

Exam Questions DP-203

Data Engineering on Microsoft Azure

<https://www.2passeasy.com/dumps/DP-203/>



NEW QUESTION 1

- (Exam Topic 3)

You have an Azure subscription.

You plan to build a data warehouse in an Azure Synapse Analytics dedicated SQL pool named pool1 that will contain staging tables and a dimensional model. Pool1 will contain the following tables.

Name	Number of rows	Update frequency	Description
Common.Date	7,300	New rows inserted yearly	Contains one row per date for the last 20 years

Table distribution types

Hash

Replicated

Round-robin

Answer Area

Common.Data:

Marketing.Web.Sessions:

Staging. Web.Sessions:

- A. Mastered
B. Not Mastered

Answer: A

Explanation:

Table distribution types

Hash

Replicated

Round-robin

Answer Area

Common.Data: Replicated

Marketing.Web.Sessions: Round-robin

Staging. Web.Sessions: Hash

NEW QUESTION 2

- (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 3

- (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 4

- (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 OPENROWSET 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 5

- (Exam Topic 3)

You plan to create an Azure Data Factory pipeline that will include a mapping data flow. You have JSON data containing objects that have nested arrays.

You need to transform the JSON-formatted data into a tabular dataset. The dataset must have one row for each item in the arrays.

Which transformation method should you use in the mapping data flow?

- A. unpivot
- B. flatten
- C. new branch
- D. alter row

Answer: B

Explanation:

Use the flatten transformation to take array values inside hierarchical structures such as JSON and unroll them into individual rows. This process is known as denormalization.

Reference:
<https://docs.microsoft.com/en-us/azure/data-factory/data-flow-flatten>

NEW QUESTION 6

- (Exam Topic 3)

You manage an enterprise data warehouse in Azure Synapse Analytics.

Users report slow performance when they run commonly used queries. Users do not report performance changes for infrequently used queries.

You need to monitor resource utilization to determine the source of the performance issues. Which metric should you monitor?

- A. Data IO percentage
- B. Local tempdb percentage
- C. Cache used percentage
- D. DWU percentage

Answer: C

Explanation:

Monitor and troubleshoot slow query performance by determining whether your workload is optimally leveraging the adaptive cache for dedicated SQL pools.

Reference:
<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-how-to-monit>

NEW QUESTION 7

- (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 Get Metadata activity that retrieves the DateTime of the files.

Does this meet the goal?

A. Yes

B. No

Answer: B

Explanation:

Instead use a serverless SQL pool to create an external table with the extra column. Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/create-use-external-tables>

NEW QUESTION 8

- (Exam Topic 3)

You have an Azure Synapse Analytics dedicated SQL pool mat contains a table named dbo.Users.

You need to prevent a group of users from reading user email addresses from dbo.Users. What should you use?

A. row-level security

B. column-level security

C. Dynamic data masking

D. Transparent Data Encryption (TDD)

Answer: B

NEW QUESTION 9

- (Exam Topic 3)

You are creating an Apache Spark job in Azure Databricks that will ingest JSON-formatted data. You need to convert a nested JSON string into a DataFrame that will contain multiple rows. Which Spark SQL function should you use?

A. explode

B. filter

C. coalesce

D. extract

Answer: A

Explanation:

Convert nested JSON to a flattened DataFrame

You can to flatten nested JSON, using only \$"column.*" and explode methods. Note: Extract and flatten

Use \$"column.*" and explode methods to flatten the struct and array types before displaying the flattened DataFrame.

Scala

```
display(DF.select($"id" as "main_id", $"name", $"batters", $"ppu", explode($"topping"))) // Exploding the topping column using explode as it is an array type
```

```
withColumn("topping_id", $"col.id") // Extracting topping_id from col using DOT form withColumn("topping_type", $"col.type") // Extracting topping_tytpe from col
```

```
using DOT form drop($"col")
```

```
select($"*", $"batters.*") // Flattened the struct type batters tto array type which is batter drop($"batters")
```

```
select($"*", explode($"batter")) drop($"batter")
```

```
withColumn("batter_id", $"col.id") // Extracting batter_id from col using DOT form withColumn("battter_type", $"col.type") // Extracting battter_type from col using
```

```
DOT form drop($"col")
```

```
)
```

Reference: <https://learn.microsoft.com/en-us/azure/databricks/kb/scala/flatten-nested-columns-dynamically>

NEW QUESTION 10

- (Exam Topic 3)

You are designing a dimension table for a data warehouse. The table will track the value of the dimension attributes over time and preserve the history of the data by adding new rows as the data changes.

Which type of slowly changing dimension (SCD) should use?

A. Type 0

B. Type 1

C. Type 2

D. Type 3

Answer: C

Explanation:

Type 2 - Creating a new additional record. In this methodology all history of dimension changes is kept in the database. You capture attribute change by adding a new row with a new surrogate key to the dimension table. Both the prior and new rows contain as attributes the natural key(or other durable identifier). Also 'effective date' and 'current indicator' columns are used in this method. There could be only one record with current indicator set to 'Y'. For 'effective date' columns, i.e. start_date and end_date, the end_date for current record usually is set to value 9999-12-31. Introducing changes to the dimensional model in type 2 could be very expensive database operation so it is not recommended to use it in dimensions where a new attribute could be added in the future.

<https://www.datawarehouse4u.info/SCD-Slowly-Changing-Dimensions.html>

NEW QUESTION 10

- (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 11

- (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 16

- (Exam Topic 3)

You have an Azure Data Factory pipeline that contains a data flow. The data flow contains the following expression.

```
source(output(
    License_plate as string,
    Make as string,
    Time as string
),
allowSchemaDrift: true,
```

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

See below answer.

Answer Area

Number of columns: 22

Number of rows: 4

NEW QUESTION 17

- (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 20

- (Exam Topic 3)

You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Table1. Table1 contains the following:

- > One billion rows
- > A clustered columnstore index
- > A hash-distributed column named Product Key
- > A column named Sales Date that is of the date data type and cannot be null Thirty million rows will be added to Table1 each month.

You need to partition Table1 based on the Sales Date column. The solution must optimize query performance and data loading.

How often should you create a partition?

- A. once per month
- B. once per year
- C. once per day
- D. once per week

Answer: B

Explanation:

Need a minimum 1 million rows per distribution. Each table is 60 distributions. 30 millions rows is added each month. Need 2 months to get a minimum of 1 million rows per distribution in a new partition.

Note: When creating partitions on clustered columnstore tables, it is important to consider how many rows belong to each partition. For optimal compression and performance of clustered columnstore tables, a minimum of 1 million rows per distribution and partition is needed. Before partitions are created, dedicated SQL pool already divides each table into 60 distributions.

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

Reference:

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

NEW QUESTION 21

- (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 22

- (Exam Topic 3)

The following code segment is used to create an Azure Databricks cluster.

```
{
  "num_workers": null,
  "autoscale": {
    "min_workers": 2,
    "max_workers": 8
  },
  "cluster_name": "MyCluster",
  "spark_version": "latest-stable-scala2.11",
  "spark_conf": {
    "spark.databricks.cluster.profile": "serverless",
    "spark.databricks.repl.allowedLanguages": "sql,python,r"
  },
  "node_type_id": "Standard_DS13_v2",
  "ssh_public_keys": [],
  "custom_tags": {
    "ResourceClass": "Serverless"
  },
  "spark_env_vars": {
    "PYSPARK_PYTHON": "/databricks/python3/bin/python3"
  },
  "autotermination_minutes": 90,
  "enable_elastic_disk": true,
  "init_scripts": []
}
```

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
The Databricks cluster supports multiple concurrent users.	<input type="radio"/>	<input type="radio"/>
The Databricks cluster minimizes costs when running scheduled jobs that execute notebooks.	<input type="radio"/>	<input type="radio"/>
The Databricks cluster supports the creation of a Delta Lake table.	<input type="radio"/>	<input type="radio"/>

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Graphical user interface, text, application Description automatically generated

Box 1: Yes

A cluster mode of 'High Concurrency' is selected, unlike all the others which are 'Standard'. This results in a worker type of Standard_DS13_v2.

Box 2: No

When you run a job on a new cluster, the job is treated as a data engineering (job) workload subject to the job workload pricing. When you run a job on an existing cluster, the job is treated as a data analytics (all-purpose) workload subject to all-purpose workload pricing.

Box 3: Yes

Delta Lake on Databricks allows you to configure Delta Lake based on your workload patterns. Reference:

<https://adatis.co.uk/databricks-cluster-sizing/> <https://docs.microsoft.com/en-us/azure/databricks/jobs>

<https://docs.databricks.com/administration-guide/capacity-planning/cmbp.html> <https://docs.databricks.com/delta/index.html>

NEW QUESTION 27

- (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 32

- (Exam Topic 3)

You are designing a folder structure for the files in an Azure Data Lake Storage Gen2 account. The account has one container that contains three years of data. You need to recommend a folder structure that meets the following requirements:

- Supports partition elimination for queries by Azure Synapse Analytics serverless SQL pool
- Supports fast data retrieval for data from the current month
- Simplifies data security management by department Which folder structure should you recommend?

- A. \YYY\MM\DD\Department\DataSource\DataFile_YYYYMMDD.parquet
- B. \Department\DataSource\YYY\MM\DataFile_YYYYMMDD.parquet
- C. \DD\MM\YYYY\Department\DataSource\DataFile_DDMMYY.parquet
- D. \DataSource\Department\YYYYMM\DataFile_YYYYMMDD.parquet

Answer: B

Explanation:

Department top level in the hierarchy to simplify security management.

Month (MM) at the leaf/bottom level to support fast data retrieval for data from the current month.

NEW QUESTION 33

- (Exam Topic 3)

You have an Azure Synapse Analytics serverless SQL pool, an Azure Synapse Analytics dedicated SQL pool, an Apache Spark pool, and an Azure Data Lake Storage Gen2 account.

You need to create a table in a lake database. The table must be available to both the serverless SQL pool and the Spark pool.

Where should you create the table, and Which file format should you use for data in the table? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Create the table in:

- ☐ The dedicated SQL pool
- ☒ The serverless SQL pool
- ☒ The Spark pool

File format:

- ☒ Apache Parquet
- ☐ Delta
- ☐ JSON

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

The dedicated SQL pool Apache Parquet

NEW QUESTION 38

- (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 41

- (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 High Concurrency cluster for the data engineers, and a High Concurrency cluster for the jobs.

Does this meet the goal?

- A. Yes
- B. No

Answer: A

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 46

- (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\YYYYY\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 47

- (Exam Topic 3)

You are performing exploratory analysis of the bus fare data in an Azure Data Lake Storage Gen2 account by using an Azure Synapse Analytics serverless SQL pool.

You execute the Transact-SQL query shown in the following exhibit.

```
SELECT
    payment_type,
    SUM(fare_amount) AS fare_total
FROM OPENROWSET (
    BULK 'csv/busfare/tripdata_2020*.csv',
    DATA_SOURCE = 'BusData',
    FORMAT = 'CSV', PARSER_VERSION = '2.0',
    FIRSTROW = 2
)
WITH (
    payment_type INT 10,
    fare_amount FLOAT 11
) AS nyc
GROUP BY payment_type
ORDER BY payment_type;
```

What do the query results include?

- A. Only CSV files in the tripdata_2020 subfolder.
- B. All files that have file names that beginning with "tripdata_2020".
- C. All CSV files that have file names that contain "tripdata_2020".
- D. Only CSV that have file names that beginning with "tripdata_2020".

Answer: D

NEW QUESTION 50

- (Exam Topic 3)

You are planning a streaming data solution that will use Azure Databricks. The solution will stream sales transaction data from an online store. The solution has the following specifications:

- * The output data will contain items purchased, quantity, line total sales amount, and line total tax amount.
- * Line total sales amount and line total tax amount will be aggregated in Databricks.
- * Sales transactions will never be updated. Instead, new rows will be added to adjust a sale.

You need to recommend an output mode for the dataset that will be processed by using Structured Streaming. The solution must minimize duplicate data.

What should you recommend?

- A. Append
- B. Update
- C. Complete

Answer: B

Explanation:

By default, streams run in append mode, which adds new records to the table. <https://docs.databricks.com/delta/delta-streaming.html>

NEW QUESTION 53

- (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 

RIGHT



LEFT

 FOR VALUES
(


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 55

- (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



%%spark



%%sql


park.sqlContext.sql ("select * from pysparkdftemptable")


jdbc



saveAsTable



synapsesql


("sqlpool1.dbo.PySparkTable", Constants.INTERNAL)
```

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Answer Area

```
pyspark_df.createOrReplaceTempView("pysparkdftemptable")


%%local



%%spark



%%sql


park.sqlContext.sql ("select * from pysparkdftemptable")


jdbc



saveAsTable



synapsesql


("sqlpool1.dbo.PySparkTable", Constants.INTERNAL)
```

NEW QUESTION 56

- (Exam Topic 3)

You have an Azure Databricks workspace named workspace1 in the Standard pricing tier. Workspace1 contains an all-purpose cluster named cluster1. You need to reduce the time it takes for cluster1 to start and scale up. The solution must minimize costs. What should you do first?

- A. Upgrade workspace1 to the Premium pricing tier.
- B. Create a cluster policy in workspace1.
- C. Create a pool in workspace1.
- D. Configure a global init script for workspace1.

Answer: C

Explanation:

You can use Databricks Pools to Speed up your Data Pipelines and Scale Clusters Quickly.

Databricks Pools, a managed cache of virtual machine instances that enables clusters to start and scale 4 times faster.

Reference:

<https://databricks.com/blog/2019/11/11/databricks-pools-speed-up-data-pipelines.html>

NEW QUESTION 57

- (Exam Topic 3)

You use Azure Stream Analytics to receive Twitter data from Azure Event Hubs and to output the data to an Azure Blob storage account.

You need to output the count of tweets from the last five minutes every minute. Which windowing function should you use?

- A. Sliding
- B. Session
- C. Tumbling
- D. Hopping

Answer: D

Explanation:

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.

Reference:

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-window-functions>

NEW QUESTION 60

- (Exam Topic 3)

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

Ingest:

- Access multiple data sources.
- Provide the ability to orchestrate workflow.
- Provide the capability to run SQL Server Integration Services packages.

Store:

Optimize storage for big data workloads. Provide encryption of data at rest. Operate with no size limits.

Prepare and Train:

- Provide a fully-managed and interactive workspace for exploration and visualization.
- Provide the ability to program in R, SQL, Python, Scala, and Java.
- Provide seamless user authentication with Azure Active Directory.

Model & Serve:

- Implement native columnar storage.
- Support for the SQL language
- Provide support for structured streaming. You need to build the data integration pipeline.

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

NOTE: Each correct selection is worth one point.

Answer Area

Architecture requirement	Technology
Ingest	<div>▼</div> <div>Logic Apps</div> <div>Azure Data Factory</div> <div>Azure Automation</div>
Store	<div>▼</div> <div>Azure Data Lake Storage</div> <div>Azure Blob storage</div> <div>Azure files</div>
Prepare and Train	<div>▼</div> <div>HDInsight Apache Spark cluster</div> <div>Azure Databricks</div> <div>HDInsight Apache Storm cluster</div>
Model and Serve	<div>▼</div> <div>HDInsight Apache Kafka cluster</div> <div>Azure Synapse Analytics</div> <div>Azure Data Lake Storage</div>

- A. Mastered
 B. Not Mastered

Answer: A

Explanation:

Graphical user interface, application, table, email Description automatically generated

NEW QUESTION 62

- (Exam Topic 3)

You have an Azure Stream Analytics job.

You need to ensure that the job has enough streaming units provisioned. You configure monitoring of the SU % Utilization metric.

Which two additional metrics should you monitor? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. Backlogged Input Events
 B. Watermark Delay
 C. Function Events
 D. Out of order Events
 E. Late Input Events

Answer: AB

Explanation:

To react to increased workloads and increase streaming units, consider setting an alert of 80% on the SU Utilization metric. Also, you can use watermark delay and backlogged events metrics to see if there is an impact.

Note: 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, by increasing the SUs.

Reference:

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-monitoring>

NEW QUESTION 63

- (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	
PARTITION BY	'Standard'
ROW_NUMBER	END AS employee_type
	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 65

- (Exam Topic 3)

You have an Azure Synapse Analytics dedicated SQL pool named Pool1. Pool1 contains a fact table named Table1. Table1 contains sales data. Sixty-five million rows of data are added to Table1 monthly.

At the end of each month, you need to remove data that is older than 36 months. The solution must minimize how long it takes to remove the data.

How should you partition Table1, and how should you remove the old data? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

Partition the data:	<div>Partition by date with one partition per day.</div> <div>Partition by date with one partition per day.</div> <div>Partition by date with one partition per month.</div> <div>Partition by product.</div>
Remove the data:	<div>Delete the old data from Table1 by using a WHERE clause.</div> <div>Delete the old data from Table1 by using a WHERE clause.</div> <div>Delete the old data from Table1 by using a JOIN.</div> <div>Switch the oldest partition to another table named Table2 and drop Table2.</div> <div>Truncate the oldest partition.</div>

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Answer Area

Partition the data:	<div>Partition by date with one partition per day.</div> <div>Partition by date with one partition per day.</div> <div>Partition by date with one partition per month.</div> <div>Partition by product.</div>
Remove the data:	<div>Delete the old data from Table1 by using a WHERE clause.</div> <div>Delete the old data from Table1 by using a WHERE clause.</div> <div>Delete the old data from Table1 by using a JOIN.</div> <div>Switch the oldest partition to another table named Table2 and drop Table2.</div> <div>Truncate the oldest partition.</div>

NEW QUESTION 68

- (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 69

- (Exam Topic 3)

You are designing an Azure Stream Analytics job to process incoming events from sensors in retail environments.

You need to process the events to produce a running average of shopper counts during the previous 15 minutes, calculated at five-minute intervals.

Which type of window should you use?

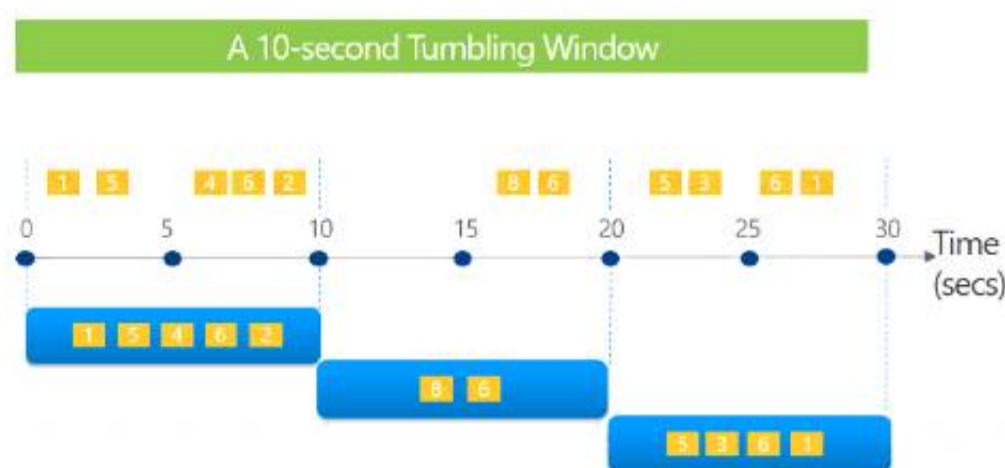
- A. snapshot
- B. tumbling
- C. hopping
- D. sliding

Answer: B

Explanation:

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.

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



```
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 72

- (Exam Topic 3)

You have an Azure Synapse Analytics serverless SQ1 pool.

You have an Azure Data Lake Storage account named aols1 that contains a public container named container1. The container 1 container contains a folder named folder 1.

You need to query the top 100 rows of all the CSV files in folder 1.

How shouk1 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

BULK

DATA_SOURCE

LOCATION

OPENROWSET

Answer Area

SELECT TOP 100 *

FROM

'https://adls1.dfs.core.windows.net/container1/folder1/*.csv',

FORMAT = 'CSV') AS rows

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Values

BULK

DATA_SOURCE

LOCATION

OPENROWSET

Answer Area

SELECT TOP 100 *

FROM

OPENROWSET

BULK

'https://adls1.dfs.core.windows.net/container1/folder1/*.csv',

FORMAT = 'CSV') AS rows

NEW QUESTION 77

- (Exam Topic 3)

You are designing a fact table named FactPurchase in an Azure Synapse Analytics dedicated SQL pool. The table contains purchases from suppliers for a retail store. FactPurchase will contain the following columns.

Name	Data type	Nullable
PurchaseKey	Bigint	No
DateKey	Int	No
SupplierKey	Int	No
StockItemKey	Int	No
PurchaseOrderID	Int	Yes
OrderedQuantity	Int	No
OrderedOuters	Int	No
ReceivedOuters	Int	No
Package	Nvarchar(50)	No
IsOrderFinalized	Bit	No
LineageKey	Int	No

FactPurchase will have 1 million rows of data added daily and will contain three years of data. Transact-SQL queries similar to the following query will be executed daily.

SELECT
SupplierKey, StockItemKey, COUNT(*) FROM FactPurchase
WHERE DateKey >= 20210101 AND DateKey <= 20210131
GROUP By SupplierKey, StockItemKey
Which table distribution will minimize query times?

- A. round-robin
- B. replicated
- C. hash-distributed on DateKey
- D. hash-distributed on PurchaseKey

Answer: D

Explanation:

Hash-distributed tables improve query performance on large fact tables, and are the focus of this article. Round-robin tables are useful for improving loading speed.
Reference:

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

NEW QUESTION 80

- (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 be 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 85

- (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 scenario, 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 Storage account that contains 100 GB of files. The files contain text and numerical values. 75% of the rows contain description data that has an average length of 1.1 MB.

You plan to copy the data from the storage account to an Azure SQL data warehouse. You need to prepare the files to ensure that the data copies quickly.

Solution: You modify the files to ensure that each row is less than 1 MB. Does this meet the goal?

- A. Yes
- B. No

Answer: A

Explanation:

When exporting data into an ORC File Format, you might get Java out-of-memory errors when there are large text columns. To work around this limitation, export only a subset of the columns.

References:

https://docs.microsoft.com/en-us/azure/sql-data-warehouse/guidance-for-loading-data

NEW QUESTION 88

- (Exam Topic 3)

You have an Azure Synapse Analytics dedicated SQL pool named Pool1 that contains an external table named Sales. Sales contains sales data. Each row in Sales

contains data on a single sale, including the name of the salesperson.

You need to implement row-level security (RLS). The solution must ensure that the salespeople can access only their respective sales.

What should you do? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

Create:

A materialized view in Pool1
A security policy for Sales
Database scoped credentials in Pool1

Restrict row access by using:

A masking rule
A table-valued function
The CONTAINS predicate

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Box 1: A security policy for sale

Here are the steps to create a security policy for Sales:

- > Create a user-defined function that returns the name of the current user:
- > CREATE FUNCTION dbo.GetCurrentUser()
- > RETURNS NVARCHAR(128)
- > AS
- > BEGIN
- > RETURN SUSER_SNAME();
- > END;
- > Create a security predicate function that filters the Sales table based on the current user:
- > CREATE FUNCTION dbo.SalesPredicate(@salesperson NVARCHAR(128))
- > RETURNS TABLE
- > WITH SCHEMABINDING
- > AS
- > RETURN SELECT 1 AS access_result
- > WHERE @salesperson = SalespersonName;
- > Create a security policy on the Sales table that uses the SalesPredicate function to filter the data:
- > CREATE SECURITY POLICY SalesFilter
- > ADD FILTER PREDICATE dbo.SalesPredicate(dbo.GetCurrentUser()) ON dbo.Sales
- > WITH (STATE = ON);

By creating a security policy for the Sales table, you ensure that each salesperson can only access their own sales data. The security policy uses a user-defined function to get the name of the current user and a security predicate function to filter the Sales table based on the current user.

Box 2: table-value function

to restrict row access by using row-level security, you need to create a table-valued function that returns a table of values that represent the rows that a user can access. You then use this function in a security policy that applies a predicate on the table.

NEW QUESTION 91

- (Exam Topic 3)

You have an Azure subscription that contains a Microsoft Purview account named MP1, an Azure data factory named DF1, and a storage account named storage. MP1 is configured

10 scan storage1. DF1 is connected to MP1 and contains 3 dataset named DS1. DS1 references 2 file in storage.

In DF1, you plan to create a pipeline that will process data from DS1.

You need to review the schema and lineage information in MP1 for the data referenced by DS1.

Which two features can you use to locate the information? Each correct answer presents a complete solution. NOTE: Each correct answer is worth one point.

- A. the Storage browser of storage1 in the Azure portal
- B. the search bar in the Azure portal
- C. the search bar in Azure Data Factory Studio
- D. the search bar in the Microsoft Purview governance portal

Answer: CD

Explanation:

> The search bar in the Microsoft Purview governance portal: This is a feature that allows you to search for assets in your data estate using keywords, filters, and facets. You can use the search bar to find the files in storage1 that are referenced by DS1, and then view their schema and lineage information in the asset details page12.

> The search bar in Azure Data Factory Studio: This is a feature that allows you to search for datasets, linked services, pipelines, and other resources in your data factory. You can use the search bar to find DS1 in DF1, and then view its schema and lineage information in the dataset details page. You can also click on the Open in Purview button to open the corresponding asset in MP13.

The two features that can be used to locate the schema and lineage information for the data referenced by DS1 are the search bar in Azure Data Factory Studio and the search bar in the Microsoft Purview governance portal.

The search bar in Azure Data Factory Studio allows you to search for the dataset DS1 and view its properties and lineage. This can help you locate information about the source and destination data stores, as well as the transformations that were applied to the data.

The search bar in the Microsoft Purview governance portal allows you to search for the storage account and view its metadata, including schema and lineage information. This can help you understand the different data assets that are stored in the storage account and how they are related to each other.

The Storage browser of storage1 in the Azure portal may allow you to view the files that are stored in the storage account, but it does not provide lineage or schema information for those files. Similarly, the search bar in the Azure portal may allow you to search for resources in the Azure subscription, but it does not provide detailed information about the data assets themselves.

References:

- > What is Azure Purview?
- > Use Azure Data Factory Studio

NEW QUESTION 96

- (Exam Topic 2)

Which Azure Data Factory components should you recommend using together to import the daily inventory data from the SQL server to Azure Data Lake Storage?

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

NOTE: Each correct selection is worth one point.

Integration runtime type:

Azure integration runtime
Azure-SSIS integration runtime
Self-hosted integration runtime

Trigger type:

Event-based trigger
Schedule trigger
Tumbling window trigger

Activity type:

Copy activity
Lookup activity
Stored procedure activity

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Box 1: Self-hosted integration runtime

A self-hosted IR is capable of running copy activity between a cloud data stores and a data store in private network.

Box 2: Schedule trigger Schedule every 8 hours Box 3: Copy activity Scenario:

- > Customer data, including name, contact information, and loyalty number, comes from Salesforce and can be imported into Azure once every eight hours. Row modified dates are not trusted in the source table.
- > Product data, including product ID, name, and category, comes from Salesforce and can be imported into Azure once every eight hours. Row modified dates are not trusted in the source table.

NEW QUESTION 100

- (Exam Topic 1)

You need to design the partitions for the product sales transactions. The solution must meet the sales transaction dataset requirements.

What should you include in the solution? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

Partition product sales transactions data by:

Sales date
Product ID
Promotion ID

Store product sales transactions data in:

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

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Box 1: Sales date

Scenario: Contoso requirements for data integration include:

➤ Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month. Boundary values must belong to the partition on the right.

Box 2: An Azure Synapse Analytics Dedicated SQL pool Scenario: Contoso requirements for data integration include:

➤ Ensure that data storage costs and performance are predictable.

The size of a dedicated SQL pool (formerly SQL DW) is determined by Data Warehousing Units (DWU). Dedicated SQL pool (formerly SQL DW) stores data in relational tables with columnar storage. This format

significantly reduces the data storage costs, and improves query performance.

Synapse analytics dedicated sql pool Reference:

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

NEW QUESTION 103

- (Exam Topic 1)

You need to ensure that the Twitter feed data can be analyzed in the dedicated SQL pool. The solution must meet the customer sentiment analytics requirements. Which three Transaction-SQL DDL commands should you run in sequence? To answer, move the appropriate commands from the list of commands to the answer area and arrange them in the correct order.

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

Commands

CREATE EXTERNAL DATA SOURCE

CREATE EXTERNAL FILE FORMAT

CREATE EXTERNAL TABLE

CREATE EXTERNAL TABLE AS SELECT

CREATE DATABASE SCOPED CREDENTIAL

Answer Area

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Scenario: Allow Contoso users to use PolyBase in an Azure Synapse Analytics dedicated SQL pool to query the content of the data records that host the Twitter feeds. Data must be protected by using row-level security (RLS). The users must be authenticated by using their own Azure AD credentials.

Box 1: CREATE EXTERNAL DATA SOURCE

External data sources are used to connect to storage accounts. Box 2: CREATE EXTERNAL FILE FORMAT

CREATE EXTERNAL FILE FORMAT creates an external file format object that defines external data stored in Azure Blob Storage or Azure Data Lake Storage.

Creating an external file format is a prerequisite for creating an external table.

Box 3: CREATE EXTERNAL TABLE AS SELECT

When used in conjunction with the CREATE TABLE AS SELECT statement, selecting from an external table imports data into a table within the SQL pool. In addition to the COPY statement, external tables are useful for loading data.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables>

NEW QUESTION 106

- (Exam Topic 1)

You need to implement versioned changes to the integration pipelines. The solution must meet the data integration requirements.

In which order should you perform the actions? To answer, move all actions from the list of actions to the answer area and arrange them in the correct order.

Actions

Publish changes.

Create a feature branch.

Merge changes.

Create a repository and a main branch.

Create a pull request.

Answer Area

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Graphical user interface, application Description automatically generated

Scenario: Identify a process to ensure that changes to the ingestion and transformation activities can be version-controlled and developed independently by multiple data engineers.

Step 1: Create a repository and a main branch

You need a Git repository in Azure Pipelines, TFS, or GitHub with your app. Step 2: Create a feature branch

Step 3: Create a pull request Step 4: Merge changes

Merge feature branches into the main branch using pull requests. Step 5: Publish changes

Reference:

<https://docs.microsoft.com/en-us/azure/devops/pipelines/repos/pipeline-options-for-git>

NEW QUESTION 108

- (Exam Topic 1)

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

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

NOTE: Each correct selection is worth one point.

Table type to store retail store data:

	▼
Hash	
Replicated	
Round-robin	

Table type to store promotional data:

	▼
Hash	
Replicated	
Round-robin	

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Graphical user interface, text, application, table Description automatically generated

Box 1: Round-robin

Round-robin tables are useful for improving loading speed.

Scenario: Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month.

Box 2: Hash

Hash-distributed tables improve query performance on large fact tables. Reference:

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

NEW QUESTION 109

- (Exam Topic 1)

You need to design a data ingestion and storage solution for the Twitter feeds. The solution must meet the customer sentiment analytics requirements.

What should you include in the solution? To answer, select the appropriate options in the answer area NOTE: Each correct selection b worth one point.

Answer Area

To increase the throughput of ingesting the Twitter feeds:

Configure Event Hubs partitions.
Enable Auto-Inflate in Event Hubs.
Use Event Hubs Dedicated.

To store the Twitter feed data, use:

An Azure Data Lake Storage Gen2 account
An Azure Databricks high concurrency cluster
An Azure General-purpose v2 storage account in the Premium tier

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Graphical user interface, text Description automatically generated

Box 1: Configure Evegnet Hubs partitions

Scenario: Maximize the throughput of ingesting Twitter feeds from Event Hubs to Azure Storage without purchasing additional throughput or capacity units.

Event Hubs is designed to help with processing of large volumes of events. Event Hubs throughput is scaled by using partitions and throughput-unit allocations.

Event Hubs traffic is controlled by TUs (standard tier). Auto-inflate enables you to start small with the minimum required TUs you choose. The feature then scales automatically to the maximum limit of TUs you need, depending on the increase in your traffic.

Box 2: An Azure Data Lake Storage Gen2 account

Scenario: Ensure that the data store supports Azure AD-based access control down to the object level. Azure Data Lake Storage Gen2 implements an access control model that supports both Azure role-based access control (Azure RBAC) and POSIX-like access control lists (ACLs).

Reference:

<https://docs.microsoft.com/en-us/azure/event-hubs/event-hubs-features> <https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control>

NEW QUESTION 112

- (Exam Topic 1)

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

- A. change feed
- B. soft delete
- C. time-based retention
- D. lifecycle management

Answer: D

Explanation:

Scenario: Purge Twitter feed data records that are older than two years.

Data sets have unique lifecycles. Early in the lifecycle, people access some data often. But the need for access often drops drastically as the data ages. Some data remains idle in the cloud and is rarely accessed once stored. Some data sets expire days or months after creation, while other data sets are actively read and modified throughout their lifetimes. Azure Storage lifecycle management offers a rule-based policy that you can use to transition blob data to the appropriate access tiers or to expire data at the end of the data lifecycle.

Reference:

<https://docs.microsoft.com/en-us/azure/storage/blobs/lifecycle-management-overview>

NEW QUESTION 116

- (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 Storage account that contains 100 GB of files. The files contain rows of text and numerical values. 75% of the rows contain description data that has an average length of 1.1 MB.

You plan to copy the data from the storage account to an enterprise data warehouse in Azure Synapse Analytics.

You need to prepare the files to ensure that the data copies quickly. Solution: You copy the files to a table that has a columnstore index. Does this meet the goal?

- A. Yes
- B. No

Answer: B

Explanation:

Instead convert the files to compressed delimited text files. Reference:

<https://docs.microsoft.com/en-us/azure/sql-data-warehouse/guidance-for-loading-data>

NEW QUESTION 119

- (Exam Topic 3)

You use Azure Data Lake Storage Gen2 to store data that data scientists and data engineers will query by using Azure Databricks interactive notebooks. Users will have access only to the Data Lake Storage folders that relate to the projects on which they work.

You need to recommend which authentication methods to use for Databricks and Data Lake Storage to provide the users with the appropriate access. The solution must minimize administrative effort and development effort.

Which authentication method should you recommend for each Azure service? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Databricks:

	▼
Azure Active Directory credential passthrough	
Azure Key Vault secrets	
Personal access tokens	

Data Lake Storage:

	▼
Azure Active Directory credential passthrough	
Shared access keys	
Shared access signatures	

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Table Description automatically generated

Box 1: Personal access tokens

You can use storage shared access signatures (SAS) to access an Azure Data Lake Storage Gen2 storage account directly. With SAS, you can restrict access to a storage account using temporary tokens with fine-grained access control.

You can add multiple storage accounts and configure respective SAS token providers in the same Spark session.

Box 2: Azure Active Directory credential passthrough

You can authenticate automatically to Azure Data Lake Storage Gen1 (ADLS Gen1) and Azure Data Lake Storage Gen2 (ADLS Gen2) from Azure Databricks clusters using the same Azure Active Directory (Azure AD) identity that you use to log into Azure Databricks. When you enable your cluster for Azure Data Lake Storage credential passthrough, commands that you run on that cluster can read and write data in Azure Data Lake Storage without requiring you to configure service principal credentials for access to storage.

After configuring Azure Data Lake Storage credential passthrough and creating storage containers, you can access data directly in Azure Data Lake Storage Gen1 using an adl:// path and Azure Data Lake Storage Gen2 using an abfs:// path:

Reference:

<https://docs.microsoft.com/en-us/azure/databricks/data/data-sources/azure/adls-gen2/azure-datalake-gen2-sas-ac> <https://docs.microsoft.com/en-us/azure/databricks/security/credential-passthrough/adls-passthrough>

NEW QUESTION 124

- (Exam Topic 3)

You have an Azure Synapse Analytics serverless SQL pool named Pool1 and an Azure Data Lake Storage Gen2 account named storage1. The AllowedBlobpublicAccess property is disabled for storage1.

You need to create an external data source that can be used by Azure Active Directory (Azure AD) users to access storage1 from Pool1.

What should you create first?

- A. an external resource pool
- B. a remote service binding
- C. database scoped credentials
- D. an external library

Answer: C

Explanation:

Security

User must have SELECT permission on an external table to read the data. External tables access underlying Azure storage using the database scoped credential defined in data source.

Note: A database scoped credential is a record that contains the authentication information that is required to connect to a resource outside SQL Server. Most credentials include a Windows user and password.

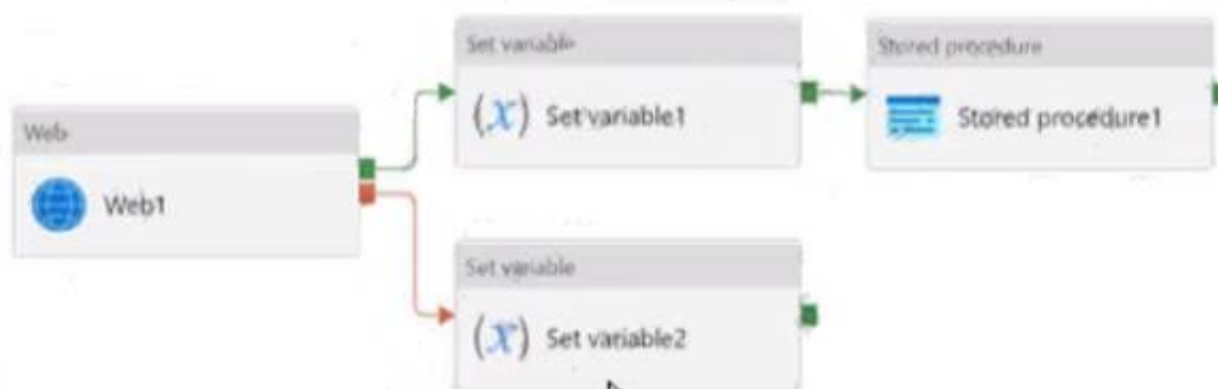
Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables> <https://docs.microsoft.com/en-us/sql/t-sql/statements/create-database-scoped-credential-transact-sql>

NEW QUESTION 125

- (Exam Topic 3)

You have an Azure Data Factory pipeline that has the activity shown in the following exhibit.



Use the drop-down menus to select the answer choice that completes each statement based on the information presented in the graphic.

Answer Area

Stored procedure1 will execute if Web1 and Set variable1 [answer choice]

complete

fail

succeed

This case are the selections for the statement Stored procedure1 will execute if Web1 and Set variable1 [answer choice]

If Web1 fails and Set variable2 succeeds, the pipeline status will be [answer choice]

Cancelled

Failed

Succeeded

This case are the selections for the statement If Web1 fails and Set variable2 succeeds, the pipeline status will be [answer choice]

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Answer Area

Stored procedure1 will execute if Web1 and Set variable1 [answer choice] succeed

If Web1 fails and Set variable2 succeeds, the pipeline status will be [answer choice] Failed

NEW QUESTION 128

- (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 132

- (Exam Topic 3)

You need to create an Azure Data Factory pipeline to process data for the following three departments at your company: Ecommerce, retail, and wholesale. The solution must ensure that data can also be processed for the entire company.

How should you complete the Data Factory data flow script? 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

all, ecommerce, retail, wholesale

dept=='ecommerce', dept=='retail', dept=='wholesale'

dept=='ecommerce', dept=='wholesale', dept=='retail'

disjoint: false

disjoint: true

ecommerce, retail, wholesale, all

Answer Area

```
CleanData
split(
    ) ~> SplitByDept@(
    )
```

- A. Mastered
B. Not Mastered

Answer: A

Explanation:

The conditional split transformation routes data rows to different streams based on matching conditions. The conditional split transformation is similar to a CASE decision structure in a programming language. The transformation evaluates expressions, and based on the results, directs the data row to the specified stream. Box 1: dept=='ecommerce', dept=='retail', dept=='wholesale'

First we put the condition. The order must match the stream labeling we define in Box 3. Syntax:

```
<incomingStream> split(
<conditionalExpression1>
<conditionalExpression2> disjoint: {true | false}
) ~> <splitTx>@(stream1, stream2, ..., <defaultStream>)
```

Box 2: discount : false

disjoint is false because the data goes to the first matching condition. All remaining rows matching the third condition go to output stream all.

Box 3: ecommerce, retail, wholesale, all Label the streams

Reference:

https://docs.microsoft.com/en-us/azure/data-factory/data-flow-conditional-split

NEW QUESTION 135

- (Exam Topic 3)

You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Contacts. Contacts contains a column named Phone. You need to ensure that users in a specific role only see the last four digits of a phone number when querying the Phone column. What should you include in the solution?

- A. a default value
- B. dynamic data masking
- C. row-level security (RLS)
- D. column encryption
- E. table partitions

Answer: B

Explanation:

Dynamic data masking helps prevent unauthorized access to sensitive data by enabling customers to designate how much of the sensitive data to reveal with minimal impact on the application layer. It's a policy-based security feature that hides the sensitive data in the result set of a query over designated database fields, while the data in the database is not changed.

Reference:

https://docs.microsoft.com/en-us/azure/azure-sql/database/dynamic-data-masking-overview

NEW QUESTION 138

- (Exam Topic 3)

You have an Azure subscription that contains an Azure Databricks workspace. The workspace contains a notebook named Notebook1. In Notebook1, you create an Apache Spark DataFrame named df_sales that contains the following columns:

- Customer
- Salesperson
- Region
- Amount

You need to identify the three top performing salespersons by amount for a region named HQ.

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.

Values

agg(col('SalesPerson'))

filter(col('SalesPerson'))

groupBy(col('SalesPerson'))

groupBy(col('TotalAmount'))

orderBy(col('TotalAmount'))

orderBy(desc('TotalAmount'))

Answer Area

```
df_sales.filter(col('Region')=='HQ').  
  
    .agg(sum('Amount').alias  
    ('TotalAmount')).  
  
    
```

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Values

agg(col('SalesPerson'))

filter(col('SalesPerson'))

groupBy(col('SalesPerson'))

groupBy(col('TotalAmount'))

orderBy(col('TotalAmount'))

orderBy(desc('TotalAmount'))

Answer Area

```
df_sales.filter(col('Region')=='HQ').  
  
    .agg(sum('Amount').alias  
    ('TotalAmount')).  
  
    
```

NEW QUESTION 141

- (Exam Topic 3)

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

- A. Last Sync Time
- B. Average Success Latency
- C. Error errors
- D. availability

Answer: A

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 146

- (Exam Topic 3)

You are deploying a lake database by using an Azure Synapse database template.

You need to add additional tables to the database. The solution must use the same grouping method as the template tables.

Which grouping method should you use?

- A. business area
- B. size
- C. facts and dimensions
- D. partition style

Answer: A

Explanation:

Business area: This is how the Azure Synapse database templates group tables by default. Each template consists of one or more enterprise templates that contain tables grouped by business areas. For example, the Retail template has business areas such as Customer, Product, Sales, and Store123. Using the same grouping method as the template tables can help you maintain consistency and compatibility with the industry-specific data model.

<https://techcommunity.microsoft.com/t5/azure-synapse-analytics-blog/database-templates-in-azure-synapse-anal>

NEW QUESTION 147

- (Exam Topic 3)

You are designing an Azure Databricks interactive cluster. The cluster will be used infrequently and will be configured for auto-termination.

You need to ensure that the cluster configuration is retained indefinitely after the cluster is terminated. The solution must minimize costs.

What should you do?

- A. Clone the cluster after it is terminated.
- B. Terminate the cluster manually when processing completes.
- C. Create an Azure runbook that starts the cluster every 90 days.
- D. Pin the cluster.

Answer: D

Explanation:

To keep an interactive cluster configuration even after it has been terminated for more than 30 days, an administrator can pin a cluster to the cluster list.

References:

<https://docs.azuredatabricks.net/clusters/clusters-manage.html#automatic-termination>

NEW QUESTION 150

- (Exam Topic 3)

You are designing a slowly changing dimension (SCD) for supplier data in an Azure Synapse Analytics dedicated SQL pool.

You plan to keep a record of changes to the available fields. The supplier data contains the following columns.

Name	Description
SupplierSystemID	Unique supplier ID in an enterprise resource planning (ERP) system
SupplierName	Name of the supplier company
SupplierAddress1	Address of the supplier company
SupplierAddress2	Second address line of the supplier company
SupplierCity	City of the supplier company
SupplierStateProvince	State or province of the supplier company
SupplierCountry	Country of the supplier company
SupplierPostalCode	Postal code of the supplier company
SupplierDescription	Free-text description of the supplier company
SupplierCategory	Category of goods provided by the supplier company

Which three additional columns should you add to the data to create a Type 2 SCD? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. surrogate primary key
- B. foreign key
- C. effective start date
- D. effective end date
- E. last modified date
- F. business key

Answer: CDF

Explanation:

Reference:

<https://docs.microsoft.com/en-us/sql/integration-services/data-flow/transformations/slowly-changing-dimension>

NEW QUESTION 154

- (Exam Topic 3)

You have two Azure Data Factory instances named ADFdev and ADFprod. ADFdev connects to an Azure DevOps Git repository.

You publish changes from the main branch of the Git repository to ADFdev. You need to deploy the artifacts from ADFdev to ADFprod.

What should you do first?

- A. From ADFdev, modify the Git configuration.
- B. From ADFdev, create a linked service.
- C. From Azure DevOps, create a release pipeline.
- D. From Azure DevOps, update the main branch.

Answer: C

Explanation:

In Azure Data Factory, continuous integration and delivery (CI/CD) means moving Data Factory pipelines from one environment (development, test, production) to another.

Note:

The following is a guide for setting up an Azure Pipelines release that automates the deployment of a data factory to multiple environments.

- In Azure DevOps, open the project that's configured with your data factory.
- On the left side of the page, select Pipelines, and then select Releases.
- Select New pipeline, or, if you have existing pipelines, select New and then New release pipeline.
- In the Stage name box, enter the name of your environment.
- Select Add artifact, and then select the git repository configured with your development data factory.

Select the publish branch of the repository for the Default branch. By default, this publish branch is adf_publish.

- Select the Empty job template. Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/continuous-integration-deployment>

NEW QUESTION 156

- (Exam Topic 3)

You plan to implement an Azure Data Lake Gen2 storage account.

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

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

- A. geo-redundant storage (GRS)
- B. zone-redundant storage (ZRS)
- C. locally-redundant storage (LRS)
- D. geo-zone-redundant storage (GZRS)

Answer: C

Explanation:

Locally redundant storage (LRS) copies your data synchronously three times within a single physical location in the primary region. LRS is the least expensive replication option

Reference:

<https://docs.microsoft.com/en-us/azure/storage/common/storage-redundancy>

NEW QUESTION 159

- (Exam Topic 3)

You have an Azure Databricks workspace and an Azure Data Lake Storage Gen2 account named storage1. 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 160

- (Exam Topic 3)

You are designing a dimension table in an Azure Synapse Analytics dedicated SQL pool.

You need to create a surrogate key for the table. The solution must provide the fastest query performance. What should you use for the surrogate key?

- A. a GUID column
- B. a sequence object
- C. an IDENTITY column

Answer: C

Explanation:

Use IDENTITY to create surrogate keys using dedicated SQL pool in AzureSynapse Analytics.

Note: A surrogate key on a table is a column with a unique identifier for each row. The key is not generated from the table data. Data modelers like to create surrogate keys on their tables when they design data warehouse models. You can use the IDENTITY property to achieve this goal simply and effectively without affecting load performance.

Reference:

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

NEW QUESTION 161

- (Exam Topic 3)

You have an Azure Synapse Analytics pipeline named Pipeline1 that contains a data flow activity named Dataflow1.

Pipeline1 retrieves files from an Azure Data Lake Storage Gen 2 account named storage1.

Dataflow1 uses the AutoResolveIntegrationRuntime integration runtime configured with a core count of 128. You need to optimize the number of cores used by Dataflow1 to accommodate the size of the files in storage1. What should you configure? To answer, select the appropriate options in the answer area.

To Pipeline1, add:

A custom activity
A Get Metadata activity
An If Condition activity

For Dataflow1, set the core count by using:

Dynamic content
Parameters
User properties

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Box 1: A Get Metadata activity

Dynamically size data flow compute at runtime

The Core Count and Compute Type properties can be set dynamically to adjust to the size of your incoming source data at runtime. Use pipeline activities like Lookup or Get Metadata in order to find the size of the source dataset data. Then, use Add Dynamic Content in the Data Flow activity properties.

Box 2: Dynamic content

Reference: <https://docs.microsoft.com/en-us/azure/data-factory/control-flow-execute-data-flow-activity>

NEW QUESTION 166

- (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	

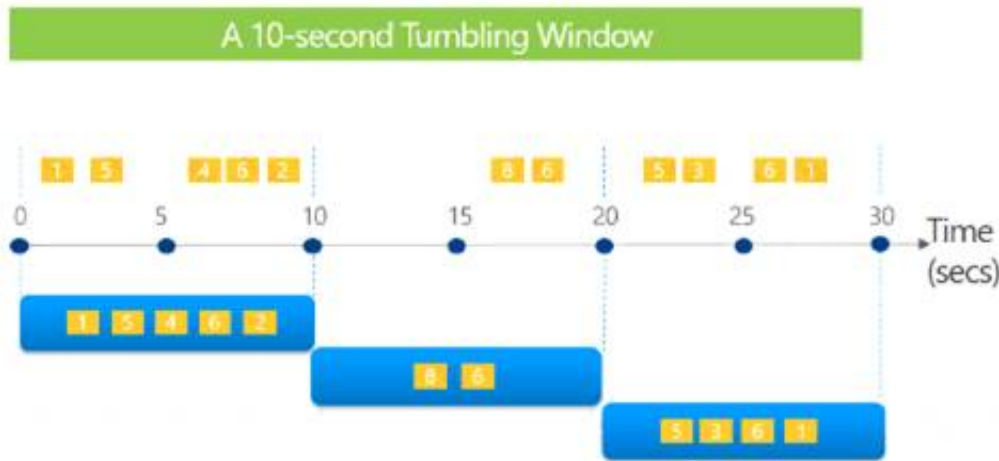
- A. Mastered
- B. Not Mastered

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



```
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 169

- (Exam Topic 3)
 You have an Azure Synapse Analytics dedicated SQL pool named Pool1. Pool1 contains a table named table1. You load 5 TB of data into table1. You need to ensure that columnstore compression is maximized for table1. Which statement should you execute?

- A. ALTER INDEX ALL on table1 REORGANIZE
- B. ALTER INDEX ALL on table1 REBUILD
- C. DBCC DBREINDEX (table1)
- D. DBCC INDEXDEFRAG (pool1,table1)

Answer: B

Explanation:
 Columnstore and columnstore archive compression
 Columnstore tables and indexes are always stored with columnstore compression. You can further reduce the size of columnstore data by configuring an additional compression called archival compression. To perform archival compression, SQL Server runs the Microsoft XPRESS compression algorithm on the data. Add or remove archival compression by using the following data compression types:
 Use COLUMNSTORE_ARCHIVE data compression to compress columnstore data with archival compression.
 Use COLUMNSTORE data compression to decompress archival compression. The resulting data continue to be compressed with columnstore compression.
 To add archival compression, use ALTER TABLE (Transact-SQL) or ALTER INDEX (Transact-SQL) with the REBUILD option and DATA COMPRESSION = COLUMNSTORE_ARCHIVE.
 Reference: <https://learn.microsoft.com/en-us/sql/relational-databases/data-compression/data-compression>

NEW QUESTION 173

- (Exam Topic 3)
 You have an Azure Synapse Analytics dedicated SQL pool named Pool1 and an Azure Data Lake Storage Gen2 account named Account1. You plan to access the files in Account1 by using an external table. You need to create a data source in Pool1 that you can reference when you create the external table. 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.

```
CREATE EXTERNAL DATA SOURCE source1
WITH
    ( LOCATION = 'https://account1. .core.windows.net',
    )
```

PUSHDOWN = ON
 TYPE = BLOB_STORAGE
 TYPE = HADOOP

blob
 dfs
 table

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Graphical user interface, diagram Description automatically generated

Box 1: blob

The following example creates an external data source for Azure Data Lake Gen2 CREATE EXTERNAL DATA SOURCE YellowTaxi WITH (LOCATION = 'https://azureopendatastorage.blob.core.windows.net/nyctlc/yellow/', TYPE = HADOOP)

Box 2: HADOOP

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables>

NEW QUESTION 176

- (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 are designing an Azure Stream Analytics solution that will analyze Twitter data.

You need to count the tweets in each 10-second window. The solution must ensure that each tweet is counted only once.

Solution: You use a hopping window that uses a hop size of 5 seconds and a window size 10 seconds. Does this meet the goal?

- A. Yes
- B. No

Answer: B

Explanation:

Instead use a tumbling window. Tumbling windows are a series of fixed-sized, non-overlapping and contiguous time intervals.

Reference:

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

NEW QUESTION 180

- (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 jobs

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 181

- (Exam Topic 3)

You have an Azure Data Factory version 2 (V2) resource named Df1. Df1 contains a linked service. You have an Azure Key vault named vault1 that contains an encryption key named key1.

You need to encrypt Df1 by using key1. What should you do first?

- A. Add a private endpoint connection to vault 1.
- B. Enable Azure role-based access control on vault 1.
- C. Remove the linked service from Df1.
- D. Create a self-hosted integration runtime.

Answer: C

Explanation:

Linked services are much like connection strings, which define the connection information needed for Data Factory to connect to external resources.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/enable-customer-managed-key> <https://docs.microsoft.com/en-us/azure/data-factory/concepts-linked-services> <https://docs.microsoft.com/en-us/azure/data-factory/create-self-hosted-integration-runtime>

NEW QUESTION 186

- (Exam Topic 3)

You are building an Azure Analytics query that will receive input data from Azure IoT Hub and write the results to Azure Blob storage.

You need to calculate the difference in readings per sensor per hour.

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 sensorId,
growth = reading -

LAG

LAST

LEAD

(reading) OVER (PARTITION BY sensorId

LIMIT DURATION

OFFSET

WHEN

(hour,1))

FROM input

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Box 1: LAG
The LAG analytic operator allows one to look up a “previous” event in an event stream, within certain constraints. It is very useful for computing the rate of growth of a variable, detecting when a variable crosses a threshold, or when a condition starts or stops being true.
Box 2: LIMIT DURATION
Example: Compute the rate of growth, per sensor: SELECT sensorId,
growth = reading
LAG(reading) OVER (PARTITION BY sensorId LIMIT DURATION(hour, 1)) FROM input
Reference:
https://docs.microsoft.com/en-us/stream-analytics-query/lag-azure-stream-analytics

NEW QUESTION 187

- (Exam Topic 3)
You use PySpark in Azure Databricks to parse the following JSON input.

```
{
  "persons": [
    {
      "name": "Keith",
      "age": 30,
      "dogs": ["Fido", "Fluffy"]
    },
    {
      "name": "Donna",
      "age": 46,
      "dogs": ["Spot"]
    }
  ]
}
```

You need to output the data in the following tabular format.

owner	age	dog
Keith	30	Fido
Keith	30	Fluffy
Donna	46	Spot

How should you complete the PySpark code? To answer, drag the appropriate values to he 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

alias

array_union

createDataFrame

explode

select

translate

Answer Area

```
dbutils.fs.put("/tmp/source.json", source_json, True)

source_df = spark.read.option("multiline", "true").json("/tmp/source.json")

persons = source_df. Value Value ("persons").alias("persons"))

persons_dogs = persons.select(col("persons.name").alias("owner"), col("persons.age").alias("age"),
explode Value ("dog"))
("persons.dogs").
display(persons_dogs)
```

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Graphical user interface, text, application Description automatically generated
Box 1: select

Box 2: explode

Box 3: alias

pyspark.sql.Column.alias returns this column aliased with a new name or names (in the case of expressions that return more than one column, such as explode).

Reference: <https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.sql.Column.alias.html> <https://docs.microsoft.com/en-us/azure/databricks/sql/language-manual/functions/explode>

NEW QUESTION 188

- (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 192

- (Exam Topic 3)

You are designing an Azure Data Lake Storage solution that will transform raw JSON files for use in an analytical workload.

You need to recommend a format for the transformed files. The solution must meet the following requirements:

- Contain information about the data types of each column in the files.
- Support querying a subset of columns in the files.
- Support read-heavy analytical workloads.
- Minimize the file size.

What should you recommend?

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

Answer: D

Explanation:

Parquet, an open-source file format for Hadoop, stores nested data structures in a flat columnar format. Compared to a traditional approach where data is stored in a row-oriented approach, Parquet file format is more efficient in terms of storage and performance.

It is especially good for queries that read particular columns from a “wide” (with many columns) table since only needed columns are read, and IO is minimized.

Reference: <https://www.clairvoyant.ai/blog/big-data-file-formats>

NEW QUESTION 194

- (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

Answer Area

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.



- 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 195

- (Exam Topic 3)

You are creating an Azure Data Factory data flow that will ingest data from a CSV file, cast columns to specified types of data, and insert the data into a table in an Azure Synapse Analytic dedicated SQL pool. The CSV file contains three columns named username, comment, and date.

The data flow already contains the following:

- > A source transformation.
- > A Derived Column transformation to set the appropriate types of data.
- > A sink transformation to land the data in the pool.

You need to ensure that the data flow meets the following requirements:

- > All valid rows must be written to the destination table.
- > Truncation errors in the comment column must be avoided proactively.
- > Any rows containing comment values that will cause truncation errors upon insert must be written to a file in blob storage.

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

- A. To the data flow, add a sink transformation to write the rows to a file in blob storage.
- B. To the data flow, add a Conditional Split transformation to separate the rows that will cause truncation errors.
- C. To the data flow, add a filter transformation to filter out rows that will cause truncation errors.
- D. Add a select transformation to select only the rows that will cause truncation errors.

Answer: AB

Explanation:

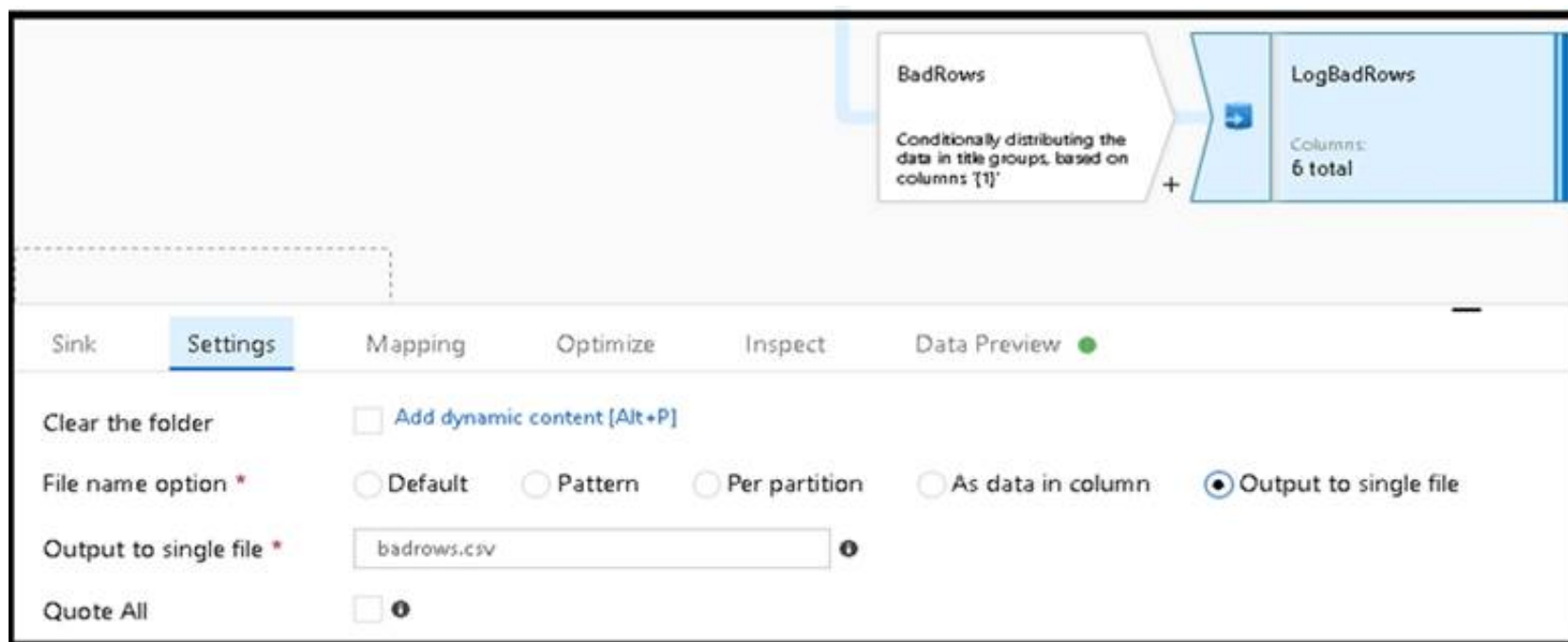
B: Example:

* 1. This conditional split transformation defines the maximum length of "title" to be five. Any row that is less than or equal to five will go into the GoodRows stream. Any row that is larger than five will go into the BadRows stream.

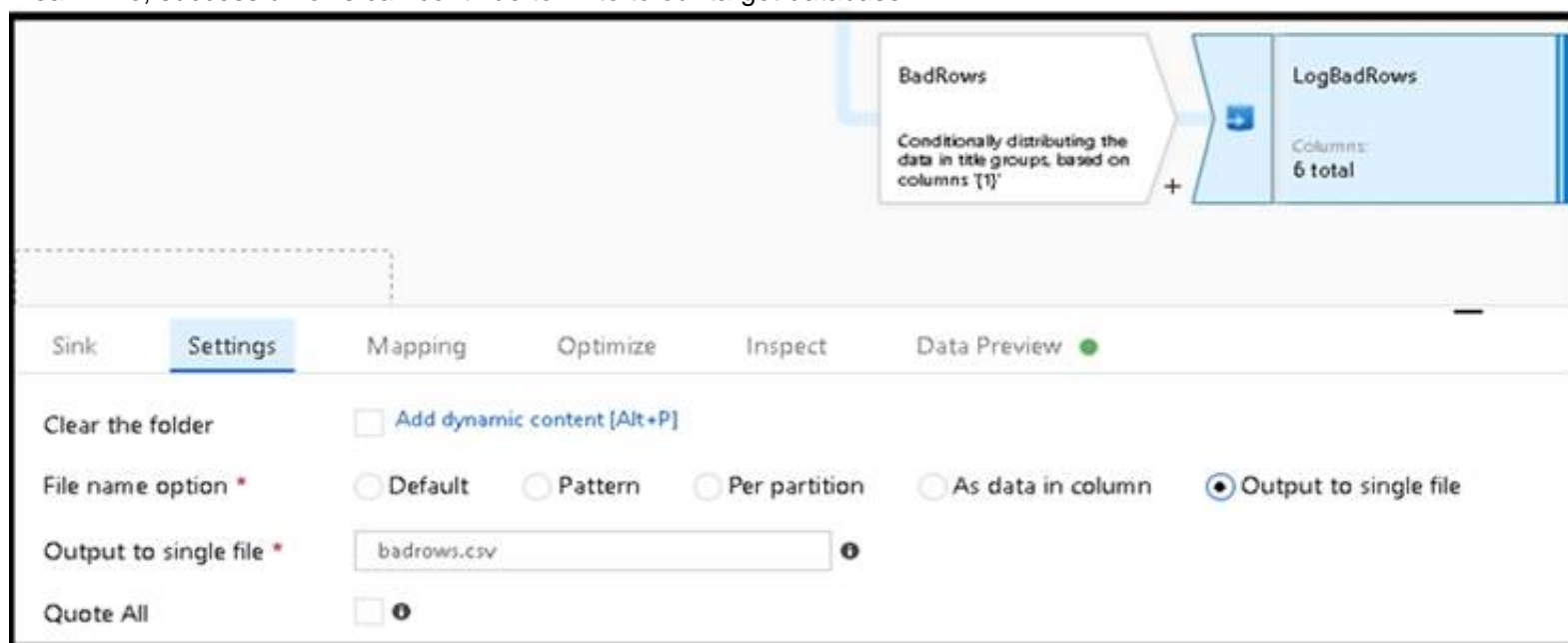


* 2. This conditional split transformation defines the maximum length of "title" to be five. Any row that is less than or equal to five will go into the GoodRows stream. Any row that is larger than five will go into the BadRows stream. A:

* 3. Now we need to log the rows that failed. Add a sink transformation to the BadRows stream for logging. Here, we'll "auto-map" all of the fields so that we have logging of the complete transaction record. This is a text-delimited CSV file output to a single file in Blob Storage. We'll call the log file "badrows.csv".



* 4. The completed data flow is shown below. We are now able to split off error rows to avoid the SQL truncation errors and put those entries into a log file. Meanwhile, successful rows can continue to write to our target database.



Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/how-to-data-flow-error-rows>

NEW QUESTION 198

- (Exam Topic 3)

You have a data warehouse in Azure Synapse Analytics.

You need to ensure that the data in the data warehouse is encrypted at rest. What should you enable?

- A. Advanced Data Security for this database
- B. Transparent Data Encryption (TDE)
- C. Secure transfer required
- D. Dynamic Data Masking

Answer: B

Explanation:

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

- Support for server encryption is currently provided through the SQL feature called Transparent Data Encryption.
- Client-side encryption of Azure SQL Database data is supported through the Always Encrypted feature. Reference:

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

NEW QUESTION 203

- (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 208

- (Exam Topic 3)

You are responsible for providing access to an Azure Data Lake Storage Gen2 account.

Your user account has contributor access to the storage account, and you have the application ID and access key.

You plan to use PolyBase to load data into an enterprise data warehouse in Azure Synapse Analytics. You need to configure PolyBase to connect the data warehouse to storage account.

Which three components should you create in sequence? To answer, move the appropriate components from the list of components to the answer area and arrange them in the correct order.

Components

Answer Area

a database scoped credential

an asymmetric key

an external data source

a database encryption key

an external file format



- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Components

a database scoped credential

an asymmetric key

an external data source

a database encryption key

an external file format

Answer Area

a database scoped credential

an external data source

an external file format

⬆

⬇

NEW QUESTION 211

- (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 214

- (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.

Values

CustomerKey

HASH

ROUND_ROBIN

REPLICATE

OrderDateKey

SalesOrderNumber

Answer Area

```
CREATE TABLE [dbo].[FactSales]
(
    [ProductKey]          int          NOT NULL
,   [OrderDateKey]       int          NOT NULL
,   [CustomerKey]        int          NOT NULL
,   [SalesOrderNumber]   nvarchar ( 20 ) NOT NULL
,   [OrderQuantity]      smallint     NOT NULL
,   [UnitPrice]          money        NOT NULL
)
WITH
(
    CLUSTERED            COLUMNSTORE    INDEX
,   DISTRIBUTION = Value ([ProductKey])
,   PARTITION ( [ Value ] RANGE RIGHT FOR VALUES
                (20170101,20180101,20190101,20200101,20210101)
                )
)
```

- 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 217

- (Exam Topic 3)

You are incrementally loading data into fact tables in an Azure Synapse Analytics dedicated SQL pool. Each batch of incoming data is staged before being loaded into the fact tables. |

You need to ensure that the incoming data is staged as quickly as possible. |

How should you configure the staging tables? To answer, select the appropriate options in the answer area.

Table distribution:

Table structure:

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Round-robin distribution is recommended for staging tables because it distributes data evenly across all the distributions without requiring a hash column. This can improve the speed of data loading and avoid data skew. Heap tables are recommended for staging tables because they do not have any indexes or partitions that can slow down the data loading process. Heap tables are also easier to truncate and reload than clustered index or columnstore index tables.

NEW QUESTION 222

- (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 225

- (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 230

- (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

	▼
Azure Active Directory (Azure AD)	
a shared access signature (SAS)	
a shared key	

 to authenticate by using

	▼
a managed identity	
a stored access policy	
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 234

- (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
<div><div>{deviceID}</div><div>{mm}/{HH}/{DD}/{MM}/{YYYY}</div><div>{regionID}/{deviceID}</div><div>{regionID}/raw</div><div>{YYYY}/{MM}/{DD}/{HH}</div><div>{YYYY}/{MM}/{DD}/{HH}/{mm}</div><div>raw/{deviceID}</div><div>raw/{regionID}</div></div>	<div>/</div> <div>Value</div> <div>/</div> <div>Value</div> <div>/</div> <div>Value</div> <div>.json</div>

- 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 236

- (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 241

- (Exam Topic 3)

You have data stored in thousands of CSV files in Azure Data Lake Storage Gen2. Each file has a header row followed by a properly formatted carriage return (/r) and line feed (/n).

You are implementing a pattern that batch loads the files daily into an enterprise data warehouse in Azure Synapse Analytics by using PolyBase.
 You need to skip the header row when you import the files into the data warehouse. Before building the loading pattern, you need to prepare the required database objects in Azure Synapse Analytics.
 Which three 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: Each correct selection is worth one point

Actions

Create a database scoped credential that uses Azure Active Directory Application and a Service Principal Key

Create an external data source that uses the abfs location

Use CREATE EXTERNAL TABLE AS SELECT (CETAS) and configure the reject options to specify reject values or percentages

Create an external file format and set the First_Row option

Answer Area

>

<

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

A picture containing timeline Description automatically generated

Step 1: Create an external data source that uses the abfs location

Create External Data Source to reference Azure Data Lake Store Gen 1 or 2 Step 2: Create an external file format and set the First_Row option.

Create External File Format.

Step 3: Use CREATE EXTERNAL TABLE AS SELECT (CETAS) and configure the reject options to specify reject values or percentages

To use PolyBase, you must create external tables to reference your external data. Use reject options.

Note: REJECT options don't apply at the time this CREATE EXTERNAL TABLE AS SELECT statement is run. Instead, they're specified here so that the database can use them at a later time when it imports data from the external table. Later, when the CREATE TABLE AS SELECT statement selects data from the external table, the database will use the reject options to determine the number or percentage of rows that can fail to import before it stops the import.

Reference:

<https://docs.microsoft.com/en-us/sql/relational-databases/polybase/polybase-t-sql-objects> <https://docs.microsoft.com/en-us/sql/t-sql/statements/create-external-table-as-select-transact-sql>

NEW QUESTION 246

- (Exam Topic 3)

You plan to create a real-time monitoring app that alerts users when a device travels more than 200 meters away from a designated location.

You need to design an Azure Stream Analytics job to process the data for the planned app. The solution must minimize the amount of code developed and the number of technologies used.

What should you include in the Stream Analytics job? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Input type:

▼

Stream

Reference

Function:

▼

Aggregate

Geospatial

Windowing

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Diagram, table Description automatically generated

Input type: Stream

You can process real-time IoT data streams with Azure Stream Analytics. Function: Geospatial

With built-in geospatial functions, you can use Azure Stream Analytics to build applications for scenarios such as fleet management, ride sharing, connected cars, and asset tracking.

Note: In a real-world scenario, you could have hundreds of these sensors generating events as a stream. Ideally, a gateway device would run code to push these events to Azure Event Hubs or Azure IoT Hubs.

Reference:

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-get-started-with-azure-stream-analytic> <https://docs.microsoft.com/en-us/azure/stream-analytics/geospatial-scenarios>

NEW QUESTION 249

- (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 250

- (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 254

- (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 257

- (Exam Topic 3)

You are implementing Azure Stream Analytics windowing functions.

Which windowing function should you use for each requirement? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

Segment the data stream into distinct time segments that repeat but do not overlap: ☐ Hopping ☐ Sliding ☐ Tumbling

Segment the data stream into distinct time segments that repeat and can overlap: ☐ Hopping ☐ Sliding ☐ Tumbling

Segment the data stream to produce an output only when an event occurs: ☐ Hopping ☐ Sliding ☐ Tumbling

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Answer Area

Segment the data stream into distinct time segments that repeat but do not overlap: ☒ Hopping ☒ Sliding ☐ Tumbling

Segment the data stream into distinct time segments that repeat and can overlap: ☐ Hopping ☒ Sliding ☐ Tumbling

Segment the data stream to produce an output only when an event occurs: ☐ Hopping ☒ Sliding ☐ Tumbling

NEW QUESTION 262

- (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 Azure Data Factor Studio, 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 265

- (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 are designing an Azure Stream Analytics solution that will analyze Twitter data.

You need to count the tweets in each 10-second window. The solution must ensure that each tweet is counted only once.

Solution: You use a tumbling window, and you set the window size to 10 seconds. Does this meet the goal?

- A. Yes
- B. No

Answer: A

Explanation:

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.

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



```
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 269

.....

THANKS FOR TRYING THE DEMO OF OUR PRODUCT

Visit Our Site to Purchase the Full Set of Actual DP-203 Exam Questions With Answers.

We Also Provide Practice Exam Software That Simulates Real Exam Environment And Has Many Self-Assessment Features. Order the DP-203 Product From:

<https://www.2passeasy.com/dumps/DP-203/>

Money Back Guarantee

DP-203 Practice Exam Features:

- * DP-203 Questions and Answers Updated Frequently
- * DP-203 Practice Questions Verified by Expert Senior Certified Staff
- * DP-203 Most Realistic Questions that Guarantee you a Pass on Your FirstTry
- * DP-203 Practice Test Questions in Multiple Choice Formats and Updatesfor 1 Year