 contains a fact table named Table1.
You need to identify the extent of the data skew in Table1. What should you do in Synapse Studio?

A. Connect to the built-in pool and run dbcc pdw_showspaceused.
B. Connect to the built-in pool and run dbcc checkalloc.
C. Connect to Pool1 and query sys.dm_pdw_node_scacus.
D. Connect to Pool1 and query sys.dm_pdw_nodes_db_partition_scacs.

Answer: A

Explanation:
A quick way to check for data skew is to use DBCC PDW_SHOWSPACEUSED. The following SQL code returns the number of table rows that are stored in each of the 60 distributions. For balanced performance, the rows in your distributed table should be spread evenly across all the distributions.
DBCC PDW_SHOWSPACEUSED('dbo.FactInternetSales'); Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribu

NEW QUESTION 11
- (Exam Topic 3)
You are designing the folder structure for an Azure Data Lake Storage Gen2 account. You identify the following usage patterns:
• Users will query data by using Azure Synapse Analytics serverless SQL pools and Azure Synapse Analytics serverless Apache Spark pods.
• Most queries will include a filter on the current year or week.
• Data will be secured by data source.
You need to recommend a folder structure that meets the following requirements:
• Supports the usage patterns
• Simplifies folder security
• Minimizes query times
Which folder structure should you recommend?
A)

```
\YYYYY\WW\DataSource\SubjectArea\FileData_YYYY_MM_DD.parquet
```

B)

```
DataSource\SubjectArea\WW\YYYY\FileData_YYYY_MM_DD.parquet
```

C)

```
\DataSource\SubjectArea\YYYY\WW\FileData_YYYY_MM_DD.parquet
```

D)

```
\DataSource\SubjectArea\YYYY-WW\FileData_YYYY_MM_DD.parquet
```

E)

```
WW\YYYY\SubjectArea\DataSource\FileData_YYYY_MM_DD.parquet
```

A. Option A
B. Option B
C. Option C
D. Option D
E. Option E

Answer: C

Explanation:
Data will be secured by data source. -> Use DataSource as top folder.
Most queries will include a filter on the current year or week -> Use \YYYY\WW\ as subfolders. Common Use Cases
A common use case is to filter data stored in a date (and possibly time) folder structure such as
/YYYY/MM/DD/ or /YYYY/MM/YYYY-MM-DD/. As new data is generated/sent/copied/moved to the storage account, a new folder is created for each specific time period. This strategy organises data into a maintainable folder structure.
Reference: https://www.serverlesssql.com/optimisation/azurestoragefilteringusingfilepath/

NEW QUESTION 13
- (Exam Topic 3)
You have an Azure Storage account that generates 200.000 new files daily. The file names have a format of (YYY)/(MM)/(DD)/|HH])/(CustornerID).csv.
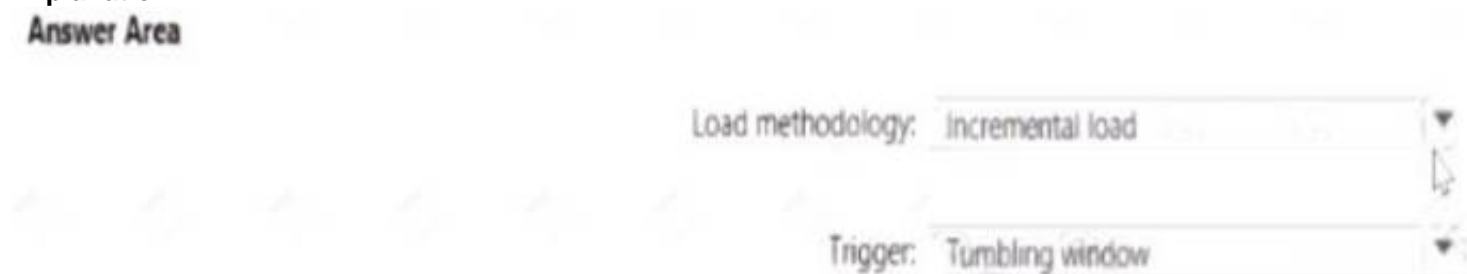You need to design an Azure Data Factory solution that will toad new data from the storage account to an Azure Data lake once hourly. The solution must minimize load times and costs.
How should you configure the solution? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

A. Mastered
B. Not Mastered

Answer: A

Explanation:

Answer Area

Load methodology:	Incremental load ▼

Trigger:	Tumbling window ▼

NEW QUESTION 17
- (Exam Topic 3)
You plan to create a dimension table in Azure Synapse Analytics that will be less than 1 GB. You need to create the table to meet the following requirements:
• Provide the fastest Query time.
• Minimize data movement during queries. Which type of table should you use?

A. hash distributed
B. heap
C. replicated
D. round-robin

Answer: C

Explanation:
A replicated table has a full copy of the table accessible on each Compute node. Replicating a table removes the need to transfer data among Compute nodes before a join or aggregation. Since the table has multiple copies, replicated tables work best when the table size is less than 2 GB compressed. 2 GB is not a hard limit.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/design-guidance-for-replicated-tab

NEW QUESTION 22
- (Exam Topic 3)
You have an Azure data factory.
You need to examine the pipeline failures from the last 180 flays. What should you use?

A. the Activity tog blade for the Data Factory resource
B. Azure Data Factory activity runs in Azure Monitor
C. Pipeline runs in the Azure Data Factory user experience
D. the Resource health blade for the Data Factory resource

Answer: B

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

NEW QUESTION 23
- (Exam Topic 3)
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 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

NEW QUESTION 25
- (Exam Topic 3)
You are building an Azure Stream Analytics job to identify how much time a user spends interacting with a feature on a webpage.
The job receives events based on user actions on the webpage. Each row of data represents an event. Each event has a type of either 'start' or 'end'.
You need to calculate the duration between start and end events.
How should you complete the query? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

```
SELECT
    [user],
    feature,
    _____▼
    DATEADD(
    DATEDIFF(
    DATEPART(
        second,
    _____▼  (Time) OVER (PARTITION BY [user], feature LIMIT DURATION(hour, 1) WHEN Event = 'start'),
    ISFIRST
    LAST
    TOPONE
        Time) as duration
    FROM input TIMESTAMP BY Time
    WHERE
        Event = 'end'
```

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Box 1: DATEDIFF
DATEDIFF function returns the count (as a signed integer value) of the specified datepart boundaries crossed between the specified startdate and enddate.
Syntax: DATEDIFF (datepart , startdate, enddate) Box 2: LAST
The LAST function can be used to retrieve the last event within a specific condition. In this example, the condition is an event of type Start, partitioning the search

by PARTITION BY user and feature. This way, every user and feature is treated independently when searching for the Start event. LIMIT DURATION limits the search back in time to 1 hour between the End and Start events.
Example: SELECT
[user], feature, DATEDIFF(
second,
LAST(Time) OVER (PARTITION BY [user], feature LIMIT DURATION(hour,
1) WHEN Event = 'start'), Time) as duration
FROM input TIMESTAMP BY Time WHERE
Event = 'end' Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-stream-analytics-query-patterns

NEW QUESTION 27
- (Exam Topic 3)
You are implementing a batch dataset in the Parquet format.
Data tiles will be produced by using Azure Data Factory and stored in Azure Data Lake Storage Gen2. The files will be consumed by an Azure Synapse Analytics serverless SQL pool.
You need to minimize storage costs for the solution. What should you do?

A. Store all the data as strings in the Parquet tiles.
B. Use OPENROWEST to query the Parquet files.
C. Create an external table mat contains a subset of columns from the Parquet files.
D. Use Snappy compression for the files.

Answer: C

Explanation:
An external table points to data located in Hadoop, Azure Storage blob, or Azure Data Lake Storage. External tables are used to read data from files or write data to files in Azure Storage. With Synapse SQL, you can use external tables to read external data using dedicated SQL pool or serverless SQL pool.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables

NEW QUESTION 32
- (Exam Topic 3)
You are implementing a star schema in an Azure Synapse Analytics dedicated SQL pool. You plan to create a table named DimProduct.
DimProduct must be a Type 3 slowly changing dimension (SCO) table that meets the following requirements:
• The values in two columns named ProductKey and ProductSourceID will remain the same.
• The values in three columns named ProductName, ProductDescription, and Color can change. You need to add additional columns to complete the following table definition.

```
CREATE TABLE [dbo].[dimproduct]
  (
      [ProductKey]          INT NOT NULL,
      [ProductSourceID]     INT NOT NULL,
      [ProductName]         NVARCHAR(100) NOT NULL,
      [ProductDescription]  NVARCHAR(2000) NOT NULL,
      [Color]               NVARCHAR(50) NOT NULL
  )
  WITH
  (
      DISTRIBUTION = REPLICATE,
      CLUSTERED COLUMNSTORE INDEX
  );
```

A)
```
[OriginalProductDescription] NVARCHAR(2000) NOT NULL
```

B)
```
[IsCurrentRow] [bit] NOT NULL
```

C)
```
[EffectiveStartDate] [datetime] NOT NULL
```

D)
```
[EffectiveEndDate] [datetime] NOT NULL
```

E)
```
[OriginalProductName] NVARCHAR(100) NULL
```

F)
```
[OriginalColor] NVARCHAR(50) NOT NULL
```

A. Option A
B. Option B
C. Option C
D. Option D
E. Option E
F. Option F

Answer: ABC

NEW QUESTION 33
- (Exam Topic 3)
You are designing database for an Azure Synapse Analytics dedicated SQL pool to support workloads for detecting ecommerce transaction fraud.
Data will be combined from multiple ecommerce sites and can include sensitive financial information such as credit card numbers.
You need to recommend a solution that meets the following requirements:

≫ Users must be able to identify potentially fraudulent transactions.

≫ Users must be able to use credit cards as a potential feature in models.

≫ Users must NOT be able to access the actual credit card numbers.
What should you include in the recommendation?

A. Transparent Data Encryption (TDE)
B. row-level security (RLS)
C. column-level encryption
D. Azure Active Directory (Azure AD) pass-through authentication

Answer: C

Explanation:
Use Always Encrypted to secure the required columns. You can configure Always Encrypted for individual database columns containing your sensitive data.
Always Encrypted is a feature designed to protect sensitive data, such as credit card numbers or national identification numbers (for example, U.S. social security numbers), stored in Azure SQL Database or SQL Server databases.
Reference:
https://docs.microsoft.com/en-us/sql/relational-databases/security/encryption/always-encrypted-database-engine

NEW QUESTION 34
- (Exam Topic 3)
You have an Azure subscription linked to an Azure Active Directory (Azure AD) tenant that contains a service principal named ServicePrincipal1. The subscription contains an Azure Data Lake Storage account named adls1. Adls1 contains a folder named Folder2 that has a URI of
https://adls1.dfs.core.windows.net/container1/Folder1/Folder2/.
ServicePrincipal1 has the access control list (ACL) permissions shown in the following table.

Resource	Permission
container1	Access – Execute
Folder1	Access – Execute
Folder2	Access – Read

You need to ensure that ServicePrincipal1 can perform the following actions:

≫ Traverse child items that are created in Folder2.

≫ Read files that are created in Folder2.
The solution must use the principle of least privilege.
Which two permissions should you grant to ServicePrincipal1 for Folder2? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.

A. Access - Read
B. Access - Write
C. Access - Execute
D. Default-Read
E. Default - Write
F. Default - Execute

Answer: DF

Explanation:
Execute (X) permission is required to traverse the child items of a folder.
There are two kinds of access control lists (ACLs), Access ACLs and Default ACLs. Access ACLs: These control access to an object. Files and folders both have Access ACLs.
Default ACLs: A "template" of ACLs associated with a folder that determine the Access ACLs for any child items that are created under that folder. Files do not have Default ACLs.
Reference:
https://docs.microsoft.com/en-us/azure/data-lake-store/data-lake-store-access-control

NEW QUESTION 37
- (Exam Topic 3)
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.

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'
    )
```

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Graphical user interface, text, application, email Description automatically generated
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

NEW QUESTION 38
- (Exam Topic 3)
You are developing a solution that will stream to Azure Stream Analytics. The solution will have both streaming data and reference data.
Which input type should you use for the reference data?

A. Azure Cosmos DB
B. Azure Blob storage
C. Azure IoT Hub
D. Azure Event Hubs

Answer: B

Explanation:
Stream Analytics supports Azure Blob storage and Azure SQL Database as the storage layer for Reference Data.
Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-use-reference-data

NEW QUESTION 41
- (Exam Topic 3)
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:
≫ Minimize the risk of unauthorized user access.
≫ Use the principle of least privilege.
≫ 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.

Use [▼] to authenticate by using [▼]

Azure Active Directory (Azure AD)	a managed identity
a shared access signature (SAS)	a stored access policy
a shared key	an Authorization header

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Text Description automatically generated with low confidence
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.
≫ Account key authentication
≫ Service principal authentication
≫ 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

NEW QUESTION 45
- (Exam Topic 3)
You have an Azure subscription that contains an Azure Synapse Analytics workspace named workspace1. Workspace1 contains a dedicated SQL pool named SQL Pool and an Apache Spark pool named sparkpool. Sparkpool1 contains a DataFrame named pyspark.df.
You need to write the contents of pyspark_df to a tabte in SQLPooM by using a PySpark notebook. How should you complete the code? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

Answer Area

```
pyspark_df.createOrReplaceTempView("pysparkdftemptable")

[          ▼]
  %%local          park.sqlContext.sql ("select * from pysparkdftemptable")
  %%spark
  %%sql          [          ▼]  ("sqlpool1.dbo.PySparkTable", Constants.INTERNAL)
                    jdbc
                    saveAsTable
                    synapsesql
```

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Answer Area

```
pyspark_df.createOrReplaceTempView("pysparkdftemptable")

[          ▼]
  %%local          park.sqlContext.sql ("select * from pysparkdftemptable")
  %%spark
  %%sql          [          ▼]  ("sqlpool1.dbo.PySparkTable", Constants.INTERNAL)
                    jdbc
                    saveAsTable
                    synapsesql
```

NEW QUESTION 50
- (Exam Topic 3)
You have the following table named Employees.

first_name	last_name	hire_date	employee_type
Jane	Doe	2019-08-23	new
Ben	Smith	2017-12-15	Standard

You need to calculate the employee_type value based on the hire_date value.
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.

Values   **Answer Area**

```
                    SELECT
                      *,
  CASE              [          ]

  ELSE                  WHEN hire_date >= '2019-01-01' THEN 'New'

  OVER              [          ]        'Standard'

  PARTITION BY          END AS employee_type

  ROW_NUMBER      FROM

                      employees
```

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Graphical user interface, text, application Description automatically generated
Box 1: CASE
CASE evaluates a list of conditions and returns one of multiple possible result expressions.
CASE can be used in any statement or clause that allows a valid expression. For example, you can use CASE in statements such as SELECT, UPDATE, DELETE and SET, and in clauses such as select_list, IN, WHERE, ORDER BY, and HAVING.
Syntax: Simple CASE expression: CASE input_expression
WHEN when_expression THEN result_expression [...n] [ELSE else_result_expression]
END
Box 2: ELSE
Reference:
https://docs.microsoft.com/en-us/sql/t-sql/language-elements/case-transact-sql

NEW QUESTION 53
- (Exam Topic 3)
You have an Azure data solution that contains an enterprise data warehouse in Azure Synapse Analytics named DW1.
Several users execute ad hoc queries to DW1 concurrently. You regularly perform automated data loads to DW1.
You need to ensure that the automated data loads have enough memory available to complete quickly and successfully when the adhoc queries run.
What should you do?

A. Hash distribute the large fact tables in DW1 before performing the automated data loads.
B. Assign a smaller resource class to the automated data load queries.
C. Assign a larger resource class to the automated data load queries.
D. Create sampled statistics for every column in each table of DW1.

Answer: C

Explanation:
The performance capacity of a query is determined by the user's resource class. Resource classes are
pre-determined resource limits in Synapse SQL pool that govern compute resources and concurrency for query execution.
Resource classes can help you configure resources for your queries by setting limits on the number of queries that run concurrently and on the compute-resources
assigned to each query. There's a trade-off between memory and concurrency.
Smaller resource classes reduce the maximum memory per query, but increase concurrency. Larger resource classes increase the maximum memory per query,
but reduce concurrency. Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/resource-classes-for-workload-ma

NEW QUESTION 55
- (Exam Topic 3)
You are developing an application that uses Azure Data Lake Storage Gen 2.
You need to recommend a solution to grant permissions to a specific application for a limited time period. What should you include in the recommendation?

A. Azure Active Directory (Azure AD) identities
B. shared access signatures (SAS)
C. account keys
D. role assignments

Answer: B

Explanation:
A shared access signature (SAS) provides secure delegated access to resources in your storage account. With a SAS, you have granular control over how a client
can access your data. For example:
What resources the client may access.
What permissions they have to those resources. How long the SAS is valid.
Reference:
https://docs.microsoft.com/en-us/azure/storage/common/storage-sas-overview

NEW QUESTION 57
- (Exam Topic 3)
You have an Azure Factory instance named DF1 that contains a pipeline named PL1.PL1 includes a tumbling window trigger.
You create five clones of PL1. You configure each clone pipeline to use a different data source.
You need to ensure that the execution schedules of the clone pipeline match the execution schedule of PL1. What should you do?

A. Add a new trigger to each cloned pipeline
B. Associate each cloned pipeline to an existing trigger.
C. Create a tumbling window trigger dependency for the trigger of PL1.
D. Modify the Concurrency setting of each pipeline.

Answer: B

NEW QUESTION 59
- (Exam Topic 3)
You are designing a real-time dashboard solution that will visualize streaming data from remote sensors that connect to the internet. The streaming data must be
aggregated to show the average value of each 10-second interval. The data will be discarded after being displayed in the dashboard.
The solution will use Azure Stream Analytics and must meet the following requirements:
≫ Minimize latency from an Azure Event hub to the dashboard.
≫ Minimize the required storage.
≫ Minimize development effort.
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

Azure Stream Analytics input type:	▼
	Azure Event Hub
	Azure SQL Database
	Azure Stream Analytics
	Microsoft Power BI

Azure Stream Analytics output type:	▼
	Azure Event Hub
	Azure SQL Database
	Azure Stream Analytics
	Microsoft Power BI

Aggregation query location:	▼
	Azure Event Hub
	Azure SQL Database
	Azure Stream Analytics
	Microsoft Power BI

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-power-bi-dashboard

NEW QUESTION 60
- (Exam Topic 3)
You have the following Azure Data Factory pipelines
• ingest Data from System 1
• Ingest Data from System2
• Populate Dimensions
• Populate facts
ingest Data from System1 and Ingest Data from System1 have no dependencies. Populate Dimensions must execute after Ingest Data from System1 and Ingest Data from System* Populate Facts must execute after the Populate Dimensions pipeline. All the pipelines must execute every eight hours.
What should you do to schedule the pipelines for execution?

A. Add an event trigger to all four pipelines.
B. Create a parent pipeline that contains the four pipelines and use an event trigger.
C. Create a parent pipeline that contains the four pipelines and use a schedule trigger.
D. Add a schedule trigger to all four pipelines.

Answer: C

Explanation:
Schedule trigger: A trigger that invokes a pipeline on a wall-clock schedule. Reference:
https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipeline-execution-triggers

NEW QUESTION 65
- (Exam Topic 3)
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: B

Explanation:
Once a database has been created by a Spark job, you can create tables in it with Spark that use Parquet as the storage format. Table names will be converted to lower case and need to be queried using the lower case name. These tables will immediately become available for querying by any of the Azure Synapse workspace Spark pools. They can also be used from any of the Spark jobs subject to permissions.
Note: For external tables, since they are synchronized to serverless SQL pool asynchronously, there will be a delay until they appear.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/metadata/table

NEW QUESTION 66
- (Exam Topic 3)
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. a shared access signature (SAS)
B. a managed identity
C. a shared key
D. an Azure Active Directory (Azure AD) user

Answer: B

Explanation:
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-exa

NEW QUESTION 71
- (Exam Topic 3)
You are designing a statistical analysis solution that will use custom proprietary1 Python functions on near real-time data from Azure Event Hubs.
You need to recommend which Azure service to use to perform the statistical analysis. The solution must minimize latency.
What should you recommend?

A. Azure Stream Analytics
B. Azure SQL Database
C. Azure Databricks
D. Azure Synapse Analytics

Answer: A

Explanation:
Reference:
https://docs.microsoft.com/en-us/azure/event-hubs/process-data-azure-stream-analytics

NEW QUESTION 76
- (Exam Topic 3)
You haw an Azure data factory named ADF1.
You currently publish all pipeline authoring changes directly to ADF1.
You need to implement version control for the changes made to pipeline artifacts. The solution must ensure that you can apply version control to the resources currently defined m the UX Authoring canvas for ADF1.
Which two actions should you perform? Each correct answer presents part of the solution NOTE: Each correct selection is worth one point.

A. Create an Azure Data Factory trigger
B. From the UX Authoring canvas, select Set up code repository
C. Create a GitHub action
D. From the UX Authoring canvas, run Publish All.
E. Create a Git repository
F. From the UX Authoring canvas, select Publish

Answer: DE

Explanation:
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/source-control

NEW QUESTION 79
- (Exam Topic 3)
You have an Azure Data Factory instance named ADF1 and two Azure Synapse Analytics workspaces named WS1 and WS2.
ADF1 contains the following pipelines:
› P1: Uses a copy activity to copy data from a nonpartitioned table in a dedicated SQL pool of WS1 to an Azure Data Lake Storage Gen2 account
› P2: Uses a copy activity to copy data from text-delimited files in an Azure Data Lake Storage Gen2 account to a nonpartitioned table in a dedicated SQL pool of WS2
You need to configure P1 and P2 to maximize parallelism and performance.
Which dataset settings should you configure for the copy activity if each pipeline? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

P1:
```
                                                    ▼
    Set the Copy method to Bulk insert
    Set the Copy method to PolyBase
    Set the Isolation level to Repeatable read
    Set the Partition option to Dynamic range
```

P2:
```
                                                    ▼
    Set the Copy method to Bulk insert
    Set the Copy method to PolyBase
    Set the Isolation level to Repeatable read
    Set the Partition option to Dynamic range
```

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Box 1: Set the Copy method to PolyBase
While SQL pool supports many loading methods including non-Polybase options such as BCP and SQL BulkCopy API, the fastest and most scalable way to load data is through PolyBase. PolyBase is a technology that accesses external data stored in Azure Blob storage or Azure Data Lake Store via the T-SQL language.
Box 2: Set the Copy method to Bulk insert
Polybase not possible for text files. Have to use Bulk insert. Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/load-data-overview

NEW QUESTION 81
- (Exam Topic 3)
You need to trigger an Azure Data Factory pipeline when a file arrives in an Azure Data Lake Storage Gen2 container.
Which resource provider should you enable?

A. Microsoft.Sql
B. Microsoft-Automation
C. Microsoft.EventGrid
D. Microsoft.EventHub

Answer: C

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. Data Factory natively integrates with Azure Event Grid, which lets you trigger pipelines on such events.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/how-to-create-event-trigger https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipeline-execution-triggers

NEW QUESTION 82
- (Exam Topic 3)
You are building an Azure Synapse Analytics dedicated SQL pool that will contain a fact table for transactions from the first half of the year 2020.
You need to ensure that the table meets the following requirements:
≫ Minimizes the processing time to delete data that is older than 10 years
≫ Minimizes the I/O for queries that use year-to-date values
How should you complete the Transact-SQL statement? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

Answer Area

```
CREATE TABLE [dbo].[FactTransaction]

(
        [TransactionTypeID]    int      NOT NULL

,       [TransactionDateID]    int      NOT NULL

,       [CustomerID]           int      NOT NULL

,       [RecipientID]          int      NOT NULL

,       [Amount]               money    NOT NU::

)

WITH

(
```

▼
CLUSTERED COLUMNSTORE INDEX
DISTRIBUTION
PARTITION
TRUNCATE_TARGET

```
        (
```

▼	RANGE RIGHT FOR VALUES
[TransactionDateID]	
[TransactionDateID], [TransactionTypeID]	
HASH([TransactionTypeID])	
ROUND_ROBIN	

```
            (20200101,20200201,20200301,20200401,20200501,20200601)
```

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Table Description automatically generated
Box 1: PARTITION
RANGE RIGHT FOR VALUES is used with PARTITION.
Part 2: [TransactionDateID] Partition on the date column.
Example: Creating a RANGE RIGHT partition function on a datetime column
The following partition function partitions a table or index into 12 partitions, one for each month of a year's worth of values in a datetime column.
CREATE PARTITION FUNCTION [myDateRangePF1] (datetime)
AS RANGE RIGHT FOR VALUES ('20030201', '20030301', '20030401',
'20030501', '20030601', '20030701', '20030801',
'20030901', '20031001', '20031101', '20031201');
Reference:
https://docs.microsoft.com/en-us/sql/t-sql/statements/create-partition-function-transact-sql

NEW QUESTION 87
- (Exam Topic 3)
You have a SQL pool in Azure Synapse that contains a table named dbo.Customers. The table contains a column name Email.
You need to prevent nonadministrative users from seeing the full email addresses in the Email column. The users must see values in a format of
aXXX@XXXX.com instead.
What should you do?

A. From Microsoft SQL Server Management Studio, set an email mask on the Email column.
B. From the Azure portal, set a mask on the Email column.
C. From Microsoft SQL Server Management studio, grant the SELECT permission to the users for all the columns in the dbo.Customers table except Email.
D. From the Azure portal, set a sensitivity classification of Confidential for the Email column.

Answer: D

Explanation:
From Microsoft SQL Server Management Studio, set an email mask on the Email column. This is because "This feature cannot be set using portal for Azure Synapse (use PowerShell or REST API) or SQL Managed Instance." So use Create table statement with Masking e.g. CREATE TABLE Membership (MemberID int IDENTITY PRIMARY KEY, FirstName varchar(100) MASKED WITH (FUNCTION = 'partial(1,"XXXXXXX",0)') NULL, . .
https://docs.microsoft.com/en-us/azure/azure-sql/database/dynamic-data-masking-overview
upvoted 24 times

NEW QUESTION 92
- (Exam Topic 3)
You have an Azure Synapse Analytics dedicated SQL pool named SA1 that contains a table named Table1. You need to identify tables that have a high percentage of deleted rows. What should you run?
A)

```
sys.pdw_nodes_column_store_segments
```

B)
```
sys.dm_db_column_store_row_group_operational_stats
```
C)
```
sys.pdw_nodes_column_store_row_groups
```

D)
```
sys.dm_db_column_store_row_group_physical_stats
```

A. Option
B. Option
C. Option
D. Option

Answer: B

NEW QUESTION 95
- (Exam Topic 3)
You have an Azure Data Lake Storage Gen2 account that contains two folders named Folder and Folder2. You use Azure Data Factory to copy multiple files from Folder1 to Folder2.

```
Operation on target Copy_sks failed: Failure happened on 'Sink' side.
ErrorCode=DelimitedTextMoreColumnsThanDefined,
'Type=Microsoft.DataTransfer.Common.Shared.HybridDeliveryException,
Message=Error found when processing 'Csv/Tsv Format Text' source
'0_2020_11_09_11_43_32.avro' with row number 53: found more columns
than expected column count 27., Source=Microsoft.DataTransfer.Common,'
```

You receive the following error.
What should you do to resolve the error.

A. Add an explicit mapping.
B. Enable fault tolerance to skip incompatible rows.
C. Lower the degree of copy parallelism
D. Change the Copy activity setting to Binary Copy

Answer: A

Explanation:
Reference:
https://knowledge.informatica.com/s/article/Microsoft-Azure-Data-Lake-Store-Gen2-target-file-names-not-gene

NEW QUESTION 100
- (Exam Topic 3)
You have an Azure Stream Analytics job that is a Stream Analytics project solution in Microsoft Visual Studio. The job accepts data generated by IoT devices in the JSON format.
You need to modify the job to accept data generated by the IoT devices in the Protobuf format.
Which three actions should you perform from Visual Studio on sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

Actions   **Answer Area**

| Change the Event Serialization Format to Protobuf in the input.json file of the job and reference the DLL. |
| Add an Azure Stream Analytics Custom Deserializer Project (.NET) project to the solution. |
| Add .NET deserializer code for Protobuf to the custom deserializer project. |
| Add .NET deserializer code for Protobuf to the Stream Analytics project. |
| Add an Azure Stream Analytics Application project to the solution. |

A. Mastered
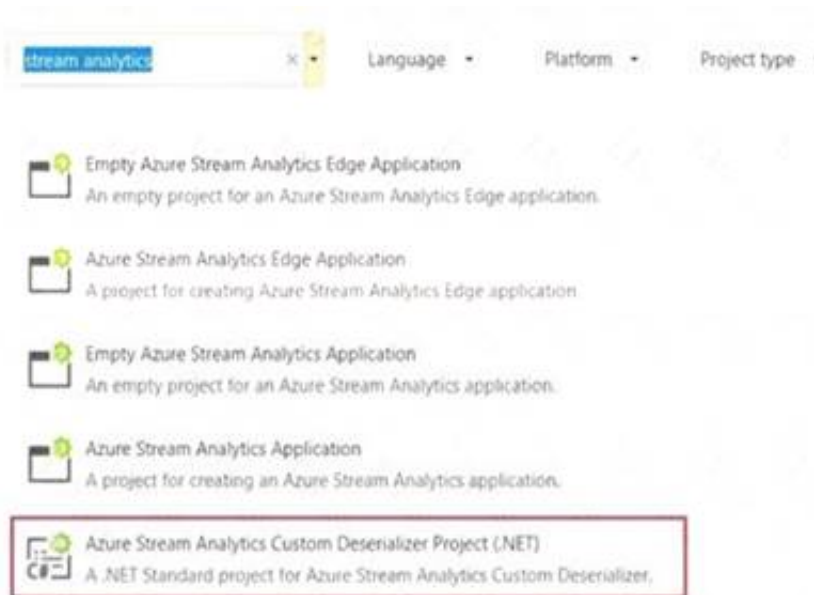B. Not Mastered

Answer: A

Explanation:
Step 1: Add an Azure Stream Analytics Custom Deserializer Project (.NET) project to the solution. Create a custom deserializer
* 1. Open Visual Studio and select File > New > Project. Search for Stream Analytics and select Azure Stream Analytics Custom Deserializer Project (.NET). Give the project a name, like Protobuf Deserializer.

* 2. In Solution Explorer, right-click your Protobuf Deserializer project and select Manage NuGet Packages from the menu. Then install the Microsoft.Azure.StreamAnalytics and Google.Protobuf NuGet packages.
* 3. Add the MessageBodyProto class and the MessageBodyDeserializer class to your project.
* 4. Build the Protobuf Deserializer project.
Step 2: Add .NET deserializer code for Protobuf to the custom deserializer project
Azure Stream Analytics has built-in support for three data formats: JSON, CSV, and Avro. With custom .NET deserializers, you can read data from other formats such as Protocol Buffer, Bond and other user defined formats for both cloud and edge jobs.
Step 3: Add an Azure Stream Analytics Application project to the solution Add an Azure Stream Analytics project

⯈ In Solution Explorer, right-click the Protobuf Deserializer solution and select Add > New Project. Under Azure Stream Analytics > Stream Analytics, choose Azure Stream Analytics Application. Name it ProtobufCloudDeserializer and select OK.

⯈ Right-click References under the ProtobufCloudDeserializer Azure Stream Analytics project. Under Projects, add Protobuf Deserializer. It should be automatically populated for you.
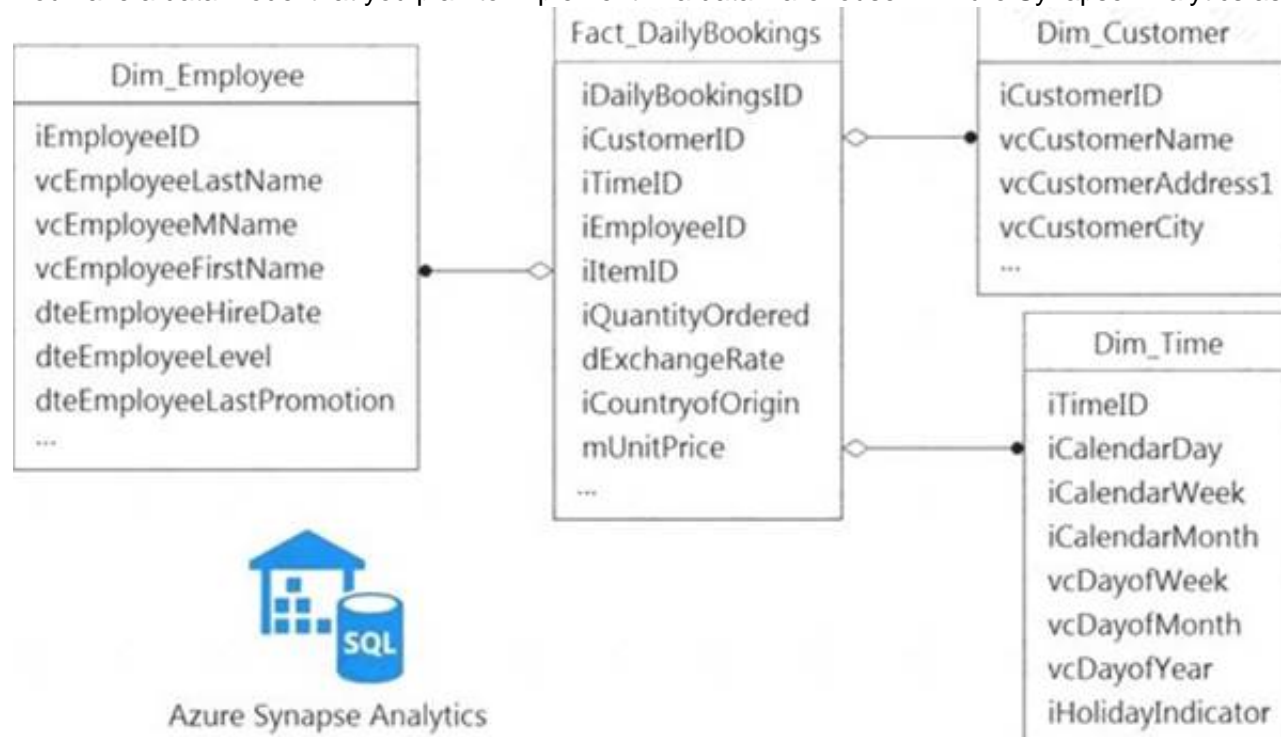Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/custom-deserializer

NEW QUESTION 105
- (Exam Topic 3)
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.
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.

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. Mastered
B. Not Mastered

Answer: A

Explanation:

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

NEW QUESTION 106
- (Exam Topic 3)
You are developing a solution using a Lambda architecture on Microsoft Azure. The data at test layer must meet the following requirements:
Data storage:
•Serve as a repository (or high volumes of large files in various formats.
•Implement optimized storage for big data analytics workloads.
•Ensure that data can be organized using a hierarchical structure. Batch processing:
•Use a managed solution for in-memory computation processing.
•Natively support Scala, Python, and R programming languages.
•Provide the ability to resize and terminate the cluster automatically. Analytical data store:
•Support parallel processing.
•Use columnar storage.

•Support SQL-based languages.
You need to identify the correct technologies to build the Lambda architecture.
Which technologies should you use? To answer, select the appropriate options in the answer area NOTE: Each correct selection is worth one point.

Architecture requirement	Technology
Data storage	▼ Azure SQL Database Azure Blob Storage Azure Cosmos DB Azure Data Lake Store
Batch processing	▼ HDInsight Spark HDInsight Hadoop Azure Databricks HDInsight Interactive Query
Analytical data store	▼ HDInsight HBase Azure SQL Data Warehouse Azure Analysis Services Azure Cosmos DB

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Data storage: Azure Data Lake Store
A key mechanism that allows Azure Data Lake Storage Gen2 to provide file system performance at object storage scale and prices is the addition of a hierarchical namespace. This allows the collection of objects/files within an account to be organized into a hierarchy of directories and nested subdirectories in the same way that the file system on your computer is organized. With the hierarchical namespace enabled, a storage account becomes capable of providing the scalability and cost-effectiveness of object storage, with file system semantics that are familiar to analytics engines and frameworks.
Batch processing: HD Insight Spark
Aparch Spark is an open-source, parallel-processing framework that supports in-memory processing to boost the performance of big-data analysis applications. HDInsight is a managed Hadoop service. Use it deploy and manage Hadoop clusters in Azure. For batch processing, you can use Spark, Hive, Hive LLAP, MapReduce.
Languages: R, Python, Java, Scala, SQL Analytic data store: SQL Data Warehouse
SQL Data Warehouse is a cloud-based Enterprise Data Warehouse (EDW) that uses Massively Parallel Processing (MPP).
SQL Data Warehouse stores data into relational tables with columnar storage. References:
https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-namespace https://docs.microsoft.com/en-us/azure/architecture/data-guide/technology-choices/batch-processing https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-overview-what-is

NEW QUESTION 109
- (Exam Topic 3)
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.

Report1: ▼

Avro
CSV
Parquet
TSV

Report2: ▼

Avro
CSV
Parquet
TSV

A. Mastered
B. Not Mastered

Answer: A

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

NEW QUESTION 114
- (Exam Topic 3)
You build an Azure Data Factory pipeline to move data from an Azure Data Lake Storage Gen2 container to a database in an Azure Synapse Analytics dedicated SQL pool.
Data in the container is stored in the following folder structure.
/in/{YYYY}/{MM}/{DD}/{HH}/{mm}
The earliest folder is /in/2021/01/01/00/00. The latest folder is /in/2021/01/15/01/45. You need to configure a pipeline trigger to meet the following requirements:

≫ Existing data must be loaded.

≫ Data must be loaded every 30 minutes.

≫ Late-arriving data of up to two minutes must he included in the load for the time at which the data
should have arrived.
How should you configure the pipeline trigger? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

Type: ▼

Event
On-demand
Schedule
Tumbling window

Additional properties: ▼

Prefix: /in/, Event: Blob created
Recurrence: 30 minutes, Start time: 2021-01-01T00:00
Recurrence: 30 minutes, Start time: 2021-01-01T00:00, Delay: 2 minutes
Recurrence: 32 minutes, Start time: 2021-01-15T01:45

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Box 1: Tumbling window
To be able to use the Delay parameter we select Tumbling window. Box 2:
Recurrence: 30 minutes, not 32 minutes
Delay: 2 minutes.
The amount of time to delay the start of data processing for the window. The pipeline run is started after the expected execution time plus the amount of delay. The delay defines how long the trigger waits past the due time before triggering a new run. The delay doesn't alter the window startTime.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/how-to-create-tumbling-window-trigger

NEW QUESTION 117
- (Exam Topic 3)
You have an Apache Spark DataFrame named temperatures. A sample of the data is shown in the following table.

Date	Temp
...	...
18-01-2021	3
19-01-2021	4
20-01-2021	2
21-01-2021	2
...	...

You need to produce the following table by using a Spark SQL query.

Year	JAN	FEB	MAR	APR	MAY
2019	2.3	4.1	5.2	7.6	9.2
2020	2.4	4.2	4.9	7.8	9.1
2021	2.6	5.3	3.4	7.9	9.5

How should you complete the query? 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.

Values

CAST
COLLATE
CONVERT
FLATTEN
PIVOT
UNPIVOT

Answer Area

```
SELECT * FROM (
  SELECT YEAR(Date) Year, MONTH(Date) Month, Temp
  FROM temperatures
  WHERE date BETWEEN DATE '2019-01-01' AND DATE '2021-08-31'
)
[_____] (
  AVG ( [_____] (Temp AS DECIMAL(4, 1)))

  FOR Month in (
    1 JAN, 2 FEB, 3 MAR, 4 APR, 5 MAY, 6 JUN,
    7 JUL, 8 AUG, 9 SEP, 10 OCT, 11 NOV, 12 DEC
        )
)
ORDER BY Year ASC
```

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Text Description automatically generated
Box 1: PIVOT
PIVOT rotates a table-valued expression by turning the unique values from one column in the expression into multiple columns in the output. And PIVOT runs aggregations where they're required on any remaining column values that are wanted in the final output.
Reference:
https://learnsql.com/cookbook/how-to-convert-an-integer-to-a-decimal-in-sql-server/ https://docs.microsoft.com/en-us/sql/t-sql/queries/from-using-pivot-and-unpivot

NEW QUESTION 118
- (Exam Topic 3)
You have an Azure Synapse Analytics SQL pool named Pool1 on a logical Microsoft SQL server named Server1.
You need to implement Transparent Data Encryption (TDE) on Pool1 by using a custom key named key1. Which five actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

Actions

Enable TDE on Pool1.

Assign a managed identity to Server1.

Configure key1 as the TDE protector for Server1.

Add key1 to the Azure key vault.

Create an Azure key vault and grant the managed identity permissions to the key vault.

Answer Area

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Graphical user interface, text, application Description automatically generated
Step 1: Assign a managed identity to Server1
You will need an existing Managed Instance as a prerequisite.
Step 2: Create an Azure key vault and grant the managed identity permissions to the vault Create Resource and setup Azure Key Vault.
Step 3: Add key1 to the Azure key vault
The recommended way is to import an existing key from a .pfx file or get an existing key from the vault. Alternatively, generate a new key directly in Azure Key Vault.
Step 4: Configure key1 as the TDE protector for Server1 Provide TDE Protector key
Step 5: Enable TDE on Pool1 Reference:
https://docs.microsoft.com/en-us/azure/azure-sql/managed-instance/scripts/transparent-data-encryption-byok-po

NEW QUESTION 123
- (Exam Topic 3)
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: A

Explanation:
Use the same definition as the EmployeeID column. Reference:
https://docs.microsoft.com/en-us/analysis-services/tabular-models/hierarchies-ssas-tabular

NEW QUESTION 125
- (Exam Topic 3)
You have an Azure Data lake Storage account that contains a staging zone.

You need to design a daily process to ingest incremental data from the staging zone, transform the data by executing an R script, and then insert the transformed data into a data warehouse in Azure Synapse Analytics.
Solution: You use an Azure Data Factory schedule trigger to execute a pipeline that executes an Azure Databricks notebook, and then inserts the data into the data warehouse.
Dow this meet the goal?

A. Yes
B. No

Answer: B

Explanation:
If you need to transform data in a way that is not supported by Data Factory, you can create a custom activity, not an Azure Databricks notebook, with your own data processing logic and use the activity in the pipeline. You can create a custom activity to run R scripts on your HDInsight cluster with R installed.
Reference:
https://docs.microsoft.com/en-US/azure/data-factory/transform-data

NEW QUESTION 126
- (Exam Topic 3)
You have an Azure Synapse Analytics dedicated SQL pool.
You need to Create a fact table named Table1 that will store sales data from the last three years. The solution must be optimized for the following query operations:
Show order counts by week.
• Calculate sales totals by region.
• Calculate sales totals by product.
• Find all the orders from a given month.
Which data should you use to partition Table1?

A. region
B. product
C. week
D. month

Answer: D

Explanation:
Table partitions enable you to divide your data into smaller groups of data. In most cases, table partitions are created on a date column.
Benefits to queries
Partitioning can also be used to improve query performance. A query that applies a filter to partitioned data can limit the scan to only the qualifying partitions. This method of filtering can avoid a full table scan and only scan a smaller subset of data. With the introduction of clustered columnstore indexes, the predicate elimination performance benefits are less beneficial, but in some cases there can be a benefit to queries.
For example, if the sales fact table is partitioned into 36 months using the sales date field, then queries that filter on the sale date can skip searching in partitions that don't match the filter.
Note: Benefits to loads
The primary benefit of partitioning in dedicated SQL pool is to improve the efficiency and performance of loading data by use of partition deletion, switching and merging. In most cases data is partitioned on a date column that is closely tied to the order in which the data is loaded into the SQL pool. One of the greatest benefits of using partitions to maintain data is the avoidance of transaction logging. While simply inserting, updating, or deleting data can be the most straightforward approach, with a little thought and effort, using partitioning during your load process can substantially improve performance.
Reference:
https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-partitio

NEW QUESTION 128
- (Exam Topic 3)
You have an Azure Stream Analytics query. The query returns a result set that contains 10,000 distinct values for a column named clusterID.
You monitor the Stream Analytics job and discover high latency. You need to reduce the latency.
Which two actions should you perform? Each correct answer presents a complete solution. NOTE: Each correct selection is worth one point.

A. Add a pass-through query.
B. Add a temporal analytic function.
C. Scale out the query by using PARTITION BY.
D. Convert the query to a reference query.
E. Increase the number of streaming units.

Answer: CE

Explanation:
C: Scaling a Stream Analytics job takes advantage of partitions in the input or output. Partitioning lets you divide data into subsets based on a partition key. A process that consumes the data (such as a Streaming Analytics job) can consume and write different partitions in parallel, which increases throughput.
E: Streaming Units (SUs) represents the computing resources that are allocated to execute a Stream Analytics job. The higher the number of SUs, the more CPU and memory resources are allocated for your job. This capacity lets you focus on the query logic and abstracts the need to manage the hardware to run your Stream Analytics job in a timely manner.
References:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-parallelization https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-streaming-unit-consumption

NEW QUESTION 133
- (Exam Topic 3)
A company has a real-time data analysis solution that is hosted on Microsoft Azure. The solution uses Azure Event Hub to ingest data and an Azure Stream Analytics cloud job to analyze the data. The cloud job is configured to use 120 Streaming Units (SU).
You need to optimize performance for the Azure Stream Analytics job.

Which two actions should you perform? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.

A. Implement event ordering.
B. Implement Azure Stream Analytics user-defined functions (UDF).
C. Implement query parallelization by partitioning the data output.
D. Scale the SU count for the job up.
E. Scale the SU count for the job down.
F. Implement query parallelization by partitioning the data input.

Answer: DF

Explanation:
Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-parallelization

NEW QUESTION 135
- (Exam Topic 3)
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. zone-redundant storage (ZRS)
B. read-access geo-redundant storage (RA-GRS)
C. locally-redundant storage (LRS)
D. geo-redundant storage (GRS)

Answer: B

Explanation:
Geo-redundant storage (with GRS or GZRS) replicates your data to another physical location in the secondary region to protect against regional outages.
However, that data is available to be read only if the customer or Microsoft initiates a failover from the primary to secondary region. When you enable read access to the secondary region, your data is available to be read at all times, including in a situation where the primary region becomes unavailable.
Reference:
https://docs.microsoft.com/en-us/azure/storage/common/storage-redundancy

NEW QUESTION 137
- (Exam Topic 3)
You plan to create a table in an Azure Synapse Analytics dedicated SQL pool.
Data in the table will be retained for five years. Once a year, data that is older than five years will be deleted. You need to ensure that the data is distributed evenly across partitions. The solution must minimize the
amount of time required to delete old data.
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.

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Box 1: HASH
Box 2: OrderDateKey
In most cases, table partitions are created on a date column.
A way to eliminate rollbacks is to use Metadata Only operations like partition switching for data management. For example, rather than execute a DELETE statement to delete all rows in a table where the order_date was in October of 2001, you could partition your data early. Then you can switch out the partition with data for an empty partition from another table.
Reference:
https://docs.microsoft.com/en-us/sql/t-sql/statements/create-table-azure-sql-data-warehouse https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/best-practices-dedicated-sql-pool

NEW QUESTION 140
- (Exam Topic 3)
You have an Azure Storage account that generates 200,000 new files daily. The file names have a format of {YYYY}/{MM}/{DD}/{HH}/{CustomerID}.csv.
You need to design an Azure Data Factory solution that will load new data from the storage account to an Azure Data Lake once hourly. The solution must minimize load times and costs.
How should you configure the solution? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

Load methodology:
- Full Load
- Incremental Load
- Load individual files as they arrive

Trigger:
- Fixed schedule
- New file
- Tumbling window

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Table Description automatically generated
Box 1: Incremental load Box 2: Tumbling window
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.
Timeline Description automatically generated

Tell me the count of tweets per time zone every 10 seconds

A 10-second Tumbling Window
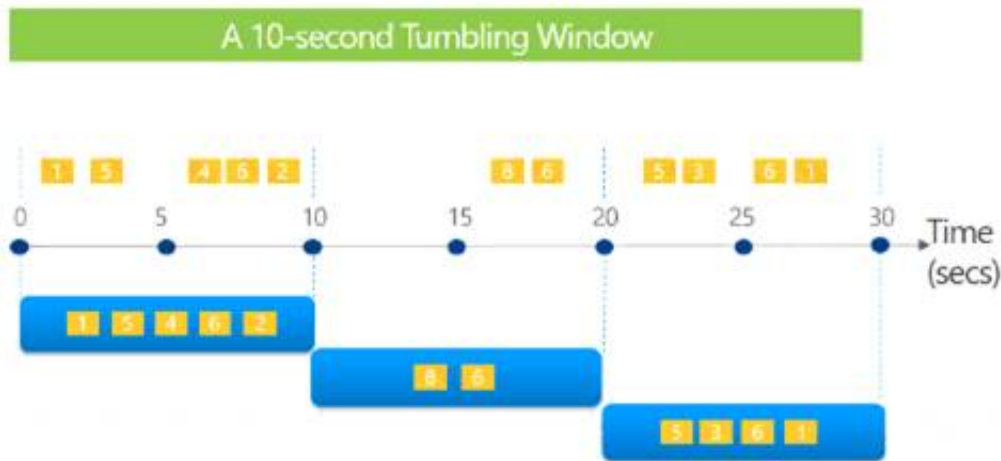
```
SELECT TimeZone, COUNT(*) AS Count
FROM TwitterStream TIMESTAMP BY CreatedAt
GROUP BY TimeZone, TumblingWindow(second,10)
```

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

NEW QUESTION 144
- (Exam Topic 3)
You have an Azure Synapse Analytics Apache Spark pool named Pool1.
You plan to load JSON files from an Azure Data Lake Storage Gen2 container into the tables in Pool1. The structure and data types vary by file.
You need to load the files into the tables. The solution must maintain the source data types. What should you do?

A. Use a Get Metadata activity in Azure Data Factory.

B. Use a Conditional Split transformation in an Azure Synapse data flow.
C. Load the data by using the OPEHROwset Transact-SQL command in an Azure Synapse Anarytics serverless SQL pool.
D. Load the data by using PySpark.

Answer: A

Explanation:
Serverless SQL pool can automatically synchronize metadata from Apache Spark. A serverless SQL pool database will be created for each database existing in serverless Apache Spark pools.
Serverless SQL pool enables you to query data in your data lake. It offers a T-SQL query surface area that accommodates semi-structured and unstructured data queries.
To support a smooth experience for in place querying of data that's located in Azure Storage files, serverless SQL pool uses the OPENROWSET function with additional capabilities.
The easiest way to see to the content of your JSON file is to provide the file URL to the OPENROWSET function, specify csv FORMAT.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/query-json-files https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/query-data-storage

NEW QUESTION 149
- (Exam Topic 3)
You have an activity in an Azure Data Factory pipeline. The activity calls a stored procedure in a data warehouse in Azure Synapse Analytics and runs daily.
You need to verify the duration of the activity when it ran last. What should you use?

A. activity runs in Azure Monitor
B. Activity log in Azure Synapse Analytics
C. the sys.dm_pdw_wait_stats data management view in Azure Synapse Analytics
D. an Azure Resource Manager template

Answer: A

Explanation:
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/monitor-visually

NEW QUESTION 154
- (Exam Topic 3)
You have an Azure Data Lake Storage Gen2 account named account1 that stores logs as shown in the following table.

Type	Designated retention period
Application	360 days
Infrastructure	60 days

You do not expect that the logs will be accessed during the retention periods.
You need to recommend a solution for account1 that meets the following requirements:
≫ Automatically deletes the logs at the end of each retention period
≫ Minimizes storage costs
What should you include in the recommendation? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

To minimize storage costs: ▼

Store the infrastructure logs and the application logs in the Archive access tier
Store the infrastructure logs and the application logs in the Cool access tier
Store the infrastructure logs in the Cool access tier and the application logs in the Archive access tier

To delete logs automatically: ▼

Azure Data Factory pipelines
Azure Blob storage lifecycle management rules
Immutable Azure Blob storage time-based retention policies

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Table Description automatically generated
Box 1: Store the infrastructure logs in the Cool access tier and the application logs in the Archive access tier For infrastructure logs: Cool tier - An online tier optimized for storing data that is infrequently accessed or
modified. Data in the cool tier should be stored for a minimum of 30 days. The cool tier has lower storage costs and higher access costs compared to the hot tier.
For application logs: Archive tier - An offline tier optimized for storing data that is rarely accessed, and that has flexible latency requirements, on the order of hours.
Data in the archive tier should be stored for a minimum of 180 days.
Box 2: Azure Blob storage lifecycle management rules
Blob storage lifecycle management offers a rule-based policy that you can use to transition your data to the desired access tier when your specified conditions are

met. You can also use lifecycle management to expire data at the end of its life.
Reference:
https://docs.microsoft.com/en-us/azure/storage/blobs/access-tiers-overview

NEW QUESTION 157
- (Exam Topic 3)
You are designing an anomaly detection solution for streaming data from an Azure IoT hub. The solution must meet the following requirements:

≫ Send the output to Azure Synapse.

≫ Identify spikes and dips in time series data.

≫ Minimize development and configuration effort. Which should you include in the solution?

A. Azure Databricks
B. Azure Stream Analytics
C. Azure SQL Database

Answer: B

Explanation:
You can identify anomalies by routing data via IoT Hub to a built-in ML model in Azure Stream Analytics. Reference:
https://docs.microsoft.com/en-us/learn/modules/data-anomaly-detection-using-azure-iot-hub/

NEW QUESTION 158
- (Exam Topic 3)
You have an Azure subscription that contains the following resources:

≫ An Azure Active Directory (Azure AD) tenant that contains a security group named Group1

≫ An Azure Synapse Analytics SQL pool named Pool1
You need to control the access of Group1 to specific columns and rows in a table in Pool1.
Which Transact-SQL commands should you use? To answer, select the appropriate options in the answer area.

To control access to the columns:	▼
	CREATE CRYPTOGRAPHIC PROVIDER
	CREATE PARTITION FUNCTION
	CREATE SECURITY POLICY
	GRANT

To control access to the rows:	▼
	CREATE CRYPTOGRAPHIC PROVIDER
	CREATE PARTITION FUNCTION
	CREATE SECURITY POLICY
	GRANT

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Text Description automatically generated
Box 1: GRANT
You can implement column-level security with the GRANT T-SQL statement. Box 2: CREATE SECURITY POLICY
Implement Row Level Security by using the CREATE SECURITY POLICY Transact-SQL statement Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/column-level-security

NEW QUESTION 159
- (Exam Topic 3)
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. Assign Azure AD security groups to Azure Data Lake Storage.
B. Configure end-user authentication for the Azure Data Lake Storage account.
C. Configure service-to-service authentication for the Azure Data Lake Storage account.
D. Create security groups in Azure Active Directory (Azure AD) and add project members.
E. Configure access control lists (ACL) for the Azure Data Lake Storage account.

Answer: ADE

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

NEW QUESTION 162
- (Exam Topic 3)
A company purchases IoT devices to monitor manufacturing machinery. The company uses an IoT appliance to communicate with the IoT devices.
The company must be able to monitor the devices in real-time. You need to design the solution.
What should you recommend?

A. Azure Stream Analytics cloud job using Azure PowerShell
B. Azure Analysis Services using Azure Portal
C. Azure Data Factory instance using Azure Portal
D. Azure Analysis Services using Azure PowerShell

Answer: A

Explanation:
Stream Analytics is a cost-effective event processing engine that helps uncover real-time insights from devices, sensors, infrastructure, applications and data quickly and easily.
Monitor and manage Stream Analytics resources with Azure PowerShell cmdlets and powershell scripting that execute basic Stream Analytics tasks.
Reference:
https://cloudblogs.microsoft.com/sqlserver/2014/10/29/microsoft-adds-iot-streaming-analytics-data-production-a

NEW QUESTION 166
- (Exam Topic 3)
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 theGet-AzOperationalInsightsSearchResult PowerShell cmdlet.
D. On the master database, execute a query against the sys.dm_pdw_nodes_os_performance_counters dynamic management view.

Answer: A

Explanation:
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

NEW QUESTION 169
- (Exam Topic 3)
You are designing a highly available Azure Data Lake Storage solution that will include geo-zone-redundant storage (GZRS).
You need to monitor for replication delays that can affect the recovery point objective (RPO). What should you include in the monitoring solution?

A. availability
B. Average Success E2E Latency
C. 5xx: Server Error errors
D. Last Sync Time

Answer: D

Explanation:
Because geo-replication is asynchronous, it is possible that data written to the primary region has not yet been written to the secondary region at the time an outage occurs. The Last Sync Time property indicates the last time that data from the primary region was written successfully to the secondary region. All writes made to the primary region before the last sync time are available to be read from the secondary location. Writes made to the primary region after the last sync time property may or may not be available for reads yet.
Reference:
https://docs.microsoft.com/en-us/azure/storage/common/last-sync-time-get

NEW QUESTION 172
- (Exam Topic 3)
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 Datiabricks 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 tiles can be queried quickly and that the data type information is retained.
What should you recommend?

A. Parquet
B. Avro
C. CSV
D. JSON

Answer: A

Explanation:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-define-outputs

NEW QUESTION 177
- (Exam Topic 3)

You are designing an Azure Synapse solution that will provide a query interface for the data stored in an Azure Storage account. The storage account is only accessible from a virtual network.
You need to recommend an authentication mechanism to ensure that the solution can access the source data. What should you recommend?

A. a managed identity
B. anonymous public read access
C. a shared key

Answer: A

Explanation:
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-exa

NEW QUESTION 178
- (Exam Topic 3)
You have an on-premises data warehouse that includes the following fact tables. Both tables have the following columns: DateKey, ProductKey, RegionKey. There are 120 unique product keys and 65 unique region keys.

Table	Comments
Sales	The table is 600 GB in size. DateKey is used extensively in the WHERE clause in queries. ProductKey is used extensively in join operations. RegionKey is used for grouping. Severity-five percent of records relate to one of 40 regions.
Invoice	The table is 6 GB in size. DateKey and ProductKey are used extensively in the WHERE clause in queries. RegionKey is used for grouping.

Queries that use the data warehouse take a long time to complete.
You plan to migrate the solution to use Azure Synapse Analytics. You need to ensure that the Azure-based solution optimizes query performance and minimizes processing skew.
What should you recommend? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Box 1: Hash-distributed
Box 2: ProductKey
ProductKey is used extensively in joins.
Hash-distributed tables improve query performance on large fact tables.
Box 3: Round-robin
Box 4: RegionKey
Round-robin tables are useful for improving loading speed.
Consider using the round-robin distribution for your table in the following scenarios:
➤ When getting started as a simple starting point since it is the default
➤ If there is no obvious joining key
➤ If there is not good candidate column for hash distributing the table
➤ If the table does not share a common join key with other tables
➤ If the join is less significant than other joins in the query
➤ When the table is a temporary staging table
Note: A distributed table appears as a single table, but the rows are actually stored across 60 distributions. The rows are distributed with a hash or round-robin algorithm.
Reference:
https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-tables-distribute

NEW QUESTION 179
- (Exam Topic 3)
You are designing an Azure Synapse Analytics dedicated SQL pool.
Groups will have access to sensitive data in the pool as shown in the following table.

Name	Enhanced access
Executives	No access to sensitive data
Analysts	Access to in-region sensitive data
Engineers	Access to all numeric sensitive data

You have policies for the sensitive data. The policies vary be region as shown in the following table.

Region	Data considered sensitive
RegionA	Financial, Personally Identifiable Information (PII)
RegionB	Financial, Personally Identifiable Information (PII), medical
RegionC	Financial, medical

You have a table of patients for each region. The tables contain the following potentially sensitive columns.

Name	Sensitive data	Description
CardOnFile	Financial	Debit/credit card number for charges
Height	Medical	Patient's height in cm
ContactEmail	PII	Email address for secure communications

You are designing dynamic data masking to maintain compliance.
For each of the following statements, select Yes if the statement is true. Otherwise, select No.
NOTE: Each correct selection is worth one point.

Statements	Yes	No
Analysts in RegionA require dynamic data masking rules for [Patients_RegionA].	○	○
Engineers in RegionC require a dynamic data masking rule for [Patients_RegionA], [Height]	○	○
Engineers in RegionB require a dynamic data masking rule for [Patients_RegionB], [Height]	○	○

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Text Description automatically generated
Reference:
https://docs.microsoft.com/en-us/azure/azure-sql/database/dynamic-data-masking-overview

NEW QUESTION 180
- (Exam Topic 3)
You have an Azure subscription that contains an Azure Synapse Analytics workspace named workspace1. Workspace1 connects to an Azure DevOps repository named repo1. Repo1 contains a collaboration branch named main and a development branch named branch1. Branch1 contains an Azure Synapse pipeline named pipeline1.
In workspace1, you complete testing of pipeline1. You need to schedule pipeline1 to run daily at 6 AM.
Which four actions should you perform in sequence? To answer, move the appropriate actions from the list of actions 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.

Actions   **Answer Area**

| Create a new branch in Repo1. |

| Merge the changes from branch1 into main. |

| Associate the schedule trigger with pipeline1. | (>)
                                                        (<)
| Switch to Synapse live mode. |

| Create a schedule trigger. |

| Publish the contents of main. |

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Timeline Description automatically generated

NEW QUESTION 183
- (Exam Topic 3)
You are designing an Azure Data Lake Storage Gen2 structure for telemetry data from 25 million devices distributed across seven key geographical regions. Each minute, the devices will send a JSON payload of metrics to Azure Event Hubs.
You need to recommend a folder structure for the data. The solution must meet the following requirements:

≫ Data engineers from each region must be able to build their own pipelines for the data of their respective region only.

≫ The data must be processed at least once every 15 minutes for inclusion in Azure Synapse Analytics serverless SQL pools.
How should you recommend completing the structure? 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.

Values   **Answer Area**

| {deviceID} |

| {mm}/{HH}/{DD}/{MM}/{YYYY} | / [Value   ] / [Value   ] / [Value   ] .json

| {regionID}/{deviceID} |

| {regionID}/raw |

| {YYYY}/{MM}/{DD}/{HH} |

| {YYYY}/{MM}/{DD}/{HH}/{mm} |

| raw/{deviceID} |

| raw/{regionID} |

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Box 1: {YYYY}/{MM}/{DD}/{HH}
Date Format [optional]: if the date token is used in the prefix path, you can select the date format in which your files are organized. Example: YYYY/MM/DD
Time Format [optional]: if the time token is used in the prefix path, specify the time format in which your files are organized. Currently the only supported value is HH.
Box 2: {regionID}/raw
Data engineers from each region must be able to build their own pipelines for the data of their respective region only.
Box 3: {deviceID} Reference:
https://github.com/paolosalvatori/StreamAnalyticsAzureDataLakeStore/blob/master/README.md

NEW QUESTION 185

- (Exam Topic 3)
You have an enterprise data warehouse in Azure Synapse Analytics named DW1 on a server named Server1. You need to verify whether the size of the transaction log file for each distribution of DW1 is smaller than 160 GB.
What should you do?

A. On the master database, execute a query against the sys.dm_pdw_nodes_os_performance_counters dynamic management view.
B. From Azure Monitor in the Azure portal, execute a query against the logs of DW1.
C. On DW1, execute a query against the sys.database_files dynamic management view.
D. Execute a query against the logs of DW1 by using theGet-AzOperationalInsightSearchResult PowerShell cmdlet.

Answer: A

Explanation:
The following query returns the transaction log size on each distribution. If one of the log files is reaching 160 GB, you should consider scaling up your instance or limiting your transaction size.
-- Transaction log size SELECT
instance_name as distribution_db, cntr_value*1.0/1048576 as log_file_size_used_GB, pdw_node_id
FROM sys.dm_pdw_nodes_os_performance_counters WHERE
instance_name like 'Distribution_%'
AND counter_name = 'Log File(s) Used Size (KB)' References:
https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-manage-monitor

NEW QUESTION 186
- (Exam Topic 3)
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?

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

Answer: B

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-overviewmana

NEW QUESTION 190
- (Exam Topic 3)
You have an Azure Data Lake Storage account that contains a staging zone.
You need to design a dairy process to ingest incremental data from the staging zone, transform the data by executing an R script, and then insert the transformed data into a data warehouse in Azure Synapse Analytics.
Solution: You use an Azure Data Factory schedule trigger to execute a pipeline that copies the data to a staging table in the data warehouse, and then uses a stored procedure to execute the R script.
Does this meet the goal?

A. Yes
B. No

Answer: A

Explanation:
If you need to transform data in a way that is not supported by Data Factory, you can create a custom activity with your own data processing logic and use the activity in the pipeline.
Note: You can use data transformation activities in Azure Data Factory and Synapse pipelines to transform and process your raw data into predictions and insights at scale.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/transform-data

NEW QUESTION 195
- (Exam Topic 3)
You have an Azure Databricks resource.
You need to log actions that relate to changes in compute for the Databricks resource. Which Databricks services should you log?

A. clusters
B. workspace
C. DBFS
D. SSHE lobs

Answer: B

Explanation:
Cloud Provider Infrastructure Logs.Databricks logging allows security and admin teams to demonstrate conformance to data governance standards within or from a Databricks workspace. Customers, especially in the regulated industries, also need records on activities like:– User access control to cloud data storage– Cloud Identity and Access Management roles– User access to cloud network and compute
Azure Databricks offers three distinct workloads on several VM Instances tailored for your data analytics workflow—the Jobs Compute and Jobs Light Compute

workloads make it easy for data engineers to build and execute jobs, and the All-Purpose Compute workload makes it easy for data scientists to explore, visualize, manipulate, and share data and insights interactively.

NEW QUESTION 199
- (Exam Topic 3)
You need to design a solution that will process streaming data from an Azure Event Hub and output the data to Azure Data Lake Storage. The solution must ensure that analysts can interactively query the streaming data.
What should you use?

A. event triggers in Azure Data Factory
B. Azure Stream Analytics and Azure Synapse notebooks
C. Structured Streaming in Azure Databricks
D. Azure Queue storage and read-access geo-redundant storage (RA-GRS)

Answer: C

Explanation:
Apache Spark Structured Streaming is a fast, scalable, and fault-tolerant stream processing API. You can use it to perform analytics on your streaming data in near real-time.
With Structured Streaming, you can use SQL queries to process streaming data in the same way that you would process static data.
Azure Event Hubs is a scalable real-time data ingestion service that processes millions of data in a matter of seconds. It can receive large amounts of data from multiple sources and stream the prepared data to Azure Data Lake or Azure Blob storage.
Azure Event Hubs can be integrated with Spark Structured Streaming to perform the processing of messages in near real-time. You can query and analyze the processed data as it comes by using a Structured Streaming query and Spark SQL.
Reference:
https://k21academy.com/microsoft-azure/data-engineer/structured-streaming-with-azure-event-hubs/

NEW QUESTION 200
- (Exam Topic 3)
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:
• After initial processing, the HR department data will be retained for seven years.
• 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.

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Answer Area

HR: Archive storage after one day and delete storage after 2.555 days. ▼

Operations: Cool storage after 180 days. ▼

NEW QUESTION 204
- (Exam Topic 3)
You are monitoring an Azure Stream Analytics job.
You discover that the Backlogged Input Events metric is increasing slowly and is consistently non-zero. You need to ensure that the job can handle all the events.
What should you do?

A. Change the compatibility level of the Stream Analytics job.
B. Increase the number of streaming units (SUs).
C. Remove any named consumer groups from the connection and use $default.
D. Create an additional output stream for the existing input stream.

Answer: B

Explanation:
Backlogged Input Events: Number of input events that are backlogged. A non-zero value for this metric implies that your job isn't able to keep up with the number of incoming events. If this value is slowly increasing or consistently non-zero, you should scale out your job. You should increase the Streaming Units.
Note: Streaming Units (SUs) represents the computing resources that are allocated to execute a Stream Analytics job. The higher the number of SUs, the more CPU and memory resources are allocated for your job.
Reference:
https://docs.microsoft.com/bs-cyrl-ba/azure/stream-analytics/stream-analytics-monitoring

NEW QUESTION 208
- (Exam Topic 3)
You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Table1. You have files that are ingested and loaded into an Azure Data Lake Storage Gen2 container named
container1.
You plan to insert data from the files into Table1 and azure Data Lake Storage Gen2 container named container1.

You plan to insert data from the files into Table1 and transform the data. Each row of data in the files will produce one row in the serving layer of Table1.
You need to ensure that when the source data files are loaded to container1, the DateTime is stored as an additional column in Table1.
Solution: You use a dedicated SQL pool to create an external table that has a additional DateTime column. Does this meet the goal?

A. Yes
B. No

Answer: A

NEW QUESTION 210
- (Exam Topic 3)
You are building an Azure Stream Analytics job to retrieve game data.
You need to ensure that the job returns the highest scoring record for each five-minute time interval of each game.
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.

```
SELECT  [                                     ▼ ]  as HighestScore
          Collect(Score)
          CollectTop(1) OVER(ORDER BY Score Desc)
          Game, MAX(Score)
          TopOne() OVER(PARTITION BY Game ORDER BY Score Desc)


FROM input TIMESTAMP BY CreatedAt

GROUP BY  [                                   ▼ ]
          Game
          Hopping(minute,5)
          Tumbling(minute,5)
          Windows(TumblingWindow(minute,5),Hopping(minute,5))
```
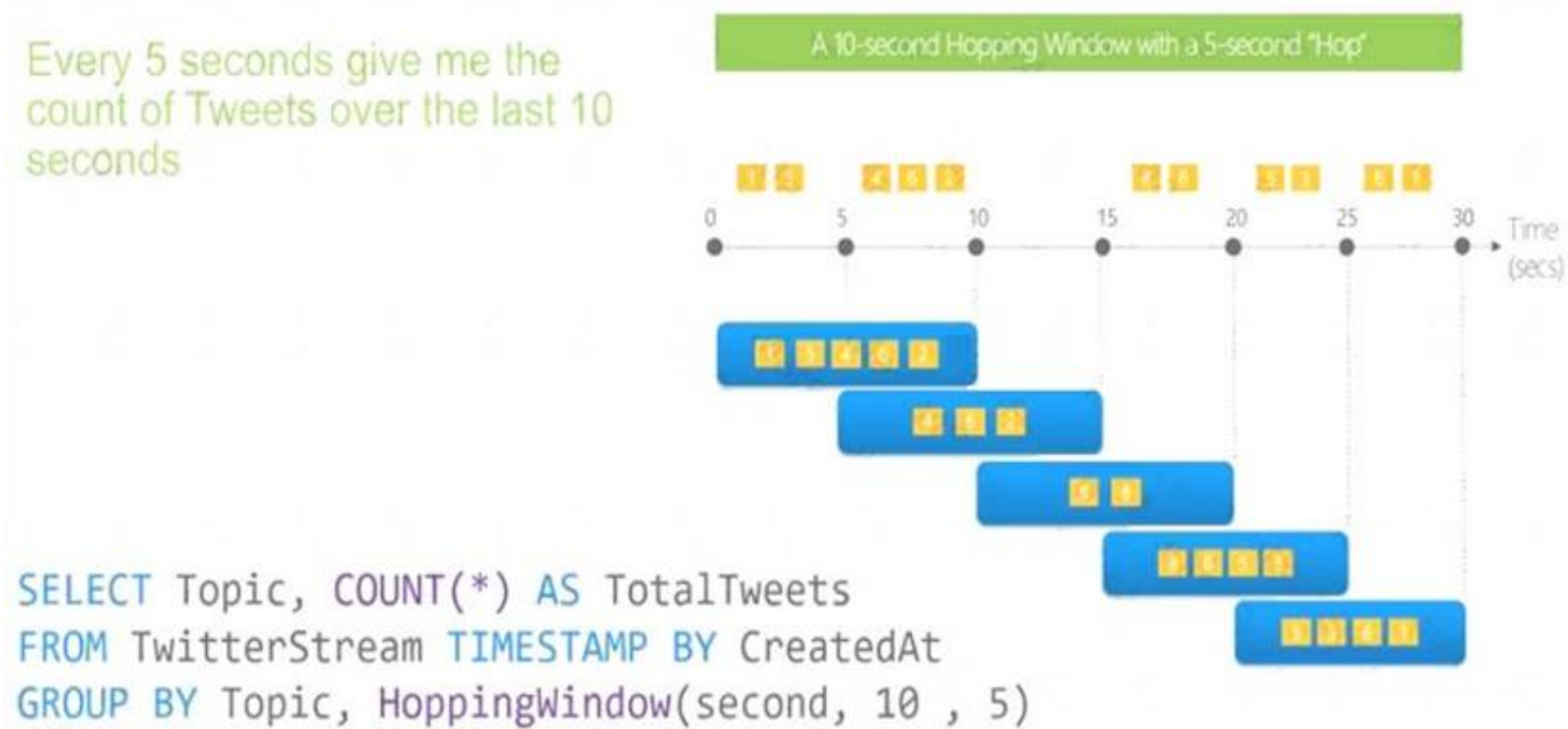
A. Mastered
B. Not Mastered

Answer: A

Explanation:
Box 1: TopOne OVER(PARTITION BY Game ORDER BY Score Desc)
TopOne returns the top-rank record, where rank defines the ranking position of the event in the window according to the specified ordering. Ordering/ranking is based on event columns and can be specified in ORDER BY clause.
Box 2: Hopping(minute,5)
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 and be emitted more often than the window size. 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.
A picture containing timeline Description automatically generated

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

NEW QUESTION 214
- (Exam Topic 3)
You have an Azure subscription that contains a logical Microsoft SQL server named Server1. Server1 hosts an Azure Synapse Analytics SQL dedicated pool named Pool1.
You need to recommend a Transparent Data Encryption (TDE) solution for Server1. The solution must meet the following requirements:

> Track the usage of encryption keys.

> Maintain the access of client apps to Pool1 in the event of an Azure datacenter outage that affects the availability of the encryption keys.

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

NOTE: Each correct selection is worth one point.

To track encryption key usage: ▼

| Always Encrypted |
| TDE with customer-managed keys |
| TDE with platform-managed keys |

To maintain client app access in
the event of a datacenter outage: ▼

| Create and configure Azure key vaults in two Azure regions. |
| Enable Advanced Data Security on Server1. |
| Implement the client apps by using a Microsoft .NET Framework data provider. |

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Box 1: TDE with customer-managed keys
Customer-managed keys are stored in the Azure Key Vault. You can monitor how and when your key vaults are accessed, and by whom. You can do this by enabling logging for Azure Key Vault, which saves information in an Azure storage account that you provide.
Box 2: Create and configure Azure key vaults in two Azure regions
The contents of your key vault are replicated within the region and to a secondary region at least 150 miles away, but within the same geography to maintain high durability of your keys and secrets.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/security/workspaces-encryption https://docs.microsoft.com/en-us/azure/key-vault/general/logging

NEW QUESTION 218
- (Exam Topic 3)
You have an enterprise data warehouse in Azure Synapse Analytics.
You need to monitor the data warehouse to identify whether you must scale up to a higher service level to accommodate the current workloads
Which is the best metric to monitor?
More than one answer choice may achieve the goal. Select the BEST answer.

A. Data 10 percentage
B. CPU percentage
C. DWU used
D. DWU percentage

Answer: D

NEW QUESTION 221
- (Exam Topic 3)
You have an Azure Data Lake Storage account that contains a staging zone.
You need to design a daily process to ingest incremental data from the staging zone, transform the data by executing an R script, and then insert the transformed data into a data warehouse in Azure Synapse Analytics.
Solution: You use an Azure Data Factory schedule trigger to execute a pipeline that executes mapping data Flow, and then inserts the data info the data warehouse.
Does this meet the goal?

A. Yes
B. No

Answer: B

Explanation:
If you need to transform data in a way that is not supported by Data Factory, you can create a custom activity, not a mapping flow,5 with your own data processing logic and use the activity in the pipeline. You can create a custom activity to run R scripts on your HDInsight cluster with R installed.
Reference:
https://docs.microsoft.com/en-US/azure/data-factory/transform-data

NEW QUESTION 222
- (Exam Topic 3)
You are designing a data mart for the human resources (MR) department at your company. The data mart will contain information and employee transactions.
From a source system you have a flat extract that has the following fields:
• EmployeeID

• FirstName
• LastName
• Recipient
• GrossArnount
• TransactionID
• GovernmentID
• NetAmountPaid
• TransactionDate

You need to design a start schema data model in an Azure Synapse analytics dedicated SQL pool for the data mart.
Which two tables should you create? Each Correct answer present part of the solution.

A. a dimension table for employee
B. a fabric for Employee
C. a dimension table far EmployeeTransaction
D. a dimension table for Transaction
E. a fact table for Transaction

Answer: AE

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

NEW QUESTION 226
- (Exam Topic 3)
You are designing a star schema for a dataset that contains records of online orders. Each record includes an order date, an order due date, and an order ship date.
You need to ensure that the design provides the fastest query times of the records when querying for arbitrary date ranges and aggregating by fiscal calendar attributes.
Which two actions should you perform? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.

A. Create a date dimension table that has a DateTime key.
B. Use built-in SQL functions to extract date attributes.
C. Create a date dimension table that has an integer key in the format of yyyymmdd.
D. In the fact table, use integer columns for the date fields.
E. Use DateTime columns for the date fields.

Answer: BD

NEW QUESTION 231
- (Exam Topic 3)
You are designing an application that will store petabytes of medical imaging data
When the data is first created, the data will be accessed frequently during the first week. After one month, the data must be accessible within 30 seconds, but files will be accessed infrequently. After one year, the data will be accessed infrequently but must be accessible within five minutes.
You need to select a storage strategy for the data. The solution must minimize costs.
Which storage tier should you use for each time frame? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

First week:
| Archive |
| Cool |
| Hot |

After one month:
| Archive |
| Cool |
| Hot |

After one year:
| Archive |
| Cool |
| Hot |

A. Mastered
B. Not Mastered

Answer: A

Explanation:
First week: Hot
Hot - Optimized for storing data that is accessed frequently. After one month: Cool
Cool - Optimized for storing data that is infrequently accessed and stored for at least 30 days.
After one year: Cool

NEW QUESTION 233
- (Exam Topic 3)
You have an Azure Synapse Analytics dedicated SQL pool.
You run PDW_SHOWSPACEUSED(dbo,FactInternetSales'); and get the results shown in the following table.

ROWS	RESERVED_SPACE	DATA_SPACE	INDEX_SPACE	UNUSED_SPACE	PDW_NODE_ID	DISTRIBUTION_ID
694	2776	616	48	2112	1	1
407	2704	576	48	2080	1	2
53	2376	512	16	1848	1	3
58	2376	512	16	1848	1	4
168	2632	528	32	2072	1	5
195	2696	536	32	2128	1	6
5995	3464	1424	32	2008	1	7
0	2232	496	0	1736	1	8
264	2576	544	40	1992	1	9
3008	3016	960	32	2024	1	10
--	--	--	--	--	--	--
1550	2832	752	48	2032	1	50
1238	2832	696	40	2096	1	51
192	2632	528	32	2072	1	52
1127	2768	680	48	2040	1	53
1244	3032	704	64	2264	1	54
409	2632	568	32	2032	1	55
0	2232	496	0	1736	1	56
1437	2832	728	40	2064	1	57
0	2232	496	0	1736	1	58
584	2632	560	32	2040	1	59
225	2768	544	40	2184	1	60

Which statement accurately describes the dbo,FactInternetSales table?

A. The table contains less than 1,000 rows.
B. All distributions contain data.
C. The table is skewed.
D. The table uses round-robin distribution.

Answer: B

Explanation:
Data skew means the data is not distributed evenly across the distributions. Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribu

NEW QUESTION 238
- (Exam Topic 3)
You configure monitoring for a Microsoft Azure SQL Data Warehouse implementation. The implementation uses PolyBase to load data from comma-separated value (CSV) files stored in Azure Data Lake Gen 2 using an external table.
Files with an invalid schema cause errors to occur. You need to monitor for an invalid schema error. For which error should you monitor?

A. EXTERNAL TABLE access failed due to internal error: 'Java exception raised on call to HdfsBridge_Connect:
Error[com.microsoft.polybase.client.KerberosSecureLogin] occurred while accessing external files.'
B. EXTERNAL TABLE access failed due to internal error: 'Java exception raised on call to HdfsBridge_Connect: Error [No FileSystem for scheme: wasbs] occurred while accessing external file.'
C. Cannot execute the query "Remote Query" against OLE DB provider "SQLNCLI11": for linked server "(null)", Query aborted- the maximum reject threshold (orows) was reached while regarding from an external source: 1 rows rejected out of total 1 rows processed.
D. EXTERNAL TABLE access failed due to internal error: 'Java exception raised on call to HdfsBridge_Connect: Error [Unable to instantiate LoginClass] occurredwhile accessing external files.'

Answer: C

Explanation:
Customer Scenario:
SQL Server 2016 or SQL DW connected to Azure blob storage. The CREATE EXTERNAL TABLE DDL points to a directory (and not a specific file) and the directory contains files with different schemas.
SSMS Error:
Select query on the external table gives the following error: Msg 7320, Level 16, State 110, Line 14
Cannot execute the query "Remote Query" against OLE DB provider "SQLNCLI11" for linked server "(null)". Query aborted-- the maximum reject threshold (0 rows) was reached while reading from an external source: 1 rows rejected out of total 1 rows processed.
Possible Reason:
The reason this error happens is because each file has different schema. The PolyBase external table DDL when pointed to a directory recursively reads all the files in that directory. When a column or data type mismatch happens, this error could be seen in SSMS.
Possible Solution:
If the data for each table consists of one file, then use the filename in the LOCATION section prepended by the directory of the external files. If there are multiple files per table, put each set of files into different directories in Azure Blob Storage and then you can point LOCATION to the directory instead of a particular file. The latter suggestion is the best practices recommended by SQLCAT even if you have one file per table.

NEW QUESTION 241
- (Exam Topic 3)
You have an enterprise data warehouse in Azure Synapse Analytics.

Using PolyBase, you create an external table named [Ext].[Items] to query Parquet files stored in Azure Data Lake Storage Gen2 without importing the data to the data warehouse.
The external table has three columns.
You discover that the Parquet files have a fourth column named ItemID.
Which command should you run to add the ItemID column to the external table?

```
A.  ALTER EXTERNAL TABLE [Ext].[Items]
       ADD [ItemID] int;

B.  DROP EXTERNAL FILE FORMAT parquetfile1;
    CREATE EXTERNAL FILE FORMAT parquetfile1
    WITH (
        FORMAT_TYPE = PARQUET,
        DATA_COMPRESSION = 'org.apache.hadoop.io.compress.SnappyCodec'
    );

C.  DROP EXTERNAL TABLE [Ext].[Items]
    CREATE EXTERNAL TABLE [Ext].[Items]
    ([ItemID] [int] NULL,
     [ItemName] nvarchar(50) NULL,
     [ItemType] nvarchar(20) NULL,
     [ItemDescription] nvarchar(250))
    WITH
    (
        LOCATION= '/Items/',
            DATA_SOURCE = AzureDataLakeStore,
            FILE_FORMAT = PARQUET,
            REJECT_TYPE = VALUE,
            REJECT_VALUE = 0
    );

D.  ALTER TABLE [Ext].[Items]
    ADD [ItemID] int;
```

A. Option A
B. Option B
C. Option C
D. Option D

Answer: C

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

NEW QUESTION 245
- (Exam Topic 3)
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: D

Explanation:
A Type 2 SCD supports versioning of dimension members. Often the source system doesn't store versions, so the data warehouse load process detects and manages changes in a dimension table. In this case, the dimension table must use a surrogate key to provide a unique reference to a version of the dimension member. It also includes columns that define the date range validity of the version (for example, StartDate and EndDate) and possibly a flag column (for example, IsCurrent) to easily filter by current dimension members.
Reference:
https://docs.microsoft.com/en-us/learn/modules/populate-slowly-changing-dimensions-azure-synapse-analytics

NEW QUESTION 249
- (Exam Topic 3)
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?

A. Yes
B. No

Answer: B

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

NEW QUESTION 254
- (Exam Topic 3)
You have an Azure Synapse Analytics dedicated SQL pool named pool1.
You plan to implement a star schema in pool1 and create a new table named DimCustomer by using the following code.

```
CREATE TABLE dbo.[DimCustomer](
    [CustomerKey] int NOT NULL,
    [CustomerSourceID] [int] NOT NULL,
    [Title] [nvarchar](8) NULL,
    [FirstName] [nvarchar](50) NOT NULL,
    [MiddleName] [nvarchar](50) NULL,
    [LastName] [nvarchar](50) NOT NULL,
    [Suffix] [nvarchar](10) NULL,
    [CompanyName] [nvarchar](128) NULL,
    [SalesPerson] [nvarchar](256) NULL,
    [EmailAddress] [nvarchar](50) NULL,
    [Phone] [nvarchar](25) NULL,
    [InsertedDate] [datetime] NOT NULL,
    [ModifiedDate] [datetime] NOT NULL,
    [HashKey] [varchar](100) NOT NULL,
    [IsCurrentRow] [bit] NOT NULL
)
WITH
(
    DISTRIBUTION = REPLICATE,
    CLUSTERED COLUMNSTORE INDEX
);
GO
```

You need to ensure that DimCustomer has the necessary columns to support a Type 2 slowly changing dimension (SCD). Which two columns should you add? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

A. [HistoricalSalesPerson] [nvarchar] (256) NOT NULL
B. [EffectiveEndDate] [datetime] NOT NULL
C. [PreviousModifiedDate] [datetime] NOT NULL
D. [RowID] [bigint] NOT NULL
E. [EffectiveStartDate] [datetime] NOT NULL

Answer: AB

NEW QUESTION 259
- (Exam Topic 3)
You develop a dataset named DBTBL1 by using Azure Databricks. DBTBL1 contains the following columns:

> SensorTypeID

> GeographyRegionID

> Year

> Month

> Day

> Hour

> Minute

> Temperature

> WindSpeed

> Other

You need to store the data to support daily incremental load pipelines that vary for each GeographyRegionID. The solution must minimize storage costs.

How should you complete the code? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

```
df.write
```

▼	▼
.bucketBy	("*")
.format	("GeographyRegionID")
.partitionBy	("GeographyRegionID", "Year", "Month", "Day")
.sortBy	("Year", "Month", "Day", "GeographyRegionID")

```
.mode ("append")
```

▼
.csv("/DBTBL1")
.json("/DBTBL1")
.parquet("/DBTBL1")
.saveAsTable("/DBTBL1")

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Graphical user interface, text, application Description automatically generated

NEW QUESTION 263
......

Thank You for Trying Our Product

* 100% Pass or Money Back

    All our products come with a 90-day Money Back Guarantee.

* One year free update

    You can enjoy free update one year. 24x7 online support.

* Trusted by Millions

    We currently serve more than 30,000,000 customers.

* Shop Securely

    All transactions are protected by VeriSign!

100% Pass Your DP-203 Exam with Our Prep Materials Via below:

https://www.certleader.com/DP-203-dumps.html