US Census – Immigration Related Statistical Analysis

BIG DATA - HADOOP

GITABASHYAN RAMAMOORTHY | Professional Diploma in Digital Transformation – Big Data with Hadoop | 28.02.2017

Tools : Apache Hadoop Framework – HDFS, MapReduce, Hive, Pig, Sqoop, MySql and MS Excel for Data Visualization

S170030600311
NIIT Limited

NIIT

US CENSUS – IMMIGRATION RELATED STATISTICAL ANALYSIS

A PROJECT REPORT

Submitted by

GITABASHYAN RAMAMOORTHY

in the partial fulfillment for the award of the course of

PROFESSIONAL DIPLOMA

in

BIG DATA WITH HADOOP

NIIT, CHENNAI

MARCH 2017

NIIT

ACKNOWLEDGEMENT

I find immense pleasure to convey my sincere and grateful thanks to **NIIT** and the management for providing necessary facilities in carrying out this project.

I greatly indebted to my Tech Mentor **Ms. Amirtha,** the batch instructor **Mr. Annu** and the SLT faculty **Mr. Sandeep** for constant support throughout the course and also for useful suggestions, constant encouragement and kind advice in bringing out this project as a success.

I express my regards and sincere thanks to the Academic Leader **Ms. Kotteswari** for providing all necessary resource to complete the project.

I extend my thanks to all staff members of NIIT for their kind co-operation for the completion of the project successfully. I am grateful to thank my family and all my friends for their valuable feedback, encouragement and suggestions.

Introduction

Big Data

Big data is being generated by everything around us at all times. Every digital process and social media exchange produces it. Systems, sensors and mobile devices transmit it. Big data is arriving from multiple sources at an alarming velocity, volume and variety. To extract meaningful value from big data, you need optimal parallel processing power, analytics capabilities and skills.

While the term "big data" is relatively new, the act of gathering and storing large amounts of information for eventual analysis is ages old. The concept gained momentum in the early 2000s when industry analyst Doug Laney articulated the now-mainstream definition of big data as the three Vs:

- **Volume.** Organizations collect data from a variety of sources, including business transactions, social media and information from sensor or machine-to-machine data. In the past, storing it would've been a problem – but new technologies (such as Hadoop) have eased the burden.

- **Velocity.** Data streams in at an unprecedented speed and must be dealt with in a timely manner. RFID tags, sensors and smart metering are driving the need to deal with torrents of data in near-real time

- **Variety.** Data comes in all types of formats – from structured, numeric data in traditional databases to unstructured text documents, email, video, audio, stock ticker data and financial transactions.

3V's of Big Data

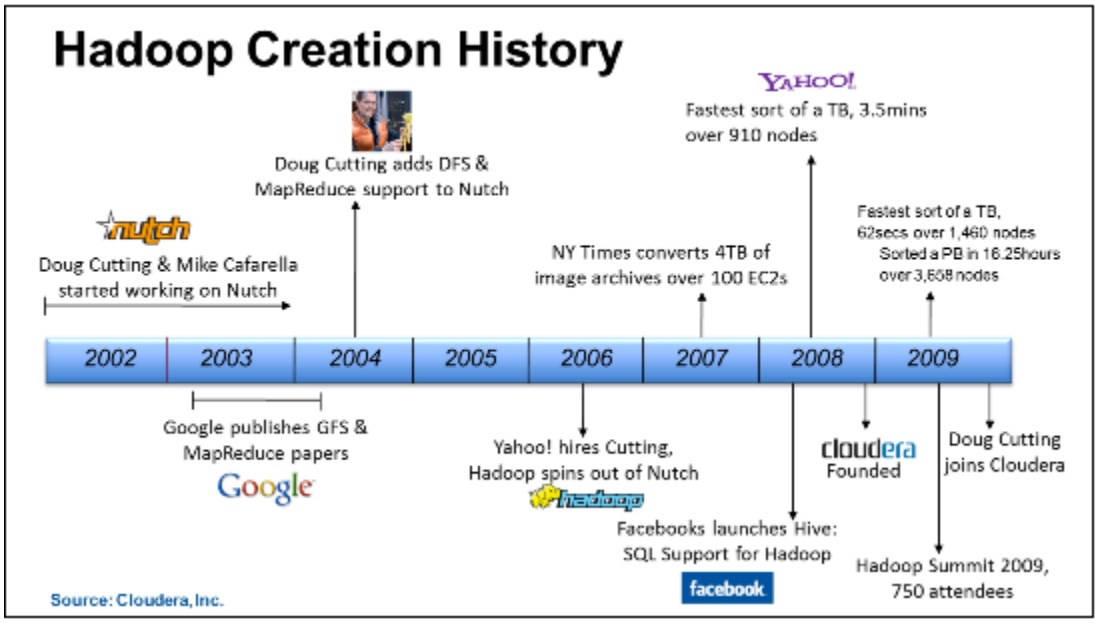
Apache Hadoop

The Apache Hadoop project develops open-source software for reliable, scalable, distributed computing.

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.

History

The genesis of Hadoop came from the Google File System paper that was published in October 2003. This paper spawned another research paper from Google – MapReduce: Simplified Data Processing on Large Clusters. Development started in the Apache Nutch project, but was moved to the new Hadoop subproject in January 2006. The first committer added to the Hadoop project was Owen O'Malley in March 2006. Hadoop 0.1.0 was released in April 2006 and continues to be evolved by the many contributors to the Apache Hadoop project. Hadoop was named after one of the founder's toy elephant.

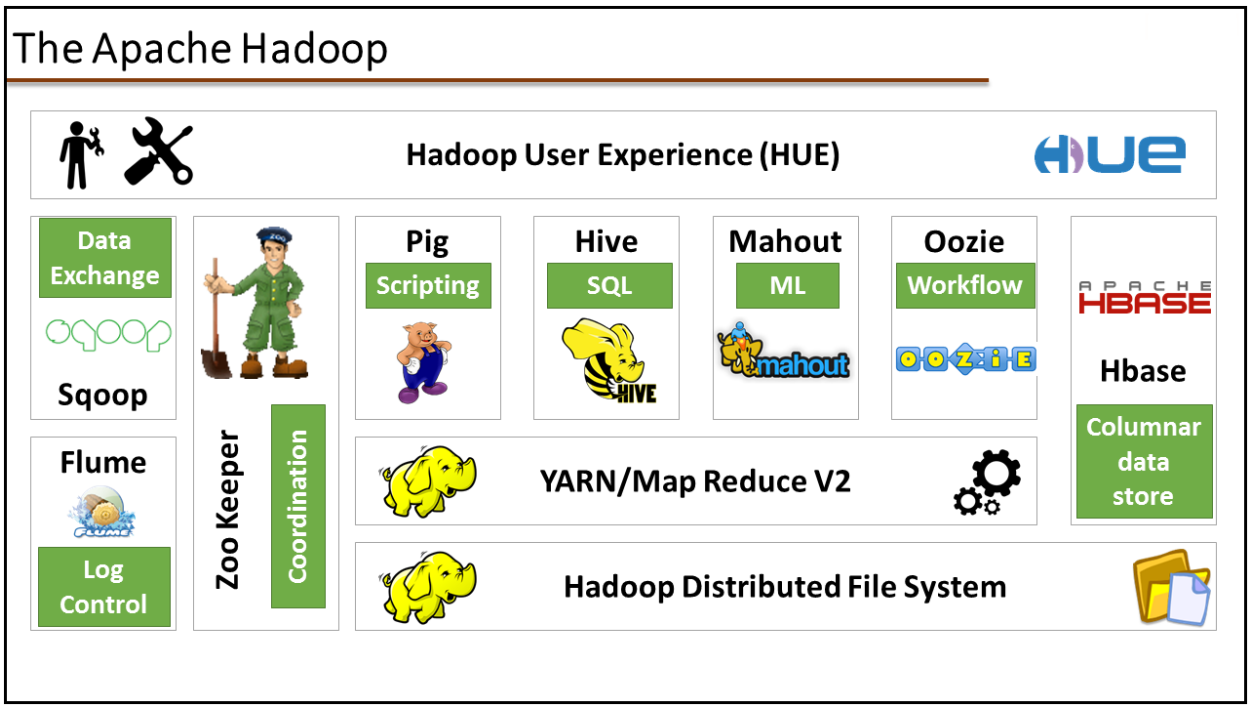


Hadoop Timeline

Benefits

Some of the reasons organizations use Hadoop is its' ability to store, manage and analyze vast amounts of structured and unstructured data quickly, reliably, flexibly and at low-cost.

- **Scalability and Performance** – distributed processing of data local to each node in a cluster enables Hadoop to store, manage, process and analyze data at petabyte scale.
- **Reliability** – large computing clusters are prone to failure of individual nodes in the cluster. Hadoop is fundamentally resilient – when a node fails processing is re-directed to the remaining nodes in the cluster and data is automatically re-replicated in preparation for future node failures.
- **Flexibility** – unlike traditional relational database management systems, you don't have to created structured schemas before storing data. You can store data in any format, including semi-structured or unstructured formats, and then parse and apply schema to the data when read.
- **Low Cost** – unlike proprietary software, Hadoop is open source and runs on low-cost commodity hardware.



Hadoop Framework

Hadoop framework and Apache projects:

- **Hadoop Common**: The common utilities that support the other Hadoop modules.
- **Hadoop Distributed File System (HDFS™)**: A distributed file system that provides high-throughput access to application data.
- **Hadoop YARN**: A framework for job scheduling and cluster resource management.
- **Hadoop MapReduce**: A YARN-based system for parallel processing of large data sets.

Hadoop-related projects at Apache include:

- **Ambar**: A web-based tool for provisioning, managing, and monitoring Apache Hadoop clusters which includes support for Hadoop HDFS, Hadoop MapReduce, Hive, HCatalog, HBase, ZooKeeper, Oozie, Pig and Sqoop. Ambari also provides a dashboard for viewing cluster health such as heatmaps and ability to view MapReduce, Pig and Hive applications visually alongwith features to diagnose their performance characteristics in a user-friendly manner.
- **Avro**: A data serialization system.
- **Cassandra**: A scalable multi-master database with no single points of failure.
- **Chukwa**: A data collection system for managing large distributed systems.
- **HBase**: A scalable, distributed database that supports structured data storage for large tables.
- **Hive**: A data warehouse infrastructure that provides data summarization and ad hoc querying.
- **Mahout**: A Scalable machine learning and data mining library.
- **Pig**: A high-level data-flow language and execution framework for parallel computation.
- **Spark**: A fast and general compute engine for Hadoop data. Spark provides a simple and expressive programming model that supports a wide range of applications, including ETL, machine learning, stream processing, and graph computation.
- **Tez**: A generalized data-flow programming framework, built on Hadoop YARN, which provides a powerful and flexible engine to execute an arbitrary DAG of tasks to process data for both batch and interactive use-cases. Tez is being adopted by Hive™, Pig™ and other frameworks in the Hadoop ecosystem, and also by other commercial software (e.g. ETL tools), to replace Hadoop™ MapReduce as the underlying execution engine.
- **ZooKeeper**: A high-performance coordination service for distributed applications.

Why Big Data Analytics?

Big data analytics examines large amounts of data to uncover hidden patterns, correlations and other insights. With today's technology, it's possible to analyze data and get answers from it almost immediately – an effort that's slower and less efficient with more traditional business intelligence solutions.

Big Data – Government

By implementing a big data platform, governments can access vast amounts of relevant information important to their daily functions. The positive effect it can have is nearly endless. It's so important because it not only allows the government to pinpoint areas that need attention, but it also gives them that information in real time. In a society that moves so quickly from one thing to the next, real-time analysis is vital. It allows governments to make faster decisions, and it allows them to monitor those decisions and quickly enact changes if necessary.

Uses of Census Data

Every country needs basic information on its residents for purposes of planning, development and improvement of the residents' quality of life. Good planning is based on reliable,up-to-date, accurate and detailed information on the state of the society in the country. This information makes it possible to plan better services, improve the quality of life and solve existing problems. Statistical information, which serves as the basis for constructing planning forecasts, is essential for the democratic process since it enables the citizens to examine the decisions made by the government and local authorities, and decide whether they serve the public they are meant to help. For these reasons official statistics are collected and published in all countries, world-wide. Thus, for example,while planning a road system, the planners use information regarding the quantity of people and number of vehicles who are likely to use the road; for if not, the advantages of constructing the road may not justify its cost. Similarly, when planning a school system, there is a need for a forecast of the number of pupils who are likely to need schools, in order to ensure that they will be built in locations with an appropriate number of pupils.

The importance of the population census

The census is a special, wide-range activity, which takes place once a decade in the entire country. Its purpose is to gather information about the general population, in order to present a full and reliable picture of the population in the country - its housing conditions and demographic, social and economic characteristics.The information collected includes data on,

- Age
- Gender
- Country of origin
- Immigration
- Marital status
- Marriage
- Parents
- Tax Status
- Education
- Income
- Employment

Immigration to the United States

Immigration to the United States is the international movement of individuals who are not natives or do not possess citizenship in order to settle, reside, study or to take-up employment in the United States. It has been a major source of population growth and cultural change throughout much of the history of the United States. The economic, social, and political aspects of immigration have caused controversy regarding ethnicity, economic benefits, jobs for non-immigrants, settlement patterns, impact on upward social mobility, crime, and voting behavior.

As for economic effects, research suggests that immigration to the United States is beneficial to the US economy. Research, with few exceptions, finds that immigration on average has positive economic effects on the native population, but is mixed as to whether low-skilled immigration adversely affects low-skilled natives. Research finds that immigration either has no impact on the crime rate or that it reduces the crime rate in the United States. Research shows that the United States excels at assimilating first- and second-generation immigrants relative to many other Western countries.

The census is extremely important for documenting the growth of immigrant communities, allocating resources for needed services, and identifying areas where civil rights enforcement may be needed. Department of Homeland Security (DHS), United States of America keeps track of all data from the Bureau of the Census, and do statistical analysis to plan its policy and funds.

Immigration Data and Statistics

Since the passage of the Homeland Security Act of 2002, the Office of Immigration Statistics (OIS) has responsibility to carry out two statutory requirements: 1) to collect and disseminate to Congress and the public data and information useful in evaluating the social, economic, environmental, and demographic impact of immigration laws; and 2) to establish standards of reliability and validity for immigration statistics collected by the Department's operational Components.

Feb, 2017

Benefit of Census

Everyone, including immigrants, benefits from investments in education, health care, and jobs that are distributed based on census information. And census data are also used in ways that are of special importance to immigrants, including:

- funding for nonprofit organizations to provide job assistance aimed at making foreign-born people economically self-sufficient;
- helping states and local agencies develop health care and other services tailored to the language and cultural diversity of immigrants, including health care and other services tailored to the language and cultural diversity of elderly people under the Older Americans Act.
- protecting the right to vote by evaluating voting practices of government subdivisions, such as states, counties, and school districts, under the Voting Rights Act;
- evaluating the effectiveness of equal opportunity employment programs and policies under the Civil Rights Act;
- allocating funds to school districts for children with limited English language proficiency.

Term in this Project

Immigrant - "immigrant" refer to persons with no U.S. citizenship at birth. This population includes naturalized citizens, lawful permanent residents, refugees and asylees, persons on certain temporary visas, and the unauthorized.

Project Outline

Title : US Census – Immigration Related Statistical Analysis

Input : Sample of US census data [JSON format]

Data Fields : Age, Education, Marital Status, Gender, Tax Filer Status, Income, Parents, Country of Birth, Citizenship, Weeks Worked

Lookup File : Age, Age Group

Analysis : Demographic Analysis – Education, Social and Economy

Calculations : Poverty, Tax per individual, Per Capita and Median Income

Purpose : Immigration Data and Statistical Analysis helps Department of Homeland Security(DHS), USA to make policies and planning like, creating job for immigrants, plan improved and extended health care, protect the right to vote for those naturalized, under Civil Rights Act. DHS ensures equal opportunity employment and allocates fund for schools and elderly people for health care, English language proficiency and social security assistance.

Project Implementation

Prerequisite: Hadoop Distributed File System access, Hive, Pig, Sqoop are installed in the node where the Hadoop is installed.

Mode: Hadoop standalone mode.

Pre-Process

Input file for the analysis should be reformatted and cleased for further processing.The input JSON format file can be converted to more usable format by,

- Hive-adding HCatalog jar file which contains JsonSerDe class from package 'org.apache.hive.hcatalog.data' and loading to hive table or in Pig by using CSVExcelStorage class from the package 'org.apache.pig.piggybank.storage' as storage in csv format.

Hive – Steps to load JSON file

1. Create database in Hive.

2. Move the jar file from Hive installed directory to HDFS, and enter hive command,

```
hive(immigration)>add jar hdfs://localhost:54310/hivejar/hive-hcatalog-core-1.2.1.jar;
```

3. Create table using the format,

```
hive(immigration)>create table table_name (field 1, field 2, field 3,…….field n)
                >row format serde 'org.apache.hive.hcatalog.data.JsonSerDe';
```

4. Load data to table by,

```
hive(immigration)>load data local inpath '/path of the input file'
                >overwrite into table table_name;
```

Pig – Steps to convert JSON file

1. Load the JSON file using

```
grunt>loadbag =load '/path of the input file' using JsonLoader('field 1,  field 3,….field n')
```

2. Store the JSON file as CSV using the command,

```
grunt>store loadbag into '/output_dir' using
>>org.apache.pig.piggybank.storage.CSVExcelStorage();
```

3. Output file can be used for further processing either in HDFS or in localfile system.

Input Files Details:

Census sample file contains,

1. Age
2. Education
3. MaritalStatus
4. Gender
5. TaxFilerStatus
6. Income
7. Parents
8. CountryOfBirth
9. Citizenship
10. WeeksWorked

The lookup file contains,

1. Age
2. AgeGroup

Demographic Analysis

- Tool : Hive
- Prerequisite : Sample data and lookup files should be loaded in Hive table.
- Sample data loaded in ‘census’ table and lookup file loaded in ‘age’ table of database ‘immigration’

1. Age group wise Immigrant Population.

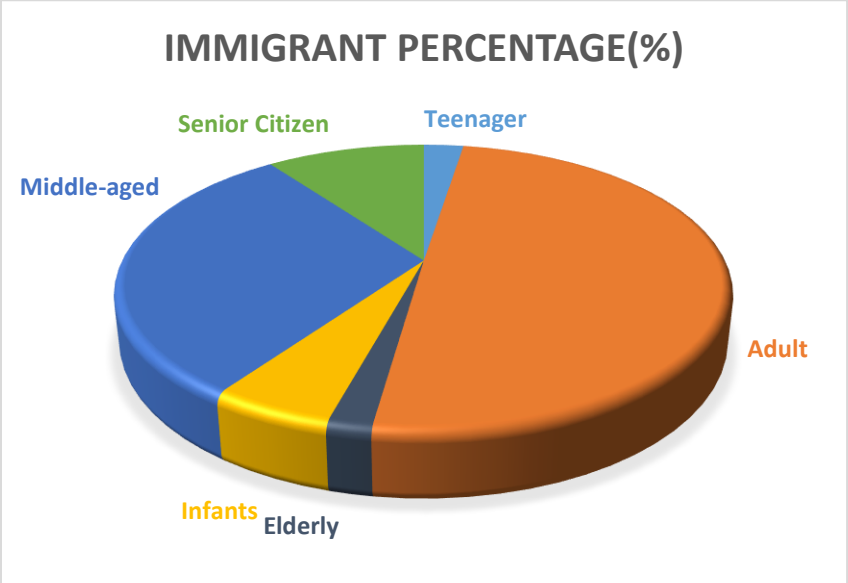
Hive Shell,

hive (immigration)> select b.agegroup as Age_Group,round((count(a.Citizenship)/t.tot)*100,2) as Immigrants_Percentage from census a, age b, (select (count(Citizenship)) as tot from census where (Citizenship=" Foreign born- Not a citizen of U S " or Citizenship= " Foreign born- U S citizen by naturalization")) t where a.age=b.age and (a.Citizenship=" Foreign born- Not a citizen of U S " or a.Citizenship= " Foreign born- U S citizen by naturalization") group by b.agegroup, t.tot;

Output:

Age Group	Immigrant Percentage(%)
Teenager	2.51
Adult	49.75
Elderly	2.01
Infants	5.53
Middle-aged	30.15
Senior Citizen	10.05

Visualization:



Immigrant Population distributed by Age Group

2. Education Group Wise Immigrant Population

```
hive (immigration)> select education,round((count(Citizenship)/t.tot)*100,2) as
Immigrants_Percentage from  census, (select (count(Citizenship)) as tot from census where
(Citizenship=" Foreign born- Not a citizen of U S " or Citizenship= " Foreign born- U S citizen
by naturalization")) t where (Citizenship=" Foreign born- Not a citizen of U S " or Citizenship= "
Foreign born- U S citizen by naturalization") group by education, t.tot order by
Immigrants_Percentage desc;
```

Output:

Row No.	Education	Percentage(%)
1	High school graduate	27.14
2	Some college but no degree	12.56
3	Bachelors degree(BA AB BS)	11.56
4	Children	5.53
5	7th and 8th grade	5.53
6	5th or 6th grade	5.03
7	9th grade	5.03
8	1st 2nd 3rd or 4th grade	4.02
9	Associates degree-academic program	4.02
10	10th grade	4.02
11	Less than 1st grade	3.52
12	11th grade	3.02
13	Associates degree-occup /vocational	3.02
14	Masters degree(MA MS MEng MEd MSW MBA)	2.01
15	Prof school degree (MD DDS DVM LLB JD)	2.01
16	12th grade no diploma	1.01
17	Doctorate degree(PhD EdD)	1.01

3. Population and Poverty:

- Tool : Pig
- Prerequisite : Sample data and lookup data available for Pig to load
- UDF : Apache DataFu Pig is a collection of user-defined functions for data analysis
- Methodology : For reusability, Pig commands are executed in BATCH MODE using .pig script file

censusPoverty.pig

```
register 'path_of_jar/datafu-1.2.0.jar';

define Median datafu.pig.stats.Median();
```

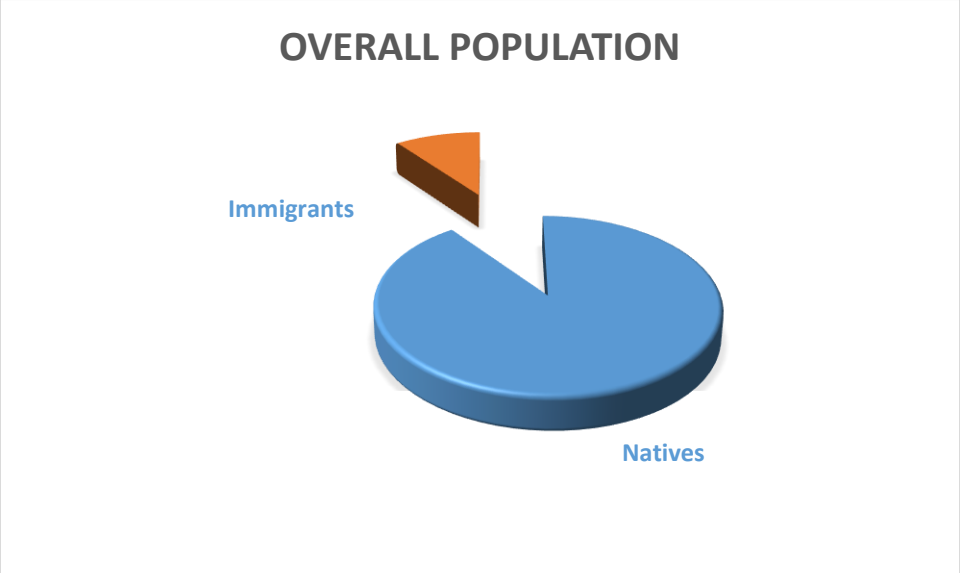
```
define BagConcat datafu.pig.bags.BagConcat();

census = load 'path_of_samplefile/sampledata'

using PigStorage(',') as (Age: int, Income:double, Citizenship:chararray);

total = foreach (group census all) generate COUNT(census) as total;

totimmi = filter census by (Citizenship == ' Foreign born- Not a citizen of U S ') OR
(Citizenship == ' Foreign born- U S citizen by naturalization');

immigrants = foreach (group totimmi all) generate COUNT(totimmi) as immigrants;

age = load 'path_of_look_up_file/agegroup' using PigStorage('\t') as (Age: int,
Agegroup:chararray);

joinage = join census by Age, age by Age;

joinedbag = foreach joinage generate $0,$11 as Agegroup,$1,$2,$3,$4,$5 as
Income,$6,$7,$8 as Citizenship,$9;

ordercensus = order joinedbag by Income asc;

medianincome = foreach (group ordercensus all) generate
FLATTEN(Median(ordercensus.Income)) as Median;

poverty = filter ordercensus by (Income<(medianincome.Median*0.60));

totalpoverty = foreach (group poverty all) generate COUNT(poverty) as tot_poverty;

                                    ---------- POVERTY IMMIGRANT----------

immipov = filter poverty by (Citizenship == ' Foreign born- Not a citizen of U S ') OR
(Citizenship == ' Foreign born- U S citizen by naturalization');

immipovcount = foreach (group immipov all) generate COUNT(immipov) as
immigrant_poverty;

immichild = filter immipov by (Agegroup == 'infants');

childpov = foreach (group immichild all) generate COUNT(immichild) as childpov;

immiold = filter immipov by (Agegroup == 'senior citizen') or (Agegroup == 'elderly');

oldpov = foreach (group immiold all) generate COUNT(immiold) as oldpov;

joinbags = cogroup total by total, immigrants by immigrants, totalpoverty by tot_poverty,
immipovcount by immigrant_poverty;

bagcon = foreach joinbags generate
BagConcat(total,immigrants,totalpoverty,immipovcount);
```

```
inters = foreach bagcon generate total.total, immigrants.immigrants, (total.total -
immigrants.immigrants) as natives, totalpoverty.tot_poverty,
immipovcount.immigrant_poverty, (totalpoverty.tot_poverty -
immipovcount.immigrant_poverty) as native_poverty, childpov.childpov, oldpov.oldpov;

interlimit = limit inters 1;

percentage = foreach interlimit generate total, natives,

        (((double)natives*100)/(double)total) as native_per,immigrants,

        (((double)immigrants*100)/(double)total) as immi_per, tot_poverty,

        (((double)tot_poverty*100)/(double)total) as poverty_per, native_poverty,

        ROUND_TO((((double)native_poverty*100)/(double)natives),2) as
nativepov_per,

        ROUND_TO((((double)native_poverty*100)/(double)tot_poverty),2) as
nativetotpov_per, immigrant_poverty,

        ROUND_TO((((double)immigrant_poverty*100)/(double)immigrants),2) as
immipoverty_per,

        ROUND_TO((((double)immigrant_poverty*100)/(double)tot_poverty),2) as
immitotpov_per,

        ROUND_TO((((double)childpov*100)/(double)immigrants),2) as
immichildpov_per,

        ROUND_TO((((double)oldpov*100)/(double)immigrants),2) as immioldpov_per;

percent = foreach percentage generate CONCAT(

        '\nTotal Sample Population \t',(chararray)$0,

        '\n\nTotal Natives Percentage\t',(chararray)$2,'% of total population',

        '\nTotal Immigrants Percentage\t',(chararray)$4,'% of total population',

        '\n\nTotal Poverty Percentage\t',(chararray)$6,'% of total population',

        '\n\nNative Poverty percentage\t',(chararray)$8,'% among natives',

        '\nNative Poverty Percentage\t',(chararray)$9,'% among total poverty',

        '\n\nImmigrant Poverty Percent.\t',(chararray)$11,'% among immigrants',

        '\nImmigrant Poverty Percent.\t',(chararray)$12,'% among total poverty',

        '\n\nInfant Poverty Percentage\t',(chararray)$13,'% among immigrants',
```

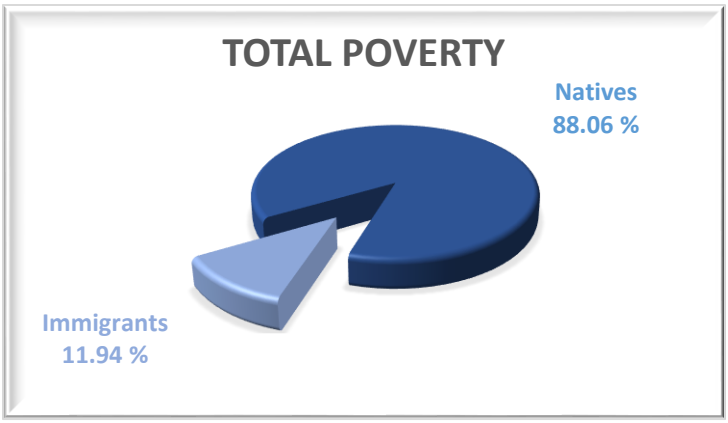
```
        '\nOldage Poverty Percentage\t',(chararray)$14,'% among immigrants\n');

store percent into 'output_path/PovertyPercentage';
```

Output: /PovertPercentage/part-r-00000 file,

Total Sample Population	2000
Total Natives Percentage	90.05% of total population
Total Immigrants Percentage	9.95% of total population
Total Poverty Percentage	20.1% of total population
Native Poverty percentage	19.66% among natives
Native Poverty Percentage	88.06% among total poverty
Immigrant Poverty Percent.	24.12% among immigrants
Immigrant Poverty Percent.	11.94% among total poverty
Infant Poverty Percentage	2.01% among immigrants
Oldage Poverty Percentage	3.02% among immigrants



Natives Vs. Immigrants [Population]



Natives Vs. Immigrants [Poverty]

4. Which country sources large educated people and contributes more to US by tax?

- Tool : Pig
- Prerequisite : Sample data available for Pig to load, no lookup file needed
- The Sample data contains Tax field which is executed in MapReduce, and will be discussed later, the MapReduce Java file is available in https://github.com/bashyan/Immigration_Data_Statistical_Analysis/blob/master/tax.java

country.pig

> **loaddata** = load 'path_of_input_file/TaxMapper' using PigStorage(',') as (education:chararray, taxfilerstatus:chararray, tax:double, country:chararray, citizenship:chararray);
>
> **immigrants** = filter loaddata by (taxfilerstatus != ' Nonfiler') and ((citizenship == ' Foreign born- Not a citizen of U S ') OR (citizenship == ' Foreign born- U S citizen by naturalization')) and (country != ' ?');
>
> **taxcal** = foreach (group immigrants by country) generate group, ROUND_TO(SUM(immigrants.tax),2) as taxes ;
>
> **listcountry** = order taxcal by taxes desc;
>
> **topcountry** = limit listcountry 1;
>
> **countrytaxper** = foreach (group listcountry all) generate CONCAT('Tax Contribution : ', (chararray)topcountry.$0, CONCAT(' contributed more in income tax and accounted for about ', (chararray)ROUND_TO((topcountry.$1*100)/SUM(listcountry.$1),2)), '% of total tax revenue from immigrants');
>
> **educated** = filter loaddata by ((education == ' Bachelors degree(BA AB BS)') or (education == ' Masters degree(MA MS MEng MEd MSW MBA)') or (education == ' Prof school degree (MD DDS DVM LLB JD)') or (education == ' Associates degree-academic program') or (education == ' Doctorate degree(PhD EdD)') or (education == ' Associates degree-occup /vocational') or (education == ' High school graduate') or (education == '

```
Some college but no degree')) and ((citizenship == ' Foreign born- Not a citizen of U S ') OR (citizenship == ' Foreign born- U S citizen by naturalization')) and (country != ' ?');

educal = foreach (group educated by country) generate group, COUNT(educated);

listedu = order educal by $1 desc;

topedu = limit listedu 1;

topeduper = foreach (group listedu all) generate CONCAT('Educated Immigrants : ', (chararray)topedu.$0, CONCAT(' sources more educated immigrants and accountes ',(chararray)((topedu.$1*100)/SUM(listedu.$1))), '% of total educated immigrant population');

store countrytaxper into '/output_path_1/CountryTaxPaid';

store topeduper into '/output_path_2/CountryEducated';
```

Output 1: /CountryTaxPaid/part-r-00000 file,

Tax Contribution : **Mexico** contributed more in income tax and accounted for about 31.77% of total tax revenue from immigrants

Output 2: /CountryEducated/part-r-00000 file,

Educated Immigrants : **Mexico** sources more educated immigrants and accounts 18% of total educated immigrant population

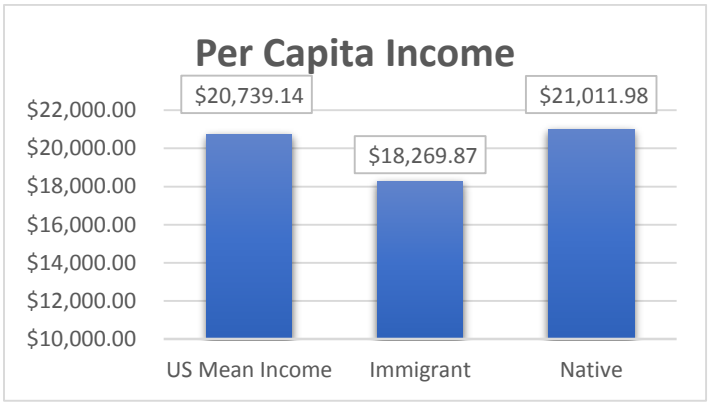
5. Socio - Economic Analysis – Tax Distribution and Per Capita Income between Natives and Immigrants, Is "Immigrant Ban" required?

- Tool : MapReduce, Eclipse
- Prerequisite : Sample data files should be in HDFS
- Methodology: Reduce Side Join is used in this process. One Mapper calculate **TAX** and other Mapper is meant for cleansing data for **Per Capita Income** calculation. The reducer calculate the tax distribution and per capita income (mean income).
- For Tax Nonfilers, tax is calculated considering the individual as 'Single' filer.
- MapReduce Java file is available in https://github.com/bashyan/Immigration_Data_Statistical_Analysis/blob/master/taxpayer.java

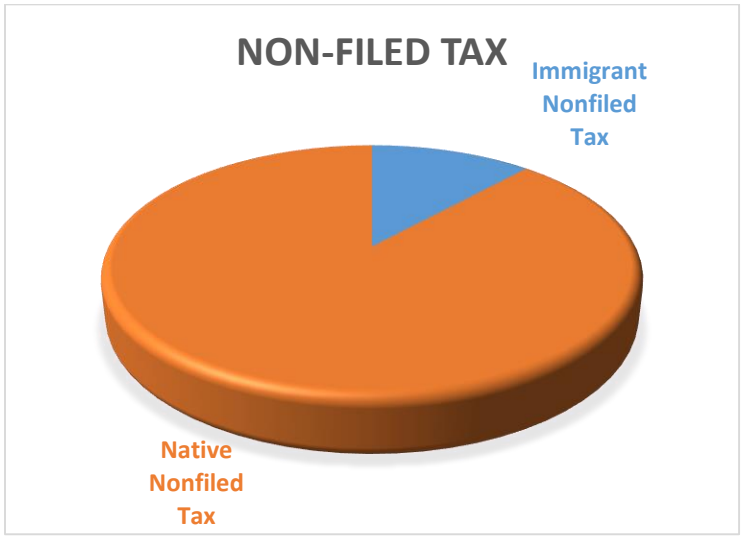
Output part-r-00000 file,

1, USA Total Tax:	$ 4001038.81
2, Total Immigrant Tax:	$ 405905.40
3, Immigrant Nonfiled Tax:	$ 87723.19
4, Total Native Tax:	$ 3595133.41
5, Native Nonfiled Tax:	$ 657909.09
6, US Mean Income:	$ 20739.14
7, Immigrant's Mean Income:	$ 18269.87
8, Native's Mean Income:	$ 21011.98

Socio-Economic Data Visualization:



Tax Distribution Visualization:



Native Vs. Immigrant [Non-Tax Filed]

Data Export to Relational Database

The census data can be used whenever needed, apart from HDFS, it is better to export data to the local database where analysts do simple query in MySQL and do analysis in future.

This can be achieved using Apache Sqoop,

- Sqoop is used to import and export data from or to Relational Databases
- HDFS file can be directly exported to RDBMS

MapReduce output is written in HDFS as **part-r-ooo..** or **part-m-ooo…** files

In this project the data with tax calculation is exported to MySQL from HDFS.

Steps for export:

1. Create database in MySQL
2. Create table with exact schema of HDFS file
3. Using Sqoop export command the data can be exported,
    - ~$ sqoop export --connect jdbc:mysql://localhost/census --username root --password '------' --table census_data --export-dir /immigrantProject/Censusdata/tax/part-m-00000;
4. In MySQL, the table can be used for further analysis
    - mysql> select Education, round(sum(tax),2) as Tax_Paid from census_data group by education order by tax_paid desc limit 3;
    - Output:

Education	Tax_Paid	
High school graduate	$	1,314,928.20
Some college but no degree	$	814,841.71
Bachelors degree(BA AB BS)	$	469,355.37

Analysis Report: from the sample census data

- In overall population foreign born accounts for 9.95% and natives accounts for 90.05%
- Adult age group with high school graduated are the large numbers of foreign born living in USA.
- Among entire population nearly 20.1% of people are living under poverty and natives are the most hit people of poverty and they account for about 88.06% of total poverty.
- Among immigrants, children and elderly age people have poor standard and are under poverty, they comprise nearly 2.01% and 3.02% respectively.
- Mexico sources more immigrants and their population comprise nearly 18 % of total educated immigrants and also they pay more income tax which generates tax revenue of about 31.77% of total tax collected from foreign born.
- The per capita income of immigrants is lesser than that of natives, as it shows that immigrants are paid lesser.
- Analysing the tax filer details, more number of immigrants and natives doesn't file their income tax.

Suggestions from the report:

- Native people standard should be improved and more jobs to be created for natives.
- Though children and old age immigrants have lesser poverty level, healthcare and education for them to be improved. They should be given **higher social security** assistance.
- The per capita income of immigrants is 14% lesser than that of natives, Civil Rights Act to be revised and **equal employment** opportunity and salary should be ensured.
- As tax revenue by non-filer shows significant amount, more number of campaigns can be arranged to create awareness among people.
- More immigrants come from Mexico, and the **diplomatic relation** to be improved and community education opportunity can be provided for Mexicans.
- The detailed analysis on tax and income shows that on an average, Immigrant pay tax 2.1% more than Native people and the per capita income of Immigrant is 13.05% lesser than Native. It signifies that Immigrant's contribution to US wealth is highly assessable and rising ban on Immigrant will **decline** the nation's revenue.
- On analysing the sample census data **without** considering **illegal** migrants, the foreign born contribute more to USA and it can be concluded that '**Immigration Ban'** is not necessary for United States of America.