Workshop #3: Using Spark

In this module of the class, you are going to deploy an HDInsight Spark Cluster and run a word counter in it.

1 EXPECTED OUTCOME

The student is going to:

- Learn about the basics on deploying Spark in Azure
- Run applications on top of Spark
- Compare the differences between Spark and MapReduce

2 ASSUMPTIONS

This workshop assumes that the student had some familiarity with the Azure portal and MapReduce.

3 APACHE SPARK

Apache Spark is an open-source processing framework that runs large-scale data applications. It can run tasks similar to the well known MapReduce, with the difference that it tries to maintain the dataset on memory as much as possible. There are multiple applications that can run on top of the Spark Core Engine, such as Spark SQL, Spark Streaming, MLib (Machine Learning) and GraphX (Graph Computation).

3.1 Deploying the basic service

Follow the tutorial *Create Apache Spark cluster on HDInsight Linux* and additionally plot the values for MAX and MIN for the temperature difference.

3.2 Word Count

Using the PySpark notebook and Spark SQL, implement a word counter, similar to the one implemented in the MapReduce assignment. Discuss with your fellow students about which dataset use for counting words. Examples of interesting data sets is the completes work of Shakespeare, or any other author that you find interesting.

3.3 Spark vs Map Reduce

Write a summary of the main differences between Map Reduce and Spark.

4 Additional Notes (READ)

After running your assignment, save the python notebooks used to execute the commands and remove the HDInsight cluster, following this *tutorial*. This system are charged by the time that is allocated, which means after finishing you need to make sure that the cluster is deleted, to avoid excessive charges.

5 Deliverables

- The example cluster running the max and min computation.
- The Word Counter application.
- Explain the main differences between MapReduce and Spark.
- Discuss with your fellow students about the difference of Map Reduce and Spark, there are many online web pages that discuss the details of both, share those with other students.
- Discuss with your fellow students about interesting applications of Spark, or ways you have used it in the past.