**Week 8 Assignment 1 - Analyze data using Apache Spark**

*Please click on the link above to submit this week's assignment.*

*NOTE: Submit either this Paper or the Hands-on assignment.*

***Paper option  :  Write a paper on the following topics.***

1. Provide a brief history of Spark?
2. How is Spark better than MapReduce?
3. What is a Spark RDD?
4. What is the meaning of a "lazy evaluation" and what are its benefits?
5. What are transformations and actions? Describe the following actions/transformations:  map, filter, flatMap, count, collect and take.
6. The state of Spark in the Cloud.*( Sample reference is:   https://databricks.com/session/state-apache-spark-cloud )*
7. What is Spark streaming and where can it be used?*( Here's a useful site:   https://databricks.com/blog/2016/07/28/structured-streaming-in-apache-spark.html )*
8. What is Spark SQL, how to use it and where to use it ?
9. How can you use R with spark?*(https://databricks.com/blog/2017/05/25/using-sparklyr-databricks.html )*
10. How does Apache Spark compare to Apache Flink?

*======================================================================================*

***Hands-on assignment***

In this assignment you will be using Apache Spark to perform data analysis.

* You may use Scala or Python to run the Spark queries.
* You must include screenshots of the queries and the results.
* You may use Azure HDInsight or a local installation of  Bitnami Hadoop or Cloudera Hadoop distribution for this assignment
* When creating HDInsight cluster, choose "Spark" for cluster type

(1) Connect to the cluster

(2) Part 1:

* 1. Put the sample data from Week 5 ( counting of the word Sentence) into a text file
  2. Upload the file into HDFS
  3. Use Spark CLI (spark-shell or pyspark) or the Jupyter/Zeppelin notebook for your commands/queries
  4. Run the Spark transformations and actions (for example, filter, map, reduce etc..)  to count the number of times the word "Sentence" appears in the file

(4) Part 2:

* 1. Using the hdfs commands, upload the Baseball data file(s) into an HDFS folder such as /temp ( or /tmp)  and/or create a Hive table using the baseball data
  2. Use the Spark command line interface (CLI) or the Jupyter or Zeppelin Notebook to answer the following questions:
     1. What is the total number of baseball players?
     2. How many players were born before 1960?
     3. How many players were born in or after   1960?
     4. How many players were born outside of the USA?
     5. How many players were born in the USA?
  3. You may use either the Spark actions/transformations or the SparkSQL API or both.

(5) Capture the screen shots of the command execution and the results

(6) Provide a write up of the commands and the results.

**IMPORTANT**:  Don't forget to shutdown the Azure cluster