

Microsof.DP-203.vDec-2024.by.Danie.173q

Number: DP-203
Passing Score: 800
Time Limit: 120
File Version: 13.0

Exam Code: DP-203
Exam Name: Data Engineering on Microsoft Azure



Case 01 - Design and develop data processing

This is a case study. Case studies are not timed separately. You can use as much exam time as you would like to complete each case. However, there may be additional case studies and sections on this exam. You must manage your time to ensure that you are able to complete all questions included on this exam in the time provided.

To answer the questions included in a case study, you will need to reference information that is provided in the case study. Case studies might contain exhibits and other resources that provide more information about the scenario that is described in the case study. Each question is independent of the other questions in this case study.

At the end of this case study, a review screen will appear. This screen allows you to review your answers and to make changes before you move to the next section of the exam. After you begin a new section, you cannot return to this section.

To start the case study

To display the first question in this case study, click the Next button. Use the buttons in the left pane to explore the content of the case study before you answer the questions. Clicking these buttons displays information such as business requirements, existing environment, and problem statements. If the case study has an All Information tab, note that the information displayed is identical to the information displayed on the subsequent tabs. When you are ready to answer a question, click the Question button to return to the question.

Overview

Litware, Inc. owns and operates 300 convenience stores across the US. The company sells a variety of packaged foods and drinks, as well as a variety of prepared foods, such as sandwiches and pizzas.

Litware has a loyalty club whereby members can get daily discounts on specific items by providing their membership number at checkout.

Litware employs business analysts who prefer to analyze data by using Microsoft Power BI, and data scientists who prefer analyzing data in Azure Databricks notebooks.

Requirements

Business Goals

Litware wants to create a new analytics environment in Azure to meet the following requirements:

See inventory levels across the stores. Data must be updated as close to real time as possible. Execute ad hoc analytical queries on historical data to identify whether the loyalty club discounts increase sales of the discounted products. Every four hours, notify store employees about how many prepared food items to produce based on historical demand from the sales data.

Technical Requirements

Litware identifies the following technical requirements:

Minimize the number of different Azure services needed to achieve the business goals. Use platform as a service (PaaS) offerings whenever possible and avoid having to provision virtual machines that must be managed by Litware. Ensure that the analytical data store is accessible only to the company's on-premises network and Azure services. Use Azure Active Directory (Azure AD) authentication whenever possible. Use the principle of least privilege when designing security. Stage Inventory data in Azure Data Lake Storage Gen2 before loading the data into the analytical data store. Litware wants to remove transient data from Data Lake Storage once the data is no longer in use. Files that have a modified date that is older than 14 days must be removed.

Limit the business analysts' access to customer contact information, such as phone numbers, because this type of data is not analytically relevant. Ensure that you can quickly restore a copy of the analytical data store within one hour in the event of corruption or accidental deletion.

Planned Environment

Litware plans to implement the following environment:

The application development team will create an Azure event hub to receive real-time sales data, including store number, date, time, product ID, customer loyalty number, price, and discount amount, from the point of sale (POS) system and output the data to data storage in Azure.

Customer data, including name, contact information, and loyalty number, comes from Salesforce, a SaaS application, 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. Daily inventory data comes from a Microsoft SQL server located on a private network. Litware currently has 5 TB of historical sales data and 100 GB of customer data. The company expects approximately 100 GB of new data per month for the next year. Litware will build a custom application named FoodPrep to provide store employees with the calculation results of how many prepared food items to produce every four hours. Litware does not plan to implement Azure ExpressRoute or a VPN between the on-premises network and Azure.

QUESTION 1

HOTSPOT

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.

Hot Area:

Answer Area

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



Answer Area:

Answer Area

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

Section:

Explanation:

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.

Case 02 - Design and develop data processing

This is a case study. Case studies are not timed separately. You can use as much exam time as you would like to complete each case. However, there may be additional case studies and sections on this exam. You must manage your time to ensure that you are able to complete all questions included on this exam in the time provided.

To answer the questions included in a case study, you will need to reference information that is provided in the case study. Case studies might contain exhibits and other resources that provide more information about the scenario that is described in the case study. Each question is independent of the other questions in this case study.

At the end of this case study, a review screen will appear. This screen allows you to review your answers and to make changes before you move to the next section of the exam. After you begin a new section, you cannot return to this section.

To start the case study

To display the first question in this case study, click the Next button. Use the buttons in the left pane to explore the content of the case study before you answer the questions. Clicking these buttons displays information such as business requirements, existing environment, and problem statements. If the case study has an All Information tab, note that the information displayed is identical to the information displayed on the subsequent tabs. When you are ready to answer a question, click the Question button to return to the question.

Overview

Contoso, Ltd. is a clothing retailer based in Seattle. The company has 2,000 retail stores across the United States and an emerging online presence.

The network contains an Active Directory forest named contoso.com. The forest it integrated with an Azure Active Directory (Azure AD) tenant named contoso.com. Contoso has an Azure subscription associated to the contoso.com Azure AD tenant.

Existing Environment

Transactional Data

Contoso has three years of customer, transactional, operational, sourcing, and supplier data comprised of 10 billion records stored across multiple on-premises Microsoft SQL Server servers. The SQL Server instances contain data from various operational systems. The data is loaded into the instances by using SQL Server Integration Services (SSIS) packages.

You estimate that combining all product sales transactions into a company-wide sales transactions dataset will result in a single table that contains 5 billion rows, with one row per transaction.

Most queries targeting the sales transactions data will be used to identify which products were sold in retail stores and which products were sold online during different time periods. Sales transaction data that is older than three years will be removed monthly.

You plan to create a retail store table that will contain the address of each retail store. The table will be approximately 2 MB. Queries for retail store sales will include the retail store addresses.

You plan to create a promotional table that will contain a promotion ID. The promotion ID will be associated to a specific product. The product will be identified by a product ID. The table will be approximately 5 GB.

Streaming Twitter Data

The ecommerce department at Contoso develops an Azure logic app that captures trending Twitter feeds referencing the company's products and pushes the products to Azure Event Hubs.

Planned Changes and Requirements

Planned Changes

Contoso plans to implement the following changes:

Load the sales transaction dataset to Azure Synapse Analytics. Integrate on-premises data stores with Azure Synapse Analytics by using SSIS packages. Use Azure Synapse Analytics to analyze Twitter feeds to assess customer sentiments about products.

Sales Transaction Dataset Requirements

Contoso identifies the following requirements for the sales transaction dataset:

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. Ensure that queries joining and filtering sales transaction records based on product ID complete as quickly as possible. Implement a surrogate key to account for changes to the retail store addresses. Ensure that data storage costs and performance are predictable. Minimize how long it takes to remove old records.

Customer Sentiment Analytics Requirements

Contoso identifies the following requirements for customer sentiment analytics:

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.

Maximize the throughput of ingesting Twitter feeds from Event Hubs to Azure Storage without purchasing additional throughput or capacity units. Store Twitter feeds in Azure Storage by using Event Hubs Capture. The feeds will be converted into Parquet files. Ensure that the data store supports Azure AD-based access control down to the object level. Minimize administrative effort to maintain the Twitter feed data records. Purge Twitter feed data records that are older than two years.

Data Integration Requirements

Contoso identifies the following requirements for data integration:

Use an Azure service that leverages the existing SSIS packages to ingest on-premises data into datasets stored in a dedicated SQL pool of Azure Synapse Analytics and transform the data. Identify a process to ensure that changes to the ingestion and transformation activities can be version-controlled and developed independently by multiple data engineers.

QUESTION 1

DRAG DROP

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 the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

Select and Place:

Actions

Merge changes

Create a pull request

Create a feature branch

Publish changes

Create a repository and a main branch



Answer Area
Vdumps

Correct Answer:

Actions

Answer Area

Create a repository and a main branch

Create a feature branch

Create a pull request

Merge changes

Publish changes



Section:

Explanation:

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>



03 - Design and develop data processing

QUESTION 1

You are monitoring an Azure Stream Analytics job by using metrics in Azure. You discover that during the last 12 hours, the average watermark delay is consistently greater than the configured late arrival tolerance. What is a possible cause of this behavior?

- A. Events whose application timestamp is earlier than their arrival time by more than five minutes arrive as inputs.
- B. There are errors in the input data.
- C. The late arrival policy causes events to be dropped.
- D. The job lacks the resources to process the volume of incoming data.

Correct Answer: D

Section:

Explanation:

Watermark Delay indicates the delay of the streaming data processing job. There are a number of resource constraints that can cause the streaming pipeline to slow down. The watermark delay metric can rise due to: Not enough processing resources in Stream Analytics to handle the volume of input events. To scale up resources, see Understand and adjust Streaming Units. Not enough throughput within the input event brokers, so they are throttled. For possible solutions, see Automatically scale up Azure Event Hubs throughput units. Output sinks are not provisioned with enough capacity, so they are throttled. The possible solutions vary widely based on

the flavor of output service being used.

Reference:

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

QUESTION 2

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

A. Yes

B. No

Correct Answer: B

Section:

Explanation:

If you need to transform data in a way that is not supported by Data Factory, you can create a custom activity with your own data processing logic and use the activity in the pipeline. Note: You can use data transformation activities in Azure Data Factory and Synapse pipelines to transform and process your raw data into predictions and insights at scale.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/transform-data>

QUESTION 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 SQL. A workload that data scientists will use to perform ad hoc analysis in Scala and R. The enterprise architecture team at your company identifies the following standards for Databricks environments:

The data engineers must share a cluster.

The job cluster will be managed by using a request process whereby data scientists and data engineers provide packaged notebooks for deployment to the cluster. All the data scientists must be assigned their own cluster that terminates automatically after 120 minutes of inactivity. Currently, there are three data scientists. You need to create the Databricks clusters for the workloads. Solution: You create a High Concurrency cluster for each data scientist, a High Concurrency cluster for the data engineers, and a Standard cluster for the jobs. Does this meet the goal?

A. Yes

B. No

Correct Answer: B

Section:

Explanation:

Need a High Concurrency cluster for the jobs.

Standard clusters are recommended for a single user. Standard can run workloads developed in any language: Python, R, Scala, and SQL. A high concurrency cluster is a managed cloud resource. The key benefits of high concurrency clusters are that they provide Apache Spark-native fine-grained sharing for maximum resource utilization and minimum query latencies.

Reference:

<https://docs.azuredatabricks.net/clusters/configure.html>

QUESTION 4

HOTSPOT

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.

Hot Area:

Answer Area

```
SELECT sensorId,
       growth = reading -
```

▼
LAG
LAST
LEAD

```
       (reading) OVER (PARTITION BY sensorId
```

▼
LIMIT DURATION
OFFSET
WHEN

```
       (hour,1))
```

```
FROM input
```

Answer Area:

Answer Area

```
SELECT sensorId,
       growth = reading -
```

▼
LAG
LAST
LEAD

```
       (reading) OVER (PARTITION BY sensorId
```

▼
LIMIT DURATION
OFFSET
WHEN

```
       (hour,1))
```

```
FROM input
```

Section:

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>

QUESTION 5

HOTSPOT

You are designing a monitoring solution for a fleet of 500 vehicles. Each vehicle has a GPS tracking device that sends data to an Azure event hub once per minute.

You have a CSV file in an Azure Data Lake Storage Gen2 container. The file maintains the expected geographical area in which each vehicle should be. You need to ensure that when a GPS position is outside the expected area, a message is added to another event hub for processing within 30 seconds. The solution must minimize cost. What should you include in the solution? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.


Hot Area:

Answer Area

Service:
An Azure Synapse Analytics Apache Spark pool
An Azure Synapse Analytics serverless SQL pool
Azure Data Factory
Azure Stream Analytics

Window:
Hopping
No window
Session
Tumbling

Analysis type:
Event pattern matching
Lagged record comparison
Point within polygon
Polygon overlap



Answer Area:

Answer Area

Service:
 An Azure Synapse Analytics Apache Spark pool
 An Azure Synapse Analytics serverless SQL pool
 Azure Data Factory
 Azure Stream Analytics

Window:
 Hopping
 No window
 Session
 Tumbling

Analysis type:
 Event pattern matching
 Lagged record comparison
 Point within polygon
 Polygon overlap



Section:

Explanation:

Box 1: Azure Stream Analytics

Box 2: Hopping

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.

Box 3: Point within polygon

Reference:

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

QUESTION 6

HOTSPOT

You have a self-hosted integration runtime in Azure Data Factory.

The current status of the integration runtime has the following configurations:

Status: Running

Type: Self-Hosted

Version: 4.4.7292.1

Running / Registered Node(s): 1/1

High Availability Enabled: False

Linked Count: 0

Queue Length: 0

Average Queue Duration: 0.00s

The integration runtime has the following node details:

Name: X-M

Status: Running
Version: 4.4.7292.1
Available Memory: 7697MB
CPU Utilization: 6%
Network (In/Out): 1.21KBps/0.83KBps
Concurrent Jobs (Running/Limit): 2/14
Role: Dispatcher/Worker
Credential Status: In Sync

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

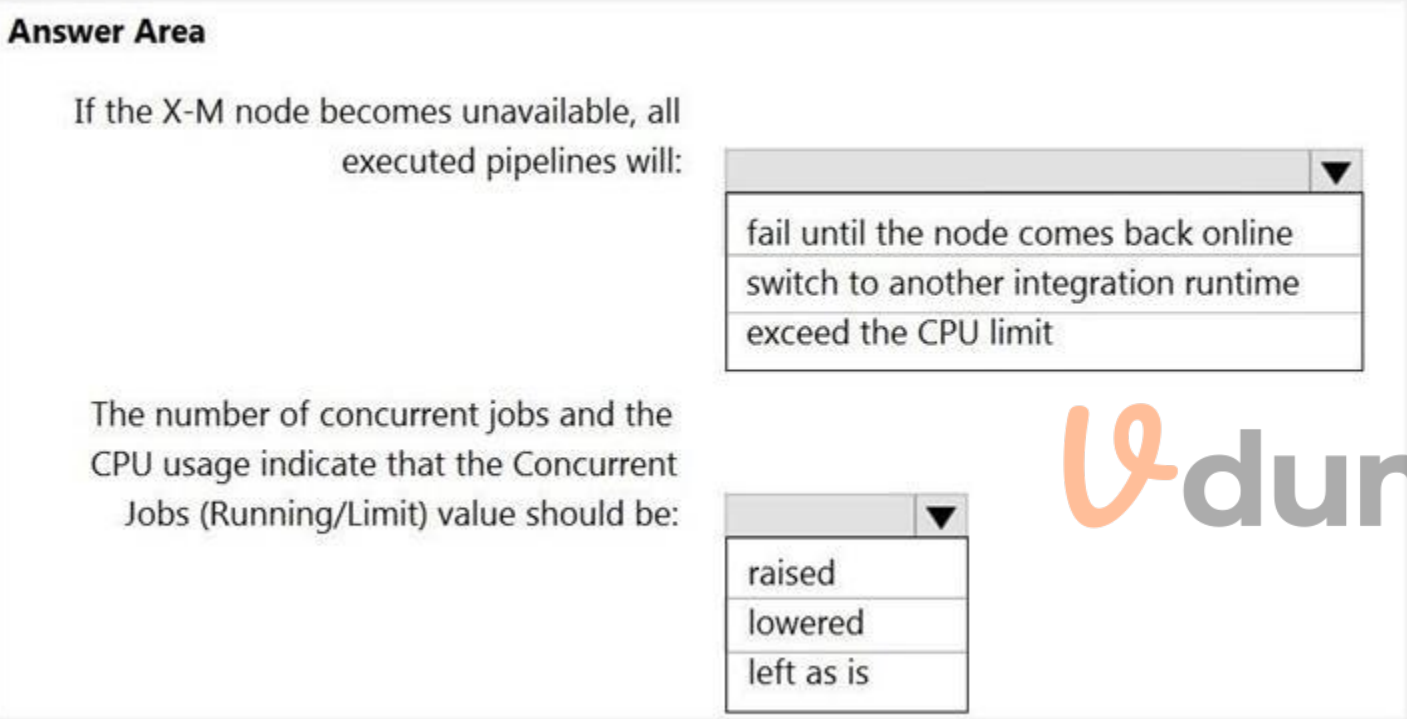
NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

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

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



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

raised
lowered
left as is

Answer Area:

Answer Area

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

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

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

raised
lowered
left as is

Section:

Explanation:

Box 1: fail until the node comes back online

We see: High Availability Enabled: False

Note: Higher availability of the self-hosted integration runtime so that it's no longer the single point of failure in your big data solution or cloud data integration with Data Factory.

Box 2: lowered

We see:

Concurrent Jobs (Running/Limit): 2/14

CPU Utilization: 6%

Note: When the processor and available RAM aren't well utilized, but the execution of concurrent jobs reaches a node's limits, scale up by increasing the number of concurrent jobs that a node can run

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/create-self-hosted-integration-runtime>

QUESTION 7

You are designing an Azure Databricks cluster that runs user-defined local processes. You need to recommend a cluster configuration that meets the following requirements:

Minimize query latency.

Maximize the number of users that can run queries on the cluster at the same time. Reduce overall costs without compromising other requirements. Which cluster type should you recommend?

- A. Standard with Auto Termination
- B. High Concurrency with Autoscaling
- C. High Concurrency with Auto Termination
- D. Standard with Autoscaling

Correct Answer: B

Section:

Explanation:

A High Concurrency cluster is a managed cloud resource. The key benefits of High Concurrency clusters are that they provide fine-grained sharing for maximum resource utilization and minimum query latencies. Databricks

chooses the appropriate number of workers required to run your job. This is referred to as autoscaling. Autoscaling makes it easier to achieve high cluster utilization, because you don't need to provision the cluster to match a workload.

Incorrect Answers:

C: The cluster configuration includes an auto terminate setting whose default value depends on cluster mode:

Standard and Single Node clusters terminate automatically after 120 minutes by default. High Concurrency clusters do not terminate automatically by default.

Reference:

<https://docs.microsoft.com/en-us/azure/databricks/clusters/configure>

QUESTION 8

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.

After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You plan to create an Azure Databricks workspace that has a tiered structure. The workspace will contain the following three workloads:

A workload for data engineers who will use Python and SQL.

A workload for jobs that will run notebooks that use Python, Scala, and SQL. A workload that data scientists will use to perform ad hoc analysis in Scala and R.

The enterprise architecture team at your company identifies the following standards for Databricks environments:

The data engineers must share a 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 Standard cluster for the jobs.

Does this meet the goal?

A. Yes

B. No

Correct Answer: B

Section:

Explanation:

We would need a High Concurrency cluster for the jobs.

Note:

Standard clusters are recommended for a single user. Standard can run workloads developed in any language: Python, R, Scala, and SQL. A high concurrency cluster is a managed cloud resource. The key benefits of high concurrency clusters are that they provide Apache Spark-native fine-grained sharing for maximum resource utilization and minimum query latencies.

Reference: <https://docs.azuredatabricks.net/clusters/configure.html>

QUESTION 9

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.

After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You plan to create an Azure Databricks workspace that has a tiered structure. The workspace will contain the following three workloads:

A workload for data engineers who will use Python and SQL.

A workload for jobs that will run notebooks that use Python, Scala, and SQL. A workload that data scientists will use to perform ad hoc analysis in Scala and R.

The enterprise architecture team at your company identifies the following standards for Databricks environments:

The data engineers must share a 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

Correct Answer: A

Section:

Explanation:

We need a High Concurrency cluster for the data engineers and the jobs. Note: Standard clusters are recommended for a single user. Standard can run workloads developed in any language: Python, R, Scala, and SQL. A high concurrency cluster is a managed cloud resource. The key benefits of high concurrency clusters are that they provide Apache Spark-native fine-grained sharing for maximum resource utilization and minimum query latencies.

Reference: <https://docs.azuredatabricks.net/clusters/configure.html>

QUESTION 10

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.

Correct Answer: D, F

Section:

Explanation:

D: Scale out the query by allowing the system to process each input partition separately. F: A Stream Analytics job definition includes inputs, a query, and output. Inputs are where the job reads the data stream from.

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

QUESTION 11

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

Correct Answer: C

Section:

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>

QUESTION 12

You plan to perform batch processing in Azure Databricks once daily. Which type of Databricks cluster should you use?

- A. High Concurrency
- B. automated

C. interactive

Correct Answer: C

Section:

Explanation:

Azure Databricks has two types of clusters: interactive and automated. You use interactive clusters to analyze data collaboratively with interactive notebooks. You use automated clusters to run fast and robust automated jobs.

Example: Scheduled batch workloads (data engineers running ETL jobs) This scenario involves running batch job JARs and notebooks on a regular cadence through the Databricks platform. The suggested best practice is to

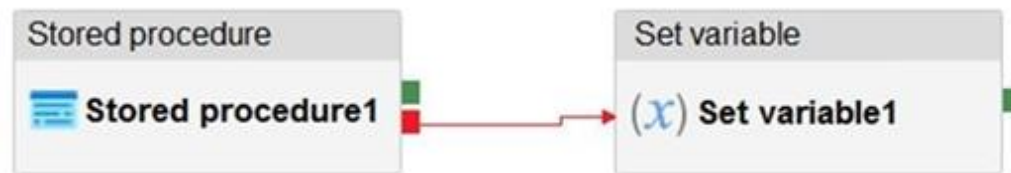
launch a new cluster for each run of critical jobs. This helps avoid any issues (failures, missing SLA, and so on) due to an existing workload (noisy neighbor) on a shared cluster.

Reference: <https://docs.databricks.com/administration-guide/cloud-configurations/aws/cmbp.html#scenario-3-scheduled-batch-workloads-data-engineers-running-etl-jobs>

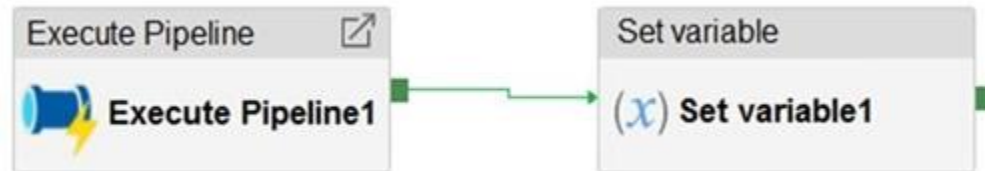
QUESTION 13

You have an Azure Data Factory instance that contains two pipelines named Pipeline1 and Pipeline2.

Pipeline1 has the activities shown in the following exhibit.



Pipeline2 has the activities shown in the following exhibit.



You execute Pipeline2, and Stored procedure1 in Pipeline1 fails.

What is the status of the pipeline runs?

- A. Pipeline1 and Pipeline2 succeeded.
- B. Pipeline1 and Pipeline2 failed.
- C. Pipeline1 succeeded and Pipeline2 failed.
- D. Pipeline1 failed and Pipeline2 succeeded.

Correct Answer: A

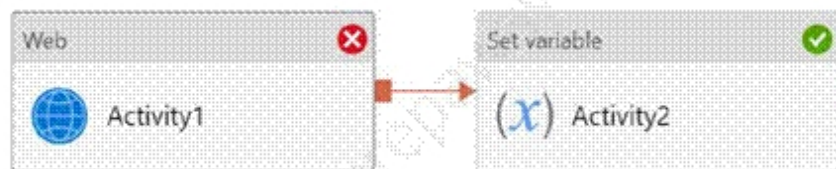
Section:

Explanation:

Activities are linked together via dependencies. A dependency has a condition of one of the following: Succeeded, Failed, Skipped, or Completed.

Consider Pipeline1:

If we have a pipeline with two activities where Activity2 has a failure dependency on Activity1, the pipeline will not fail just because Activity1 failed. If Activity1 fails and Activity2 succeeds, the pipeline will succeed. This scenario is treated as a try-catch block by Data Factory.



The failure dependency means this pipeline reports success.

Note:

If we have a pipeline containing Activity1 and Activity2, and Activity2 has a success dependency on Activity1, it will only execute if Activity1 is successful. In this scenario, if Activity1 fails, the pipeline will fail.

Reference:

<https://datasavvy.me/category/azure-data-factory/>



QUESTION 14

You have an Azure Data Factory that contains 10 pipelines.

You need to label each pipeline with its main purpose of either ingest, transform, or load. The labels must be available for grouping and filtering when using the monitoring experience in Data Factory. What should you add to each pipeline?

- A. a resource tag
- B. a correlation ID
- C. a run group ID
- D. an annotation

Correct Answer: D

Section:

Explanation:

Annotations are additional, informative tags that you can add to specific factory resources: pipelines, datasets, linked services, and triggers. By adding annotations, you can easily filter and search for specific factory resources.

Reference:

<https://www.cathrinewilhelmsen.net/annotations-user-properties-azure-data-factory/>

QUESTION 15

You are designing a statistical analysis solution that will use custom proprietary Python functions on near real-time data from Azure Event Hubs. You need to recommend which Azure service to use to perform the statistical analysis. The solution must minimize latency. What should you recommend?

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



Correct Answer: C

Section:

Explanation:

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

QUESTION 16

You need to implement a Type 3 slowly changing dimension (SCD) for product category data in an Azure Synapse Analytics dedicated SQL pool.

You have a table that was created by using the following Transact-SQL statement. Which two columns should you add to the table? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. [EffectiveStartDate] [datetime] NOT NULL,
- B. [CurrentProductCategory] [nvarchar] (100) NOT NULL,
- C. [EffectiveEndDate] [datetime] NULL,
- D. [ProductCategory] [nvarchar] (100) NOT NULL,
- E. [OriginalProductCategory] [nvarchar] (100) NOT NULL,

Correct Answer: B, E

Section:

Explanation:

A Type 3 SCD supports storing two versions of a dimension member as separate columns. The table includes a column for the current value of a member plus either the original or previous value of the member. So Type 3 uses additional columns to track one key instance of history, rather than storing additional rows to track each change like in a Type 2 SCD.

This type of tracking may be used for one or two columns in a dimension table. It is not common to use it for many members of the same table. It is often used in combination with Type 1 or Type 2 members.

CustomerID	FirstName	LastName	CurrentEmail	OriginalEmail	CompanyName	InsertedDate	ModifiedDate
2	Keith	Harris	keith0@aw.com	keith0@aw.com	Progressive Sports	2021-03-20	2021-03-20
3	Donna	Carreras	donna0@aw.com	donna0@aw.com	A Bike Store	2021-03-20	2021-03-20

CustomerID	FirstName	LastName	CurrentEmail	OriginalEmail	CompanyName	InsertedDate	ModifiedDate
2	Keith	Harris	keith0@aw.com	keith0@aw.com	Progressive Sports	2021-03-20	2021-03-20
3	Donna	Carreras	dc3@aw.com	donna0@aw.com	A Bike Store	2021-03-20	2021-03-22

Reference:

<https://k21academy.com/microsoft-azure/azure-data-engineer-dp203-q-a-day-2-live-session-review/>

QUESTION 17

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 10 seconds and a window size of 10 seconds. Does this meet the goal?

- A. Yes
- B. No

Correct Answer: B

Section:

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>

QUESTION 18

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

Correct Answer: B

Section:

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>

QUESTION 19

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.

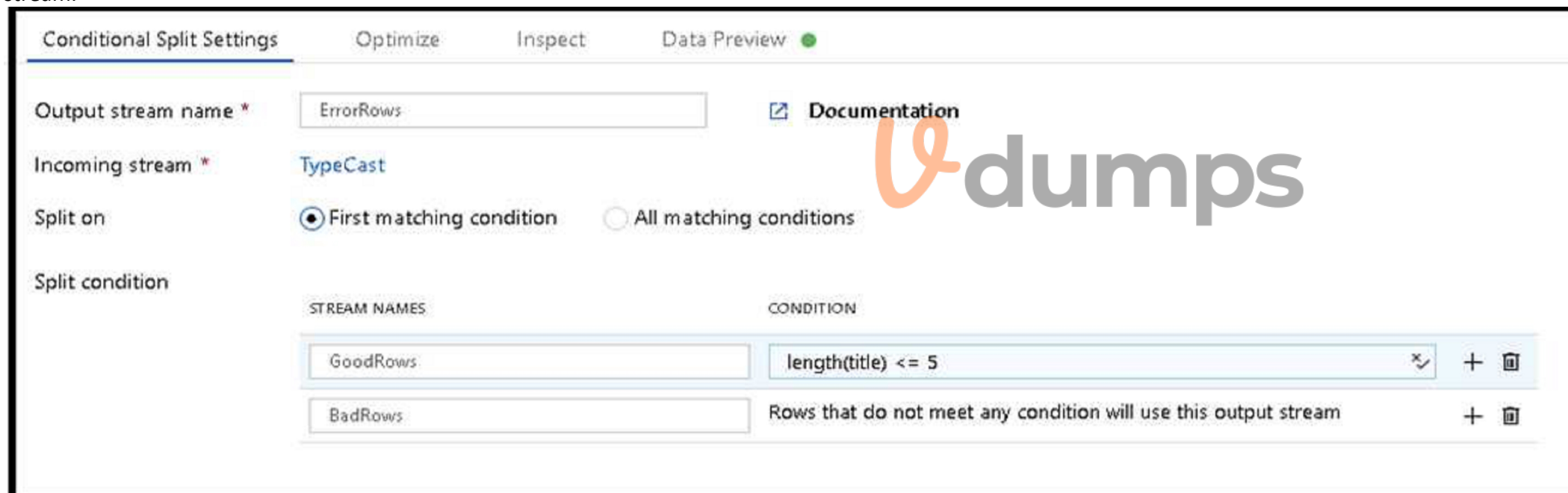
Correct Answer: A, B

Section:

Explanation:

B: Example:

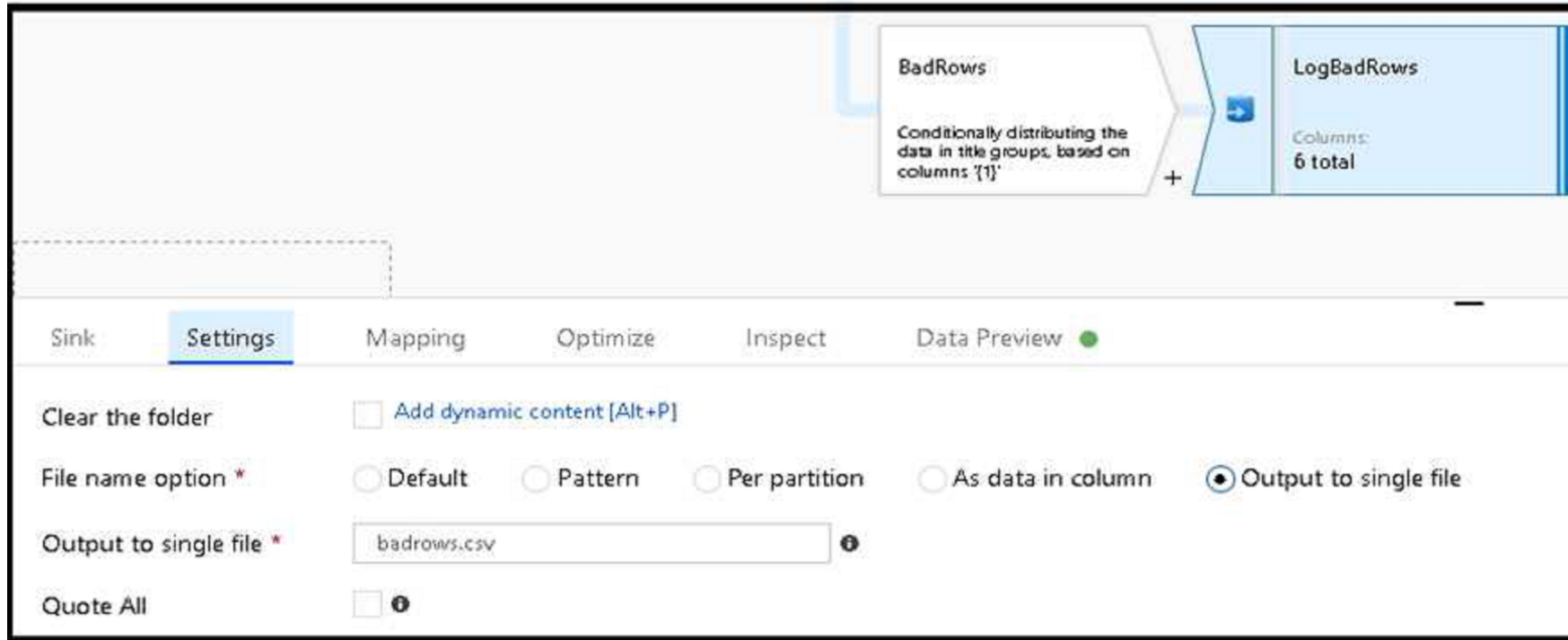
1. This conditional split transformation defines the maximum length of "title" to be five. Any row that is less than or equal to five will go into the GoodRows stream. Any row that is larger than five will go into the BadRows stream.



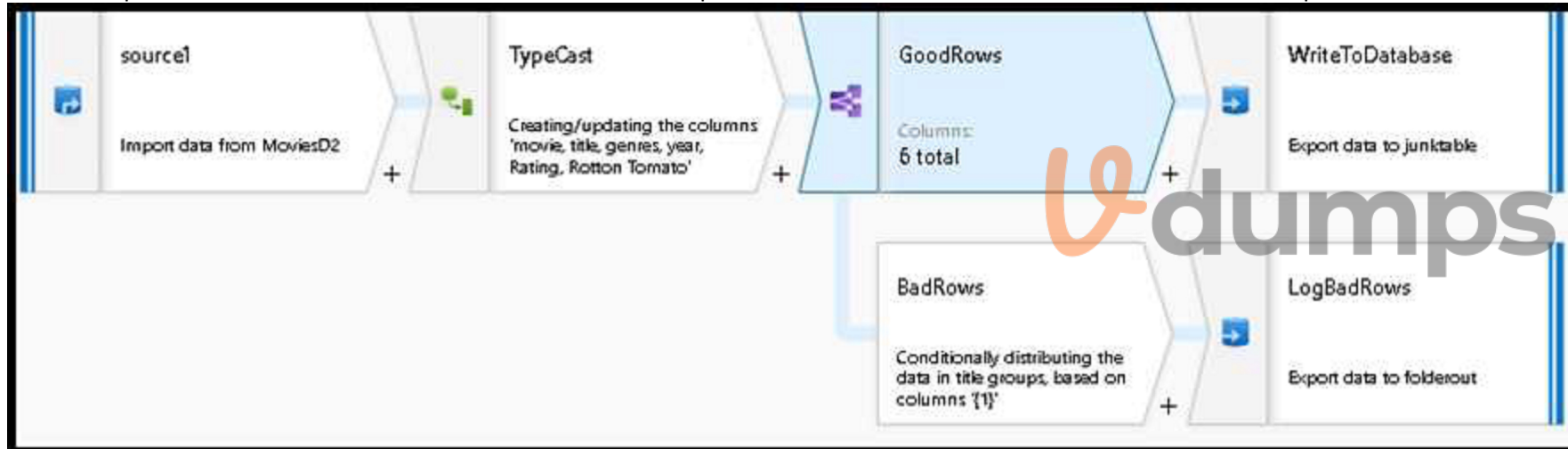
2. This conditional split transformation defines the maximum length of "title" to be five. Any row that is less than or equal to five will go into the GoodRows stream. Any row that is larger than five will go into the BadRows stream.

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>

QUESTION 20

You have an Azure Storage account and a data warehouse in Azure Synapse Analytics in the UK South region. You need to copy blob data from the storage account to the data warehouse by using Azure Data Factory. The solution must meet the following requirements:

Ensure that the data remains in the UK South region at all times. Minimize administrative effort. Which type of integration runtime should you use?

- A. Azure integration runtime
- B. Azure-SSIS integration runtime
- C. Self-hosted integration runtime

Correct Answer: A

Section:

Explanation:

IR type	Public network	Private network
Azure	Data Flow Data movement Activity dispatch	
Self-hosted	Data movement Activity dispatch	Data movement Activity dispatch
Azure-SSIS	SSIS package execution	SSIS package execution

Incorrect Answers:

C: Self-hosted integration runtime is to be used On-premises.

Reference:

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

QUESTION 21

You have an Azure Stream Analytics job that receives clickstream data from an Azure event hub. You need to define a query in the Stream Analytics job. The query must meet the following requirements: Count the number of clicks within each 10-second window based on the country of a visitor. Ensure that each click is NOT counted more than once. How should you define the Query?

- A. SELECT Country, Avg(*) AS Average
FROM ClickStream TIMESTAMP BY CreatedAt
GROUP BY Country, SlidingWindow(second, 10)
- B. SELECT Country, Count(*) AS Count
FROM ClickStream TIMESTAMP BY CreatedAt
GROUP BY Country, TumblingWindow(second, 10)
- C. SELECT Country, Avg(*) AS Average
FROM ClickStream TIMESTAMP BY CreatedAt
GROUP BY Country, HoppingWindow(second, 10, 2)
- D. SELECT Country, Count(*) AS Count
FROM ClickStream TIMESTAMP BY CreatedAt
GROUP BY Country, SessionWindow(second, 5, 10)



Correct Answer: B

Section:

Explanation:

Tumbling window functions are used to segment a data stream into distinct time segments and perform a function against them, such as the example below. The key differentiators of a Tumbling window are that they repeat, do not overlap, and an event cannot belong to more than one tumbling window. Example:

Incorrect Answers:

A: Sliding windows, unlike Tumbling or Hopping windows, output events only for points in time when the content of the window actually changes. In other words, when an event enters or exits the window. Every window has at least one event, like in the case of Hopping windows, events can belong to more than one sliding window. C: 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, so 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. D: Session windows group events that arrive at similar times, filtering out periods of time where there is no data.

Reference:

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

QUESTION 22

You need to schedule an Azure Data Factory pipeline to execute when a new file arrives in an Azure Data Lake Storage Gen2 container. Which type of trigger should you use?

- A. on-demand
- B. tumbling window
- C. schedule
- D. event

Correct Answer: D

Section:

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.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/how-to-create-event-trigger>

QUESTION 23

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.



Correct Answer: C

Section:

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>

QUESTION 24

You are developing a solution that will stream to Azure Stream Analytics. The solution will have both streaming data and reference data. Which input type should you use for the reference data?

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

Correct Answer: B

Section:

Explanation:

Stream Analytics supports Azure Blob storage and Azure SQL Database as the storage layer for Reference Data.

Reference:
<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-use-reference-data>

QUESTION 25

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?

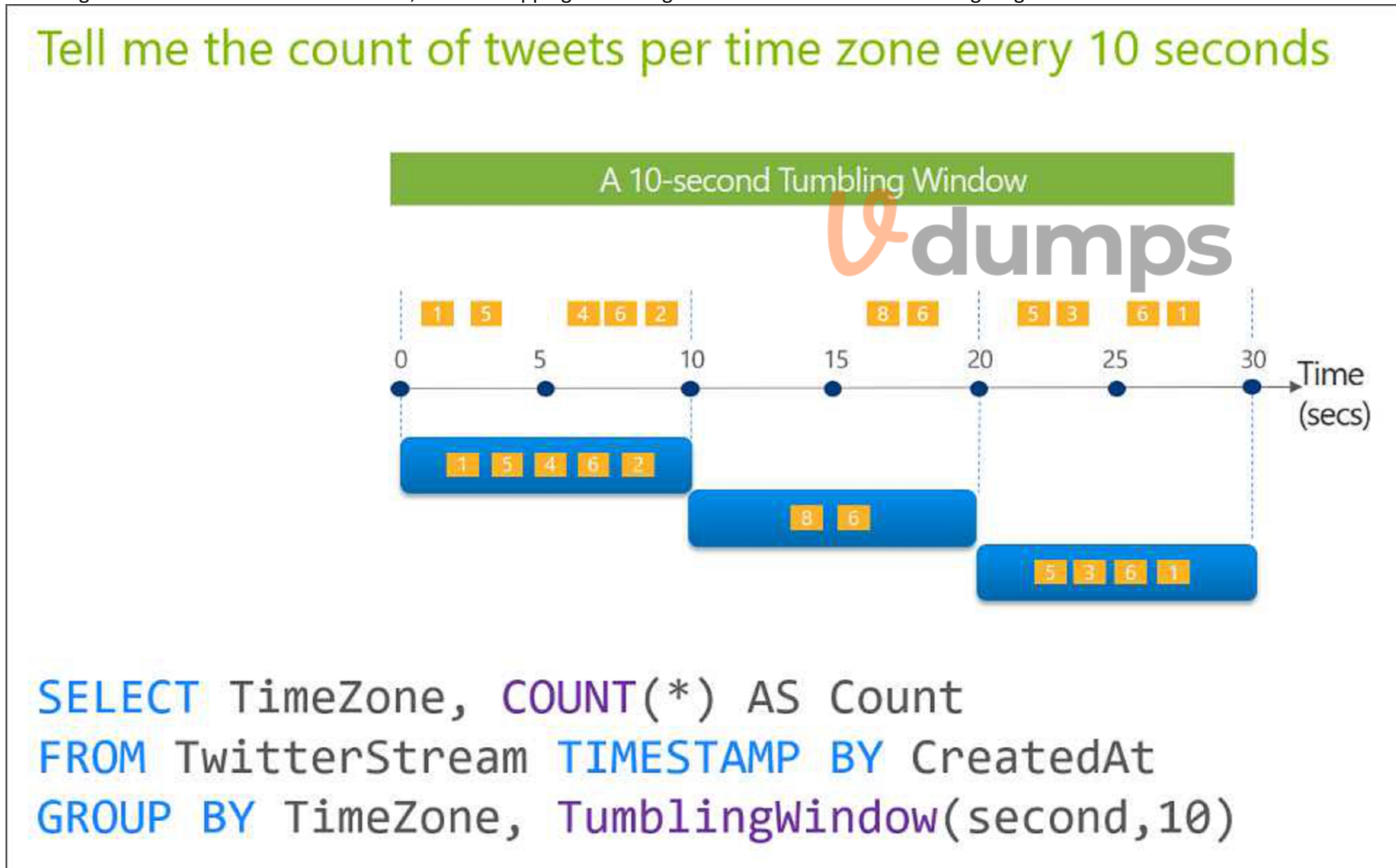
- A. snapshot
- B. tumbling
- C. hopping
- D. sliding

Correct Answer: B

Section:

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.



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

QUESTION 26

You are designing an Azure Databricks table. The table will ingest an average of 20 million streaming events per day. You need to persist the events in the table for use in incremental load pipeline jobs in Azure Databricks. The solution must minimize storage costs and incremental load times. What should you include in the solution?

- A. Partition by DateTime fields.
- B. Sink to Azure Queue storage.
- C. Include a watermark column.
- D. Use a JSON format for physical data storage.

Correct Answer: B

Section:

Explanation:

The Databricks ABS-AQS connector uses Azure Queue Storage (AQS) to provide an optimized file source that lets you find new files written to an Azure Blob storage (ABS) container without repeatedly listing all of the files. This provides two major advantages:

Lower latency: no need to list nested directory structures on ABS, which is slow and resource intensive. Lower costs: no more costly LIST API requests made to ABS.

Reference:

<https://docs.microsoft.com/en-us/azure/databricks/spark/latest/structured-streaming/aqs>

QUESTION 27

You have an Azure Databricks workspace named workspace1 in the Standard pricing tier. You need to configure workspace1 to support autoscaling all-purpose clusters. The solution must meet the following requirements: Automatically scale down workers when the cluster is underutilized for three minutes. Minimize the time it takes to scale to the maximum number of workers. Minimize costs. What should you do first?

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



Correct Answer: B

Section:

Explanation:

For clusters running Databricks Runtime 6.4 and above, optimized autoscaling is used by all-purpose clusters in the Premium plan. Optimized autoscaling:

Scales up from min to max in 2 steps.

Can scale down even if the cluster is not idle by looking at shuffle file state. Scales down based on a percentage of current nodes.

On job clusters, scales down if the cluster is underutilized over the last 40 seconds. On all-purpose clusters, scales down if the cluster is underutilized over the last 150 seconds. The spark.databricks.aggressiveWindowDownS Spark configuration property specifies in seconds how often a cluster makes down-scaling decisions. Increasing the value causes a cluster to scale down more slowly. The maximum value is 600.

Note: Standard autoscaling

Starts with adding 8 nodes. Thereafter, scales up exponentially, but can take many steps to reach the max. You can customize the first step by setting the spark.databricks.autoscaling.standardFirstStepUp Spark configuration property. Scales down only when the cluster is completely idle and it has been underutilized for the last 10 minutes. Scales down exponentially, starting with 1 node.

Reference:

<https://docs.databricks.com/clusters/configure.html>

QUESTION 28

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?

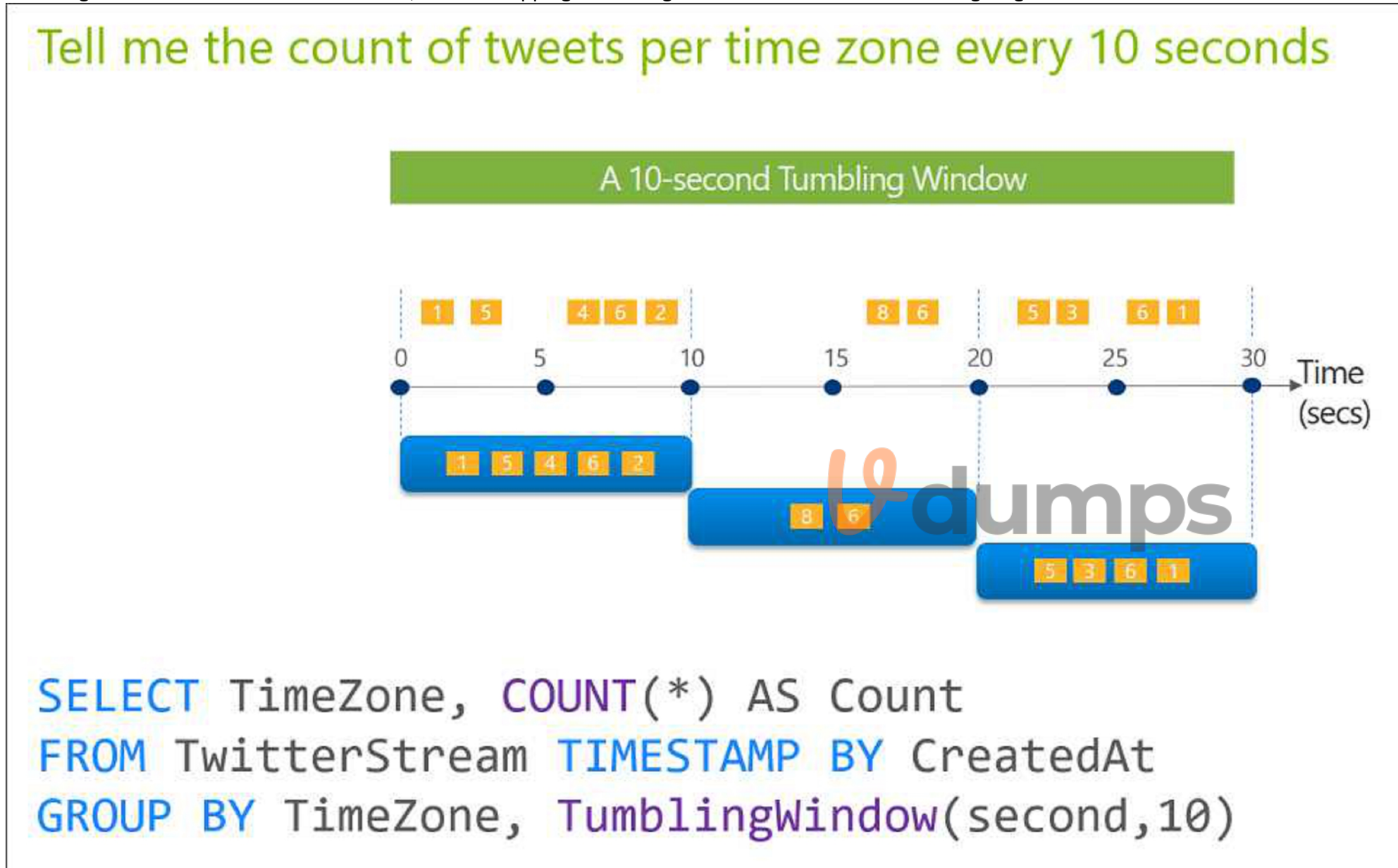
- A. Yes
- B. No

Correct Answer: A

Section:

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.



Reference:

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

QUESTION 29

HOTSPOT

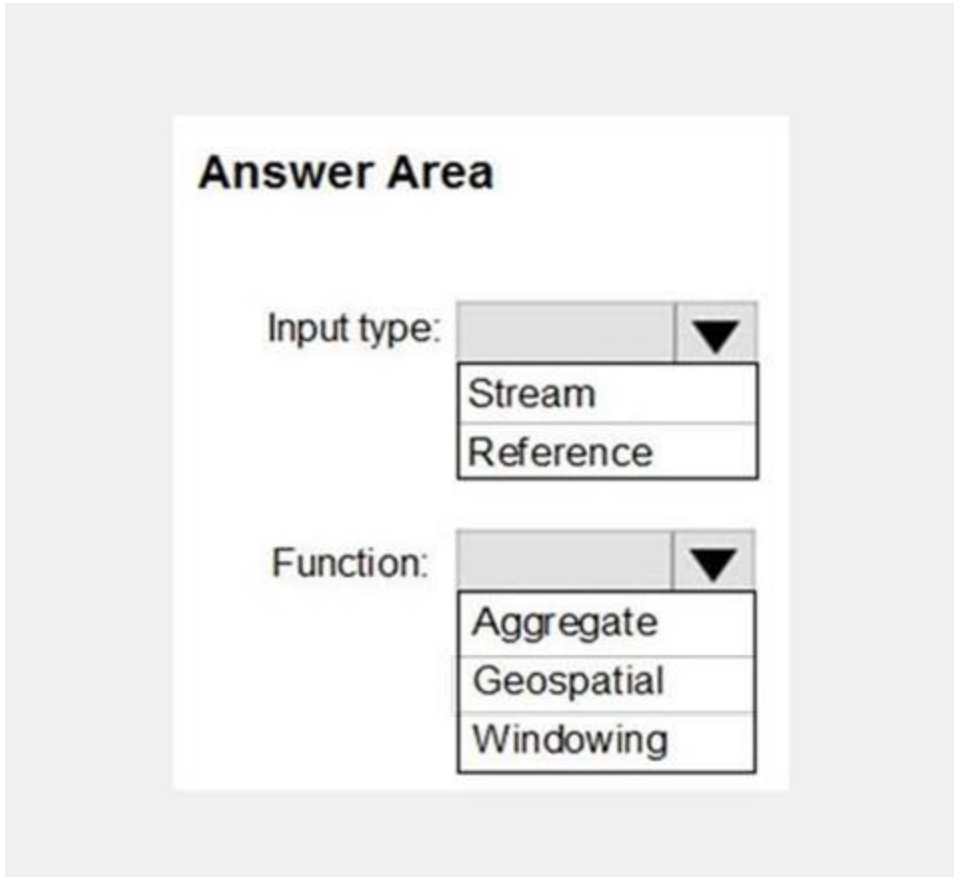
You plan to create a real-time monitoring app that alerts users when a device travels more than 200 meters away from a designated location.

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.

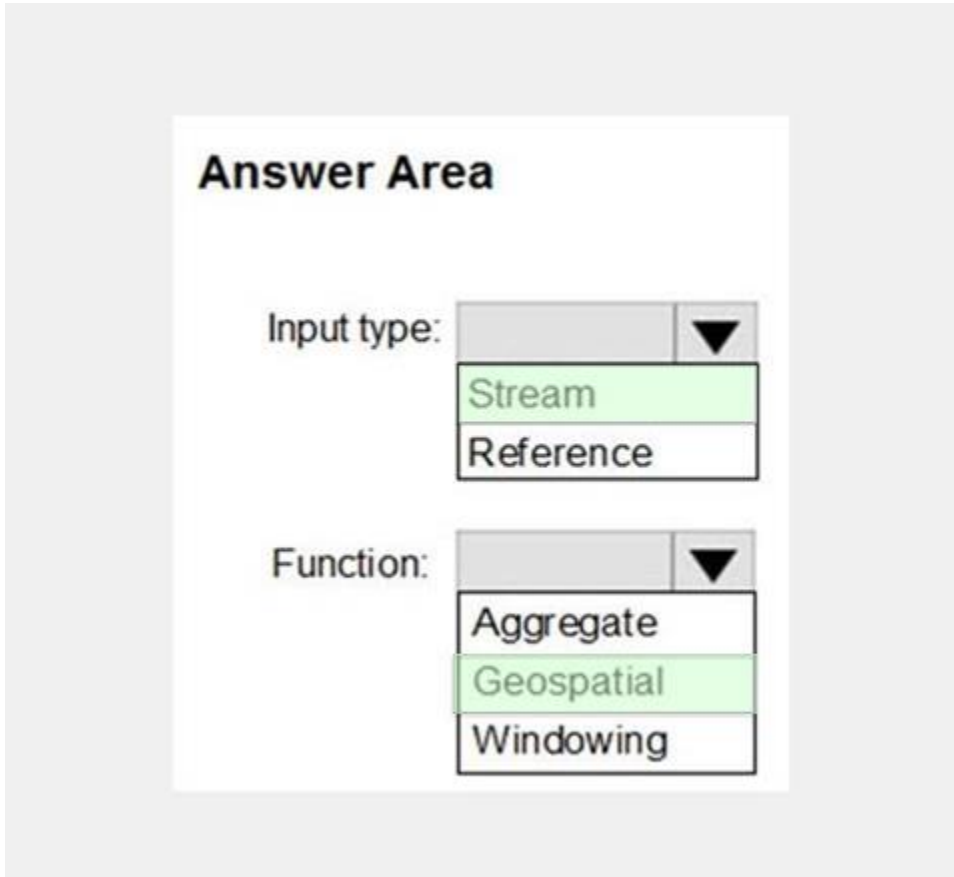
What should you include in the Stream Analytics job? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:



Answer Area:



Section:

Explanation:

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-analytics-to-process-data-from-iot-devices>

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

QUESTION 30

HOTSPOT

You are processing streaming data from vehicles that pass through a toll booth.

You need to use Azure Stream Analytics to return the license plate, vehicle make, and hour the last vehicle passed during each 10-minute window.

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

NOTE: Each correct selection is worth one point.

Hot Area:



Answer Area

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



Answer Area:

Answer Area

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



Section:

Explanation:

Box 1: MAX

The first step on the query finds the maximum time stamp in 10-minute windows, that is the time stamp of the last event for that window. The second step joins the results of the first query with the original stream to find the event that match the last time stamps in each window.

Query:

```
WITH LastInWindow AS
(
  SELECT
  MAX(Time) AS LastEventTime
  FROM
```

```
Input TIMESTAMP BY Time
GROUP BY
TumblingWindow(minute, 10)
)
SELECT
Input.License_plate,
Input.Make,
Input.Time
FROM
Input TIMESTAMP BY Time
INNER JOIN LastInWindow
ON DATEDIFF(minute, Input, LastInWindow) BETWEEN 0 AND 10
AND Input.Time = LastInWindow.LastEventTime
```

Box 2: TumblingWindow

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

Box 3: DATEDIFF

DATEDIFF is a date-specific function that compares and returns the time difference between two DateTime fields, for more information, refer to date functions.

Reference:

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

QUESTION 31

HOTSPOT

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

Ingest:

Access multiple data sources.

Provide the ability to orchestrate workflow.

Provide the capability to run SQL Server Integration Services packages.

Store:

Optimize storage for big data workloads.

Provide encryption of data at rest.

Operate with no size limits.

Prepare and Train:

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

Model & Serve:

Implement native columnar storage.

Support for the SQL language

Provide support for structured streaming.

You need to build the data integration pipeline.

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

NOTE: Each correct selection is worth one point.

Hot Area:



Answer Area

Architecture requirement

Technology

Ingest

	▼
Logic Apps	
Azure Data Factory	
Azure Automation	

Store

	▼
Azure Data Lake Storage	
Azure Blob storage	
Azure files	

Prepare and Train

	▼
HDInsight Apache Spark cluster	
Azure Databricks	
HDInsight Apache Storm cluster	

Model and Serve

	▼
HDInsight Apache Kafka cluster	
Azure Synapse Analytics	
Azure Data Lake Storage	

Answer Area:

Answer Area

Architecture requirement

Technology

Ingest

	▼
Logic Apps	
Azure Data Factory	
Azure Automation	

Store

	▼
Azure Data Lake Storage	
Azure Blob storage	
Azure files	

Prepare and Train

	▼
HDInsight Apache Spark cluster	
Azure Databricks	
HDInsight Apache Storm cluster	

Model and Serve

	▼
HDInsight Apache Kafka cluster	
Azure Synapse Analytics	
Azure Data Lake Storage	

Section:

Explanation:

Ingest: Azure Data Factory

Azure Data Factory pipelines can execute SSIS packages.

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

Store: Data Lake Storage

Data Lake Storage Gen1 provides unlimited storage.

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

Prepare and Train: Azure Databricks

Azure Databricks provides enterprise-grade Azure security, including Azure Active Directory integration. With Azure Databricks, you can set up your Apache Spark environment in minutes, autoscale and collaborate on shared projects in an interactive workspace. Azure Databricks supports Python, Scala, R, Java and SQL, as well as data science frameworks and libraries including TensorFlow, PyTorch and scikit-learn.

Model and Serve: Azure Synapse Analytics

Azure Synapse Analytics/ SQL Data Warehouse stores data into relational tables with columnar storage. Azure SQL Data Warehouse connector now offers efficient and scalable structured streaming write support for SQL Data Warehouse. Access SQL Data Warehouse from Azure Databricks using the SQL Data Warehouse connector.

Note: Note: As of November 2019, Azure SQL Data Warehouse is now Azure Synapse Analytics.

Reference:

<https://docs.microsoft.com/bs-latn-ba/azure/architecture/data-guide/technology-choices/pipeline-orchestration-data-movement>

<https://docs.microsoft.com/en-us/azure/azure-databricks/what-is-azure-databricks>

QUESTION 32

DRAG DROP

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.

Select and Place:

Values	Answer Area
	SELECT
	*,
CASE	[]
ELSE	WHEN hire_date >= '2019-01-01' THEN 'New'
OVER	[] 'Standard'
PARTITION BY	END AS employee_type
ROW_NUMBER	FROM
	employees



Correct Answer:

Values	Answer Area
[]	SELECT
[]	*,
[]	CASE
OVER	WHEN hire_date >= '2019-01-01' THEN 'New'
PARTITION BY	ELSE [] 'Standard'
ROW_NUMBER	END AS employee_type
	FROM
	employees

Section:

Explanation:

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>

QUESTION 33

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 session window that uses a timeout size of 10 seconds. Does this meet the goal?

- A. Yes
- B. No

Correct Answer: A

Section:

QUESTION 34

You use Azure Stream Analytics to receive data from Azure Event Hubs and to output the data to an Azure Blob Storage account. You need to output the count of records received from the last five minutes every minute. Which windowing function should you use?

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

Correct Answer: D

Section:

Explanation:

QUESTION 35

You plan to perform batch processing in Azure Databricks once daily. Which type of Databricks cluster should you use?

- A. High Concurrency
- B. interactive
- C. automated

Correct Answer: C

Section:

Explanation:



Automated Databricks clusters are the best for jobs and automated batch processing.

Reference:

<https://docs.microsoft.com/en-us/azure/databricks/clusters/create>

QUESTION 36

You have the following Azure Data Factory pipelines:

Ingest Data from System1

Ingest Data from System2

Populate Dimensions

Populate Facts

Ingest Data from System1 and Ingest Data from System2 have no dependencies. Populate Dimensions must execute after Ingest Data from System1 and Ingest Data from System2. Populate Facts must execute after 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. Add a schedule trigger to all four pipelines.
- C. Create a parent pipeline that contains the four pipelines and use a schedule trigger.
- D. Create a parent pipeline that contains the four pipelines and use an event trigger.

Correct Answer: C

Section:

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>



QUESTION 37

HOTSPOT

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.

Hot Area:

Answer Area

Load methodology:

	▼
Full Load	
Incremental Load	
Load individual files as they arrive	

Trigger:

	▼
Fixed schedule	
New file	
Tumbling window	

Answer Area:



Answer Area

Load methodology:

	▼
Full Load	
Incremental Load	
Load individual files as they arrive	

Trigger:

	▼
Fixed schedule	
New file	
Tumbling window	

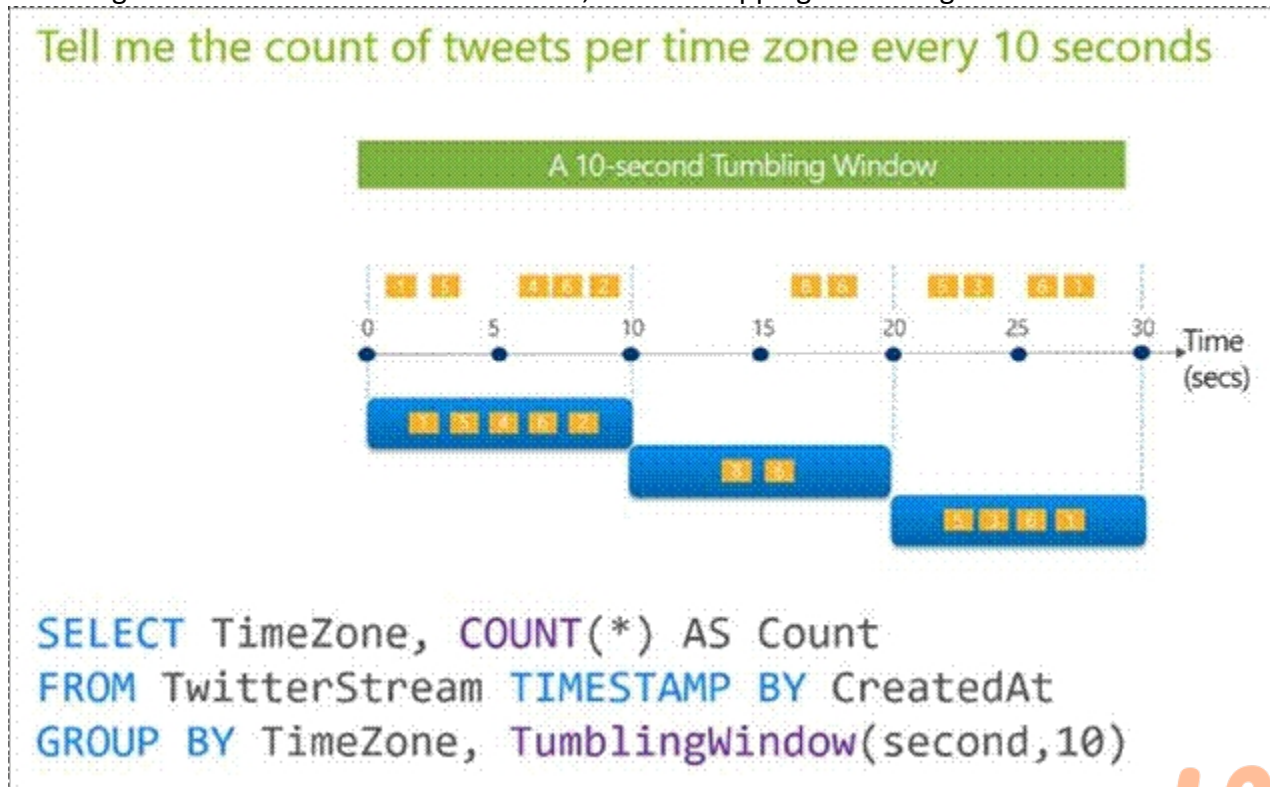
Section:

Explanation:

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.



Reference:

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



QUESTION 38

DRAG DROP

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.

Select and Place:

Components

- a database scoped credential
- an asymmetric key
- an external data source
- a database encryption key
- an external file format



Answer Area



Correct Answer:

Components

-
-
-
- a database encryption key
- an external file format



Answer Area

- an asymmetric key
- a database scoped credential
- an external data source



Section:

Explanation:

Step 1: an asymmetric key

A master key should be created only once in a database. The Database Master Key is a symmetric key used to protect the private keys of certificates and asymmetric keys in the database.

Step 2: a database scoped credential

Create a Database Scoped Credential. A Database Scoped Credential is a record that contains the authentication information required to connect an external resource. The master key needs to be created first before creating the database scoped credential.

Step 3: an external data source

Create an External Data Source. External data sources are used to establish connectivity for data loading using Polybase.

Reference:

<https://www.sqlservercentral.com/articles/access-external-data-from-azure-synapse-analytics-using-polybase>

QUESTION 39

You are creating a new notebook in Azure Databricks that will support R as the primary language but will also support Scala and SQL. Which switch should you use to switch between languages?

- A. %<language>
- B. @<Language >

C. \[<language >]

D. \(<language >)

Correct Answer: A

Section:

Explanation:

To change the language in Databricks' cells to either Scala, SQL, Python or R, prefix the cell with '%', followed by the language. %python //or r, scala, sql

Reference:

<https://www.theta.co.nz/news-blogs/tech-blog/enhancing-digital-twins-part-3-predictive-maintenance-with-azure-databricks>

QUESTION 40

You have an Azure Data Factory pipeline that performs an incremental load of source data to an Azure Data Lake Storage Gen2 account. Data to be loaded is identified by a column named LastUpdatedDate in the source table.

You plan to execute the pipeline every four hours.

You need to ensure that the pipeline execution meets the following requirements:

Automatically retries the execution when the pipeline run fails due to concurrency or throttling limits. Supports backfilling existing data in the table. Which type of trigger should you use?

A. event

B. on-demand

C. schedule

D. tumbling window

Correct Answer: D

Section:

Explanation:

In case of pipeline failures, tumbling window trigger can retry the execution of the referenced pipeline automatically, using the same input parameters, without the user intervention. This can be specified using the property "retryPolicy" in the trigger definition.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/how-to-create-tumbling-window-trigger>

QUESTION 41

You are designing a solution that will copy Parquet files stored in an Azure Blob storage account to an Azure Data Lake Storage Gen2 account. The data will be loaded daily to the data lake and will use a folder structure of {Year}/{Month}/{Day}/. You need to design a daily Azure Data Factory data load to minimize the data transfer between the two accounts. Which two configurations should you include in the design? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point

A. Specify a file naming pattern for the destination.

B. Delete the files in the destination before loading the data.

C. Filter by the last modified date of the source files.

D. Delete the source files after they are copied.

Correct Answer: A, C

Section:

Explanation:

Copy only the daily files by using filtering.

Reference: <https://docs.microsoft.com/en-us/azure/data-factory/connector-azure-data-lake-storage>

QUESTION 42

You plan to build a structured streaming solution in Azure Databricks. The solution will count new events in five-minute intervals and report only events that arrive during the interval. The output will be sent to a Delta Lake table.

Which output mode should you use?

- A. update
- B. complete
- C. append

Correct Answer: C

Section:

Explanation:

Append Mode: Only new rows appended in the result table since the last trigger are written to external storage. This is applicable only for the queries where existing rows in the Result Table are not expected to change.

Incorrect Answers:

B: Complete Mode: The entire updated result table is written to external storage. It is up to the storage connector to decide how to handle the writing of the entire table. A: Update Mode: Only the rows that were updated in the result table since the last trigger are written to external storage. This is different from Complete Mode in that Update Mode outputs only the rows that have changed since the last trigger. If the query doesn't contain aggregations, it is equivalent to Append mode.

Reference:

<https://docs.databricks.com/getting-started/spark/streaming.html>

QUESTION 43

Note: This question-is part of a series of questions that present the same scenario. Each question-in the series contains a unique solution that might meet the stated goals. Some question-sets might have more than one correct solution, while others might not have a correct solution.

After you answer a question-in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen. You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Table1. You have files that are ingested and loaded into an Azure Data Lake Storage Gen2 container named container1. You plan to insert data from the files in container1 into Table1 and transform the data. Each row of data in the files will produce one row in the serving layer of Table1. You need to ensure that when the source data files are loaded to container1, the DateTime is stored as an additional column in Table1. Solution: In an Azure Synapse Analytics pipeline, you use a data flow that contains a Derived Column transformation. Does this meet the goal?

- A. Yes
- B. No

Correct Answer: A

Section:

Explanation:

Use the derived column transformation to generate new columns in your data flow or to modify existing fields.

Reference:

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

QUESTION 44

Note: This question-is part of a series of questions that present the same scenario. Each question-in the series contains a unique solution that might meet the stated goals. Some question-sets might have more than one correct solution, while others might not have a correct solution.

After you answer a question-in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen. You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Table1. You have files that are ingested and loaded into an Azure Data Lake Storage Gen2 container named container1. You plan to insert data from the files in container1 into Table1 and transform the data. Each row of data in the files will produce one row in the serving layer of Table1. You need to ensure that when the source data files are loaded to container1, the DateTime is stored as an additional column in Table1. Solution: You use a dedicated SQL pool to create an external table that has an additional DateTime column. Does this meet the goal?

- A. Yes
- B. No

Correct Answer: B

Section:

Explanation:

Instead use the derived column transformation to generate new columns in your data flow or to modify existing fields.

Reference:

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

QUESTION 45

Note: This question-is part of a series of questions that present the same scenario. Each question-in the series contains a unique solution that might meet the stated goals. Some question-sets might have more than one correct solution, while others might not have a correct solution.

After you answer a question-in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen. You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Table1. You have files that are ingested and loaded into an Azure Data Lake Storage Gen2 container named container1. You plan to insert data from the files in container1 into Table1 and transform the data. Each row of data in the files will produce one row in the serving layer of Table1. You need to ensure that when the source data files are loaded to container1, the DateTime is stored as an additional column in Table1. Solution: You use an Azure Synapse Analytics serverless SQL pool to create an external table that has an additional DateTime column. Does this meet the goal?

A. Yes

B. No

Correct Answer: B

Section:

Explanation:

Instead use the derived column transformation to generate new columns in your data flow or to modify existing fields.

Reference:

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

QUESTION 46

Note: This question-is part of a series of questions that present the same scenario. Each question-in the series contains a unique solution that might meet the stated goals. Some question-sets might have more than one correct solution, while others might not have a correct solution.

After you answer a question-in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen. You plan to create an Azure Databricks workspace that has a tiered structure. The workspace will contain the following three workloads:

A workload for data engineers who will use Python and SQL.

A workload for jobs that will run notebooks that use Python, Scala, and SQL. A workload that data scientists will use to perform ad hoc analysis in Scala and R. The enterprise architecture team at your company identifies the following standards for Databricks environments:

The data engineers must share a duster.

The job duster will be managed by using a request process whereby data scientists and data engineers provide packaged notebooks for deployment to the cluster. All the data scientists must be assigned their own cluster that terminates automatically after 120 minutes of inactivity. Currently, there are three data scientists. You need to create the Databricks clusters for the workloads. Solution: You create a Standard cluster for each data scientist, a Standard cluster for the data engineers, and a High Concurrency cluster for the jobs. Does this meet the goal?

A. Yes

B. No

Correct Answer: B

Section:

Explanation:

We need a High Concurrency cluster for the data engineers and the jobs. Note: Standard clusters are recommended for a single user. Standard can run workloads developed in any language: Python, R, Scala, and SQL. A high concurrency cluster is a managed cloud resource. The key benefits of high concurrency clusters are that they provide Apache Spark-native fine-grained sharing for maximum resource utilization and minimum query latencies.

Reference:

<https://docs.azuredatabricks.net/clusters/configure.html>

QUESTION 47

DRAG DROP

You have an Azure Synapse Analytics workspace named WS1.

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

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



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

How should you complete the Transact-SQL statement? To answer, drag the appropriate values to the correct targets. Each value may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point.

Select and Place:

Values

Answer Area

- opendatasource
- openjson
- openquery
- openrowset

```
select*  
  
FROM  
    (   
        BULK 'https://contoso.blob.core.windows.net/contosodw' ,  
        FORMAT= 'CSV' ,  
        fieldterminator = '0x0b' ,  
        fieldquote = '0x0b' ,  
        rowterminator = '0x0b'   
    )   
with (id varchar(50),  
    contextdateeventTime varchar(50) '$.context.data.eventTime' ,  
    contextdatasamplingRate varchar(50) '$.context.data.samplingRate' ,  
    contextdataisSynthetic varchar(50) '$.context.data.isSynthetic' .  
    contextsessionisFirst varchar(50) '$.context.session.isFirst' ,  
    contextsession varchar(50) '$.context.session.id' ,  
    contextcustomdimensions varchar(max) '$.context.custom.dimensions'   
  
    ) as q  
cross apply (contextcustomdimensions)  
  
with ( ProfileType varchar(50) '$.customerInfo.ProfileType' ,  
    RoomName varchar(50) '$.customerInfo.RoomName' ,  
    CustomerName varchar(50) '$.customerInfo.CustomerName' ,  
    UserName varchar(50) '$.customerInfo.UserName'   
    )
```

Correct Answer:

Values**Answer Area**

```

select*

FROM
   (
    BULK 'https://contoso.blob.core.windows.net/contosodw',
    FORMAT= 'CSV',
    fieldterminator = '0x0b',
    fieldquote = '0x0b',
    rowterminator = '0x0b'
  )
  with (id varchar(50),
    contextdateeventTime varchar(50) '$.context.data.eventTime',
    contextdatasamplingRate varchar(50) '$.context.data.samplingRate',
    contextdataisSynthetic varchar(50) '$.context.data.isSynthetic',
    contextsessionisFirst varchar(50) '$.context.session.isFirst',
    contextsession varchar(50) '$.context.session.id',
    contextcustomdimensions varchar(max) '$.context.custom.dimensions'
  ) as q
  cross apply  (contextcustomdimensions)
  with ( ProfileType varchar(50) '$.customerInfo.ProfileType',
    RoomName varchar(50) '$.customerInfo.RoomName',
    CustomerName varchar(50) '$.customerInfo.CustomerName',
    UserName varchar(50) '$.customerInfo.UserName'
  )

```

Section:**Explanation:**

Box 1: openrowset

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

Example:

```

SELECT *
FROM OPENROWSET(
  BULK 'csv/population/population.csv',
  DATA_SOURCE = 'SqlOnDemandDemo',
  FORMAT = 'CSV', PARSER_VERSION = '2.0',
  FIELDTERMINATOR = ',',
  ROWTERMINATOR = '\n'

```

Box 2: openjson

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

```

SELECT book.* FROM
OPENROWSET(BULK N't:\books\books.json', SINGLE_CLOB) AS json CROSS APPLY OPENJSON(BulkColumn)
WITH( id nvarchar(100), name nvarchar(100), price float,
pages_i int, author nvarchar(100)) AS book

```

Reference:

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

QUESTION 48

DRAG DROP

You have an Apache Spark DataFrame named temperatures. A sample of the data is shown in the following table.

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

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

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

How should you complete the query? To answer, drag the appropriate values to the correct targets. Each value may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point.



Select and Place:

Values **Answer Area**

- CAST
- COLLATE
- CONVERT
- FLATTEN
- PIVOT
- UNPIVOT

```
SELECT * FROM (  
  SELECT YEAR(Date) Year, MONTH(Date) Month, Temp  
  FROM temperatures  
  WHERE date BETWEEN DATE '2019-01-01' AND DATE '2021-08-31'  
)  
 [ ] (  
  AVG ( [ ] (Temp AS DECIMAL(4, 1)))  
  FOR Month in (  
    1 JAN, 2 FEB, 3 MAR, 4 APR, 5 MAY, 6 JUN,  
    7 JUL, 8 AUG, 9 SEP, 10 OCT, 11 NOV, 12 DEC  
  )  
)  
ORDER BY Year ASC
```

Correct Answer:

Values Answer Area

COLLATE

CONVERT

FLATTEN

UNPIVOT

```
SELECT * FROM (
  SELECT YEAR(Date) Year, MONTH(Date) Month, Temp
  FROM temperatures
  WHERE date BETWEEN DATE '2019-01-01' AND DATE '2021-08-31'
)
PIVOT (
  AVG ( CAST (Temp AS DECIMAL(4, 1)))
  FOR Month in (
    1 JAN, 2 FEB, 3 MAR, 4 APR, 5 MAY, 6 JUN,
    7 JUL, 8 AUG, 9 SEP, 10 OCT, 11 NOV, 12 DEC
  )
)
ORDER BY Year ASC
```

Section:

Explanation:

Box 1: PIVOT

PIVOT rotates a table-valued expression by turning the unique values from one column in the expression into multiple columns in the output. And PIVOT runs aggregations where they're required on any remaining column values that are wanted in the final output.

Incorrect Answers:

UNPIVOT carries out the opposite operation to PIVOT by rotating columns of a table-valued expression into column values.

Box 2: CAST

If you want to convert an integer value to a DECIMAL data type in SQL Server use the CAST() function.

Example:

```
SELECT
CAST(12 AS DECIMAL(7,2) ) AS decimal_value;
```

Here is the result:

decimal_value

12.00

Reference:

<https://learnsql.com/cookbook/how-to-convert-an-integer-to-a-decimal-in-sql-server/>

<https://docs.microsoft.com/en-us/sql/t-sql/queries/from-using-pivot-and-unpivot>

QUESTION 49

HOTSPOT

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

Vdumps

```

{
  "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.

Hot Area:

 Vdumps

Answer Area		
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"/>

Answer Area:

Answer Area		
Statements	Yes	No
The Databricks cluster supports multiple concurrent users.	<input checked="" type="radio"/>	<input type="radio"/>
The Databricks cluster minimizes costs when running scheduled jobs that execute notebooks.	<input type="radio"/>	<input checked="" type="radio"/>
The Databricks cluster supports the creation of a Delta Lake table.	<input checked="" type="radio"/>	<input type="radio"/>

Section:

Explanation:

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>



QUESTION 50

HOTSPOT

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.

Hot Area:

Answer Area

```
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		

 Vdumps

Answer Area:

Answer Area

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

▼ FOR VALUES

RIGHT
LEFT

(▼)

20090101,20121231
20100101,20110101,20120101
20090101,20100101,20110101,20120101

dumps

Section:

Explanation:

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-ver15>

QUESTION 51

HOTSPOT

You have an Azure SQL database named Database1 and two Azure event hubs named HubA and HubB. The data consumed from each source is shown in the following table.

Source	Data
Database1	Driver's name Driver's license number
HubA	Ride route Ride distance Ride duration
HubB	Ride fare Ride payment

You need to implement Azure Stream Analytics to calculate the average fare per mile by driver.
 How should you configure the Stream Analytics input for each source? To answer, select the appropriate options in the answer area.
 NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

HubA: ▼

- Stream
- Reference

HubB: ▼

- Stream
- Reference

Database1: ▼

- Stream
- Reference



Answer Area:

Answer Area

HubA:	<input type="text"/>	▼
	Stream	
	Reference	
HubB:	<input type="text"/>	▼
	Stream	
	Reference	
Database1:	<input type="text"/>	▼
	Stream	
	Reference	

Section:

Explanation:

HubA: Stream

HubB: Stream

Database1: Reference

Reference data (also known as a lookup table) is a finite data set that is static or slowly changing in nature, used to perform a lookup or to augment your data streams. For example, in an IoT scenario, you could store metadata about sensors (which don't change often) in reference data and join it with real time IoT data streams. Azure Stream Analytics loads reference data in memory to achieve low latency stream processing

Reference:












<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-use-reference-data>

QUESTION 52

HOTSPOT

You configure version control for an Azure Data Factory instance as shown in the following exhibit.



 **Connections**
 Linked services
 Integration runtimes
 **Source control**
 **Git configuration**
 ARM template
 Parameterization template
Author
 Triggers
 Global parameters
Security
 Customer managed key
 Managed private endpoints

Git repository

Git repository information associated with your data factory. [CI/CD best practices](#)

 Setting
  Disconnect

Repository type	Azure DevOps Git
Azure DevOps Account	CONTOSO
Project name	Data
Repository name	dwh_batchetl
Collaboration branch	main
Publish branch	adf_publish
Root folder	/



Use the drop-down menus to select the answer choice that completes each statement based on the information presented in the graphic.
 NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Azure Resource Manager (ARM) templates for the pipeline assets are stored in [answer choice]

	▼
/	
adf_publish	
main	
Parameterization template	

A Data Factory Azure Resource Manager (ARM) template named contososales can be found in [answer choice]

	▼
/	
/contososales	
/dwh_batchetl/adf_publish/contososales	
/main	

Answer Area:

Answer Area

Azure Resource Manager (ARM) templates for the pipeline assets are stored in [answer choice]

	▼
/	
adf_publish	
main	
Parameterization template	

A Data Factory Azure Resource Manager (ARM) template named contososales can be found in [answer choice]

	▼
/	
/contososales	
/dwh_batchetl/adf_publish/contososales	
/main	

Section:

Explanation:

Box 1: adf_publish

The Publish branch is the branch in your repository where publishing related ARM templates are stored and updated. By default, it's adf_publish.

Box 2: / dwh_batchetl/adf_publish/contososales

Note: RepositoryName (here dwh_batchetl): Your Azure Repos code repository name. Azure Repos projects contain Git repositories to manage your source code as your project grows. You can create a new repository or use an existing repository that's already in your project.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/source-control>

QUESTION 53

HOTSPOT

You are designing an Azure Stream Analytics solution that receives instant messaging data from an Azure Event Hub. You need to ensure that the output from the Stream Analytics job counts the number of messages per time zone every 15 seconds. How should you complete the Stream Analytics query? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Select TimeZone, count (*) AS MessageCount

FROM MessageStream

	▼	CreatedAt
LAST		
OVER		
SYSTEM.TIMESTAMP()		
TIMESTAMP BY		

GROUP BY TimeZone,

	▼	(second, 15)
HOPPINGWINDOW		
SESSIONWINDOW		
SLIDINGWINDOW		
TUMBLINGWINDOW		

Answer Area:

Answer Area

Select TimeZone, count (*) AS MessageCount

FROM MessageStream

	▼
LAST	
OVER	
SYSTEM.TIMESTAMP()	
TIMESTAMP BY	

CreatedAt

GROUP BY TimeZone,

	▼
HOPPINGWINDOW	
SESSIONWINDOW	
SLIDINGWINDOW	
TUMBLINGWINDOW	

(second, 15)

Section:

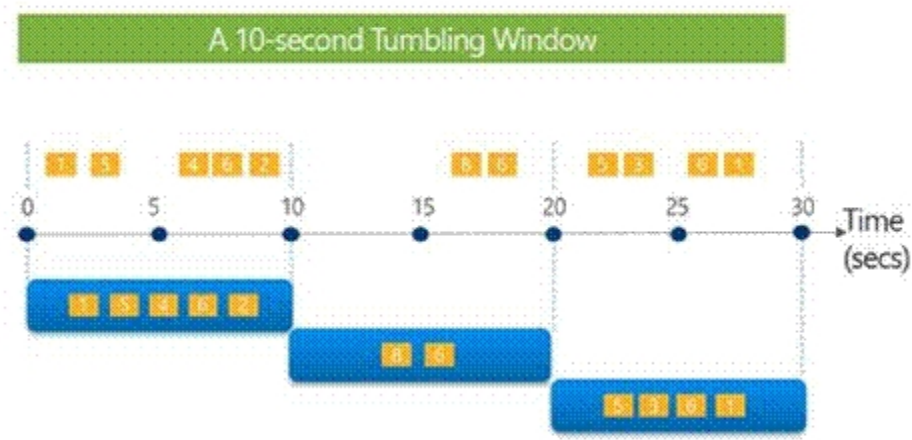
Explanation:

Box 1: timestamp by

Box 2: TUMBLINGWINDOW

Tumbling window functions are used to segment a data stream into distinct time segments and perform a function against them, such as the example below. The key differentiators of a Tumbling window are that they repeat, do not overlap, and an event cannot belong to more than one tumbling window.

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/azure/stream-analytics/stream-analytics-window-functions>

QUESTION 54

HOTSPOT

You are building an Azure Stream Analytics job to identify how much time a user spends interacting with a feature on a webpage.

The job receives events based on user actions on the webpage. Each row of data represents an event. Each event has a type of either 'start' or 'end'.

You need to calculate the duration between start and end events. How should you complete the query? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

```
SELECT
  [user],
  feature,
  

|            |
|------------|
| ▼          |
| DATEADD (  |
| DATEDIFF ( |
| DATEPART ( |


  second,
  

|         |
|---------|
| ▼       |
| ISFIRST |
| LAST    |
| TOPONE  |


  (Time) OVER (PARTITION BY [user], feature LIMIT DURATION(hour, 1) WHEN Event = 'start'),
  Time) as duration
FROM input TIMESTAMP BY Time
WHERE
  Event = 'end'
```

Answer Area:



Answer Area

```
SELECT
  [user],
  feature,
  

|            |
|------------|
| ▼          |
| DATEADD (  |
| DATEDIFF ( |
| DATEPART ( |


  second,
  

|         |
|---------|
| ▼       |
| ISFIRST |
| LAST    |
| TOPONE  |


  (Time) OVER (PARTITION BY [user], feature LIMIT DURATION(hour, 1) WHEN Event = 'start'),
  Time) as duration
FROM input TIMESTAMP BY Time
WHERE
  Event = 'end'
```

Section:

Explanation:

Box 1: DATEDIFF

DATEDIFF function returns the count (as a signed integer value) of the specified datepart boundaries crossed between the specified startdate and enddate. Syntax: DATEDIFF (datepart , startdate, enddate)

Box 2: LAST

The LAST function can be used to retrieve the last event within a specific condition. In this example, the condition is an event of type Start, partitioning the search by PARTITION BY user and feature. This way, every user and

feature is treated independently when searching for the Start event. LIMIT DURATION limits the search back in time to 1 hour between the End and Start events.

Example:

```
SELECT  
[user],  
feature,  
DATEDIFF(  
second,  
LAST(Time) OVER (PARTITION BY [user], feature LIMIT DURATION(hour, 1) WHEN Event = 'start'),  
Time) as duration  
FROM input TIMESTAMP BY Time
```

WHERE
Event = 'end'

Reference:

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

QUESTION 55

DRAG DROP

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.

Select and Place:

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@(  
_____ )
```



Correct Answer:

Values

- all, ecommerce, retail, wholesale
- _____
- dept=='ecommerce', dept=='wholesale', dept=='retail'
- _____
- disjoint: true
- _____

Answer Area

```
CleanData  
split(  
dept=='ecommerce', dept=='retail',  
dept=='wholesale'  
disjoint: false  
) ~> SplitByDept@(  
ecommerce, retail, wholesale, all )
```

Section:

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>

QUESTION 56

DRAG DROP

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

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

You create the following components:

A destination table in Azure Synapse

An Azure Blob storage container

A service principal

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



Select and Place:

Actions

Answer Area

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

Correct Answer:

Actions

Write the results to Data Lake Storage.
Drop the data frame.
Perform transformations on the data frame.

Answer Area

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

Section:**Explanation:**

<https://docs.databricks.com/data/data-sources/azure/azure-datalake-gen2.html><https://docs.microsoft.com/en-us/azure/databricks/scenarios/databricks-extract-load-sql-data-warehouse>

QUESTION 57**HOTSPOT**

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.

Hot Area:

Answer Area

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

Answer Area:

Answer Area

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

Section:

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>

QUESTION 58

HOTSPOT

You are designing a real-time dashboard solution that will visualize streaming data from remote sensors that connect to the internet. The streaming data must be aggregated to show the average value of each 10-second interval. The data will be discarded after being displayed in the dashboard.

The solution will use Azure Stream Analytics and must meet the following requirements:

Minimize latency from an Azure Event hub to the dashboard.

Minimize the required storage.

Minimize development effort.

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

NOTE: Each correct selection is worth one point

Hot Area:

Answer Area

Azure Stream Analytics input type:

Azure Stream Analytics output type:

Aggregation query location:

Answer Area:

Answer Area

Azure Stream Analytics input type:

	▼
Azure Event Hub	
Azure SQL Database	
Azure Stream Analytics	
Microsoft Power BI	

Azure Stream Analytics output type:

	▼
Azure Event Hub	
Azure SQL Database	
Azure Stream Analytics	
Microsoft Power BI	

Aggregation query location:

	▼
Azure Event Hub	
Azure SQL Database	
Azure Stream Analytics	
Microsoft Power BI	

Section:

Explanation:

Reference:

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

QUESTION 59

DRAG DROP

You have an Azure Stream Analytics job that is a Stream Analytics project solution in Microsoft Visual Studio. The job accepts data generated by IoT devices in the JSON format.

You need to modify the job to accept data generated by the IoT devices in the Protobuf format.

Which three actions should you perform from Visual Studio on sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

Select and Place:

Actions

Change the Event Serialization Format to Protobuf in the input.json file of the job and reference the DLL.

Add an Azure Stream Analytics Custom Deserializer Project (.NET) project to the solution.

Add .NET deserializer code for Protobuf to the custom deserializer project.

Add .NET deserializer code for Protobuf to the Stream Analytics project.

Add an Azure Stream Analytics Application project to the solution.

Answer Area**Correct Answer:****Actions**

Change the Event Serialization Format to Protobuf in the input.json file of the job and reference the DLL.

Add .NET deserializer code for Protobuf to the Stream Analytics project.

Answer Area

Add an Azure Stream Analytics Custom Deserializer Project (.NET) project to the solution.

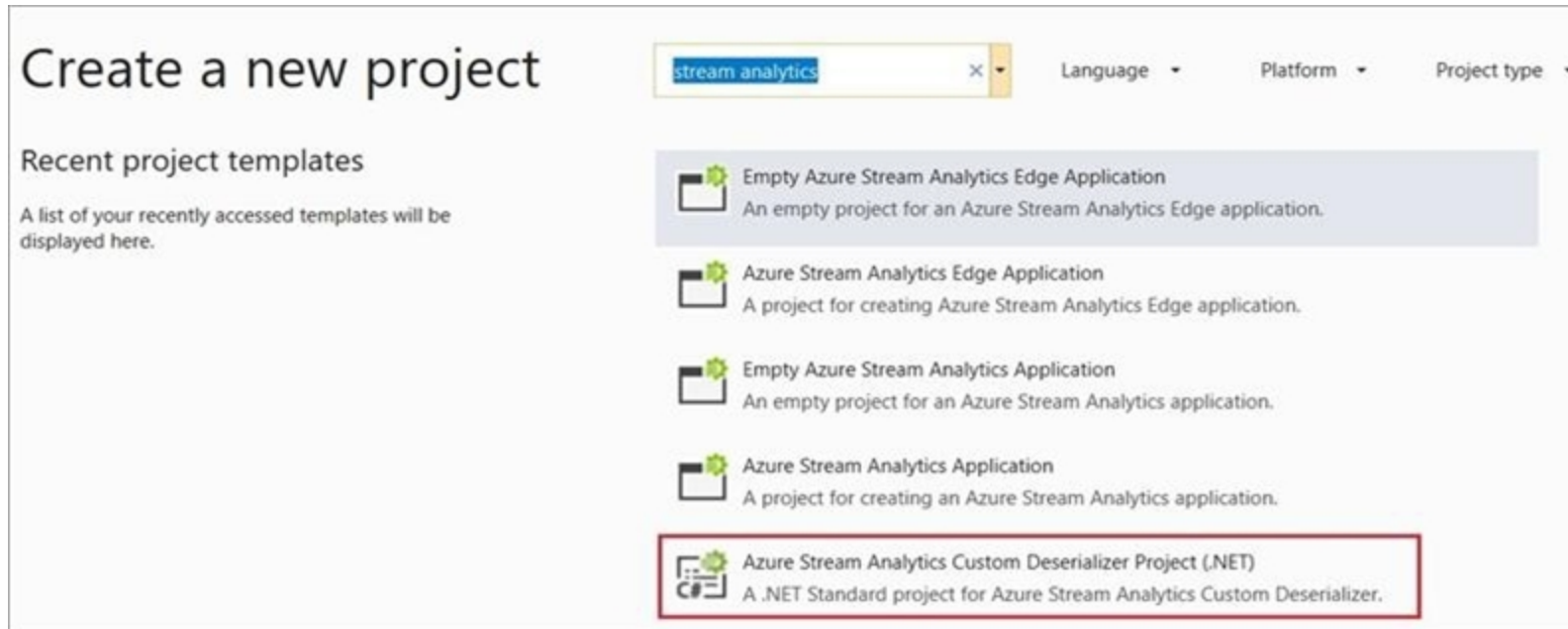
Add .NET deserializer code for Protobuf to the custom deserializer project.

Add an Azure Stream Analytics Application project to the solution.

Section:**Explanation:**

Step 1: Add an Azure Stream Analytics Custom Deserializer Project (.NET) project to the solution. Create a custom deserializer

1. Open Visual Studio and select File > New > Project. Search for Stream Analytics and select Azure Stream Analytics Custom Deserializer Project (.NET). Give the project a name, like Protobuf Deserializer.



2. In Solution Explorer, right-click your Protobuf Deserializer project and select Manage NuGet Packages from the menu. Then install the Microsoft.Azure.StreamAnalytics and Google.Protobuf NuGet packages.

3. Add the MessageBodyProto class and the MessageBodyDeserializer class to your project.

4. Build the Protobuf Deserializer project.

Step 2: Add .NET deserializer code for Protobuf to the custom deserializer project Azure Stream Analytics has built-in support for three data formats: JSON, CSV, and Avro. With custom .NET deserializers, you can read data from other formats such as Protocol Buffer, Bond and other user defined formats for both cloud and edge jobs.

Step 3: Add an Azure Stream Analytics Application project to the solution Add an Azure Stream Analytics project

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

2. Right-click References under the ProtobufCloudDeserializer Azure Stream Analytics project. Under Projects, add Protobuf Deserializer. It should be automatically populated for you.

Reference:

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



QUESTION 60

HOTSPOT

You have an Azure Data Factory instance named ADF1 and two Azure Synapse Analytics workspaces named WS1 and WS2.

ADF1 contains the following pipelines:

P1: Uses a copy activity to copy data from a nonpartitioned table in a dedicated SQL pool of WS1 to an Azure Data Lake Storage Gen2 account P2: Uses a copy activity to copy data from text-delimited files in an Azure Data Lake Storage Gen2 account to a nonpartitioned table in a dedicated SQL pool of WS2

You need to configure P1 and P2 to maximize parallelism and performance.

Which dataset settings should you configure for the copy activity if each pipeline? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

P1:

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

P2:

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

Answer Area:

Answer Area

P1:

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

P2:

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

Section:

Explanation:

Box 1: Set the Copy method to PolyBase

While SQL pool supports many loading methods including non-Polybase options such as BCP and SQL BulkCopy API, the fastest and most scalable way to load data is through PolyBase. PolyBase is a technology that accesses external data stored in Azure Blob storage or Azure Data Lake Store via the T-SQL language.

Box 2: Set the Copy method to Bulk insert

Polybase not possible for text files. Have to use Bulk insert.

Reference:

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

QUESTION 61

HOTSPOT

You are building an Azure Stream Analytics job to retrieve game data.

You need to ensure that the job returns the highest scoring record for each five-minute time interval of each game.

How should you complete the Stream Analytics query? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

SELECT

	▼
Collect(Score)	
CollectTop(1) OVER(ORDER BY Score Desc)	
Game, MAX(Score)	
TopOne() OVER(PARTITION BY Game ORDER BY Score Desc)	

 as HighestScore

FROM input TIMESTAMP BY CreatedAt

GROUP BY

	▼
Game	
Hopping(minute,5)	
Tumbling(minute,5)	
Windows(TumblingWindow(minute,5),Hopping(minute,5))	

Answer Area:

Answer Area

SELECT

	▼
Collect(Score)	
CollectTop(1) OVER(ORDER BY Score Desc)	
Game, MAX(Score)	
TopOne() OVER(PARTITION BY Game ORDER BY Score Desc)	

as HighestScore

FROM input TIMESTAMP BY CreatedAt

GROUP BY

	▼
Game	
Hopping(minute,5)	
Tumbling(minute,5)	
Windows(TumblingWindow(minute,5),Hopping(minute,5))	

Section:

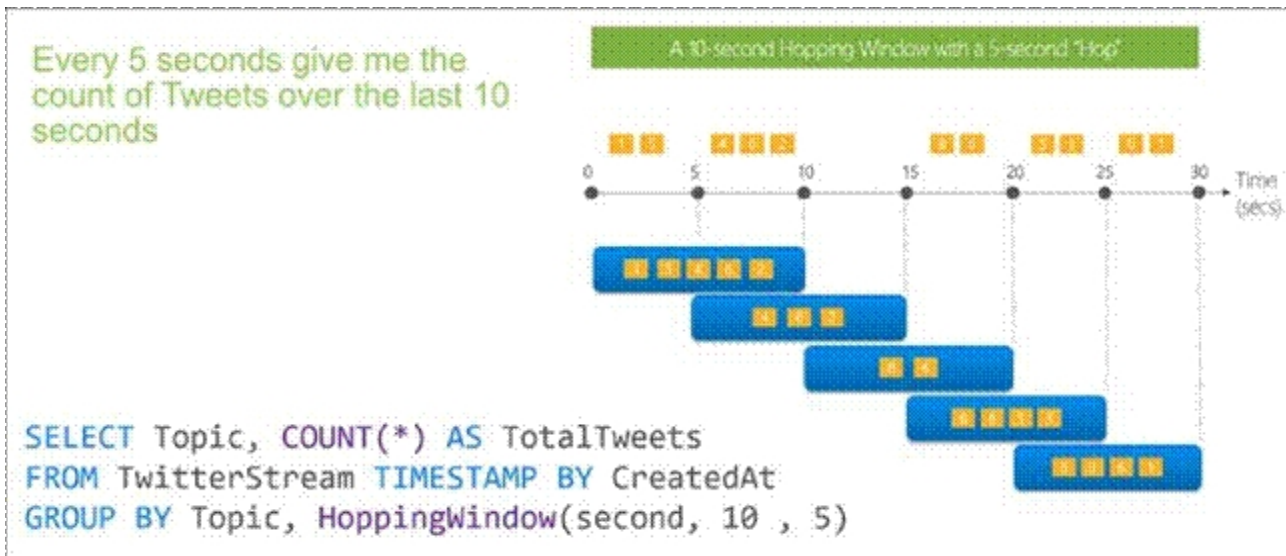
Explanation:

Box 1: TopOne OVER(PARTITION BY Game ORDER BY Score Desc)

TopOne returns the top-rank record, where rank defines the ranking position of the event in the window according to the specified ordering. Ordering/ranking is based on event columns and can be specified in ORDER BY clause.

Box 2: Hopping(minute,5)

Hopping window functions hop forward in time by a fixed period. It may be easy to think of them as Tumbling windows that can overlap and be emitted more often than the window size. Events can belong to more than one Hopping window result set. To make a Hopping window the same as a Tumbling window, specify the hop size to be the same as the window size.



Reference:

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

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

QUESTION 62

HOTSPOT

You are building an Azure Data Factory solution to process data received from Azure Event Hubs, and then ingested into an Azure Data Lake Storage Gen2 container.

The data will be ingested every five minutes from devices into JSON files. The files have the following naming pattern.

`/{deviceType}/in/{YYYY}/{MM}/{DD}/{HH}/{deviceId}_{YYYY}{MM}{DD}{HH}{mm}.json`

You need to prepare the data for batch data processing so that there is one dataset per hour per deviceType. The solution must minimize read times.

How should you configure the sink for the copy activity? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Vdumps

Answer Area

Parameter:

	▼
@pipeline(),TriggerTime	
@pipeline(),TriggerType	
@trigger().outputs.windowStartTime	
@trigger().startTime	

Naming pattern:

	▼
/{deviceId}/out/{YYYY}/{MM}/{DD}/{HH}.json	
/{YYYY}/{MM}/{DD}/{deviceType}.json	
/{YYYY}/{MM}/{DD}/{HH}.json	
/{YYYY}/{MM}/{DD}/{HH}_{deviceType}.json	

Copy behavior:

	▼
Add dynamic content	
Flatten hierarchy	
Merge files	

Answer Area:

Answer Area

Parameter:

	▼
@pipeline(),TriggerTime	
@pipeline(),TriggerType	
@trigger().outputs.windowStartTime	
@trigger().startTime	

Naming pattern:

	▼
/{deviceId}/out/{YYYY}/{MM}/{DD}/{HH}.json	
/{YYYY}/{MM}/{DD}/{deviceType}.json	
/{YYYY}/{MM}/{DD}/{HH}.json	
/{YYYY}/{MM}/{DD}/{HH}_{deviceType}.json	

Copy behavior:

	▼
Add dynamic content	
Flatten hierarchy	
Merge files	

Section:

Explanation:

Box 1: @trigger().startTime

startTime: A date-time value. For basic schedules, the value of the startTime property applies to the first occurrence. For complex schedules, the trigger starts no sooner than the specified startTime value.

Box 2: /{YYYY}/{MM}/{DD}/{HH}_{deviceType}.json

One dataset per hour per deviceType.

Box 3: Flatten hierarchy

- FlattenHierarchy: All files from the source folder are in the first level of the target folder. The target files have autogenerated names.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipeline-execution-triggers>

<https://docs.microsoft.com/en-us/azure/data-factory/connector-file-system>

QUESTION 63

DRAG DROP

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.

Select and Place:

Values	Answer Area
{deviceID}	/ [Value] / [Value] / [Value] .json
{mm}/{HH}/{DD}/{MM}/{YYYY}	
{regionID}/{deviceID}	
{regionID}/raw	
{YYYY}/{MM}/{DD}/{HH}	
{YYYY}/{MM}/{DD}/{HH}/{mm}	
raw/{deviceID}	
raw/{regionID}	

Correct Answer:

Values	Answer Area
	/ {YYYY}/{MM}/{DD}/{HH} / {regionID}/raw / {deviceID} .json
{mm}/{HH}/{DD}/{MM}/{YYYY}	
{regionID}/{deviceID}	
{YYYY}/{MM}/{DD}/{HH}/{mm}	
raw/{deviceID}	
raw/{regionID}	

Section:

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>

QUESTION 64

HOTSPOT

You are implementing an Azure Stream Analytics solution to process event data from devices.

The devices output events when there is a fault and emit a repeat of the event every five seconds until the fault is resolved. The devices output a heartbeat event every five seconds after a previous event if there are no faults present.

A sample of the events is shown in the following table.

DeviceID	EventType	EventTime
78cc5ht9-w357-684r-w4fr-kr16h6p9874e	HeartBeat	2020-12-01T19:00.000Z
78cc5ht9-w357-684r-w4fr-kr16h6p9874e	HeartBeat	2020-12-01T19:05.000Z
78cc5ht9-w357-684r-w4fr-kr16h6p9874e	TemperatureSensorFault	2020-12-01T19:07.000Z

You need to calculate the uptime between the faults.

How should you complete the Stream Analytics SQL query? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

```
SELECT
```

```
DeviceID,
```

```
MIN(EventTime) as StartTime,
```

```
MAX(EventTime) as EndTime,
```

```
DATEDIFF(second, MIN(EventTime), MAX(EventTime)) AS duration_in_seconds
```

```
FROM input TIMESTAMP BY EventTime
```

```
WHERE EventType='HeartBeat'
```

```
WHERE LAG(EventType, 1) OVER (LIMIT DURATION(second,5)) <> EventType
```

```
WHERE IsFirst(second,5) = 1
```

```
GROUP BY
```

```
DeviceID
```

```
,SessionWindow(second, 5, 50000) OVER (PARTITION BY DeviceID)
```

```
,TumblingWindow(second,5)
```

```
HAVING DATEDIFF(second, MIN(EventTime), MAX(EventTime)) > 5
```

Vdumps

Answer Area:

Answer Area

```
SELECT
DeviceID,
MIN(EventTime) as StartTime,
MAX(EventTime) as EndTime,
DATEDIFF(second, MIN(EventTime), MAX(EventTime)) AS duration_in_seconds
FROM input TIMESTAMP BY EventTime
```

```
WHERE EventType='HeartBeat'
WHERE LAG(EventType, 1) OVER (LIMIT DURATION(second,5)) <> EventType
WHERE IsFirst(second,5) = 1
```

GROUP BY

DeviceID

```
,SessionWindow(second, 5, 50000) OVER (PARTITION BY DeviceID)
,TumblingWindow(second,5)
HAVING DATEDIFF(second, MIN(EventTime), MAX(EventTime)) > 5
```



Section:

Explanation:

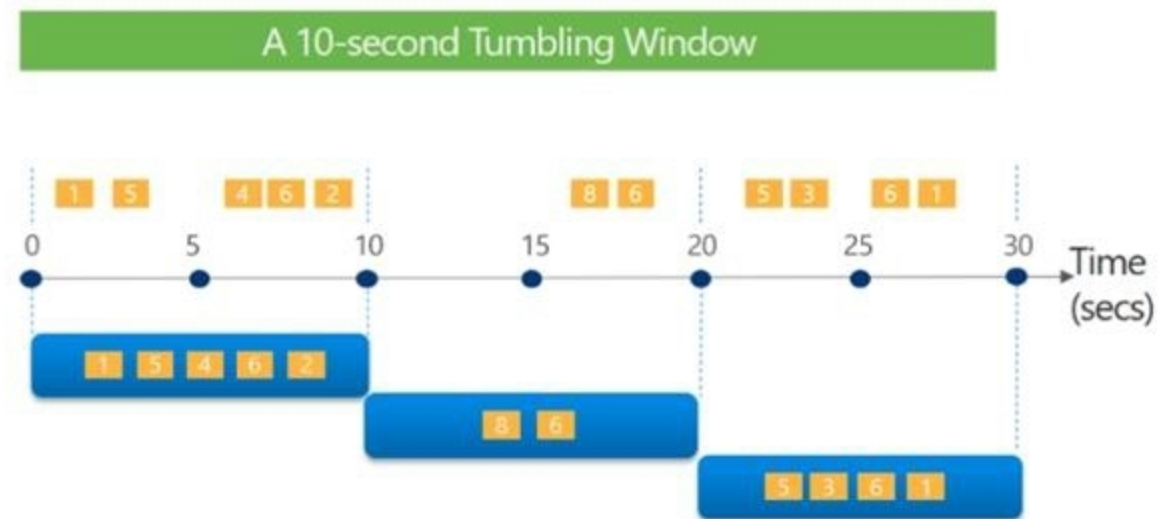
Box 1: WHERE EventType='HeartBeat'

Box 2: ,TumblingWindow(Second, 5)

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)
```

Incorrect Answers:

,SessionWindow.. : Session windows group events that arrive at similar times, filtering out periods of time where there is no data.

Reference:

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

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

Case 01 - Monitor and optimize data storage and data processing

This is a case study. Case studies are not timed separately. You can use as much exam time as you would like to complete each case. However, there may be additional case studies and sections on this exam. You must manage your time to ensure that you are able to complete all questions included on this exam in the time provided.

To answer the questions included in a case study, you will need to reference information that is provided in the case study. Case studies might contain exhibits and other resources that provide more information about the scenario that is described in the case study. Each question is independent of the other questions in this case study.

At the end of this case study, a review screen will appear. This screen allows you to review your answers and to make changes before you move to the next section of the exam. After you begin a new section, you cannot return to this section.

To start the case study

To display the first question in this case study, click the Next button. Use the buttons in the left pane to explore the content of the case study before you answer the questions. Clicking these buttons displays information such as business requirements, existing environment, and problem statements. If the case study has an All Information tab, note that the information displayed is identical to the information displayed on the subsequent tabs. When you are ready to answer a question, click the Question button to return to the question.

Overview

Litware, Inc. owns and operates 300 convenience stores across the US. The company sells a variety of packaged foods and drinks, as well as a variety of prepared foods, such as sandwiches and pizzas.

Litware has a loyalty club whereby members can get daily discounts on specific items by providing their membership number at checkout.

Litware employs business analysts who prefer to analyze data by using Microsoft Power BI, and data scientists who prefer analyzing data in Azure Databricks notebooks.

Requirements

Business Goals

Litware wants to create a new analytics environment in Azure to meet the following requirements:

See inventory levels across the stores. Data must be updated as close to real time as possible. Execute ad hoc analytical queries on historical data to identify whether the loyalty club discounts increase sales of the discounted products. Every four hours, notify store employees about how many prepared food items to produce based on historical demand from the sales data.

Technical Requirements

Litware identifies the following technical requirements:

Minimize the number of different Azure services needed to achieve the business goals. Use platform as a service (PaaS) offerings whenever possible and avoid having to provision virtual machines that must be managed by Litware. Ensure that the analytical data store is accessible only to the company's on-premises network and Azure services. Use Azure Active Directory (Azure AD) authentication whenever possible. Use the principle of least privilege when designing security. Stage Inventory data in Azure Data Lake Storage Gen2 before loading the data into the analytical data store. Litware wants to remove transient data from Data Lake Storage once the data is no longer in use. Files that have a modified date that is older than 14 days must be removed.

Limit the business analysts' access to customer contact information, such as phone numbers, because this type of data is not analytically relevant. Ensure that you can quickly restore a copy of the analytical data store within one hour in the event of corruption or accidental deletion.

Planned Environment

Litware plans to implement the following environment:

The application development team will create an Azure event hub to receive real-time sales data, including store number, date, time, product ID, customer loyalty number, price, and discount amount, from the point of sale (POS) system and output the data to data storage in Azure.

Customer data, including name, contact information, and loyalty number, comes from Salesforce, a SaaS application, 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. Daily inventory data comes from a Microsoft SQL server located on a private network. Litware currently has 5 TB of historical sales data and 100 GB of customer data. The company expects approximately 100 GB of new data per month for the next year. Litware will build a custom application named FoodPrep to provide store employees with the calculation results of how many prepared food items to produce every four hours. Litware does not plan to implement Azure ExpressRoute or a VPN between the on-premises network and Azure.

QUESTION 1

What should you do to improve high availability of the real-time data processing solution?

- A. Deploy a High Concurrency Databricks cluster.
- B. Deploy an Azure Stream Analytics job and use an Azure Automation runbook to check the status of the job and to start the job if it stops.
- C. Set Data Lake Storage to use geo-redundant storage (GRS).
- D. Deploy identical Azure Stream Analytics jobs to paired regions in Azure.



Correct Answer: D

Section:

Explanation:

Guarantee Stream Analytics job reliability during service updates

Part of being a fully managed service is the capability to introduce new service functionality and improvements at a rapid pace. As a result, Stream Analytics can have a service update deploy on a weekly (or more frequent) basis. No matter how much testing is done there is still a risk that an existing, running job may break due to the introduction of a bug. If you are running mission critical jobs, these risks need to be avoided. You can reduce this risk by following Azure's paired region model.

Scenario: The application development team will create an Azure event hub to receive real-time sales data, including store number, date, time, product ID, customer loyalty number, price, and discount amount, from the point of sale (POS) system and output the data to data storage in Azure

Reference:

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

02 - Monitor and optimize data storage and data processing

QUESTION 1

You have an Azure Synapse Analytics dedicated SQL pool named Pool1 and a database named DB1. DB1 contains a fact table named Table1. You need to identify the extent of the data skew in Table1.

What should you do in Synapse Studio?

- A. Connect to the built-in pool and run sys.dm_pdw_nodes_db_partition_stats.
- B. Connect to Pool1 and run DBCC CHECKALLOC.
- C. Connect to the built-in pool and run DBCC CHECKALLOC.
- D. Connect to Pool1 and query sys.dm_pdw_nodes_db_partition_stats.

Correct Answer: D

Section:

Explanation:

Microsoft recommends use of sys.dm_pdw_nodes_db_partition_stats to analyze any skewness in the data. Reference:

<https://docs.microsoft.com/en-us/sql/relational-databases/system-dynamic-management-views/sys-dm-db-partition-stats-transact-sql> <https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/cheat-sheet>

QUESTION 2

You have an Azure Synapse Analytics dedicated SQL pool that contains a large fact table. The table contains 50 columns and 5 billion rows and is a heap. Most queries against the table aggregate values from approximately 100 million rows and return only two columns. You discover that the queries against the fact table are very slow. Which type of index should you add to provide the fastest query times?

- A. nonclustered columnstore
- B. clustered columnstore
- C. nonclustered
- D. clustered

Correct Answer: B

Section:

Explanation:

Clustered columnstore indexes are one of the most efficient ways you can store your data in dedicated SQL pool. Columnstore tables won't benefit a query unless the table has more than 60 million rows.

Reference: <https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/best-practices-dedicated-sql-pool>

QUESTION 3

You create an Azure Databricks cluster and specify an additional library to install. When you attempt to load the library to a notebook, the library is not found. You need to identify the cause of the issue. What should you review?

- A. notebook logs
- B. cluster event logs
- C. global init scripts logs
- D. workspace logs

Correct Answer: C

Section:

Explanation:

Cluster-scoped Init Scripts: Init scripts are shell scripts that run during the startup of each cluster node before the Spark driver or worker JVM starts. Databricks customers use init scripts for various purposes such as installing custom libraries, launching background processes, or applying enterprise security policies. Logs for Cluster-scoped init scripts are now more consistent with Cluster Log Delivery and can be found in the same root folder as driver and executor logs for the cluster.

Reference: <https://databricks.com/blog/2018/08/30/introducing-cluster-scoped-init-scripts.html>

QUESTION 4

You have an Azure data factory.

You need to examine the pipeline failures from the last 60 days. What should you use?

- A. the Activity log blade for the Data Factory resource

- B. the Monitor & Manage app in Data Factory
- C. the Resource health blade for the Data Factory resource
- D. Azure Monitor

Correct Answer: D

Section:

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>

QUESTION 5

You are monitoring an Azure Stream Analytics job.

The Backlogged Input Events count has been 20 for the last hour.

You need to reduce the Backlogged Input Events count.

What should you do?

- A. Drop late arriving events from the job.
- B. Add an Azure Storage account to the job.
- C. Increase the streaming units for the job.
- D. Stop the job.

Correct Answer: C

Section:

Explanation:

General symptoms of the job hitting system resource limits include:

If the backlog event metric keeps increasing, it's an indicator that the system resource is constrained (either because of output sink throttling, or high CPU). 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 nonzero, you should scale out your job: adjust Streaming Units.

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

QUESTION 6

You are designing an Azure Databricks interactive cluster. The cluster will be used infrequently and will be configured for auto-termination. You need to use that the cluster configuration is retained indefinitely after the cluster is terminated. The solution must minimize costs. What should you do?

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

Correct Answer: A

Section:

Explanation:

Azure Databricks retains cluster configuration information for up to 70 all-purpose clusters terminated in the last 30 days and up to 30 job clusters recently terminated by the job scheduler. To keep an all-purpose cluster configuration even after it has been terminated for more than 30 days, an administrator can pin a cluster to the cluster list.

Reference:

<https://docs.microsoft.com/en-us/azure/databricks/clusters/>

QUESTION 7

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.

Correct Answer: C

Section:

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-management>

QUESTION 8

You have an Azure Synapse Analytics dedicated SQL pool named Pool1 and a database named DB1. DB1 contains a fact table named Table1. You need to identify the extent of the data skew in Table1. What should you do in Synapse Studio?

- A. Connect to the built-in pool and run DBCC PDW_SHOWSPACEUSED.
- B. Connect to the built-in pool and run DBCC CHECKALLOC.
- C. Connect to Pool1 and query sys.dm_pdw_node_status.
- D. Connect to Pool1 and query sys.dm_pdw_nodes_db_partition_stats.



Correct Answer: D

Section:

Explanation:

Microsoft recommends use of sys.dm_pdw_nodes_db_partition_stats to analyze any skewness in the data.

Reference:

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

QUESTION 9

You have a SQL pool in Azure Synapse. You discover that some queries fail or take a long time to complete. You need to monitor for transactions that have rolled back. Which dynamic management view should you query?

- A. sys.dm_pdw_request_steps
- B. sys.dm_pdw_nodes_tran_database_transactions
- C. sys.dm_pdw_waits
- D. sys.dm_pdw_exec_sessions

Correct Answer: B

Section:

Explanation:

You can use Dynamic Management Views (DMVs) to monitor your workload including investigating query execution in SQL pool. If your queries are failing or taking a long time to proceed, you can check and monitor if you

have any transactions rolling back. Example:

```
-- Monitor rollback
```

```
SELECT  
SUM(CASE WHEN t.database_transaction_next_undo_lsn IS NOT NULL THEN 1 ELSE 0 END), t.pdw_node_id, nod.[type] FROM sys.dm_pdw_nodes_tran_database_transactions t  
JOIN sys.dm_pdw_nodes nod ON t.pdw_node_id = nod.pdw_node_id GROUP BY t.pdw_node_id, nod.[type]
```

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-manage-monitor#monitor-transaction-log-rollback>

QUESTION 10

You are monitoring an Azure Stream Analytics job.

You discover that the Backlogged Input Events metric is increasing slowly and is consistently non-zero. You need to ensure that the job can handle all the events.

What should you do?

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

Correct Answer: B

Section:

Explanation:

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

Reference:

<https://docs.microsoft.com/bs-cyrl-ba/azure/stream-analytics/stream-analytics-monitoring>



QUESTION 11

You are designing an inventory updates table in an Azure Synapse Analytics dedicated SQL pool. The table will have a clustered columnstore index and will include the following columns:

Table	Comment
EventDate	One million records are added to the table each day
EventTypeID	The table contains 10 million records for each event type.
WarehouseID	The table contains 100 million records for each warehouse.
ProductCategoryTypeID	The table contains 25 million records for each product category type.

You identify the following usage patterns:

Analysts will most commonly analyze transactions for a warehouse. Queries will summarize by product category type, date, and/or inventory event type.

You need to recommend a partition strategy for the table to minimize query times.

On which column should you partition the table?

- A. EventTypeID
- B. ProductCategoryTypeID
- C. EventDate
- D. WarehouseID

Correct Answer: D

Section:

Explanation:

The number of records for each warehouse is big enough for a good partitioning. Note: Table partitions enable you to divide your data into smaller groups of data. In most cases, table partitions are created on a date column. 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 distributed databases.

QUESTION 12

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

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

Correct Answer: C, D

Section:

Explanation:

QUESTION 13

A company purchases IoT devices to monitor manufacturing machinery. The company uses an Azure IoT Hub to communicate with the IoT devices. The company must be able to monitor the devices in real-time. You need to design the solution.

What should you recommend?

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



Correct Answer: D

Section:

Explanation:

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

QUESTION 14

You have a SQL pool in Azure Synapse.

A user reports that queries against the pool take longer than expected to complete. You determine that the issue relates to queried columnstore segments. You need to add monitoring to the underlying storage to help diagnose the issue. Which two metrics should you monitor? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

- A. Snapshot Storage Size
- B. Cache used percentage
- C. DWU Limit
- D. Cache hit percentage

Correct Answer: B, D

Section:

Explanation:

D: Cache hit percentage: $(\text{cache hits} / \text{cache miss}) * 100$ where cache hits is the sum of all columnstore segments hits in the local SSD cache and cache miss is the columnstore segments misses in the local SSD cache summed across all nodes

B: $(\text{cache used} / \text{cache capacity}) * 100$ where cache used is the sum of all bytes in the local SSD cache across all nodes and cache capacity is the sum of the storage capacity of the local SSD cache across all nodes Incorrect Answers:

C: DWU limit: Service level objective of the data warehouse.

Reference: <https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-concept-resource-utilization-query-activity>

QUESTION 15

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. DWU percentage
- B. Cache hit percentage
- C. DWU limit
- D. Data IO percentage

Correct Answer: B

Section:

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-monitor-cache>

QUESTION 16

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. SSH
- E. jobs

Correct Answer: B

Section:

Explanation:

Databricks provides access to audit logs of activities performed by Databricks users, allowing your enterprise to monitor detailed Databricks usage patterns. There are two types of logs:

Workspace-level audit logs with workspace-level events. Account-level audit logs with account-level events.

Reference: <https://docs.databricks.com/administration-guide/account-settings/audit-logs.html>

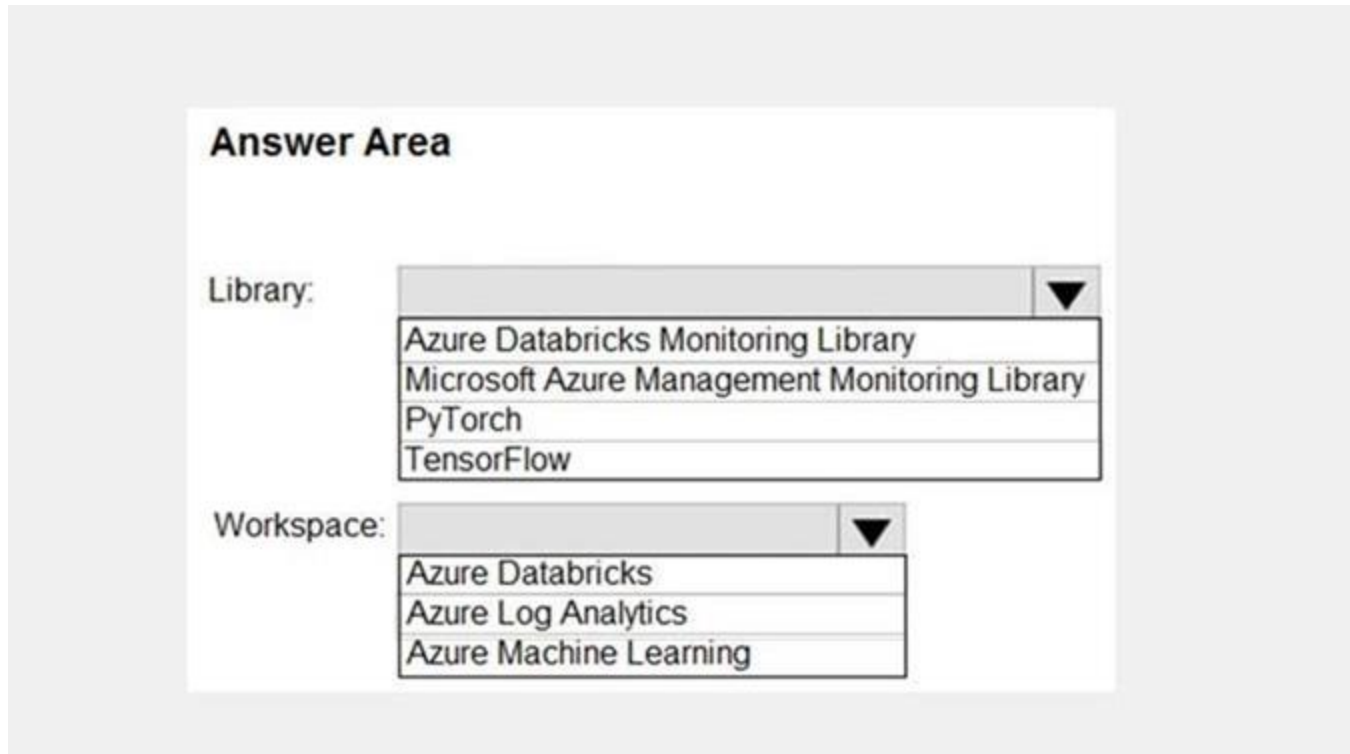
QUESTION 17

HOTSPOT

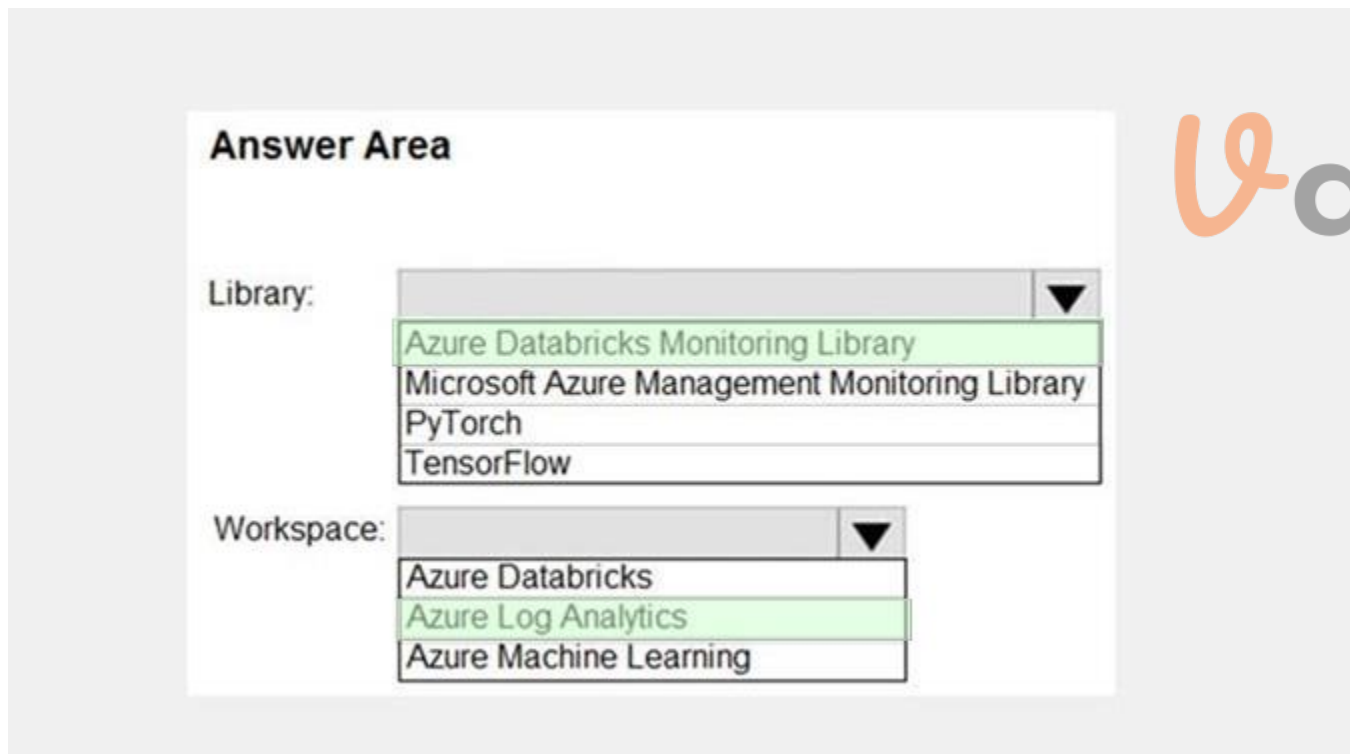
You need to collect application metrics, streaming query events, and application log messages for an Azure Databrick cluster. Which type of library and workspace should you implement? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:



Answer Area:



Section:

Explanation:

You can send application logs and metrics from Azure Databricks to a Log Analytics workspace. It uses the Azure Databricks Monitoring Library, which is available on GitHub.

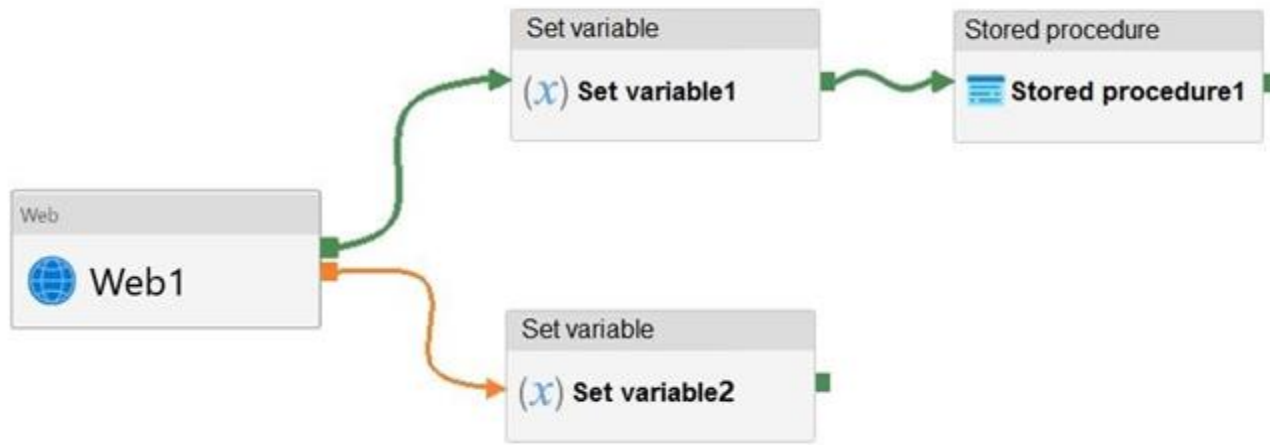
Reference:

<https://docs.microsoft.com/en-us/azure/architecture/databricks-monitoring/application-logs>

QUESTION 18

HOTSPOT

You have an Azure Data Factory pipeline that has the activities 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.
 NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

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

- complete
- fail
- succeed

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

- Canceled
- Failed
- Succeeded

Answer Area:

Answer Area

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

- complete
- fail
- succeed

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

- Canceled
- Failed
- Succeeded

Section:

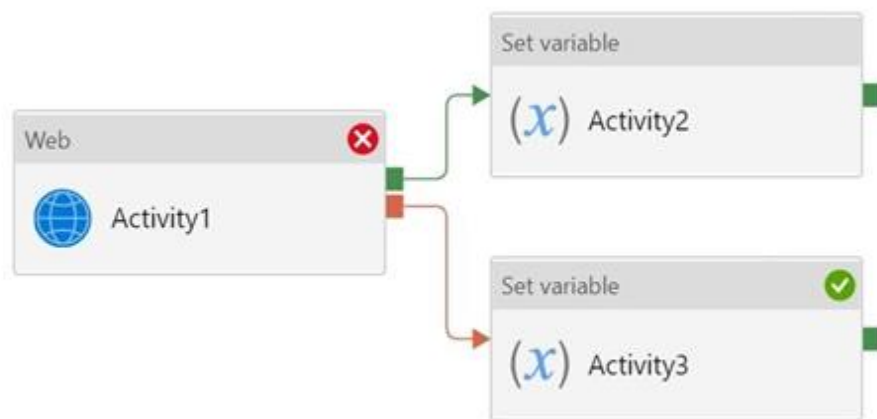
Explanation:

Box 1: succeed

Box 2: failed

Example:

Now let's say we have a pipeline with 3 activities, where Activity1 has a success path to Activity2 and a failure path to Activity3. If Activity1 fails and Activity3 succeeds, the pipeline will fail. The presence of the success path alongside the failure path changes the outcome reported by the pipeline, even though the activity executions from the pipeline are the same as the previous scenario.



Activity1 fails, Activity2 is skipped, and Activity3 succeeds. The pipeline reports failure.

Reference:

<https://datasavvy.me/2021/02/18/azure-data-factory-activity-failures-and-pipeline-outcomes/>

QUESTION 19

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

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

Correct Answer: D

Section:

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>

QUESTION 20

You configure monitoring from an Azure Synapse Analytics implementation. The implementation uses PolyBase to load data from comma-separated value (CSV) files stored in Azure Data Lake Storage Gen2 using an external table. Files with an invalid schema cause errors to occur.

You need to monitor for an invalid schema error.

For which error should you monitor?

- A. EXTERNAL TABLE access failed due to internal error: 'Java exception raised on call to HdfsBridge_Connect: Error [com.microsoft.polybase.client.KerberosSecureLogin] occurred while accessing external file.'
- B. Cannot execute the query "Remote Query" against OLE DB provider "SQLNCLI11" for linked server "(null)". Query aborted- the maximum reject threshold (0 rows) was reached while reading from an external source: 1 rows rejected outof total 1 rows processed.

- C. EXTERNAL TABLE access failed due to internal error: 'Java exception raised on call to HdfsBridge_Connect: Error [Unable to instantiate LoginClass] occurred while accessing external file.'
- D. EXTERNAL TABLE access failed due to internal error: 'Java exception raised on call to HdfsBridge_Connect: Error [No FileSystem for scheme: wasbs] occurred while accessing external file.'

Correct Answer: B

Section:

Explanation:

Error message: Cannot execute the query "Remote Query"

Possible Reason:

The reason this error happens is because each file has different schema. The PolyBase external table DDL when pointed to a directory recursively reads all the files in that directory. When a column or data type mismatch happens, this error could be seen in SSMS.

Reference:

<https://docs.microsoft.com/en-us/sql/relational-databases/polybase/polybase-errors-and-possible-solutions>

QUESTION 21

You have an Azure Synapse Analytics dedicated SQL pool.

You run PDW_SHOWSPACEUSED('dbo.FactInternetSales'); and get the results shown in the following table.

ROWS	RESERVED_SPACE	DATA_SPACE	INDEX_SPACE	UNUSED_SPACE	PDW_NODE_ID	DISTRIBUTION_ID
694	2776	616	48	2112	1	1
407	2704	576	48	2080	1	2
53	2376	512	16	1848	1	3
58	2376	512	16	1848	1	4
188	2632	528	32	2072	1	5
195	2696	536	32	2128	1	6
5995	3464	1424	32	2008	1	7
0	2232	496	0	1736	1	8
264	2576	544	40	1992	1	9
3008	3016	960	32	2024	1	10
--	--	--	--	--	--	--
1550	2832	752	48	2032	1	50
1238	2832	696	40	2096	1	51
192	2632	528	32	2072	1	52
1127	2768	680	48	2040	1	53
1244	3032	704	64	2264	1	54
409	2632	568	32	2032	1	55
0	2232	496	0	1736	1	56
1417	2832	728	40	2064	1	57
0	2232	496	0	1736	1	58
384	2632	568	32	2040	1	59
225	2768	544	40	2184	1	60



Which statement accurately describes the dbo.FactInternetSales table?

- A. All distributions contain data.
- B. The table contains less than 10,000 rows.
- C. The table uses round-robin distribution.
- D. The table is skewed.

Correct Answer: D

Section:

QUESTION 22

You have two fact tables named Flight and Weather. Queries targeting the tables will be based on the join between the following columns.

Table	Column
Flight	ArrivalAirportID ArrivalDateTime
Weather	AirportID ReportDateTime

You need to recommend a solution that maximizes query performance. What should you include in the recommendation?

- A. In the tables use a hash distribution of ArrivalDateTime and ReportDateTime.
- B. In the tables use a hash distribution of ArrivalAirportID and AirportID.
- C. In each table, create an IDENTITY column.
- D. In each table, create a column as a composite of the other two columns in the table.

Correct Answer: B

Section:

Explanation:

Hash-distribution improves query performance on large fact tables. Incorrect Answers:

A: Do not use a date column for hash distribution. All data for the same date lands in the same distribution. If several users are all filtering on the same date, then only 1 of the 60 distributions do all the processing work.

QUESTION 23

A company purchases IoT devices to monitor manufacturing machinery. The company uses an Azure IoT Hub to communicate with the IoT devices. The company must be able to monitor the devices in real-time. You need to design the solution.

What should you recommend?

- A. Azure Data Factory instance using Azure Portal
- B. Azure Data Factory instance using Azure PowerShell
- C. Azure Stream Analytics cloud job using Azure Portal
- D. Azure Data Factory instance using Microsoft Visual Studio



Correct Answer: A

Section:

Explanation:

QUESTION 24

HOTSPOT

You have an Azure event hub named retailhub that has 16 partitions. Transactions are posted to retailhub. Each transaction includes the transaction ID, the individual line items, and the payment details. The transaction ID is used as the partition key.

You are designing an Azure Stream Analytics job to identify potentially fraudulent transactions at a retail store. The job will use retailhub as the input. The job will output the transaction ID, the individual line items, the payment details, a fraud score, and a fraud indicator.

You plan to send the output to an Azure event hub named fraudhub.

You need to ensure that the fraud detection solution is highly scalable and processes transactions as quickly as possible.

How should you structure the output of the Stream Analytics job? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Number of partitions:

	▼
1	
8	
16	
32	

Partition key:

	▼
Fraud indicator	
Fraud score	
Individual line items	
Payment details	
Transaction ID	

Answer Area:



Answer Area

Number of partitions:

	▼
1	
8	
16	
32	

Partition key:

	▼
Fraud indicator	
Fraud score	
Individual line items	
Payment details	
Transaction ID	

Section:

Explanation:

Box 1: 16

For Event Hubs you need to set the partition key explicitly.

An embarrassingly parallel job is the most scalable scenario in Azure Stream Analytics. It connects one partition of the input to one instance of the query to one partition of the output.

Box 2: Transaction ID

Reference:

<https://docs.microsoft.com/en-us/azure/event-hubs/event-hubs-features#partitions>

QUESTION 25

You have a partitioned table in an Azure Synapse Analytics dedicated SQL pool. You need to design queries to maximize the benefits of partition elimination. What should you include in the Transact-SQL queries?

- A. JOIN
- B. WHERE
- C. DISTINCT
- D. GROUP BY

Correct Answer: B

Section:

QUESTION 26

You implement an enterprise data warehouse in Azure Synapse Analytics.

You have a large fact table that is 10 terabytes (TB) in size.

Incoming queries use the primary key SaleKey column to retrieve data as displayed in the following table:

SaleKey	CityKey	CustomerKey	StockItemKey	InvoiceDateKey	Quantity	UnitPrice	TotalExcludingTax
49309	90858	70	69	10/22/13	8	16	128
49313	55710	126	69	10/22/13	2	16	32
49343	44710	234	68	10/22/13	10	16	160
49352	66109	163	70	10/22/13	4	16	64
49488	65312	230	70	10/22/13	8	16	128
49646	85877	271	70	10/24/13	1	16	16
49798	41238	288	69	10/24/13	1	16	16

You need to distribute the large fact table across multiple nodes to optimize performance of the table. Which technology should you use?

- A. hash distributed table with clustered index
- B. hash distributed table with clustered Columnstore index
- C. round robin distributed table with clustered index
- D. round robin distributed table with clustered Columnstore index
- E. heap table with distribution replicate

Correct Answer: B

Section:

Explanation:

Hash-distributed tables improve query performance on large fact tables. Columnstore indexes can achieve up to 100x better performance on analytics and data warehousing workloads and up to 10x better data compression than traditional rowstore indexes. Incorrect Answers:

C, D: Round-robin tables are useful for improving loading speed.

Reference: <https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-tables-distribute> <https://docs.microsoft.com/en-us/sql/relational-databases/indexes/columnstore-indexes-query-performance>

QUESTION 27

You have several Azure Data Factory pipelines that contain a mix of the following types of activities:

Wrangling data flow

Notebook

Copy Jar

Which two Azure services should you use to debug the activities? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point

- A. Azure Synapse Analytics
- B. Azure HDInsight
- C. Azure Machine Learning
- D. Azure Data Factory
- E. Azure Databricks

Correct Answer: B, D

Section:

QUESTION 28

HOTSPOT

You have an on-premises data warehouse that includes the following fact tables. Both tables have the following columns: DateKey, ProductKey, RegionKey. There are 120 unique product keys and 65 unique region keys.

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

Queries that use the data warehouse take a long time to complete.

You plan to migrate the solution to use Azure Synapse Analytics. You need to ensure that the Azure-based solution optimizes query performance and minimizes processing skew.

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

NOTE: Each correct selection is worth one point

Hot Area:

Answer Area

Table	Distribution type	Distribution column
Sales:	<div style="border: 1px solid #ccc; padding: 2px;">▼</div> <div style="border: 1px solid #ccc; padding: 2px;">Hash-distributed</div> <div style="border: 1px solid #ccc; padding: 2px;">Round-robin</div>	<div style="border: 1px solid #ccc; padding: 2px;">▼</div> <div style="border: 1px solid #ccc; padding: 2px;">DateKey</div> <div style="border: 1px solid #ccc; padding: 2px;">ProductKey</div> <div style="border: 1px solid #ccc; padding: 2px;">RegionKey</div>
Invoices:	<div style="border: 1px solid #ccc; padding: 2px;">▼</div> <div style="border: 1px solid #ccc; padding: 2px;">Hash-distributed</div> <div style="border: 1px solid #ccc; padding: 2px;">Round-robin</div>	<div style="border: 1px solid #ccc; padding: 2px;">▼</div> <div style="border: 1px solid #ccc; padding: 2px;">DateKey</div> <div style="border: 1px solid #ccc; padding: 2px;">ProductKey</div> <div style="border: 1px solid #ccc; padding: 2px;">RegionKey</div>

Answer Area:

Answer Area

Table	Distribution type	Distribution column
-------	-------------------	---------------------

Sales:

Hash-distributed	DateKey
Round-robin	ProductKey
	RegionKey

Invoices:

Hash-distributed	DateKey
Round-robin	ProductKey
	RegionKey

Vdumps

Section:

Explanation:

Box 1: Hash-distributed

Box 2: ProductKey

ProductKey is used extensively in joins.

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

Box 3: Hash-distributed

Box 4: RegionKey

Round-robin tables are useful for improving loading speed.

Consider using the round-robin distribution for your table in the following scenarios:

When getting started as a simple starting point since it is the default if there is no obvious joining key

If there is not good candidate column for hash distributing the table

If the table does not share a common join key with other tables

If the join is less significant than other joins in the query

When the table is a temporary staging table

Note: A distributed table appears as a single table, but the rows are actually stored across 60 distributions. The rows are distributed with a hash or round-robin algorithm.

Reference:

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

Exam F

QUESTION 1

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)

Correct Answer: C

Section:

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/>

QUESTION 2

You have an Azure Synapse Analytics dedicated SQL pool named pool1. You plan to implement a star schema in pool1 and create a new table named DimCustomer by using the following code.

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



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

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

Correct Answer: A, B

Section:

QUESTION 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

Correct Answer: D

Section:

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>

QUESTION 4

HOTSPOT

You have an Azure subscription that contains an Azure Synapse Analytics dedicated SQL pool. You plan to deploy a solution that will analyze sales data and include the following:

- A table named Country that will contain 195 rows
 - A table named Sales that will contain 100 million rows
 - A query to identify total sales by country and customer from the past 30 days
- You need to create the tables. The solution must maximize query performance. How should you complete the script? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

```
CREATE TABLE [dbo].[Sales]
(
    [OrderDate] date NOT NULL
,   [CustomerId] int NOT NULL
,   [CountryId] int NOT NULL
,   [Total] money NOT NULL
)
WITH
(
    DISTRIBUTION = HASH([CustomerId])
,   CLUSTERED COLUMNSTORE INDEX
)
CREATE TABLE [dbo].[Country]
```

Answer Area:

Answer Area

```
CREATE TABLE [dbo].[Sales]
(
    [OrderDate] date NOT NULL
,   [CustomerId] int NOT NULL
,   [CountryId] int NOT NULL
,   [Total] money NOT NULL
)
WITH
(
    DISTRIBUTION = HASH([CustomerId])
,   CLUSTERED COLUMNSTORE INDEX
)
CREATE TABLE [dbo].[Country]
```



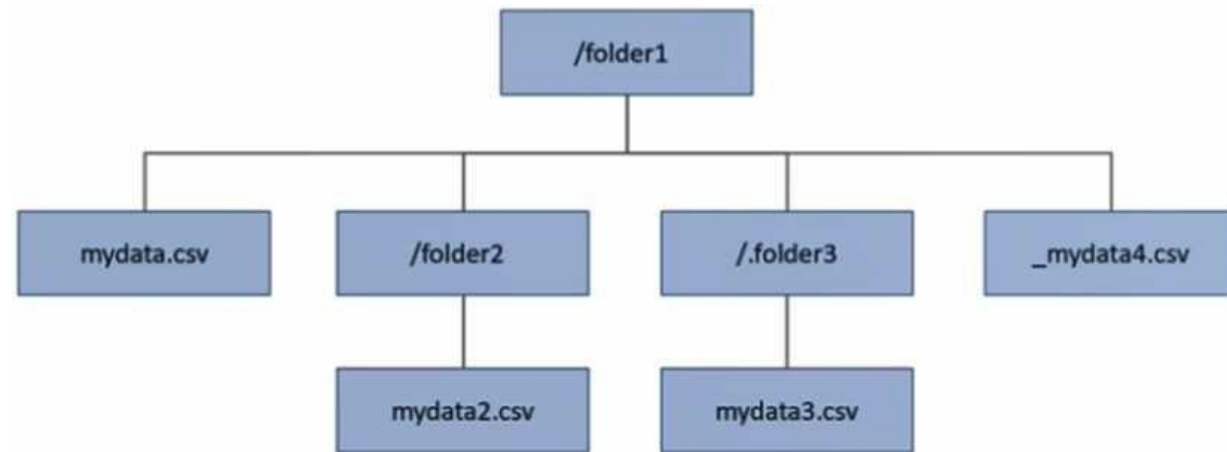
Section:

Explanation:

QUESTION 5

HOTSPOT

You have an Azure Data Lake Storage Gen2 account that contains a container named container1. You have an Azure Synapse Analytics serverless SQL pool that contains a native external table named dbo.Table1. The source data for dbo.Table1 is stored in container1. The folder structure of container1 is shown in the following exhibit.



The external data source is defined by using the following statement.

```

CREATE EXTERNAL DATA SOURCE DataLake
WITH
(
  LOCATION = 'https://mydatalake.dfs.core.windows.net/container1/folder1/**'
  , CREDENTIAL = DataLakeCred
);
  
```

For each of the following statements, select Yes if the statement is true. Otherwise, select No. NOTE: Each correct selection is worth one point.

Hot Area:

Statements	Yes	No
When selecting all the rows in dbo.Table1, data from the mydata2.csv file will be returned.	<input type="radio"/>	<input type="radio"/>
When selecting all the rows in dbo.Table1, data from the mydata3.csv file will be returned.	<input type="radio"/>	<input type="radio"/>
When selecting all the rows in dbo.Table1, data from the _mydata4.csv file will be returned.	<input type="radio"/>	<input type="radio"/>

Answer Area:

Statements	Yes	No
When selecting all the rows in dbo.Table1, data from the mydata2.csv file will be returned.	<input checked="" type="radio"/>	<input type="radio"/>
When selecting all the rows in dbo.Table1, data from the mydata3.csv file will be returned.	<input checked="" type="radio"/>	<input type="radio"/>
When selecting all the rows in dbo.Table1, data from the _mydata4.csv file will be returned.	<input type="radio"/>	<input checked="" type="radio"/>

Section:

Explanation:

Box 1: Yes

In the serverless SQL pool you can also use recursive wildcards /logs/** to reference Parquet or CSV files in any sub-folder beneath the referenced folder.

Box 2: Yes

Box 3: No

Reference: <https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-externaltables>

QUESTION 6

You have an Azure Data Factory pipeline named Pipeline1!. Pipelinel contains a copy activity that sends data to an Azure Data Lake Storage Gen2 account. Pipeline 1 is executed by a schedule trigger. You change the copy activity sink to a new storage account and merge the changes into the collaboration branch. After Pipelinel executes, you discover that data is NOT copied to the new storage account. You need to ensure that the data is copied to the new storage account. What should you do?

- A. Publish from the collaboration branch.
- B. Configure the change feed of the new storage account.
- C. Create a pull request.
- D. Modify the schedule trigger.

Correct Answer: A

Section:

Explanation:

CI/CD lifecycle

A development data factory is created and configured with Azure Repos Git. All developers should have permission to author Data Factory resources like pipelines and datasets. A developer creates a feature branch to make a change. They debug their pipeline runs with their most recent changes. After a developer is satisfied with their changes, they create a pull request from their feature branch to the main or collaboration branch to get their changes reviewed by peers.

After a pull request is approved and changes are merged in the main branch, the changes get published to the development factory.

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

QUESTION 7

You have an Azure Databricks workspace that contains a Delta Lake dimension table named Table1. Table1 is a Type 2 slowly changing dimension (SCD) table. You need to apply updates from a source table to Table1. Which Apache Spark SQL operation should you use?

- A. CREATE
- B. UPDATE
- C. MERGE
- D. ALTER

Correct Answer: C

Section:

Explanation:

The Delta provides the ability to infer the schema for data input which further reduces the effort required in managing the schema changes. The Slowly Changing Data(SCD) Type 2 records all the changes made to each key in the dimensional table. These operations require updating the existing rows to mark the previous values of the keys as old and then inserting new rows as the latest values. Also, Given a source table with the updates and the target table with dimensional data, SCD Type 2 can be expressed with the merge.

Example:

```
// Implementing SCD Type 2 operation using merge function
customersTable
.as("customers")
.merge(
stagedUpdates.as("staged_updates"),
"customers.customerId = mergeKey")
.whenMatched("customers.current = true AND customers.address <> staged_updates.address") .updateExpr(Map(
"current" -> "false",
"endDate" -> "staged_updates.effectiveDate"))
.whenNotMatched()
.insertExpr(Map(
"customerid" -> "staged_updates.customerId",
"address" -> "staged_updates.address",
"current" -> "true",
"effectiveDate" -> "staged_updates.effectiveDate",
"endDate" -> "null"))
.execute()
}
```



Reference:

<https://www.projectpro.io/recipes/what-is-slowly-changing-data-scd-type-2-operation-delta-tabledatabricks>

QUESTION 8

You have 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 in the UX Authoring canvas for ADF1. Which two actions should you perform? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

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

Correct Answer: B, D

Section:

QUESTION 9

You have an Azure Synapse Analytics dedicated SQL pool.

You need to create a fact table named Table1 that will store sales data from the last three years. The solution must be optimized for the following query operations:

Show order counts by week.

- Calculate sales totals by region.
- Calculate sales totals by product.
- Find all the orders from a given month.

Which data should you use to partition Table1?

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

Correct Answer: C

Section:

QUESTION 10

HOTSPOT

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.

Hot Area:



Create the table in:

- The dedicated SQL pool
- The serverless SQL pool
- The Spark pool

File format:

- Apache Parquet
- Delta
- JSON

Answer Area:

Create the table in:

- The dedicated SQL pool
- The serverless SQL pool
- The Spark pool

File format:

- Apache Parquet
- Delta
- JSON

Section:

Explanation:

QUESTION 11

HOTSPOT

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 table in SQLPoolM 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.

Hot Area:

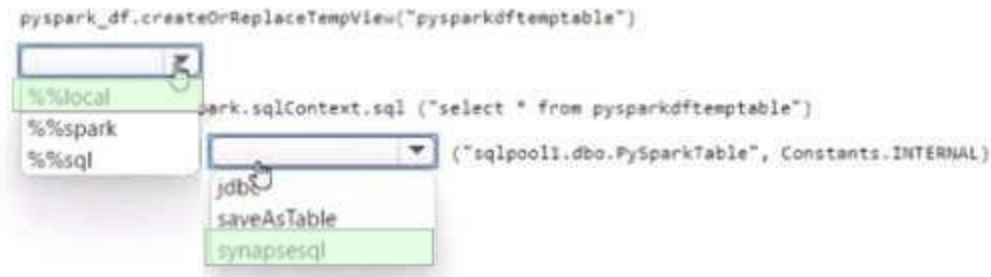
Answer Area

```
pyspark_df.createOrReplaceTempView("pysparkdftemptable")
spark.sqlContext.sql ("select * from pysparkdftemptable")
spark.saveAsTable ("sqlpool1.dbo.PySparkTable", Constants.INTERNAL)
```

Answer Area:



Answer Area



Section:

Explanation:

QUESTION 12

You are implementing a star schema in an Azure Synapse Analytics dedicated SQL pool. You plan to create a table named DimProduct.

DimProduct must be a Type 3 slowly changing dimension (SCD) table that meets the following requirements:

- The values in two columns named ProductKey and ProductSourceID will remain the same.
- The values in three columns named ProductName, ProductDescription, and Color can change. You need to add additional columns to complete the following table definition.

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



- A. [OriginalProductDescription] NVARCHAR(2000) NOT NULL
- B. [IsCurrentRow] [bit] NOT NULL
- C. [EffectiveStartDate] [datetime] NOT NULL
- D. [EffectiveEndDate] [datetime] NOT NULL
- E. [OriginalProductName] NVARCHAR(100) NULL
- F. [OriginalColor] NVARCHAR(50) NOT NULL

Correct Answer: A, B, C

Section:

QUESTION 13

You have an Azure Data Lake Storage Gen2 account that contains two folders named Folder and 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 use Azure Data Factory to copy multiple files from Folder1 to Folder2. 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

Correct Answer: A

Section:

Explanation:

Reference: <https://knowledge.informatica.com/s/article/Microsoft-Azure-Data-Lake-Store-Gen2-target-file-names-not-generating-as-expected-using-create-at-runtime-option>

QUESTION 14

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



Correct Answer: A

Section:

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-analytics/ba-p/2929112>

QUESTION 15

HOTSPOT

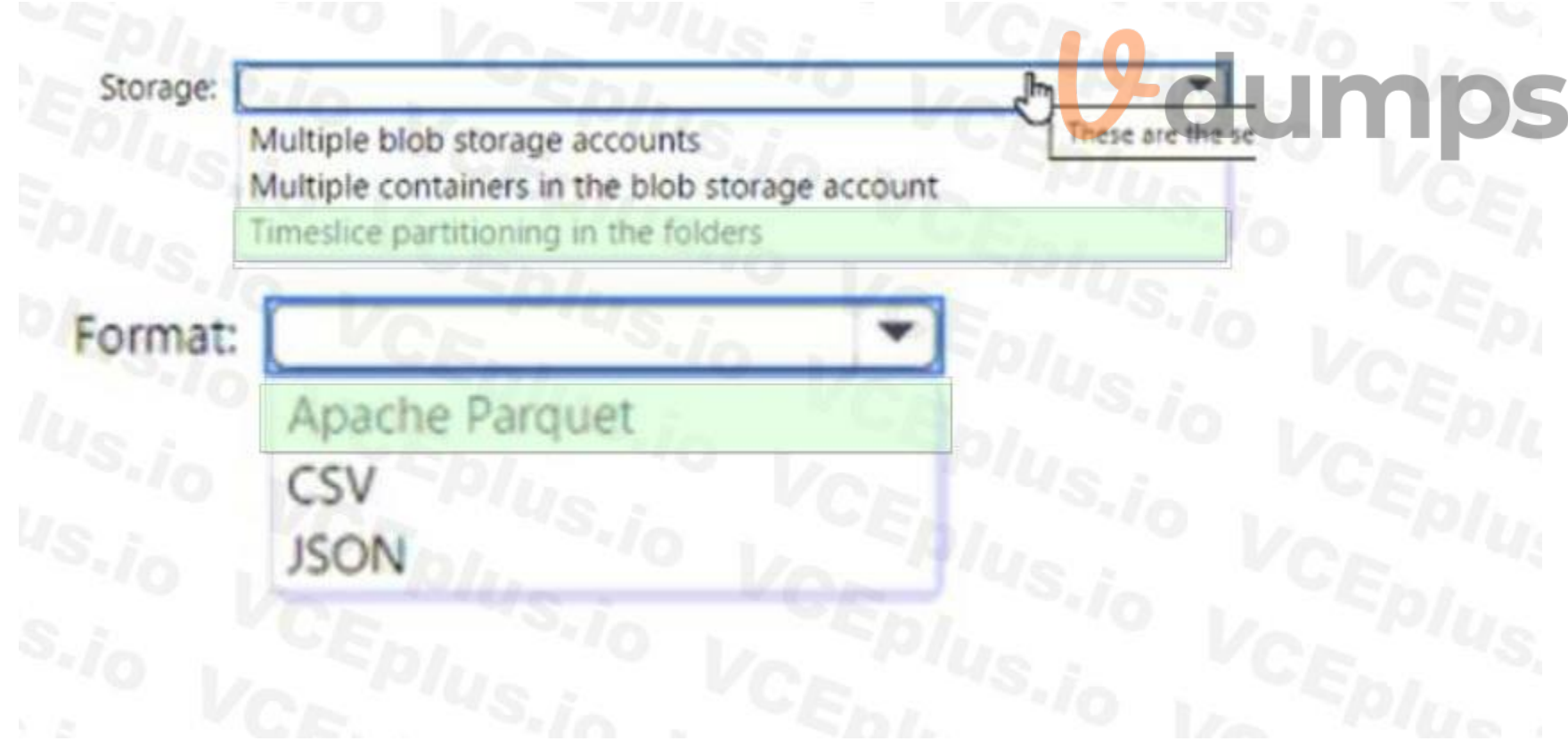
You have an Azure Blob storage account that contains a folder. The folder contains 120,000 files. Each file contains 62 columns. Each day, 1,500 new files are added to the folder.

You plan to incrementally load five data columns from each new file into an Azure Synapse Analytics workspace. You need to minimize how long it takes to perform the incremental loads. What should you use to store the files and format?

Hot Area:



Answer Area:



Section:

Explanation:

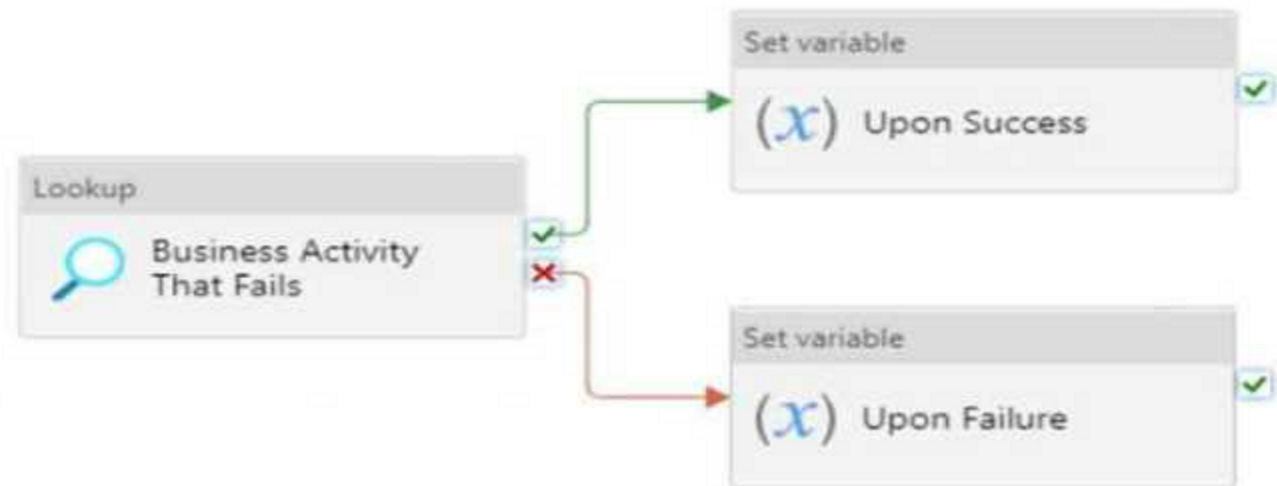
Box 1 = timeslice partitioning in the folders

This means that you should organize your files into folders based on a time attribute, such as year, month, day, or hour. For example, you can have a folder structure like /yyyy/mm/dd/file.csv. This way, you can easily identify and load only the new files that are added each day by using a time filter in your Azure Synapse pipeline¹². Timeslice partitioning can also improve the performance of data loading and querying by reducing the number of files that need to be scanned

This is because Parquet is a columnar file format that can efficiently store and compress data with many columns. Parquet files can also be partitioned by a time attribute, which can improve the performance of incremental loading and querying by reducing the number of files that need to be scanned¹²³. Parquet files are supported by both dedicated SQL pool and serverless SQL pool in Azure Synapse Analytics².

QUESTION 16

You have the Azure Synapse Analytics pipeline shown in the following exhibit.



You need to add a set variable activity to the pipeline to ensure that after the pipeline's completion, the status of the pipeline is always successful. What should you configure for the set variable activity?

- A. a success dependency on the Business Activity That Fails activity
- B. a failure dependency on the Upon Failure activity
- C. a skipped dependency on the Upon Success activity
- D. a skipped dependency on the Upon Failure activity



Correct Answer: B

Section:

Explanation:

A failure dependency means that the activity will run only if the previous activity fails. In this case, setting a failure dependency on the Upon Failure activity will ensure that the set variable activity will run after the pipeline fails and set the status of the pipeline to successful.
<https://www.validexamdumps.com>

QUESTION 17

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

Correct Answer: B

Section:

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.

QUESTION 18

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

Correct Answer: C

Section:

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>

QUESTION 19

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.

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

B.

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

C.

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

D.

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

E.

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

Correct Answer: C

Section:

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/>

QUESTION 20

You are designing an Azure Synapse Analytics workspace.

You need to recommend a solution to provide double encryption of all the data at rest. Which two components should you include in the recommendation? Each correct answer presents part of the solution NOTE: Each correct selection is worth one point.

- A. an X509 certificate
- B. an RSA key
- C. an Azure key vault that has purge protection enabled
- D. an Azure virtual network that has a network security group (NSG)
- E. an Azure Policy initiative

Correct Answer: A, D

Section:

Explanation:

QUESTION 21

You have an Azure Synapse Analytics dedicated SQL pool named Pool1 that contains a table named Sales. Sales has row-level security (RLS) applied. RLS uses the following predicate filter.

```
CREATE FUNCTION Security.fn_securitypredicate(@SalesRep AS sysname)
    RETURNS TABLE
WITH SCHEMABINDING
AS
    RETURN SELECT 1 AS fn_securitypredicate_result
    WHERE @SalesRep = USER_NAME() OR USER_NAME() = 'Manager';
```

A user named SalesUser1 is assigned the db_datareader role for Pool1.

A user named SalesUser1 is assigned the db_datareader role for Pool1. Which rows in the Sales table are returned when SalesUser1 queries the table?

- A. only the rows for which the value in the User_Name column is SalesUser1
- B. all the rows
- C. only the rows for which the value in the SalesRep column is Manager
- D. only the rows for which the value in the SalesRep column is SalesUser1

Correct Answer: A

Section:

QUESTION 22

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

Correct Answer: C

Section:

Explanation:

Use IDENTITY to create surrogate keys using dedicated SQL pool in Azure Synapse 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-datawarehouse-tables-identity>

QUESTION 23

HOTSPOT

You have an Azure Synapse Analytics dedicated SQL pool.

You need to create a table named FactInternetSales that will be a large fact table in a dimensional model. FactInternetSales will contain 100 million rows and two columns named SalesAmount and OrderQuantity. Queries executed on FactInternetSales will aggregate the values in SalesAmount and OrderQuantity from the last year for a specific product. The solution must minimize the data size and query execution time. How should you complete the code? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

```
CREATE TABLE [dbo].[FactInternetSales]
(
  [ProductKey] int NOT NULL
  , [OrderDateKey] int NOT NULL
  , [CustomerKey] int NOT NULL
  , [PromotionKey] int NOT NULL
  , [SalesOrderNumber] nvarchar(20) NOT NULL
  , [OrderQuantity] smallint NOT NULL
  , [UnitPrice] money NOT NULL
  , [SalesAmount] money NOT NULL
)
WITH
(
  CLUSTERED COLUMNSTORE INDEX
  ( CLUSTERED INDEX ([OrderDateKey])
  ( HEAP
  ( INDEX on [ProductKey]
  , DISTRIBUTION =
  );
```

(CLUSTERED COLUMNSTORE INDEX
(CLUSTERED INDEX ([OrderDateKey])
(HEAP
(INDEX on [ProductKey]

Hash([OrderDateKey])
Hash([ProductKey])
REPLICATE
ROUND_ROBIN

Answer Area:

Answer Area

```
CREATE TABLE [dbo].[FactInternetSales]
(
  [ProductKey] int NOT NULL
, [OrderDateKey] int NOT NULL
, [CustomerKey] int NOT NULL
, [PromotionKey] int NOT NULL
, [SalesOrderNumber] nvarchar(20) NOT NULL
, [OrderQuantity] smallint NOT NULL
, [UnitPrice] money NOT NULL
, [SalesAmount] money NOT NULL
)
WITH
(
  ( CLUSTERED COLUMNSTORE INDEX
  ( CLUSTERED INDEX ([OrderDateKey])
  ( HEAP
  ( INDEX on [ProductKey]
, DISTRIBUTION =
);
```

(CLUSTERED COLUMNSTORE INDEX
(CLUSTERED INDEX ([OrderDateKey])
(HEAP
(INDEX on [ProductKey]

Hash([OrderDateKey])
Hash([ProductKey])
REPLICATE
ROUND_ROBIN



Section:

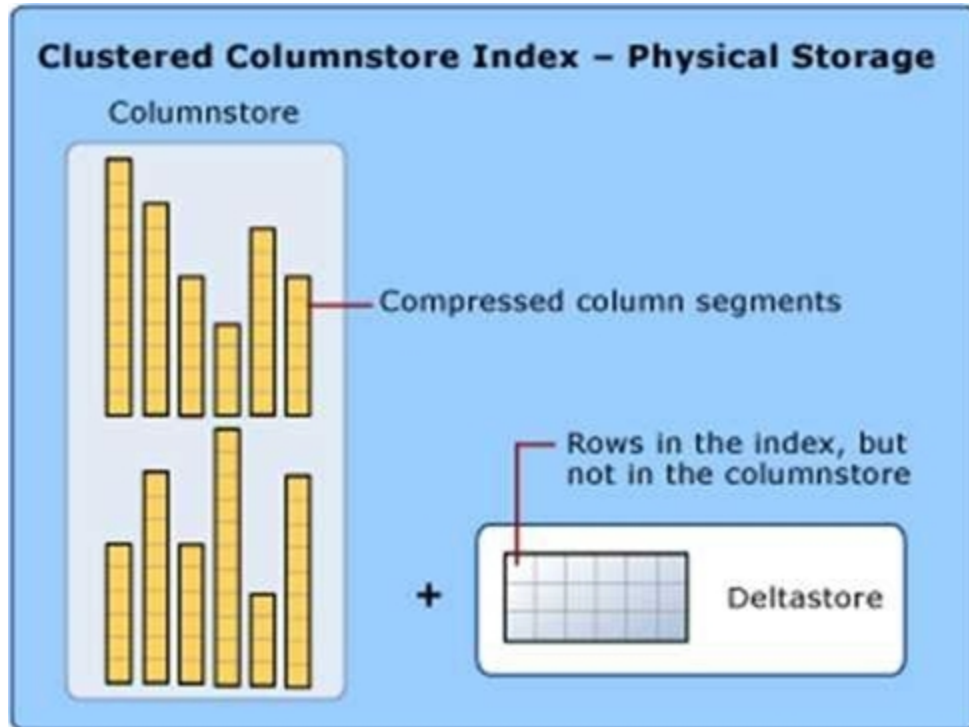
Explanation:

Box 1: (CLUSTERED COLUMNSTORE INDEX
CLUSTERED COLUMNSTORE INDEX

Columnstore indexes are the standard for storing and querying large data warehousing fact tables. This index uses column-based data storage and query processing to achieve gains up to 10 times the query performance in your data warehouse over traditional row-oriented storage. You can also achieve gains up to 10 times the data compression over the uncompressed data size. Beginning with SQL Server 2016 (13.x) SP1, columnstore indexes enable operational analytics: the ability to run performant real-time analytics on a transactional workload.

Note: Clustered columnstore index

A clustered columnstore index is the physical storage for the entire table.



To reduce fragmentation of the column segments and improve performance, the columnstore index might store some data temporarily into a clustered index called a deltastore and a B-tree list of IDs for deleted rows. The deltastore operations are handled behind the scenes. To return the correct query results, the clustered columnstore index combines query results from both the columnstore and the deltastore.

Box 2: HASH([ProductKey])

A hash distributed table distributes rows based on the value in the distribution column. A hash distributed table is designed to achieve high performance for queries on large tables. Choose a distribution column with data that distributes evenly
Reference:

<https://docs.microsoft.com/en-us/sql/relational-databases/indexes/columnstore-indexes-overview>

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

QUESTION 24

DRAG DROP

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. You need to design the table storage for pool1. The solution must meet the following requirements:

Maximize the performance of data loading operations to Staging.WebSessions. Minimize query times for reporting queries against the dimensional model. Which type of table distribution should you use for each table? To answer, drag the appropriate table distribution types to the correct tables.

Name	Number of rows	Update frequency	Description
Common.Date	7,300	New rows inserted yearly	<ul style="list-style-type: none"> Contains one row per date for the last 20 years Contains columns named Year, Month, Quarter, and IsWeekend
Marketing.WebSessions	1,500,500,000	Hourly inserts and updates	Fact table that contains counts of and updates sessions and page views, including foreign key values for date, channel, device, and medium
Staging.WebSessions	300,000	Hourly truncation and inserts	Staging table for web session data, truncation and including descriptive fields for inserts channel, device, and medium

Each table distribution type may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point.

Select and Place:

Table distribution types	Answer Area
Hash	Common.Data: <input type="text"/>
Replicated	Marketing.Web.Sessions: <input type="text"/>
Round-robin	Staging. Web.Sessions: <input type="text"/>

Correct Answer:

Table distribution types	Answer Area
<input type="text"/>	Common.Data: <input type="text" value="Replicated"/>
<input type="text"/>	Marketing.Web.Sessions: <input type="text" value="Hash"/>
<input type="text"/>	Staging. Web.Sessions: <input type="text" value="Round-robin"/>

Section:

Explanation:

Box 1: Replicated

The best table storage option for a small table is to replicate it across all the Compute nodes. Box 2: Hash

Hash-distribution improves query performance on large fact tables. Box 3: Round-robin

Round-robin distribution is useful for improving loading speed.

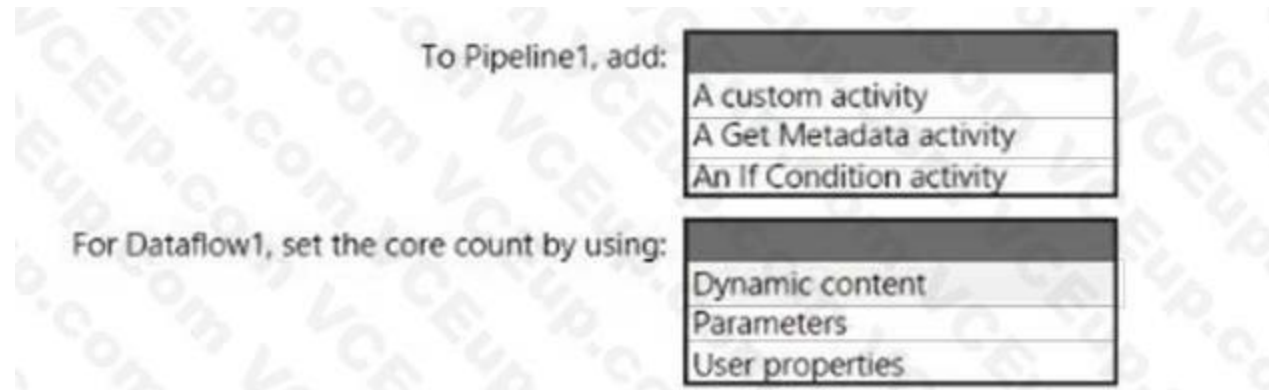
Reference: <https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-datawarehouse-tables-distribute>

QUESTION 25

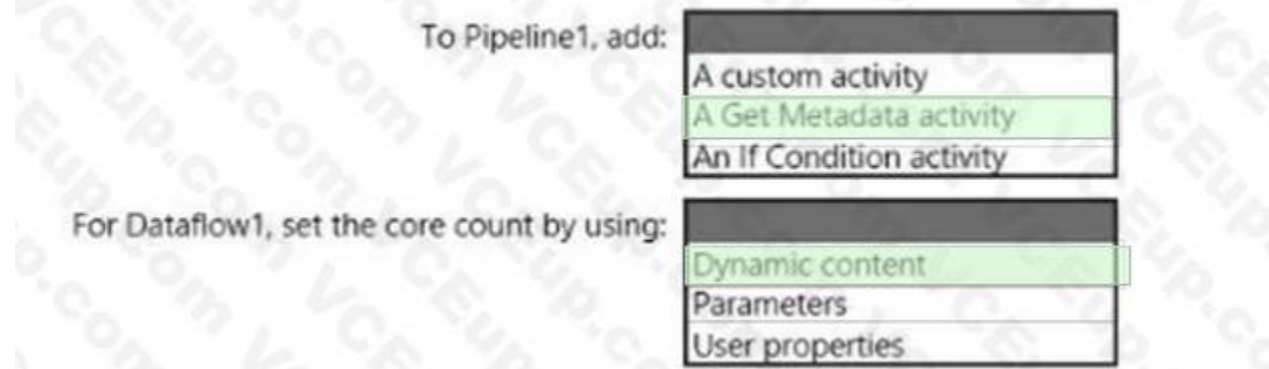
HOTSPOT

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.

Hot Area:



Answer Area:



Section:

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 dat

a. 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-flowactivity>

QUESTION 26

HOTSPOT

You have an Azure subscription that contains an Azure Synapse Analytics dedicated SQL pool named Pool1 and an Azure Data Lake Storage account named storage1. Storage1 requires secure transfers. You need to create an external data source in Pool1 that will be used to read .orc files in storage1. How should you complete the code? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

```
CREATE EXTERNAL DATA SOURCE AzureDataLakeStore
WITH
( Location1 'abfs://data@newyorktaxidataset.dfs.core.windows.net' ,
credential = ADLS_credential ,
TYPE -
);
```

Answer Area:

Answer Area

```
CREATE EXTERNAL DATA SOURCE AzureDataLakeStore
WITH
( Location1 'abfs://data@newyorktaxidataset.dfs.core.windows.net' ,
credential = ADLS_credential ,
TYPE -
);
```



Section:

Explanation:

<https://docs.microsoft.com/en-us/sql/t-sql/statements/create-external-data-source-transactsql?view=azure-sqldw-latest&preserve-view=true&tabs=dedicated>

QUESTION 27

DRAG DROP

You have an Azure subscription that contains an Azure Synapse Analytics workspace named workspace1. Workspace1 connects to an Azure DevOps repository named repo1. Repo1 contains a collaboration branch named main and a development branch named branch1. Branch1 contains an Azure Synapse pipeline named pipeline1. In workspace1, you complete testing of pipeline1. You need to schedule pipeline1 to run daily at 6 AM.

Which four actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order. NOTE: More than one order of answer choices is correct. You will receive credit for any of the correct orders you select.

Select and Place:

Actions

- Create a new branch in Repo1.
- Merge the changes from branch1 into main.
- Associate the schedule trigger with pipeline1.
- Switch to Synapse live mode.
- Create a schedule trigger.
- Publish the contents of main.

Answer Area

>

<

Correct Answer:

Actions

- Create a new branch in Repo1.
-
-
- Switch to Synapse live mode.
-
-

Answer Area

- Create a schedule trigger.
- Associate the schedule trigger with pipeline1.
- Merge the changes from branch1 into main.
- Publish the contents of main.

>

<

Section:

Explanation:

QUESTION 28

DRAG DROP

You have an Azure Data Lake Storage Gen 2 account named storage1. You need to recommend a solution for accessing the content in storage1. The solution must meet the following requirements:

List and read permissions must be granted at the storage account level. Additional permissions can be applied to individual objects in storage1. Security principals from Microsoft Azure Active Directory (Azure AD), part of Microsoft Entra, must be used for authentication. What should you use? To answer, drag the appropriate components to the correct requirements. Each component may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content. NOTE: Each correct selection is worth one point.

Select and Place:

Components

- Access control lists (ACLs)
- Role-based access control (RBAC) roles
- Shared access signatures (SAS)
- Shared account keys

Answer Area

To grant permissions at the storage account level:

To grant permissions at the object level:

Correct Answer:

Components

-
-
- Shared access signatures (SAS)
- Shared account keys

Answer Area

To grant permissions at the storage account level:

To grant permissions at the object level:

Section:

Explanation:

Box 1: Role-based access control (RBAC) roles

List and read permissions must be granted at the storage account level. Security principals from Microsoft Azure Active Directory (Azure AD), part of Microsoft Entra, must be used for authentication.

Role-based access control (Azure RBAC)

Azure RBAC uses role assignments to apply sets of permissions to security principals. A security principal is an object that represents a user, group, service principal, or managed identity that is defined in Azure Active Directory (AD). A permission set can give a security principal a "coarse-grain" level of access such as read or write access to all of the data in a storage account or all of the data in a container.

Box 2: Access control lists (ACLs)

Additional permissions can be applied to individual objects in storage1. Access control lists (ACLs)

ACLs give you the ability to apply "finer grain" level of access to directories and files. An ACL is a permission construct that contains a series of ACL entries. Each ACL entry associates security principal with an access level.

Reference: <https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-controlmodel>

QUESTION 29

HOTSPOT

You have an Azure subscription that contains an Azure Databricks workspace named databricks1 and an Azure Synapse Analytics workspace named synapse1. The synapse1 workspace contains an Apache Spark pool named pool1.

You need to share an Apache Hive catalog of pool1 with databricks1. What should you do? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

Hot Area:

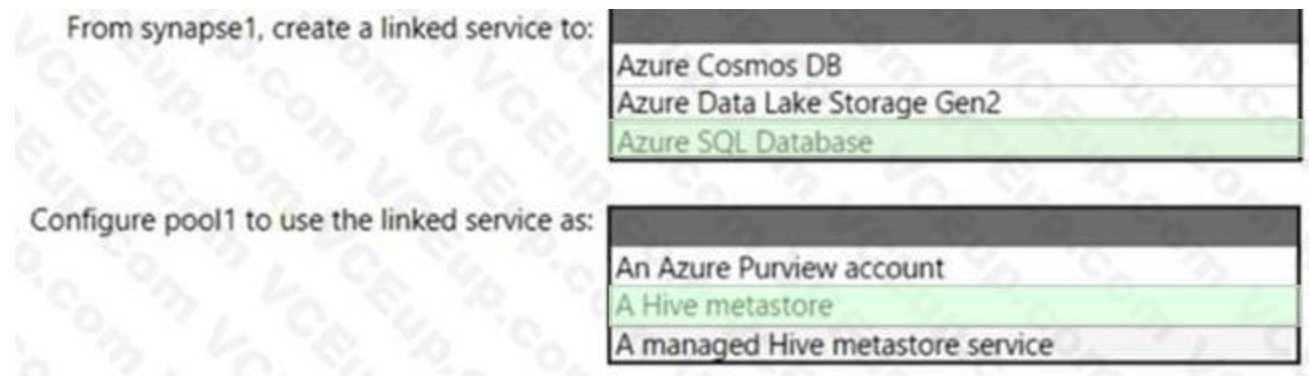
From synapse1, create a linked service to:

- Azure Cosmos DB
- Azure Data Lake Storage Gen2
- Azure SQL Database

Configure pool1 to use the linked service as:

- An Azure Purview account
- A Hive metastore
- A managed Hive metastore service

Answer Area:



Section:

Explanation:

Box 1: Azure SQL Database

Use external Hive Metastore for Synapse Spark Pool

Azure Synapse Analytics allows Apache Spark pools in the same workspace to share a managed HMS (Hive Metastore) compatible metastore as their catalog.

Set up linked service to Hive Metastore

Follow below steps to set up a linked service to the external Hive Metastore in Synapse workspace. Open Synapse Studio, go to Manage > Linked services at left, click New to create a new linked service. Set up Hive Metastore linked service

Choose Azure SQL Database or Azure Database for MySQL based on your database type, click Continue. Provide Name of the linked service. Record the name of the linked service, this info will be used to configure Spark shortly. You can either select Azure SQL Database/Azure Database for MySQL for the external Hive Metastore from Azure subscription list, or enter the info manually. Provide User name and Password to set up the connection.

Test connection to verify the username and password.

Click Create to create the linked service.

Box 2: A Hive Metastore

Reference: <https://docs.microsoft.com/en-us/azure/synapse-analytics/spark/apache-spark-externalmetastore>

QUESTION 30

DRAG DROP

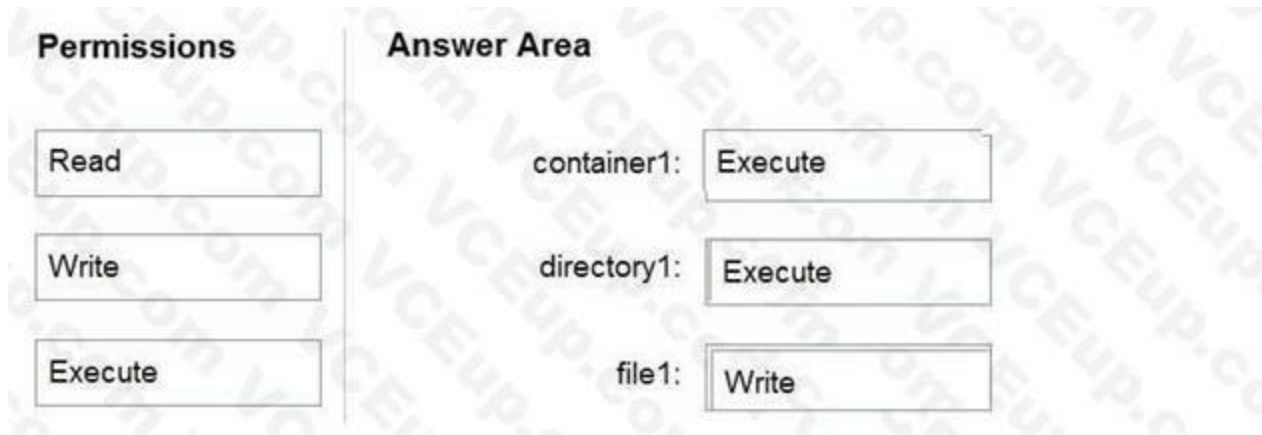
You have an Azure subscription that contains an Azure Data Lake Storage Gen2 account named storage1. Storage1 contains a container named container1. Container1 contains a directory named directory1. Directory1 contains a file named file1.

You have an Azure Active Directory (Azure AD) user named User1 that is assigned the Storage Blob Data Reader role for storage1. You need to ensure that User1 can append data to file1. The solution must use the principle of least privilege. Which permissions should you grant? To answer, drag the appropriate permissions to the correct resources. Each permission 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.

Select and Place:

Permissions	Answer Area
Read	container1: Permission
Write	directory1: Permission
Execute	file1: Permission

Correct Answer:



Section:

Explanation:

Box 1: Execute

If you are granting permissions by using only ACLs (no Azure RBAC), then to grant a security principal read or write access to a file, you'll need to give the security principal Execute permissions to the root folder of the container, and to each folder in the hierarchy of folders that lead to the file. Box 2: Execute

On Directory: Execute (X): Required to traverse the child items of a directory Box 3: Write

On file: Write (W): Can write or append to a file.

Reference:

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

QUESTION 31

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

Correct Answer: B

Section:

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-partition>

QUESTION 32

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

Correct Answer: A

Section:

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-columnsdynamically>

QUESTION 33

You have an Azure subscription that contains an Azure Synapse Analytics dedicated SQL pool named Pool1. Pool1 receives new data once every 24 hours. You have the following function.

```
create function dbo.udfFtoC(F decimal)
return decimal
as
begin
return (F - 32) * 5.0 / 9
end
```



You have the following query.

```
select avg_date, sensorid, avg_f, dbo.udfFtoC(avg_temperature) as avg_c from SensorTemps
where avg_date = @parameter
```

The query is executed once every 15 minutes and the @parameter value is set to the current date. You need to minimize the time it takes for the query to return results. 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 index on the avg_f column.
- B. Convert the avg_c column into a calculated column.
- C. Create an index on the sensorid column.
- D. Enable result set caching.
- E. Change the table distribution to replicate.

Correct Answer: B, D

Section:

QUESTION 34

You have an Azure Data Factory pipeline named pipeline1 that is invoked by a tumbling window trigger named Trigger1. Trigger1 has a recurrence of 60 minutes. You need to ensure that pipeline1 will execute only if the

previous execution completes successfully. How should you configure the self-dependency for Trigger1?

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

Correct Answer: D

Section:

Explanation:

Tumbling window self-dependency properties

In scenarios where the trigger shouldn't proceed to the next window until the preceding window is successfully completed, build a self-dependency. A self-dependency trigger that's dependent on the success of earlier runs of itself within the preceding hour will have the properties indicated in the following code.

Example code:

```
"name": "DemoSelfDependency",
"properties": {
  "runtimeState": "Started",
  "pipeline": {
    "pipelineReference": {
      "referenceName": "Demo",
      "type": "PipelineReference"
    }
  },
  "type": "TumblingWindowTrigger",
  "typeProperties": {
    "frequency": "Hour",
    "interval": 1,
    "startTime": "2018-10-04T00:00:00Z",
    "delay": "00:01:00",
    "maxConcurrency": 50,
    "retryPolicy": {
      "intervalInSeconds": 30
    },
    "dependsOn": [
      {
        "type": "SelfDependencyTumblingWindowTriggerReference",
        "size": "01:00:00",
        "offset": "-01:00:00"
      }
    ]
  }
}
```

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/tumbling-window-trigger-dependency>

QUESTION 35

You have an Azure subscription that contains an Azure Synapse Analytics dedicated SQL pool named SQLPool1. SQLPool1 is currently paused.



You need to restore the current state of SQLPool1 to a new SQL pool. What should you do first?

- A. Create a workspace.
- B. Create a user-defined restore point.
- C. Resume SQLPool1.
- D. Create a new SQL pool.

Correct Answer: B

Section:

Explanation:

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-restore-active-paused-dw>

QUESTION 36

HOTSPOT

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.

Hot Area:

Create:

Restrict row access by using:

Answer Area:

Create:

Restrict row access by using:

Section:

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.

QUESTION 37

You have an Azure SQL database named DB1 and an Azure Data Factory data pipeline named pipeline. From Data Factory, you configure a linked service to DB1.

In DB1, you create a stored procedure named SP1. SP1 returns a single row of data that has four columns. You need to add an activity to pipeline to execute SP1. The solution must ensure that the values in the columns are stored as pipeline variables. Which two types of activities can you use to execute SP1? (Refer to Data Engineering on Microsoft Azure documents or guide for Answers/Explanation available at Microsoft.com)

- A. Stored Procedure
- B. Lookup
- C. Script
- D. Copy

Correct Answer: A, B

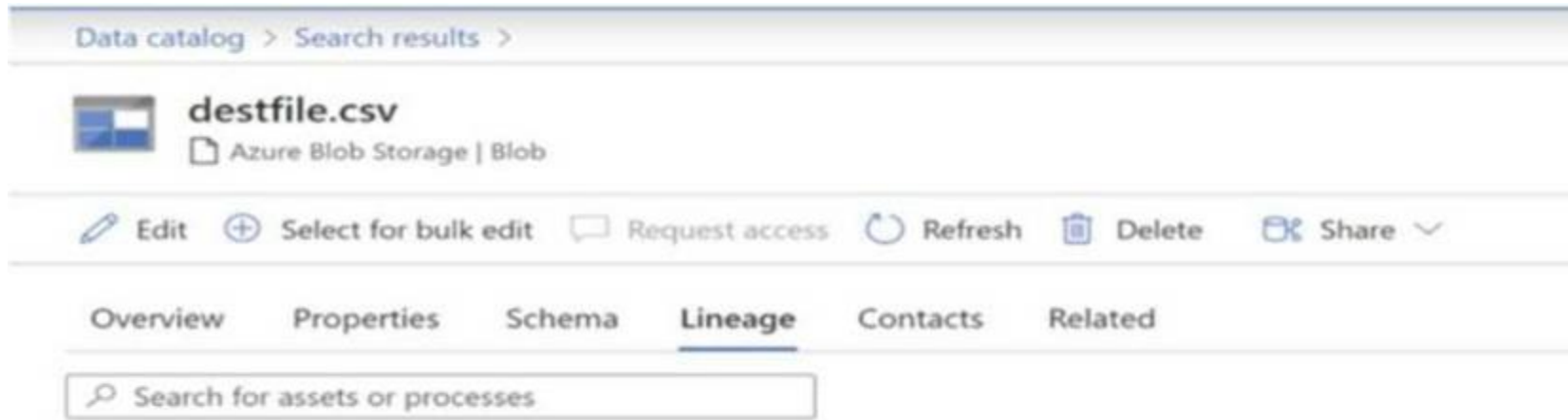
Section:

Explanation:

the two types of activities that you can use to execute SP1 are Stored Procedure and Lookup. A Stored Procedure activity executes a stored procedure on an Azure SQL Database or Azure Synapse Analytics or SQL Server1. You can specify the stored procedure name and parameters in the activity settings1. A Lookup activity retrieves a dataset from any data source that returns a single row of data with four columns2. You can use a query to execute a stored procedure as the source of the Lookup activity2. You can then store the values in the columns as pipeline variables by using expressions2. <https://learn.microsoft.com/en-us/azure/data-factory/transform-data-using-stored-procedure>

QUESTION 38

You have a Microsoft Purview account. The Lineage view of a CSV file is shown in the following exhibit.



How is the data for the lineage populated?

- A. manually
- B. by scanning data stores
- C. by executing a Data Factory pipeline

Correct Answer: B

Section:

Explanation:

According to Microsoft Purview Data Catalog lineage user guide¹, data lineage in Microsoft Purview is a core platform capability that populates the Microsoft Purview Data Map with data movement and transformations across systems². Lineage is captured as it flows in the enterprise and stitched without gaps irrespective of its source².

QUESTION 39

HOTSPOT

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.

Hot Area:

Answer Area

Partition the data:

- Partition by date with one partition per day.
- Partition by date with one partition per day.
- Partition by date with one partition per month.
- Partition by product.

Remove the data:

- Delete the old data from Table1 by using a WHERE clause.
- Delete the old data from Table1 by using a WHERE clause.
- Delete the old data from Table1 by using a JOIN.
- Switch the oldest partition to another table named Table2 and drop Table2.
- Truncate the oldest partition.

Answer Area:

Answer Area

Partition the data:

- Partition by date with one partition per day.
- Partition by date with one partition per month.
- Partition by product.

Remove the data:

- Delete the old data from Table1 by using a WHERE clause.
- Delete the old data from Table1 by using a JOIN.
- Switch the oldest partition to another table named Table2 and drop Table2.
- Truncate the oldest partition.

Section:

Explanation:

QUESTION 40

HOTSPOT

You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Sales.Orders. Sales.Orders contains a column named SalesRep.

You plan to implement row-level security (RLS) for Sales.Orders. You need to create the security policy that will be used to implement RLS. The solution must ensure that sales representatives only see rows for which the value of the SalesRep column matches their username. How should you complete the code? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

```
CREATE SCHEMA Security;
GO
CREATE FUNCTION Security.tvf_securitypredicate(@SalesRep AS nvarchar(50))
RETURNS TABLE
WITH
AS
RETURN SELECT 1 AS tvf_securitypredicate_result
WHERE @SalesRep = USER_NAME();
GO
CREATE SECURITY POLICY SalesFilter
```

WITH options:

- SCHEMABINDING
- ENCRYPTION
- RETURNS NULL ON NULL INPUT
- SCHEMABINDING

AS options:

- ADD FILTER PREDICATE Security.tvf_securitypredicate(SalesRep)
- ADD BLOCK PREDICATE Security.tvf_securitypredicate(SalesRep)
- ADD BLOCK PREDICATE tvf_securitypredicate_result
- ADD FILTER PREDICATE Security.tvf_securitypredicate(SalesRep)

Answer Area:

Answer Area

```
CREATE SCHEMA Security;  
GO  
CREATE FUNCTION Security.tvf_securitypredicate(@SalesRep AS nvarchar(50))  
RETURNS TABLE  
WITH SCHEMABINDING  
ENCRYPTION  
RETURNS NULL ON NULL INPUT  
AS  
RETURN SELECT 1 AS tvf_securitypredicate_result  
WHERE @SalesRep = USER_NAME();  
GO  
CREATE SECURITY POLICY SalesFilter  
ADD FILTER PREDICATE Security.tvf_securitypredicate(SalesRep)  
ADD BLOCK PREDICATE Security.tvf_securitypredicate(SalesRep)  
ADD BLOCK PREDICATE tvf_securitypredicate_result  
ADD FILTER PREDICATE Security.tvf_securitypredicate(SalesRep)
```

Section:
Explanation:

QUESTION 41
DRAG DROP

You are batch loading a table in an Azure Synapse Analytics dedicated SQL pool. You need to load data from a staging table to the target table. The solution must ensure that if an error occurs while loading the data to the target table, all the inserts in that batch are undone. How should you complete the Transact-SQL code? To answer, drag the appropriate values to the correct targets. Each value may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.

NOTE Each correct selection is worth one point.

Select and Place:



Values

BEGIN DISTRIBUTED TRANSACTION

BEGIN TRAN

COMMIT TRAN

ROLLBACK TRAN

SET RESULT_SET_CACHING ON

Answer Area

```

BEGIN TRY
  INSERT INTO dbo.Table1 (col1, col2, col3)
  SELECT col1, col2, col3 FROM stage.Table1;
END TRY
BEGIN CATCH
  IF @@TRANCOUNT > 0
  BEGIN
    ;
  END
END CATCH;
IF @@TRANCOUNT >0
BEGIN
  COMMIT TRAN;
END

```

Correct Answer:

Values

BEGIN DISTRIBUTED TRANSACTION

COMMIT TRAN

SET RESULT_SET_CACHING ON

Answer Area

BEGIN TRAN

```

BEGIN TRY
  INSERT INTO dbo.Table1 (col1, col2, col3)
  SELECT col1, col2, col3 FROM stage.Table1;
END TRY
BEGIN CATCH
  IF @@TRANCOUNT > 0
  BEGIN
    ROLLBACK TRAN ;
  END
END CATCH;
IF @@TRANCOUNT >0
BEGIN
  COMMIT TRAN;
END

```



Section:

Explanation:

QUESTION 42

You have an Azure subscription that contains an Azure SQL database named DB1 and a storage account named storage1. The storage1 account contains a file named File1.txt. File1.txt contains the names of selected tables in DB1. You need to use an Azure Synapse pipeline to copy data from the selected tables in DB1 to the files in storage1. The solution must meet the following requirements:

- The Copy activity in the pipeline must be parameterized to use the data in File1.txt to identify the source and destination of the copy.
- Copy activities must occur in parallel as often as possible. Which two pipeline activities should you include in the pipeline? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

- A. If Condition
- B. ForEach
- C. Lookup
- D. Get Metadata

Correct Answer: A, D

Section:

Explanation:

QUESTION 43

You have an Azure data factory that connects to a Microsoft Purview account. The data factory is registered in Microsoft Purview. You update a Data Factory pipeline. You need to ensure that the updated lineage is available in Microsoft Purview. What should you do first?

- A. Locate the related asset in the Microsoft Purview portal.
- B. Execute the pipeline.
- C. Disconnect the Microsoft Purview account from the data factory.
- D. Execute an Azure DevOps build pipeline.

Correct Answer: B

Section:



QUESTION 44

You plan to use an Apache Spark pool in Azure Synapse Analytics to load data to an Azure Data Lake Storage Gen2 account. You need to recommend which file format to use to store the data in the Data Lake Storage account. The solution must meet the following requirements:

- Column names and data types must be defined within the files loaded to the Data Lake Storage account.
- Data must be accessible by using queries from an Azure Synapse Analytics serverless SQL pool.
- Partition elimination must be supported without having to specify a specific partition. What should you recommend?

- A. Delta Lake
- B. JSON
- C. CSV
- D. ORC

Correct Answer: D

Section:

QUESTION 45

HOTSPOT

You have two Azure SQL databases named DB1 and DB2.

DB1 contains a table named Table1. Table1 contains a timestamp column named LastModifiedOn. LastModifiedOn contains the timestamp of the most recent update for each individual row. DB2 contains a table named Watermark. Watermark contains a single timestamp column named WatermarkValue. You plan to create an Azure Data Factory pipeline that will incrementally upload into Azure Blob Storage all the rows in Table1 for which the LastModifiedOn column contains a timestamp newer than the most recent value of the WatermarkValue column in Watermark.

You need to identify which activities to include in the pipeline. The solution must meet the following requirements:

- Minimize the effort to author the pipeline.
- Ensure that the number of data integration units allocated to the upload operation can be controlled. What should you identify? To answer, select the appropriate options in the answer area.

Hot Area:

Answer Area

To retrieve the watermark value, use:

Lookup
Filter
Get Metadata
Lookup

To perform the upload, use:

Copy data
Copy data
Custom
Data flow

Answer Area:

Answer Area

To retrieve the watermark value, use:

Lookup
Filter
Get Metadata
Lookup

To perform the upload, use:

Copy data
Copy data
Custom
Data flow

Section:

Explanation:

QUESTION 46

You have an Azure Synapse Analytics dedicated SQL pool.

You plan to create a fact table named Table1 that will contain a clustered columnstore index.

You need to optimize data compression and query performance for Table1.

What is the minimum number of rows that Table1 should contain before you create partitions?

- A. 100,000
- B. 600,000
- C. 1 million
- D. 60 million

Correct Answer: A

Section:

QUESTION 47

You have an Azure subscription that contains an Azure Data Factory data pipeline named Pipeline1, a Log Analytics workspace named LA1, and a storage account named account1.

You need to retain pipeline-run data for 90 days. The solution must meet the following requirements:

* The pipeline-run data must be removed automatically after 90 days.

* Ongoing costs must be minimized.

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

- A. Configure Pipeline1 to send logs to LA1.
- B. From the Diagnostic settings (classic) settings of account1, set the retention period to 90 days.
- C. Configure Pipeline1 to send logs to account1.
- D. From the Data Retention settings of LA1, set the data retention period to 90 days.

Correct Answer: A, B

Section:

QUESTION 48

You have an Azure data factory named ADM.

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 in the Azure Data Factory Studio 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 Azure Data Factory Studio, select Publish.
- C. From the Azure Data Factory Studio, run Publish All.
- D. Create a Git repository.
- E. Create a GitHub action.
- F. From the Azure Data Factory Studio, select up code repository.

Correct Answer: D, F

Section:

QUESTION 49

HOTSPOT

You have an Azure Synapse Analytics pipeline named pipeline1 that has concurrency set to 1.

To run pipeline 1, you create a new trigger as shown in the following exhibit.



Type *
Schedule

Start date * ⓘ
01/03/2023, 9:59:33 pm

Time zone * ⓘ
Coordinated Universal Time (UTC)

Recurrence * ⓘ
Every 1 Day(s)

Advanced recurrence options

Execute at these times ⓘ

Hours: 10 X 12 X 17 X

Minutes: 30 X 45 X

Use the drop-down menus to select the answer choice that completes each statement based on the information presented in the [graphic].
NOTE: Each correct selection is worth one point.

Hot Area:
Answer Area

The trigger will run at [answer choice].

If the previous execution of pipeline1 is still running when the trigger fires next, the new triggered execution will [answer choice].

Answer Area:
Answer Area

The trigger will run at [answer choice].

If the previous execution of pipeline1 is still running when the trigger fires next, the new triggered execution will [answer choice].

Section:
Explanation:

QUESTION 50

HOTSPOT

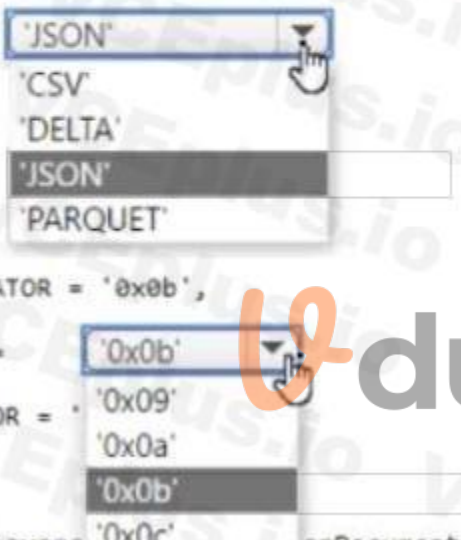
You have an Azure Synapse serverless SQL pool.

You need to read JSON documents from a file by using the OPENROWSET function. How should you complete the query? To answer, select the appropriate options in the answer area. **NOTE:** Each correct selection is worth one point.

Hot Area:

Answer Area

```
SELECT *  
FROM OPENROWSET  
(  
    BULK  
    'https://sourcedatalake.blob.core.windows.net/public/docs.json',  
    FORMAT =  
        'JSON',  
    FIELDTERMINATOR = '0x0b',  
    FIELDQUOTE =  
        '0x0b',  
    ROWTERMINATOR =  
        '0x0b',  
    WITH (jsondoc nvarchar(1000) onDocuments)
```



Answer Area:

```

SELECT *
FROM OPENROWSET
(
    BULK
    'https://sourcedatalake.blob.core.windows.net/public/docs.json',
    FORMAT = 'JSON',
    FIELDTERMINATOR = '\b',
    FIELDQUOTE = '\b',
    ROWTERMINATOR = '\b'
)
WITH (jsondoc nvarchar(1000))
SELECT jsondoc

```

Section:
Explanation:

QUESTION 51

HOTSPOT

You have an Azure Data Factory pipeline shown the following exhibit.



The execution log for the first pipeline run is shown in the following exhibit.

Activity runs

Pipeline run ID 87f89922-14fa-468f-b13f-2f86760614ff

All status ▾

Showing 1 - 2 items

Activity name ↑↓	Activity type ↑↓	Run start ↑↓	Duration ↑↓	Status ↑↓
Web_GetIP	Web	Nov 10, 2022, 11:11:36 a	00:00:02	Failed
Exec_COPY_BLOB	Execute Pipeline	Nov 10, 2022, 11:11:25 a	00:00:11	Succeeded

The execution log for the second pipeline run is shown in the following exhibit.

Activity runs

Pipeline run ID: a7b5b522-cfaf-4c09-b3a9-fb42986be984

All status ▾

Showing 1 - 3 items

Activity name ↑↓	Activity type ↑↓	Run start ↑↓	Duration ↑↓	Status ↑↓
Set status	Set variable	Nov 10, 2022, 11:13:17 a	00:00:01	✔ Succeeded
Web_GetIP	Web	Nov 10, 2022, 11:12:59 a	00:00:16	✔ Succeeded
Exec_COPY_BLOB	Execute Pipeline	Nov 10, 2022, 11:12:48 a	00:00:11	⊘ Skipped

For each of the following statements, select Yes if the statement is true. Otherwise, select No. NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Statements	Yes	No
The <code>retry</code> property of the Web_GetIP activity is set to 1.	<input type="radio"/>	<input type="radio"/>
The <code>waitOnCompletion</code> property of the Exec_COPY_BLOB activity is set to true.	<input type="radio"/>	<input type="radio"/>
The Exec_COPY_BLOB activity was skipped during the second run due to pipeline dependencies.	<input type="radio"/>	<input type="radio"/>

Answer Area:

Answer Area

Statements	Yes	No
The <code>retry</code> property of the Web_GetIP activity is set to 1.	<input type="radio"/>	<input checked="" type="radio"/>
The <code>waitOnCompletion</code> property of the Exec_COPY_BLOB activity is set to true.	<input type="radio"/>	<input checked="" type="radio"/>
The Exec_COPY_BLOB activity was skipped during the second run due to pipeline dependencies.	<input type="radio"/>	<input checked="" type="radio"/>

Section:

Explanation:

QUESTION 52

You are designing 2 solution that will use tables in Delta Lake on Azure Databricks. You need to minimize how long it takes to perform the following:

*Queries against non-partitioned tables

* Joins on non-partitioned columns

Which two options should you include in the solution? Each correct answer presents part of the solution. (Choose Correct Answer and Give Explanation and Reference to Support the answers based from Data Engineering on Microsoft Azure)

- A. Z-Ordering
- B. Apache Spark caching
- C. dynamic file pruning (DFP)
- D. the clone command

Correct Answer: A, C

Section:

Explanation:

- A. Z-Ordering
- B. Apache Spark caching
- C. dynamic file pruning (DFP)
- D. the clone command

Answer: AB

Explanation:

According to the information I found on the web, two options that you should include in the solution to minimize how long it takes to perform queries and joins on non-partitioned tables are:

Z-Ordering: This is a technique to colocate related information in the same set of files. This colocality is automatically used by Delta Lake in data-skipping algorithms. This behavior dramatically reduces the amount of data that Delta Lake on Azure Databricks needs to read. Apache Spark caching: This is a feature that allows you to cache data in memory or on disk for faster access. Caching can improve the performance of repeated queries and joins on the same data. You can cache Delta tables using the `CACHE TABLE` or `CACHE LAZY` commands.

To minimize the time it takes to perform queries against non-partitioned tables and joins on nonpartitioned columns in Delta Lake on Azure Databricks, the following options should be included in the solution:

1. Z-Ordering: Z-Ordering improves query performance by co-locating data that share the same column values in the same physical partitions. This reduces the need for shuffling data across nodes during query execution. By using Z-Ordering, you can avoid full table scans and reduce the amount of data processed. 2. Apache Spark caching: Caching data in memory can improve query performance by reducing the amount of data read from disk.

This helps to speed up subsequent queries that need to access the same data. When you cache a table, the data is read from the data source and stored in memory.

Subsequent queries can then read the data from memory, which is much faster than reading it from disk.

Reference:

Delta Lake on Databricks: <https://docs.databricks.com/delta/index.html>

Best Practices for Delta Lake on Databricks: <https://databricks.com/blog/2020/05/14/best-practicesfor-delta-lake-on-databricks.html>

QUESTION 53

You are building a data flow in Azure Data Factory that upserts data into a table in an Azure Synapse Analytics dedicated SQL pool. You need to add a transformation to the data flow. The transformation must specify logic indicating when a row from the input data must be upserted into the sink.

Which type of transformation should you add to the data flow?

- A. join
- B. select
- C. surrogate key
- D. alter row



Correct Answer: D

Section:

Explanation:

The alter row transformation allows you to specify insert, update, delete, and upsert policies on rows based on expressions. You can use the alter row transformation to perform upserts on a sink table by matching on a key column and setting the appropriate row policy.

QUESTION 54

HOTSPOT

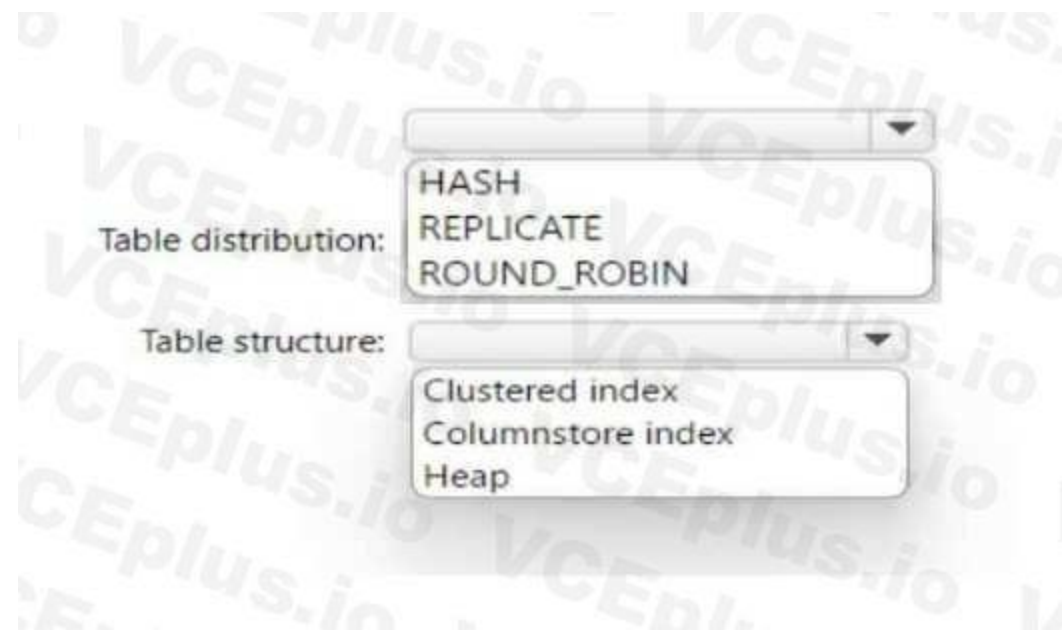
You are incrementally loading data into fact tables in an Azure Synapse Analytics dedicated SQL pool.

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.

Hot Area:



Answer Area:



Section:

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.

QUESTION 55

DRAG DROP

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 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.

Select and Place:

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
```

Correct Answer:

Values

DATA_SOURCE

LOCATION

Answer Area

```
SELECT TOP 100 *
FROM OPENROWSET (
BULK [ ] 'https://adls1.dfs.core.windows.net/container1/folder1/*.csv',
FORMAT = 'CSV') AS rows
```

Section:

Explanation:

QUESTION 56

HOTSPOT

You have an Azure Synapse Analytics dedicated SQL pool.

You need to monitor the database for long-running queries and identify which queries are waiting on resources.

Which dynamic management view should you use for each requirement? To answer, select the appropriate options in the answer area.

NOTE: Each correct answer is worth one point.

Hot Area:

Answer Area

Monitor the database for long-running queries:

- sys.dm_pdw_exec_requests
- sys.dm_pdw_exec_requests
- sys.dm_pdw_sql_requests
- sys.dm_pdw_exec_sessions

Identify which queries are waiting on resources:

- sys.dm_pdw_lock_waits
- sys.dm_pdw_waits
- sys.dm_pdw_lock_waits
- sys.resource_governor_workload_groups

Answer Area:

Answer Area

Monitor the database for long-running queries:

- sys.dm_pdw_exec_requests
- sys.dm_pdw_sql_requests
- sys.dm_pdw_exec_sessions

Identify which queries are waiting on resources:

- sys.dm_pdw_waits
- sys.dm_pdw_lock_waits
- sys.resource_governor_workload_groups

Section:

Explanation:

QUESTION 57

HOTSPOT

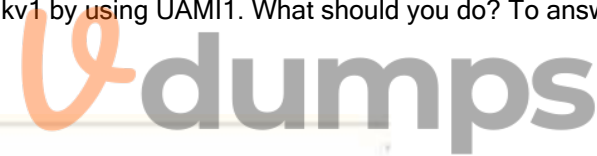
You have an Azure subscription that contains the resources shown in the following table.

Name	Type	Description
ws1	Azure Synapse Analytics workspace	None
kv1	Azure Key Vault	None
UAMI1	User-assigned managed identity	Associated with ws1
sp1	Apache Spark pool in Azure Synapse Analytics	Associated with ws1

You need to ensure that you can Spark notebooks in ws1. The solution must ensure secrets from kv1 by using UAMI1. What should you do? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:



Answer Area

In the Azure portal:

- Add a role-based access control (RBAC) role to kv1.
- Add a role-based access control (RBAC) role to ws1.
- Create a linked service to kv1.

In Synapse Studio:

- Add a role-based access control (RBAC) role to kv1.
- Add a role-based access control (RBAC) role to ws1.
- Create a linked service to kv1.

Answer Area:

Answer Area

In the Azure portal:

- Add a role-based access control (RBAC) role to kv1.
- Add a role-based access control (RBAC) role to ws1.
- Create a linked service to kv1.

In Synapse Studio:

- Add a role-based access control (RBAC) role to kv1.
- Add a role-based access control (RBAC) role to ws1.
- Create a linked service to kv1.

Section:

Explanation:

QUESTION 58

You have an Azure Synapse Analytics dedicated SQL pod. You need to create a pipeline that will execute a stored procedure in the dedicated SQL pool and use the returned result set as the input (or a downstream activity). The solution must minimize development effort. Which Type of activity should you use in the pipeline?

- A. Notebook
- B. U-SQL
- C. Script
- D. Stored Procedure

Correct Answer: D
Section:

QUESTION 59

You have an Azure subscription that contains an Azure Synapse Analytics workspace named ws1 and an Azure Cosmos D6 database account named Cosmos1. Cosmos1 contains a container named container1 and ws1 contains a serverless1 SQL pool. You need to ensure that you can query the data in container1 by using the serverless1 SQL pool. Which three actions should you perform? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

- A. Enable Azure Synapse Link for Cosmos1
- B. Disable the analytical store for container1.
- C. In ws1, create a linked service that references Cosmos1
- D. Enable the analytical store for container1
- E. Disable indexing for container1

Correct Answer: A, C, D
Section:

QUESTION 60

HOTSPOT

You are developing an Azure Synapse Analytics pipeline that will include a mapping data flow named Dataflow1. Dataflow1 will read customer data from an external source and use a Type 1 slowly changing dimension (SCD) when loading the data into a table named DimCustomer1 in an Azure Synapse Analytics dedicated SQL pool.

You need to ensure that Dataflow1 can perform the following tasks:

- * Detect whether the data of a given customer has changed in the DimCustomer table.
- Perform an upsert to the DimCustomer table.

Which type of transformation should you use for each task? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

Hot Area:

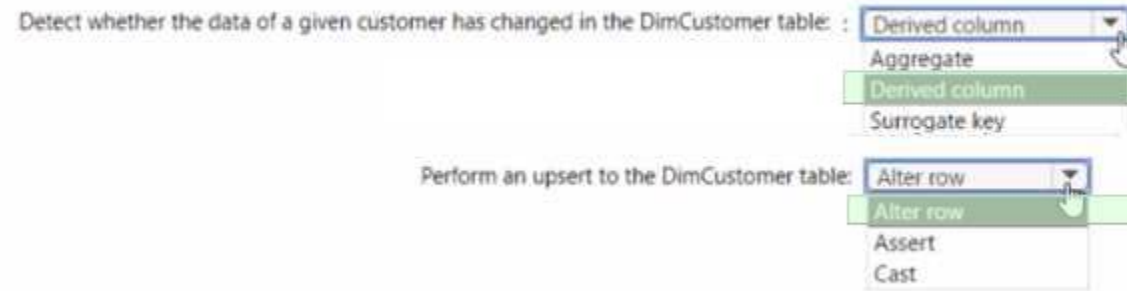
Answer Area

Detect whether the data of a given customer has changed in the DimCustomer table: :
Derived column
Aggregate
Derived column
Surrogate key

Perform an upsert to the DimCustomer table:
Alter row
Alter row
Assert
Cast

Answer Area:

Answer Area



Section:

Explanation:

QUESTION 61

You have an Azure data factory named ADM that contains a pipeline named Pipeline1. Pipeline1 must execute every 30 minutes with a 15-minute offset.

You need to create a trigger for Pipeline1. The trigger must meet the following requirements:

- Backfill data from the beginning of the day to the current time.
- If Pipeline1 fails, ensure that the pipeline can re-execute within the same 30-minute period.
- Ensure that only one concurrent pipeline execution can occur.
- Minimize development and configuration effort

Which type of trigger should you create?

- A. schedule
- B. event-based
- C. manual
- D. tumbling window

Correct Answer: D

Section:

QUESTION 62

Note: The 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 Data Lake Storage account that contains a staging zone.

You need to design a dairy process to ingest incremental data from the staging zone, transform the data by executing an R script, and then insert the transformed data into a data warehouse in Azure Synapse Analytics.

Solution: You use an Azure Data Factory schedule trigger to execute a pipeline that executes a mapping data flow. and then inserts the data into the data warehouse.

Does this meet the goal?

- A. Yes
- B. No

Correct Answer: B

Section:

Explanation:

QUESTION 63

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



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

Correct Answer: C

Section:

QUESTION 64

You have two Azure Blob Storage accounts named account1 and account2?

You plan to create an Azure Data Factory pipeline that will use scheduled intervals to replicate newly created or modified blobs from account1 to account?

You need to recommend a solution to implement the pipeline. The solution must meet the following requirements:

* Ensure that the pipeline only copies blobs that were created or modified since the most recent replication event.

* Minimize the effort to create the pipeline.

What should you recommend?

- A. Create a pipeline that contains a flowlet.
- B. Create a pipeline that contains a Data Flow activity.
- C. Run the Copy Data tool and select Metadata-driven copy task.
- D. Run the Copy Data tool and select Built-in copy task.

Correct Answer: A

Section:

QUESTION 65

HOTSPOT

You have an Azure data factory.

You execute a pipeline that contains an activity named Activity1. Activity1 produces the following output.

```
{
  ...
  "dataRead": 1208,
  "dataWritten": 1208,
  "filesRead": 1,
  "filesWritten": 1,
  "sourcePeakConnections": 3,
  "sinkPeakConnections": 2,
  "copyDuration": 13,
  "throughput": 0.147,
  "effectiveIntegrationRuntime": "AutoResolveIntegrationRuntime (West Central US)",
  "usedDataIntegrationUnits": 4,
  "reportLineageToPurview": {
    "status": "Succeeded",
    "durationInSeconds": "4"
  }
  ...
}
```

For each of the following statements select Yes if the statement is true. Otherwise, select No.

NOTE: Each correct selection is worth one point.



Hot Area:

Answer Area

Statements	Yes	No
Activity1 is a Copy activity.	<input type="radio"/>	<input type="radio"/>
Activity1 is executed by using a self-hosted integration runtime.	<input type="radio"/>	<input type="radio"/>
The data factory that executed the pipeline is connected to Microsoft Purview.	<input type="radio"/>	<input type="radio"/>

Answer Area:

Answer Area

Statements	Yes	No
Activity1 is a Copy activity.	<input checked="" type="radio"/>	<input type="radio"/>
Activity1 is executed by using a self-hosted integration runtime.	<input checked="" type="radio"/>	<input type="radio"/>
The data factory that executed the pipeline is connected to Microsoft Purview.	<input type="radio"/>	<input checked="" type="radio"/>

Section:

Explanation:

QUESTION 66

HOTSPOT

You have an Azure data factory that has the Git repository settings shown in the following exhibit.



Git repository

Git repository information associated with your data factory. [CI/CD best practices](#)

[Edit](#) [Overwrite live mode](#) [Disconnect](#) [Import resources](#)

Repository type	Azure DevOps Git
Azure DevOps Account	
Project name	ADFDemo
Repository name	ADFDemo
Collaboration branch	main
Publish branch	adf_publish
Root folder	/
Last published commit	23b144ac4aa7daf16f2fe7c2ab0eb303a8e4ed65
Publish (from ADF Studio)	Enabled

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

NOTE: Each correct answer is worth one point.

Hot Area:

Answer Area

Changes to pipelines will be saved in Azure DevOps [answer choice].

To publish changes by using Azure Data Factory Studio, the changes must first be saved in the [answer choice].

Answer Area:

Answer Area

Changes to pipelines will be saved in Azure DevOps [answer choice].

To publish changes by using Azure Data Factory Studio, the changes must first be saved in the [answer choice].

Section:

Explanation:

QUESTION 67

HOTSPOT

You have an Azure subscription that contains the resources shown in the following table.



Name	Type	Description
ws1	Azure Synapse Analytics workspace	Contains a pipeline named pipeline1
storage1	Azure Data Lake Storage account	Used to store Apache Parquet data files that include LIST and STRUCT data types
SQL1	Azure Synapse Analytics dedicated SQL pool	None

You need to ingest the Parquet files from storage1 to SQL1 by using pipeline1. The solution must meet the following requirements:

- * Minimize complexity.
- * Ensure that additional columns in the files are processed as strings.
- * Ensure that files containing additional columns are processed successfully.

How should you configure pipeline1? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Ingest the data from storage1 to SQL1 by using:

A copy activity
A copy activity
A custom activity
A data flow activity
An Execute SSIS package activity

In Source settings, enable:

Allow schema drift
Allow schema drift
Infer drifted column types
Sampling
Validate schema

Answer Area:

Answer Area

Ingest the data from storage1 to SQL1 by using:

A copy activity
A copy activity
A custom activity
A data flow activity
An Execute SSIS package activity

In Source settings, enable:

Allow schema drift
Allow schema drift
Infer drifted column types
Sampling
Validate schema

Section:

Explanation:

QUESTION 68

You have an Azure data factor/ connected to a Git repository that contains the following branches:

- * mam: Collaboration branch
- * abc: Feature branch
- * xyz: Feature branch

You save changes to a pipeline in the xyz branch.

You need to publish the changes to the live service

What should you do first?

- A. Push the code to a remote origin.
- B. Publish the data factory.
- C. Create a pull request to merge the changes into the abc branch.
- D. Create a pull request to merge the changes into the main branch.

Correct Answer: D

Section:

QUESTION 69

You have an Azure Synapse Analytics dedicated SQL pool named Pcol1. Pool1 contains a table named table1.

You load 5 TB of data into table1.

You need to ensure that column store compression is maximized for table1.

Which statement should you execute?

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

Correct Answer: B

Section:

QUESTION 70

DRAG DROP

you have a project in Azure DevOps that contains a repository named Repo1. Repo1 contains a branch named main.

You create a new Azure Synapse workspace named Workspace1.

You need to create data processing pipelines in Workspace1. The solution must meet the following requirements:

- * Pipeline artifacts must be stored in Repo1.
- * Source control must be provided for pipeline artifacts.
- * All development must be performed in a feature branch.

which four actions should you perform in sequence in Synapse Studio? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

Select and Place:



Actions

- Set the main branch as the collaboration branch.
- Create pipeline artifacts and save them in the main branch.
- Configure a code repository and select **Repo1**.
- Create a new branch.
- Create pipeline artifacts and save them in the new branch.
- Create a pull request to merge the contents of the main branch into the new branch.

Answer Area

Correct Answer:

Actions

Set the main branch as the collaboration branch.
Create pipeline artifacts and save them in the main branch.

**Answer Area**

Configure a code repository and select Repo1 .
Create a new branch.
Create pipeline artifacts and save them in the new branch.
Create a pull request to merge the contents of the main branch into the new branch.

**Section:****Explanation:**

Configure a code repository and select Repo1.

Create a new branch.

Create pipeline artifacts and save them in the new branch.

Create a pull request to merge the contents of the main branch into the new branch.

QUESTION 71

You have an Azure Data Factory pipeline named pipeline1 that includes a Copy activity named Copy1. Copy1 has the following configurations:

* The source of Copy1 is a table in an on-premises Microsoft SQL Server instance that is accessed by using a linked service connected via a self-hosted integration runtime.

* The sink of Copy1 uses a table in an Azure SQL database that is accessed by using a linked service connected via an Azure integration runtime.

You need to maximize the amount of compute resources available to Copy1. The solution must minimize administrative effort.

What should you do?

- A. Scale up the data flow runtime of the Azure integration runtime.
- B. Scale up the data flow runtime of the Azure integration runtime and scale out the self-hosted integration runtime.
- C. Scale out the self-hosted integration runtime.

Correct Answer: A

Section:**QUESTION 72****HOTSPOT**

You have an Azure subscription that contains an Azure Cosmos DB analytical store and an Azure Synapse Analytics workspace named WS 1. WS1 has a serverless SQL pool name Pool1.

You execute the following query by using Pool1.

```

WITH IDENTITY = 'SHARED /
SECRET = 'fed4347479872423433563653456345ddfa==';

SELECT clientID AS ClientID,
       client AS ClientName
FROM OPENROWSET
(
    PROVIDER = 'CosmosDB',
    CONNECTION = 'Account=account1;Database=database1',
    OBJECT = 'clients',
    SERVER_CREDENTIAL = 'AccountCred'
)
WITH

```

```

(
    clientID int,
    client varchar(50),
    streetAddress varchar(100)
) AS c;

```



For each of the following statements, select Yes if the statement is true. Otherwise, select No.
 NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Statements	Yes	No
The query returns three columns.	<input type="radio"/>	<input type="radio"/>
The container being queried is named <code>clients</code> .	<input type="radio"/>	<input type="radio"/>
Authentication is performed by using an account key.	<input type="radio"/>	<input type="radio"/>

Answer Area:

Answer Area

Statements	Yes	No
The query returns three columns.	<input type="radio"/>	<input checked="" type="radio"/>
The container being queried is named <code>clients</code> .	<input checked="" type="radio"/>	<input type="radio"/>
Authentication is performed by using an account key.	<input checked="" type="radio"/>	<input type="radio"/>

Section:

Explanation:

QUESTION 73

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 Data Lake Storage account that contains a staging zone.

You need to design a daily process to ingest incremental data from the staging zone, transform the data by executing an R script and then insert the transformed data into a data warehouse in Azure Synapse Analytics.

Solution: You use an Azure Data Factory schedule trigger to execute a pipeline that executes a mapping data flow, and then inserts the data into the data warehouse.

Does this meet the goal?

A. Yes

B. NO

Correct Answer: A

Section:

QUESTION 74

HOTSPOT

You have a trigger in Azure Data Factory configured as shown in the following exhibit.

Name *

Description

Type *

Start date * ⓘ

Time zone * ⓘ

ⓘ This time zone observes daylight savings. Trigger will auto-adjust for one hour difference.

Recurrence * ⓘ
 Every

Specify an end date

End On * ⓘ

Annotations

Status ⓘ
 Started Stopped



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

Hot Area:

Answer Area

If the trigger was published on May 12, 2023, at 9:00 AM, the first execution will occur on [answer choice].

- May 12, 2023, at 8:15 PM
- May 12, 2023, at 9:00 AM
- May 12, 2023, at 9:15 AM
- May 12, 2023, at 8:15 PM
- May 12, 2025, at 5:30 PM

The last expected execution time of the pipeline will occur on [answer choice].

- May 12, 2025, at 5:30 PM
- May 12, 2025, at 8:15 PM
- May 12, 2025, at 5:30 PM
- May 11, 2025, at 5:15 PM

Answer Area:

Answer Area

If the trigger was published on May 12, 2023, at 9:00 AM, the first execution will occur on [answer choice].

- May 12, 2023, at 8:15 PM
- May 12, 2023, at 9:00 AM
- May 12, 2023, at 9:15 AM
- May 12, 2023, at 8:15 PM
- May 12, 2023, at 5:30 PM

The last expected execution time of the pipeline will occur on [answer choice].

- May 12, 2025, at 5:30 PM
- May 12, 2025, at 8:15 PM
- May 12, 2025, at 5:30 PM
- May 11, 2025, at 5:15 PM

Section:

Explanation:

QUESTION 75

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 during the last five minutes every five minutes. Each tweet must only be counted once.

Which windowing function should you use?

- A. a five-minute Tumbling window
- B. a five-minute Sliding window
- C. a five-minute Hopping window that has a one-minute hop
- D. a five-minute Session window

Correct Answer: A

Section:

QUESTION 76

DRAG DROP

You have an Azure subscription that contains an Azure data factory.

You are editing an Azure Data Factory activity JSON.

The script needs to copy a file from Azure Blob Storage to multiple destinations. The solution must ensure that the source and destination files have consistent folder paths.

How should you complete the 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

Select and Place:



Values

FlattenHierarchy

ForEach

MergeFiles

PreserveHierarchy

Switch

Until

Answer Area

```
{
  "name": "Pipeline1",
  "properties": {
    "activities": [
      {
        "name": "Activity1",
        "type": ,
        "typeProperties": {
          "isSequential": "true",
          "items": {
            "value": "@pipeline
() .parameters.mySinkDatasetFolderPath",
            "type": "Expression"},
          "activities" [
            {
              "name": "MyCopyActivity",
              "type": "Copy",
              "typeProperties": {
                "source": {
                  "type": "BlobSource",
                  "recursive": "false" },
                "sink": {
                  "type": "BlobSink",
                  "CopyBehavior": 
                }
              }
            }
          ]
        }
      }
    ]
  }
}
```

Correct Answer:

Values**Answer Area**

```

{
  "name": "Pipeline1",
  "properties": {
    "activities": [
      {
        "name": "Activity1",
        "type": "ForEach",
        "typeProperties": {
          "isSequential": "true",
          "items": {
            "value": "@pipeline
() .parameters.mySinkDatasetFolderPath",
            "type": "Expression"},
          "activities" [
            {
              "name": "MyCopyActivity",
              "type": "Copy",
              "typeProperties": {
                "source": {
                  "type": "BlobSource",
                  "recursive": "false" },
                "sink": {
                  "type": "BlobSink",
                  "CopyBehavior": "Switch"
                }
              }
            }
          ]
        }
      }
    ]
  }
}

```

Section:**Explanation:****QUESTION 77**

You have an Azure subscription that contains the resources shown in the following table.

Name	Type	Description
storage1	Azure Blob storage account	Contains publicly accessible TSV files that do NOT have a header row
WS1	Azure Synapse Analytics workspace	Contains a serverless SQL pool

You need to read the TSV files by using ad-hoc queries and the openrowset function. The solution must assign a name and override the inferred data type of each column. What should you include in the openrowset function?

- A. the with clause
- B. the rowsoptions bulk option
- C. the datafiletype bulk option
- D. the DATA_source parameter

Correct Answer: B

Section:

QUESTION 78

You have an Azure subscription that contains an Azure Synapse Analytics account and a Microsoft Purview account.

You create a pipeline named Pipeline1 for data ingestion to a dedicated SQL pool.

You need to generate data lineage from Pipeline1 to Microsoft Purview.

Which two activities generate data lineage? Each correct answer presents a complete solution.

NOTE: Each correct selection is worth one point.

- A. Web
- B. Copy
- C. WebHook
- D. Dataflow
- E. Validation

Correct Answer: A, D

Section:

