

Instructions for Hadoop Data Analysis

CS555: Big Data Computing Lab

2nd August 2021

Problem Statement: **what are the top 5 categories with maximum number of videos uploaded.**

1. create project directory : (youtube)

- a. Browse to home directory

cd /home/iitp

- b. Create project directory

mkdir youtube

2. Create source file: (Top5_categories.java)

- a. Browse into youtube directory

cd /home/iitp/youtube

- b. Create file

nano Top5_categories.java

paste the lines from the source code provided

Note:- To save file: Press- CTRL + o followed by Enter button

To Exit Press:- CTRL + x from nano editor

3. Create input directory (inputdata) for input files

cd /home/iitp/youtube

mkdir inputdata

4. Copy the input file (youtubedata.txt) into inputdata folder

cd /home/iitp/youtube/inputdata

nano youtubedata.txt

paste the lines from the provided input file

5. Start all hadoop services

- a. Browse to hadoop installation sbin sub-directory

cd /home/iitp/hadoop-2.6.0/sbin

- b. start all services

./start-all.sh

Note:- Enter password when prompted

- 6. Create input directory on HDFS
 - a. browse to hadoop installation bin folder

cd /home/iitp/hadoop-2.6.0/bin

- b. create directory (youtube)

./hadoop fs -mkdir /youtube

- c. create subdirectory (inputdata) inside youtube on HDFS

./hadoop fs -mkdir /youtube/inputdata

- 7. Copy the input text file from local directory to HDFS
 - a. browse to hadoop installation bin folder

cd /home/iitp/hadoop-2.6.0/bin

- b. Copy from Local

./hadoop dfs -put /home/iitp/youtube/inputdata/youtubedata.txt /youtube/inputdata/

- 8. Compile the Source Code
 - a. export the Hadoop classpath

export HADOOP_CLASSPATH=/usr/lib/jvm/java-1.8.0-openjdk-amd64/lib/tools.jar

- b. browse to bin folder of hadoop installation

cd /home/iitp/hadoop-2.6.0/bin

- c. Compile

./hadoop com.sun.tools.javac.Main /home/iitp/youtube/Top5_categories.java

- 9. Create Jar file
 - a. Browse to youtube directory on your VM

cd /home/iitp/youtube

jar cf youtube1.jar Top5_categories*.class

10. Running the program

- a. browse to the bin directory of hadoop installation

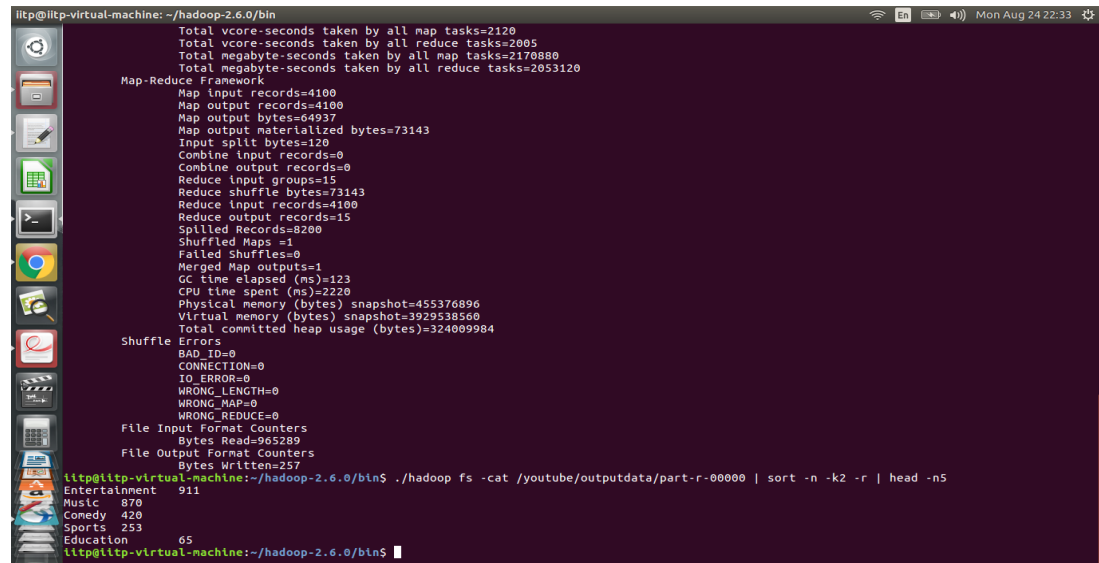
cd /home/iitp/hadoop-2.6.0/bin

- b. Running in terminal

**./hadoop jar /home/iitp/youtube/youtube1.jar Top5_categories /youtube/inputdata/
/youtube/outputdata**

- c. Finding outputs

./hadoop fs -cat /youtube/outputdata/part-r-00000 | sort -n -k2 -r | head -n5



The screenshot shows a terminal window with the following content:

```
iitp@iitp-virtual-machine: ~/hadoop-2.6.0/bin
Total vcore-seconds taken by all map tasks=2120
Total vcore-seconds taken by all reduce tasks=2005
Total megabyte-seconds taken by all map tasks=2170880
Total megabyte-seconds taken by all reduce tasks=2053120
Map-Reduce Framework
  Map input records=4100
  Map output records=4100
  Map output bytes=64937
  Map output materialized bytes=73143
  Input split bytes=120
  Combine input records=0
  Combine output records=0
  Reduce input groups=15
  Reduce shuffle bytes=73143
  Reduce input records=4100
  Reduce output records=15
  Spilled Records=8200
  Shuffled Maps =1
  Failed Shuffles=0
  Merged Map outputs=1
  GC time elapsed (ms)=123
  CPU time spent (ms)=2220
  Physical memory (bytes) snapshot=455376896
  Virtual memory (bytes) snapshot=3929538560
  Total committed heap usage (bytes)=324009984
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=965289
File Output Format Counters
  Bytes Written=257
iitp@iitp-virtual-machine:~/hadoop-2.6.0/bin$ ./hadoop fs -cat /youtube/outputdata/part-r-00000 | sort -n -k2 -r | head -n5
Entertainment  911
Music          870
Comedy         420
Sports         253
Education      65
iitp@iitp-virtual-machine:~/hadoop-2.6.0/bin$
```