

## Some setups

**Assume the Project2 directory is in the shared\_folder that Hadoop can access from local machine:**

shared\_folder/Project2

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

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

```
hadoop/bin/hdfs dfs -mkdir /user/Project2/
```

```
hadoop/bin/hdfs dfs -mkdir /user/Project2/data/
```

**Create the output folder for all outputs in Hadoop**

```
hadoop/bin/hdfs dfs -mkdir /user/Project2/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	Fully Working	Further Details are below

Q1\_1 – Create dataset

**Go to Q1\_1 directory**

```
cd /home/ds503/shared_folder/Project2/Q1_1
```

**Compile the code**

```
javac Q1_1.java
```

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

```
java Q1_1
```

**Move points.csv and rectangles.csv to Hadoop**

```
cd ~
```

```
hadoop/bin/hdfs dfs -put shared_folder/Project2/Q1_1/data/*.csv /user/Project2/data/
```

## Query 1\_2

### Go to Q1\_2 directory

```
cd /home/ds503/shared_folder/Project2/Q1_2
```

### Compile the code

```
hadoop com.sun.tools.javac.Main Q1_2.java
```

### Generate jar file (called q1\_2.jar)

```
jar cf q1_2.jar Q1_2*.class
```

### Run Hadoop Job

```
hadoop jar ./q1_2.jar Q1_2 /user/Project2/data /user/Project2/output/q1_2 1 3 3 20
```

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

```
hdfs dfs -rm -r /user/Project2/output/q1_2
```

### View the output temporarily

```
cd ~
```

```
hadoop/bin/hdfs dfs -cat /user/Project2/output/q1_2/part-r-00000
```

## Query 2\_2

### Go to Q2\_2 directory

```
cd /home/ds503/shared_folder/Project2/Q2_2
```

### Compile the code

```
hadoop com.sun.tools.javac.Main Q2_2.java
```

### Generate jar file (called q2\_2.jar)

```
jar cf q2_2.jar Q2_2*.class
```

### Run Hadoop Job <points-path> <output-path> <r> <k>

```
hadoop jar ./q2_2.jar Q2_2 /user/Project2/data/points.csv /user/Project2/output/q2_2 20 5
```

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

```
hdfs dfs -rm -r /user/Project2/output/q2_2
```

### View the output temporarily

```
cd ~
```

```
hadoop/bin/hdfs dfs -cat /user/Project2/output/q2_2/part-r-00000
```

Q3\_1 – Create initial k points file

**Go to Q3\_1 directory**

```
cd /home/ds503/shared_folder/Project2/Q3_1
```

**Compile the code**

```
javac Q3_1.java
```

**Run the java file (to generate the k points file in “data” folder)**

```
java Q3_1 {number of k points}
```

**Move k-points.csv to Hadoop**

```
cd ~
```

```
hadoop/bin/hdfs dfs -put shared_folder/Project2/Q3_1/data/*.csv /user/Project2/data/
```

Query 3\_2

**Go to Q3\_2 directory**

```
cd /home/ds503/shared_folder/Project2/Q3_2
```

**Compile the code**

```
hadoop com.sun.tools.javac.Main Q3_2.java
```

**Generate jar file (called q3\_2.jar)**

```
jar cf q3_2.jar Q3_2*.class
```

**Run Hadoop Job (parameter: <point path> <k-points path> <output path>)**

```
hadoop jar ./q3_2.jar Q3_2 /user/Project2/data/points.csv /user/Project2/data/k-points.csv  
/user/Project2/output/q3_2
```

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

```
hdfs dfs -rm -r /user/Project2/output/q3_2
```

**View the output temporarily**

```
cd ~
```

```
hadoop/bin/hdfs dfs -cat /user/Project2/output/q3_2/iteration_{number_of_iteration_here}/part-r-  
00000
```

ex:

```
hadoop/bin/hdfs dfs -cat /user/Project2/output/q3_2/iteration_0/part-r-00000
```

```
hadoop/bin/hdfs dfs -cat /user/Project2/output/q3_2/iteration_1/part-r-00000
```