**Analysis on Census Data**

**Introduction**

This Project is based on Census. A census is the procedure of systematically acquiring and recording information about the members of a given population. It is a regularly occurring and official count of a particular population. The term is used mostly in connection with national population and housing censuses; other common censuses include agriculture, business, and traffic censuses.

**Objective**

The objective of this project is to rely on data for decision-making and promote civic engagement: state and local governments; social service agencies; planners; foundations; and, child and family welfare, education, and other vital services.

**Requirements Specifications**

This project deals with Census, we have to handle huge volume of data (which will rise tremendously). Here we are having two kinds of data.

* Census data which contains details of people (such as age, education, marital status, gender, income, tax filler, parents, country of birth, citizen, work etc.)
* Age group data which contains the details of age (such as age, and category).
* Secondary table for Tax analysis, Pension and Scholarship.

**Analysis**

Being a census analyzing project, we are going to implement this project with the help HADOOP, an open source Java-based programming framework. There are many Ecosystem tools in HADOOP from there we used **Pig,** **Hive and Sqoop.**

**Technologies**

* **Map Reduce:** Hadoop Map Reduce is a software framework for easily writing applications which process vast amounts of data (multi-terabyte data-sets) in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner
* **Pig:** Pig is a high-level platform for creating programs. The language for this platform is called Pig Latin. It can be extended using User Defined Functions (UDFs) which the user can write in Java, Python, JavaScript, Ruby or Groovy and then call directly from the language.
* **Hive:** Hive gives an SQL-like interface to query data stored in various databases and file systems that integrate with Hadoop. The traditional SQL queries must be implemented in the Map Reduce Java API to execute SQL applications and queries over a distributed data.
* **Sqoop:** Sqoop is a tool designed to transfer data between Hadoop and relational databases or mainframes. You can use Sqoop to import data from a relational database management system (RDBMS) such as MySQL or Oracle or a mainframe into the Hadoop Distributed File System (HDFS), transform the data in Hadoop MapReduce, and then export the data back into an RDBMS.

**Use cases:**

Project tasks are divided into different use cases based on analysis.

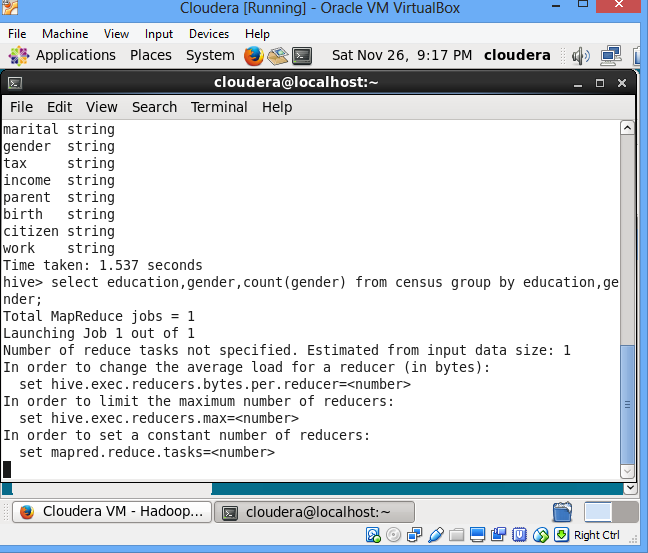
**Education and Employment**

The scope of this project is to develop the country. Education plays a major role in this. So here we are collecting data on education details used to measure the well-being of people. It’s allowing decision-makers to target developed country and resources to organization. Under this category we have taken three tasks

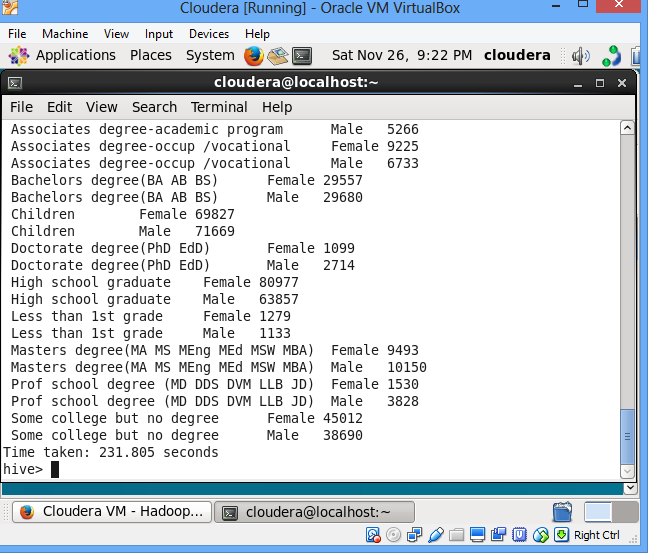
1. Total count of male/female based on education.
2. Total count of employed/unemployed based on education.
3. Total count for people in age range of 18-25 based on education.
4. Degree wise count for employability

**Using Hive**

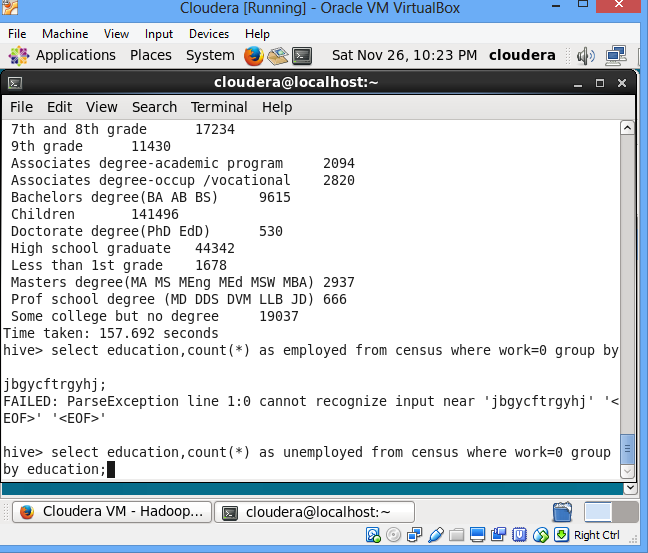
**Execution Step:** Total count of male/female based on education.



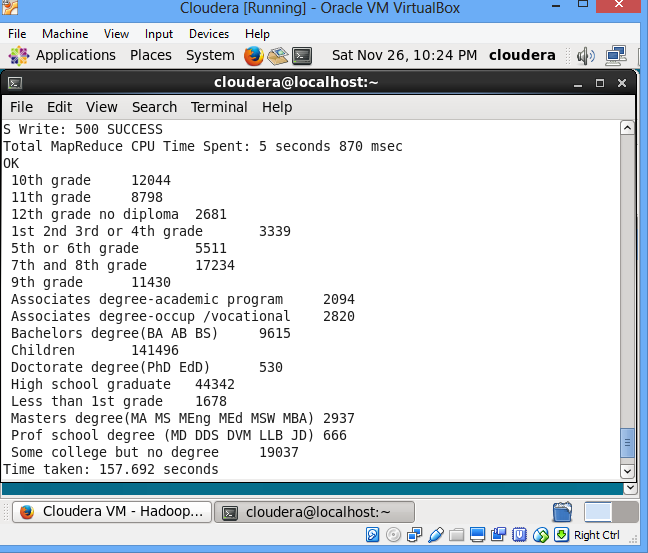
**Output**



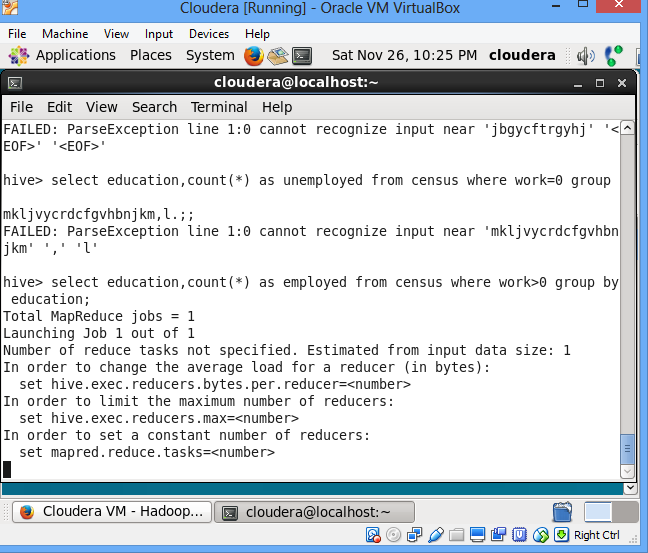
**Execution Step:** Total count of unemployed based on education.



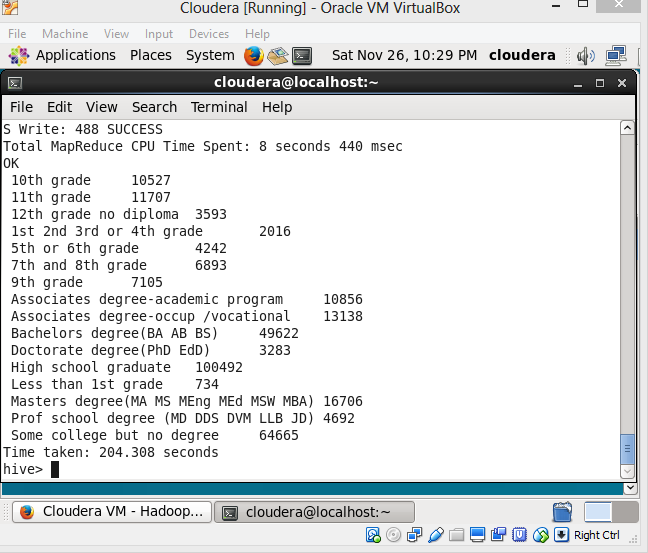
**Output (Unemployed):**



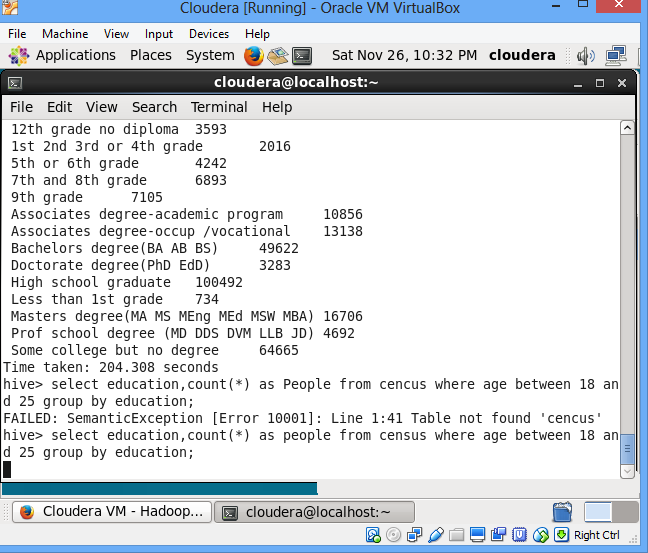
**Execution Step:** Total count of employed based on education.



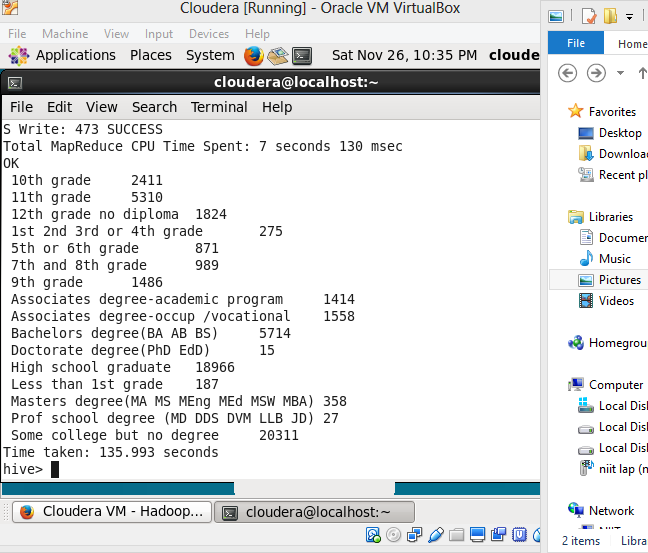
**Output (employed):**



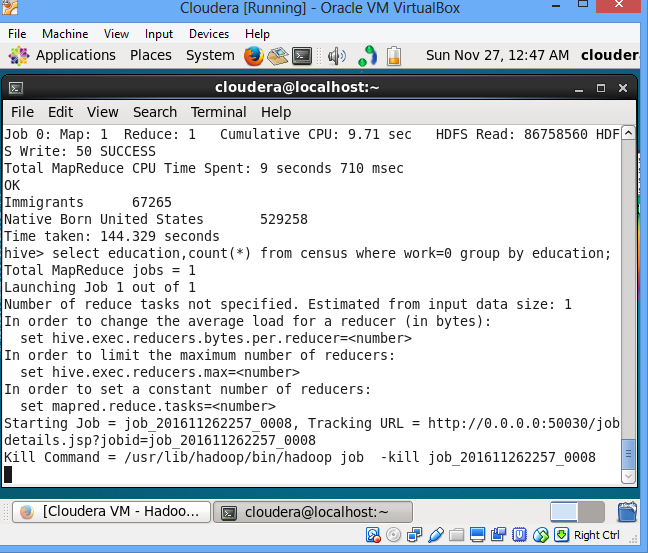
**Execution Step:** Total count for people in age range of 18-25 based on education.



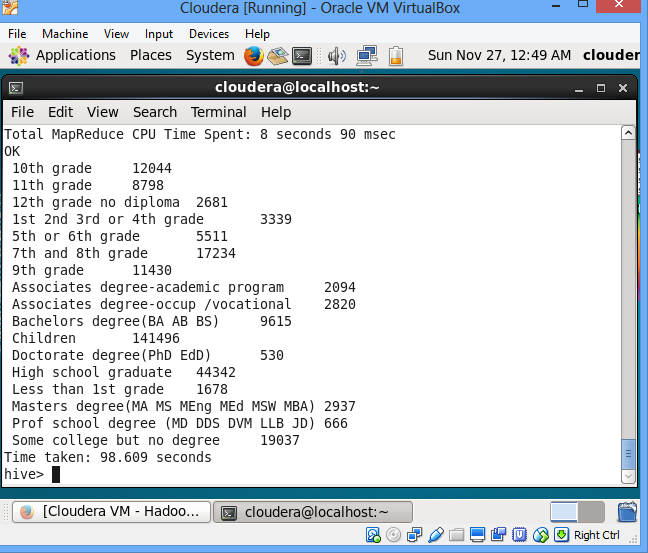
**Output:**



**Execution Step:** Degree wise count for employability



**Output:**

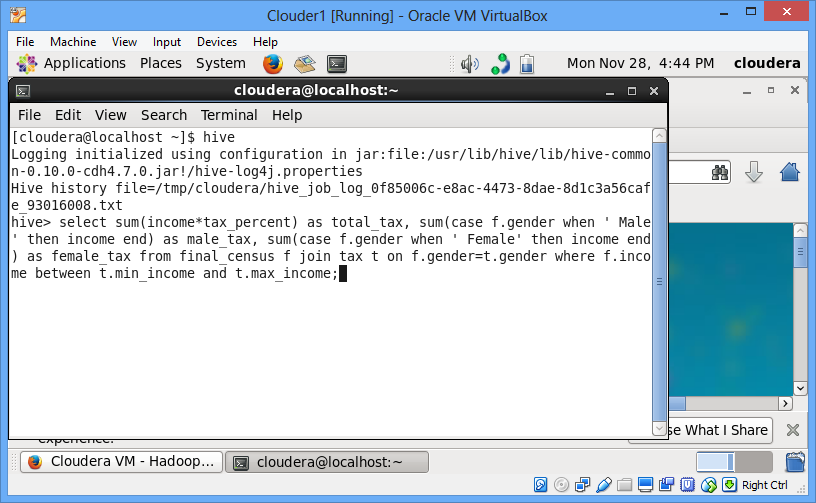


**Tax and Income:**

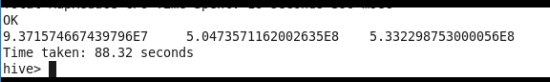
The purpose of taxes is to raise revenue to fund government. Money provided by taxation has been used by states and their functional equivalents throughout history to carry out many functions. Governments also use taxes to fund welfare and public services. These services can include education systems, pensions for the elderly, unemployment benefits. Under this we have taken three tasks

1. Tax analysis total and gender wise
2. Per Capita Income (PCI) analysis consolidated, gender wise and category wise
3. Non-US citizen(s) tax filer status

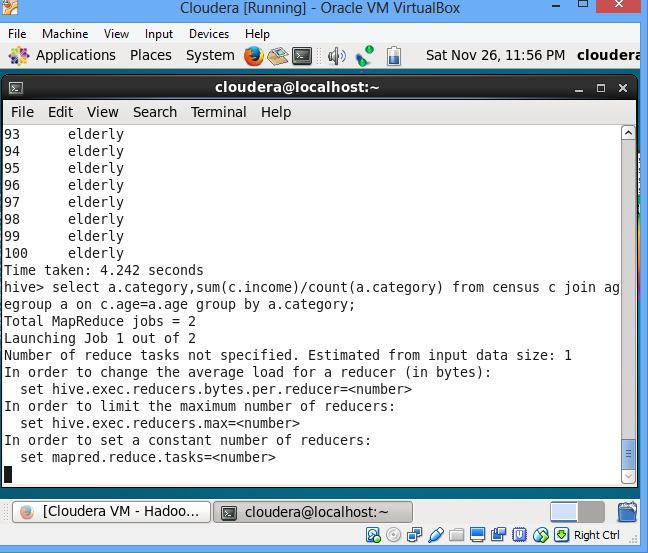
**Execution Step:** Tax analysis total and gender wise



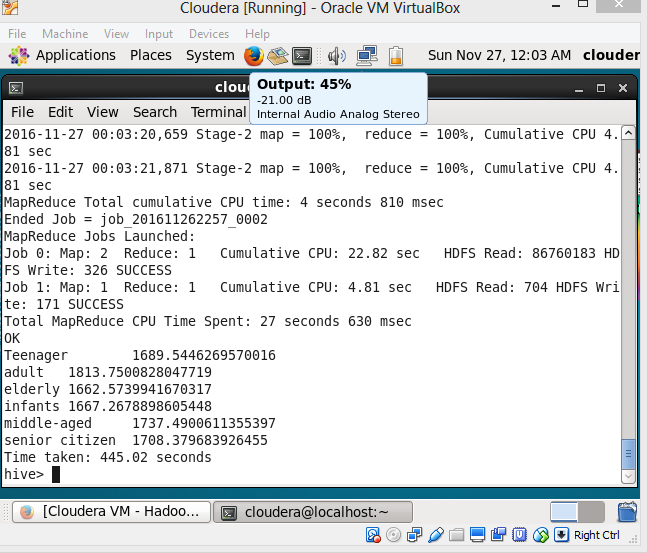
**Output:**



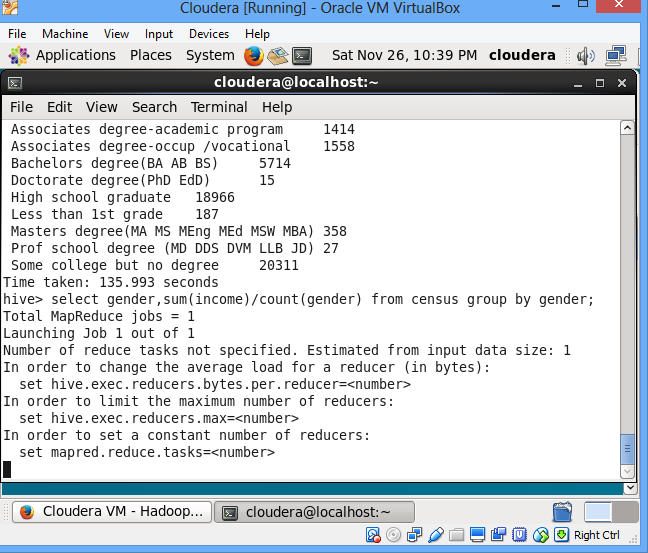
**Execution Step:** Per Capita Income (PCI) analysis category-wise



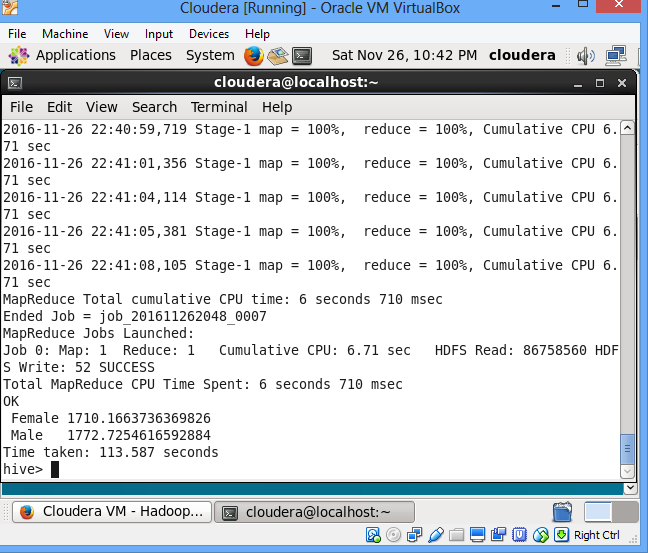
**Output:**



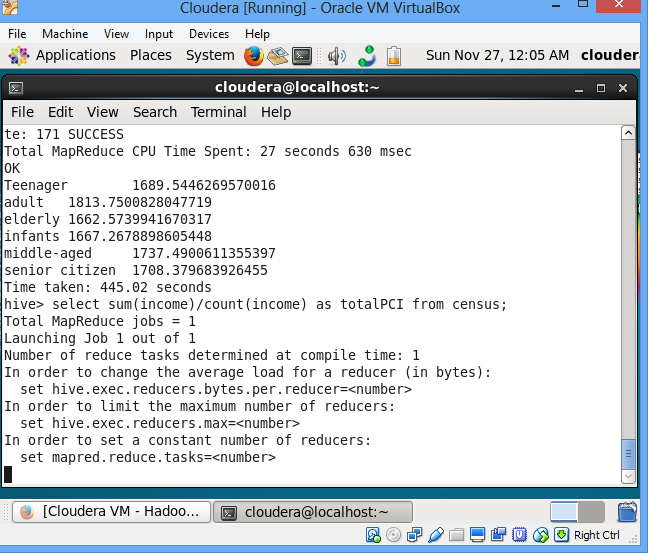
**Execution Step:** Per Capita Income (PCI) analysis gender-wise

****

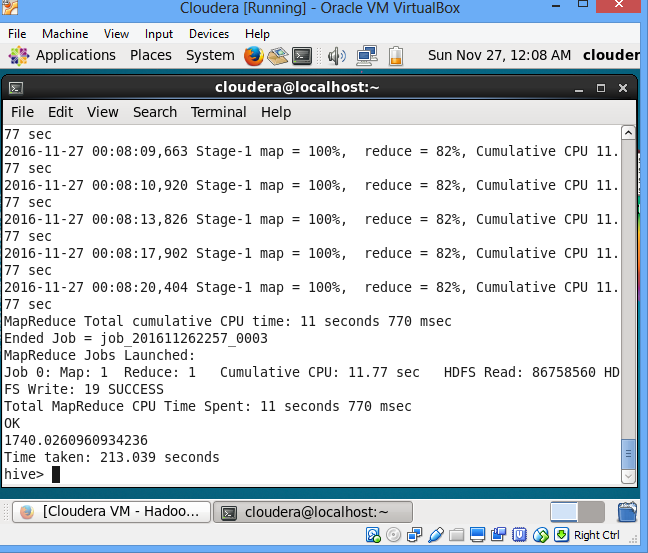
**Output:**

****

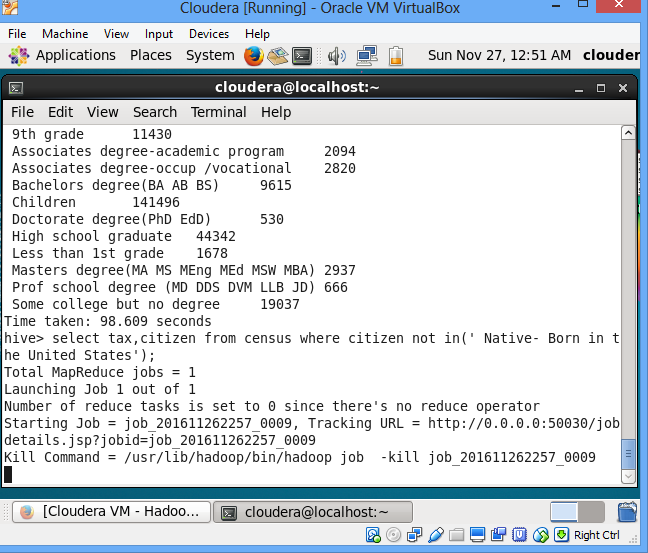
**Execution Step:** Per Capita Income (PCI) analysis consolidated



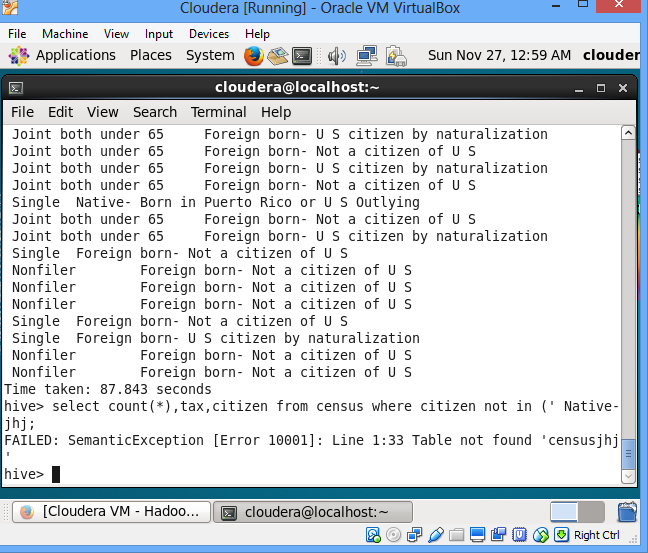
**Output:**

****

**Execution Step:** Non-US citizen(s) tax filer status



**Output:**



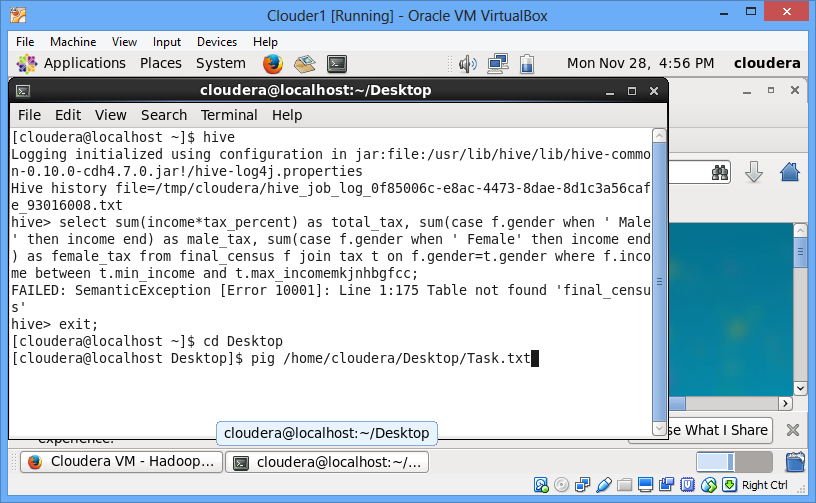
**Welfare and Budget:**

Welfare is largely provided by the government from tax income, and to a lesser extent by charities, informal social groups, religious groups, and inter-governmental organizations. Under this category we have taken

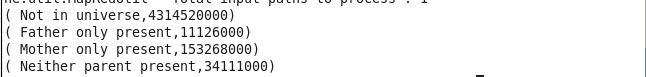
1. Total amount dispensed on scholarship in current year
2. For given age range employable female widowed and divorced count
3. Total amount dispensed on pension in x year(s)
4. Citizens and immigrants count for employed lot

**Execution Step (Using Pig):** Total amount dispensed on scholarship in current year

**Execution Step:**



**Output:**



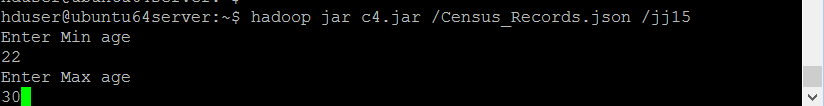
**Execution Step (Using Map Reduce):** Total amount dispensed on pension in x year(s)



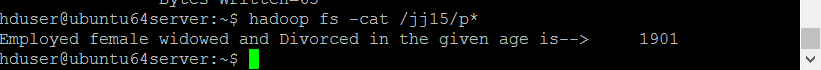
**Output:**



**Execution Step:** For given age range employable female widowed and divorced count



**Output:**



**Using Pig:**

step1 = load '/user/cloudera/Census\_Records.json' using JsonLoader('Age:int,Education:chararray,Marital:chararray,Gender:chararray,Tax:chararray,Income:float,Parent:chararray,Birth:chararray,Citizen:chararray,Work:int');

step2 = foreach step1 generate Age,Gender,Work,Marital;

step3 = filter step2 by ((Gender==' Female' and work>0) and (Marital==' Widowed' or Marital==' Divorced') and (age>21 and age<60));

step4 = group step3 by age;

step5 = foreach step4 generate group,COUNT(d.age);

dump step5;

**Output:**



**Population & Immigration**

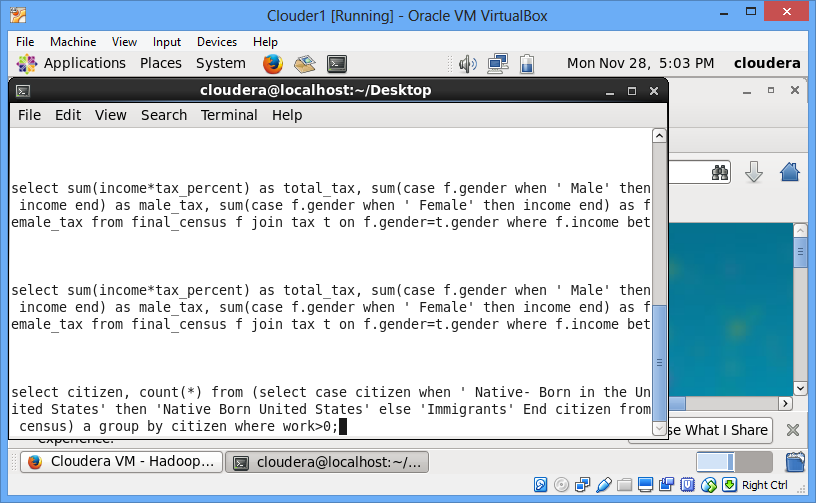
As of today's date, the world population is estimated by the United States Census Bureau to be 7.465 billion. Population growth increased significantly as the Industrial Revolution gathered pace from 1700 onwards.

In 2016, similar to the overall foreign-born population, 47 percent of the 2 million Indian immigrants residing in the United States were naturalized U.S. citizens.

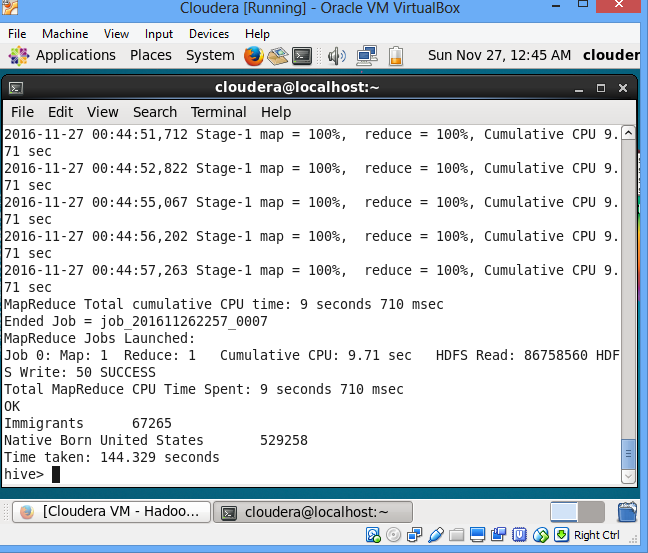
Under this category we have taken three tasks

1. Citizens and immigrants count for employed lot
2. Country of birth wise count for US citizenship by naturalization
3. Total number of Male/Female

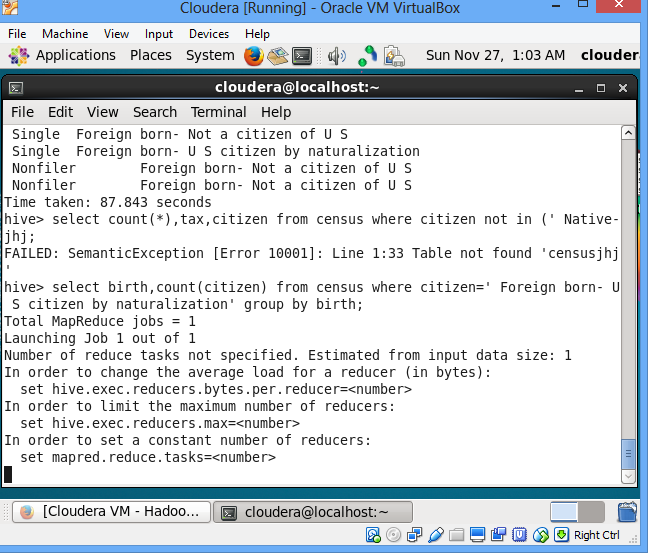
**Execution Step:** Citizens and immigrants count for employed lot



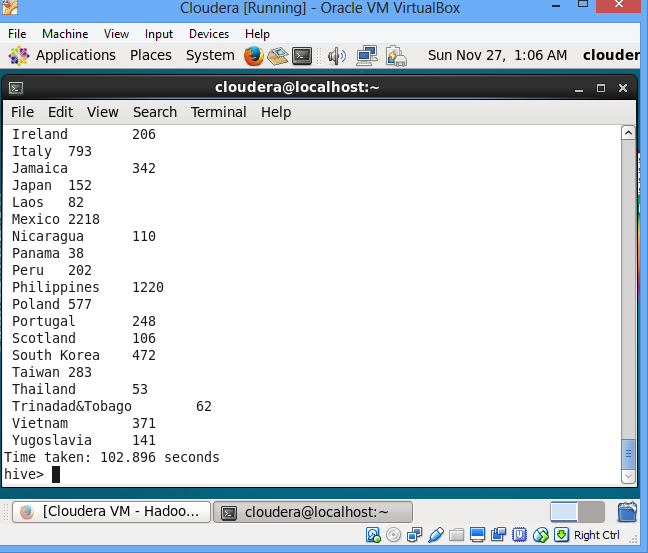
**Output:**



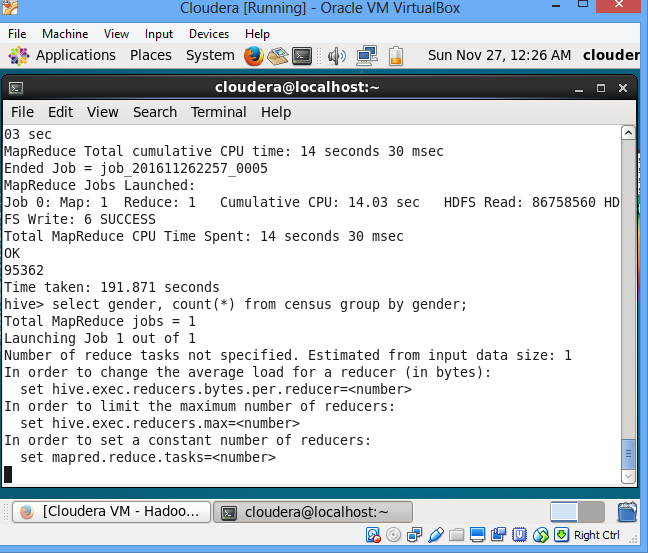
**Execution Step:** Country of birth wise count for US citizenship by naturalization



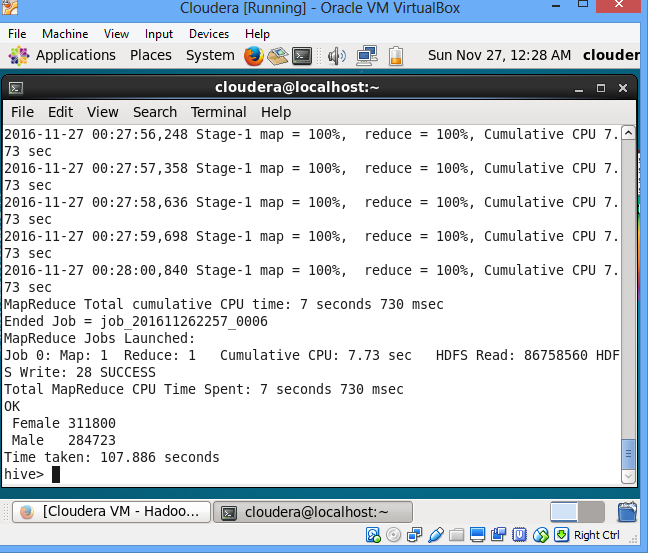
**Output:**



**Execution Step:** Total number of Male/Female



**Output:**



**Execution Step:** Customer base analysis

step1 = load '/user/cloudera/Census\_Records.json' using JsonLoader('Age:int,Education:chararray,Marital:chararray,Gender:chararray,Tax:chararray,Income:float,Parent:chararray,Birth:chararray,Citizen:chararray,Work:int');

step2 = foreach step1 generate Age,Gender,Work,Marital;

step3 = filter step2 by ((Gender==' Female' and work==0 and Marital==' Widowed') and (age>21 and age<60)) ;

step4 = group step3 by age;

step5 = foreach step4 generate group,COUNT(d.age);

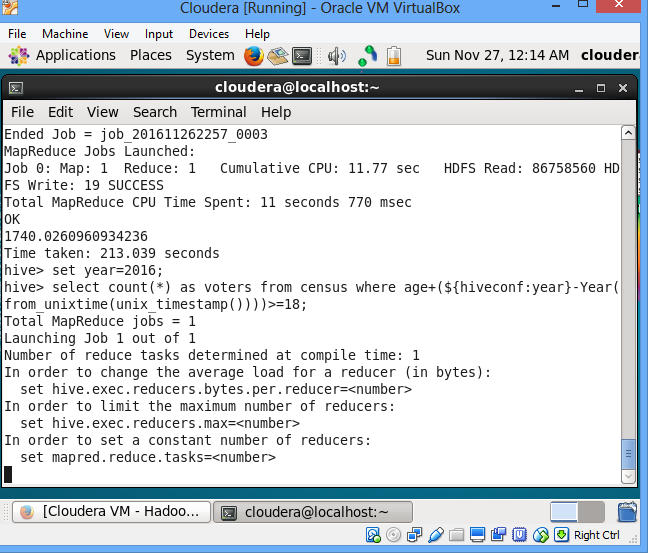
dump step5;

**Future Plan**

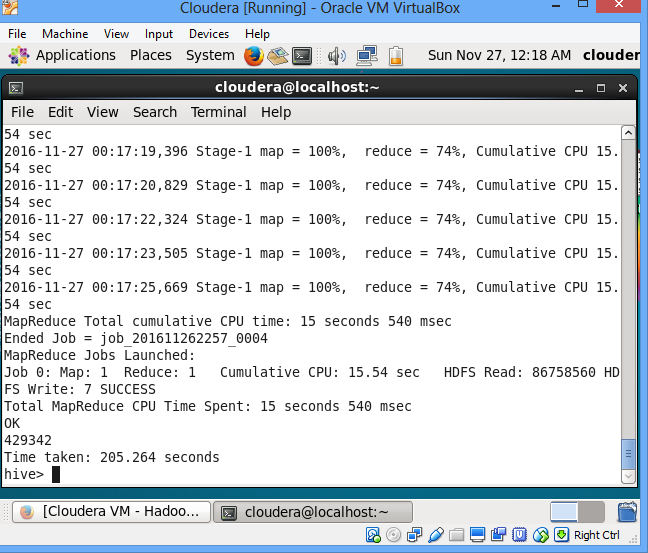
From this we can able to find how many citizens are there eligible for voting in x year and how many senior citizens are in x-year. Under this category we have taken two tasks.

1. Voter(s) count in x year(s)
2. Senior Citizen(s) count in x year(s)

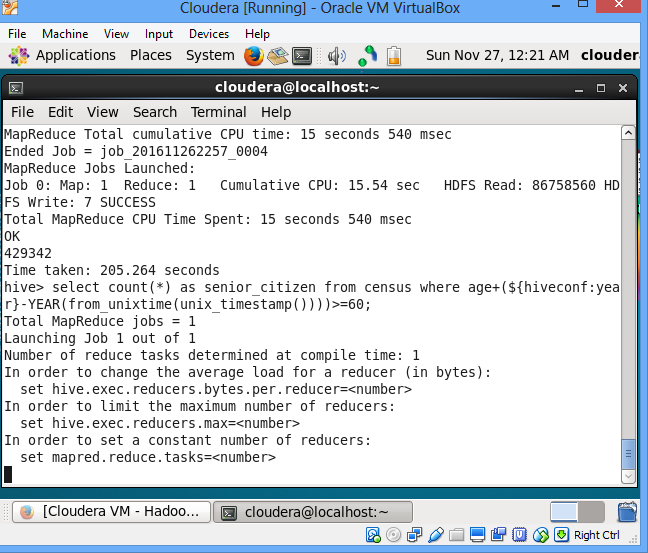
**Execution Step:** Voter(s) count in x year(s)



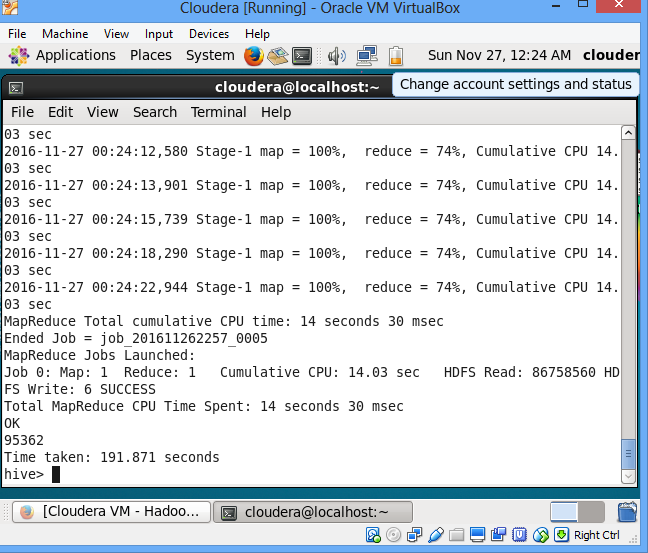
**Output:**



**Execution Step:** Senior Citizen(s) count in x year(s)



**Output:**



**Extras: Healthcare**

This is my future outcome. Here we will analysis the number of employee worked for more than 38 weeks and will conduct medical camp for them. This will lead to a health and wealth country.

**Conclusion**

The census is thus an extremely useful source of knowledge and the information available through all over the world "contributing to a revolutionary expansion of global economic, sociological and demographic knowledge".