**Documentation of Split by - Map reduce**

-Prepared by Vignesh.R

The previous section discussed the use of Reducers, which receive keys and the list of values for those keys. The keys are sorted from the perspective of the Reducer. Although for simplicity we assume a single Reducer, it is not a good choice in practice for performance reasons. In practice, it is common to use tens or even hundreds of Reducers. The Hadoop API guarantees that the Reducer processes the key-value pairs received by it, sorted by the key. This does not imply that Reducer-1 will have keys which are lower in sort order to keys received by Reducer-0. By default, the keys are randomly allocated to reach the Reducer from the Mapper. However, this random allocation is consistent. The same key is always routed to the same Reducer, regardless of the Mapper it comes from. Within the Reducer, the key and value pairs received from the Mapper are processed in a sorted (by Reducer input key or Mapper output key) order.

**Creating .jar file for Splitout and exporting it:-**

**Step 1:**

Open Eclipse and Click on File > New > Java Project.

**Step 2:**

Give the name ‘Splitout’ as project name and click ‘Finish’.

**Step 3:**

Right click on ‘Splitout’ project and select ‘Properties’. Click ‘Java Build Path’ and switch to Libraries tab and click on ‘Add external JARs’.

**Step 4:**

Select all the JAR files in usr >> lib >> hadoop directory to add them.

**Step 5:**

Again add all jar files in usr >> lib >> hadoop >> client directory and press OK.

**Step 6:**

**Right click on src, New >>Class.**

**Step 7:**

Enter the project name as ‘SplitoutMRJob’ and click ‘Finish’.

**Step 8:**

Open browser and copy and paste the Java Source code of Splitout program from the link given. The packages are automatically generated by the Eclipse.

**Website link:**<https://github.com/Apress/pro-apache-hadoop/blob/master/prohadoop/src/main/java/org/apress/prohadoop/c5/SplitByMonthMRJob.java>

**Website link:**<https://github.com/Apress/pro-apache-hadoop/blob/master/prohadoop/src/main/java/org/apress/prohadoop/utils/AirlineDataUtils.java>

**Website link:**<https://github.com/Apress/pro-apache-hadoop/blob/master/prohadoop/src/main/java/org/apress/prohadoop/c6/MonthDoWOnlyWritable.java>

**Website link:**<https://github.com/Apress/pro-apache-hadoop/blob/master/prohadoop/src/main/java/org/apress/prohadoop/c6/MonthDoWWritable.java>

**Website link:**<https://github.com/Apress/pro-apache-hadoop/blob/master/prohadoop/src/main/java/org/apress/prohadoop/c6/DelaysWritable.java>

**Step 9:**

Right click on the SplitoutMRJava project and select Export >> Java >> JAR file. Then click on ‘Next’.

**Step 10:**

Name the JAR file and click ‘Finish’.

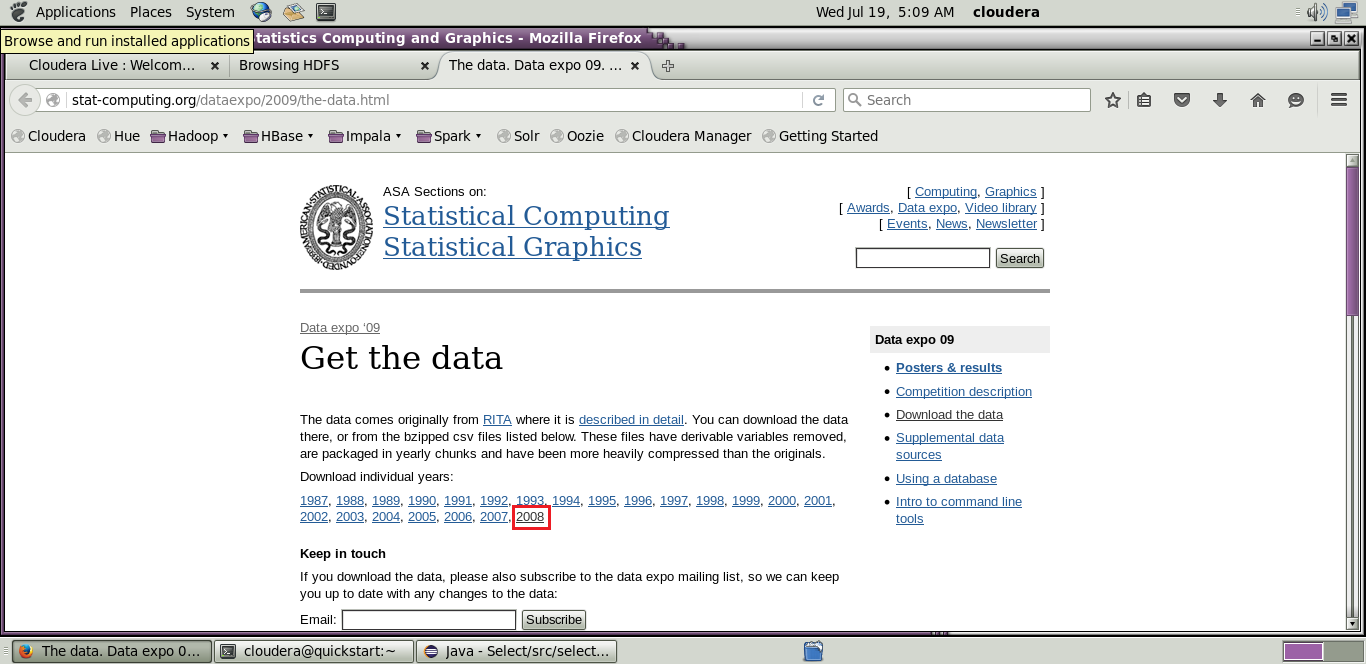
**Creating a file for Mapreduce job to work on:**

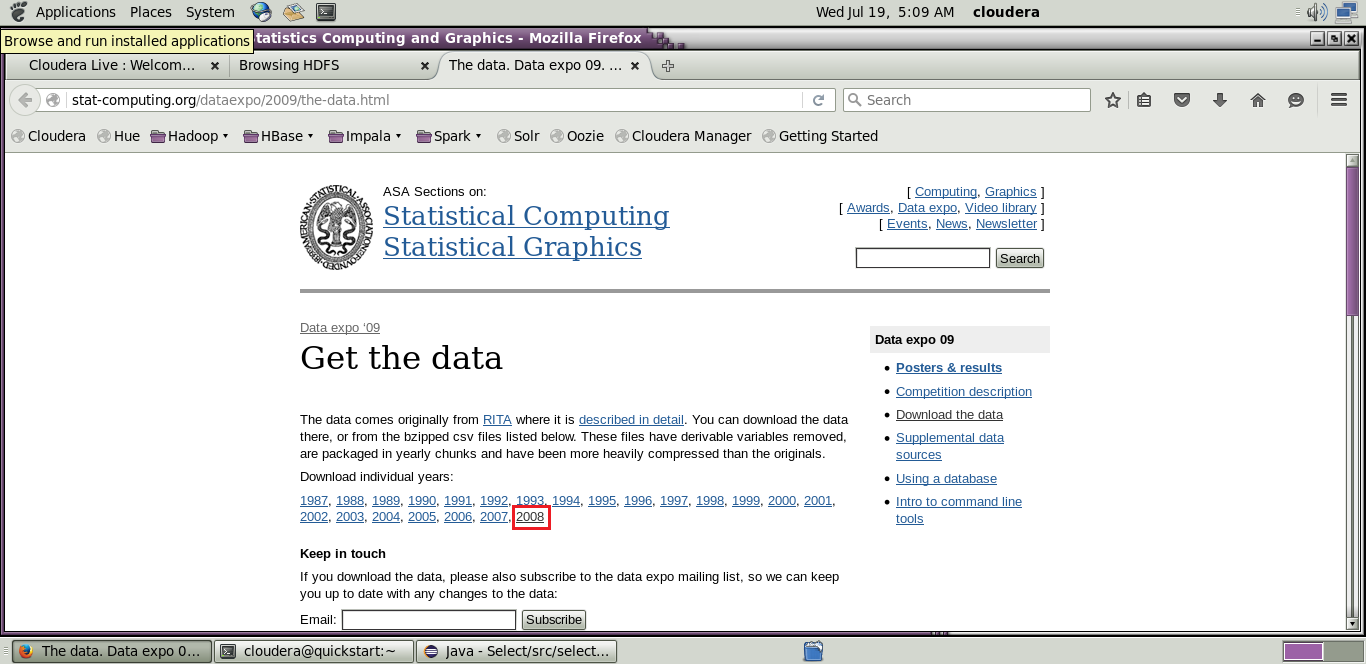
**Step 11:**Make a new Directory using the following command.

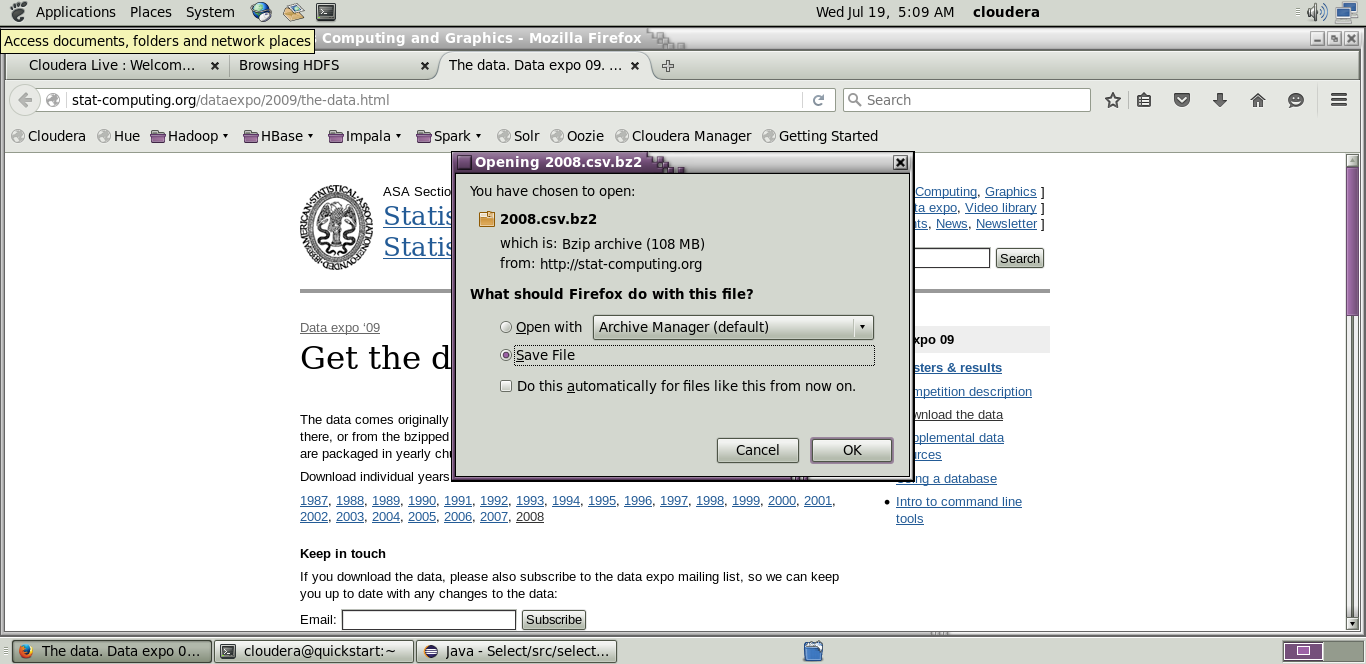
**Code:hadoop fs -mkdir /airline**

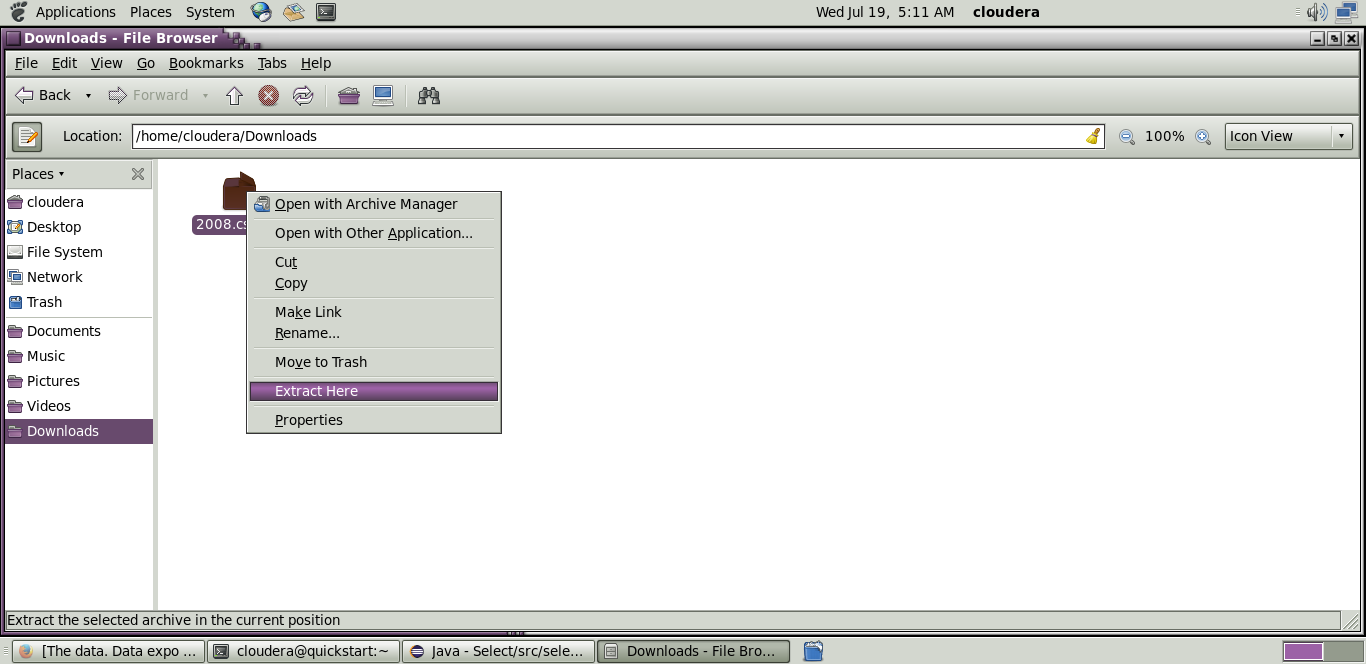
**Step 12:**Use the following link to download the csv file.

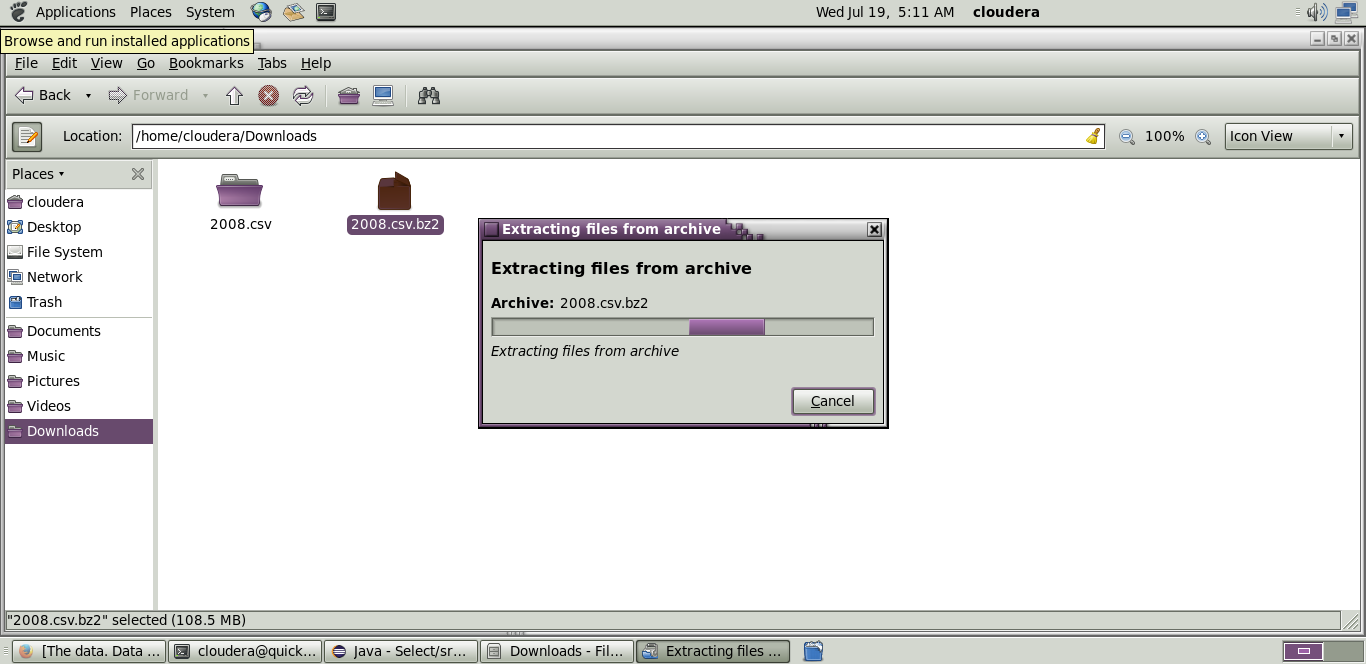
**Website link:**

[http://stat-computing.org/dataexpo/2009/the-data.html](http://stat-computing.org/dataexpo/2009/the-data.html )

[](http://stat-computing.org/dataexpo/2009/the-data.html )







**Step 13:** Copy the downloaded file to the new directory created in HDFS.

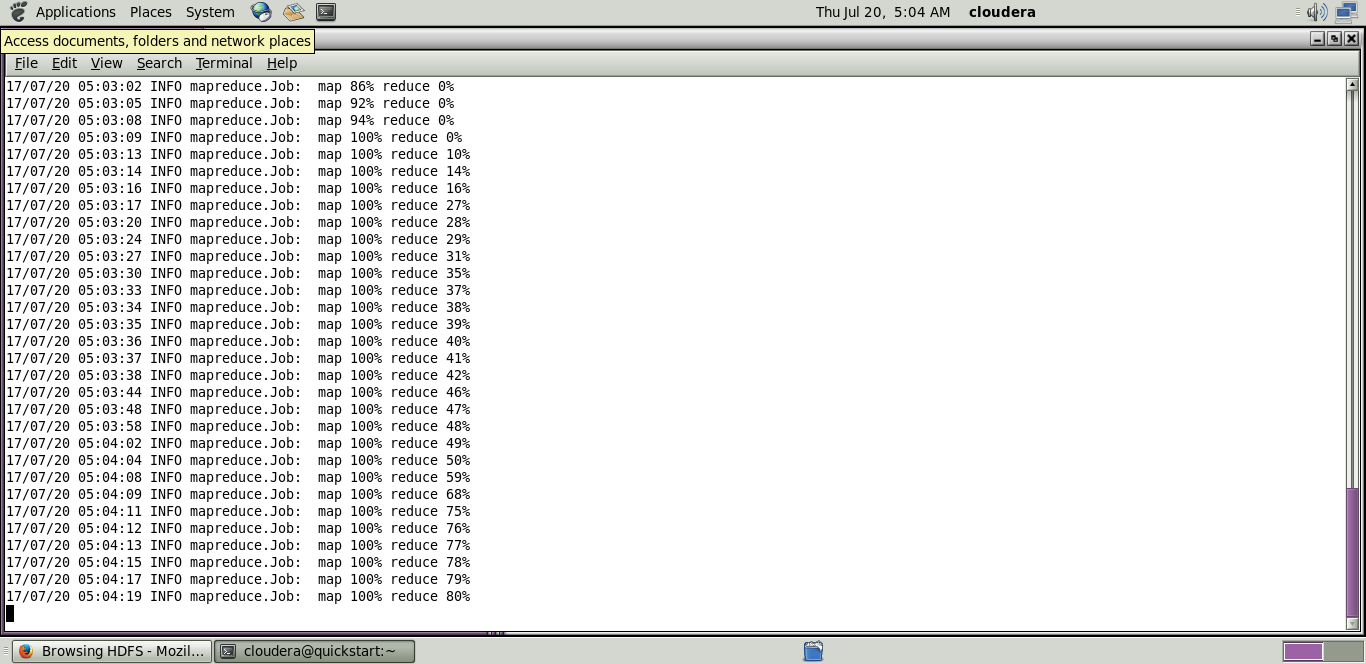
**Code: hadoop fs -put /home/cloudera/Downloads/2008.csv/airline**

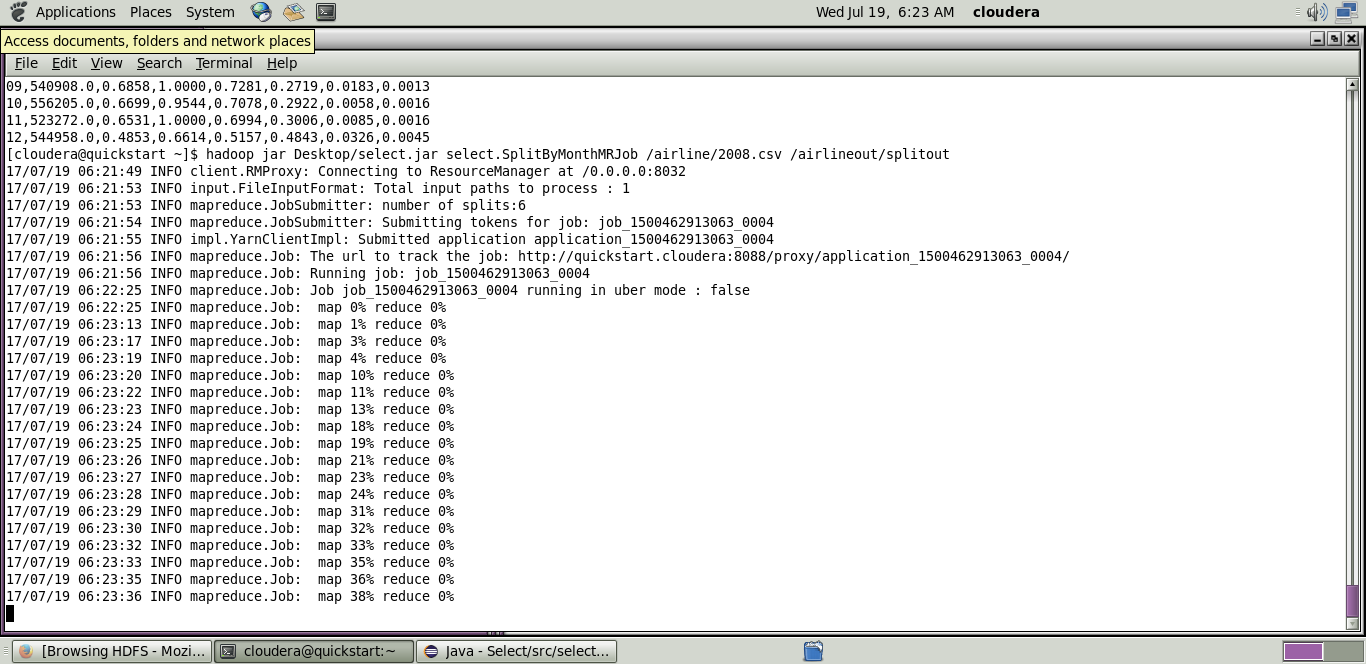
**Initializing mapreduce job:**

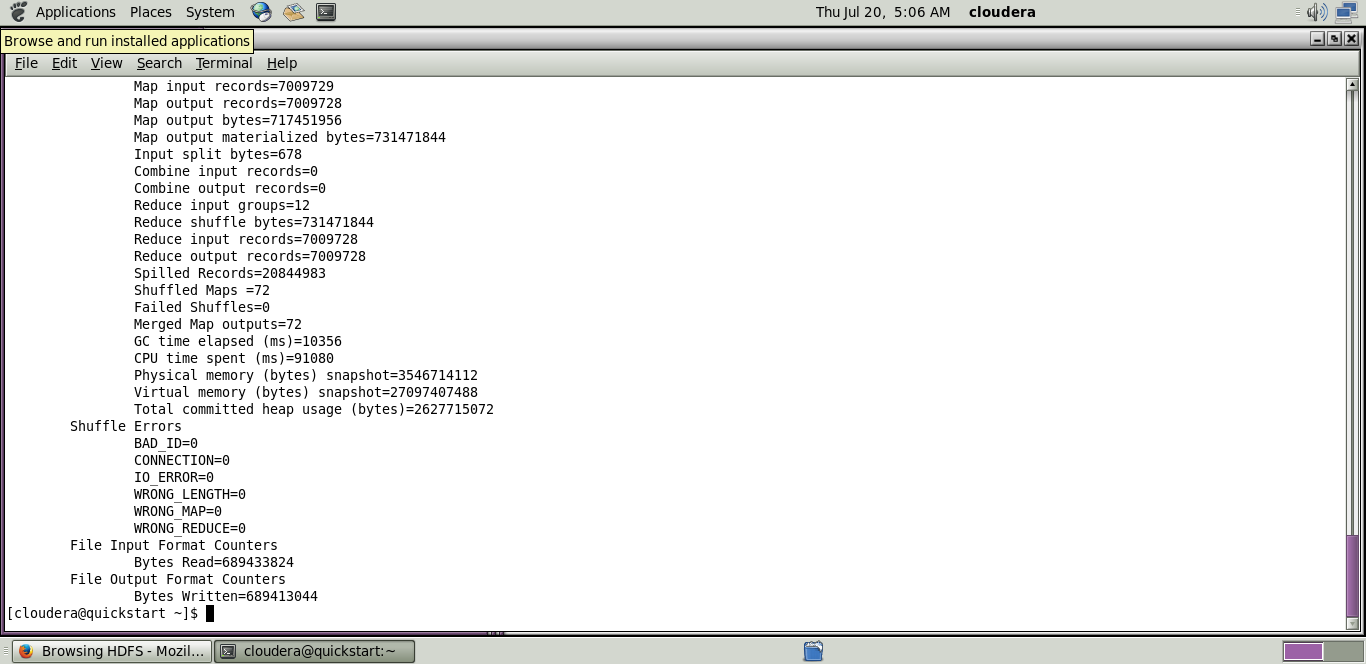
**Step 14:**

Initialize the mapreduce job by giving the following command and wait for sometime.

**Code:** hadoop jar Desktop/select.jar select.SplitByMonthMRJob /airline/2008.csv /airlineout/splitout







Now wait for about 50-70 seconds while the mapreduce job is being performed for the data created earlier.

**Output mapreduce job:**

**Step 15:**

The output directory of the mapreduce program is listed using the following command.

**Code:** hadoop fs -ls/airlineout/splitout/\*

**Step 16:**

The final output of the mapreduce program is found using the following command.

**Code:**hadoop fs -cat /airlineout/splitout/\*