Database e Data Mining

Practice N. 4

The goal of the practice is the implementation of the equivalent “Hello Word” program for the MapReduce world: “Word Count”. This is the simplest example of MapReduce job: *count the number of occurrences of each word in a given text file*.

Tools:

- Eclipse
- Cygwin

1. Create the Project with Eclipse

- Create a new **Java Project**
- Set a name for the project
- Select **JavaSE-1.6** as JRE
- Go on the Library tab and click on Import **External jars**
- Import the following jars
  - **hadoop-common.jar**
  - **hadoop-core.jar**
  - **hadoop-hdfs**
- Click on **Finish**
- Select the **src folder** of the new created Java Project
- Right-click on the src folder and select **Import**
- Select **File System**
- In **From Directory**, select **stub/src/it**
- Select it on the left and click **Finish**
- Right-click on the imported package
- Select **Refactor** and then **Rename**
- Remove trailing “**_stub**” from the package name and rename it

2. Write the Job

- Complete all **TODO** in:
  - WordCountMapper.java
  - WordCountReducer.java
  - WordCount.java

3. Esport the Job JAR file

- Select **File>Export** from the menu
- Select **JAR File** and click on **Next**
- Select package and files (if not already selected)
- Click on Browse to **Select the export destination** and give the following name to the JAR
  - **tuo_nome**-wc.jar (e.g., luigi-wc.jar)
- Click on **Finish**

4. Upload the Job JAR file

- Open a **shell** (**Cygwin terminal** for window users)
- Upload the jar file to the remote server

  -
    ```
    scp tuo_nome-wc.jar master2013@dbdmgmtr.polito.it:/home/master2013
    ```

  -
    ```
    password: master2013!
    ```

5. Submit the Job

- Launch the job with

  -
    ```
    hadoop jar <tuo_nome-wc.jar> <fully.qualified.class.Name> <Parameters>
    ```
  - E.g.,
    ```
    hadoop jar luigi-wc.jar it.polito.dbdmg.hadoop.wordcount.WordCount 1 data/divina_commedia luigi/luigi-wc-out
    ```

6. Get the results

- Visualize the results with

  -
    ```
    hadoop fs -cat tuo_nome/tuo_nome-wc-out/part-r-* | less
    ```

  -
    ```
    E.g., hadoop fs -cat luigi/luigi-wc-out/part-r-* | less
    ```

7. Remove results

- Remove an old result folder with
  - `hadoop fs -rm -r `**`tuo_nome`**`/`**`tuo_nome`**`-wc-out`
  - E.g., `hadoop fs -rm -r luigi/luigi-wc-out`