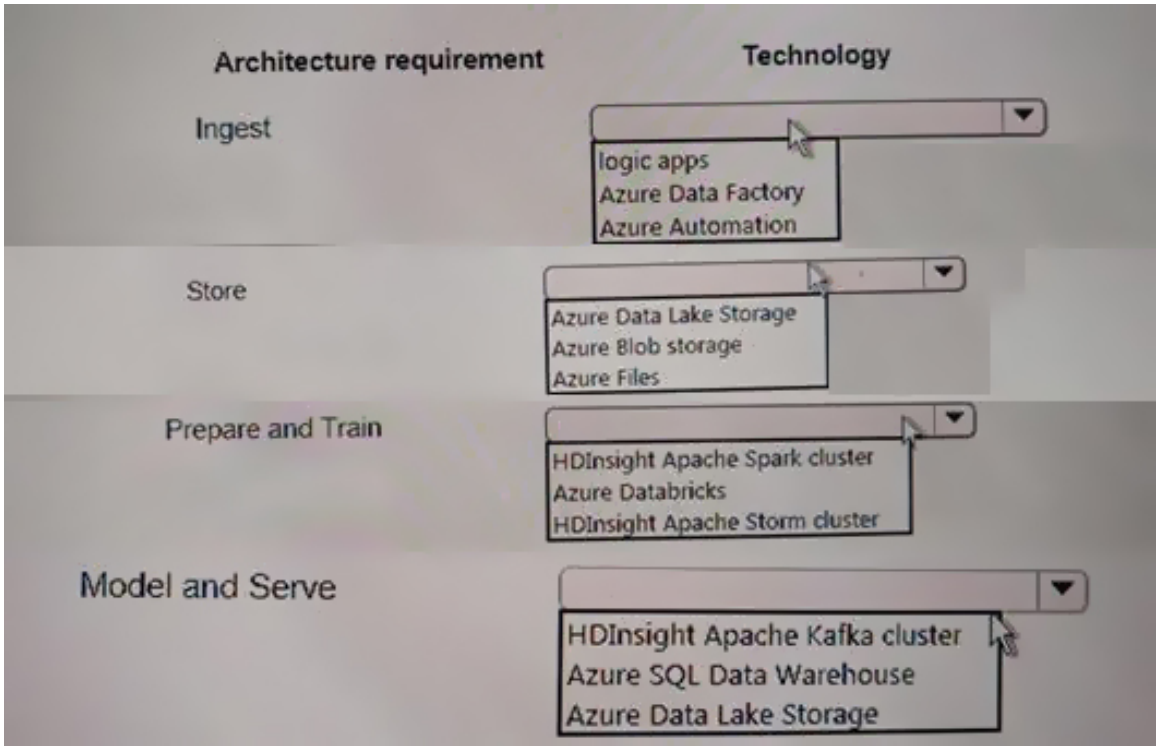
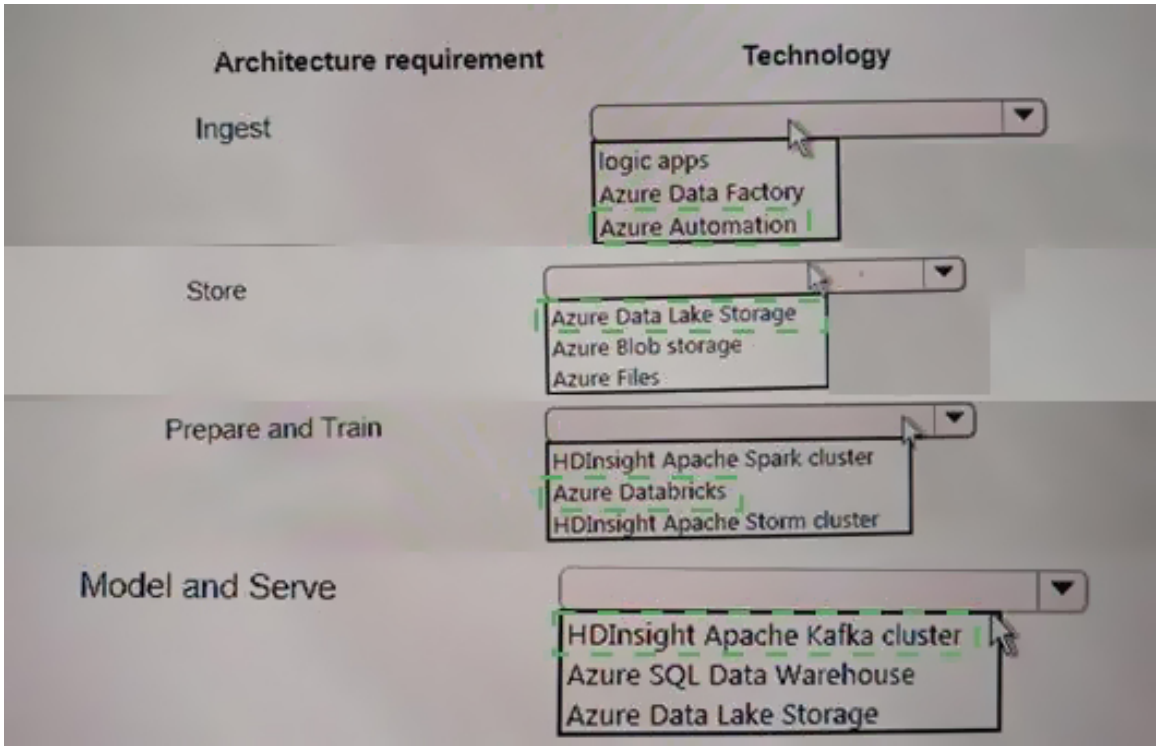
area.



Correct Answer:



QUESTION 58
You manage a process that performs analysis of daily web traffic logs on an HDInsight cluster. Each of 250 web servers generates approximately gigabytes (GB) of log data each day. All log data is stored in a single folder in Microsoft Azure Data Lake Storage Gen 2.

You need to improve the performance of the process.

Which two changes should you make? Each correct answer presents a complete solution.

NOTE: Each correct selection is worth one point.

A. Combine the daily log files for all servers into one file
B. Increase the value of the mapreduce.map.memory parameter
C. Move the log files into folders so that each day's logs are in their own folder
D. Increase the number of worker nodes
E. Increase the value of the hive.tez.container.size parameter

Correct Answer: AC
Explanation:
A: Typically, analytics engines such as HDInsight and Azure Data Lake Analytics have a per-file overhead. If you store your data as many small files, this can negatively affect performance. In general, organize your data into larger sized files for better performance (256MB to 100GB in size). Some engines and applications might have trouble efficiently processing files that are greater than 100GB in size.

C: For Hive workloads, partition pruning of time-series data can help some queries read only a subset of the data which improves performance.

Those pipelines that ingest time-series data, often place their files with a very structured naming for files and folders. Below is a very common example we see for data that is structured by date:

\DataSet\YYYY\MM\DD\datafile_YYYY_MM_DD.tsv

Notice that the datetime information appears both as folders and in the filename.

References:
https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-performance-tuning-guidance

QUESTION 59
You develop data engineering solutions for a company.

A project requires the deployment of data to Azure Data Lake Storage.

You need to implement role-based access control (RBAC) so that project members can manage the Azure Data Lake Storage resources.

Which three actions should you perform? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

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

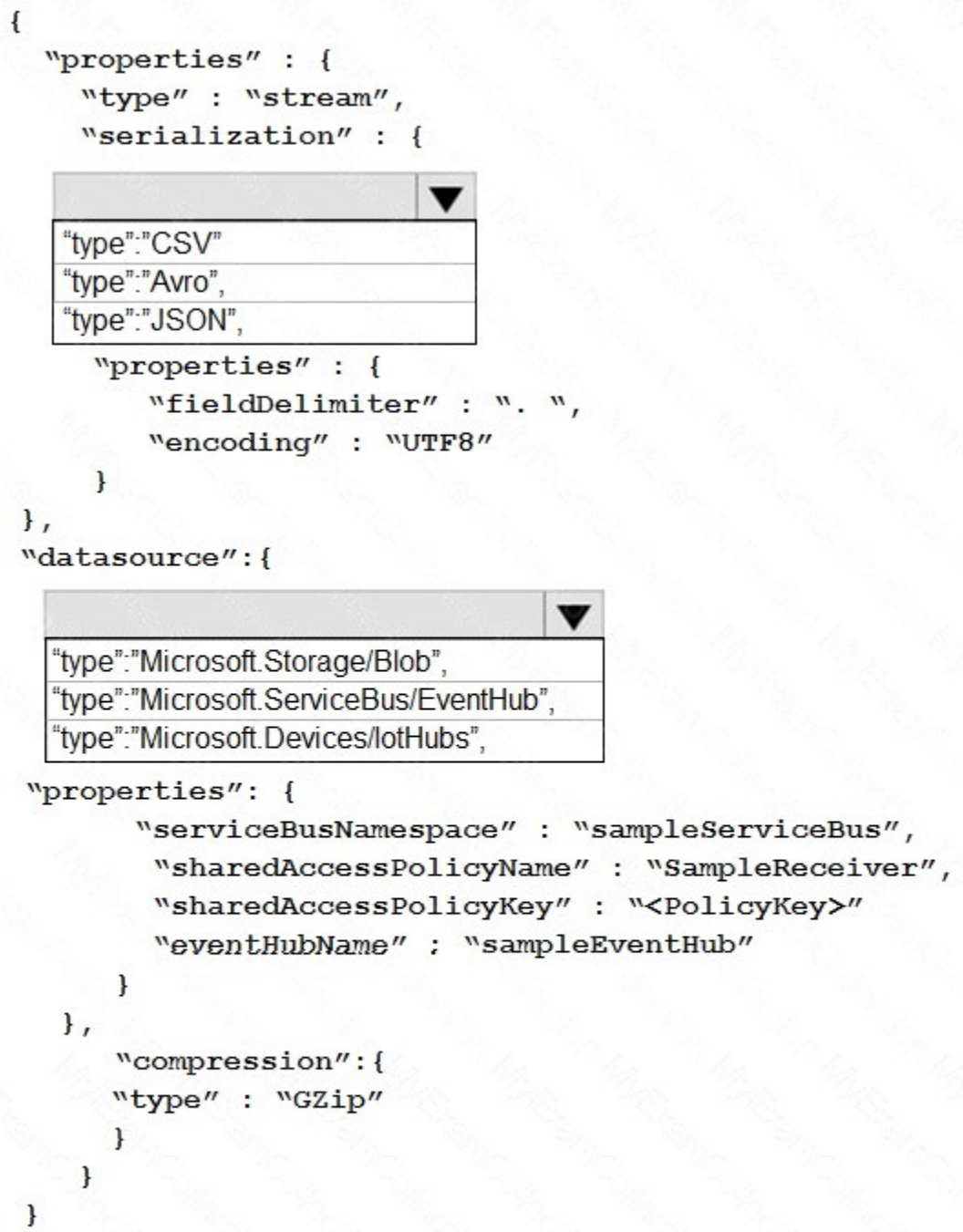
Correct Answer: ADE

QUESTION 60
HOTSPOT
A company plans to analyze a continuous flow of data from a social media platform by using Microsoft Azure Stream Analytics. The incoming data is formatted as one record per row.

You need to create the input stream.

How should you complete the REST API segment? To answer, select the appropriate configuration in the answer area.

NOTE: Each correct selection is worth one point.



Correct Answer:

```
{
  "properties" : {
    "type" : "stream",
    "serialization" : {
```

[▼]
- "type":"CSV"
- "type":"Avro",
- "type":"JSON",

```
      "properties" : {
         "fieldDelimiter" : ". ",
         "encoding" : "UTF8"
      }
},
"datasource":{
```

[▼]
- "type":"Microsoft.Storage/Blob",
- "type":"Microsoft.ServiceBus/EventHub",
- "type":"Microsoft.Devices/IotHubs",

```
"properties": {
      "serviceBusNamespace" : "sampleServiceBus",
       "sharedAccessPolicyName" : "SampleReceiver",
       "sharedAccessPolicyKey" : "<PolicyKey>"
       "eventHubName" : "sampleEventHub"
    }
  },
    "compression":{
    "type" : "GZip"
    }
  }
}
```