CENSUS DATA ANALYSIS

Project Author – Sakshi Chawla

PROJECT OBJECTIVE – In this project, we are working on a Census large Dataset in which we will be extracting useful information related to Education Sector, number of pension payers and their paid pension amount, number of tax payers and the tax amount paid by them and planning for the Socio-Economic structure. Census data which is provided to us contains the US population data so analysis or count is also done on the basis of US Citizenship.

SCOPE – Can be used by Education, Finance & Socio-Economic Department to get the following count/information.

- Count of Male/Female according to the education/degree.
- Count of people in Age-Group 18-25 years to get a % of youth.
- Tax Amount & Pension Amount
- Count of Widowed & Divorced Female can be given employment.
- Tax Filer Status of Non-US Citizens
- Count of over All Base Customers present.

DATASETS REQUIRED – We are given a Census Data in JSON format. So there is one Primary Dataset – Census_Records.json. Apart from this, we have created 3 secondary tables to get the desired output. So we are using below 4 Datasets.

- Census_Records.json which contains following fields – Age, Education, Marital_status, Gender, TaxFilerStatus, Parents, CountryOfBirth, Citizenship, Weeks_Worked
- Tax – Secondary table created in MySQL which has following fields – MinAmt, MaxAmt, Gender, Tax-percentage
- Pension – Secondary table created in MySQL which has following fields – MinAmt, MaxAmt, PensionAmt
- Scholarship – Secondary table created in MySQL which has following fields – Parent, ScholarshipAmt

TECHNOLOGY USED –

- Apache Hadoop
- Advance Map-Reduce Programming in Java.
- PIG Programming
- Hive

SOFTWARE USED –

- Eclipse
- Oracle Virtual Box
- Cloudera

PROJECT DESCRIPTION - We are given a census dataset and some operations to be performed on that dataset to get the desired result. Below are the scenarios in which this data analysis can be used.

Use-Case 1:

Number of Schools/Colleges required to be opened in future, based on count. – From the count of Male/Female based on education, we can analyze and get a count of how many schools/colleges we will be requiring in near future to accommodate all the students.

Task 1 - Total count of male/female based on education -

Output - Hive

```
Total MapReduce CPU Time Spent: 22 seconds 610 msec
OK
10th grade      Female 12187
10th grade      Male   10384
11th grade      Female 10815
11th grade      Male   9690
12th grade no diploma   Female 2970
12th grade no diploma   Male   3304
1st 2nd 3rd or 4th grade        Female 2764
1st 2nd 3rd or 4th grade        Male   2591
5th or 6th grade        Female 4992
5th or 6th grade        Male   4761
7th and 8th grade       Female 12609
7th and 8th grade       Male   11518
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
Bachelors degree(BA AB BS)      Female 29557
Bachelors degree(BA AB BS)      Male   29680
Children        Female 69827
Children        Male   71669
Doctorate degree(PhD EdD)       Female 1099
Doctorate degree(PhD EdD)       Male   2714
High school graduate    Female 80977
High school graduate    Male   63857
Less than 1st grade     Female 1279
Less than 1st grade     Male   1133
Masters degree(MA MS MEng MEd MSW MBA)  Female 9493
Masters degree(MA MS MEng MEd MSW MBA)  Male   10150
Prof school degree (MD DDS DVM LLB JD)  Female 1530
Prof school degree (MD DDS DVM LLB JD)  Male   3828
Some college but no degree      Female 45012
Some college but no degree      Male   38690
Time taken: 156.749 seconds
hive>
```

Output: Pig

```
(( 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)
(( 12th grade no diploma, Male),3304)
(( 12th grade no diploma, Female),2970)
(( 1st 2nd 3rd or 4th grade, Male),2591)
(( 1st 2nd 3rd or 4th grade, Female),2764)
(( Doctorate degree(PhD EdD), Male),2714)
(( Doctorate degree(PhD EdD), Female),1099)
(( Bachelors degree(BA AB BS), Male),29680)
(( Bachelors degree(BA AB BS), Female),29557)
(( Some college but no degree, Male),38690)
(( Some college but no degree, Female),45012)
(( Associates degree-academic program, Male),5266)
(( Associates degree-academic program, Female),7684)
(( Associates degree-occup /vocational, Male),6733)
(( Associates degree-occup /vocational, Female),9225)
(( Masters degree(MA MS MEng MEd MSW MBA), Male),10150)
(( Masters degree(MA MS MEng MEd MSW MBA), Female),9493)
(( Prof school degree (MD DDS DVM LLB JD), Male),3828)
(( Prof school degree (MD DDS DVM LLB JD), Female)
[cloudera@localhost ~]$
```

*t1 (~/Desktop/mydata) - gedit

Output - Map Reduce

```
[cloudera@localhost Desktop]$ hadoop fs -cat  /user/cloudera/edcu/part-r-00000
10th grade      Male 10384
10th grade      Female 12187
11th grade      Male 9690
11th grade      Female 10815
12th grade no diploma   Male 3304
12th grade no diploma   Female 2970
1st 2nd 3rd or 4th grade        Male 2591
1st 2nd 3rd or 4th grade        Female 2764
5th or 6th grade        Male 4761
5th or 6th grade        Female 4992
7th and 8th grade       Male 11518
7th and 8th grade       Female 12609
```

Use Case 2: Selling Course Certification – From the count of employed/unemployed people, we can decide to sell the certification course to the employed/unemployed.

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

Output: Hive

```
K
10th grade     EMployed-->     12044.0 UnEMployed-->   10527.0
11th grade     EMployed-->     8798.0  UnEMployed-->   11707.0
12th grade no diploma  EMployed-->     2681.0  UnEMployed-->   3593.0
1st 2nd 3rd or 4th grade       EMployed-->     3339.0  UnEMployed-->   2016.0
5th or 6th grade       EMployed-->     5511.0  UnEMployed-->   4242.0
7th and 8th grade      EMployed-->     17234.0 UnEMployed-->   6893.0
9th grade      EMployed-->     11430.0 UnEMployed-->   7105.0
Associates degree-academic program     EMployed-->     2094.0  UnEMployed-->  1
856.0
Associates degree-occup /vocational    EMployed-->     2820.0  UnEMployed-->  1
138.0
Bachelors degree(BA AB BS)     EMployed-->     9615.0  UnEMployed-->   49622.0
Children       EMployed-->     141496.0        UnEMployed-->   NULL
Doctorate degree(PhD EdD)      EMployed-->     530.0   UnEMployed-->   3283.0
High school graduate   EMployed-->     44342.0 UnEMployed-->   100492.0
Less than 1st grade    EMployed-->     1678.0  UnEMployed-->   734.0
Masters degree(MA MS MEng MEd MSW MBA) EMployed-->     2937.0  UnEMployed-->  1
706.0
Prof school degree (MD DDS DVM LLB JD) EMployed-->     666.0   UnEMployed-->  4
92.0
Some college but no degree     EMployed-->     19037.0 UnEMployed-->   64665.0
ime taken: 135.667 seconds
ive>
```

Output:Pig-Employed

```
( 9th grade,7105)
( 10th grade,10527)
( 11th grade,11707)
( 5th or 6th grade,4242)
( 7th and 8th grade,6893)
( Less than 1st grade,734)
( High school graduate,100492)
( 12th grade no diploma,3593)
( 1st 2nd 3rd or 4th grade,2016)
( Doctorate degree(PhD EdD),3283)
( Bachelors degree(BA AB BS),49622)
( Some college but no degree,64665)
( Associates degree-academic program,10856)
( Associates degree-occup /vocational,13138)
( Masters degree(MA MS MEng MEd MSW MBA),16706)
( Prof school degree (MD DDS DVM LLB JD),4692)
[cloudera@localhost ~]$
```

Output:Pig-Unemployed:

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

Use Case 3: Make better the Employability Percentage

Can be used by websites like naukri.com to approach this kind of people for jobs.

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

Output: Hive

```
OK
Education-->     10th grade     Total Count-->  2411
Education-->     11th grade     Total Count-->  5310
Education-->     12th grade no diploma  Total Count-->  1824
Education-->     1st 2nd 3rd or 4th grade       Total Count-->  275
Education-->     5th or 6th grade       Total Count-->  871
Education-->     7th and 8th grade      Total Count-->  989
Education-->     9th grade      Total Count-->  1486
Education-->     Associates degree-academic program     Total Count-->  1414
Education-->     Associates degree-occup /vocational    Total Count-->  1558
Education-->     Bachelors degree(BA AB BS)     Total Count-->  5714
Education-->     Doctorate degree(PhD EdD)      Total Count-->  15
Education-->     High school graduate   Total Count-->  18966
Education-->     Less than 1st grade    Total Count-->  187
Education-->     Masters degree(MA MS MEng MEd MSW MBA) Total Count-->  358
Education-->     Prof school degree (MD DDS DVM LLB JD) Total Count-->  27
Education-->     Some college but no degree     Total Count-->  20311
Time taken: 29.134 seconds
hive>
```

Output – Map Reduce

```
Enter the minimum age
18
Enter the maximum age
16
Maximum age range limit can't be less than minimum age range limit set by you
Enter valid Maximum age limit
Enter the maximum age
14
Enter the maximum age
25

[cloudera@localhost Desktop]$ hadoop fs -cat /user/cloudera/etask3/part-r-00000
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
```

<u>Use Case 4</u> – Start up a new project for public service

1. **Consolidated Tax analysis can be used by Government to start up a new project for public service like starting up a Metro Rail service in a city or making a bus service more convenient.**

2. **Gender Wise Tax Analysis can be used for Female Welfare, Female tax amount should be dedicated for female welfare only.**

Task 4 - Tax analysis total and gender wise

<u>Output - Hive</u>

```
Female 1710.1663736369826
Male   1772.7254616592884
Time taken: 28.998 seconds
hive>
```

<u>Use Case</u> 5 - Can be used to check PCI state wise/region wise and the states with the lowest PCI should take some measures to improve that.

Task 5 - Per Capita Income (PCI) analysis consolidated, gender wise and category wise

Output: Hive: Category wise

```
age group-->    Teenager        sum of income-->        1689.5446269570016
age group-->    adult   sum of income-->        1813.7500828047719
age group-->    elderly sum of income-->        1662.5739941670317
age group-->    infants sum of income-->        1667.2678898605448
age group-->    middle-aged     sum of income-->        1737.4900611355397
age group-->    senior citizen  sum of income-->        1708.379683926455
Time taken: 66.15 seconds
hive>
```

Output - Hive - Total PCI

```
TotalPCI-->      1740.0260960934236
Time taken: 29.013 seconds
hive>
```

<u>Use Case</u> 6 – Increasing Pension amount – By looking at the total amount of pension given in last x years we can check, if the pension amount can be increased if the budget allows.

<u>Task 6 -</u> Total amount dispensed on pension in x year(s)

Output - Hive

```
Total MapReduce CPU Time Spent: 20 seconds 10 msec
OK
16455420
Time taken: 87.405 seconds
```

Use case 7 -

Task 7 - Total amount dispensed on scholarship in current year

Input: Secondary table: Pension

```
Father only present,500
Mother only present,700
Neither parent present,700
Not in universe,1000
```

Output: Pig

```
a = 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');
b = load '/user/cloudera/scholar2' 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;
```

```
cloudera@localhost:~
File  Edit  View  Search  Terminal  Help
2016-11-27 02:33:02,425 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Success!
2016-11-27 02:33:02,429 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Ke
y [pig.schematuple] was not set... will not generate code.
2016-11-27 02:33:02,436 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input paths to process : 1
2016-11-27 02:33:02,436 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
( Not in universe,431452000)
( Father only present,2781500)
( Mother only present,26821900)
( Neither parent present,3411100)
[cloudera@localhost ~]$ d
```

Use Case 8 – Female Welfare

This can be used to approach these people to give jobs in Govt. schools

Task8: For given age range employable female widowed and divorced count

Data Validation: Yes

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

Output: Map reduce

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

Use Case 9: Aadhar Card

Task9 - Voter(s) count in x year(s)

Output: Hive

```
OK
Total_Voters Count-->   437549
Time taken: 31.156 seconds
hive>
```

Use Case 10: Discount given to senior citizen on Rail booking- Government can check that how many total senior citizens are there and for how many of them discount was availed and the total discount amount availed.

Task 10 - Senior Citizen(s) count in x year(s) -

Output: Hive

```
OK
 Total_Senior_Citizen in given year-->  100079
Time taken: 30.949 seconds
hive>
```

Use Case 11: Check Employability status/ Percentage -

Task 11 - Total number of Male/Female

```
OK
gender-->        Female Total count-->  311800
gender-->        Male   Total count-->  284723
Time taken: 29.985 seconds
hive>
```

Use Case 12:

Task 12 - Citizens and immigrants count for employed lot

Output: Hive

```
OK
CitizenShip-->  Immigrants      Total Count-->  67265
CitizenShip-->  Native Born United States       Total Count-->  529258
Time taken: 26.96 seconds
hive> _
```

Use Case 13: Literacy Rate – We can calculate the literacy rate from the below count.

Task 13 - Degree wise count for Employability

Output:MapReduce

```
hduser@ubuntu64server:~$ hadoop fs -cat /kk1/p*
 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
 9th grade       7105
 Associates degree-academic program      10856
 Associates degree-occup /vocational     13138
 Bachelors degree(BA AB BS)      49622
 Children        0
 Doctorate degree(PhD EdD)       3283
 High school graduate    100492
 Less than 1st grade     734
 Masters degree(MA MS MEng MEd MSW MBA)  16706
 Prof school degree (MD DDS DVM LLB JD)  4692
 Some college but no degree      64665
hduser@ubuntu64server:~$
```

Output: Hive

```
OK
Education-->     10th grade     Total Count-->  12044
Education-->     11th grade     Total Count-->  8798
Education-->     12th grade no diploma  Total Count-->  2681
Education-->     1st 2nd 3rd or 4th grade       Total Count-->  3339
Education-->     5th or 6th grade       Total Count-->  5511
Education-->     7th and 8th grade      Total Count-->  17234
Education-->     9th grade      Total Count-->  11430
Education-->     Associates degree-academic program     Total Count-->  2094
Education-->     Associates degree-occup /vocational    Total Count-->  2820
Education-->     Bachelors degree(BA AB BS)     Total Count-->  9615
Education-->     Children       Total Count-->  141496
Education-->     Doctorate degree(PhD EdD)      Total Count-->  530
Education-->     High school graduate   Total Count-->  44342
Education-->     Less than 1st grade    Total Count-->  1678
Education-->     Masters degree(MA MS MEng MEd MSW MBA) Total Count-->  2937
Education-->     Prof school degree (MD DDS DVM LLB JD) Total Count-->  666
Education-->     Some college but no degree     Total Count-->  19037
Time taken: 28.947 seconds
hive>
```

t1 (~/Desktop/mydata) - gedit

Output: Advanced Map Reduce

```
hduser@ubuntu64server:~$ hadoop fs -cat /kk1/p*
 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
 9th grade       7105
 Associates degree-academic program      10856
 Associates degree-occup /vocational     13138
 Bachelors degree(BA AB BS)      49622
 Children        0
 Doctorate degree(PhD EdD)       3283
 High school graduate    100492
 Less than 1st grade     734
 Masters degree(MA MS MEng MEd MSW MBA)  16706
 Prof school degree (MD DDS DVM LLB JD)  4692
 Some college but no degree      64665
hduser@ubuntu64server:~$
```

Use Case 14: My product is PlayStation, which will be more liked by the males in the age group 15 – 30 years So we have to analyze that if we launch this product, how successful will it be. So for this we have to find the number of males in the age group of 15 – 30 years. So I have done that analysis.

Task 14 - Customer base analysis

```
a = load '/user/cloudera/Census.json' using
JsonLoader('age:int,edu:chararray,mar:chararray,gen:chararray,tax:chararray,income:long,parent:
chararray,country:chararray,citizen:chararray,ww:int');
b = foreach 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 = foreach j generate group,COUNT(d.age);
dump k;
```

```
(15,2549)
(16,2295)
(17,2381)
(18,2085)
(19,2230)
(20,2099)
(21,2071)
(22,2198)
(23,2435)
(24,2560)
(25,2565)
(26,2360)
(27,2452)
(28,2403)
(29,2515)
(30,2634)
```

Task 15 - Non-US citizen(s) tax filer status

<u>Output - Hive</u>

```
37      Bachelors degree(BA AB BS)      Female TaxFilerStatus-->        Joint both under 65    CitizenShip-->   Foreign born- Not a citizen of U S     624.68  0
3       Children        Female TaxFilerStatus-->        Nonfiler       CitizenShip-->   Foreign born- Not a citizen of U S     1205.3  0
25      Some college but no degree      Male    TaxFilerStatus-->        Single CitizenShip-->   Foreign born- Not a citizen of U S     3442.4  52
28      Bachelors degree(BA AB BS)      Female TaxFilerStatus-->        Single CitizenShip-->   Foreign born- U S citizen by naturalization    1741.48 52
53      1st 2nd 3rd or 4th grade        Female TaxFilerStatus-->        Nonfiler       CitizenShip-->   Foreign born- Not a citizen of U S     1370.33 28
38      7th and 8th grade       Female TaxFilerStatus-->        Head of household      CitizenShip-->   Foreign born- Not a citizen of U S     1219.11 0
14      Children        Female TaxFilerStatus-->        Nonfiler       CitizenShip-->   Foreign born- Not a citizen of U S     677.03  0
2       Children        Female TaxFilerStatus-->        Nonfiler       CitizenShip-->   Foreign born- Not a citizen of U S     862.44  0
3       Children        Female TaxFilerStatus-->        Nonfiler       CitizenShip-->   Foreign born- Not a citizen of U S     1804.74 0
46      5th or 6th grade        Female TaxFilerStatus-->        Joint both under 65    CitizenShip-->   Foreign born- U S citizen by naturalization    2688.61 52
43      Some college but no degree      Female TaxFilerStatus-->        Joint both under 65    CitizenShip-->   Foreign born- Not a citizen of U S     2239.83 52
41      5th or 6th grade        Female TaxFilerStatus-->        Joint both under 65    CitizenShip-->   Foreign born- Not a citizen of U S     938.99  52
26      11th grade      Male    TaxFilerStatus-->       Joint both under 65    CitizenShip-->   Foreign born- Not a citizen of U S     1199.34 52
22      Some college but no degree      Male    TaxFilerStatus-->       Joint both under 65    CitizenShip-->   Native- Born abroad of American Parent(s)      1900.14 52
12      Children        Female TaxFilerStatus-->        Nonfiler       CitizenShip-->   Foreign born- U S citizen by naturalization    1532.61 0
52      12th grade no diploma   Male    TaxFilerStatus-->       Nonfiler       CitizenShip-->   Native- Born in Puerto Rico or U S Outlying    1140.64 0
25      Some college but no degree      Male    TaxFilerStatus-->        Single CitizenShip-->   Foreign born- Not a citizen of U S     1740.9  52
46      Some college but no degree      Male    TaxFilerStatus-->        Joint both under 65    CitizenShip-->   Foreign born- Not a citizen of U S     1031.19 52
48      High school graduate    Female TaxFilerStatus-->        Joint both under 65    CitizenShip-->   Foreign born- U S citizen by naturalization    740     52
35      High school graduate    Female TaxFilerStatus-->        Nonfiler       CitizenShip-->   Foreign born- Not a citizen of U S     1584.92 0
26      9th grade       Male    TaxFilerStatus-->       Joint both under 65    CitizenShip-->   Foreign born- Not a citizen of U S     1171.52 52
28      12th grade no diploma   Male    TaxFilerStatus-->       Joint both under 65    CitizenShip-->   Foreign born- Not a citizen of U S     1140.07 52
43      Some college but no degree      Male    TaxFilerStatus-->        Single CitizenShip-->   Native- Born abroad of American Parent(s)      1019.25 36
24      High school graduate    Female TaxFilerStatus-->        Joint both under 65    CitizenShip-->   Foreign born- U S citizen by naturalization    852.49  52
31      High school graduate    Male    TaxFilerStatus-->        Joint both under 65    CitizenShip-->   Foreign born- U S citizen by naturalization    648.87  26
39      12th grade no diploma   Female TaxFilerStatus-->        Joint both under 65    CitizenShip-->   Foreign born- Not a citizen of U S     1432.86 0
53      High school graduate    Female TaxFilerStatus-->        Joint both under 65    CitizenShip-->   Foreign born- U S citizen by naturalization    2590.42 26
19      5th or 6th grade        Female TaxFilerStatus-->        Joint both under 65    CitizenShip-->   Foreign born- Not a citizen of U S     1329.61 0
49      High school graduate    Female TaxFilerStatus-->        Single CitizenShip-->   Native- Born in Puerto Rico or U S Outlying    1198.34 52
23      High school graduate    Female TaxFilerStatus-->        Joint both under 65    CitizenShip-->   Foreign born- Not a citizen of U S     2632.78 52
38      Some college but no degree      Female TaxFilerStatus-->        Joint both under 65    CitizenShip-->   Foreign born- U S citizen by naturalization    1386.91 52
32      Some college but no degree      Male    TaxFilerStatus-->        Single CitizenShip-->   Foreign born- Not a citizen of U S     1230.37 0
46      1st 2nd 3rd or 4th grade        Female TaxFilerStatus-->        Nonfiler       CitizenShip-->   Foreign born- Not a citizen of U S     762.63  0
37      7th and 8th grade       Male    TaxFilerStatus-->       Nonfiler       CitizenShip-->   Foreign born- Not a citizen of U S     1195.92 0
24      High school graduate    Female TaxFilerStatus-->        Nonfiler       CitizenShip-->   Foreign born- Not a citizen of U S     1761.07 0
24      7th and 8th grade       Male    TaxFilerStatus-->       Single CitizenShip-->   Foreign born- Not a citizen of U S     2938.54 52
51      Masters degree(MA MS MEng MEd MSW MBA)  Male    TaxFilerStatus-->        Single CitizenShip-->   Foreign born- U S citizen by naturalization    672.59  52
5       Children        Male    TaxFilerStatus-->       Nonfiler       CitizenShip-->   Foreign born- Not a citizen of U S     649.07  0
26      5th or 6th grade        Female TaxFilerStatus-->        Nonfiler       CitizenShip-->   Foreign born- Not a citizen of U S     1287.85 0
Time taken: 38.064 seconds
Hive>
```

Task 16: Country of birth wise count for US citizenship by naturalization

Output: Hive

```
OK
 ?      3113
 Cambodia       75
 Canada 770
 China  430
 Columbia       397
 Cuba   1251
 Dominican-Republic     379
 Ecuador        192
 El-Salvador    227
 England        496
 France 87
 Germany        1054
 Greece 300
 Guatemala      98
 Haiti  144
 Holand-Netherlands     28
 Honduras       87
 Hong Kong      99
 Hungary        187
 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: 27.363 seconds
hive>
```

Extra Use Case – We are focusing on the Age Group of 0-15 years' kids whose parent status is 'Not in Universe'. By getting the number of kids in each age group, Organizations donating for children can use this data and can have an estimate of total amount that needs to be spend on this.

For example – Below we are having a total count of 1200 kids, for each if we plan to spend Rs. 5000/- , the organization can should have Rs. 60,00,000 budget to achieve this.

cloudera@localhost:~

File Edit View Search Terminal Help

```
2016-11-27 19:41:35,676 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.03
sec
2016-11-27 19:41:36,687 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.03
sec
2016-11-27 19:41:37,700 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.03
sec
2016-11-27 19:41:38,711 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.03
sec
2016-11-27 19:41:39,738 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.03
sec
MapReduce Total cumulative CPU time: 5 seconds 30 msec
Ended Job = job_201611271905_0002
MapReduce Jobs Launched:
Job 0: Map: 1  Reduce: 1   Cumulative CPU: 5.03 sec   HDFS Read: 86758562 HDFS W
rite: 135 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 30 msec
OK
 Both parents present   116318
 Father only present    5563
 Mother only present    38317
 Neither parent present 4873
 Not in universe        431452
Time taken: 39.188 seconds
hive>
```

parent (~/Desktop) - gedit

File Edit View Search Tools Documents Help

Open Save Undo

parent COB

```
a = load '/user/cloudera/Census_Records.json' using JsonLoader
('age:int,edu:chararray,mar:chararray,gen:chararray, tax:chararray,
income:long,parent:chararray,country:chararray,citizen:chararray,ww:int');
b = foreach a generate age, parent;
c = filter b by ((parent==' Not in universe') and (age>0 and age<16));
d = group c by age;
e = foreach d generate group, COUNT(c.age);
dump e;
```

Plain Text Tab Width: 8 Ln 6, Col 8 INS

```
cloudera@localhost:~
File  Edit  View  Search  Terminal  Help
2016-11-27 19:30:59,522 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Success!
2016-11-27 19:30:59,597 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Ke
y [pig.schematuple] was not set... will not generate code.
2016-11-27 19:30:59,818 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input paths to process : 1
2016-11-27 19:30:59,819 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(1,76)
(2,97)
(3,84)
(4,98)
(5,86)
(6,75)
(7,106)
(8,72)
(9,57)
(10,67)
(11,73)
(12,77)
(13,51)
(14,77)
(15,183)
[cloudera@localhost ~]$
```

Conclusion – With the census data we can analyze a lot of social and economic issues and can try to improve the Socio-Economic Structure of the country.