E2

"An Edu-Eco Project"

ABSTRACT

Growth of a Country depends on various Factors, which includes Economic, Educational, Budget Plans, and Sufficient Resources. There must always an effective mechanism used to analyse the current population in order to understand the requirements in terms of various factors

Oliver Prasanna R

Big Data Developer

Objective:

Growth of a Country depends on various Factors, which includes Economic, Educational, Budget Plans, and Sufficient Resources etc. There must always an effective mechanism used to analyse the current population in order to understand the requirements in terms of various factors. Data collected from the population are of different types. To handle such large different types of data, we call them as "**Big Data**", Hadoop Technique is required.

Outcome of this Project:

To generate reports and hence,

- Understand Economic Requirements in terms of Education, Family Support, and Senior Citizens.
- Understand Education need for Various Category of People in terms of Children, Bachelors, and Masters.
- Understand Current Trends Citizen Vs Immigrants
- Understand Employment Opportunities
- Understand Requirements for Female in terms of Economic and Education

Abstract: Given the Census Population Data with Various Fields

Age	Education	Martial Status	Gender	Tax	Income	Parent	COB	City	Week Worked

List of Activities:

Education

- Total count of male/female based on education
- Total count of employed/unemployed based on education
- Total count for people in age range of 18-25 based on education

Finance

- Tax analysis total and gender wise - hive
- Per Capita Income (PCI) analysis consolidated, gender wise and category wise - done - hive

Social

- Total amount dispensed on pension in x year(s)
- Total amount dispensed on scholarship in current year
- For given age range employable female widowed and divorced count

Planning:

- Voter(s) count in x year(s)
- Senior Citizen(s) count in x year(s)
- Total number of Male/Female
- Citizens and immigrants count for employed lot

Miscellaneous

- Degree wise count for employability - done

- Customer base analysis – done
- Non-US citizen(s) tax filer status - done
- Country of birth wise count for US citizenship by naturalisation - done

Hardware Requirements:

- 8 GB RAM
- 64 Bit OS

Software Requirements:

- Oracle Virtual Box
- Eclipse
- Ubuntu
- Hadoop
- Putty

Assumptions:

- Oracle Virtual Box – Configurations are set correctly.
- Ubuntu is lying on the Virtual Box and it is powered on
- Putty is configured with the IP address of Ubuntu.
- Hadoop Folder must be extracted and all the services of the hadoop is running. Configuration to be made in the XML are set.
- Confirmation Box Below that Everything is Set Right.

```
hduser@ubuntu64server:~$ jps
2034 NameNode
2114 DataNode
4755 Jps
2441 NodeManager
2203 ResourceManager
hduser@ubuntu64server:~$
```

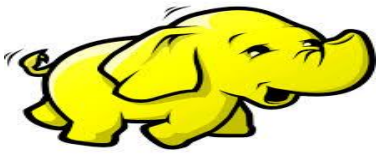
A Brief Introduction about Various Technologies

Big Data

Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate to deal with them.

4 V's of Big Data:

Hadoop



- **Apache Hadoop** is 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.
- All the modules in Hadoop are designed with a fundamental assumption that hardware failures are a common occurrence and should be automatically handled by the framework.

HIVE

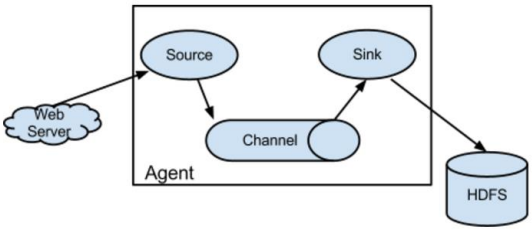
- **Apache Hive** is a data warehouse infrastructure built on top of Hadoop for providing data summarization, query, and analysis
- Hive gives an SQL-like interface to query data stored in various databases and file systems that integrate with Hadoop.

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, Apache Tez, or Apache Spark. Pig Latin abstracts the programming from the Java Map Reduce idiom into a notation which makes Map Reduce programming high level, similar to that of SQL for RDBMSs
- Pig Latin can be extended using User Defined Functions (UDFs) which the user can write in Java, Python, JavaScript, Ruby or Groovy[2] and then call directly from the language.

- **HBase** is an open source, non-relational, distributed database modelled after Google's Big Table and is written in Java. It is developed as part of Apache Software Foundation's Apache Hadoop project and runs on top of HDFS (Hadoop Distributed File System), providing Big Table-like capabilities for Hadoop.
- It provides a fault-tolerant way of storing large quantities of sparse data

- Sqoop is a tool designed to transfer data between Hadoop and relational database servers. It is used to import data from relational databases such as MySQL, Oracle to Hadoop HDFS, and export from Hadoop file system to relational databases.



- Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data. It has a simple and flexible architecture based on streaming data flows. It is robust and fault tolerant with tunable reliability mechanisms and many failover and recovery mechanisms. It uses a simple extensible data model that allows for online analytic application.

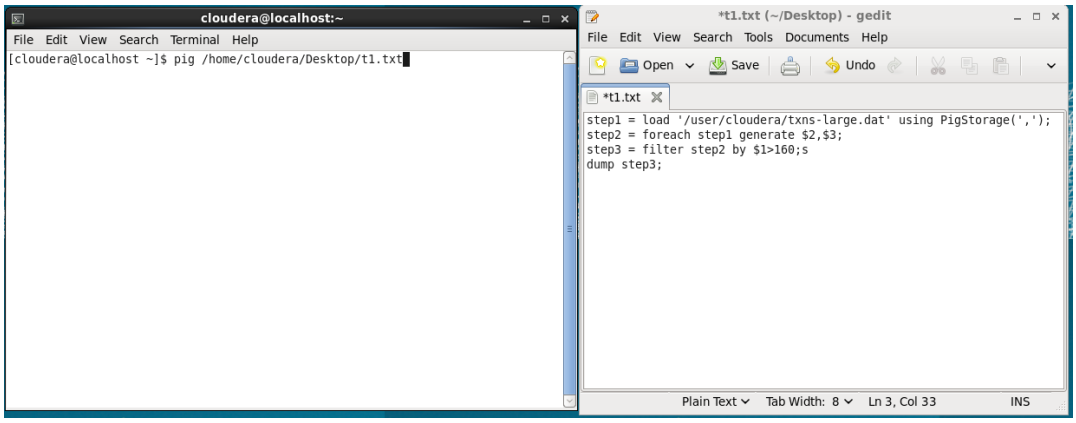
Interfaces:

Map Reduce:

MyRecordReader.java | MyMapper.java | MyReducer.java | MyDriver.java

```
1
2  import java.io.IOException;
7
8
9
10 public class MyMapper extends Mapper<MyKey, MyValue, Text, Text> {
11
12     public void map(MyKey inpK, MyValue inpV, Context c) throws IOException, InterruptedException{
13         int val = c.getConfiguration().getInt("c",0);
14         String citizen=inpK.getCitizen();
15         int ccode=inpV.getAmt();
16
17         if(val==1 && citizen.trim().equals("Male"))
18             c.write(new Text(citizen),new Text("1"));
19
20         if(val==2 && ccode>0 && !citizen.trim().equals("Male"))
21             c.write(new Text(citizen),new Text("1"));
22
23
24
25     }
26 }
27
28
```

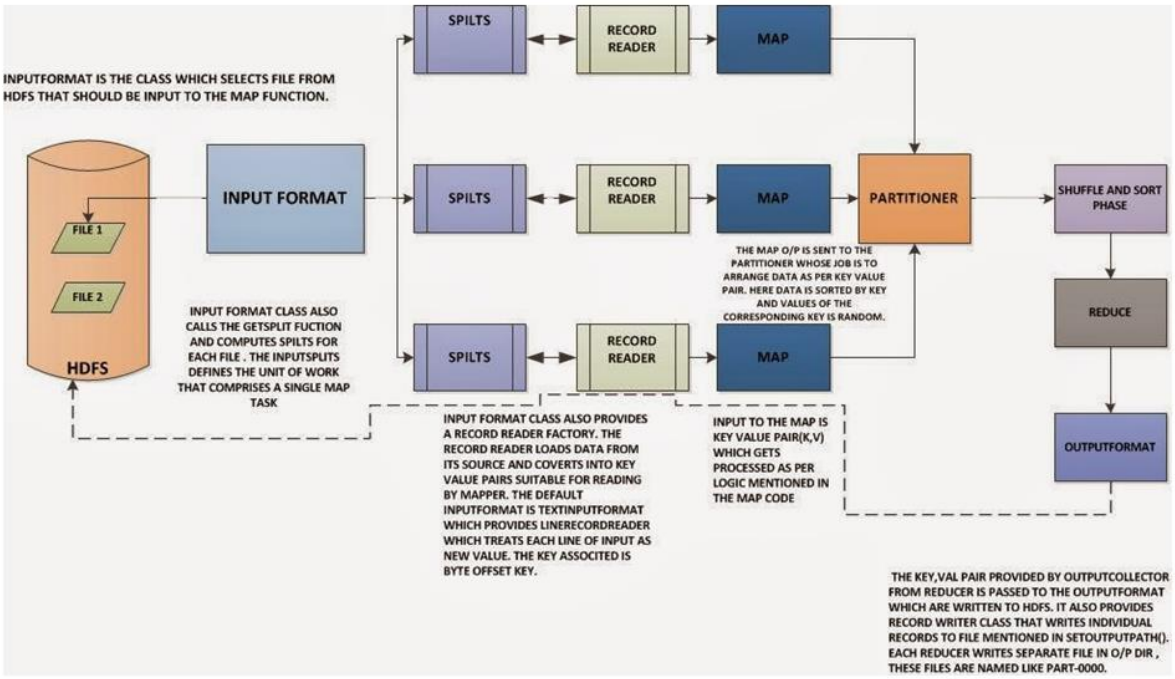
PIG:



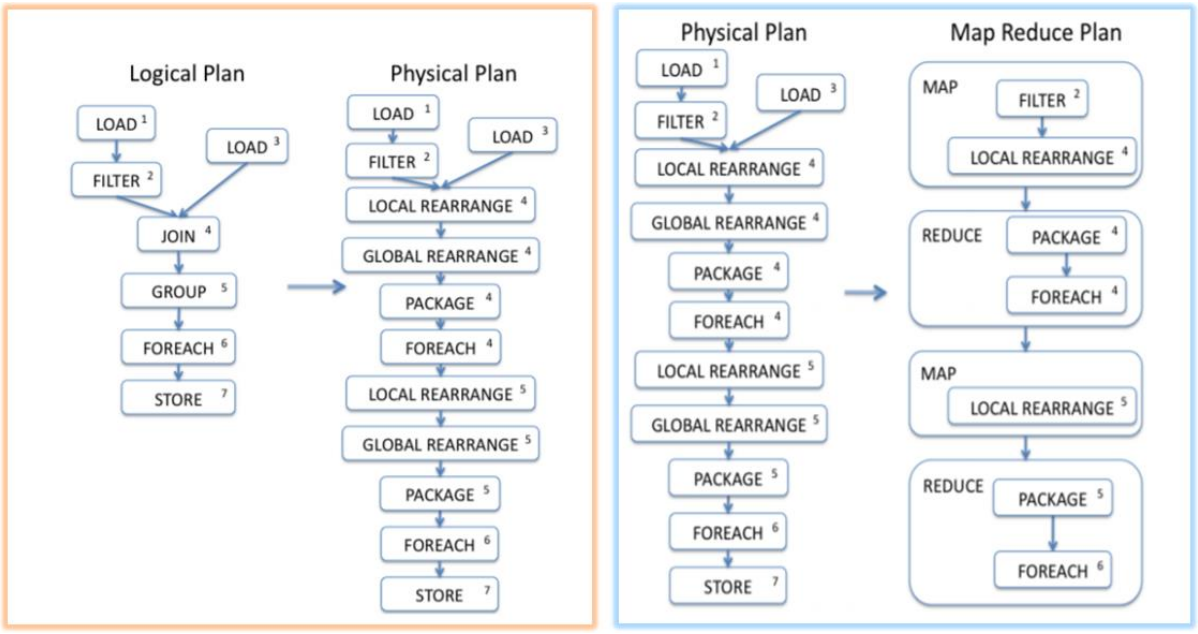
HIVE:

```
hive> select * from ecom1 where amt>160;
```

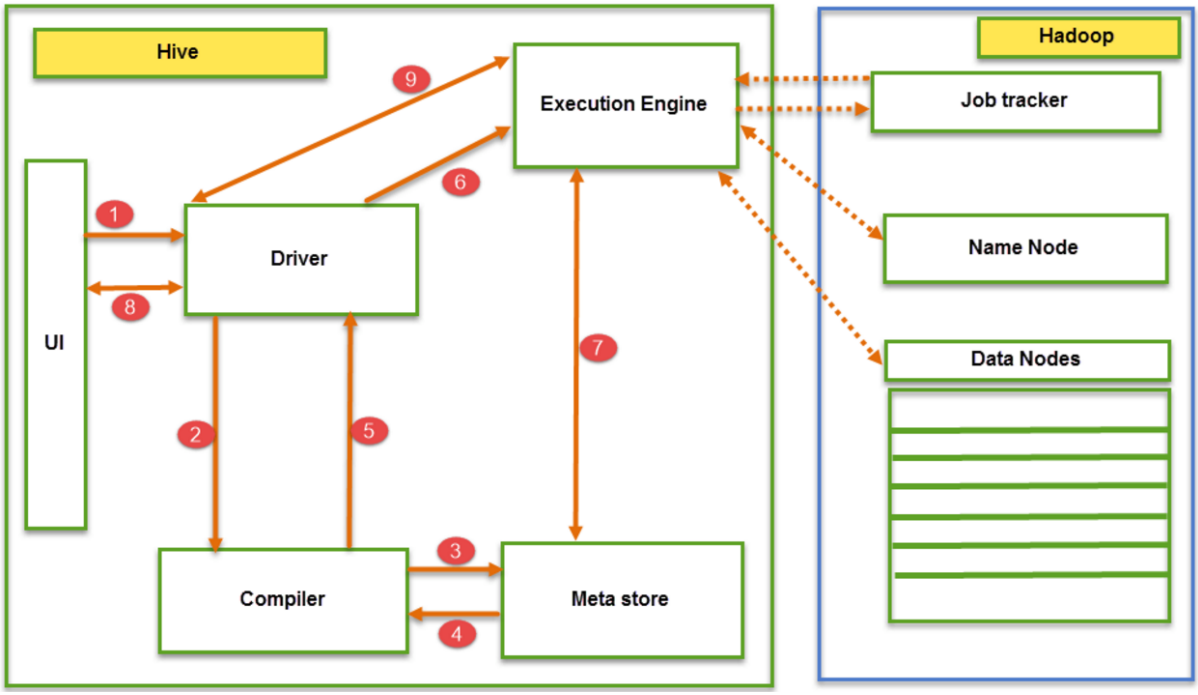
Data Flow – Map Reduce Technique



PIG DATA FLOW MECHANISM:



HIVE DATA FLOW MECHANISM:



Use – Case Generation

Miscellaneous:

Task 1: Degree wise count for employability

Use Case: ***Gateway to Success!***

- **Input :** Census_Records.json File
- **Key**: Education, **Value**: Week Worked [For Map Reduce Technique]
- **Output**: Education based Employers.
- **Concept used:** Advanced Map Reduce, HIVE and PIG.

Description: Recent Market has wide variety of Technologies lies with it. Most of the younger generation were not aware of these technologies. They may know the existence of the technology, but not a complete practical Exposure. We are here to identify the set of Younger generation who is graduate or who will be a graduate in future and educate them for the employability.

Why this Report: To generate a list of customers, basically students for education.

Progress: Organization will have an opportunity to promote their new set of products to the various set of customers and hence can provide service they required.

HIVE:

```
hive> select edu,COUNT(*) from final_census1 where ww=0 group by edu;
```

```
Total MapReduce CPU Time Spent: 4 seconds 440 msec
OK
 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
```

Map Reduce:

```
hduser@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
```

PIG:

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

b = foreach a generate $1,$9;

c = filter b by ww==0;

d = group c by $0;

e = foreach d generate group,COUNT(c.$0);

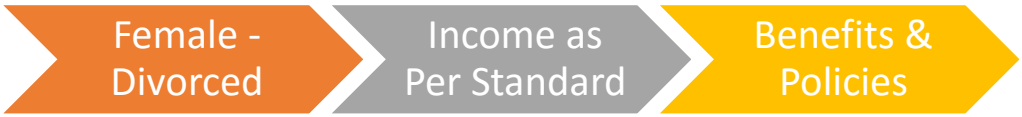
dump e;

```
( Children,141496)
( 9th grade,11430)
( 10th grade,12044)
( 11th grade,8798)
( 5th or 6th grade,5511)
```

Task 2: Customer base analysis - done

Use Case: To Serve In Need

- **Input**: Census_Records.json File
- **Output**: Set of Records based on the Condition specified by the Developer.
- **Concept used:** HIVE and PIG.



Description: Government brings up new policies and benefits to the people. To increase more flexibility and to help Female Population, Government had decided to bring a new Policy. The new Policy is for the Female Widowers whose age in a particular Range. Based on their Income, An Standard set of Amount will be given to her family on a monthly basis for a certain period of time.

Why this Report: To generate a list of Female Population who are widowed.

Progress: Country Progress comparing with other is enhanced and people happiness is achieved at its best.

PIG:

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

b = foreach a generate age, gen, income,mar;

d = filter b by ((gen==' Female' and mar==' Divorced') and (age>45 and age<60)) ;

dump d;
```

```
(51, Female,3067, Divorced)
(56, Female,531, Divorced)
(47, Female,1068, Divorced)
(47, Female,1543, Divorced)
(46, Female,2772, Divorced)
(51, Female,282, Divorced)
```

HIVE :

```
select age, gen, income, mar from final_census1 where gen=' Female' and age between 45 and 60
and mar=' Divorced';
```

```
45      Female 1116.64  Divorced
48      Female 1078.72  Divorced
52      Female 1858.26  Divorced
48      Female 1031.39  Divorced
55      Female 1785.28  Divorced
```

Task 3: Non-US citizen(s) tax filer status

Use Case Alert – Tax – Not US Based

- **Input**: Census_Records.json
- **Key**: Citizen, **Value**: Tax Profile Status
- **Output**: Set of User Information based on Tax Profiler Status.
- **Concept used:** Advanced Map Reduce, HIVE and PIG.

Description: Tax payable to the government is a duly responsibility for every People who stay in a country. Tax payable is irrespective of whether the Individual is a Citizen of US or not US Based. This needs an immediate arrest. Government decided to list a set of Individual to identify the tax profile status and based on that necessary action to be taken.

Why this Report: To generate a list of Individuals based on Non –US Citizenship and their status.

Progress: To create a good profit to the Government, and hence Government can use this for providing new services to their People.

HIVE:

```
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
```

Advanced Map Reduce:

```
 Native- Born in the United States       Joint both under 65
 Native- Born in the United States       Joint both under 65
 Native- Born in the United States       Single
 Native- Born in the United States       Joint both under 65
 Native- Born in the United States       Joint both 65+
 Native- Born in the United States       Single
 Native- Born in the United States       Single
 Native- Born in the United States       Single
```

PIG:

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

b = foreach a generate $0,$1,$4,$5,$8 as citizen;

c = filter b by citizen!=' Native- Born in the United States';

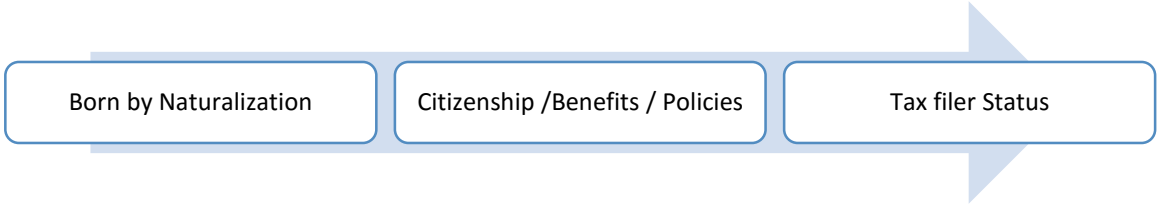
dump c;

```
(63, High school graduate, Joint both under 65,2590.42, Foreign born- U S citize
n by naturalization)
(19, 5th or 6th grade, Joint both under 65,1329.61, Foreign born- Not a citizen
of U S )
(49, High school graduate, Single,1198.34, Native- Born in Puerto Rico or U S Ou
tlying)
(23, High school graduate, Joint both under 65,2632.78, Foreign born- Not a citi
zen of U S )
(38, Some college but no degree, Joint both under 65,1386.91, Foreign born- U S
citizen by naturalization)
(82, Some college but no degree, Single,1230.37, Foreign born- Not a citizen of
U S )
```

Task 4: Country of birth wise count for US citizenship by naturalisation

Use Case: An Action to Control

- **Input**: Census_Records.json File
- **Output**: Total count of Naturalisation Count Based on Country.
- **Concept used:** Advanced Map Reduce, HIVE and PIG.



Description: Naturalization is the process by which U.S. citizenship is granted to a foreign citizen or national after he or she fulfils the requirements established by Congress in the Immigration and Nationality Act. It is also an important factor for the Government to analyse the count of

Citizenship given and identify their specific Countries. This will help the Government to track their tax filer status and their Citizenship validity.

Why this Report: To generate a list of Individuals by Naturalization given Citizenship.

Progress: To maintain a good Control over various services offered by the Government and hence it should to go the right people.

HIVE:

```
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
```

PIG:

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

b = foreach a generate $7,$8;

c = filter b by citizen==' Foreign born- U S citizen by naturalization';

d = group c by $0;

e = foreach d generate group, COUNT (c.$0);

dump e;
```

```
( Ecuador,192)
( England,496)
( Germany,1054)
( Hungary,187)
( Ireland,206)
( Jamaica,342)
( Vietnam,371)
( Cambodia,75)
( Columbia,397)
( Honduras,87)
( Portugal,248)
```

Education:

Task 1: Total count of male/female based on education. : If it bachelor, how many male/female

Use Case: Education for Betterment

- **Input**: Census_Records.json
- **Output**: Education Ratio
- **Concept used:** HIVE and PIG.

Description: Education should be one of the most primary requirement for every Individuals of a Country. Education is not only for the Individual betterment, it is also for the Country's Growth. We can generate list of people along with their education. There could be some people who were not able to do the studies or not able to continue education due to their economic issues.

Why this Report: To Create plan for education for every categories of People.

Progress: Country People are more benefitted about the plan offered by the Government. Education will therefore help People for getting a Job and increase their source of Income.

PIG:

step1 = load '/user/cloudera/Census_Records.json' using JsonLoader('Age:int,Education:chararray,MartialStatus:chararray,Gender:chararray,TaxFilerStatus:chararray,Income:float,Parents:chararray,CountryOfBirth:chararray,Citizenship:chararray,WeeksWorked:chararray');

step2 = foreach step1 generate $1 as Edu,$3 as Gen;

step3 = group step2 by ($0, $1);

step4 = foreach step3 generate group, COUNT(step2.Gen);

dump step4;

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

HIVE:

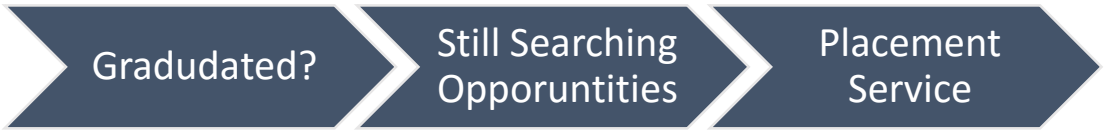
Select edu, gen, COUNT (*) Total from final_census1 group by edu, gen;

```
9th grade       Female 9780
9th grade       Male   8755
Associates degree-academic program      Female 7684
Associates degree-academic program      Male   5266
Associates degree-occup /vocational     Female 9225
Associates degree-occup /vocational     Male   6733
```

Task 2 .Total count of employed/unemployed based on education

Use Case: Understand Industry Requirements

- **Input**: Census_Records.json File
- **Key**: Education , **Value**: Weeks Worked
- **Output**: Categories of Education list along with the Employability Count.
- **Concept used:** Advanced Map Reduce, HIVE and PIG.



Description: People who do well in education, will be able to find the Job Opportunities based on their level of education. There are some categories of people who has done a basic level of education or even higher level. They may not aware to the different Job Openings in the market.

Why this Report: To Create Job Openings for all Categories of People.

Progress: Individual who struggles to find different find Job Opportunities will be benefitted by this type of services. This will definitely create a betterment in all Individual's Economic growth and there by Country's Growth Increased.

Advanced Map Reduce:

```
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
```

PIG: [Employed]

```
step1 = load '/user/cloudera/Census_Records.json' using
JsonLoader('Age:int,Education:chararray,MartialStatus:chararray,Gender:chararray,TaxFilerStatus:ch
ararray,Income:float,Parents:chararray,CountryOfBirth:chararray,Citizenship:chararray,WeeksWorke
d:int');

step2 = foreach step1 generate $1 as Edu, $9 as ww;

step3 = filter step2 by $1>0;
```

step4 = group step3 by $0;

step5 = foreach step4 generate group, COUNT ($1);

dump step5;

```
( 9th grade,7105)
( 10th grade,10527)
( 11th grade,11707)
( 5th or 6th grade,4242)
( 7th and 8th grade,6893)
```

PIG: [Unemployed]

step1 = load '/user/cloudera/Census_Records.json' using JsonLoader('Age:int,Education:chararray,MartialStatus:chararray,Gender:chararray,TaxFilerStatus:chararray,Income:float,Parents:chararray,CountryOfBirth:chararray,Citizenship:chararray,WeeksWorked:int');

step2 = foreach step1 generate $1 as Edu, $9 as ww;

step3 = filter step2 by $1==0;

step4 = group step3 by $0;

step5 = foreach step4 generate group, COUNT ($1);

dump step5;

```
( Children,141496)
( 9th grade,11430)
( 10th grade,12044)
( 11th grade,8798)
( 5th or 6th grade,5511)
```

HIVE:

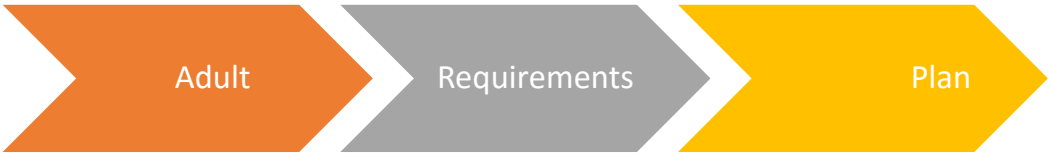
Select edu, SUM (CASE when ww <=0 then '1' else null END) as Employed, SUM (CASE when ww >0 then '1' else null END) as Unemployed from final_census1 group by edu;

```
10th grade      12044.0 10527.0
11th grade      8798.0  11707.0
12th grade no diploma   2681.0  3593.0
1st 2nd 3rd or 4th grade        3339.0  2016.0
5th or 6th grade        5511.0  4242.0
```

Task 3: Total count for people in age range of 18-25 based on education

Use Case: Necessity of an Adult

- **Input**: Census_Records.json
- **Output**: Education count between Age of a particular Range.
- **Concept used:** Advanced Map Reduce, HIVE and PIG.



Description: Population in a country is categorized into infants, Teenager, Adult, Middle-aged, Senior-Citizens and Elder Persons. The Phase where the responsibility place an important role is Adult. The Adult stage has various requirements in terms of education & economic.

Why this Report: To generate a general list based on the education in the current population and age falls between 18-25

Progress: This will make the government to bring in new features. Offers for Higher Education, Finding better Job Opportunities, Marriage Plans, Loan schemes etc.

HIVE:

```
hive> select edu ,COUNT(*) as Total_Peoples  from final_census1 where age betwee
n 18 and 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
```

PIG:

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

b = foreach a generate age, edu;

c = filter b by age>17 and age<26;

j = group c by edu;

d = foreach j generate group, COUNT (c.age);
```

dump d;

```
( 9th grade,1486)
( 10th grade,2411)
( 11th grade,5310)
( 5th or 6th grade,871)
( 7th and 8th grade,989)
```

Finance:

Use Case 1: Tax Trending

- **Input**: Census_Records.json and Tax_1 [Secondary table]
- **Output**: Tax Analysis Status
- **Concept used:** HIVE

Description: It is very Important for the government to Visualize and predict the total tax payable by the population based on their income. This will help the Government to plan their budget on the benefits they bring to the Country. Hence Government should take a list of all People Tax Trending status and calculate the Income which can be obtained from the same.

Why this Report: To Plan the Government Budget and also to alert customers to pay their tax on time.

Progress: Government can bring out multiple plans for the welfare and growth based on the budget.

HIVE:

```
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>
```

Task 2: Per Capita Income (PCI) analysis consolidated, gender wise and category wise

Use Case: Developed vs. Developing

- **Input**: Census_Records.json
- **Output**: PCI status based on Total Population, Category and Gender Wise.
- **Concept used:** HIVE and PIG.

Description: Average income /per capita income which is the total income divided the population. It serves as an important tool of comparing different nations and classifying them as rich countries or low income countries. It also gives us an information about what an average person is likely to earn and therefore is an important indicator of development. World Bank has specified the criterion for categorising countries into developed, developing and under developed countries. It is based on assumption that in countries with higher capita income, people will have standard of living and would be in a position to not only manage their basic necessities but other things require to lead a life.

Why this Report: To Identify the Country Economic status – where do they stand?

Progress: Government should take this opportunity to identify category wise the total CPI and hence bring out plans to meet the gap and hence help the people to meet their basic needs.

Gender wise:

HIVE:

```
hive> select gen,sum(income)/count(gen) from final_census1 group by gen;
```

```
Total MapReduce CPU Time Spent: 4 seconds 930 msec
OK
 Female 1710.1663740321533
 Male   1772.725461619967
Time taken: 28.881 seconds
hive>
```

PIG:

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

b = foreach a generate gen, income;

c = group b by gen;

d = foreach c generate group, SUM (b.income)/COUNT (b.gen);

dump d;
```

```
( Male,1772.725461619967)
( Female,1710.1663740321533)
[cloudera@localhost Desktop]$
```

Category Wise:

HIVE :

```
hive> select a.cat,sum(f.income)/count(a.cat) from final_census1 f join agegroup
 a on f.age=a.age group by a.cat;
```

```
Teenager        1689.544626489068
adult   1813.7500834414673
elderly 1662.5739936698362
infants 1667.2678895748925
middle-aged     1737.4900613918726
senior citizen  1708.37968406245
Time taken: 59.816 seconds
```

PIG :

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

b = load '/user/cloudera/agegroup1.dat' using PigStorage ('\t') as (age: int, cat: chararray);

c = join a by age, b by age;

d = group c by a.age;

e = foreach c generate $5 as income, $11 as cat;

f = group e by cat;

g = foreach f generate group, SUM(e.income)/COUNT(e.cat);

dump g;

```
(adult,1813.7500834414673)
(elderly,1662.5739936698362)
(infants,1667.2678895748925)
(Teenager,1689.544626489068)
(middle-aged,1737.4900613918726)
(senior citizen,1708.37968406245)
```

Total CPI :

HIVE :

```
hive> select sum(income)/count(income) as TotalCPI from final_census1;
```

```
Total MapReduce CPU Time Spent: 4 seconds 730 msec
OK
1740.0260962813627
Time taken: 24.69 seconds
```

Social:

Task 1: Total amount dispensed on pension in x year(s): after how many years how many attending the age: 65

Use Case: Service them back!

- **Input**: Amount [Custom Input from the User]
- **Key**: Year, **Value**: Income
- **Output**: Total Amount to spend on Pension over specific years.
- **Concept used:** Advanced Map Reduce

Description: There is always a need for the elderly Citizens in terms of Economic Support. It is always duly a responsibility for the Government to Identify their Country's Elderly Citizens and provide them economic support in the form of pension. This will help Government to plan how much of income will spent on Pension to Elderly Citizens and hence plan their Budget.

Why this Report: Generate a list of People who will get Pension this year or after some specific years.

Progress: Government should take this opportunity to identify the elderly citizen in needs and provide them support from the early stage itself.

Advanced Map Reduce:

```
[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
```

2. Total amount dispensed on scholarship in current year

Use Case: Calculation on Services

- **Input**: Census_Records.json File.
- **Output**: Set of Amount spend on Scholarship for a specific year.
- **Concept used:** PIG

Description: Most of the People would suffer in either not able to continue to their economic issues. This could be because of their Parents income is very low, some child have only parent (either mother or father), some not in universe etc. It could be also a responsibility of the Government to identify the category of people who would be in need of money for their studies and economic support.

Why this Report: Generate a list of People who will get Scholarship this year or after some specific years.

Progress: Government should take this opportunity to identify the citizen in needs and provide them support from the early stage itself. This factor can be also changed to a loan if the citizen requirement is more and hence plan can various benefits / Policies.

PIG:

t1.txt

```
a = load '/user/cloudera/Census_Records.json' using
JsonLoader('Age:int,Education:chararray,MartialStatus:chararray,Gender:chararray,TaxFilerStatus:ch
ararray,Income:float,Parents:chararray,CountryOfBirth:chararray,Citizenship:chararray,WeeksWorke
d:chararray');

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;
```

Secondary table: scholar1:

Father only present, 2000

Mother only present, 4000

Neither parent present, 7000

Not in universe, 10000

```
[cloudera@localhost Desktop]$ pig /home/cloudera/Desktop/t1
```

```
( Not in universe,4314520000)
( Father only present,11126000)
( Mother only present,153268000)
( Neither parent present,34111000)
```

Task 3: For given age range employable female widowed and divorced count

Use Case: Female – Backbone to the Country's Welfare

- **Input**: Census_Records.json File
- **Key**: Age, **Value**: Gender
- **Output**: Total Count of Employable Female Widowed and Divorced.
- **Data Validation**: Yes.
  Constraint: User Input Can be Only Numbers
- **Concept used:** Advanced Map Reduce, HIVE and PIG.

Description: Government brings lot of Policies and Benefits to the Female Population for their Country, which include Maternity offers, Marriage plans etc. Nevertheless to note, Economic growth is an important factor to be considered for a Female, especially under Widowed and Divorced Category. These two categories of Female Population will be in need for more monetary support

Why this Report: Generate a list of Employable Female Count whose status in widowed and divorced and whose income did not meet the standards.

Progress: Government should take this opportunity to identify the Female who are in need of economic support and hence the education: employable ratio increased.

Basic Map Reduce:

```
hduser@ubuntu64server:~$ hadoop jar c4.jar /Census_Records.json /jj15
Enter Min age
22
Enter Max age
30
```

```
hduser@ubuntu64server:~$ hadoop fs -cat /jj15/p*
Employed female widowed and Divorced in the given age is-->     1901
hduser@ubuntu64server:~$
```

HIVE:

Select count (age) from final_census1 where age between 22 and 30 and gen=' Female' and mar in (' Divorced',' Widowed') and ww>0;

```
Total MapReduce CPU Time Spent: 5 seconds 10 msec
OK
1901
Time taken: 27.161 seconds
```

PIG:

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

b = foreach a generate $0, $2, $3, $9;

c = filter b by ($0>=22 and $0<=30) and ($2==' Female') and ($3>0) and ($1==' Divorced' OR $1==' Widowed');

d = group c by $2;

e = foreach d generate group, COUNT (c.$0);

dump e;

```
( Female,1901)
[cloudera@localhost Desktop]$
```

Planning:

Task 1.Voter(s) count in x year(s)

Use Case: Eligibility for Vote and Adult Support Requirements.

- **Input**: Census_Records.json
- **Key**: Year, **Value**: Gender
- **Output**: Vote Eligibility Count after Specific years.
- **Concept used:** HIVE and PIG.

Description: It is always a healthy factor for a Government to identify the Potential to Vote. Hence these the people who will reach the age of 18 and has rights to vote and eligible for Adult support benefits of the Government. There by we will have an overview of how amount will be spent on Scholarship / Marriage support / Loan scheme etc.

Why this Report: To generate from the list of Population who will vote after some “x” years.

Progress: Government can plan their future budget and hence can plan better things for the welfare of the Country and cast the people vote eligibility factor.

HIVE:

```
hive> set year=2016;
hive> select COUNT(*) as Total_Voters from final_census1 where age+(${hiveconf:y
ear}-YEAR(from_unixtime(unix_timestamp())))>=18;
```

```
Total MapReduce CPU Time Spent: 7 seconds 230 msec
OK
429342
```

PIG:

```
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)
```

Task 2: Senior Citizen(s) count in x year(s)

Use Case: Welcome Future Senior Citizens!

- **Input**: Census_Records.json File
- **Key**: Year , **Value**: Gender
- **Output**: Set of Senior Citizen in specific years.
- **Concept used:** HIVE and PIG.

Description: Government has various benefits for the Senior citizens. This can vary from year to year. The current year, Government has spent 40L of rupees, in next year it can vary to 60L, depends on the age factor. So it is always an important responsibility for the Government to predict the amount to spend for Senior Citizens after a span of years.

Why this Report: To generate from the list of Population who will be our Future Senior Citizens after some “x” years.

Progress: Government can plan their future budget and hence can plan better things for the welfare of the Country and plan budget Factor Vs Senior Citizens

HIVE:

```
hive> set year=2019;
hive> select COUNT(*) as Total_Senior_Citizen from final_census1 where age+(${hi
veconf:year}-YEAR(from_unixtime(unix_timestamp())))>=60;
```

```
Total MapReduce CPU Time Spent: 7 seconds 370 msec
OK
109713
Time taken: 33.374 seconds
```

PIG:

```
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)
```

Task 3: Total number of Male/Female

Use Case: Employability Ratio: Male Vs. Female: Requirements of Soldiers at Army.

- **Input**: Census_Records.json File
- **Key**: Gender, **Value**: Count
- **Output**: Employability Ratio
- **Data Validation**: Yes.
  Constraint: User Input Can be Only Numbers
- **Concept used:** Advanced Map Reduce, HIVE and PIG.

Description: It is always a Government Responsibility to get a random count of Male Vs Female Ratio with respect to Employability Ratio, So that Better Opportunities can be made to them in the form of providing services to the Country via Soldiers. Many Adults who are much interested to get into Army, still not able to find the way to get into it.

Why this Report: To generate from the list of Population to Identify Employability Factors.

Progress: Government can plan a Campaign and check eligibility test from the sample set of data given as report.

HIVE:

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

```
Female 311800
Male   284723
```

PIG:

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

b = foreach a generate gen;

c = group b by gen;

d = foreach c generate group, COUNT (b.gen);

dump d;
```

```
( Male,284723)
( Female,311800)
```

Advanced Map Reduce:

```
hduser@ubuntu64server:~$ hadoop jar /home/hduser/Plan_Task3.jar /Oliver/fin_cens
us /2811_3
Total count for
1.Male
2.Female
rt
Please enter only numbers
```

```
hduser@ubuntu64server:~$ hadoop fs -cat /2811_2/p*
 Male    284723
```

Task 4: Citizens and immigrants count for employed lot

Use Case: Building Requirements for OWN

- **Input**: Census_Records.json File
- **Key**: Citizen, **Value** : Income
- **Output**: Employability List for Citizens and Immigrants.
- **Data Validation**: Yes.
  Constraint: User Input Can be Only Numbers
- **Concept used:** Advanced Map Reduce, HIVE and PIG.

Description: Immigrants entry is always for a specific country with respect to multiple requirements especially for Job. It has to be identified the employability ratio for Own Citizens Vs. Immigrants.

Why this Report: To generate from the list of Population to Increase Employability Factors for OWN Citizens.

Progress: Government to plan various schemes in providing priority the Job Offers based on the standard for our Citizens

HIVE:

```
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
```

PIG:

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

b = foreach a generate $7,$8;

c = filter b by citizen==' Foreign born- U S citizen by naturalization';

d = group c by $0;
```

e = foreach d generate group, COUNT(c.$0);

dump e;

```
( Ecuador,192)
( England,496)
( Germany,1054)
( Hungary,187)
( Ireland,206)
( Jamaica,342)
( Vietnam,371)
( Cambodia,75)
( Columbia,397)
( Honduras,87)
( Portugal,248)
```

Advanced Map Reduce:

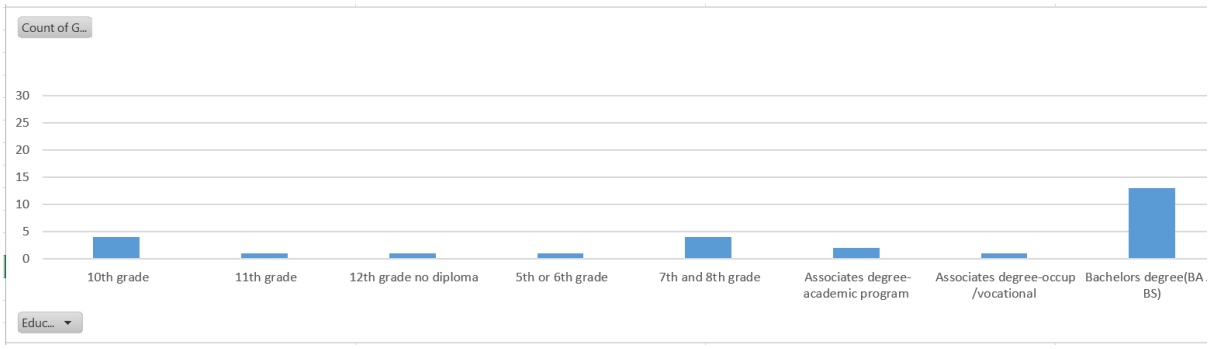
```
Employable count for
1.Citizenship
2.Immigrants
fsdf
Please enter only numbers
hduser@ubuntu64server:~$
```

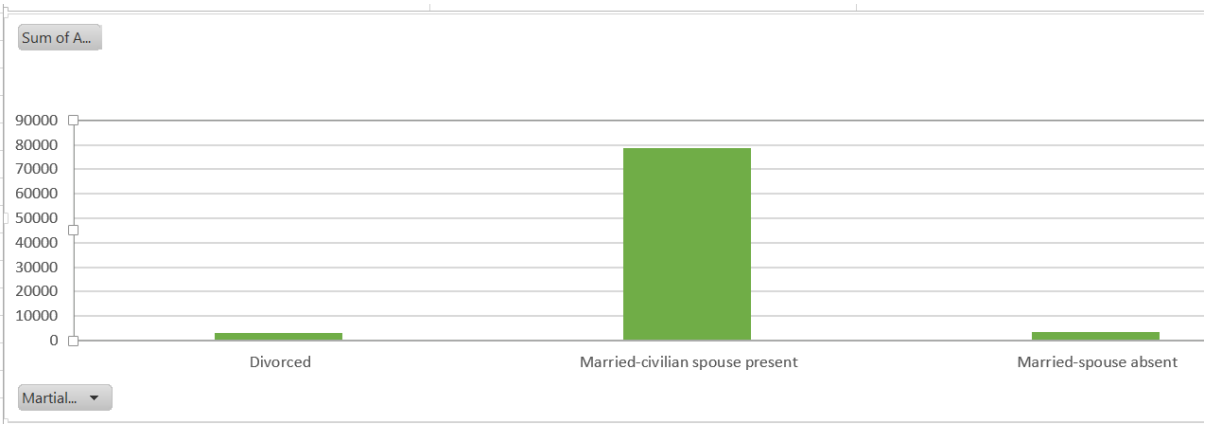
```
hduser@ubuntu64server:~$ hadoop fs -cat /2711_15/part-r-00000
 Native- Born in the United States       271487
```

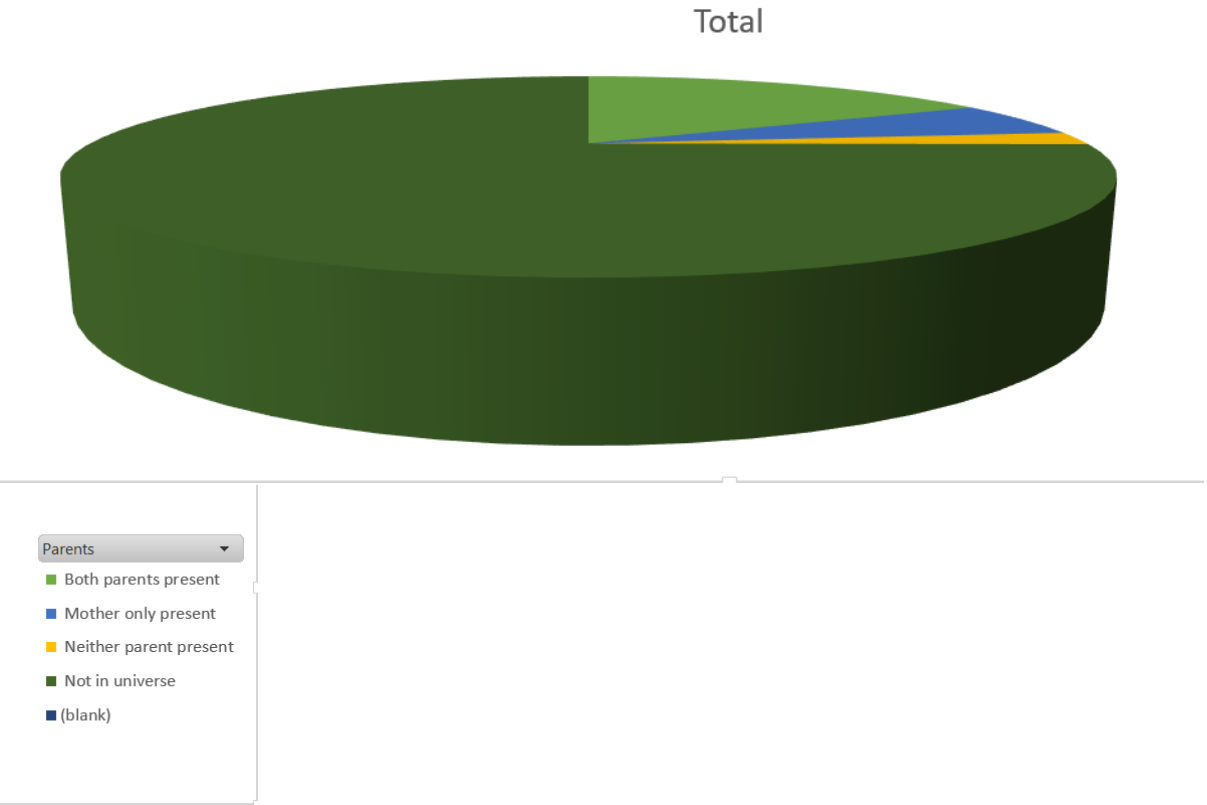
```
hduser@ubuntu64server:~$ hadoop fs -cat /2711_14/part-r-00000
 Foreign born- Not a citizen of U S      22427
 Foreign born- U S citizen by naturalization     10998
 Native- Born abroad of American Parent(s)       3219
 Native- Born in Puerto Rico or U S Outlying     2140
```

Analysis:

Education Based:







Conclusion:

Thus from the given set of Census data provided to us, we have done a complete "Edu-Edo" Report Generation on Various Factors. As an enhancement, the Government should bring new policies and benefits for the welfare to the Country.