In this assignment, you will practice performing data storage in Hadoop using data that contains information about product sales transactions. In the first part of the assignment, you will be challenged to build a Hadoop Docker *container* and ingest data into Hadoop data storage. In the second part of the assignment, you will explain what each Java program does and then apply the MapReduce framework to an input file to calculate aggregate sales by country.

Prior to beginning this activity, review the submission instructions below to ensure that you collect the required screenshots as you progress through the activity.

**Reference**

[Taylor, David. “Hadoop & MapReduce Examples: Create First Program in Java.” *Guru99*. 2021. https://www.guru99.com/create-your-first-hadoop-program.html.](https://www.guru99.com/create-your-first-hadoop-program.html)

**To complete this assignment, follow these steps:**

Before you begin the steps of the activity below, be sure that you have the Hadoop *containers* that you created in [Activity 18.1](https://classroom.emeritus.org/courses/10605/assignments/246290) running within Docker.

**Part 1: Ingesting Data into the HDFS**

For the first part of this assignment, you will ingest the SalesData.csv file into Hadoop data storage.

1. Download and extract the [testprogram.zip](https://classroom.emeritus.org/courses/10605/files/3007175/download) file on your machine. Open the command line interface on your machine and navigate to the extracted folder. Provide a screenshot to show that you unzipped the testprogram.zip file.
2. Change the *directory* to the parent of the testprogram and perform the Docker copy command to copy the testprogram folder into the home *directory* of the Hadoop namenode.  
   docker cp testprogram/ namenode:/home  
   Provide a screenshot to show that you copied the testprogram folder into the home *directory* of the Hadoop namenode.
3. Open the namenode CLI in Docker. Inside the home *directory*, use the HDFS command below to create an input folder called inputMapReduce.  
   hdfs dfs -mkdir /inputMapReduce  
   Navigate to the home *directory* of the testprogram folder. Use the command below to perform an HDFS copy of the SalesData.csv file into the Hadoop data storage.  
   hdfs dfs -copyFromLocal SalesData.csv /inputMapReduce  
   Provide a screenshot to show that you successfully copied the SalesData.csv file into the inputMapReduce folder.
4. Review the copied file using the HDFS cat command.  
   hdfs dfs -cat /inputMapReduce/SalesData.csv  
   Provide a screenshot to show that the file has been copied using the HDFS cat command.

**Part 2: Performing MapReduce: Aggregation Sales by Country**

For the second part of this assignment, you will use the code files under the testprogram folder on the namenode using the Docker CLI.

1. Navigate and review the source code in the testprogram folder and describe the *functions* within each file listed below:
   * SalesCountryDriver.java
   * SalesMapper.java
   * SalesCountryReducer.java
2. In your submission file, provide a description of each of the three Java files listed above that are used to apply the MapReduce framework on the Hadoop database.
3. In the namenode CLI, run the commands below to set up an executable path in your system environment:  
   export JAVA\_HOME=/usr/lib/jvm/java-8-openjdk-amd64/jre/  
   export CLASSPATH="$HADOOP\_HOME/share/hadoop/mapreduce/hadoop-mapreduce-client-core-3.2.1.jar:$HADOOP\_HOME/share/hadoop/mapreduce/hadoop-mapreduce-client-common-3.2.1.jar:$HADOOP\_HOME/share/hadoop/common/hadoop-common-3.2.1.jar:~/testprogram/SalesCountry/\*:$HADOOP\_HOME/lib/\*"  
   export HDFS\_NAMENODE\_USER="root"  
   export HDFS\_DATANODE\_USER="root"  
   export HDFS\_SECONDARYNAMENODE\_USER="root"  
   export YARN\_RESOURCEMANAGER\_USER="root"  
   export YARN\_NODEMANAGER\_USER="root"  
   Once you have completed this, run the command below to confirm that you have correctly defined the environment variables.  
   $echo JAVA\_HOME  
   Provide a screenshot to show that you have successfully defined the environment variables in the namenode CLI.
4. Navigate to the /home/testprogram folder in the namenode CLI and compile a Java program using the command below. This will create a SalesCountry folder with *class* files within it.  
   javac -d . SalesMapper.java SalesCountryReducer.java SalesCountryDriver.java  
   Provide a screenshot to show that you successfully compiled the Java files in the SalesCountry folder.
5. From the testprogram folder, create a jar file from the Java code compiled in the previous step. To achieve this, run the command below:  
   jar cfm ProductSalePerCountry.jar Manifest.txt SalesCountry/\*.class  
   Provide a screenshot to show that you successfully created a jar file from the compiled Java code.
6. Perform a MapReduce operation using the command below:  
   $HADOOP\_HOME/bin/hadoop jar ProductSalePerCountry.jar /inputMapReduce /mapreduce\_output\_sales  
   Provide a screenshot to show that you successfully ran the MapReduce operation to distribute the analysis of the data.
7. Review the SalesCountry data output in the part-00000 file by running the command below:  
   $HADOOP\_HOME/bin/hdfs dfs -cat /mapreduce\_output\_sales/part-00000  
   Provide a screenshot to show that you successfully visualized the content of the part-00000 file inside the mapreduce\_output\_sales folder.

**Submission Instructions:**

Your submission for this assignment should be a Word document that includes the following screenshots, each labeled for the step that the screenshot represents:

**Part 1: Ingesting Data into the HDFS**

1. Provide a screenshot to show that you unzipped the testprogram.zip file.
2. Provide a screenshot to show that you copied the testprogram folder into the home *directory* of the Hadoop namenode.
3. Provide a screenshot to show that you successfully copied the SalesData.csv file into the inputMapReduce folder.
4. Provide a screenshot to show that the file has been copied using the HDFS cat command.

**Part 2: Performing MapReduce — Aggregation Sales by Country**

1. Provide a description of the three Java files listed below that are used to apply the MapReduce framework on the Hadoop database:
   * SalesCountryDriver.java
   * SalesMapper.java
   * SalesCountryReducer.java
2. Provide a screenshot to show that you have successfully defined the environment variables in the namenode CLI.
3. Provide a screenshot to show that you successfully compiled the Java files in the SalesCountry folder.
4. Provide a screenshot to show that you successfully created a jar file from the compiled Java code.
5. Provide a screenshot to show that you successfully ran the MapReduce operation to distribute the analysis of the data.
6. Provide a screenshot to show that you successfully visualized the content of the part-00000 file inside the mapreduce\_output\_sales folder.