Comandi Hadoop KMeans

Dopo aver effettuato l'accesso al cluster creare una cartella KMeansProject in cui caricare il dataset, il file

Manifest e i file di codice, e spostarsi al suo interno:

1. `mkdir KMeansProject`

Poi si deve esportare il CLASSPATH:

2. `export CLASSPATH="$HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-client-core2.9.1.jar:$HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-client-common-2.9.1.jar:$HADOOP_HOME/share/hadoop/common/hadoop-common-2.9.1.jar:~/KMeansProject/KMeans/*:$HADOOP_HOME/lib/*"`

Dopodiché si devono compilare i file di codice e creare il file .jar:

3. `javac -d . Element.java KMeansMapper.java KMeansCombiner.java KMeansReducer.java KMeansDriver.java`
4. `jar cfm KMeans.jar Manifest.txt KMeans/*.class`

Ora dovremo creare nel filesystem di Hadoop la cartella in cui andremo a caricare il dataset:

5. `hdfs dfs -mkdir kmeanDirectory`
6. `hdfs dfs -put <nome_file_dataset> kmeanDirectory`

Per verificare:

- `hdfs dfs -ls kmeanDirectory`

Una volta compilati i file e caricato il dataset possiamo eseguire il programma:

7. `$HADOOP_HOME/bin/hadoop jar KMeans.jar <input_dir> <output_dir> <num_centers> <n_parameters> <loop> <max_num> <split_char>`
   ($HADOOP_HOME/bin/hadoop jar KMeans.jar kmeanDirectory output_mean 5 3 20 50 t **oppure** $HADOOP_HOME/bin/hadoop jar KMeans.jar kmeanDirectory output_mean 5 3 20 50 \,)

Dove:

- `Input_dir` : cartella del dataset
- `output_dir` : cartella in cui verrà messo il risultato
- `num_centers` : numero di centri da usare
- `n_parameters` : numero di parametri da usare
- `loop` : numero di cicli che devono essere effettuati
- `max_num` : numero massimo utilizzato per la scelta casuale dei centri (grandezza dataset). Se 0 si usano le prime n righe del dataset
- `split_char` : carattere usato per la concatenazione dei parametri nel dataset. Usare 't' nel caso di tabulazioni, gli altri caratteri devono essere preceduti da '\' (es. \,)

Quando l'esecuzione è ultimata è possibile scaricare il file contenente i centri e quello contenente il

risultato dell'elaborazione:

- `hdfs dfs -get output_mean/part-r-00000` (risulatato elaborazione)

- `hdfs dfs -get centers/cent.txt` (risultato centri)

Prima di rieseguire nuovamente il programma assicurarsi di eliminare il precedente risultato:

- `hdfs dfs -rm -r output_mean`