## Week 5 Assignment: Diving into SparkSQL with Scala, Python, and R

### Objective: Gain experience in querying datasets using SparkSQL across multiple languages - Scala, Python, and R.

#### **1. Environment Initialization**

* Navigate to the required directory and start your Docker containers:
* cd bellevue-bigdata  
  cd hadoop-hive-spark-hbase  
  docker-compose up -d
* Access the master container:
* docker-compose exec master bash
* Load the grades.csv into HDFS:

hdfs dfs -mkdir /data

* hdfs dfs -put /data/grades.csv /data/grades.csv

#### **2. SparkSQL with Scala**

* Enter the Spark shell:
* spark-shell
* Run the following SparkSQL commands in Scala:
* val df = spark.read.format("csv").option("header", "true").load("/data/grades.csv")  
  df.createOrReplaceTempView("df")  
    
  spark.sql("SHOW TABLES").show()  
  spark.sql("SELECT \* FROM df WHERE Final > 50").show()  
  spark.sql("SELECT \* FROM grades").show()
* Run 3 other SQL queries in the Spark Shell:
* Exit the Spark shell:

:quit

**Deliverable:**

* Screenshot of the results obtained from the SparkSQL commands in Scala.
* Screenshot of your 3 other SQL query results.

#### **3. SparkSQL with Python (PySpark)**

* Enter the PySpark environment:  
    
  pyspark
* Run the following SparkSQL commands in Python:
* df = spark.read.format('csv').option('header', 'true').load('/data/grades.csv')  
  df.show()  
    
  df.createOrReplaceTempView('df')  
  spark.sql('SHOW TABLES').show()  
  spark.sql('SELECT \* FROM df WHERE Final > 50').show()  
  spark.sql('SELECT \* FROM grades').show()
* Run 3 other SQL queries in the PySpark Shell:
* Exit the Spark shell:

exit()

**Deliverable:**

* Screenshot of the results obtained from the SparkSQL commands in Python.
* Screenshot of your 3 other SQL query results

#### **3. SparkSQL with Custom Data Set**

1. **Data Loading into Spark**: Use Spark to load the dataset from Assignment 3. You might find methods like spark.read.csv or spark.read.text useful, depending on the dataset format.
2. **SQL Queries**: Once you’ve loaded the data into Spark, please run three SQL queries on this dataset. Remember to first create a temporary view of your data in Spark using createOrReplaceTempView (for Scala) or a similar method in PySpark, so you can query it using SparkSQL.
3. **Language Selection**: You have the flexibility to use either Scala or PySpark for this exercise. Please choose whichever you’re more comfortable with.

**Deliverable:**

* Your Scala or PySpark Code.
* Screenshot of your 3 SQL query results.

## Shutting Down

Ensure all Docker containers are turned off with docker-compose down for each directory. If you’re using google cloud, please shut down your virtual machine to preserve cloud costs.