CS522 – Big Data

PROJECT

Quoc Anh Khuong – 985391

Hien Vo – 985433

Maharishi University of Management

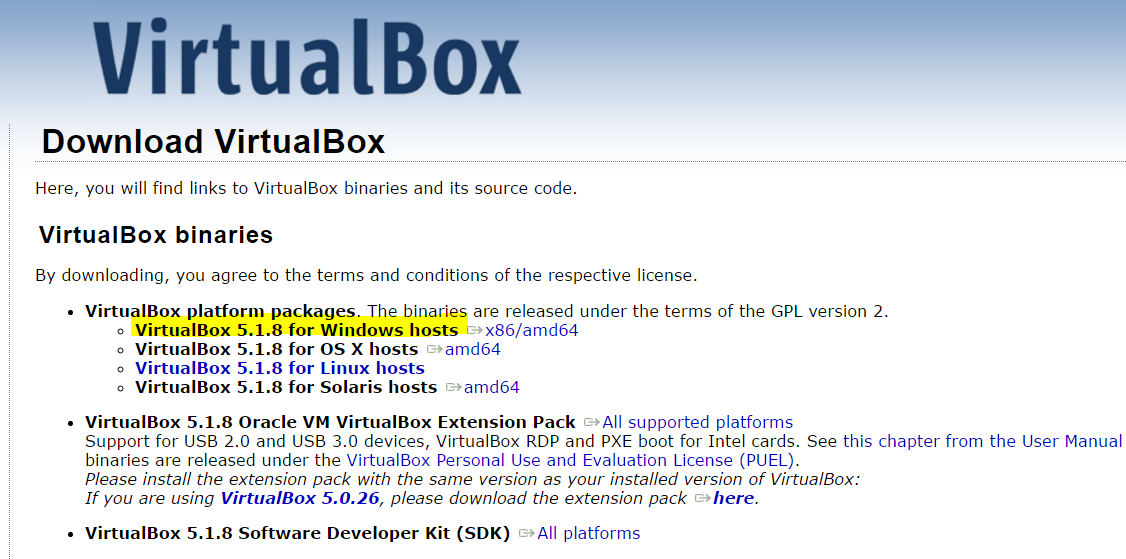
Department of Computer Science

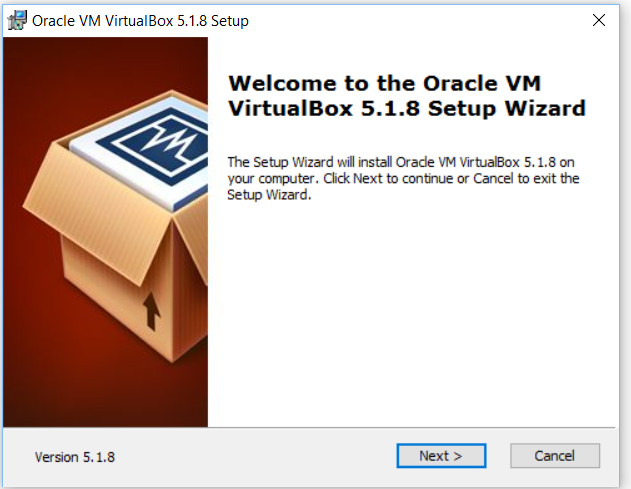
Oct 16, 2016

**A. PART 1: SETUP AND RUN WordCount Sample**

1. Download & Install Virtual Box

<https://www.virtualbox.org/wiki/Downloads>

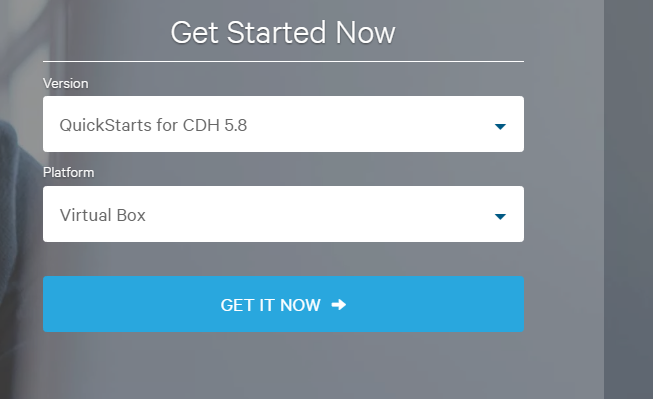


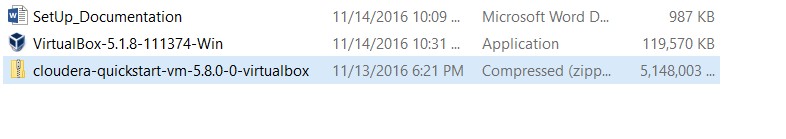


1. Download & Import Cloudera 5.8

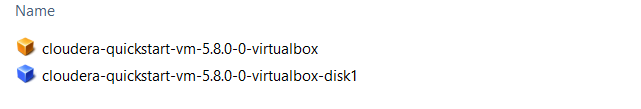
2.1 Download

<http://www.cloudera.com/downloads/quickstart_vms/5-8.html>



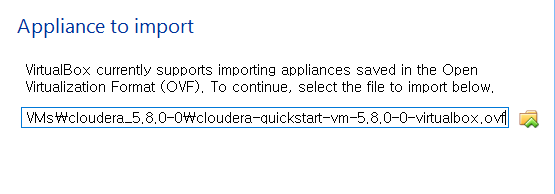


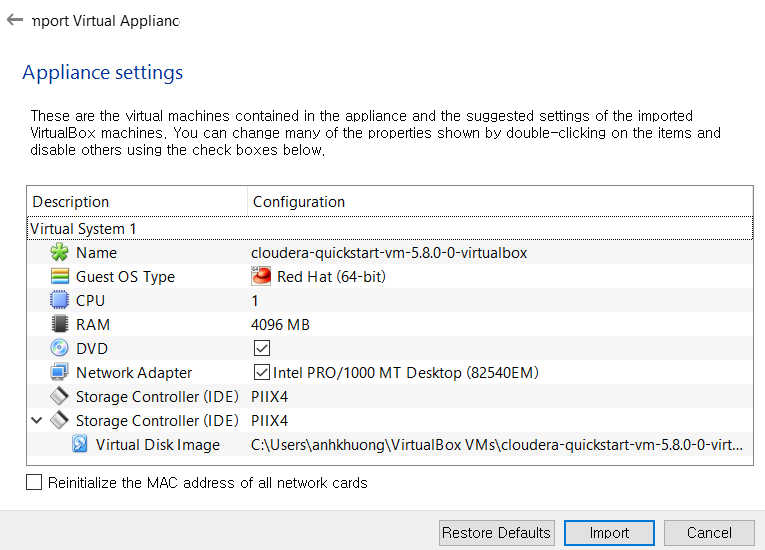
Unzip cloudera

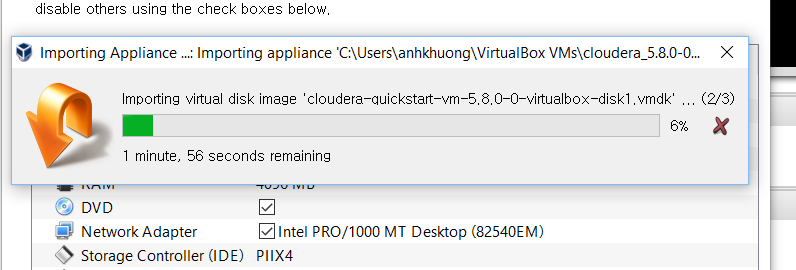


2.2 Import Cloudera to VirtualBox

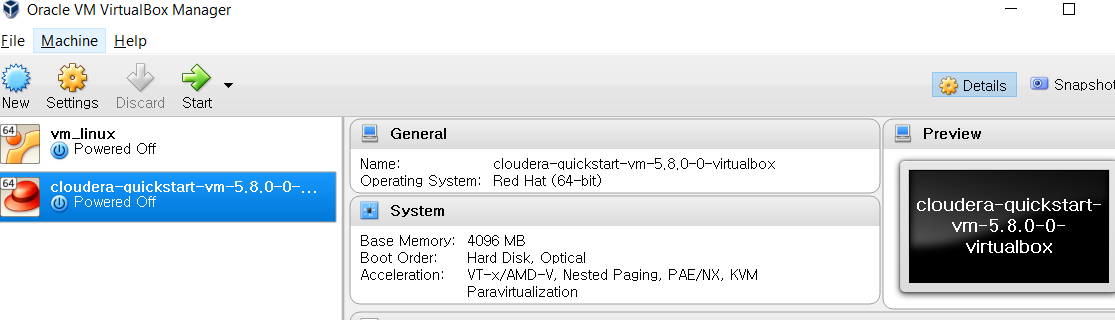
VirtualBox: File -> Import Appliance…



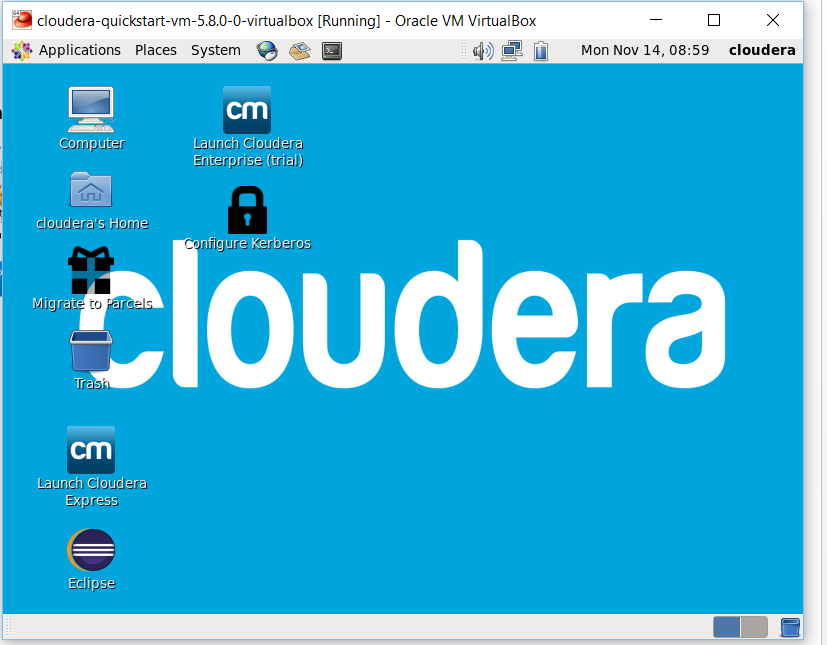




Finish import cloudera image

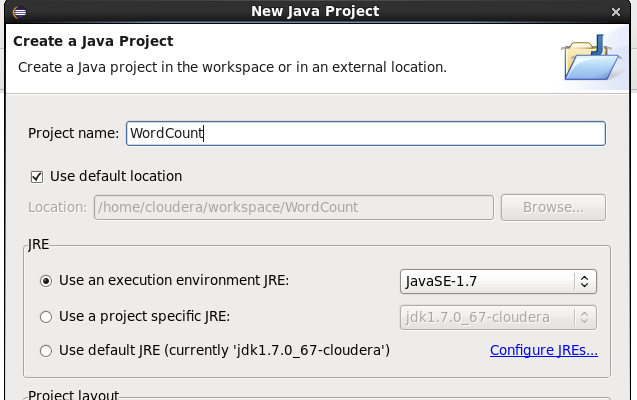


Run

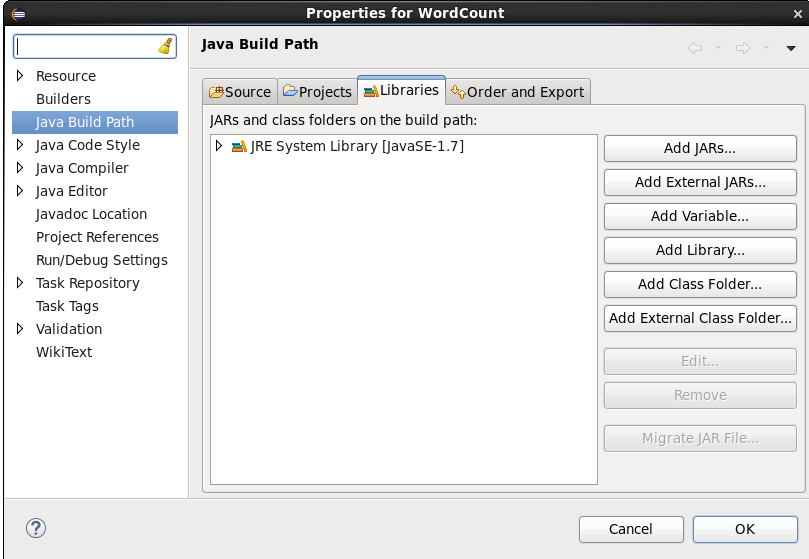


3. WordCount Project

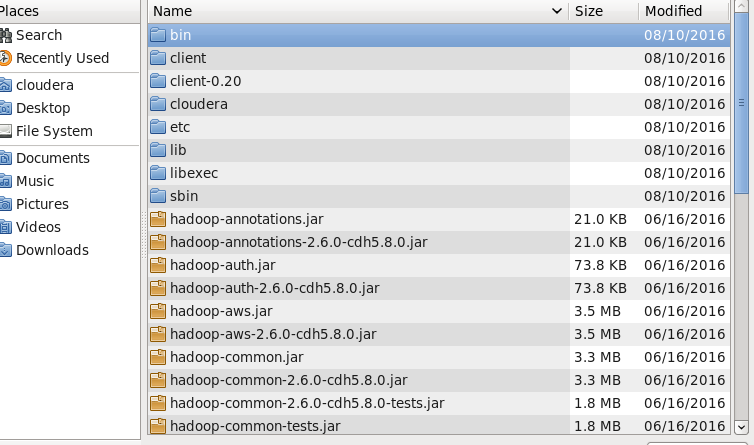
3.1 Create Project in Eclipse



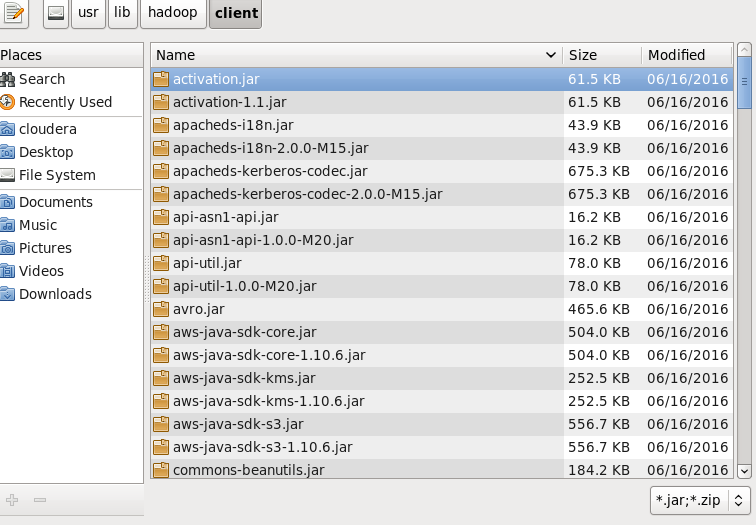
3.2 Get references Hadoop libraries



* Click “Add External JARs…” and select all jar files in /usr/lib/hadoop

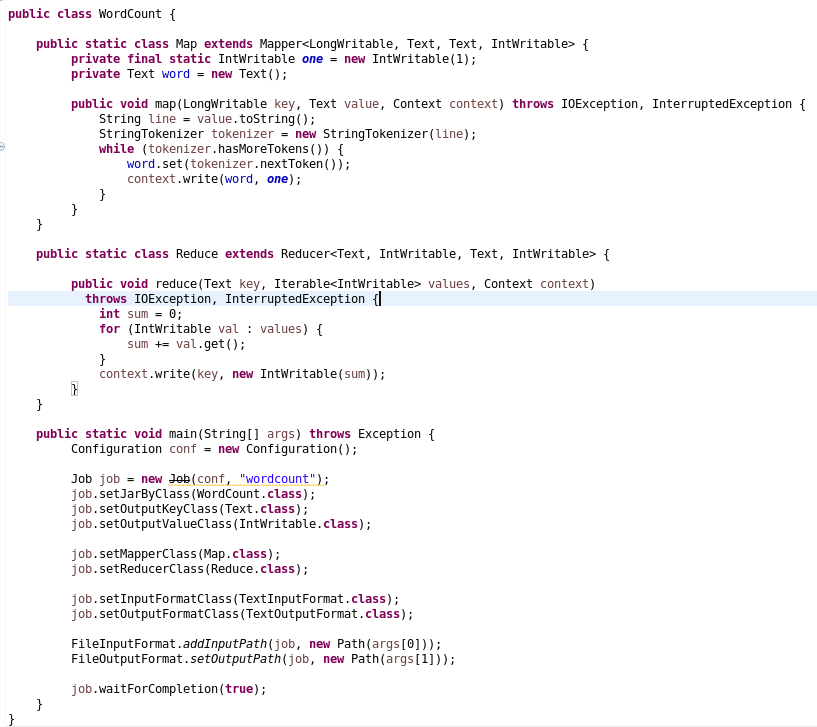


* Click “Add External JARs…” and select all jar files in /usr/lib/hadoop/client



3.2 Implementation of WordCount

Find Hadoop WordCount example on Internet and add to the project as WordCount class



- Build and export jar file “wordcount.jar”

- Create input data file “data.txt” in Linux file system and convert it to hdfs by

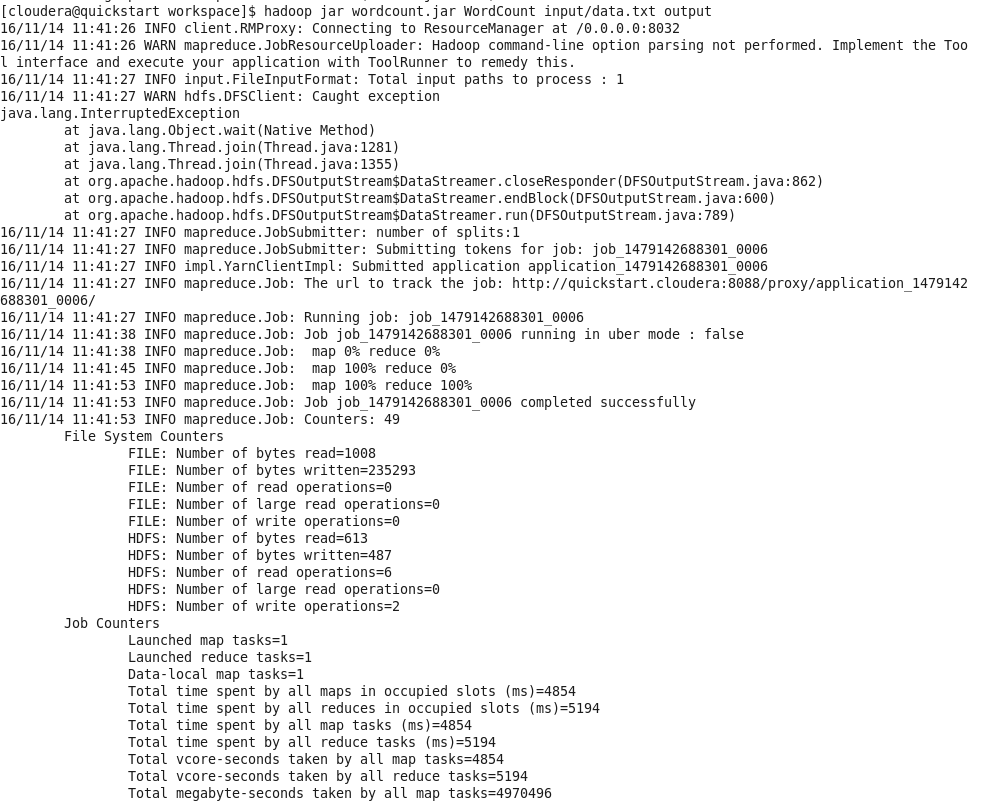
command:

$hadoop fs -mkdir input

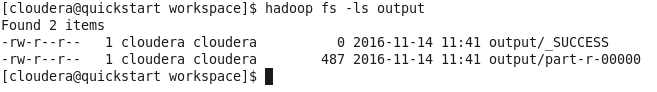
$hadoop fs -put /home/cloudera/data.txt input/

- Run program by commands

$hadoop jar wordcount.jar WordCount input/data.txt output



Result:



Check result in part-r-00000 file by command:

$hadoop -cat output/part-r-00000

**B. PARTS 2, 3 and 4**

- Source files

Part2: source file: Pairs.java; Jar file: pairs.jar

Part3: source file: Stripes.java; Jar file: Stripes.jar

Part4: source file: Hybrid.java; Jar file: Hyrid.jar

- For all parts 2, 3 and 4:

+ Batch file to put “input.txt” to hdfs: hdfd\_input.sh

+ Batch file to run program: run.sh

+ Batch file to get “output.txt” from hdfs result: hdfs\_output.sh

**How to run Program:**

# Create hdfs input file:

$./hdfs\_input.sh

# Run program

$./run.sh

# Export result to output.txt

$./hdfs\_output.sh