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.

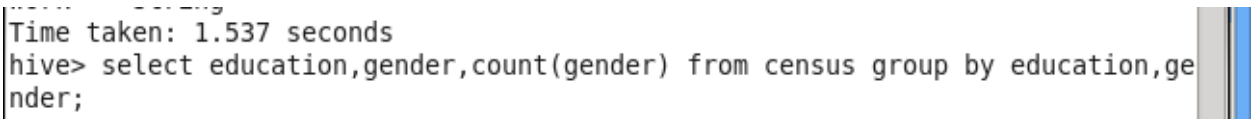
Education and Employment

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

i) Total count of male/female based on education.
ii) Total count of employed/unemployed based on education.
iii) Total count for people in age range of 18-25 based on education.
iv) Degree wise count for employability

Using Hive

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



Output

```
Associates degree-academic program      Male   5266
Associates degree-occup /vocational     Female 9225
Associates degree-occup /vocational     Male   6733
Bachelors degree(BA AB BS)      Female 29557
Bachelors degree(BA AB BS)      Male   29680
Children        Female 69827
Children        Male   71669
```

Execution Step: Total count of unemployed based on education.

```
hive> select education,count(*) as unemployed from census where work=0 group
by education;
```

Output (Unemployed):

```
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
```

Execution Step: Total count of employed based on education.

```
hive> select education,count(*) as employed from census where work>0 group by
 education;
```

Output (employed):

```
OK
10th grade      10527
11th grade      11707
12th grade no diploma   3593
1st 2nd 3rd or 4th grade        2016
5th or 6th grade        4242
7th and 8th grade       6893
```

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

```
Time taken: 204.308 seconds
hive> select education,count(*) as People from cencus where age between 18 an
d 25 group by education;
```

Output:

```
11th grade      5316
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
```

Execution Step: Degree wise count for employability

```
hive> select education,count(*) from census where work=0 group by education;
Total MapReduce jobs = 1
Launching Job 1 out of 1
```

Output:

```
10th grade      12044
11th grade      8798
12th grade no diploma  2681
1st 2nd 3rd or 4th grade        3339
```

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

i) Tax analysis total and gender wise
ii) Per Capita Income (PCI) analysis consolidated, gender wise and category wise
iii) Non-US citizen(s) tax filer status

Execution Step: Tax analysis total and gender wise

```
hive> select sum(income*tax_percent) as total_tax, sum(case f.gender when ' Male
' then income end) as male_tax, sum(case f.gender when ' Female' then income end
) as female_tax from final_census f join tax t on f.gender=t.gender where f.inco
me between t.min_income and t.max_income;
```

Output:

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

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

```
hive> select a.category,sum(c.income)/count(a.category) from census c join ag
egroup a on c.age=a.age group by a.category;
```

Output:

```
Teenager        1689.5446269570016
adult   1813.7500828047719
elderly 1662.5739941670317
infants 1667.2678898605448
middle-aged     1737.4900611355397
senior citizen  1708.379683926455
```

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

```
Time taken: 135.993 seconds
hive> select gender,sum(income)/count(gender) from census group by gender;
Total MapReduce jobs = 1
```

Output:

```
OK
 Female 1710.1663736369826
 Male   1772.7254616592884
```

Execution Step: Per Capita Income (PCI) analysis consolidated

```
hive> select sum(income)/count(income) as totalPCI from census;
Total MapReduce jobs = 1
```

Output:

```
Total MapReduce CPU Time Spent: 11 seconds 770 msec
OK
1740.0260960934236
Time taken: 213.039 seconds
```

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

```
Time taken: 98.609 seconds
hive> select tax,citizen from census where citizen not in(' Native- Born in t
he United States');
```

Output:

```
Joint both under 65     Foreign born- U S citizen by naturalization
Joint both under 65     Foreign born- Not a citizen of U S
Joint both under 65     Foreign born- U S citizen by naturalization
Joint both under 65     Foreign born- Not a citizen of U S
Single  Native- Born in Puerto Rico or U S Outlying
Joint both under 65     Foreign born- Not a citizen of U S
```

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

i) Total amount dispensed on scholarship in current year
ii) For given age range employable female widowed and divorced count
iii) Total amount dispensed on pension in x year(s)
iv) Citizens and immigrants count for employed lot

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

Execution Step:

```
[cloudera@localhost Desktop]$ pig /home/cloudera/Desktop/Task.txt
```

Output:

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

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

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

Output:

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

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

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

Output:

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

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:

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

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

i) Citizens and immigrants count for employed lot

ii) Country of birth wise count for US citizenship by naturalization

iii) Total number of Male/Female

Execution Step: Citizens and immigrants count for employed lot

```
select citizen, count(*) from (select case citizen when ' Native- Born in the Un
ited States' then 'Native Born United States' else 'Immigrants' End citizen from
 census) a group by citizen where work>0;
```

Output:

```
OK
Immigrants      67265
Native Born United States       529258
```

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

```
hive> select birth,count(citizen) from census where citizen=' Foreign born- U
 S citizen by naturalization' group by birth;
Total MapReduce jobs = 1
```

Output:

```
 Ireland        206
 Italy  793
 Jamaica        342
 Japan  152
 Laos   82
```

Execution Step: Total number of Male/Female

```
Time taken: 191.871 seconds
hive> select gender, count(*) from census group by gender;
Total MapReduce jobs = 1
```

Output:

```
S Write: 28 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 730 msec
OK
 Female 311800
 Male   284723
```

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.

i) Voter(s) count in x year(s)
ii) Senior Citizen(s) count in x year(s)

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

```
hive> set year=2016;
hive> select count(*) as voters from census where age+(${hiveconf:year}-Year(
from_unixtime(unix_timestamp())))>=18;
Total MapReduce jobs = 1
```

Output:

```
Total MapReduce CPU Time Spent: 15 seconds 540 msec
OK
429342
```

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

```
hive> select count(*) as senior_citizen from census where age+(${hiveconf:yea
r}-YEAR(from_unixtime(unix_timestamp())))>=60;
Total MapReduce jobs = 1
```

Output:

```
Total MapReduce CPU Time Spent: 14 seconds 30 msec
OK
95362
```

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".