**CENSUS DATA ANALYSIS BY GOVERNMENT DEPARTMENT**

**Developer:** Ganesh S

**Objective:**

A census of population may be defined as the total process of collecting, compiling and publishing demographic, economic and social data pertaining, at a specific time or times, to all persons in a country or delimited territory. This project analysis for various departments are education, finance, social, planning, miscellaneous.

**Given Data:**

Two Datasets are as below used for this project-

* Census Data
* Age Data

**Hardware Requirements:**

* Minimum 8GB RAM.
* Next Generation Processor Chips like Intel I7.
* Windows 64 Bit System

**Software Requirements:**

* Virtual Box
* Eclipse
* Ubuntu Terminal (for Map Reduce)
* Cloudera OS (for HIVE) SYSTEM Technology Used:
* **Apache Hadoop:** Apache Hadoop an open-source software framework used for distributed storage and processing of very large data sets. It consists of computer clusters built from commodity hardware.
* **Map Reduce Program:** 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.
* **Apache Hive:** Apache Hive is data warehouse infrastructure built on top of Apache Hadoop for providing data summarization, ad-hoc query, and analysis of large datasets. It provides a mechanism to project structure onto the data in Hadoop and to query that data using a SQL-like language called HiveQL (HQL).
* **Apache Pig:** Apache Pig is a high-level platform for creating programs that run on Apache Hadoop. The language for this platform is called Pig Latin. Pig can execute its Hadoop jobs in Map Reduce.

**Project Description:**

**Use Case 1: Government Jobs**

All government Jobs vacancies filed by the recruitment process like written examination, interview, group discussion and physical tests (if necessary).

To apply for these vacancies, candidates should eligible for the posts as per eligibility criteria stipulated by that department like qualification, age limit and other eligibility conditions.

**For example:** Nominate the government job who are all the eligible persons based on education to collect the data from the total count of male and female candidate ,total count of employed /unemployed persons and total count for people in age range of 18-25 based on education.

1.1 Total count of male/female based on education. If it bachelor, how many male/female

**PIG:**

a = load '/user/cloudera/Census1.json' using JsonLoader('Age:int,Education:chararray,MartialStatus:chararray,Gender:chararray,TaxFilerStatus:chararray,Income:float,Parents:chararray,CountryOfBirth:chararray,Citizenship:chararray,WeeksWorked:chararray');

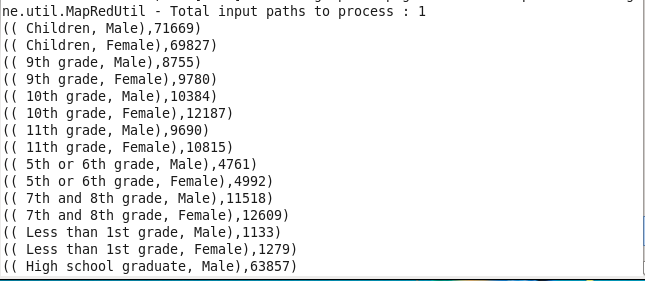
b = for each step1 generate $1 as Edu, $3 as Gen;

c = group step2 by ($0, $1);

d = for each step3 generate group, COUNT (step2.Gen);

dump d;

**Screen shot 1.1.1**



1.2 Total count of employed/unemployed based on education. : if week work is 0 then unemployed

**PIG:**

**Employed Counts:**

a = load '/user/cloudera/Census1.json' using JsonLoader('Age:int,Education:chararray,MartialStatus:chararray,Gender:chararray,TaxFilerStatus:chararray,Income:float,Parents:chararray,CountryOfBirth:chararray,Citizenship:chararray,WeeksWorked:int');

b = for each step1 generate $1 as Edu, $9 as ww;

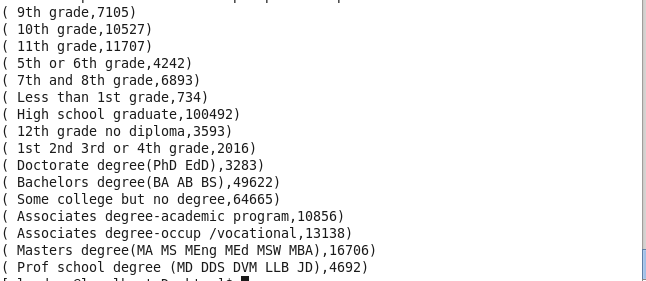
c = filter step2 by $1>0;

d = group step3 by $0;

e = for each step4 generate group, COUNT ($1);

dump e;

**Screen shot 1.2.1**



**PIG:**

**Unemployed Counts:**

a = load '/user/cloudera/Census1.json' using JsonLoader('Age:int,Education:chararray,MartialStatus:chararray,Gender:chararray,TaxFilerStatus:chararray,Income:float,Parents:chararray,CountryOfBirth:chararray,Citizenship:chararray,WeeksWorked:int');

b = for each step1 generate $1 as Edu, $9 as ww;

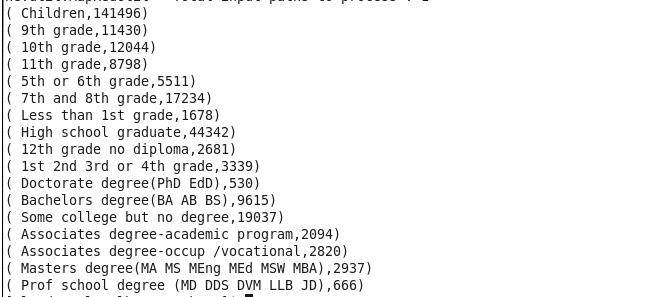
c = filter step2 by $1==0;

d = group step3 by $0;

e = for each step4 generate group, COUNT ($1);

dump e;

**Screen shot 1.2.2**



1.3 Total count for people in age range of 18-25 based on education.

**PIG:**

a = load '/user/cloudera/Census1.json' using JsonLoader('age:int,edu:chararray,mar:chararray,gen:chararray,tax:chararray,income:chararray,parent:chararray,country:chararray,citizen:chararray,ww:int');

b = for each a generate age, edu;

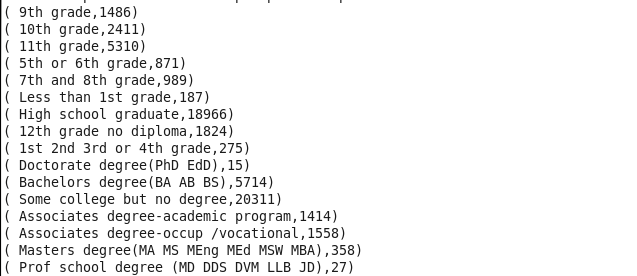
c = filter b by age>=18 and age<=26;

j = group c by edu;

d = for each j generate group, COUNT (c.age);

dump d;

**Screen shot 1.3.1**

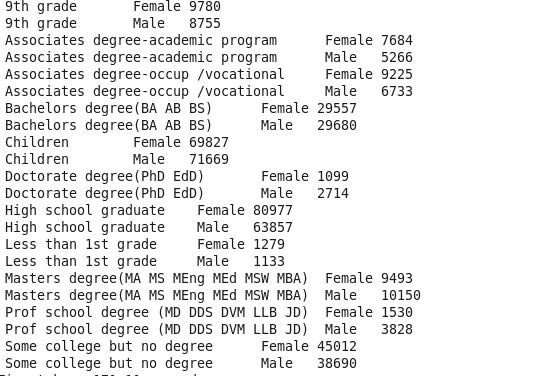


**Hive:**

1.4 Total count of male/female based on education.

Select edu, gen, COUNT (\*) Total from final\_census group by edu, gen;

**Screen shot 1.4.1**

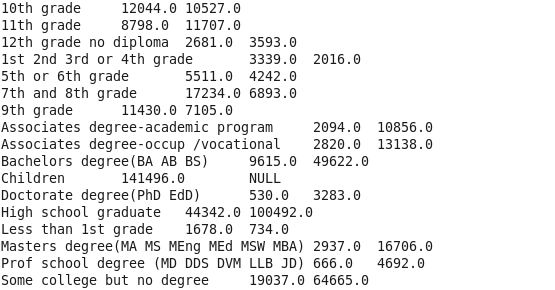


1.5 Total count of employed/unemployed based on education.

Select edu, SUM (CASE when ww <=0 then '1' else null END) as Employed, SUM (CASE when ww >0 then '1' else null END) as Unemployed

from final\_census group by edu;

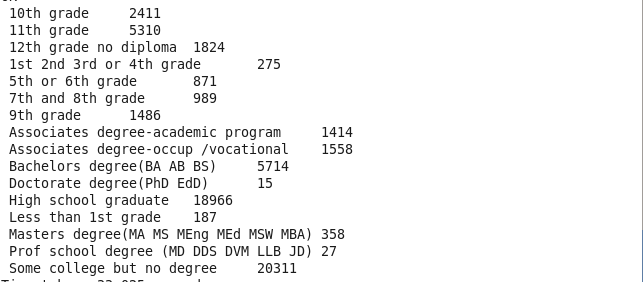
**Screen shot 1.5.1**



1.6 Total count for people in age range of 18-25 based on education.

Select edu, COUNT (\*) as Total\_Peoples from final\_census where age between 18 and 25;

**Screen shot 1.6.1**



**Use Case 2: Growth of the country**

The former British Prime Minister William Gladstone expressed the importance of finance for the economy in 1858 as follows: *"*Finance is, as it were, the stomach of the country, from which all the other organs take their tone.*"*

**For Example:** We are analyze the growth of the country based on the total tax analysis ,gender wise and Per Capita Income(PCI) analysis consolidated, gender wise and category wise. It is very helpful to increase the growth of the country.

2.1 Tax analysis total and gender wise

Select SUM (income\*tax\_pct) as Total\_Tax, SUM (CASE f.gen when 'Male' then income END) as TAX\_Male, SUM (CASE f.gender when 'Female' then income END) as Tax\_Female from final\_census f join gen\_wise\_tax t on (f.gen=t.gen) where f.income between t.minamount and t.maxamount;

**Screen shot 2.1.1**



2.2 Per Capita Income (PCI) analysis consolidated, gender wise and category wise

**Gender wise:**

Select gen, sum (income)/count (gen) from final\_census group by gen;

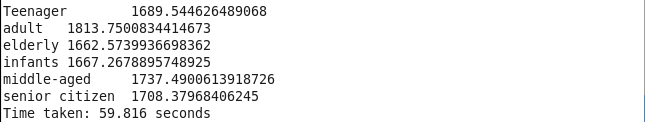
**Screen shot 2.2.1**



**Category wise:**

Select a.cat, sum (f.income)/ (a.cat) from final\_census f join age group a on f.age=a.age group by a.cat;

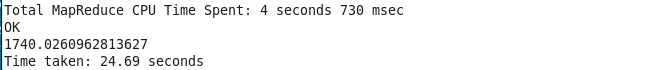
**Screen shot 2.2.2**



**Total CPI:**

Select sum (income)/count (income) as total cpi from final\_census;

**Screen shot 2.2.3**



**Use case 3: Government planning to increase the scholarship to the citizens.**

Government benefits from one of these six federal programs: **Social Security**, **Medicare**, **Medicaid**, **welfare** (**TANF**),unemployment benefits, and **food stamps** (**SNAP**).

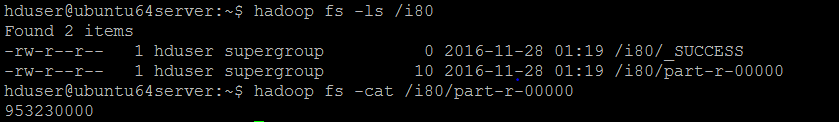
Here, We are analysis the data of how much amount dispensed in the pension for above 60 years old peoples and then analysis the total amount dispensed in the scholarship of citizens based upon the criteria of father only present people, mother only present peoples, neither parent present and parents are not in the universe.

Next to analysis the details about how many peoples are employed in female widow and divorced. Based upon these details we are sending the information to the government to increase the benefits of those citizens.

3.1 Total amount dispensed on pension in x year(s)

**Screen shot 3.1.1**





3.2 Total amount dispensed on scholarship in current year

**PIG:**

PIG File: t1.txt

a = load '/user/cloudera/Census1.json' using JsonLoader('Age:int,Education:chararray,MartialStatus:chararray,Gender:chararray,TaxFilerStatus:chararray,Income:float,Parents:chararray,CountryOfBirth:chararray,Citizenship:chararray,WeeksWorked:chararray');

b = load '/user/cloudera/scholar1' using Pig Storage (',') as (status: chararray, schamt: int);

c = join a by Parents, b by status;

d = for each c generate $6 as parent, $11 as Schamt;

e = group d by $0;

f = for each e generate group, SUM (d.Schamt);

dump f;

Secondary table: scholar1:

Father only present, 2000

Mother only present, 4000

Neither parent present, 7000

Not in universe, 10000

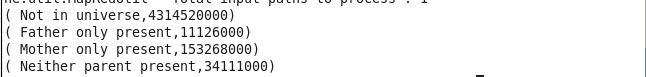
Run:

**Screen shot 3.2.1**



Output:

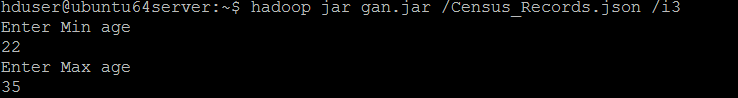
**Screen shot 3.2.2**



3.3 For given age employable female widowed and divorced count – done

**Map Reduce**

**Screen shot 3.3.1**



**Output:**

**Screen shot 3.3.2**



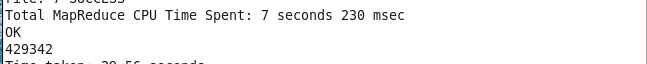
**Use Case 4: Government to check whether who all citizens of our country**

In this use case to count the voters, the total count of the senior citizenship and who are all the immigrants count for the employed lot .To count the total number of male and female. It is very useful to generate list of voting list.

4.1 Voter(s) count in x year(s)

Select COUNT (\*) as Total\_Voters from final\_census where age+(${hiveconf:year}-YEAR(from\_unixtime(unix\_timestamp())))>=18;

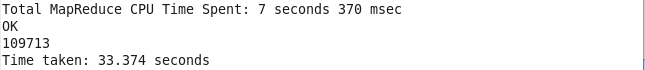
**Screen shot 4.1.1**



4.2Senior Citizen(s) count in x year(s)

Select COUNT (\*) as Total\_Senior\_Citizen from final\_census where age+(${hiveconf:year}-YEAR(from\_unixtime(unix\_timestamp())))>=60;

**Screen shot 4.2.1**



4.3Total number of Male/Female

Select gen, COUNT (\*) as Total from final\_census group by gen;

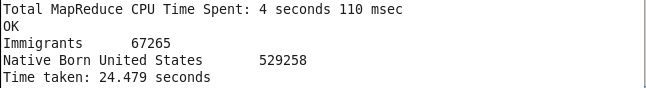
**Screen shot 4.3.1**



4.4 Citizens and immigrants count for employed lot

Select citizen, COUNT (\*) from (select CASE citizen when ' Native- Born in the United States' then 'Native Born United States' else 'Immigrants' END citizen from final\_census) a group by citizen;

**Screen shot 4.4.1**



**USE CASE 5: Tax Filer Status**

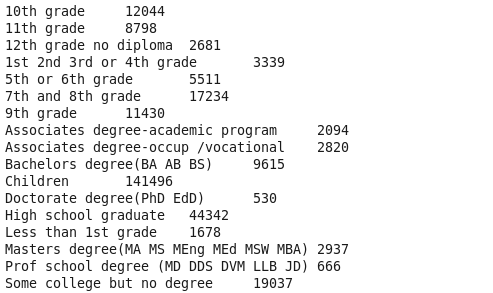
To find the who are all non-us citizens (not born in US).To give the details to the government who all are non-us citizens to pay the tax.

**Miscellaneous:**

5.1Degree wise count for employability

Select edu, count (\*) from final\_census where ww=’0’ group by edu;

**Screen shot 5.1.1**



5.2 Customer base analysis

**PIG:**

a = load '/user/cloudera/Census. Son' using JsonLoader('age:int,edu:chararray,mar:chararray,gen:chararray,tax:chararray,income:long,parent:chararray,country:chararray,citizen:chararray,ww:int');

b = for each a generate age, gen, income;

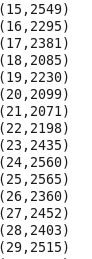
d = filter b by ((gen==' Male' and income>1500) and (age>14 and age<31));

j = group d by age;

k = for each j generate group, COUNT (d.age);

dump k;

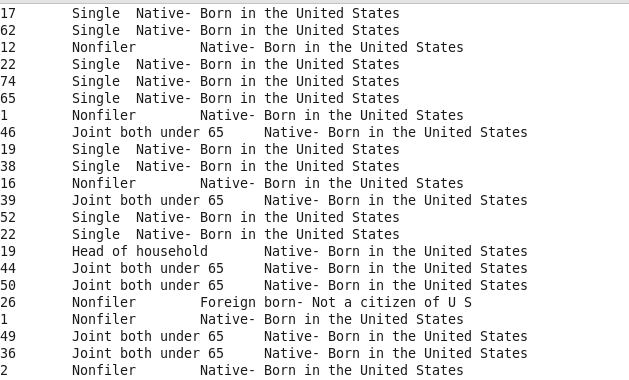
**Screen shot 5.2.1**



5.3 Non-US citizen(s) tax filer status

Select age, tax, citizen from final\_census where citizen not in (‘native born in the United States’);

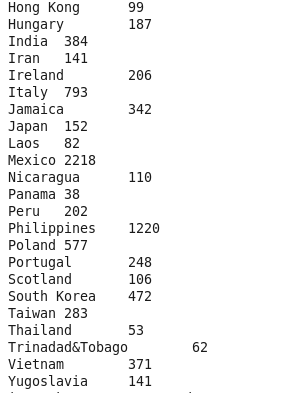
**Screen shot 5.3.1**



5.4 Country of birth wise count for US citizenship by naturalization

Select cntry, count (citizen) from final\_census where citizen=’ Foreign born- U S citizen by naturalization ‘group by cntry;

**Screen shot 5.4.1**



**Conclusion:**

This project is used to increase the chances of government jobs to the citizens and also increases the benefits of employed female widow divorced peoples.