Hola Hadoop

0. Peligro!

Please please

0. Peligro!

… please

Peligro!

… please

Peligro!

… please be careful of what you are doing!

- Think twice before:

  rm

  mv

  cp

  kill

  emacs/vim/… configuration files

Peligro!

… please.

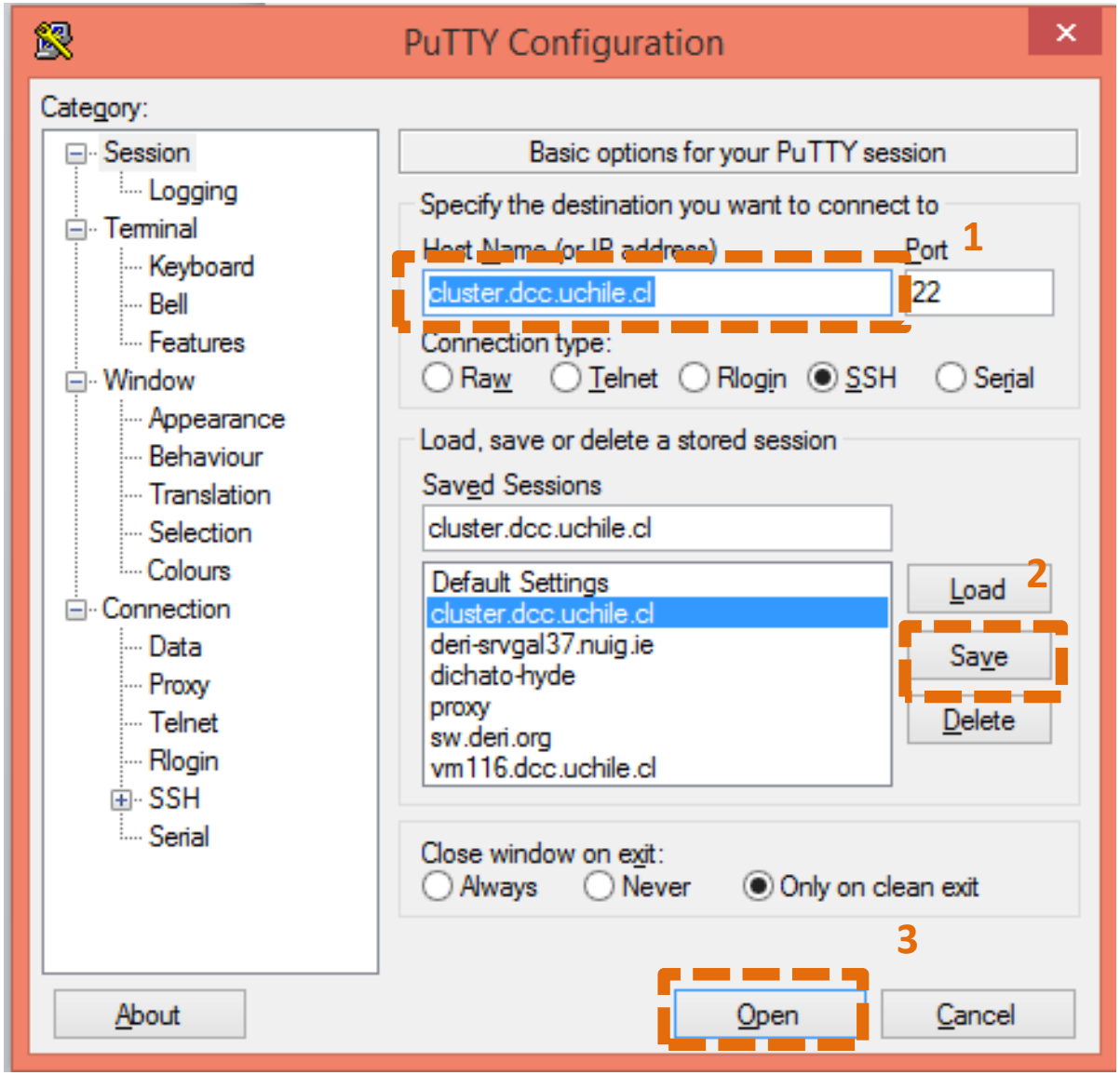
- cluster.dcc.uchile.cl

1. Download tools

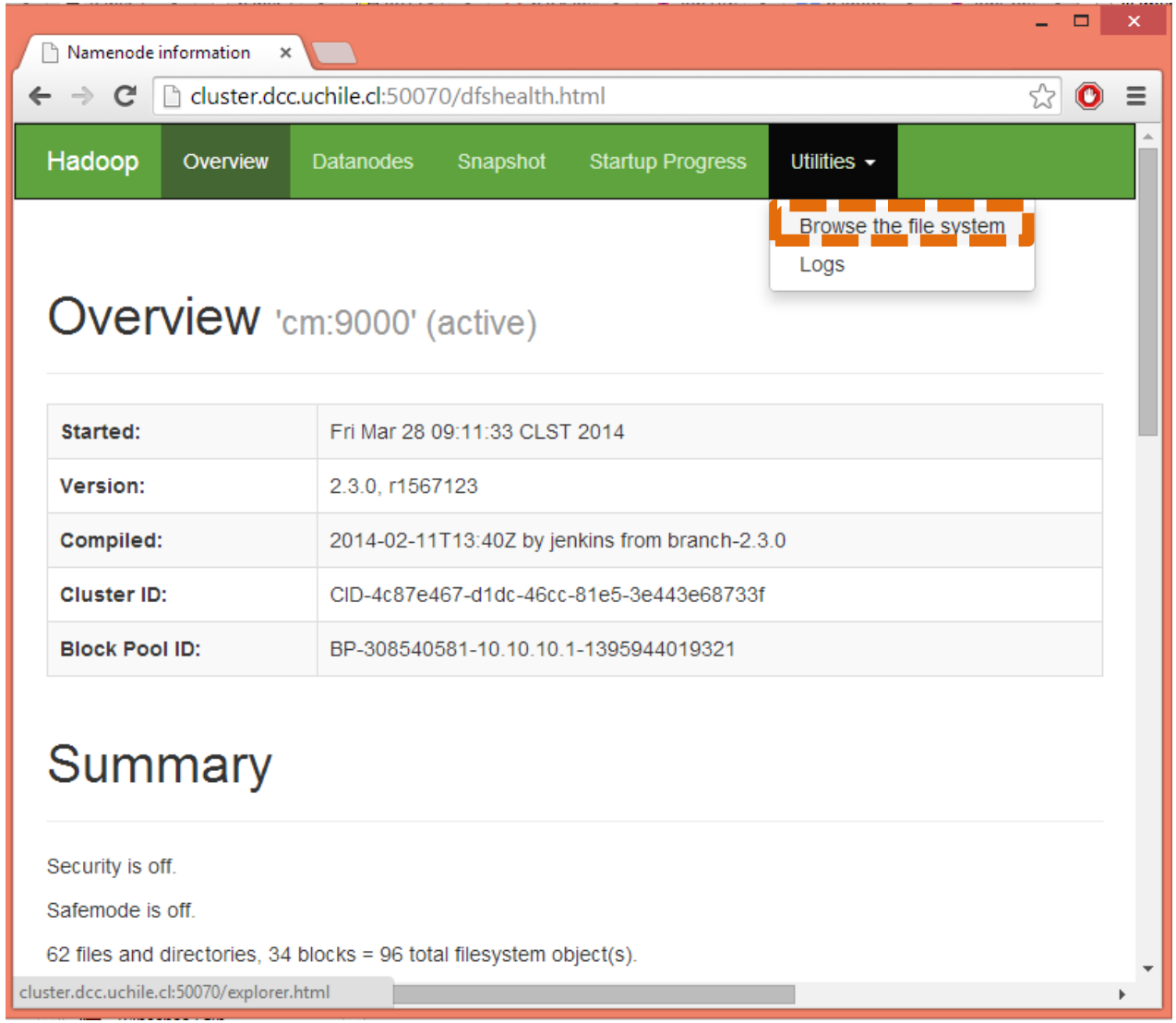
- http://aidanhogan.com/teaching/cc5212-1-2015/tools/

- Unzip them somewhere you can find them

2. Log-in PuTTy



3. Open DFS Browser



http://cluster.dcc.uchile.cl:50070/

4. PuTTy: See state of DFS

- hdfs dfsadmin -report

5. PuTTy: Create folder

- hdfs dfs -ls /
- hdfs dfs -ls /uhadoop
- hdfs dfs -mkdir /uhadoop/[username]
  - [username] = first letter first name, last name (e.g., “ahogan”)

6. PuTTy: Upload Data

- cd /data/2014/uhadoop/shared/
- Then
  - hdfs dfs -copyFromLocal /data/2014/uhadoop/shared/es-wiki-abstracts.txt /uhadoop/[username]/

    OR

  - hdfs dfs -copyFromLocal /data/2014/uhadoop/shared/es-wiki-abstracts.txt.gz /uhadoop/[username]/

Note on namespace

- If you need to disambiguate local/remote files
- HDFS file
  - hdfs://cm:9000/uhadoop/…
- Local file
  - file:///data/hadoop/...

7. Let's Build Our First MapReduce Job

- Hint: Use Monday's slides for "inspiration"
  - http://aidanhogan.com/teaching/cc5212-1-2016/
  - Also copied in `CitationCount.java`

1. Implement map(.,.,.,.) method

2. Implement reduce(.,.,.,.) method

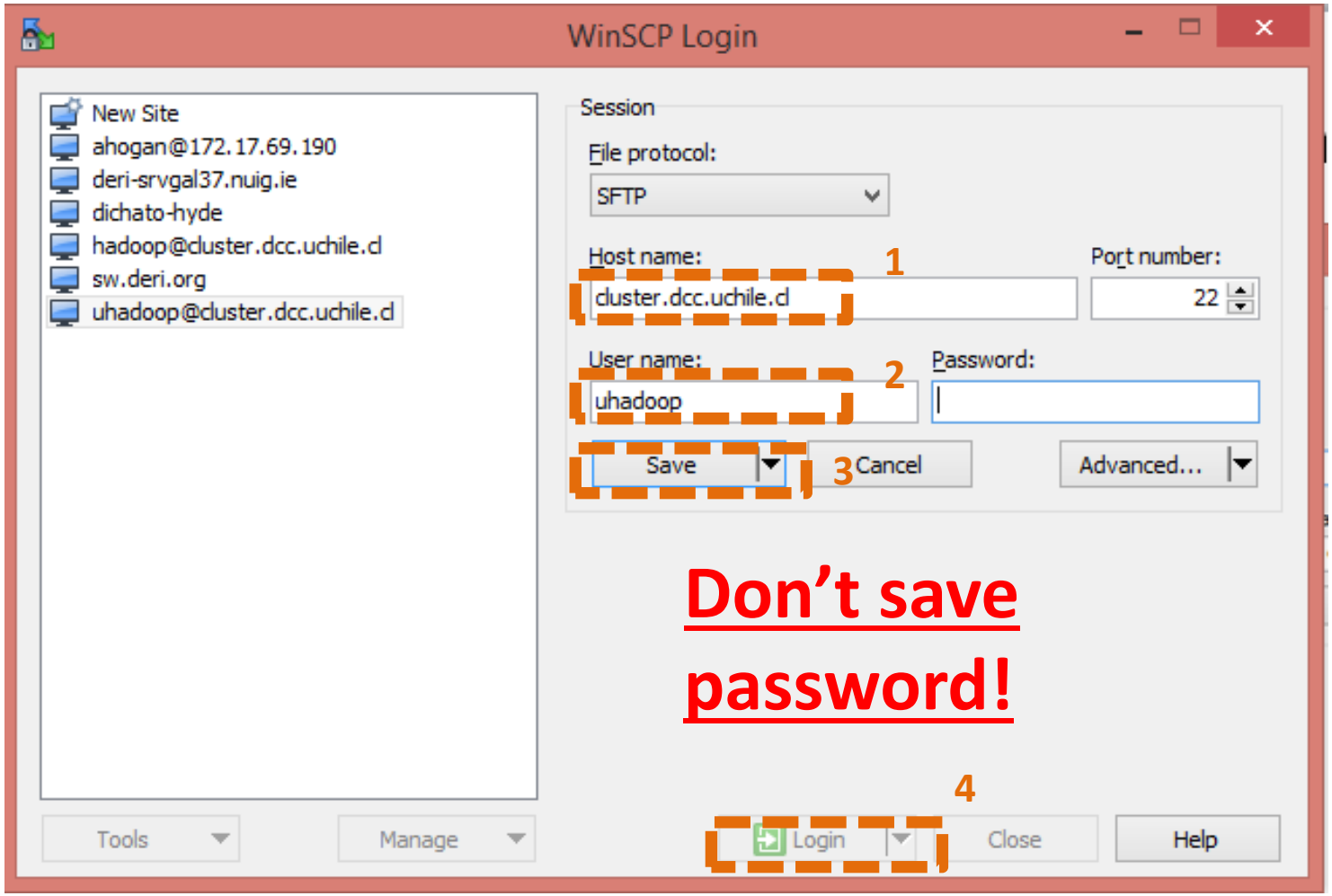
3. Implement main(.) method

8. Eclipse: Build jar

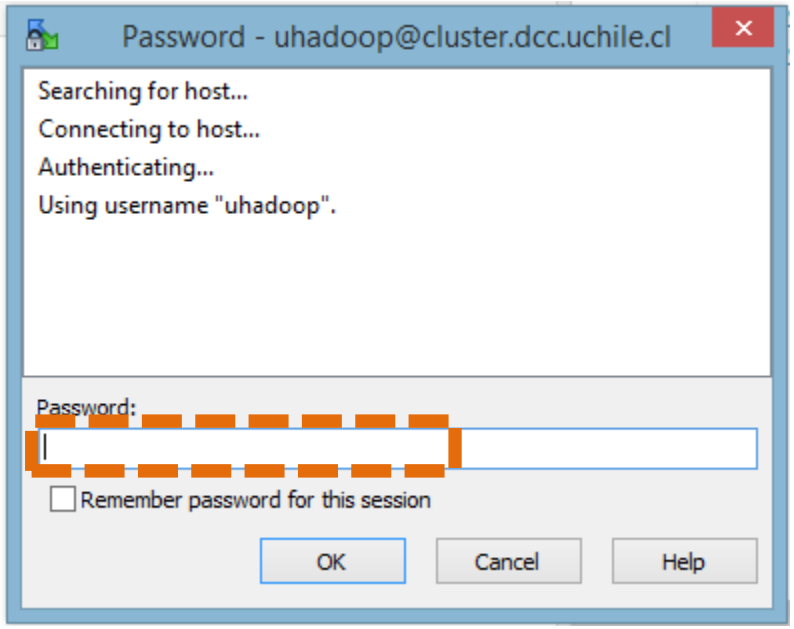
Right Click build.xml > dist

(Might need to make a dist folder)

9. WinSCP: Copy .jar to Master Server

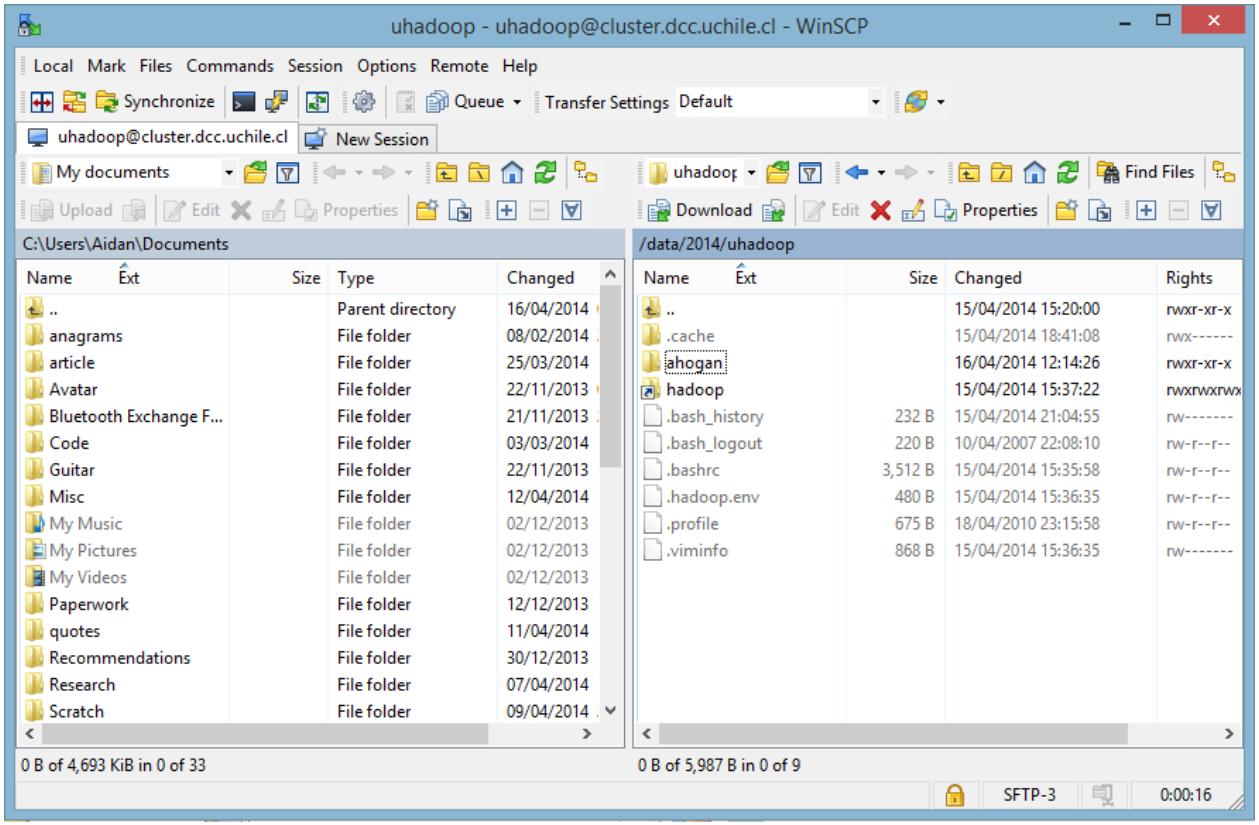


9. WinSCP: Copy .jar to Master Server



9. WinSCP: Copy .jar to Master Server

- Create dir: /data/2014/uhadoop/[username]/
- Copy your mdp-lab4.jar into it



10. PuTTy: Run Job

- hadoop jar /data/2014/uhadoop/[username]/mdp-lab4.jar WordCount /uhadoop/[username]/es-wiki-abstracts.txt[.gz] /uhadoop/[username]/wc/

All one command!

11. PuTTY: Look at output

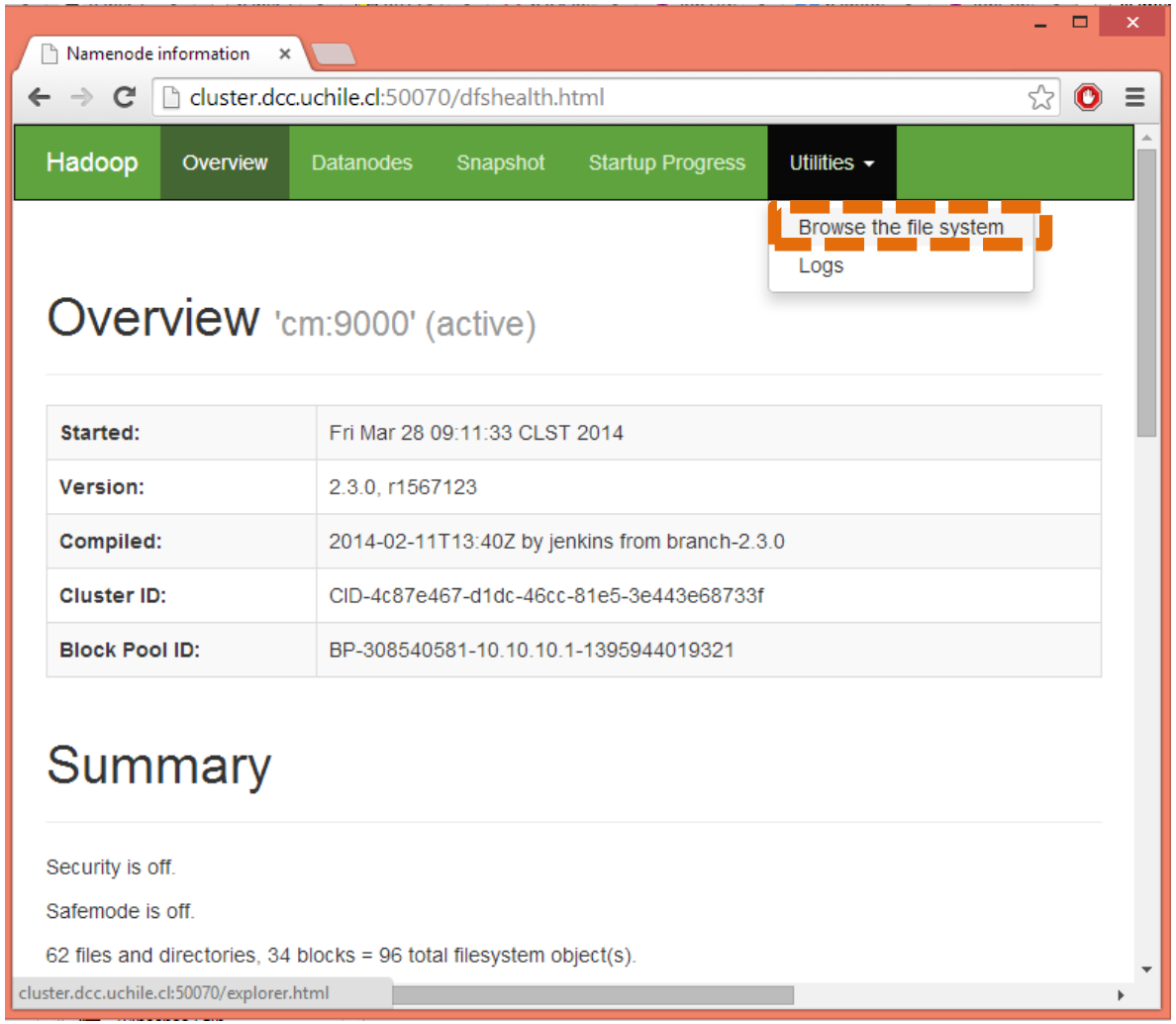
- hdfs dfs -ls /uhadoop/[username]/wc/
- hdfs dfs -cat /uhadoop/[username]/wc/part-r-00000 | more

All one command!

- hdfs dfs -cat /uhadoop/[username]/wc/part-r-00000 | grep -P "^de\t" | more

Look for “de” … 4916432 occurrences in local run

12. Look at output through browser



http://cluster.dcc.uchile.cl:50070/


HELLO-WORLD