Business Context:

Apache Spark is a parallel processing framework that supports in-memory processing to boost the performance of big-data analytic applications. Apache Spark in Azure HDInsight is the Microsoft implementation of Apache Spark in the cloud

In this exercise, you deploy a Jupyter Notebook on a Spark HDI cluster and import a dataset into it

Learning Outcomes:

1) Deploy Spark on an HDI cluster
2) Run a Jupyter notebook in an HDI cluster
3) Import a dataset into the Jupyter notebook

Problem statement

1) Create an Apache Spark Cluster in HDInsight in the East US region
2) Create a Jupyter Notebook in the Spark cluster
3) Import the following dataset into the Jupyter notebook using Spark SQL https://github.com/MicrosoftLearning/20775_Performing-Data-Engineering-on-Microsoft-HDInsight/blob/master/Allfiles/Demofiles/Mod04/clidata/hvac/HVAC.csv
4) Write a query to display the values **tempdiff=targettemp-actualtemp** and

   date= 1st June 2016
5) Display the above values as a line graph

Note:

- Other required values can be set at your discretion.
- Submission of this assessment shall be done in the form of a pdf document containing the labeled screenshots as outlined in the marks distribution section.

Marks Distribution:

1) Screenshot of creation of Spark Cluster	15 marks
2) Screenshot of creation of Jupyter Notebook	10 marks
3) Screenshot of import of dataset	5 marks
4) Screenshot of result of given query	15 marks
5) Screenshot of line graph	5 marks