CC_project 1 -Readme file

2.1: create (/user/[redacted] directory in hdfs)

1.kmeans.java file is created.

2.compile and produce .jar file as follow:

Change path to kmeans.java and execute below commands

>export HADOOP_CLASSPATH=${JAVA_HOME} /lib/tools.jar

>hadoop com.sun.tools.javac.Main KMeans.java

>jar cf kmeans.jar KMeans*.class

3.Once jar is created. make directory in hdfs file system using below commands

> hdfs dfs -mkdir /user

> hdfs dfs -mkdir /user/[redacted]

> hdfs dfs -mkdir Kmeans_java

> hdfs dfs -mkdir Kmeans_java/input

Put data.hdfs file and centroid file in hdfs

> hdfs dfs -put data.hdfs Kmeans_java/input

> hdfs dfs -put centroids Kmeans_java/input

Herer data.hdfs and centroids are file paths,Kmeans_java/input hdfs file path

4.Execute below command to run the kmeans java version on Hadoop.

>hadoop jar kmeans.jar KMeans Kmeans_java/input/centroids Kmeans_java/input/data.hdfs Kmeans_java/kmeans_ouput

Program runs and output will be stored in input centroid file as we are iterating and storing the final centroids in centroids file at input location.

5.view output by running below command:

> hdfs dfs -cat Kmeans_java/input/centroids

Kmeans.jar – path to kmeans jar file

2.3: Mapper.py file created. Change location to mapper.py file

1.Execute below commands to create directories and put file into the hdfs directories. (/users/[redacted])

>hdfs dfs -mkdir Kmeans_python

>hdfs dfs -mkdir Kmeans_python/input

> hdfs dfs -put data.hdfs Kmeans_python/input

> hdfs dfs -put centroids Kmeans_python/input

2.execute below command for file execution using Hadoop streaming:

>**hadoop jar /usr/localCellar/hadoop/3.1.2/libexec/share/hadoop/tools/lib/hadoop-streaming-3.1.2.jar input Kmeans_python/input/data.hdfs -output Kmeans_python/output -mapper "python mapper.py" -file mapper.py**

Highlighter is the Path to Hadoop streaming jar file. Python 3.7.3 version

3.view output by executing below command:

>hdfs dfs -cat Kmeans_python/output/part-00000

3.1: pyspark_transormation_Action.py (RDD without dataframes)

1.Set up: start pyspark : **./bin/pyspark**
2.Load purchase,book into hdfs using below commands:
>hdfs dfs -mkdir Pyspark
>hdfs dfs -put purchase Pyspark
>hdfs dfs -put book Pyspark

3.change location to "pysprak_transormation_Action.py"
4.Execute below command:
> spark-submit pyspark_transormation_Action.py
Or
>python pyspark_transormation_Action.py

3.2: Pyspark_SQL.py(RDD using data frames and SQL)

1.Change location to "Pyspark_SQL.py"
2.execute below command:
> spark-submit Pyspark_SQL.py
Or
>python Pyspark_SQL.py