**Documentation of Select Clause - Map reduce**

-Prepared by Vignesh.R

SELECT: Selects a subset of columns from a large number of columns in a table

Problem Definition: SELECT Clause

From the airline dataset, we want to produce the output for the entire dataset, which comprises only the following attributes:

* Flight date in MM/DD/YYYYformat (for example, 01/13/1987, 03/28/1988)
* Day of the week
* Departure time
* Arrival time
* Origin airport code
* Destination airport code
* Total flight distance in miles
* Actual flight time
* Scheduled flight time
* Departure delay
* Arrival delay

This is a simple SELECT clause on the table with some basic computations:

Flight date is a combination of the Month, Day of Month, and Year columns formatted in specified manner.

The rest of the fields are all simple as-is projections of the corresponding fields.

**Creating.jarfile for Select clause and exporting it:-**

**Step 1:**

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

**Step 2:**

Give the name ‘Select’ as your project name and click ‘Finish’.

**Step 3:**

Right click on ‘Select’ 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 ‘SelectClauseMRJob’ and click ‘Finish’.

**Step 8:**

Open browser and copy and paste the Java Source code of SelectClause 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/SelectClauseMRJob.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 SelectClauseMRJava project and select Export >> Java >> JAR file. Then click on ‘Next’.

**Step 10:**

Name the JAR file and click ‘Finish’.

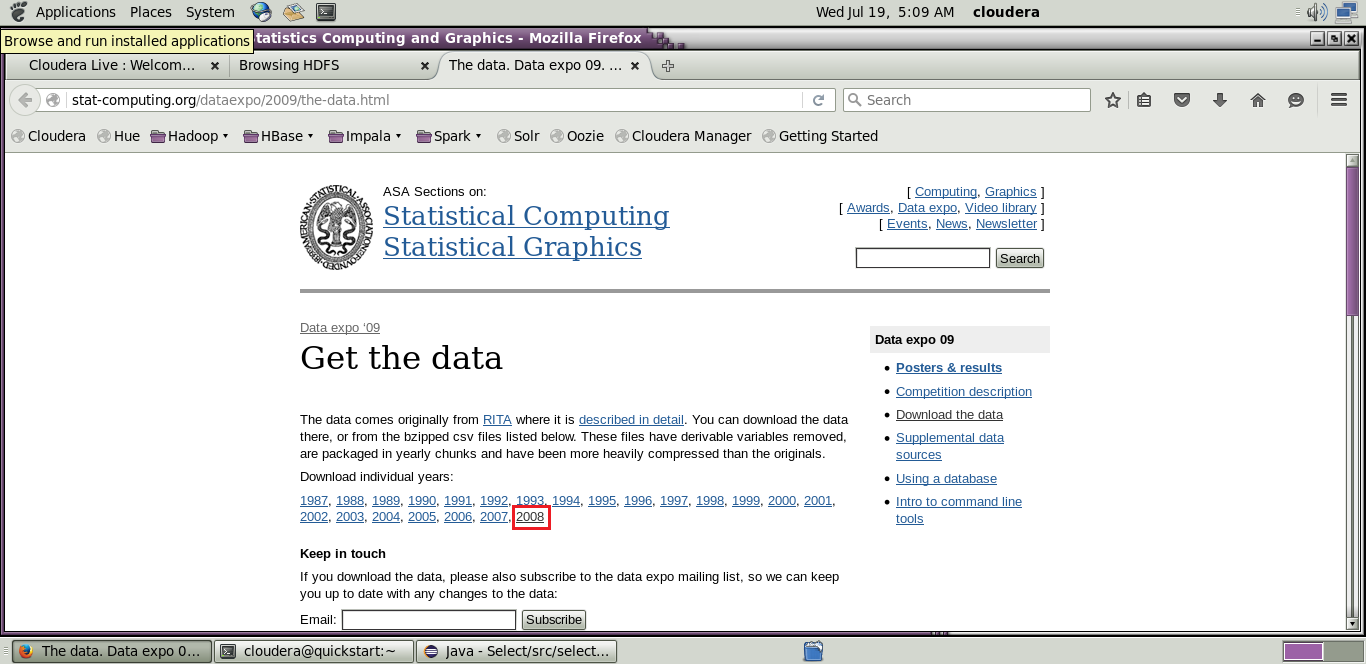
**Creating a file for Mapreduce job to work on:**

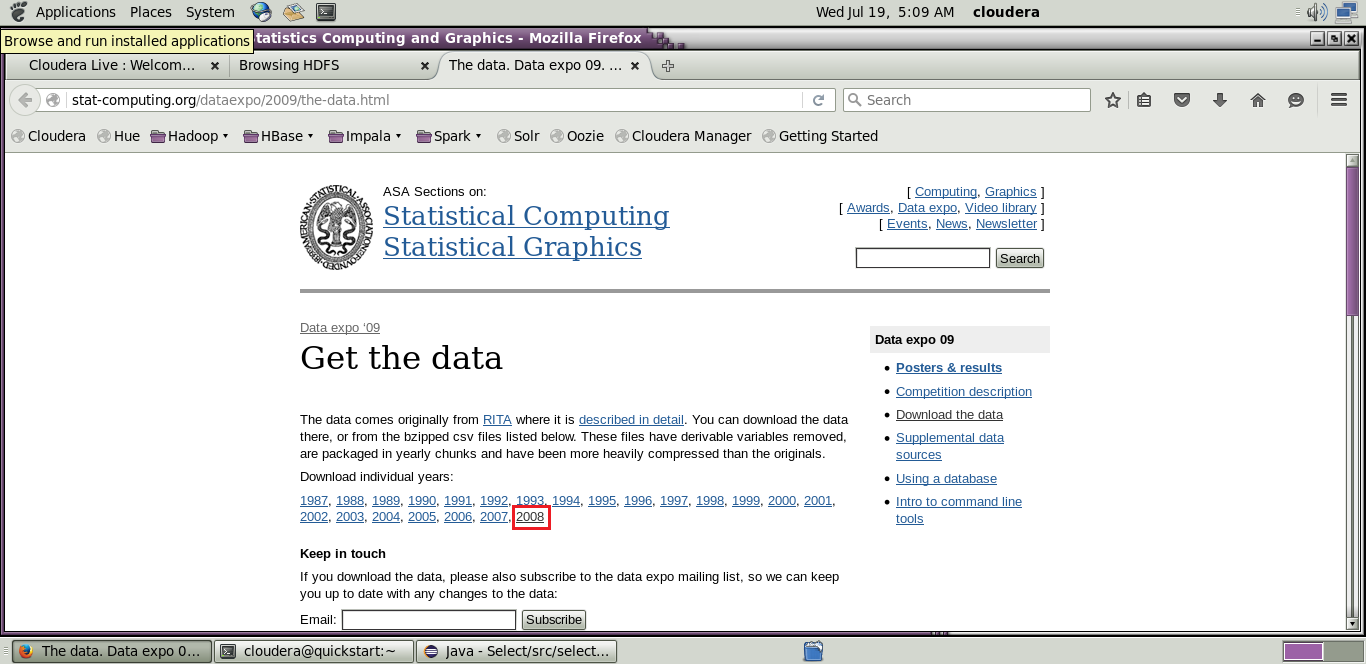
**Step 11:**Make a new Directory using the following command.

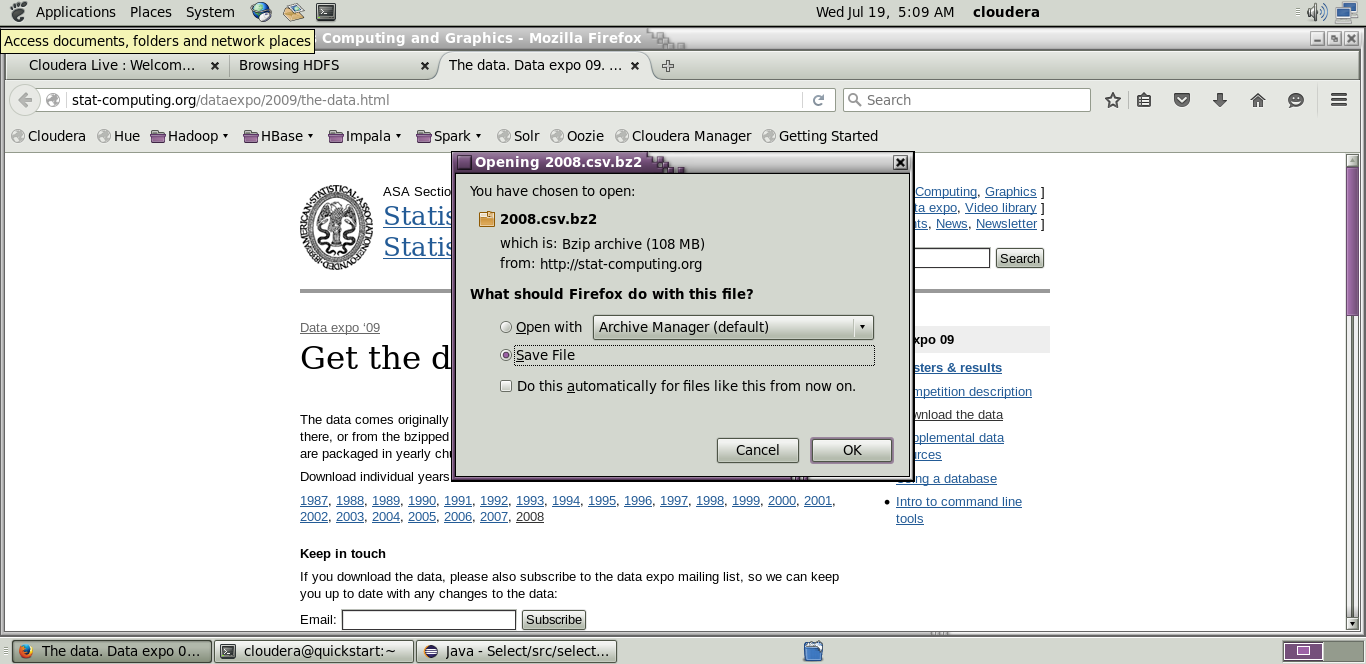
**Code:hadoop fs -mkdir /airline**

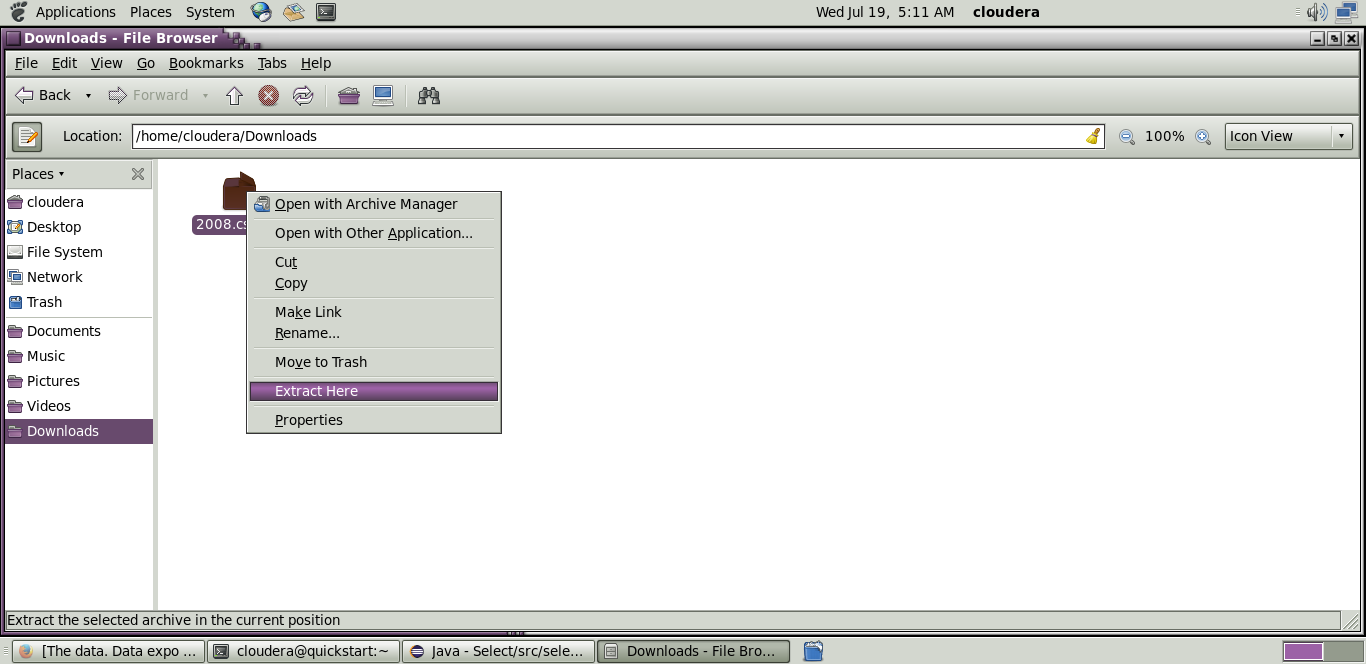
**Step 12:**Use the following link to download the csv file.

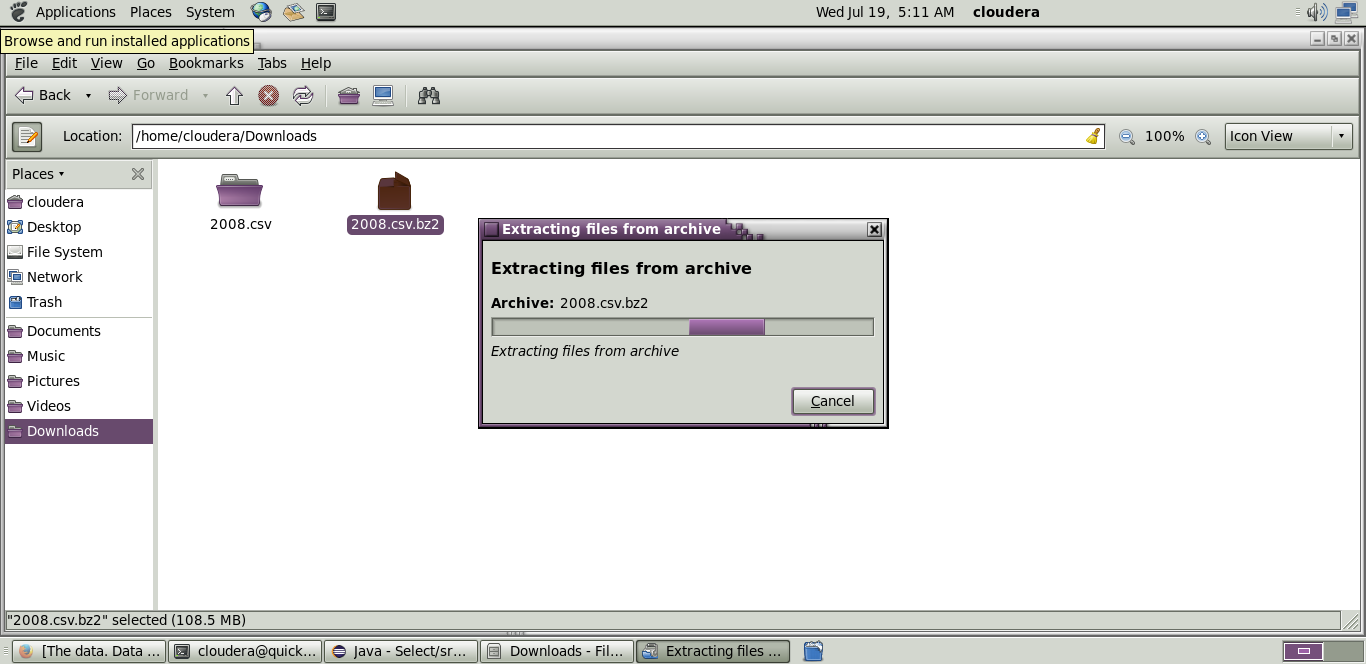
**Website link:**

[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.

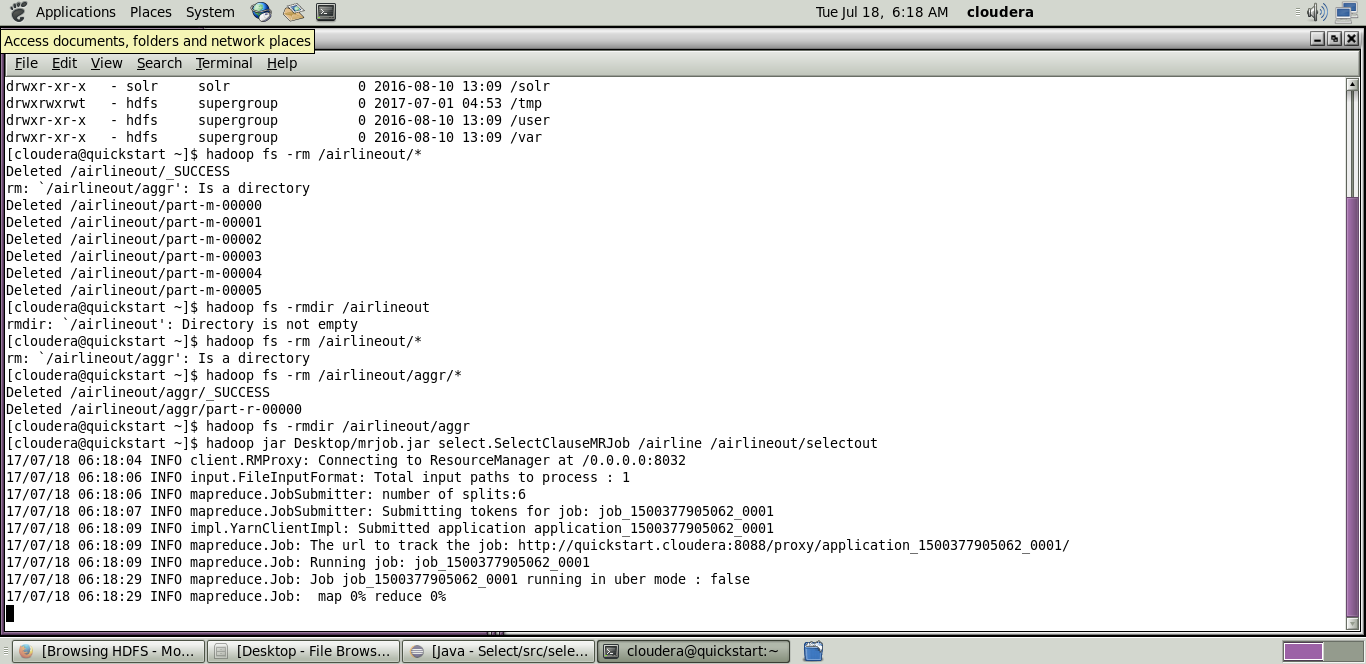
**Code:hadoop fs -put /home/cloudera/Downloads/2008.csv/airline**

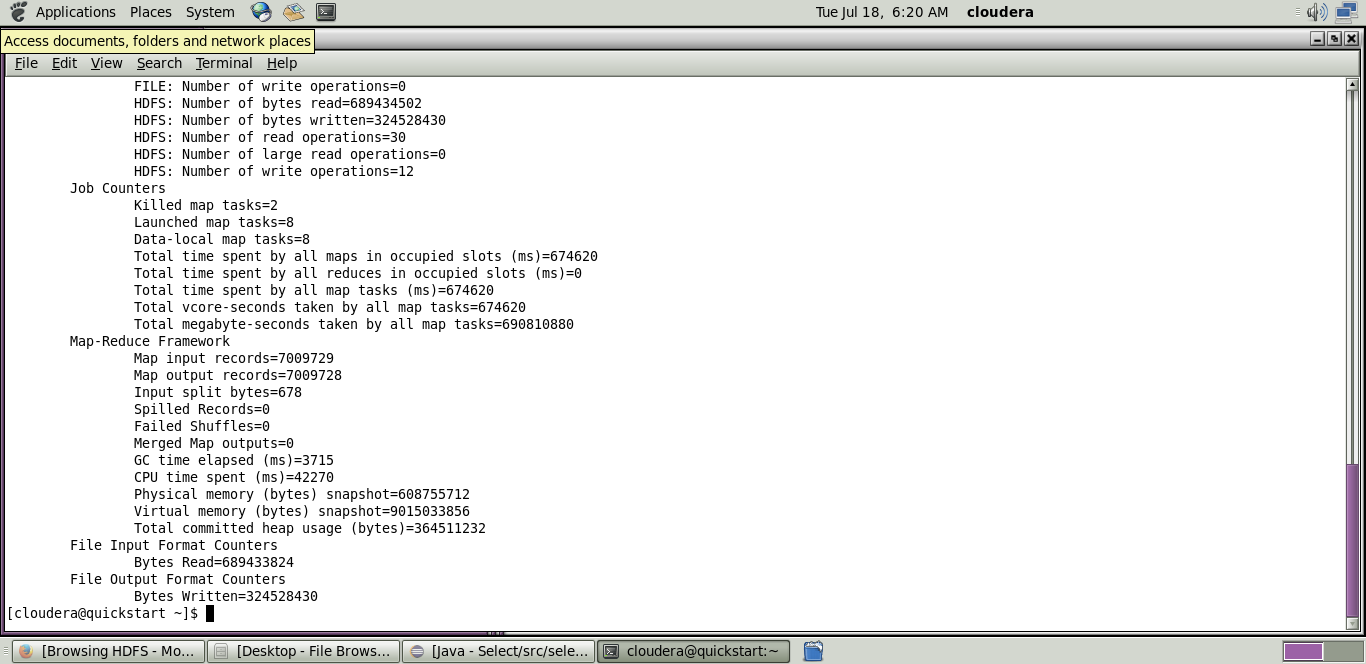
**Initializing mapreduce job:**

**Step 14:**

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

**Code:**hadoop jar Desktop/select.jar select.SeleceClauseMRJob /airline/2008.csv /airlineout/selectout





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/selectout/\*

**Step 16:**

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

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

