**Documentation of Aggregation with combiner - Map reduce**

-Prepared by Vignesh.R

When Combiner is used, we have to follow the constraints that its use requires, which usually leads to modifying the output key and value classes of the Mapper, Combiner, and Reducer. The key constraint enforced by the Combiner is that the Combiner class’s input and output key and value classes must match the corresponding output key and value classes of the Mapper.

The Combiner output is also sent to the Reducer. The Reducer does not know whether the input it is receiving is coming directly from the Mapper or processed by a Combiner en route, so the output key and value classes of the Mapper should match the input key and value classes of the Combiner. Similarly, the Output key and value classes of the Combiner must match the input key and value classes of the Reducer. This naturally implies that the Combiner output key and value class must be identical to the output key and value classes of the Mapper, so we have to rewrite the Mapper and Reducer a bit. Earlier, the Aggregation Mapper simply had to send an IntWritable marker along with the key, which is month. In the Reducer, we checked for the existence of the marker and incremented the relevant variable by 1 of the respective marker (for example, ARRIVAL\_ON\_TIME). But with the user of the Combiner, it is not enough to send this marker. We need to send the aggregation (SUM) value along with the marker type, which is the partial sum computed at the end of each Mapper. The output of Aggregation Mapper is MapWritable.

**Creating .jar file for Aggregation with combiner and exporting it:-**

**Step 1:**

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

**Step 2:**

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

**Step 3:**

Right click on ‘Aggregationwithcombiner’ 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 ‘AggregationWithCombinerMRJob’ and click ‘Finish’.

**Step 8:**

Open browser and copy and paste the Java Source code of Aggregation with combiner 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/AggregationWithCombinerMRJob.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 AggregationWithCombiner Java project and select Export >> Java >> JAR file. Then click on ‘Next’.

**Step 10:**

Name the JAR file and click ‘Finish’.

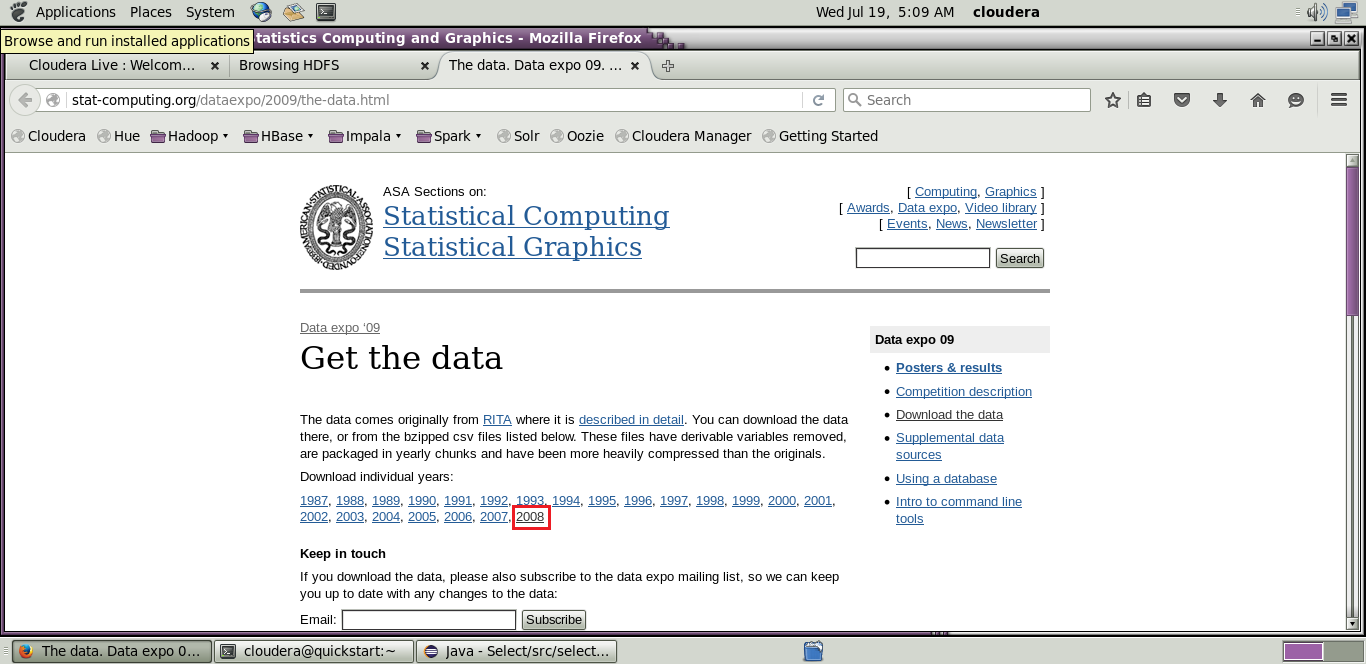
**Creating a file for Mapreduce job to work on:**

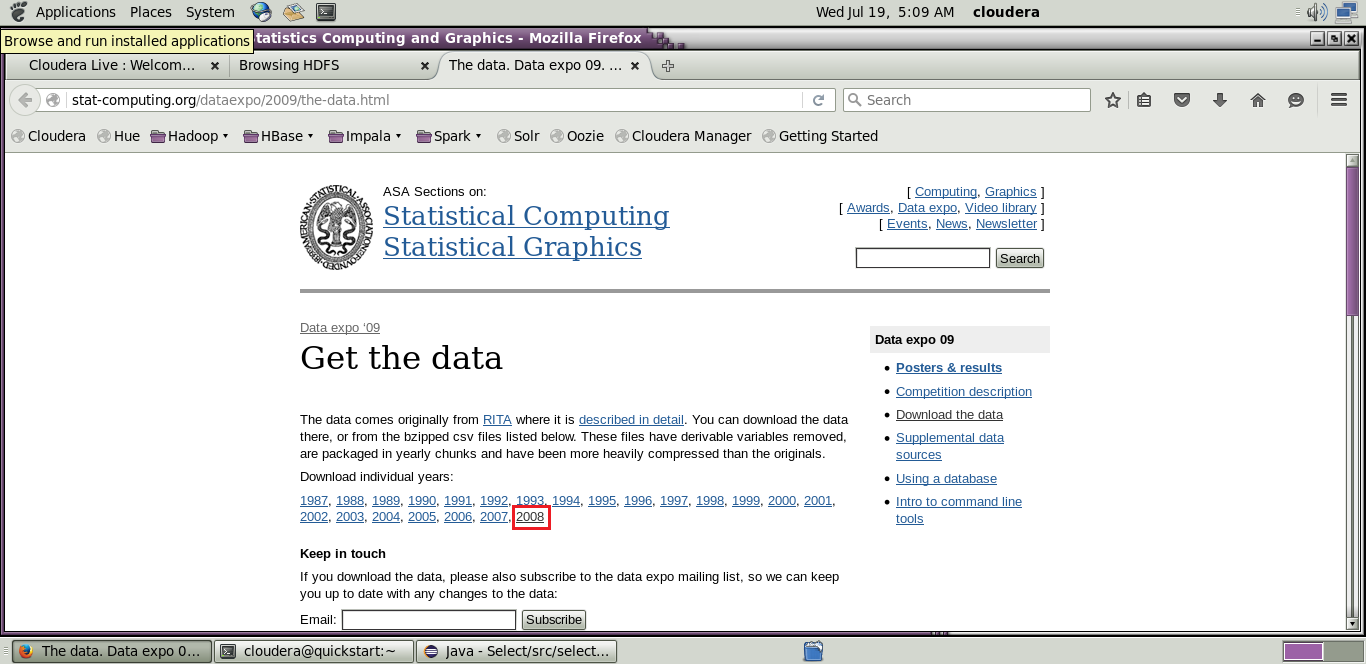
**Step 11:**Make a new Directory using the following command.

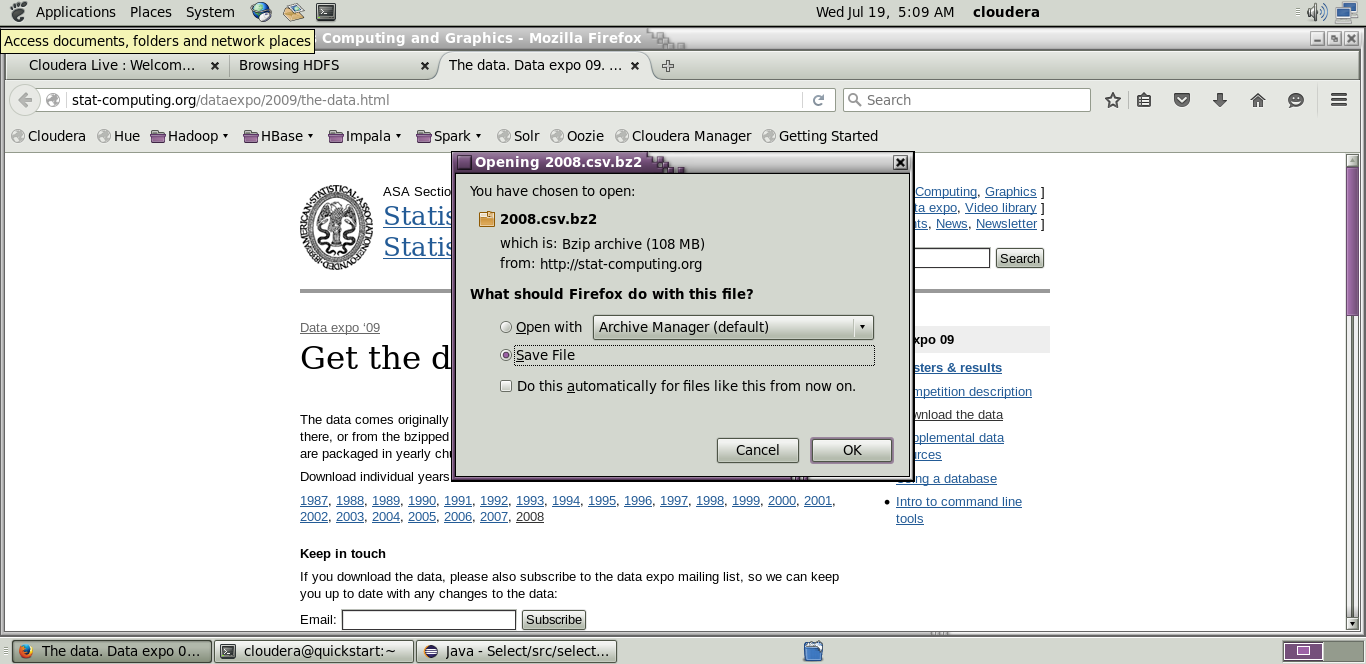
**Code:hadoop fs -mkdir /airline**

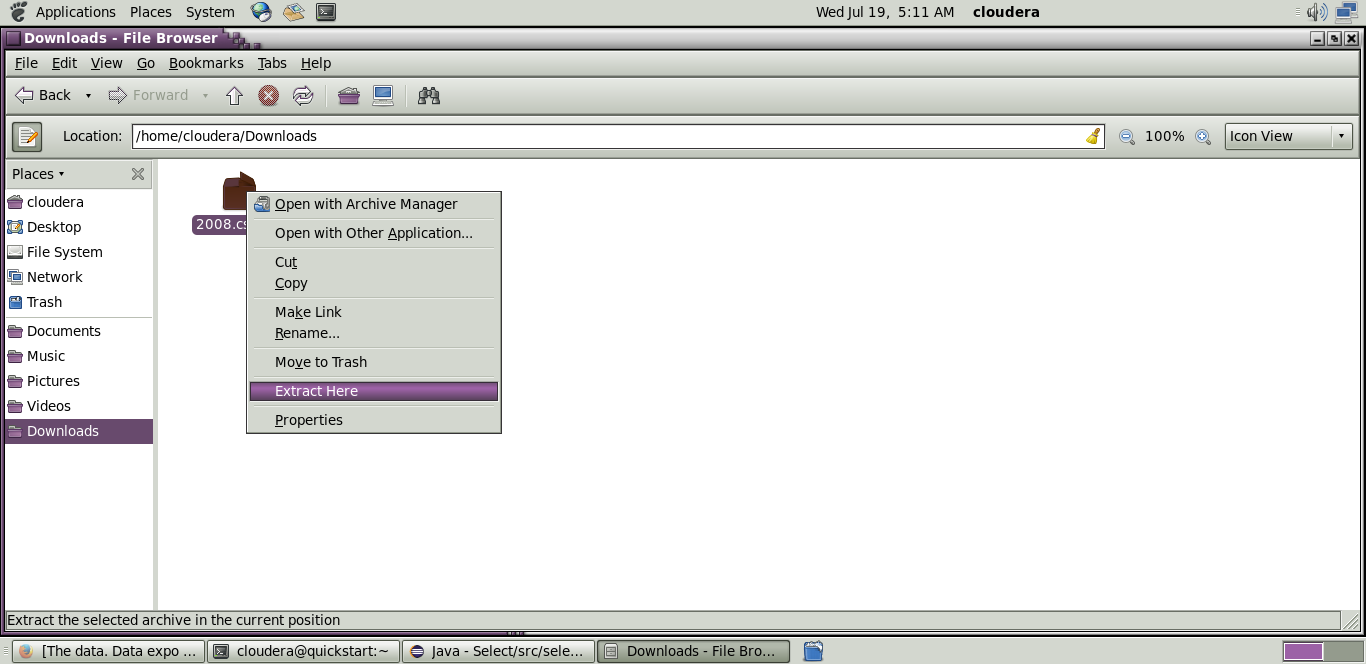
**Step 12:**Use the following link to download the csv file.

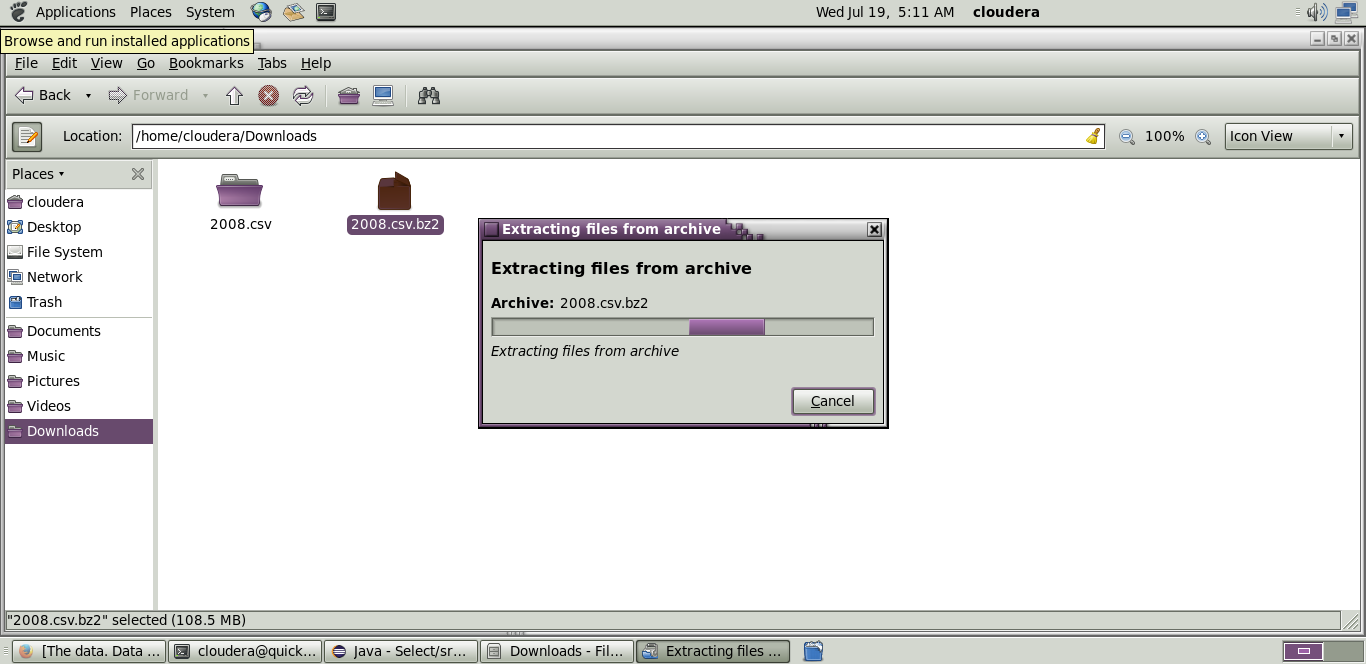
**Website link:**

[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.

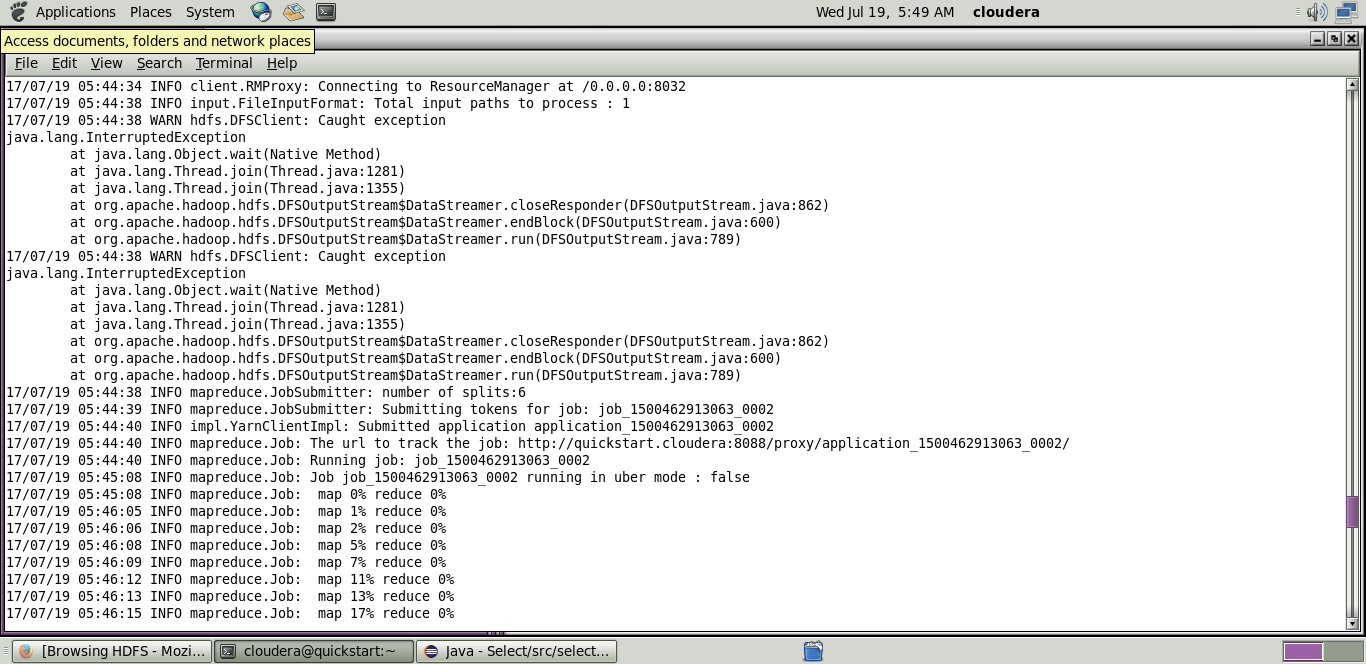
**Code: hadoop fs -put /home/cloudera/Downloads/2008.csv/airline**

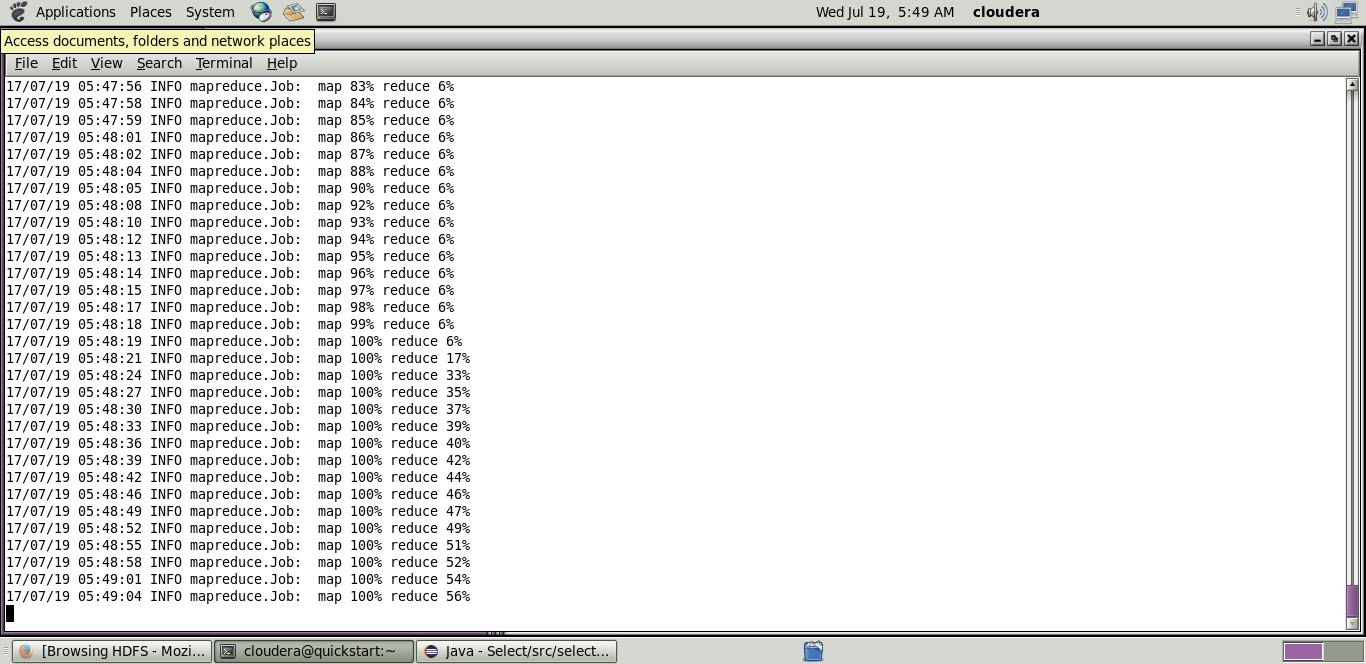
**Initializing mapreduce job:**

**Step 14:**

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

**Code:**hadoop jar Desktop/select.jar select.AggregationWithCombinerMRJob /airline/2008.csv /airlineout/aggrcombout





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/aggrcombout/\*

**Step 16:**

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

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