Some setups

Assume the Project1 directory is in the shared_folder that Hadoop can access from local machine:
shared_folder/Project1

Assume the users are already inside the container

Command to create the project directory in Hadoop:

```
hadoop/bin/hdfs dfs -mkdir /user/Project1/
```

```
hadoop/bin/hdfs dfs -mkdir /user/Project1/data/
```

Create the output folder for all outputs in Hadoop

```
hadoop/bin/hdfs dfs -mkdir /user/Project1/output
```

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
Q3.1	Fully Working	Further Details are below
Q3.2	Fully Working	Further Details are below
Q3.3	Fully Working	Further Details are below
Q4.1	Fully Working	Further Details are below
Q4.2	Fully Working	Further Details are below
Q4.3	Fully Working	Further Details are below

Q1

Go to Q1 directory

```
cd /home/ds503/shared_folder/Project1/Q1
```

Compile the code

```
javac Q1.java
```

Run the java file (to generated two datasets in “data” folder)

```
java Q1
```

Q2

Move customers.csv and transactions.csv to Hadoop

```
hadoop/bin/hdfs dfs -put shared_folder/Project1/data/*.csv /user/Project1/data/
```

Remove Hadoop data content *(if needed for new datasets)*

```
hdfs dfs -rm -r /user/Project1/data
```

Query 3.1

Go to Query_3_1 directory

```
cd /home/ds503/shared_folder/Project1/Query_3_1
```

Compile the code

```
hadoop com.sun.tools.javac.Main Query_3_1.java
```

Generate jar file (called query_3_1.jar)

```
jar cf query_3_1.jar Query_3_1*.class
```

Run Hadoop Job

```
hadoop jar ./query_3_1.jar Query_3_1 /user/Project1/data /user/Project1/output/query_3_1
```

Remove the output file *(if needed for a rerun)*

```
hdfs dfs -rm -r /user/Project1/output/query_3_1
```

View the output temporarily

```
cd ~
```

```
hadoop/bin/hdfs dfs -cat /user/Project1/output/query_3_1/part-r-00000
```

Query 3.2

Go to Query_3_2 directory

```
cd /home/ds503/shared_folder/Project1/Query_3_2
```

Compile the code

```
hadoop com.sun.tools.javac.Main Query_3_2.java
```

Generate jar file (called query_3_2.jar)

```
jar cf query_3_2.jar Query_3_2*.class
```

Run Hadoop Job

```
hadoop jar ./query_3_2.jar Query_3_2 /user/Project1/data/customers.csv /user/Project1/data /user/Project1/output/query_3_2
```

Remove the output file *(if needed for a rerun)*

```
hdfs dfs -rm -r /user/Project1/output/query_3_2
```

View the output temporarily

```
cd ~
```

```
hadoop/bin/hdfs dfs -cat /user/Project1/output/query_3_2/part-r-00000
```

Query 3.3

Go to Query_3_3 directory

```
cd /home/ds503/shared_folder/Project1/Query_3_3
```

Compile the code

```
hadoop com.sun.tools.javac.Main Query_3_3.java
```

Generate jar file (called query_3_3.jar)

```
jar cf query_3_3.jar Query_3_3*.class
```

Run Hadoop Job

```
hadoop jar ./query_3_3.jar Query_3_3 /user/Project1/data/customers.csv /user/Project1/data/transactions.csv /user/Project1/output/query_3_3
```

Remove the output file *(if needed for a rerun)*

```
hdfs dfs -rm -r /user/Project1/output/query_3_3
```

View the output temporarily

```
cd ~
```

```
hadoop/bin/hdfs dfs -cat /user/Project1/output/query_3_3/part-r-00000
```

4.1

```
cd ~
```

Move the pig script to hadoop

```
hadoop/bin/hdfs dfs -put shared_folder/Project1/query_4_1.pig /user/Project1
```

Remove the output *(if needed for a rerun)*

```
hdfs dfs -rm -r /user/Project1/output/query_4_1
```

Run the pig script

```
pig -x mapreduce hdfs:/user/Project1/query_4_1.pig
```

View the output temporarily

```
hadoop/bin/hdfs dfs -cat /user/Project1/output/query_4_1/part-r-00000
```

4.2

```
cd ~
```

Move the pig script to hadoop

```
hadoop/bin/hdfs dfs -put shared_folder/Project1/query_4_2.pig /user/Project1
```

Remove the output *(if needed for a rerun)*

```
hdfs dfs -rm -r /user/Project1/output/query_4_2
```

Run the pig script

```
pig -x mapreduce hdfs:/user/Project1/query_4_2.pig
```

View the output temporarily

```
hadoop/bin/hdfs dfs -cat /user/Project1/output/query_4_2/part-r-00000
```

4.3

```
cd ~
```

Move the pig script to hadoop

```
hadoop/bin/hdfs dfs -put shared_folder/Project1/query_4_3.pig /user/Project1
```

Remove the output *(if needed for a rerun)*

```
hdfs dfs -rm -r /user/Project1/output/query_4_3
```

Run the pig script

```
pig -x mapreduce hdfs:/user/Project1/query_4_3.pig
```

View the output temporarily

```
hadoop/bin/hdfs dfs -cat /user/Project1/output/query_4_3/part-r-00000
```