## 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:
* hadoop fs -put /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("/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()

**Deliverable:** Screenshot of the results obtained from the SparkSQL commands in Scala.

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

* Exit the Spark shell and enter the PySpark environment:
* exit()  
  pyspark
* Run the following SparkSQL commands in Python:
* df = spark.read.format('csv').option('header', 'true').load('/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()

**Deliverable:** Screenshot of the results obtained from the SparkSQL commands in Python.

#### **4. SparkSQL with R (SparkR)**

* Exit the PySpark environment and enter SparkR:
* exit()  
  sparkR
* Run the following commands in SparkR:
* df <- as.DataFrame(list("One", "Two", "Three", "Four"), "This is as example")  
  head(df)  
    
  df <- read.df("/grades.csv", "csv", header="true")  
  head(df)

**Deliverable:** Screenshot of the results obtained from the SparkR commands.

## 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.