Some setups

**Assume the transactions.csv dataset is presented in the Hadoop HDFS, located at:**user/Project1/data/transactions.csv

**Assume the Project3 directory is in the shared\_folder that is shared between the container and local machine**

**Assume the users are already inside the container and have Hadoop, Spark running**

**Command to create the project directory in Hadoop:**

hadoop/bin/hdfs dfs -mkdir /user/Project3/

hadoop/bin/hdfs dfs -mkdir /user/Project3/data/

**Set Hadoop\_ClassPath by running the following command:**  
export HADOOP\_CLASSPATH=/usr/lib/jvm/java-8-openjdk-amd64/lib/tools.jar

|  |  |  |
| --- | --- | --- |
| Question | Status (Select one) Fully Working/ Partially Working/ Not Working | Comment |
| Q1 | Fully Working | Further Details are below |
| Q2 | Fully Working | Further Details are below |

**Problem 1:**

**Assume the transactions.csv dataset is presented in the Hadoop HDFS, located at:**user/Project1/data/transactions.csv

**Command to run the pyspark script:**

spark-submit shared\_folder/Project3/problem1.py

**Problem 2:**

How cellID is calculated: cell\_id = row \* 500 + col

**Problem 2-1:**

**Go to Q2\_1 directory**

cd /home/ds503/shared\_folder/Project3/Q2\_1

**Compile the code**

javac Q2\_1.java

**Run the java file (to generated dataset in “data” folder)**

java Q2\_1

**Move points.csv to Hadoop**

cd ~

hadoop/bin/hdfs dfs -put shared\_folder/Project3/Q2\_1/data/\*.csv /user/Project3/data/

**Problem 2 (2.2 + 2.3):**

**Command to run the pyspark script:**

spark-submit shared\_folder/Project3/problem2.py