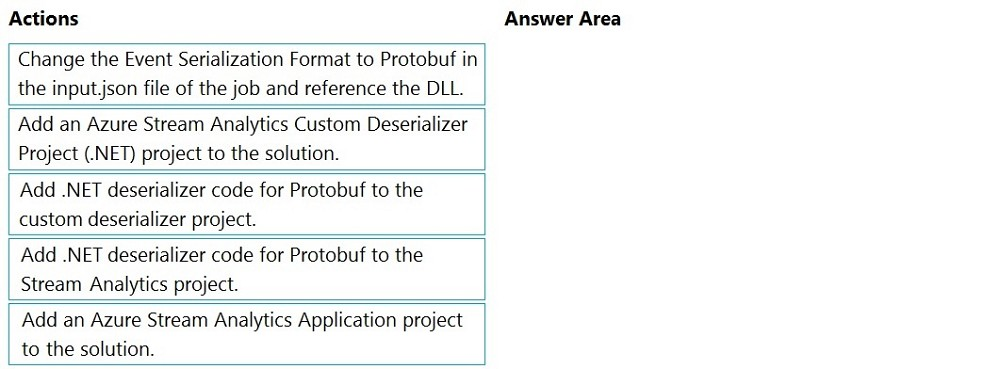


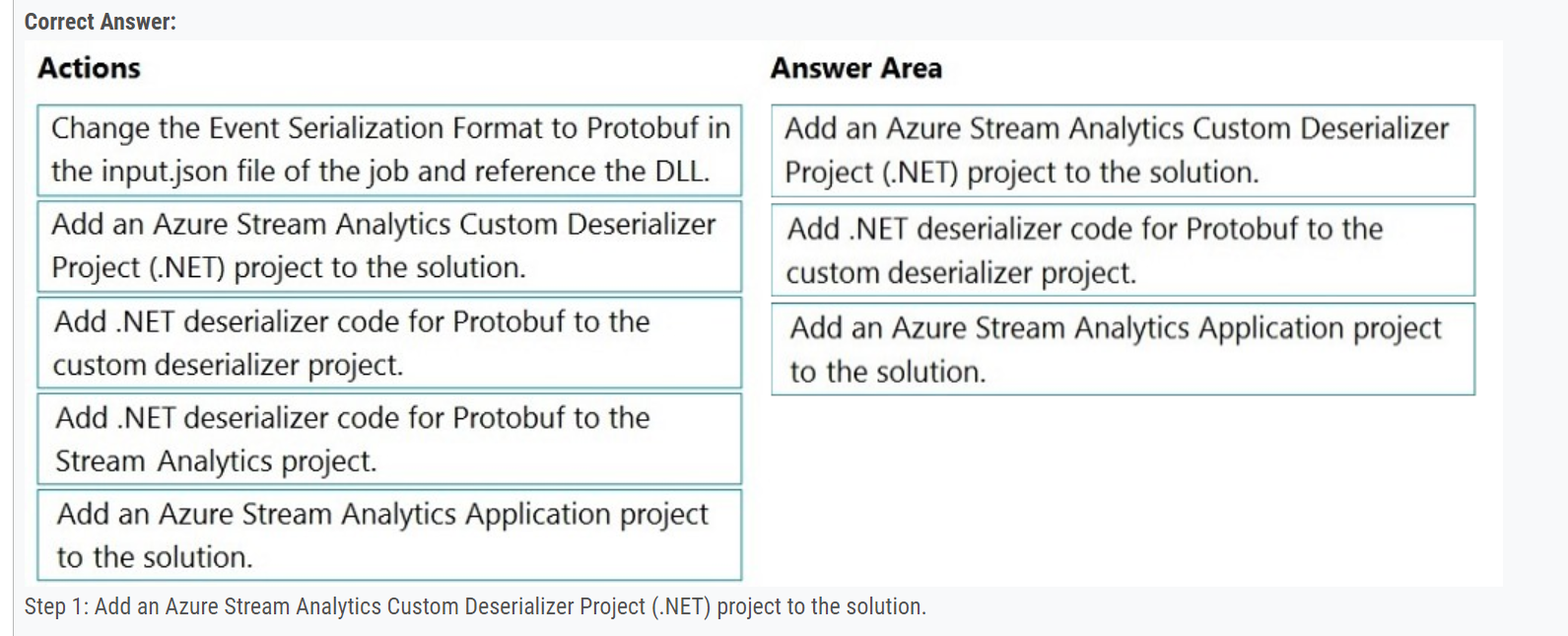
DRAG DROP -  
You have an Azure Stream Analytics job that is a Stream Analytics project solution in Microsoft Visual Studio. The job accepts data generated by IoT devices in the JSON format.  
You need to modify the job to accept data generated by the IoT devices in the Protobuf format.  
Which three 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:



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>

**Correct Answer:**



**CI/CD PROCESS:**

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.  
1. In Azure DevOps, open the project that's configured with your data factory.  
2. On the left side of the page, select Pipelines, and then select Releases.  
3. Select New pipeline, or, if you have existing pipelines, select New and then New release pipeline.  
4. In the Stage name box, enter the name of your environment.  
5. 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.  
6. Select the Empty job template.  
Reference:  
<https://docs.microsoft.com/en-us/azure/data-factory/continuous-integration-deployment>

SELECT System.Timestamp() AS WindowEnd, TollId, COUNT(\*)

FROM Input TIMESTAMP BY EntryTime

GROUP BY TollId, HoppingWindow(Duration(hour, 1), Hop(minute, 5), Offset(millisecond, -1))

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.

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

[Previous Questions](https://www.examtopics.com/exams/microsoft/dp-203/view/19/)[Next Questions](https://www.examtopics.com/exams/microsoft/dp-203/view/21/)

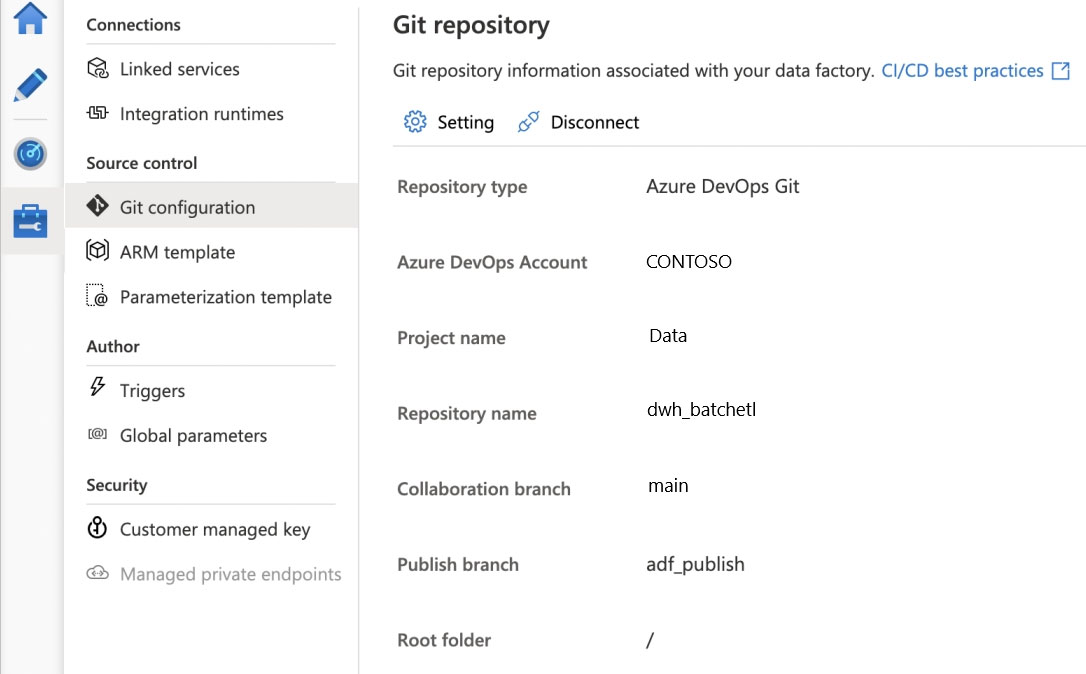
You are designing an Azure Stream Analytics solution that will analyze Twitter data.  
You need to count the tweets in each 10-second window. The solution must ensure that each tweet is counted only once.  
Solution: You use a tumbling window, and you set the window size to 10 seconds.  
Does this meet the goal?

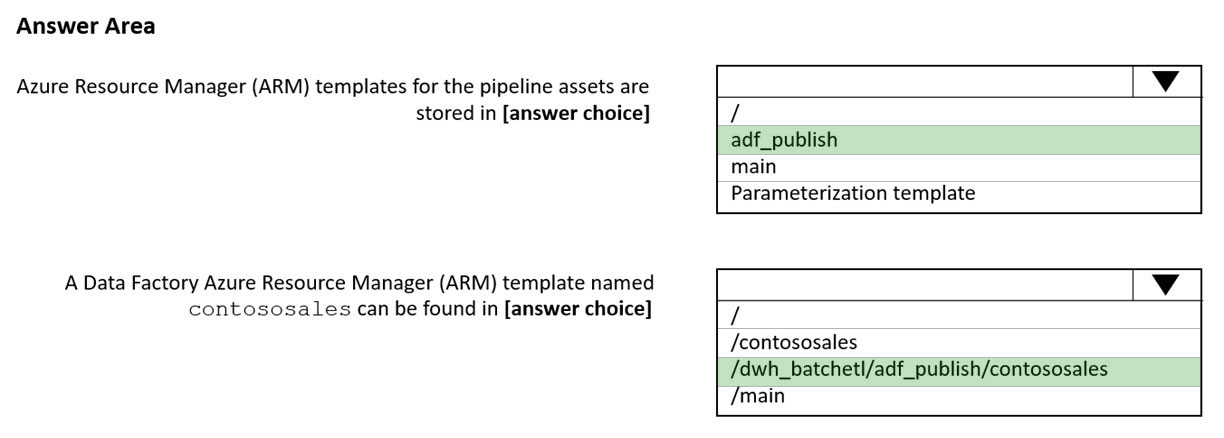
* A. Yes
* B. No

[Tumbling Window (Azure Stream Analytics) - Stream Analytics Query | Microsoft Learn](https://learn.microsoft.com/en-us/stream-analytics-query/tumbling-window-azure-stream-analytics)

[Session Window (Azure Stream Analytics) - Stream Analytics Query | Microsoft Learn](https://learn.microsoft.com/en-us/stream-analytics-query/session-window-azure-stream-analytics)

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>

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


**Correct Answer:** **  
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>

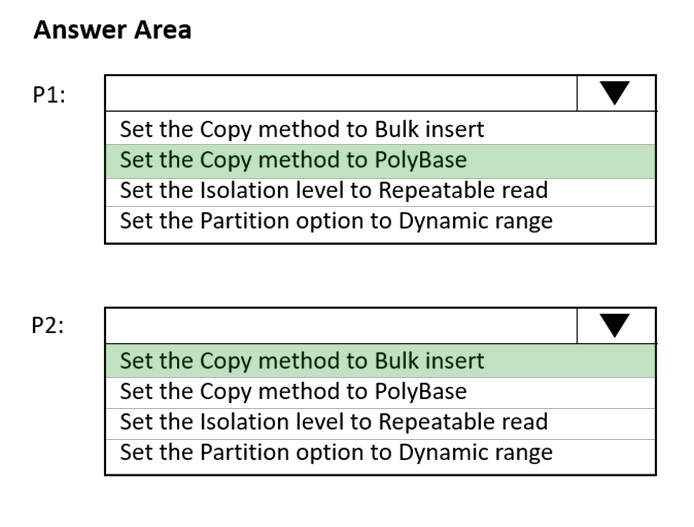
SELECT System.Timestamp() as WindowEndTime, TimeZone, COUNT(\*) AS Count

FROM TwitterStream TIMESTAMP BY CreatedAt

GROUP BY TimeZone, TumblingWindow(second,10)

QUESTION 42:

✑ 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

**rrect Answer:** **

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>

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

DWU used: DWU limit \* DWU percentage  
DWU used represents only a high-level representation of usage across the SQL pool and is not meant to be a comprehensive indicator of utilization. To determine whether to scale up or down, consider all factors which can be impacted by DWU such as concurrency, memory, tempdb, and adaptive cache capacity. We recommend running your workload at different DWU settings to determine what works best to meet your business objectives.  
Azure Synapse Analytics monitor metric "DWU used"  
Incorrect:  
\* CPU percentage. CPU utilization across all nodes for the data warehouse.  
\* DWU percentage: Maximum between CPU percentage and Data IO percentage  
\* Data IO percentage: IO Utilization across all nodes for the data warehouse

You have an Azure Synapse Analytics dedicated SQL pool named SA1 that contains a table named Table1.  
You need to identify tables that have a high percentage of deleted rows.  
What should you run?

* C. sys.pdw\_nodes\_column\_store\_row\_groups

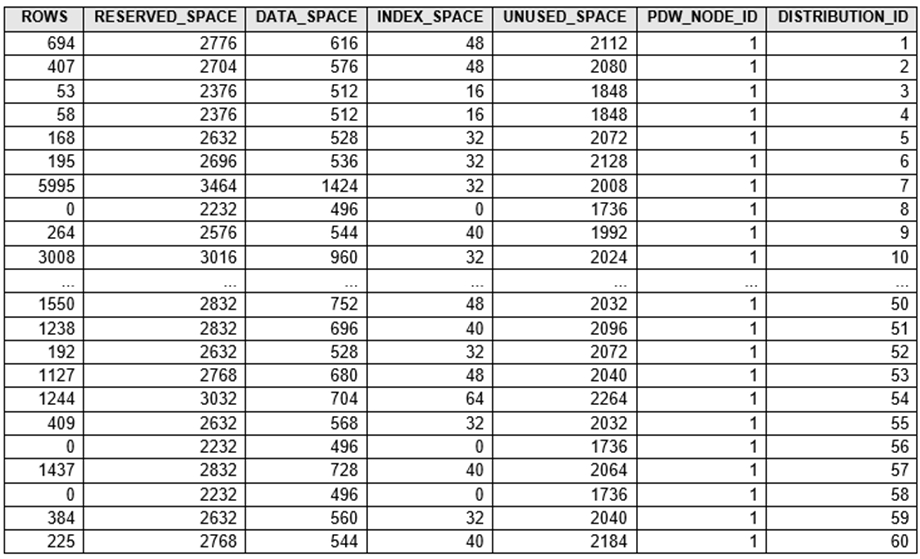
Use sys.pdw\_nodes\_column\_store\_row\_groups to determine which row groups have a high percentage of deleted rows and should be rebuilt.  
Note: sys.pdw\_nodes\_column\_store\_row\_groups provides clustered columnstore index information on a per-segment basis to help the administrator make system management decisions in Azure Synapse Analytics. sys.pdw\_nodes\_column\_store\_row\_groups has a column for the total number of rows physically stored  
(including those marked as deleted) and a column for the number of rows marked as deleted.  
Incorrect:  
Not A: You can join sys.pdw\_nodes\_column\_store\_segments with other system tables to determine the number of columnstore segments per logical table.  
Not B: Use sys.dm\_db\_column\_store\_row\_group\_operational\_stats to track the length of time a user query must wait to read or write to a compressed rowgroup or partition of a columnstore index, and identify rowgroups that are encountering significant I/O activity or hot spots.

**Question: When an activity in a Data Factory pipeline fails, does the entire pipeline fail?**

[Azure Data Factory Activity Failures and Pipeline Outcomes – Data Savvy](https://datasavvy.me/2021/02/18/azure-data-factory-activity-failures-and-pipeline-outcomes/)

Read this for Databricks audit logs:

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

You have an Azure Synapse Analytics dedicated SQL pool.  
You run PDW\_SHOWSPACEUSED('dbo.FactInternetSales'); and get the results shown in the following table.  


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. TWO Limit
* D. Cache hit percentage

[Hide Solution](https://www.examtopics.com/exams/microsoft/dp-203/view/41/) [Discussion   **3**](https://www.examtopics.com/exams/microsoft/dp-203/view/41/)

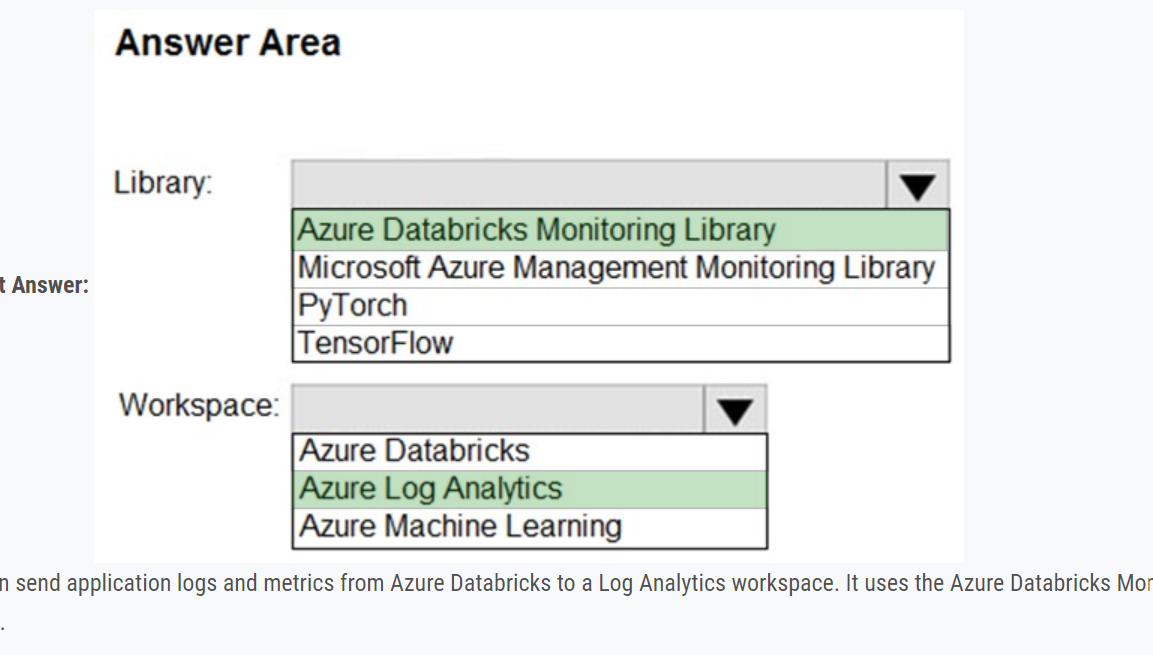
D: Cache hit percentage: (cache hits / cache miss) \* 100 where cache hits is the sum of all columnstore segments hits in the local SSD cache and cache miss is the columnstore segments misses in the local SSD cache summed across all nodes  
B: (cache used / cache capacity) \* 100 where cache used is 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 Asnwers:  
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:

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?

* C. Assign a larger resource class to the automated data load queries.

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>



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:

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?

ANSWER:

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>

What are streaming SU’s do?

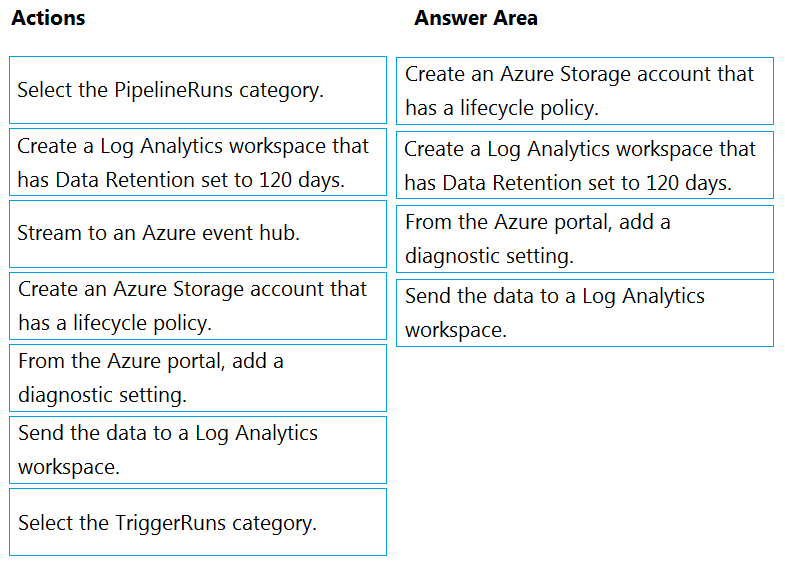
Backlogged Input Events: Number of input events that are backlogged. A non-zero value for this metric implies that your job isn't able to keep up with the number of incoming events. If this value is slowly increasing or consistently non-zero, you should scale out your job. You should increase the Streaming 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>

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 in not found.  
You need to identify the cause of the issue.  
What should you review?

* C. global init scripts logs

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>

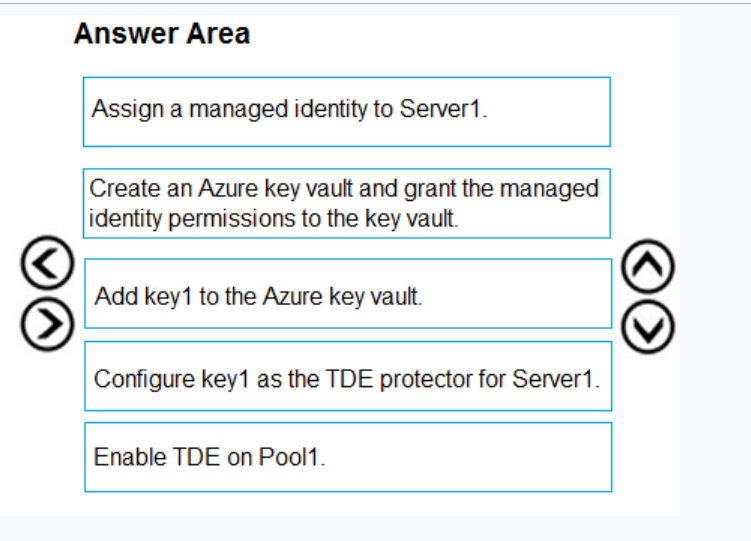
DRAG DROP -  
You have an Azure data factory.  
You need to ensure that pipeline-run data is retained for 120 days. The solution must ensure that you can query the data by using the Kusto query language.  
Which four actions should you perform in sequence? To answer, move the appropriate actions from 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:

**  
Step 1: Create an Azure Storage account that has a lifecycle policy  
To automate common data management tasks, Microsoft created a solution based on Azure Data Factory. The service, Data Lifecycle Management, makes frequently accessed data available and archives or purges other data according to retention policies. Teams across the company use the service to reduce storage costs, improve app performance, and comply with data retention policies.  
Step 2: Create a Log Analytics workspace that has Data Retention set to 120 days.  
Data Factory stores pipeline-run data for only 45 days. Use Azure Monitor if you want to keep that data for a longer time. With Monitor, you can route diagnostic logs for analysis to multiple different targets, such as a Storage Account: Save your diagnostic logs to a storage account for auditing or manual inspection. You can use the diagnostic settings to specify the retention time in days.  
Step 3: From Azure Portal, add a diagnostic setting.  
Step 4: Send the data to a log Analytics workspace,  
Event Hub: A pipeline that transfers events from services to Azure Data Explorer.  
Keeping Azure Data Factory metrics and pipeline-run data.  
Configure diagnostic settings and workspace.  
Create or add diagnostic settings for your data factory.  
1. In the portal, go to Monitor. Select Settings > Diagnostic settings.  
2. Select the data factory for which you want to set a diagnostic setting.  
3. If no settings exist on the selected data factory, you're prompted to create a setting. Select Turn on diagnostics.  
4. Give your setting a name, select Send to Log Analytics, and then select a workspace from Log Analytics Workspace.  
5. Select Save.  
Reference:  
https://docs.microsoft.com/en-us/azure/data-factory/monitor-using-azure-monitor

Box 2: CREATE SECURITY POLICY -  
Implement RLS by using the CREATE SECURITY POLICY Transact-SQL statement, and predicates created as inline table-valued functions.  
Reference:  
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/column-level-security <https://docs.microsoft.com/en-us/sql/relational-databases/security/row-level-security>

QUESTION

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



QUESTION:

You plan to create an Azure Synapse Analytics dedicated SQL pool.  
You need to minimize the time it takes to identify queries that return confidential information as defined by the company's data privacy regulations and the users who executed the queues.  
Which two components should you include in the solution? Each correct answer presents part of the solution.  
NOTE: Each correct selection is worth one point.

* A. sensitivity-classification labels applied to columns that contain confidential information
* B. resource tags for databases that contain confidential information
* C. audit logs sent to a Log Analytics workspace
* D. dynamic data masking for columns that contain confidential information