EPL646 – Advanced Topics in Databases
Advanced Hadoop

http://www.cs.ucy.ac.cy/~dzeina/courses/epl646/labs/lab.html

Calculate time

- How to calculate the time

```
long begin = System.currentTimeMillis();
job.waitForCompletion(true);
long end = System.currentTimeMillis();
long second = (end - begin) / 1000;
System.err.println(job.getJobName() + " takes " +
second + " seconds");
```

Task1: N-Gram

- Change the code in WordCount so that it counts how many times each set of five consecutive words appears
- You can find the code of WordCount from the solution of the previous lab
- If you don't have the datasets you can download them from the previous lab

Task1: N-Gram

- Function ***map*** will have as input:
  - key = line offset (we can ignore it)
  - value = a whole line from one of the input files
- Function ***map*** will have as output:
  - key = five words
  - value = 1
- Function ***reduce*** will have as input:
  - key = five words
  - value = [a list of number 1]
- The list will as many 1 as there are appearances of the five consecutive words in our data
- Function *Reduce* will have as final output:
  - key = five words
  - value = the sum of all 1 (i.e. the same as WordCount)

Task2: Anagram

- An anagram is a word that can be created by the movement of the letters of another word
- E.g.
  - Refills→fillers
  - Relayed→layered
  - Rentals→antlers
  - Rebuild→builder
- You must find the anagrams in a huge input file. How would you do it?

```
public static boolean isAnagram(String first, String second) {
    // Checks that the two inputs are anagrams, by checking they have all the same characters.
    // Left as exercise for the user...
}
```

Task2: Anagram

- **Hadoopifying…**
  - (**input**) <k1, v1> →
  - map → <k2, v2> →
  - combine → <k2, v2>→
  - reduce → <k3, v3> (**output**)

- Download
  - /usr/share/dict/words or /usr/dict/words

Jar file configuration

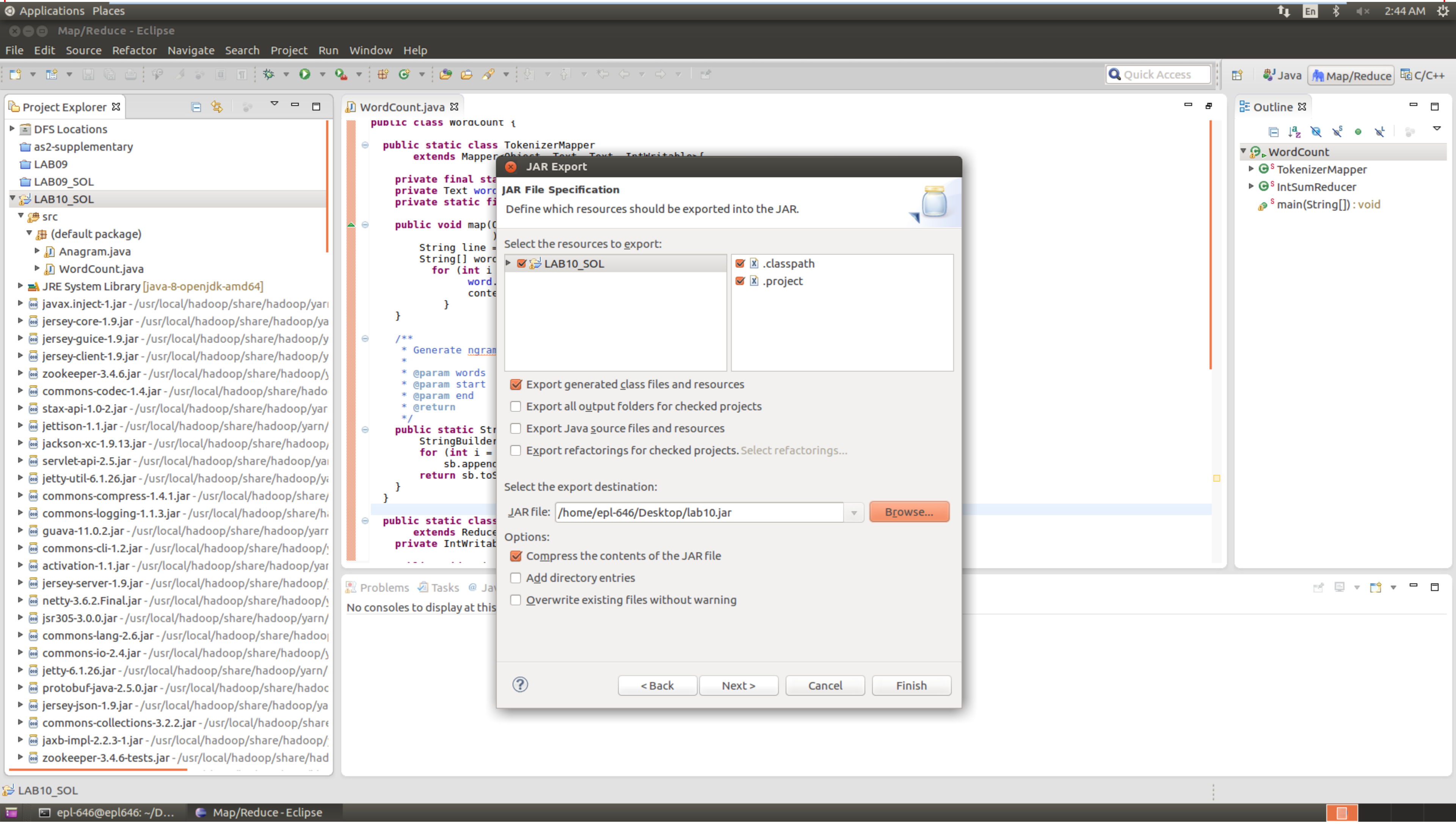
You need to set the jar by class parameter:

```
Configuration conf = new Configuration();
Job job = Job.getInstance(conf, "word count");
job.setJarByClass(WordCount.class);
```

Else you will get:

java.lang.RuntimeException: java.lang.ClassNotFoundException:

Export the .jar file



JAR Export

JAR File Specification

Define which resources should be exported into the JAR.

Select the resources to export:

- ☑ LAB10_SOL
- ☑ .classpath
- ☑ .project

☑ Export generated class files and resources

☐ Export all output folders for checked projects

☐ Export Java source files and resources

☐ Export refactorings for checked projects. Select refactorings...

Select the export destination:

JAR file: /home/epl-646/Desktop/lab10.jar

Options:

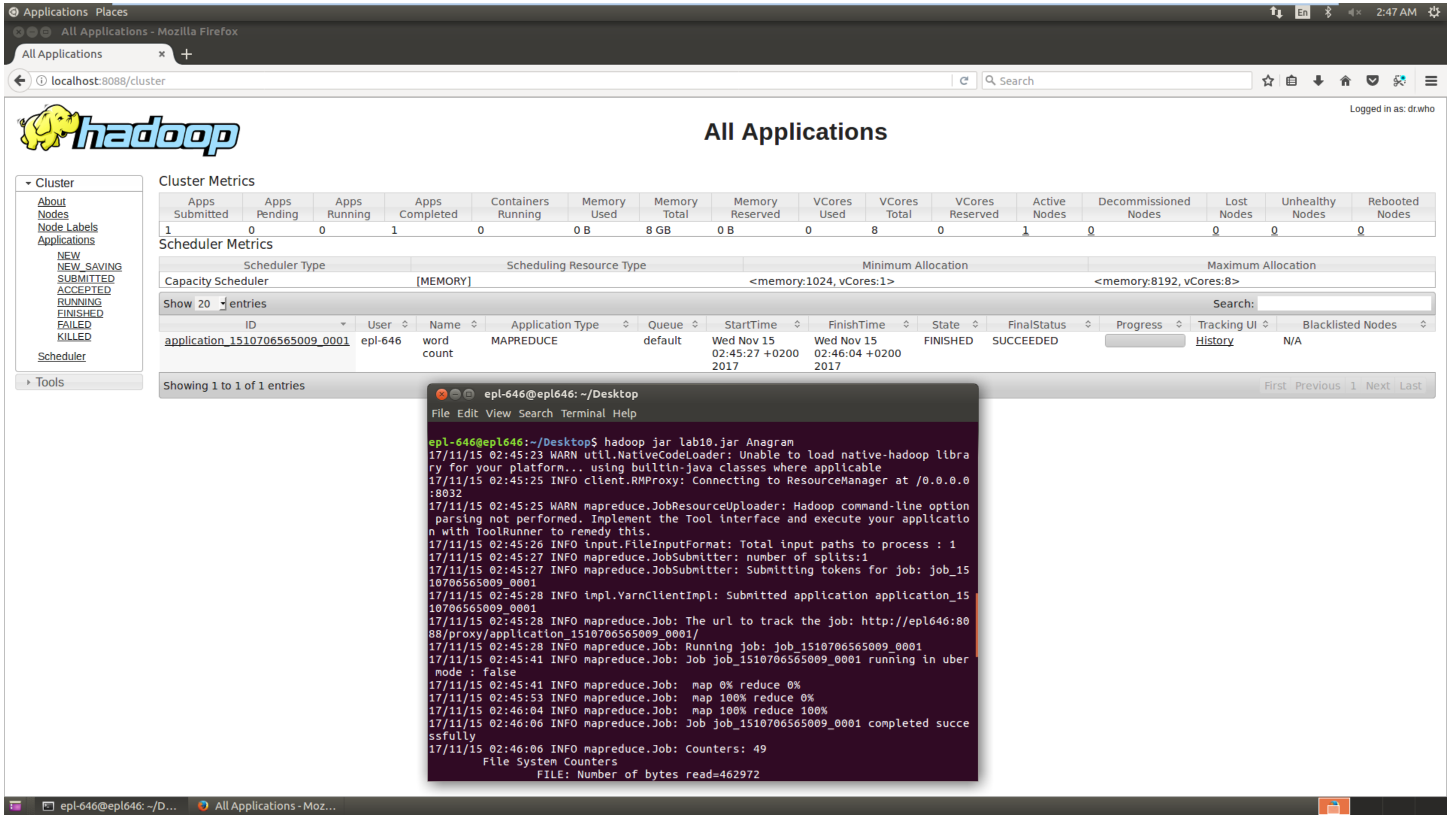
☑ Compress the contents of the JAR file

☐ Add directory entries

☐ Overwrite existing files without warning

Applications Places
2:44 AM
Map/Reduce - Eclipse
File Edit Source Refactor Navigate Search Project Run Window Help
Quick Access
Java
Map/Reduce
C/C++
Project Explorer
DFS Locations
as2-supplementary
LAB09
LAB09_SOL
LAB10_SOL
src
(default package)
Anagram.java
WordCount.java
JRE System Library [java-8-openjdk-amd64]
javax.inject-1.jar - /usr/local/hadoop/share/hadoop/yar
jersey-core-1.9.jar - /usr/local/hadoop/share/hadoop/ya
jersey-guice-1.9.jar - /usr/local/hadoop/share/hadoop/y
jersey-client-1.9.jar - /usr/local/hadoop/share/hadoop/y
zookeeper-3.4.6.jar - /usr/local/hadoop/share/hadoop/
commons-codec-1.4.jar - /usr/local/hadoop/share/hado
stax-api-1.0-2.jar - /usr/local/hadoop/share/hadoop/yar
jettison-1.1.jar - /usr/local/hadoop/share/hadoop/yarn/
jackson-xc-1.9.13.jar - /usr/local/hadoop/share/hadoop/
servlet-api-2.5.jar - /usr/local/hadoop/share/hadoop/ya
jetty-util-6.1.26.jar - /usr/local/hadoop/share/hadoop/y
commons-compress-1.4.1.jar - /usr/local/hadoop/share/
commons-logging-1.1.3.jar - /usr/local/hadoop/share/h
guava-11.0.2.jar - /usr/local/hadoop/share/hadoop/yar
commons-cli-1.2.jar - /usr/local/hadoop/share/hadoop/
activation-1.1.jar - /usr/local/hadoop/share/hadoop/yar
jersey-server-1.9.jar - /usr/local/hadoop/share/hadoop/
netty-3.6.2.Final.jar - /usr/local/hadoop/share/hadoop/
jsr305-3.0.0.jar - /usr/local/hadoop/share/hadoop/yarn/
commons-lang-2.6.jar - /usr/local/hadoop/share/hadoo
commons-io-2.4.jar - /usr/local/hadoop/share/hadoop/
jetty-6.1.26.jar - /usr/local/hadoop/share/hadoop/yarn/
protobuf-java-2.5.0.jar - /usr/local/hadoop/share/hadoo
jersey-json-1.9.jar - /usr/local/hadoop/share/hadoop/ya
commons-collections-3.2.2.jar - /usr/local/hadoop/share
jaxb-impl-2.2.3-1.jar - /usr/local/hadoop/share/hadoop/
zookeeper-3.4.6-tests.jar - /usr/local/hadoop/share/had
WordCount.java
Outline
WordCount
TokenizerMapper
IntSumReducer
main(String[]) : void
Problems
Tasks
No consoles to display at this
Browse...
< Back
Next >
Cancel
Finish
LAB10_SOL
epl-646@epl646: ~/D...
Map/Reduce - Eclipse

Run the jar file



All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved	Active Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes
1	0	0	1	0	0 B	8 GB	0 B	0	8	0	1	0	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation
Capacity Scheduler	[MEMORY]	<memory:1024, vCores:1>	<memory:8192, vCores:8>

Show 20 entries Search:

ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus	Progress	Tracking UI	Blacklisted Nodes
application_1510706565009_0001	epl-646	word count	MAPREDUCE	default	Wed Nov 15 02:45:27 +0200 2017	Wed Nov 15 02:46:04 +0200 2017	FINISHED	SUCCEEDED		History	N/A

Showing 1 to 1 of 1 entries First Previous 1 Next Last

epl-646@epl646: ~/Desktop

File Edit View Search Terminal Help

```
epl-646@epl646:~/Desktop$ hadoop jar lab10.jar Anagram
17/11/15 02:45:23 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
17/11/15 02:45:25 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
17/11/15 02:45:25 WARN mapreduce.JobResourceUploader: Hadoop command-line option
 parsing not performed. Implement the Tool interface and execute your applicatio
n with ToolRunner to remedy this.
17/11/15 02:45:26 INFO input.FileInputFormat: Total input paths to process : 1
17/11/15 02:45:27 INFO mapreduce.JobSubmitter: number of splits:1
17/11/15 02:45:27 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_15
10706565009_0001
17/11/15 02:45:28 INFO impl.YarnClientImpl: Submitted application application_15
10706565009_0001
17/11/15 02:45:28 INFO mapreduce.Job: The url to track the job: http://epl646:80
88/proxy/application_1510706565009_0001/
17/11/15 02:45:28 INFO mapreduce.Job: Running job: job_1510706565009_0001
17/11/15 02:45:41 INFO mapreduce.Job: Job job_1510706565009_0001 running in uber
 mode : false
17/11/15 02:45:41 INFO mapreduce.Job:  map 0% reduce 0%
17/11/15 02:45:53 INFO mapreduce.Job:  map 100% reduce 0%
17/11/15 02:46:04 INFO mapreduce.Job:  map 100% reduce 100%
17/11/15 02:46:06 INFO mapreduce.Job: Job job_1510706565009_0001 completed succe
ssfully
17/11/15 02:46:06 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=462972
```

- If the node is unhealthy you may need to execute the following command and then restart hadoop:

```
chown -r epl-646:epl-646  /app/hadoop/
```

Questions?

http://www.cs.ucy.ac.cy/~dzeina/courses/epl646/labs/lab.html

University of Cyprus

DMSL
@UCY