CENSUS DATA ANALYSIS

Project Author : Niyaz Ahmed

Date : 29/11/2016

Objective : To analyze the data of the Census Details and Age group so that the E-Commerce Company, Government Sector, Private Company can Analyze them and apply business methodology so that it helps them in future growth in terms of Education Sector, Sales, Customer Satisfaction and many more.

Test Data :

1) Census Data (Census_Records.json)

Age	Education	Martial	Gender	Tax	Income	Parent	Country	Citizen	Week Work

2) Age Group (agegroup.dat)

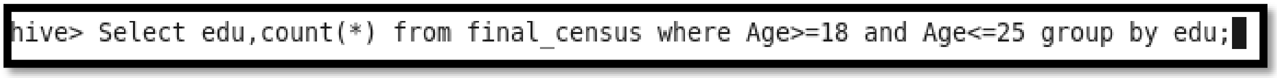
Age	Group

Use Case 1: Education Data

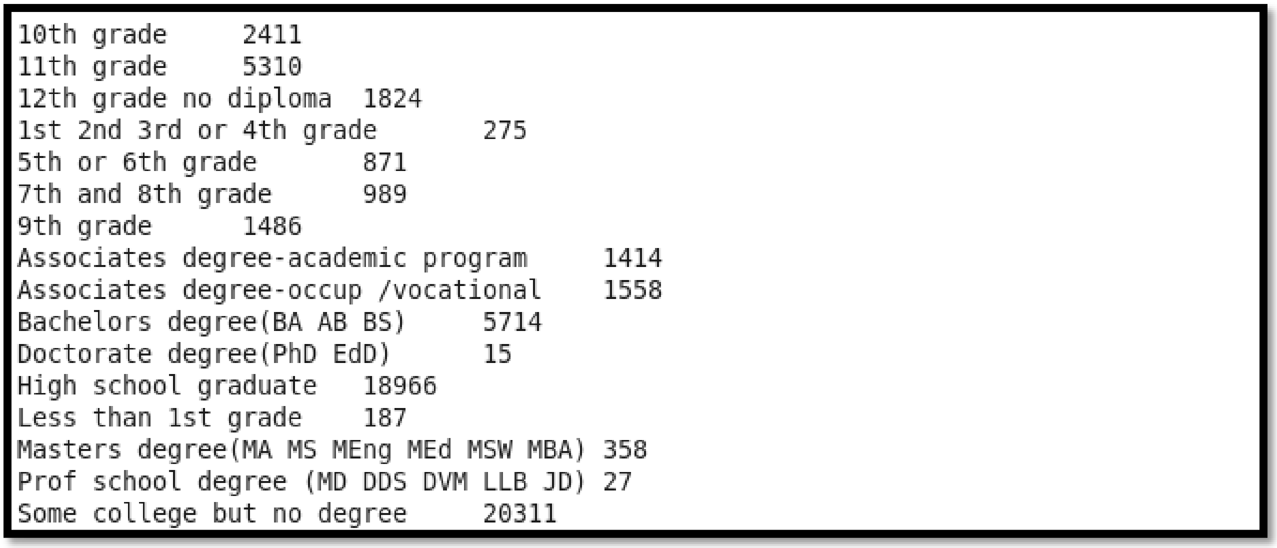
- Using the census we can find the number of people within a range of age. For eg : 18 to 24 years. We can target that category of courses and books to them
- Depending upon the count of Male and Female a specific count can be attained which will be useful deciding the product category and needs of product depending upon the market need
- A separate courses and book can be made available depending upon the number of employed and unemployed person. For eg : A person who is employed we can sell them courses which will help them in preparing for further higher technologies.
- **Tasks** included are: Task1 to Find count of Male and Female based on education. Task2 to Find count of employed and unemployed based on education and Task3 to Find count for people in age range based on education.

OUTPUT :

Screenshot of Task 3:

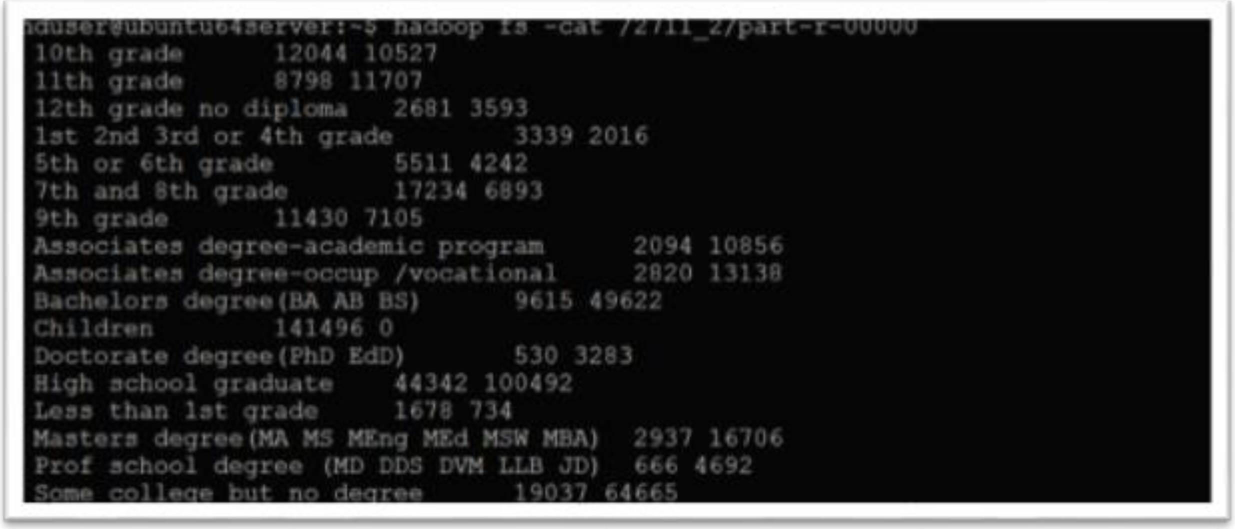


```
hive> Select edu,count(*) from final_census where Age>=18 and Age<=25 group by edu;
```



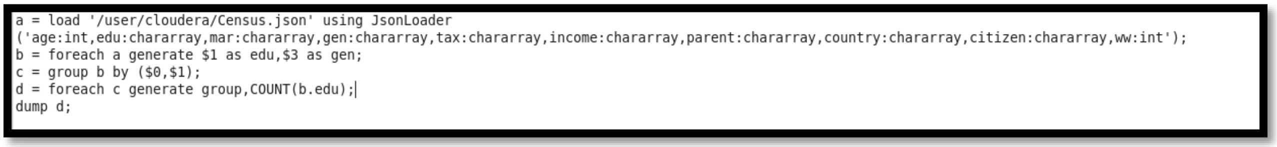
```
10th grade     2411
11th grade     5310
12th grade no diploma  1824
1st 2nd 3rd or 4th grade       275
5th or 6th grade       871
7th and 8th grade      989
9th grade      1486
Associates degree-academic program     1414
Associates degree-occup /vocational    1558
Bachelors degree(BA AB BS)     5714
Doctorate degree(PhD EdD)      15
High school graduate   18966
Less than 1st grade    187
Masters degree(MA MS MEng MEd MSW MBA) 358
Prof school degree (MD DDS DVM LLB JD) 27
Some college but no degree     20311
```

Screenshot of Task 2 :

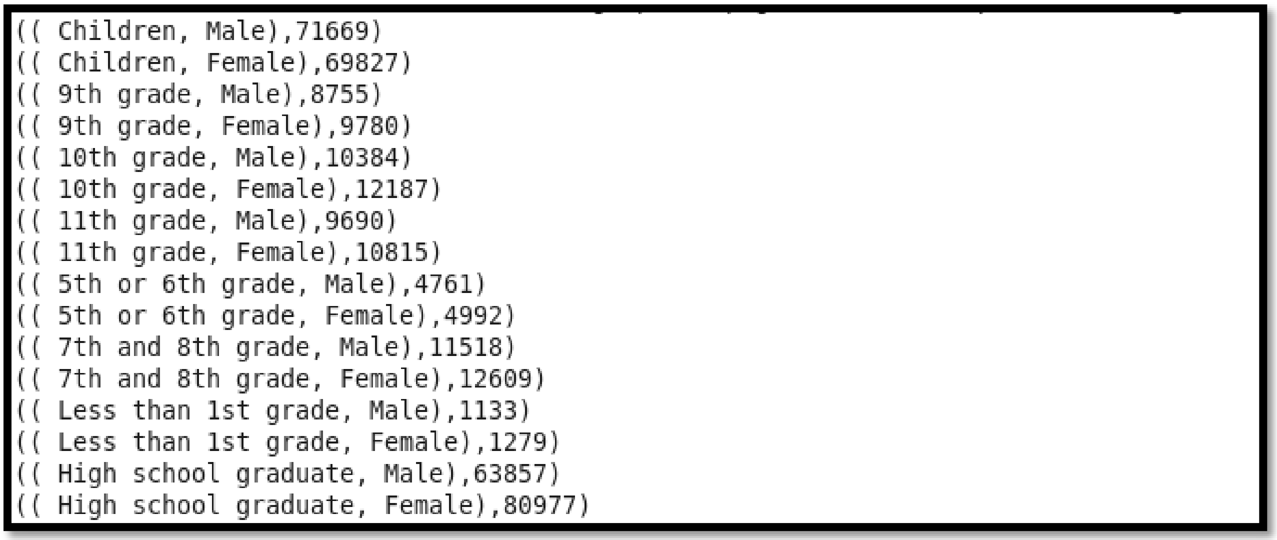


```
hduser@ubuntu64server:~$ hadoop fs -cat /2711_2/part-r-00000
 10th grade       12044 10527
 11th grade       8798 11707
 12th grade no diploma   2681 3593
 1st 2nd 3rd or 4th grade        3339 2016
 5th or 6th grade        5511 4242
 7th and 8th grade       17234 6893
 9th grade       11430 7105
 Associates degree-academic program      2094 10856
 Associates degree-occup /vocational     2820 13138
 Bachelors degree(BA AB BS)      9615 49622
 Children        141496 0
 Doctorate degree(PhD EdD)       530 3283
 High school graduate    44342 100492
 Less than 1st grade     1678 734
 Masters degree(MA MS MEng MEd MSW MBA)  2937 16706
 Prof school degree (MD DDS DVM LLB JD)  666 4692
 Some college but no degree      19037 64665
```

Screenshot of Task 1 :



```
a = load '/user/cloudera/Census.json' using JsonLoader
('age:int,edu:chararray,mar:chararray,gen:chararray,tax:chararray,income:chararray,parent:chararray,country:chararray,citizen:chararray,ww:int');
b = foreach a generate $1 as edu,$3 as gen;
c = group b by ($0,$1);
d = foreach c generate group,COUNT(b.edu);
dump d;
```



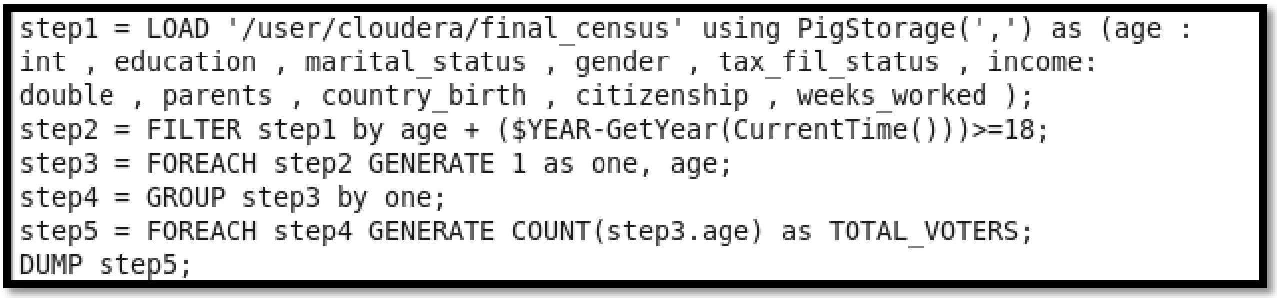
```
(( Children, Male),71669)
(( Children, Female),69827)
(( 9th grade, Male),8755)
(( 9th grade, Female),9780)
(( 10th grade, Male),10384)
(( 10th grade, Female),12187)
(( 11th grade, Male),9690)
(( 11th grade, Female),10815)
(( 5th or 6th grade, Male),4761)
(( 5th or 6th grade, Female),4992)
(( 7th and 8th grade, Male),11518)
(( 7th and 8th grade, Female),12609)
(( Less than 1st grade, Male),1133)
(( Less than 1st grade, Female),1279)
(( High school graduate, Male),63857)
(( High school graduate, Female),80977)
```

Use Case 2 : ‘Weaponizing’ Data

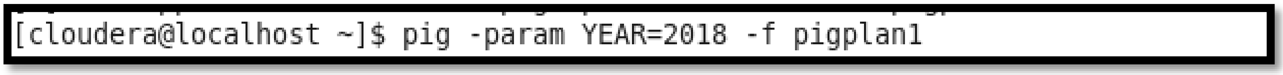
- **The eligible population** refers specifically to citizens 18 years of age and older or the citizen voting-age population.
- By the Census Data we can come up with a number which will say how many people are entering the eligible age to vote in a particular year.
- So a political party can come up with a mindset or scheme to attract those people to their political party who falls under that criteria.
- The same criteria also goes for the Senior citizen, where how many people are entering into the senior age category and plans can be made according to it.
- **Tasks** included are: Task 9 to Find the count of voters in x years and Task10 to Find the count of Senior Citizen in x years and Tasks 16 i.e Country of birth wise count for US Citizenship by Naturalization.

OUTPUT :

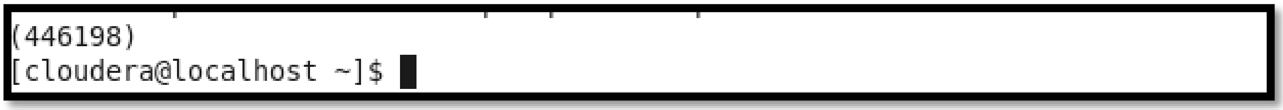
Screenshot of Task 9 :



```
step1 = LOAD '/user/cloudera/final_census' using PigStorage(',') as (age :
int , education , marital_status , gender , tax_fil_status , income:
double , parents , country_birth , citizenship , weeks_worked );
step2 = FILTER step1 by age + ($YEAR-GetYear(CurrentTime()))>=18;
step3 = FOREACH step2 GENERATE 1 as one, age;
step4 = GROUP step3 by one;
step5 = FOREACH step4 GENERATE COUNT(step3.age) as TOTAL_VOTERS;
DUMP step5;
```

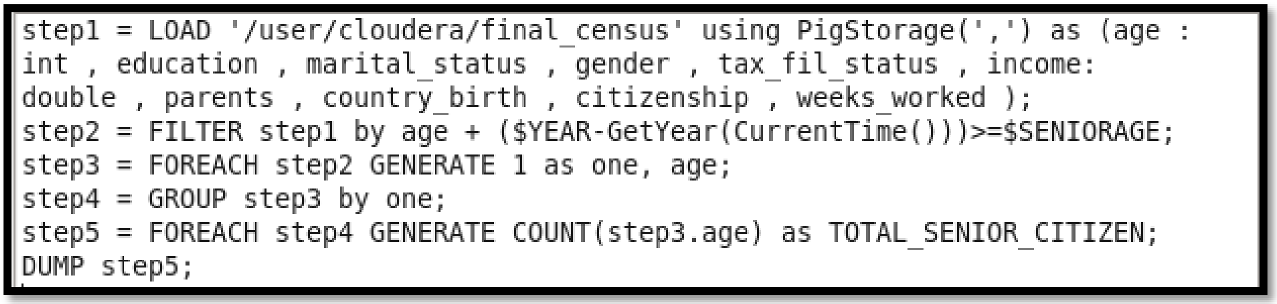


```
[cloudera@localhost ~]$ pig -param YEAR=2018 -f pigplan1
```

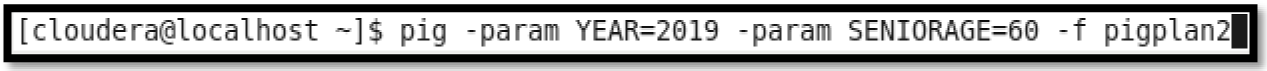


```
(446198)
[cloudera@localhost ~]$ █
```

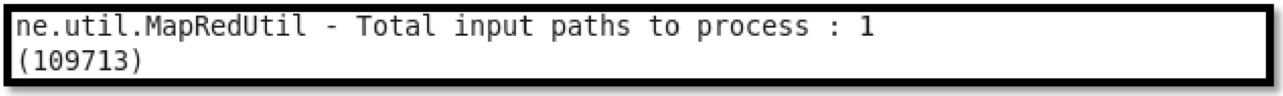
Screenshot of Task 10 :



```
step1 = LOAD '/user/cloudera/final_census' using PigStorage(',') as (age :
int , education , marital_status , gender , tax_fil_status , income:
double , parents , country_birth , citizenship , weeks_worked );
step2 = FILTER step1 by age + ($YEAR-GetYear(CurrentTime()))>=$SENIORAGE;
step3 = FOREACH step2 GENERATE 1 as one, age;
step4 = GROUP step3 by one;
step5 = FOREACH step4 GENERATE COUNT(step3.age) as TOTAL_SENIOR_CITIZEN;
DUMP step5;
```

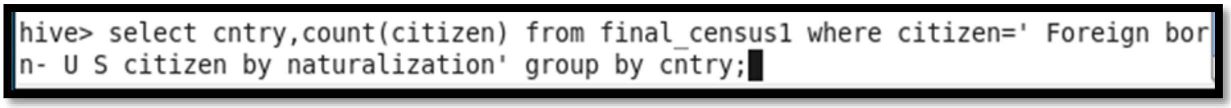


```
[cloudera@localhost ~]$ pig -param YEAR=2019 -param SENIORAGE=60 -f pigplan2
```

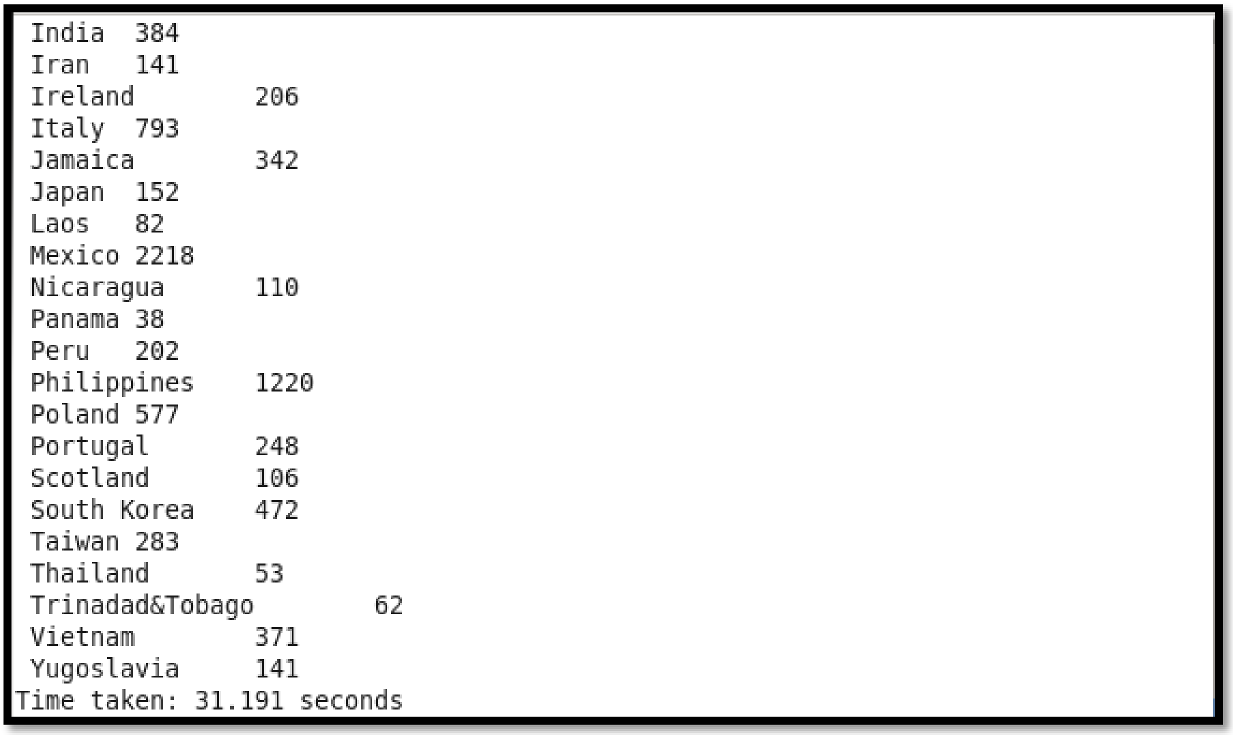


```
ne.util.MapRedUtil - Total input paths to process : 1
(109713)
```

Screenshot of Task 16 :



```
hive> select cntry,count(citizen) from final_census1 where citizen=' Foreign bor
n- U S citizen by naturalization' group by cntry;
```



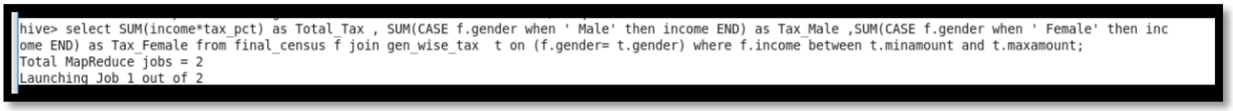
```
 India  384
 Iran   141
 Ireland        206
 Italy  793
 Jamaica        342
 Japan  152
 Laos   82
 Mexico 2218
 Nicaragua      110
 Panama 38
 Peru   202
 Philippines    1220
 Poland 577
 Portugal       248
 Scotland       106
 South Korea    472
 Taiwan 283
 Thailand       53
 Trinadad&Tobago        62
 Vietnam        371
 Yugoslavia     141
Time taken: 31.191 seconds
```

Use Case 3 : Financial Statement Analysis

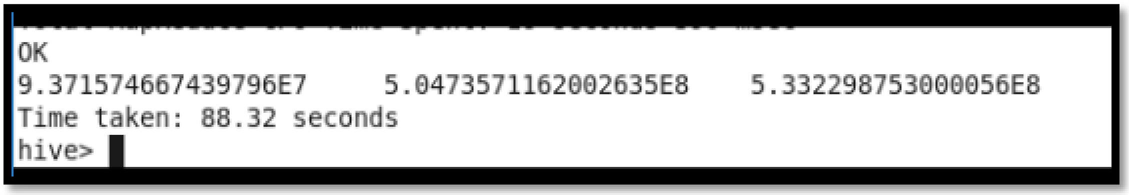
- The most important benefit with financial statement analysis is that it provides an idea to the investors about deciding on investing their funds.
- As being an investor we can get the Per Capita Income (PCI) of the country and decide on the investment part.
- We can get a total report of all the tax analysis done on gender wise or in total.
- **Tasks** includes are Tasks 4 : Tax analysis total and gender wise and Task 5 : Per Capita Income (PCI) analysis consolidated gender wise and category wise and Tasks 15 to Find the Non-US citizen tax filer status.

OUTPUT :

Screenshot of Task 4 :

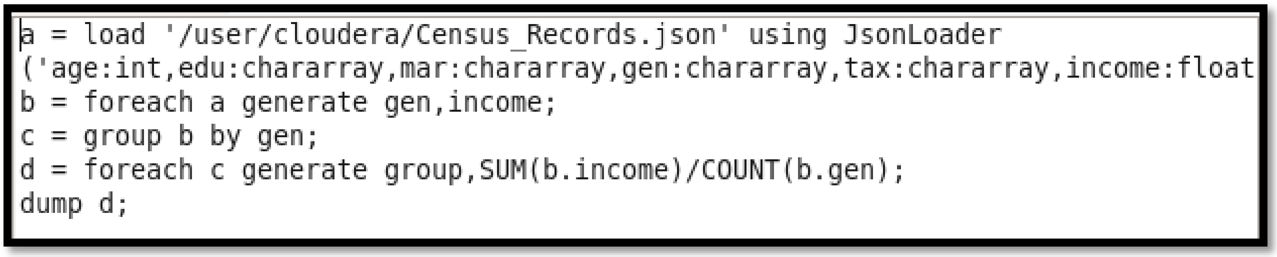


```
hive> select SUM(income*tax_pct) as Total_Tax , SUM(CASE f.gender when ' Male' then income END) as Tax_Male ,SUM(CASE f.gender when ' Female' then inc
ome END) as Tax_Female from final_census f join gen_wise_tax  t on (f.gender= t.gender) where f.income between t.minamount and t.maxamount;
Total MapReduce jobs = 2
Launching Job 1 out of 2
```

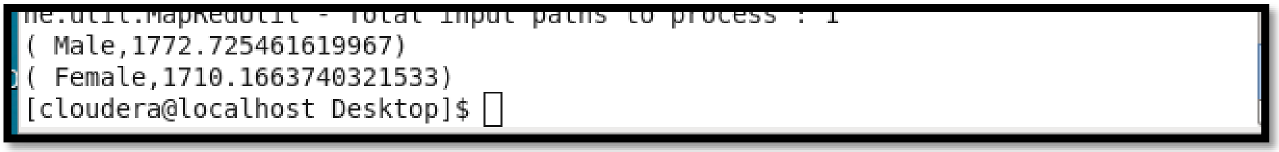


```
OK
9.371574667439796E7     5.0473571162002635E8    5.332298753000056E8
Time taken: 88.32 seconds
hive>
```

Screenshot of Task 5 :

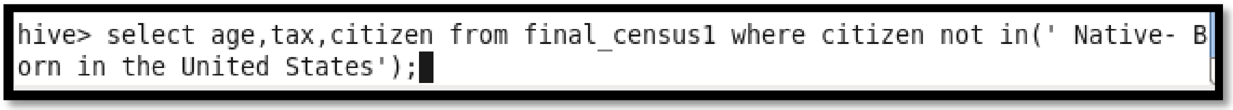


```
a = load '/user/cloudera/Census_Records.json' using JsonLoader
('age:int,edu:chararray,mar:chararray,gen:chararray,tax:chararray,income:float
b = foreach a generate gen,income;
c = group b by gen;
d = foreach c generate group,SUM(b.income)/COUNT(b.gen);
dump d;
```

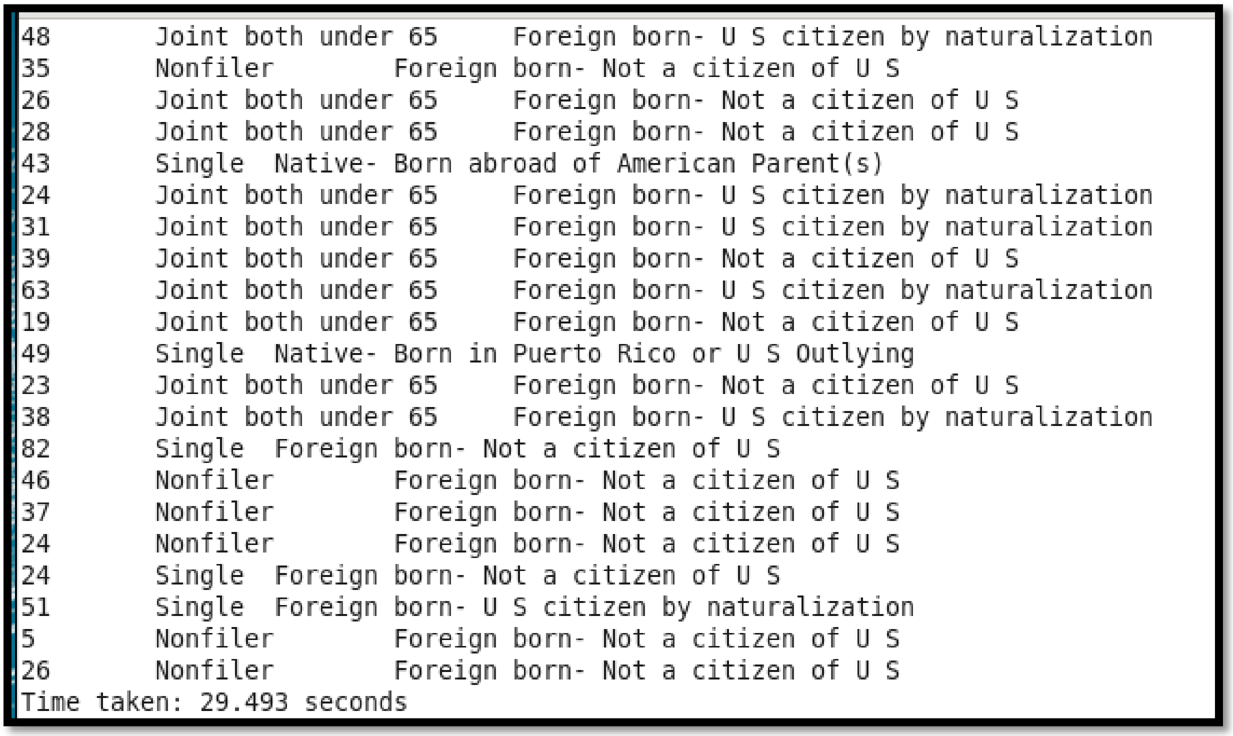


```
ne.util.MapRedUtil - Total input paths to process : 1
( Male,1772.725461619967)
( Female,1710.1663740321533)
[cloudera@localhost Desktop]$ 
```

Screenshot of Tasks 15 :



```
hive> select age,tax,citizen from final_census1 where citizen not in(' Native- B
orn in the United States');
```



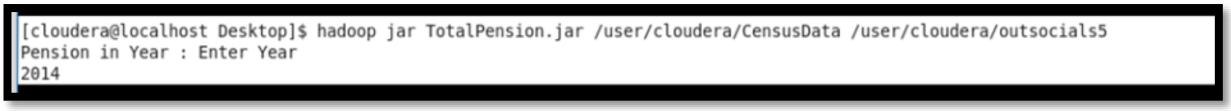
```
48       Joint both under 65     Foreign born- U S citizen by naturalization
35       Nonfiler        Foreign born- Not a citizen of U S
26       Joint both under 65     Foreign born- Not a citizen of U S
28       Joint both under 65     Foreign born- Not a citizen of U S
43       Single  Native- Born abroad of American Parent(s)
24       Joint both under 65     Foreign born- U S citizen by naturalization
31       Joint both under 65     Foreign born- U S citizen by naturalization
39       Joint both under 65     Foreign born- Not a citizen of U S
63       Joint both under 65     Foreign born- U S citizen by naturalization
19       Joint both under 65     Foreign born- Not a citizen of U S
49       Single  Native- Born in Puerto Rico or U S Outlying
23       Joint both under 65     Foreign born- Not a citizen of U S
38       Joint both under 65     Foreign born- U S citizen by naturalization
82       Single  Foreign born- Not a citizen of U S
46       Nonfiler        Foreign born- Not a citizen of U S
37       Nonfiler        Foreign born- Not a citizen of U S
24       Nonfiler        Foreign born- Not a citizen of U S
24       Single  Foreign born- Not a citizen of U S
51       Single  Foreign born- U S citizen by naturalization
5        Nonfiler        Foreign born- Not a citizen of U S
26       Nonfiler        Foreign born- Not a citizen of U S
Time taken: 29.493 seconds
```

Use Case 4 : EcoSocialism

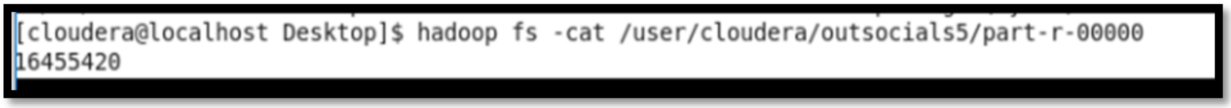
- Ecosocialism is a vision of a transformed society in harmony with nature, and the development of practices that can attain it.
- We can calculate the total amount dispensed on pension in x years, so that all people who are attending a pension after certain year, for them a fixed amount can be fixed.
- We can also find the Total amount to be given to the students depending upon their Parental Status. Different amount can be fixed to different scenario.
- Scholarship funding can be categorized based upon :
  1. Both parent present
  2. Only one parent present
  3. No parent but guardian
  4. No one in the universe
- For women who fall under the widow category or divorce category, special treatment or plan can be made by offering them job so that they can run their basic expenditure smoothly.
- **Tasks** included are Task 6 to find the Total amount dispensed on pension in x years , Task 7 to find the Total amount dispensed on scholarship in current years and Task 8 for given range employable female widowed and divorced count.
- **Data Validation :** Yes.
- **Constraint :** User Input Can be Only Numbers.

OUTPUT :

Screenshot of Task 6 :

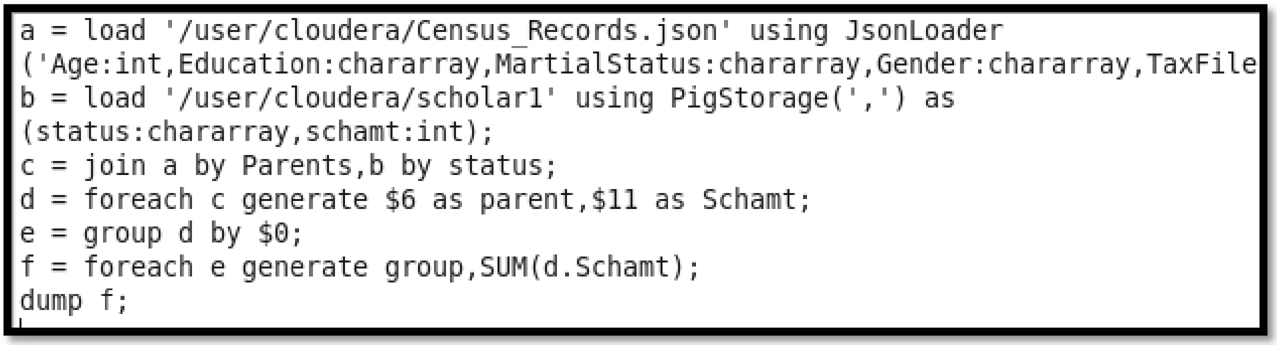


```
[cloudera@localhost Desktop]$ hadoop jar TotalPension.jar /user/cloudera/CensusData /user/cloudera/outsocials5
Pension in Year : Enter Year
2014
```

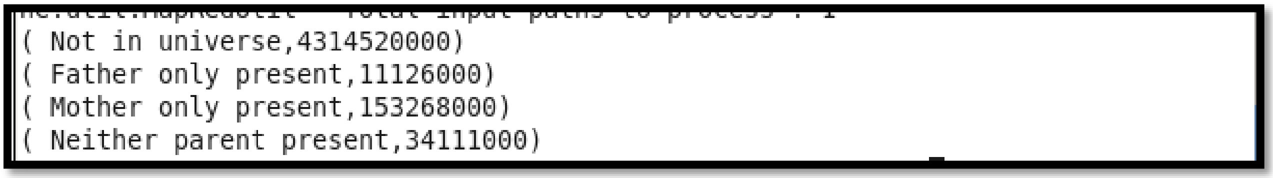


```
[cloudera@localhost Desktop]$ hadoop fs -cat /user/cloudera/outsocials5/part-r-00000
16455420
```

Screenshot for Task 7 :

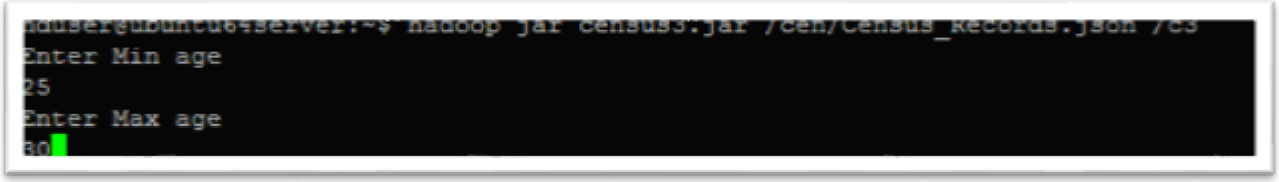


```
a = load '/user/cloudera/Census_Records.json' using JsonLoader
('Age:int,Education:chararray,MartialStatus:chararray,Gender:chararray,TaxFile
b = load '/user/cloudera/scholar1' using PigStorage(',') as
(status:chararray,schamt:int);
c = join a by Parents,b by status;
d = foreach c generate $6 as parent,$11 as Schamt;
e = group d by $0;
f = foreach e generate group,SUM(d.Schamt);
dump f;
```

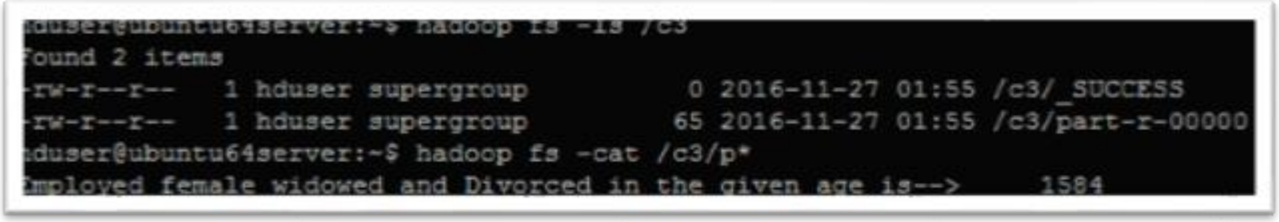


```
ne.util.MapRedUtil - Total input paths to process : 1
( Not in universe,4314520000)
( Father only present,11126000)
( Mother only present,153268000)
( Neither parent present,34111000)
```

Screenshot for Task 8 :



```
hduser@ubuntu64server:~$ hadoop jar census3.jar /cen/Census_Records.json /c3
Enter Min age
25
Enter Max age
30
```



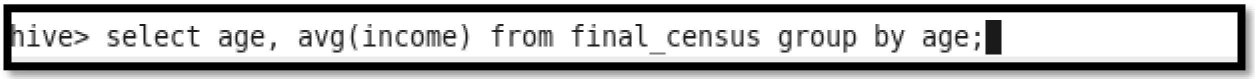
```
hduser@ubuntu64server:~$ hadoop fs -ls /c3
Found 2 items
-rw-r--r--   1 hduser supergroup          0 2016-11-27 01:55 /c3/_SUCCESS
-rw-r--r--   1 hduser supergroup         65 2016-11-27 01:55 /c3/part-r-00000
hduser@ubuntu64server:~$ hadoop fs -cat /c3/p*
Employed female widowed and Divorced in the given age is-->      1584
```

Use Case 5 : Customer Targeting and Product Targeting

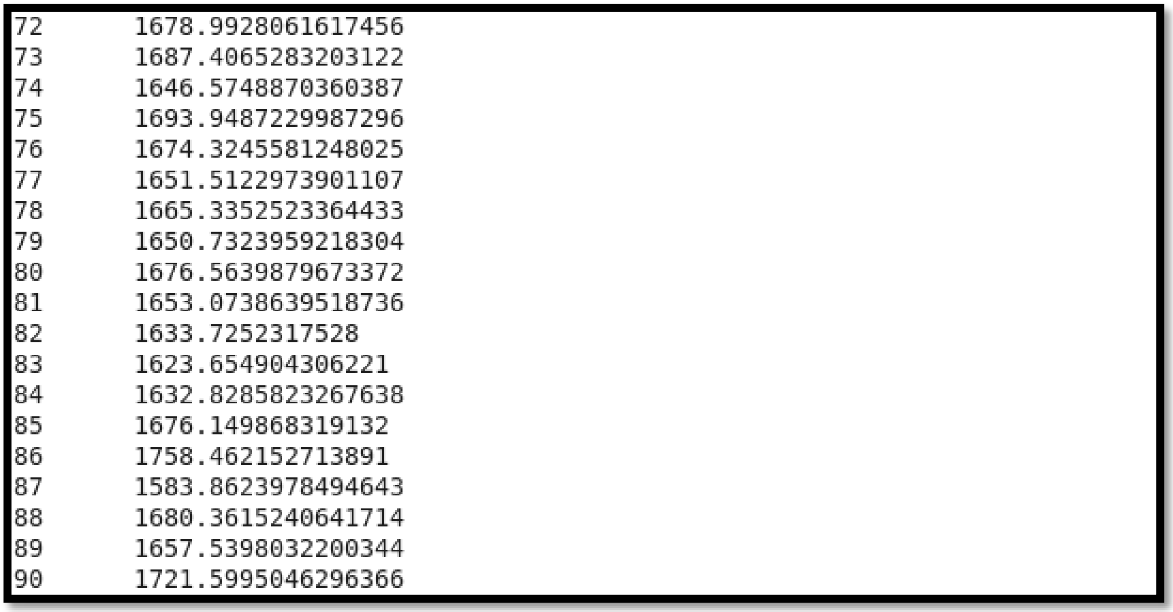
- Customer targeting is the business process that defines which customers to market to. For each direct marketing campaign, be it email or direct mail or contacted via telephone, there is a decision to be made on who will, and who will not receive the campaign.
- Depending upon the count of total number of male or female we can target that set of population for our product selling and target them.
- Customer base analysis can also done. Depending upon the company conditions or plan results from the data can be collected and thus implementation can be done on the output we have got.
- For eg : We can calculate an average amount a user spends on a particular day or month. So depending upon the amount we can create our new product on that price range only so that it suits all the customer budget.
- **Tasks** included are Task 11 to find the Total number of Male/Female and Task 14 Customer base analysis.

OUTPUT :

Screenshot of Task 14 :

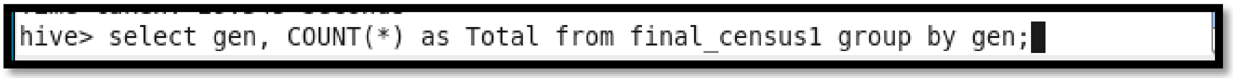


```
hive> select age, avg(income) from final_census group by age;
```

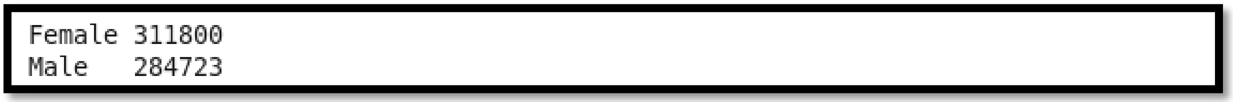


```
72      1678.9928061617456
73      1687.4065283203122
74      1646.5748870360387
75      1693.9487229987296
76      1674.3245581248025
77      1651.5122973901107
78      1665.3352523364433
79      1650.7323959218304
80      1676.5639879673372
81      1653.0738639518736
82      1633.7252317528
83      1623.654904306221
84      1632.8285823267638
85      1676.149868319132
86      1758.462152713891
87      1583.8623978494643
88      1680.3615240641714
89      1657.5398032200344
90      1721.5995046296366
```

Screenshot of Task 11 :



```
hive> select gen, COUNT(*) as Total from final_census1 group by gen;
```



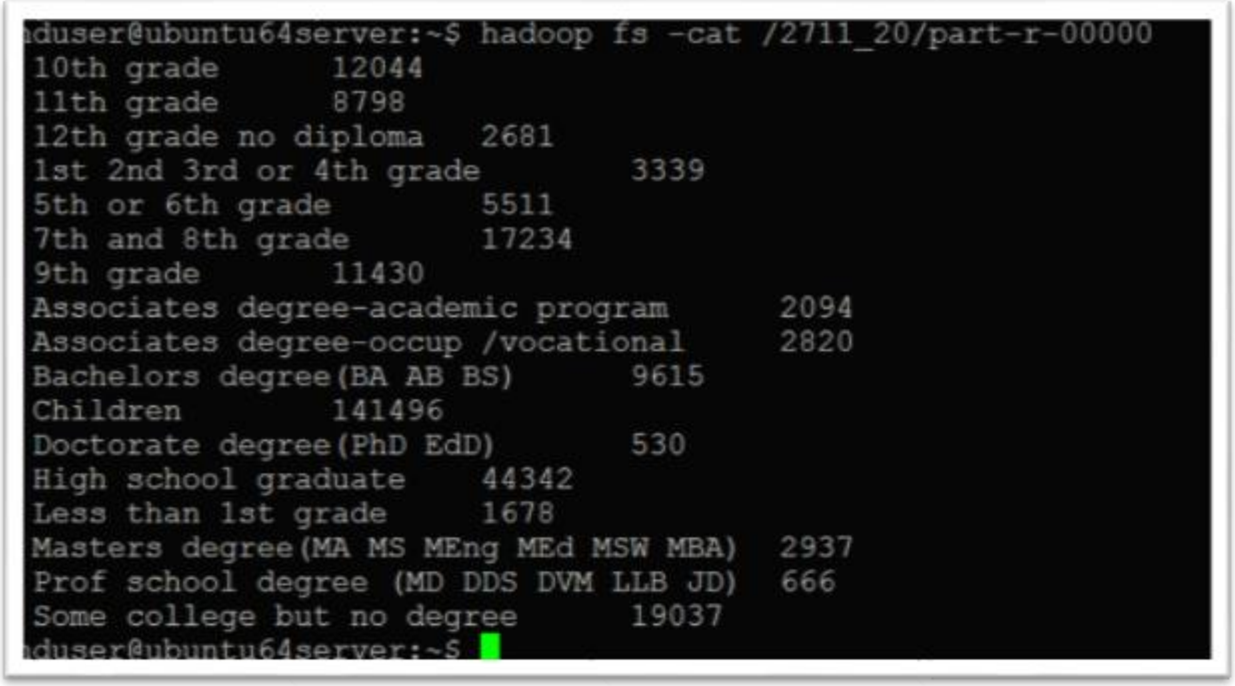
```
Female 311800
Male   284723
```

Use Case 6 : Career Junction

- India Career Portal is a web-based software for colleges to help improve their placement performance by automating the placement activities and taking it online.
- Improve opportunities for students of the college by providing a regularly updated knowledge center on career guidance, resume and interview preparation and tools to practice and improve skills.
- For every degree we have total number of candidates. We can now contact them for further studies or for Job opportunities.
- **Tasks** completed are Task 13 i.e Degree wise count for employability and Task 12 i.e Citizen and immigrants count for employed lot.

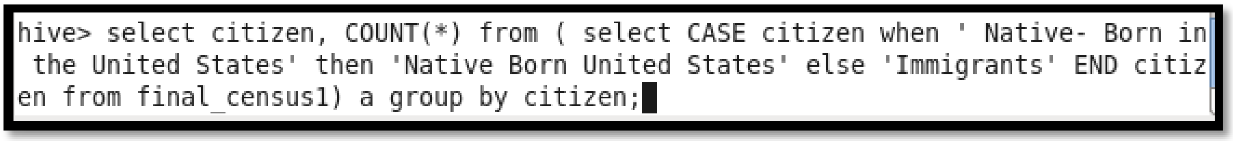
OUTPUT :

Screenshot of Task 13 :

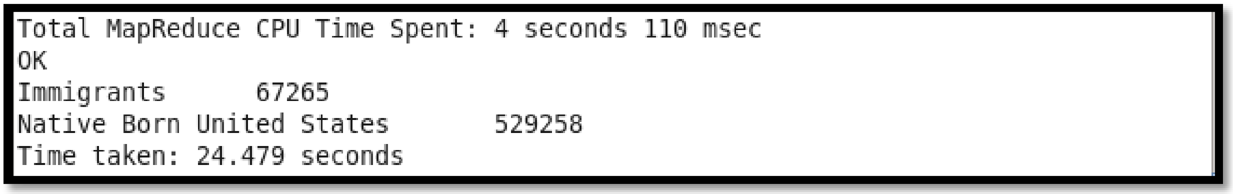


```
duser@ubuntu64server:~$ hadoop fs -cat /2711_20/part-r-00000
10th grade      12044
11th grade      8798
12th grade no diploma   2681
1st 2nd 3rd or 4th grade        3339
5th or 6th grade        5511
7th and 8th grade       17234
9th grade       11430
Associates degree-academic program      2094
Associates degree-occup /vocational     2820
Bachelors degree(BA AB BS)      9615
Children        141496
Doctorate degree(PhD EdD)       530
High school graduate    44342
Less than 1st grade     1678
Masters degree(MA MS MEng MEd MSW MBA)  2937
Prof school degree (MD DDS DVM LLB JD)  666
Some college but no degree      19037
duser@ubuntu64server:~$
```

Screenshot for Task 12 :



```
hive> select citizen, COUNT(*) from ( select CASE citizen when ' Native- Born in
 the United States' then 'Native Born United States' else 'Immigrants' END citiz
en from final_census1) a group by citizen;
```



```
Total MapReduce CPU Time Spent: 4 seconds 110 msec
OK
Immigrants      67265
Native Born United States       529258
Time taken: 24.479 seconds
```

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**.**

Java MapReduce Program: Hadoop MapReduce 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 MapReduce.

SOFTWARE Used:

1) Virtual Box
2) Eclipse
3) Ubuntu Terminal (for MapReduce)
4) Cloudera OS (for HIVE)

SYSTEM REQUIREMENT:

- Minimum 50 Gb of HardDrive Space.
- Minimum 4 Gb RAM.
- Next Generation Processor Chips like Intel I3 and so on.

CONCLUSION : Thus a Complete Analysis of Census Data happens which helps in maintaining the needs as per requirements.