CSE3025 - Large Scale Data Processing

Lab - 2 - 11/12/2017

PREPARED BY - HARGUR PARTAP SINGH BEDI (15BCE1257)

Simple MapReduce program in Hadoop

1. Start the hadoop cluster by typing the command

$ bash hadoop/bin/start-all.sh

2. Copy a text file for counting the count of each distinct word in it, using command

$ hadoop fs -copyFromLocal <source file absolute location> <destination file location>

Program: WordCount.java

```
//package com.wordCount;
import java.io.IOException;
import java.util.StringTokenizer;
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

public class WordCount{
	public static class Map extends Mapper<Object, Text, Text, IntWritable>{
		private final static IntWritable one = new IntWritable(1);
		private Text word = new Text();

		public void map(Object key, Text value, Context context) throws
IOException, InterruptedException{
			StringTokenizer st = new StringTokenizer(value.toString());
			while(st.hasMoreTokens()){
				word.set(st.nextToken());
				context.write(word,one);
			}
		}
	}
```

```
    public static class Reduce extends Reducer<Text, IntWritable, Text,
IntWritable>{
        public void reduce(Text key, Iterable<IntWritable> values, Context
context) throws IOException,InterruptedException{
            int sum = 0;
            for(IntWritable val: values){
                sum += val.get();
            }
            context.write(key, new IntWritable(sum));
        }
    }

    public static void main(String[] args) throws Exception{
        Configuration conf = new Configuration();
        Job job = new Job(conf,"wordcount");
        job.setJarByClass(WordCount.class);
        job.setMapperClass(Map.class);
        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(IntWritable.class);
        job.setReducerClass(Reduce.class);
        //job.setInputFormatClass(TextInputFormat.class);
        //job.setOutputFormatClass(TextOutputFormat.class);
        FileInputFormat.addInputPath(job, new Path(args[0]));
        FileOutputFormat.setOutputPath(job, new Path(args[1]));
        job.waitForCompletion(true);
    }
}
```

3. Now run the WordCount.java file using command

$ java WordCount.java

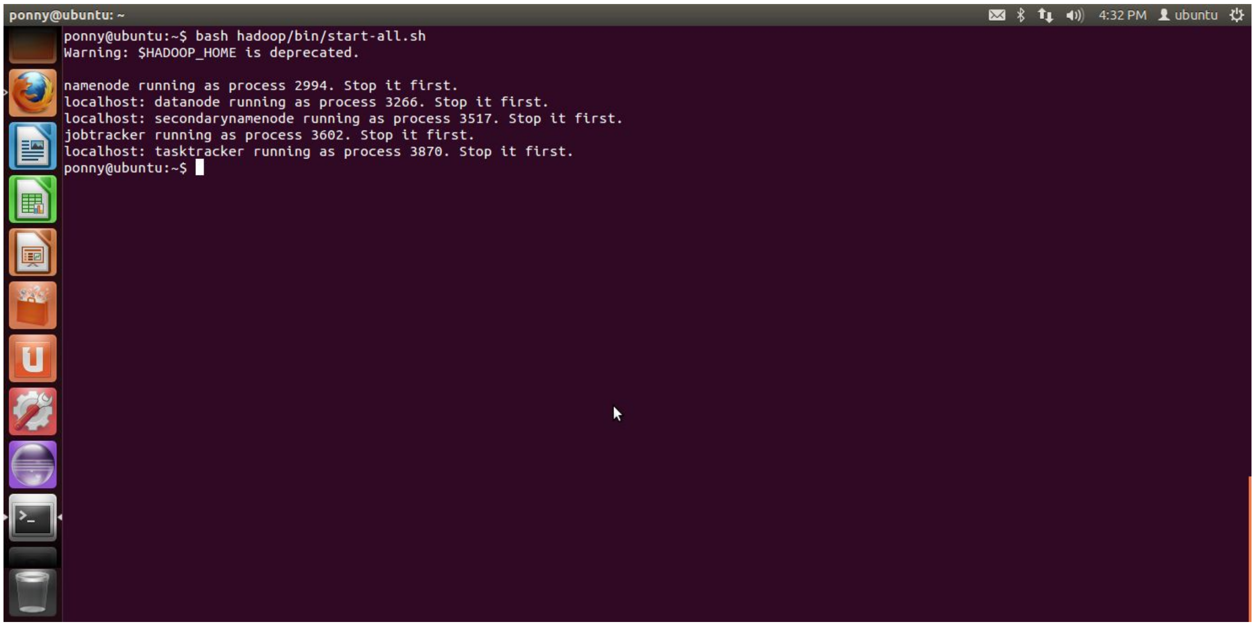
4. Make a jar file out of it using command

$ jar cf wc.jar WordCount*.class

5. Now run the wordcount mapreduce program on the text file uploaded on Hadoop using command

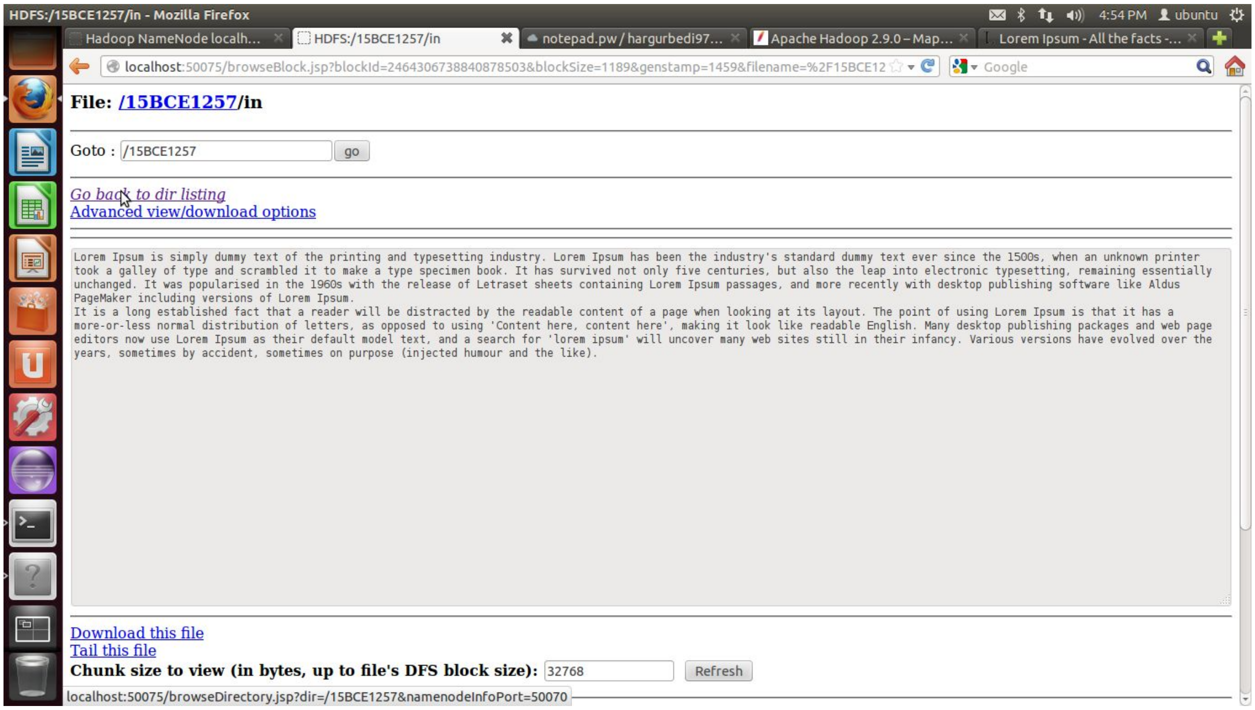
$ hadoop jar wc.jar WordCount <input textfile hadoop location> <output file location>

Output:



```
ponny@ubuntu:~$ bash hadoop/bin/start-all.sh
Warning: $HADOOP_HOME is deprecated.

namenode running as process 2994. Stop it first.
localhost: datanode running as process 3266. Stop it first.
localhost: secondarynamenode running as process 3517. Stop it first.
jobtracker running as process 3602. Stop it first.
localhost: tasktracker running as process 3870. Stop it first.
ponny@ubuntu:~$
```



HDFS:/15BCE1257/in - Mozilla Firefox
4:54 PM
ubuntu
Hadoop NameNode localh...
HDFS:/15BCE1257/in
notepad.pw / hargurbedi97...
Apache Hadoop 2.9.0 – Map...
Lorem Ipsum - All the facts -...
Google

localhost:50075/browseBlock.jsp?blockId=2464306738840878503&blockSize=1189&genstamp=1459&filename=%2F15BCE12

File: /15BCE1257/in

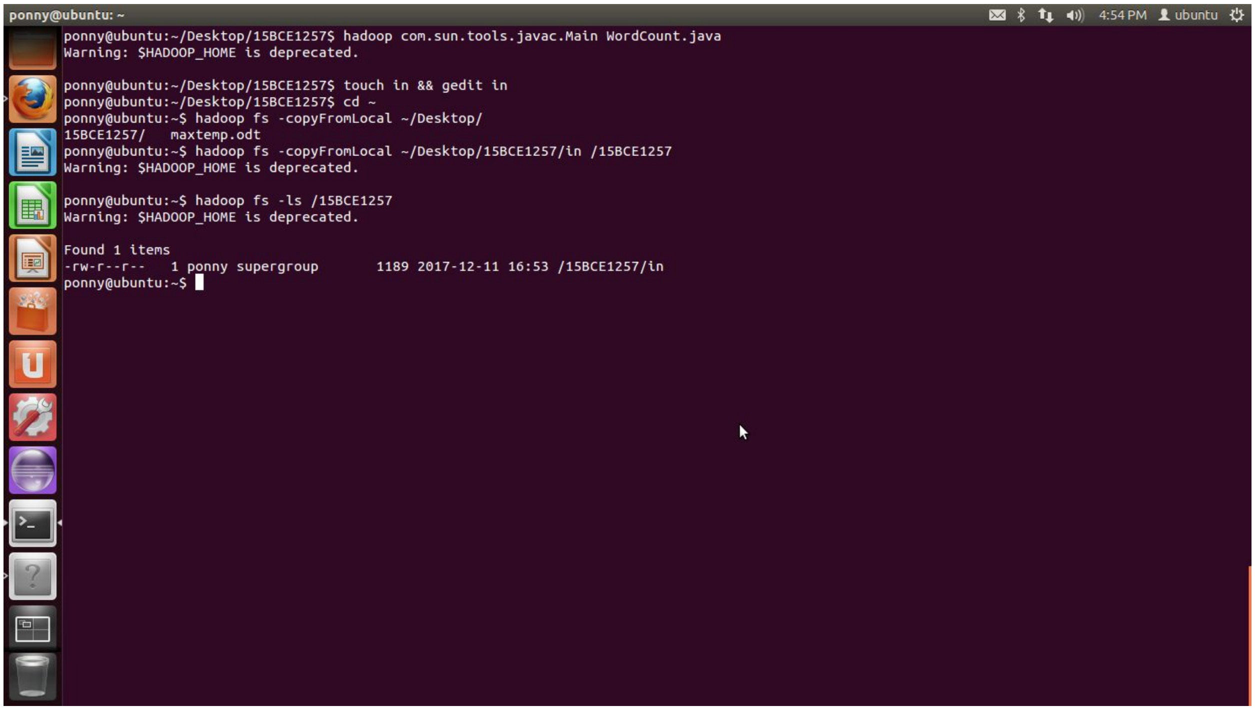
Goto : /15BCE1257 go

Go back to dir listing
Advanced view/download options

```
Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer
took a galley of type and scrambled it to make a type specimen book. It has survived not only five centuries, but also the leap into electronic typesetting, remaining essentially
unchanged. It was popularised in the 1960s with the release of Letraset sheets containing Lorem Ipsum passages, and more recently with desktop publishing software like Aldus
PageMaker including versions of Lorem Ipsum.
It is a long established fact that a reader will be distracted by the readable content of a page when looking at its layout. The point of using Lorem Ipsum is that it has a
more-or-less normal distribution of letters, as opposed to using 'Content here, content here', making it look like readable English. Many desktop publishing packages and web page
editors now use Lorem Ipsum as their default model text, and a search for 'lorem ipsum' will uncover many web sites still in their infancy. Various versions have evolved over the
years, sometimes by accident, sometimes on purpose (injected humour and the like).
```

Download this file
Tail this file
Chunk size to view (in bytes, up to file's DFS block size): 32768 Refresh

localhost:50075/browseDirectory.jsp?dir=/15BCE1257&namenodeInfoPort=50070

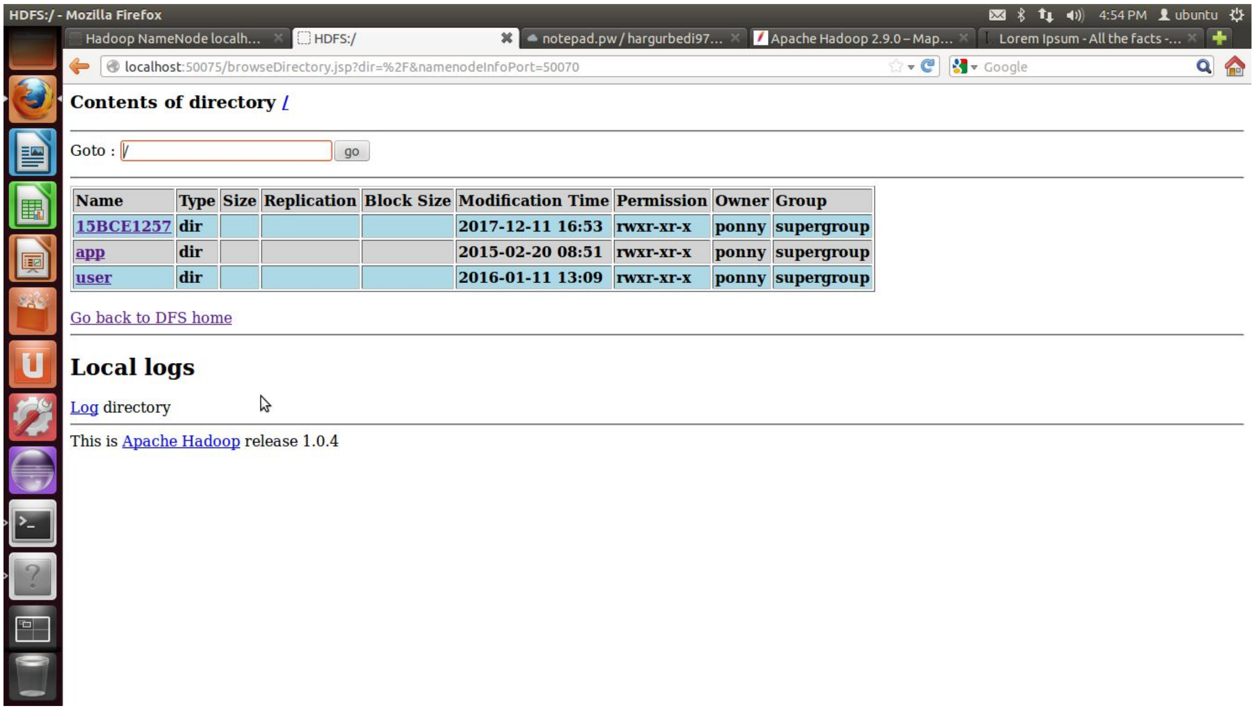


```
ponny@ubuntu:~/Desktop/15BCE1257$ hadoop com.sun.tools.javac.Main WordCount.java
Warning: $HADOOP_HOME is deprecated.

ponny@ubuntu:~/Desktop/15BCE1257$ touch in && gedit in
ponny@ubuntu:~/Desktop/15BCE1257$ cd ~
ponny@ubuntu:~$ hadoop fs -copyFromLocal ~/Desktop/
15BCE1257/   maxtemp.odt
ponny@ubuntu:~$ hadoop fs -copyFromLocal ~/Desktop/15BCE1257/in /15BCE1257
Warning: $HADOOP_HOME is deprecated.

ponny@ubuntu:~$ hadoop fs -ls /15BCE1257
Warning: $HADOOP_HOME is deprecated.

Found 1 items
-rw-r--r--   1 ponny supergroup       1189 2017-12-11 16:53 /15BCE1257/in
ponny@ubuntu:~$ 
```


HDFS:/ - Mozilla Firefox
4:54 PM
ubuntu
Hadoop NameNode localh...
HDFS:/
notepad.pw / hargurbedi97...
Apache Hadoop 2.9.0 – Map...
Lorem Ipsum - All the facts -...
localhost:50075/browseDirectory.jsp?dir=%2F&namenodeInfoPort=50070
Google

Contents of directory /

Goto : / go

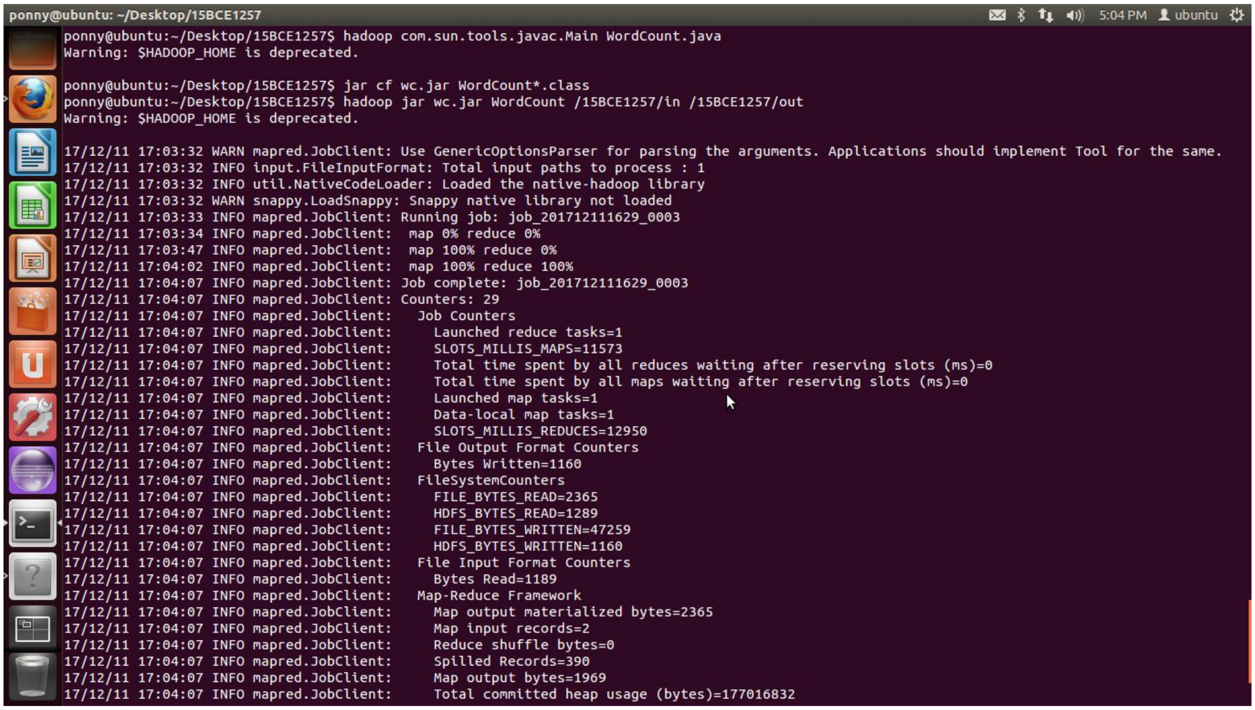
Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
15BCE1257	dir				2017-12-11 16:53	rwxr-xr-x	ponny	supergroup
app	dir				2015-02-20 08:51	rwxr-xr-x	ponny	supergroup
user	dir				2016-01-11 13:09	rwxr-xr-x	ponny	supergroup

Go back to DFS home

Local logs

Log directory

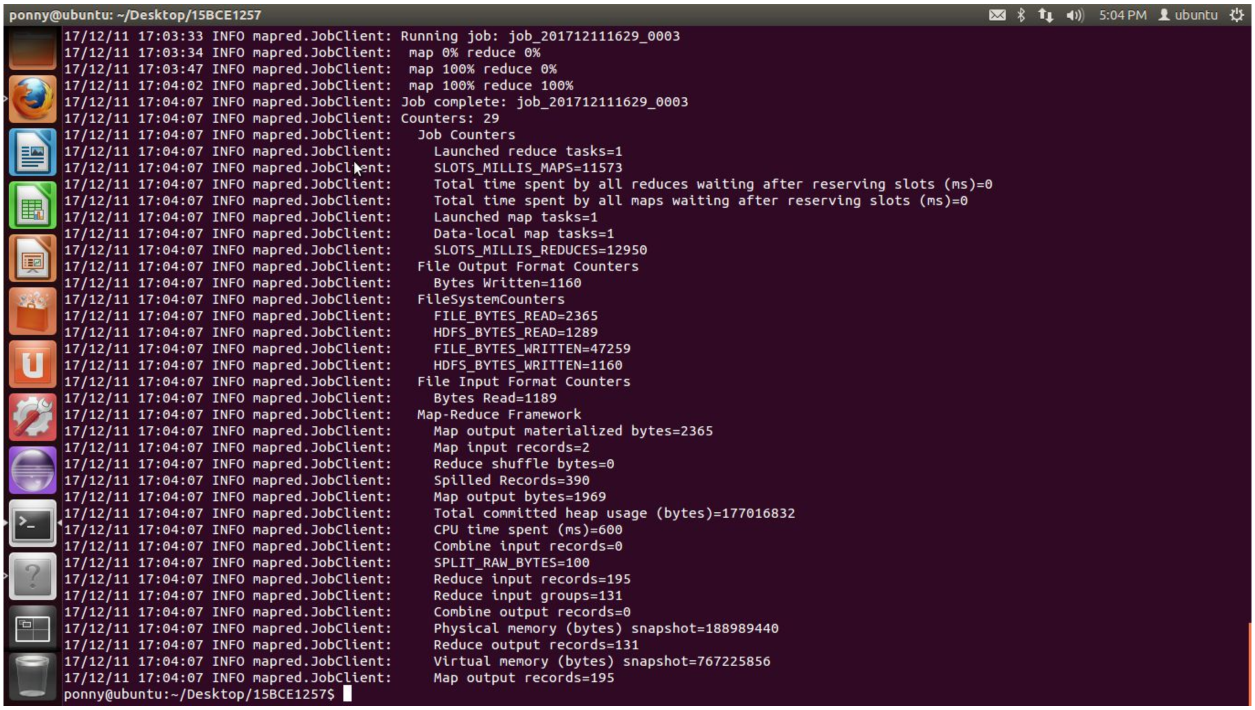
This is Apache Hadoop release 1.0.4



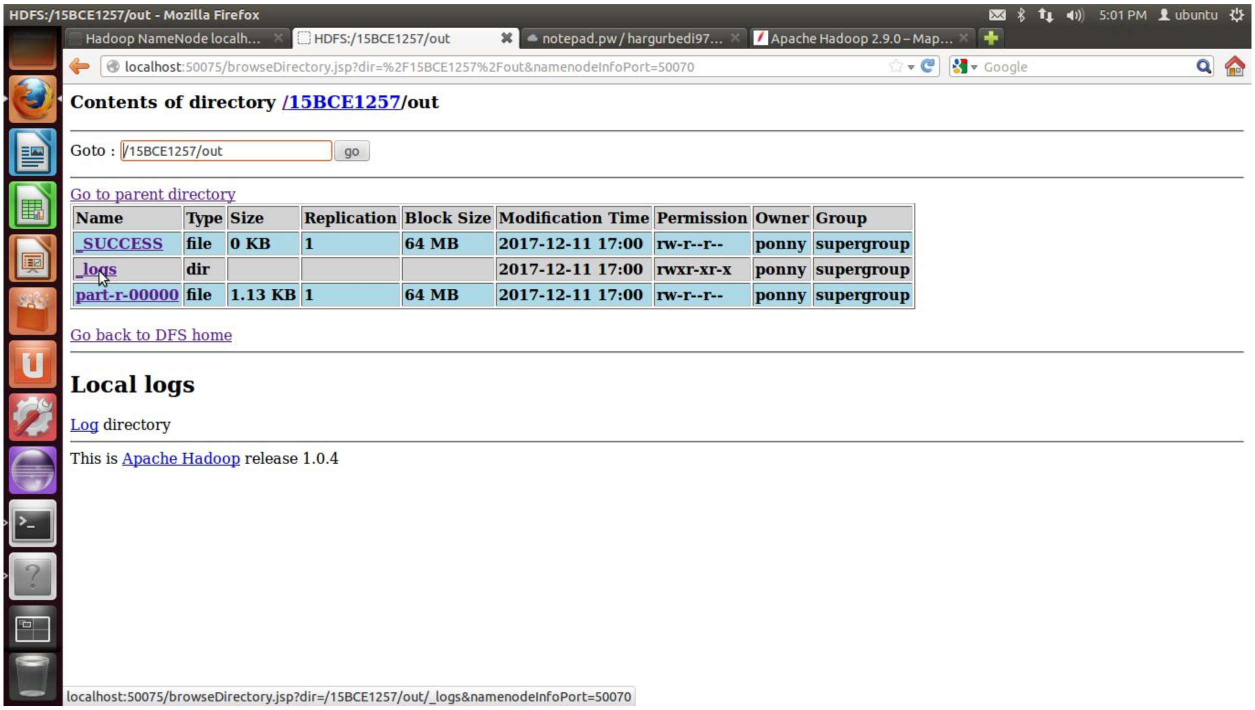
```
ponny@ubuntu:~/Desktop/15BCE1257$ hadoop com.sun.tools.javac.Main WordCount.java
Warning: $HADOOP_HOME is deprecated.

ponny@ubuntu:~/Desktop/15BCE1257$ jar cf wc.jar WordCount*.class
ponny@ubuntu:~/Desktop/15BCE1257$ hadoop jar wc.jar WordCount /15BCE1257/in /15BCE1257/out
Warning: $HADOOP_HOME is deprecated.

17/12/11 17:03:32 WARN mapred.JobClient: Use GenericOptionsParser for parsing the arguments. Applications should implement Tool for the same.
17/12/11 17:03:32 INFO input.FileInputFormat: Total input paths to process : 1
17/12/11 17:03:32 INFO util.NativeCodeLoader: Loaded the native-hadoop library
17/12/11 17:03:32 WARN snappy.LoadSnappy: Snappy native library not loaded
17/12/11 17:03:33 INFO mapred.JobClient: Running job: job_201712111629_0003
17/12/11 17:03:34 INFO mapred.JobClient:  map 0% reduce 0%
17/12/11 17:03:47 INFO mapred.JobClient:  map 100% reduce 0%
17/12/11 17:04:02 INFO mapred.JobClient:  map 100% reduce 100%
17/12/11 17:04:07 INFO mapred.JobClient: Job complete: job_201712111629_0003
17/12/11 17:04:07 INFO mapred.JobClient: Counters: 29
17/12/11 17:04:07 INFO mapred.JobClient:   Job Counters
17/12/11 17:04:07 INFO mapred.JobClient:     Launched reduce tasks=1
17/12/11 17:04:07 INFO mapred.JobClient:     SLOTS_MILLIS_MAPS=11573
17/12/11 17:04:07 INFO mapred.JobClient:     Total time spent by all reduces waiting after reserving slots (ms)=0
17/12/11 17:04:07 INFO mapred.JobClient:     Total time spent by all maps waiting after reserving slots (ms)=0
17/12/11 17:04:07 INFO mapred.JobClient:     Launched map tasks=1
17/12/11 17:04:07 INFO mapred.JobClient:     Data-local map tasks=1
17/12/11 17:04:07 INFO mapred.JobClient:     SLOTS_MILLIS_REDUCES=12950
17/12/11 17:04:07 INFO mapred.JobClient:   File Output Format Counters
17/12/11 17:04:07 INFO mapred.JobClient:     Bytes Written=1160
17/12/11 17:04:07 INFO mapred.JobClient:   FileSystemCounters
17/12/11 17:04:07 INFO mapred.JobClient:     FILE_BYTES_READ=2365
17/12/11 17:04:07 INFO mapred.JobClient:     HDFS_BYTES_READ=1289
17/12/11 17:04:07 INFO mapred.JobClient:     FILE_BYTES_WRITTEN=47259
17/12/11 17:04:07 INFO mapred.JobClient:     HDFS_BYTES_WRITTEN=1160
17/12/11 17:04:07 INFO mapred.JobClient:   File Input Format Counters
17/12/11 17:04:07 INFO mapred.JobClient:     Bytes Read=1189
17/12/11 17:04:07 INFO mapred.JobClient:   Map-Reduce Framework
17/12/11 17:04:07 INFO mapred.JobClient:     Map output materialized bytes=2365
17/12/11 17:04:07 INFO mapred.JobClient:     Map input records=2
17/12/11 17:04:07 INFO mapred.JobClient:     Reduce shuffle bytes=0
17/12/11 17:04:07 INFO mapred.JobClient:     Spilled Records=390
17/12/11 17:04:07 INFO mapred.JobClient:     Map output bytes=1969
17/12/11 17:04:07 INFO mapred.JobClient:     Total committed heap usage (bytes)=177016832
```



```
17/12/11 17:03:33 INFO mapred.JobClient: Running job: job_201712111629_0003
17/12/11 17:03:34 INFO mapred.JobClient:  map 0% reduce 0%
17/12/11 17:03:47 INFO mapred.JobClient:  map 100% reduce 0%
17/12/11 17:04:02 INFO mapred.JobClient:  map 100% reduce 100%
17/12/11 17:04:07 INFO mapred.JobClient: Job complete: job_201712111629_0003
17/12/11 17:04:07 INFO mapred.JobClient: Counters: 29
17/12/11 17:04:07 INFO mapred.JobClient:   Job Counters
17/12/11 17:04:07 INFO mapred.JobClient:     Launched reduce tasks=1
17/12/11 17:04:07 INFO mapred.JobClient:     SLOTS_MILLIS_MAPS=11573
17/12/11 17:04:07 INFO mapred.JobClient:     Total time spent by all reduces waiting after reserving slots (ms)=0
17/12/11 17:04:07 INFO mapred.JobClient:     Total time spent by all maps waiting after reserving slots (ms)=0
17/12/11 17:04:07 INFO mapred.JobClient:     Launched map tasks=1
17/12/11 17:04:07 INFO mapred.JobClient:     Data-local map tasks=1
17/12/11 17:04:07 INFO mapred.JobClient:     SLOTS_MILLIS_REDUCES=12950
17/12/11 17:04:07 INFO mapred.JobClient:   File Output Format Counters
17/12/11 17:04:07 INFO mapred.JobClient:     Bytes Written=1160
17/12/11 17:04:07 INFO mapred.JobClient:   FileSystemCounters
17/12/11 17:04:07 INFO mapred.JobClient:     FILE_BYTES_READ=2365
17/12/11 17:04:07 INFO mapred.JobClient:     HDFS_BYTES_READ=1289
17/12/11 17:04:07 INFO mapred.JobClient:     FILE_BYTES_WRITTEN=47259
17/12/11 17:04:07 INFO mapred.JobClient:     HDFS_BYTES_WRITTEN=1160
17/12/11 17:04:07 INFO mapred.JobClient:   File Input Format Counters
17/12/11 17:04:07 INFO mapred.JobClient:     Bytes Read=1189
17/12/11 17:04:07 INFO mapred.JobClient:   Map-Reduce Framework
17/12/11 17:04:07 INFO mapred.JobClient:     Map output materialized bytes=2365
17/12/11 17:04:07 INFO mapred.JobClient:     Map input records=2
17/12/11 17:04:07 INFO mapred.JobClient:     Reduce shuffle bytes=0
17/12/11 17:04:07 INFO mapred.JobClient:     Spilled Records=390
17/12/11 17:04:07 INFO mapred.JobClient:     Map output bytes=1969
17/12/11 17:04:07 INFO mapred.JobClient:     Total committed heap usage (bytes)=177016832
17/12/11 17:04:07 INFO mapred.JobClient:     CPU time spent (ms)=600
17/12/11 17:04:07 INFO mapred.JobClient:     Combine input records=0
17/12/11 17:04:07 INFO mapred.JobClient:     SPLIT_RAW_BYTES=100
17/12/11 17:04:07 INFO mapred.JobClient:     Reduce input records=195
17/12/11 17:04:07 INFO mapred.JobClient:     Reduce input groups=131
17/12/11 17:04:07 INFO mapred.JobClient:     Combine output records=0
17/12/11 17:04:07 INFO mapred.JobClient:     Physical memory (bytes) snapshot=188989440
17/12/11 17:04:07 INFO mapred.JobClient:     Reduce output records=131
17/12/11 17:04:07 INFO mapred.JobClient:     Virtual memory (bytes) snapshot=767225856
17/12/11 17:04:07 INFO mapred.JobClient:     Map output records=195
ponny@ubuntu:~/Desktop/15BCE1257$ 
```



HDFS:/15BCE1257/out - Mozilla Firefox
5:01 PM ubuntu
Hadoop NameNode localh...
HDFS:/15BCE1257/out
notepad.pw / hargurbedi97...
Apache Hadoop 2.9.0 – Map...
localhost:50075/browseDirectory.jsp?dir=%2F15BCE1257%2Fout&namenodeInfoPort=50070
Google

Contents of directory /15BCE1257/out

Goto : /15BCE1257/out go

Go to parent directory

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
_SUCCESS	file	0 KB	1	64 MB	2017-12-11 17:00	rw-r--r--	ponny	supergroup
_logs	dir				2017-12-11 17:00	rwxr-xr-x	ponny	supergroup
part-r-00000	file	1.13 KB	1	64 MB	2017-12-11 17:00	rw-r--r--	ponny	supergroup

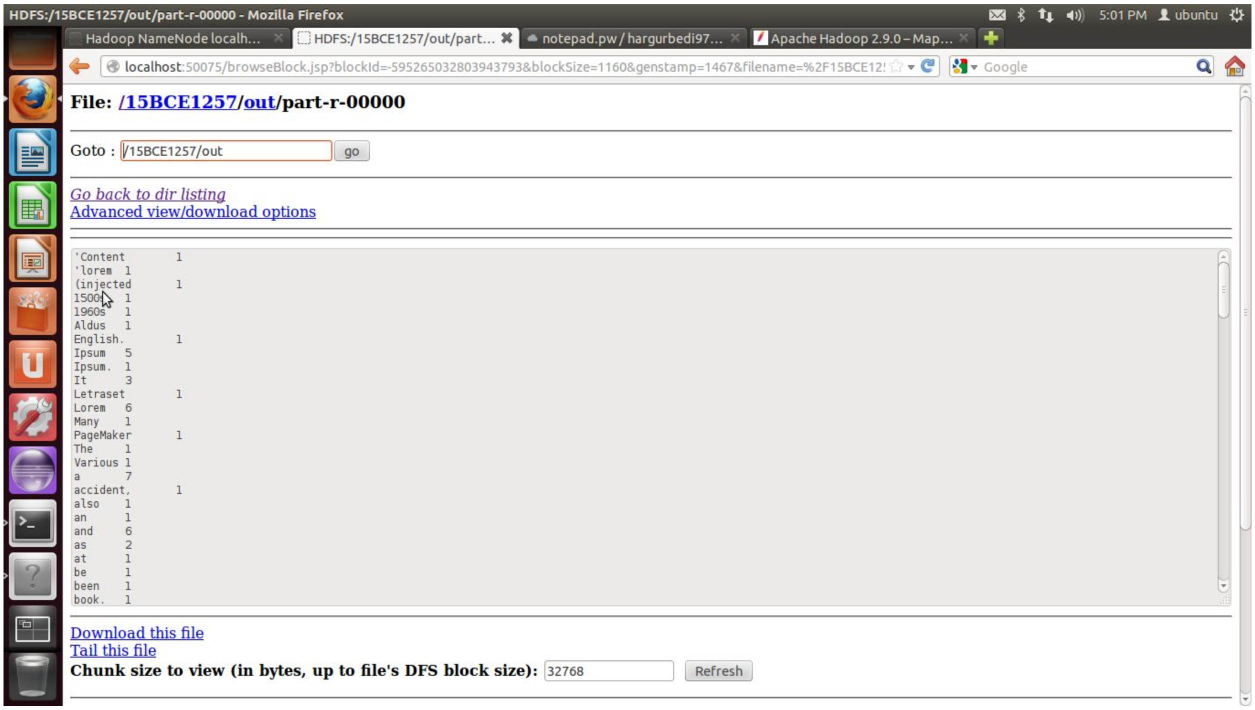
Go back to DFS home

Local logs

Log directory

This is Apache Hadoop release 1.0.4

localhost:50075/browseDirectory.jsp?dir=/15BCE1257/out/_logs&namenodeInfoPort=50070



HDFS:/15BCE1257/out/part-r-00000 - Mozilla Firefox
5:01 PM ubuntu
Hadoop NameNode localh...
HDFS:/15BCE1257/out/part...
notepad.pw / hargurbedi97...
Apache Hadoop 2.9.0 – Map...
localhost:50075/browseBlock.jsp?blockId=-595265032803943793&blockSize=1160&genstamp=1467&filename=%2F15BCE12
Google

File: /15BCE1257/out/part-r-00000

Goto : /15BCE1257/out go

Go back to dir listing
Advanced view/download options

```
'Content	1
'lorem	1
(injected	1
1500s	1
1960s	1
Aldus	1
English.	1
Ipsum	5
Ipsum.	1
It	3
Letraset	1
Lorem	6
Many	1
PageMaker	1
The	1
Various	1
a	7
accident,	1
also	1
an	1
and	6
as	2
at	1
be	1
been	1
book.	1
```

Download this file
Tail this file
Chunk size to view (in bytes, up to file's DFS block size): 32768 Refresh