JSS MAHAVIDYAPEETHA
JSS SCIENCE AND TECHNOLOGY UNIVERSITY
SRI JAYACHAMARAJENDRA COLLEGE OF ENGINEERING
JSS Technical Institutions Campus, Mysuru – 570006

"Aadhaar Data Analysis Using Hadoop"

Mini project report submitted in partial fulfillment of curriculum prescribed for the Big Data Analytics (CS620) course for the award of the degree of

BACHELOR OF ENGINEERING
IN
COMPUTER SCIENCE AND ENGINEERING

By

Dinaraj Singh
(01JST17CS141)

Rahul Singh
(01JST18CS098)

Sanajy C B
(01JST18CS119)

Prajwalan M
(01JST19CS413)

Under the Guidance of
Prof. Shruthi N M
Assistant Professor,
Dept.of CS & E,
SJCE, JSS STU Mysore

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
JUNE 2021

JSS MAHAVIDYAPEETHA

JSS SCIENCE AND TECHNOLOGY UNIVERSITY

SRI JAYACHAMARAJENDRA COLLEGE OF ENGINEERING
JSS Technical Institutions Campus, Mysuru – 570006

CERTIFICATE

This is to certify that the work entitled **“*Aadhaar Data Analysis Using Hadoop*”** is a bonafied work carried out **by Dinaraj Singh, Rahul Singh, Sanjay C B and Prajwalan M** in partial fulfillment of the award of the degree of **Bachelor of Engineering in Computer Science and Engineering of SJCE, JSS Science and Technology, Mysuru during the year 2021**. It is certified that all corrections / suggestions indicated during CIE have been incorporated in the report. The mini project report has been approved as it satisfies the academic requirements in respect of mini project work prescribed for the **Big Data Analytics (CS620)** course.

Course in Charge and Guide
Prof. Shruthi N M
Assistant Professor,
Dept.of CS & E,
SJCE, JSS STU Mysore

Place: **Mysore**

Date :

DECLARATION

We, the undersigned, solemnly declare that the project report "***Aadhaar Data Analysis using HADOOP***" is based on our work carried out during the course of our study under the supervision of ***Prof. Shruthi N M*** **(Assistant Professor, Dept. of CS & E, SJCE, JSS STU Mysuru)**. We assert the statements made and conclusions are drawn are an outcome of our research work. We further certify that

I. The work contained in the report is original and has been done by us under the general supervision of our supervisor.

II. The work has not been submitted to any other Institution for any other degree/diploma/certificate in this university or any other University of India or abroad.

III. We have followed the guidelines provided by the university in writing the report.

IV. Whenever we have used materials (data, theoretical analysis, and text) from other sources, we have given due credit to them in the text of the report and given their details in the references.

Dinaraj Singh
(01JST17CS141)

Rahul Singh
(01JST18CS098)

Sanajy C B
(01JST18CS119)

Prajwalan M
(01JST19CS413)

ABSTRACT

Aadhaar provides an essential details about an individual with 12-digit unique identification number that contains all the details about an individual, including demographic and biometric information of every resident Indian individual. Aadhaar is a big data which need to be stored and managed securely and safely. Several processing techniques and privacy measures have introduced to process such huge confidential data. However, identifying individual details which may be used by different sectors is not linked or updated with Aadhaar data.

In order to update essential details of an individual along with existing database of an Aadhaar for use by crime department, health care centre and professionals, several algorithms, tools, techniques used in big data analytics have been discussed in this survey paper. This is useful for hospitals for retrieving blood donor details, crime investigation and professionals for retrieving the details about residents along with their Aadhaar details.

ACKNOWLEDGEMENT

We would like to express our deepest appreciation to all those who provided us with the possibility to complete this report. A special gratitude we give to our project guide, **Prof Shruthi N M** whose contribution in stimulating suggestions and encouragement, helped us to coordinate our project.

PROJECT TEAM DETAILS

Dinraj Singh
(01JST17CS141)

Rahul Singh
(01JST18CS098)

Sanajy C B
(01JST18CS119)

Prajwalan M
(01JST19CS413)

Table of Contents

1. INTRODUCTION

Big Data is a collection of data that is huge in volume, yet growing exponentially with time. It is data with a huge size and complexity that none of the traditional data management tools can be used to store it or process it efficiently. With the help of big data tools, we look to create an Aadhar analysis system by performing analysis on big data which satisfies all the characteristics of the same (Velocity, Veracity, Volume, Variety).

a. Introduction to Problem Domain:

(i) The world has largest democracy, India is the second largest nation in terms of population, with 1.3 billion population.
(ii) Among these, 99% of adult population enrolled for Aadhar, the unique identity provided by the Government of India for diverse purposes.
(iii) The government maintains the Aadhar related data in digital format. https://data.uidai.gov.in/uiddatacatalog/dataCatalogHome.do website provides the access to Aadhar card related data set.
(iv) The Public can access some of the sources of these data and they can analyze to extract useful information and generate reports.
(v)So the amount of data generated by Aadhar is very huge. Similarly, all the data collected for this unique identity is not in structured data.
(vi) The purpose of Hadoop is storing and processing large amount of the data. So this project uses the Hadoop for processing Aadhar data.
(vii) The input data is processed using MapReduce and then result is loaded into Hadoop Distributed File System (HDFS).
(viii) Final reports generated using Tableau (Business Intelligence Software).

b. Aim of the project:

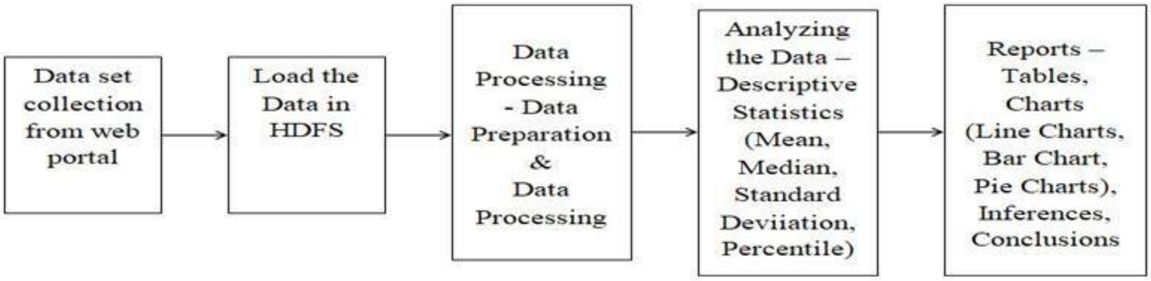
To analyse the Aadhar dataset using Hadoop and make some analysis based on gender, age and state wise**.**

c. Objective:

To analyze the Aadhar data using Hadoop to extract meaningful knowledge for the purpose of better decision-making by the central and state government.

d. Proposed System:

The proposed system concentrates on analyzingAadhar related data using Hadoop for the purpose of better decision making by the Government of India. The proposed system architecture is shown in the figure.


Data set collection from web portal
Load the Data in HDFS
Data Processing - Data Preparation & Data Processing
Analyzing the Data – Descriptive Statistics (Mean, Median, Standard Deviiation, Percentile)
Reports – Tables, Charts (Line Charts, Bar Chart, Pie Charts), Inferences, Conclusions

Step 1: Data Preparation

Data Selection: The required data set is collected from the government web portal.

Data Loading: The collected data set loaded into Hadoop Distributed File System environment.

Data Pre processing: The collected data set might consist of missing values and noisy data. If analysis is performed on this data, it may lead to wrong results. So to avoid this, data pre processing is done on the data set.

Step 2: Data Analysis

Data Analysis: Now the collected data set is ready for data analysis. Descriptive statistics like mean, median, mode, percentile are applied.

Step 3: Results

Report Generation: After the data analysis, the analyzed results need to be visualized. Tableau can be used for this purpose. Bar charts, Line charts and Pie charts are generated along with the table format.

e. Application:

It will be easy to classify the data for Government...

2. TOOLS AND TECHNOLOGY USED

The tools and frameworks used to set up the entire project are as follows:

Hadoop: Hadoop is an open-source software framework for storing data and running applications on clusters of commodity hardware. It provides massive storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs.

MapReduce: Map Reduce is a programming framework which can do parallel processing on nodes in a cluster. It takes input and gives the output in form of key-value pairs. After collecting the data, it has to be processed so that meaningful information can be extracted out of it which can serve as decision support system.The benefit of using this is that there is a need to write much fewer lines of code which reduces overall time needed for development and testing of code.

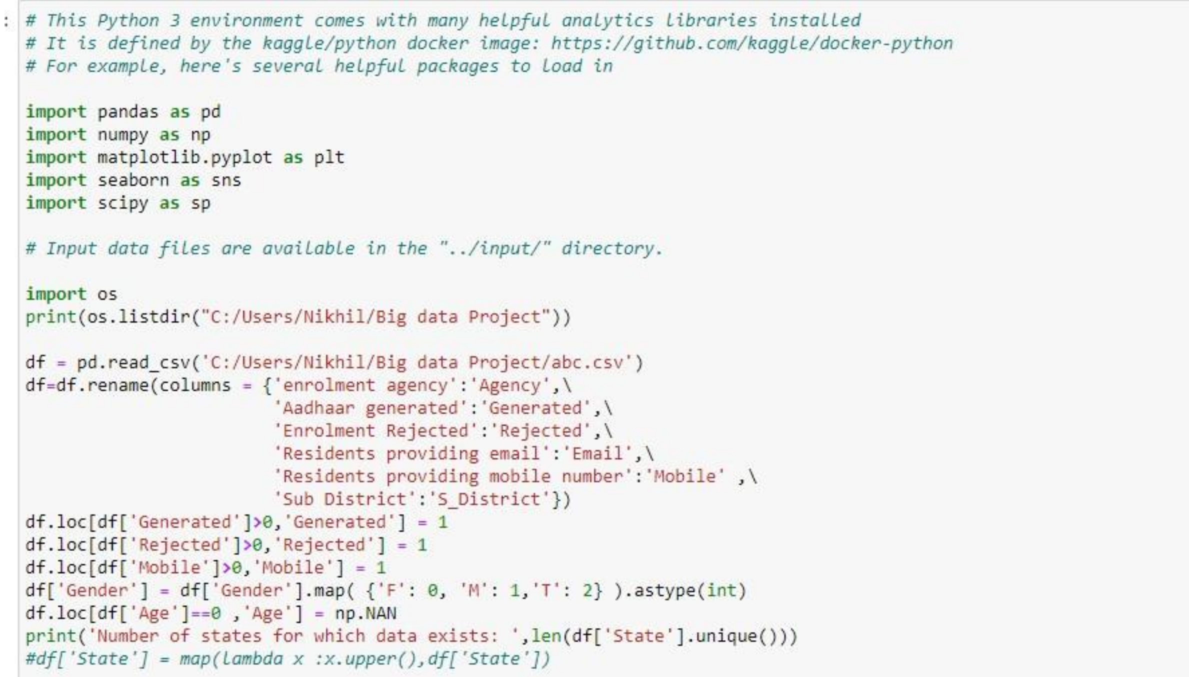
Excel: In Excel Aadhar dataset is stored in the CSV file format.

Tableau :Tableau helps people see and understand data. Our visual analytics platform is transforming the way people use data to solve problems.

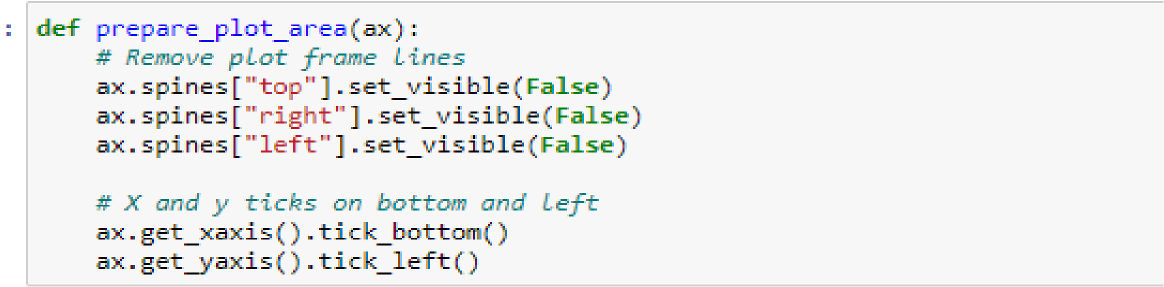
3. SYSTEM DESIGN IMPLEMENTATION

Just python for taking data and MATPLOTLIB for plotting the graph and we have not designed any frontend because for our project we can’t build any frontend design…

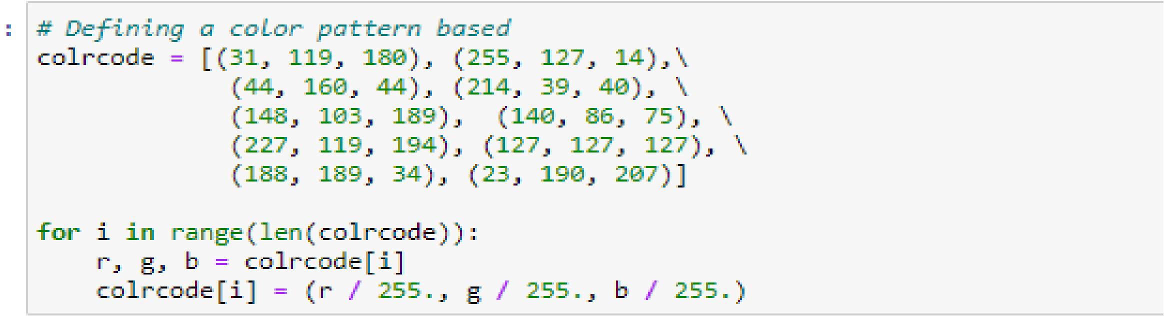
Code



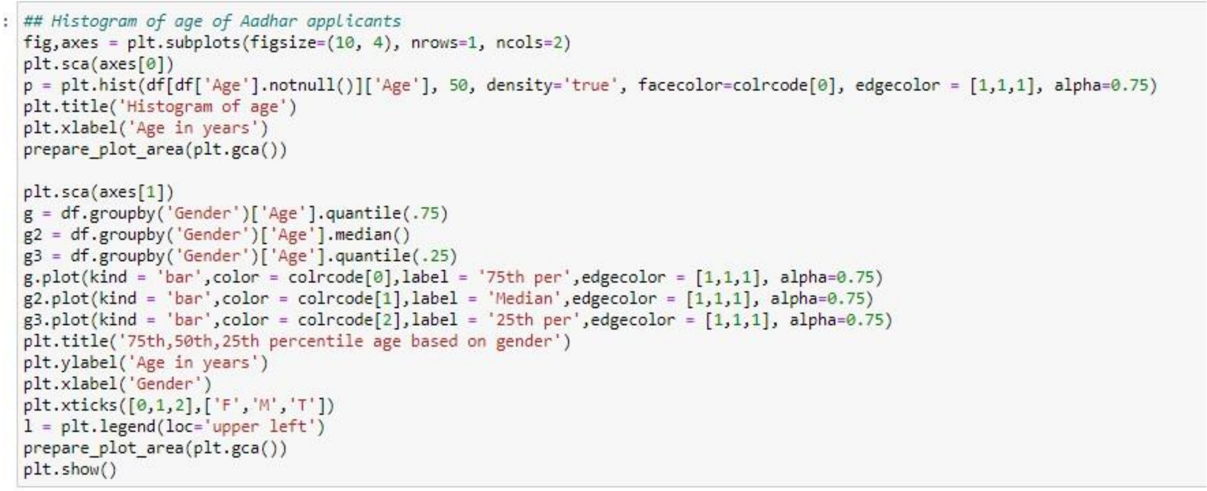
```
: # This Python 3 environment comes with many helpful analytics libraries installed
  # It is defined by the kaggle/python docker image: https://github.com/kaggle/docker-python
  # For example, here's several helpful packages to load in

  import pandas as pd
  import numpy as np
  import matplotlib.pyplot as plt
  import seaborn as sns
  import scipy as sp

  # Input data files are available in the "../input/" directory.

  import os
  print(os.listdir("C:/Users/Nikhil/Big data Project"))

  df = pd.read_csv('C:/Users/Nikhil/Big data Project/abc.csv')
  df=df.rename(columns = {'enrolment agency':'Agency',\
                          'Aadhaar generated':'Generated',\
                          'Enrolment Rejected':'Rejected',\
                          'Residents providing email':'Email',\
                          'Residents providing mobile number':'Mobile' ,\
                          'Sub District':'S_District'})
  df.loc[df['Generated']>0,'Generated'] = 1
  df.loc[df['Rejected']>0,'Rejected'] = 1
  df.loc[df['Mobile']>0,'Mobile'] = 1
  df['Gender'] = df['Gender'].map( {'F': 0, 'M': 1,'T': 2} ).astype(int)
  df.loc[df['Age']==0 ,'Age'] = np.NAN
  print('Number of states for which data exists: ',len(df['State'].unique()))
  #df['State'] = map(lambda x :x.upper(),df['State'])
```



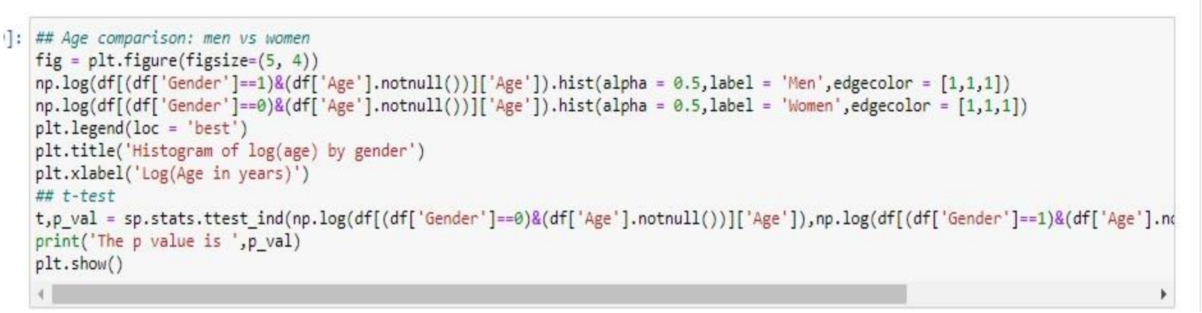
```
def prepare_plot_area(ax):
    # Remove plot frame lines
    ax.spines["top"].set_visible(False)
    ax.spines["right"].set_visible(False)
    ax.spines["left"].set_visible(False)

    # X and y ticks on bottom and left
    ax.get_xaxis().tick_bottom()
    ax.get_yaxis().tick_left()
```



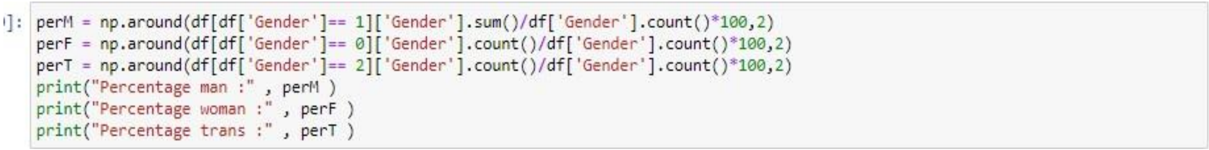
```
# Defining a color pattern based
colrcode = [(31, 119, 180), (255, 127, 14),\
             (44, 160, 44), (214, 39, 40), \
             (148, 103, 189),  (140, 86, 75), \
             (227, 119, 194), (127, 127, 127), \
             (188, 189, 34), (23, 190, 207)]

for i in range(len(colrcode)):
    r, g, b = colrcode[i]
    colrcode[i] = (r / 255., g / 255., b / 255.)
```



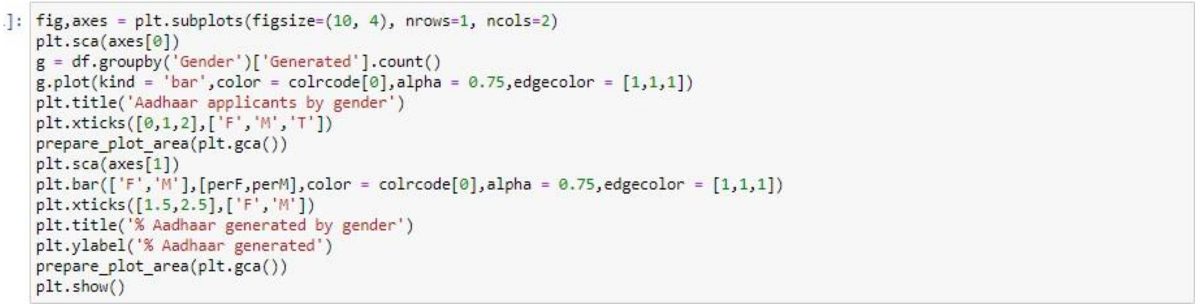
```
## Histogram of age of Aadhar applicants
fig,axes = plt.subplots(figsize=(10, 4), nrows=1, ncols=2)
plt.sca(axes[0])
p = plt.hist(df[df['Age'].notnull()]['Age'], 50, density='true', facecolor=colrcode[0], edgecolor = [1,1,1], alpha=0.75)
plt.title('Histogram of age')
plt.xlabel('Age in years')
prepare_plot_area(plt.gca())

plt.sca(axes[1])
g = df.groupby('Gender')['Age'].quantile(.75)
g2 = df.groupby('Gender')['Age'].median()
g3 = df.groupby('Gender')['Age'].quantile(.25)
g.plot(kind = 'bar',color = colrcode[0],label = '75th per',edgecolor = [1,1,1], alpha=0.75)
g2.plot(kind = 'bar',color = colrcode[1],label = 'Median',edgecolor = [1,1,1], alpha=0.75)
g3.plot(kind = 'bar',color = colrcode[2],label = '25th per',edgecolor = [1,1,1], alpha=0.75)
plt.title('75th,50th,25th percentile age based on gender')
plt.ylabel('Age in years')
plt.xlabel('Gender')
plt.xticks([0,1,2],['F','M','T'])
l = plt.legend(loc='upper left')
prepare_plot_area(plt.gca())
plt.show()
```



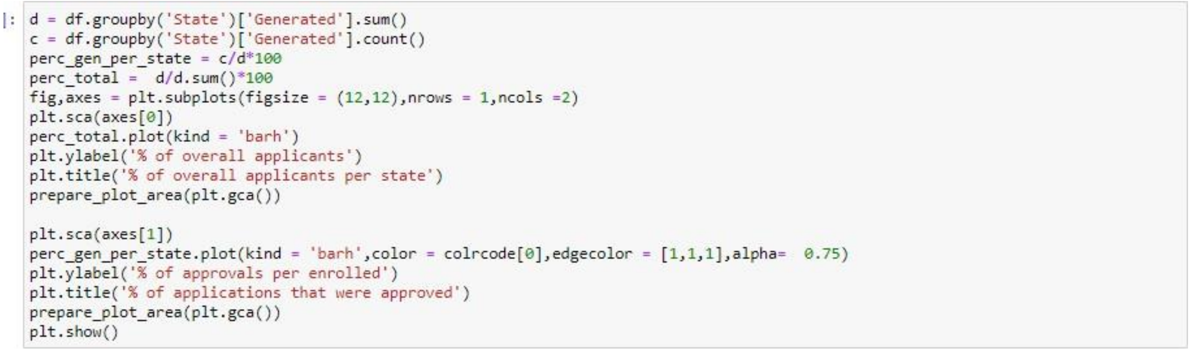
```
## Age comparison: men vs women
fig = plt.figure(figsize=(5, 4))
np.log(df[(df['Gender']==1)&(df['Age'].notnull())]['Age']).hist(alpha = 0.5,label = 'Men',edgecolor = [1,1,1])
np.log(df[(df['Gender']==0)&(df['Age'].notnull())]['Age']).hist(alpha = 0.5,label = 'Women',edgecolor = [1,1,1])
plt.legend(loc = 'best')
plt.title('Histogram of log(age) by gender')
plt.xlabel('Log(Age in years)')
## t-test
t,p_val = sp.stats.ttest_ind(np.log(df[(df['Gender']==0)&(df['Age'].notnull())]['Age']),np.log(df[(df['Gender']==1)&(df['Age'].nc
print('The p value is ',p_val)
plt.show()
```



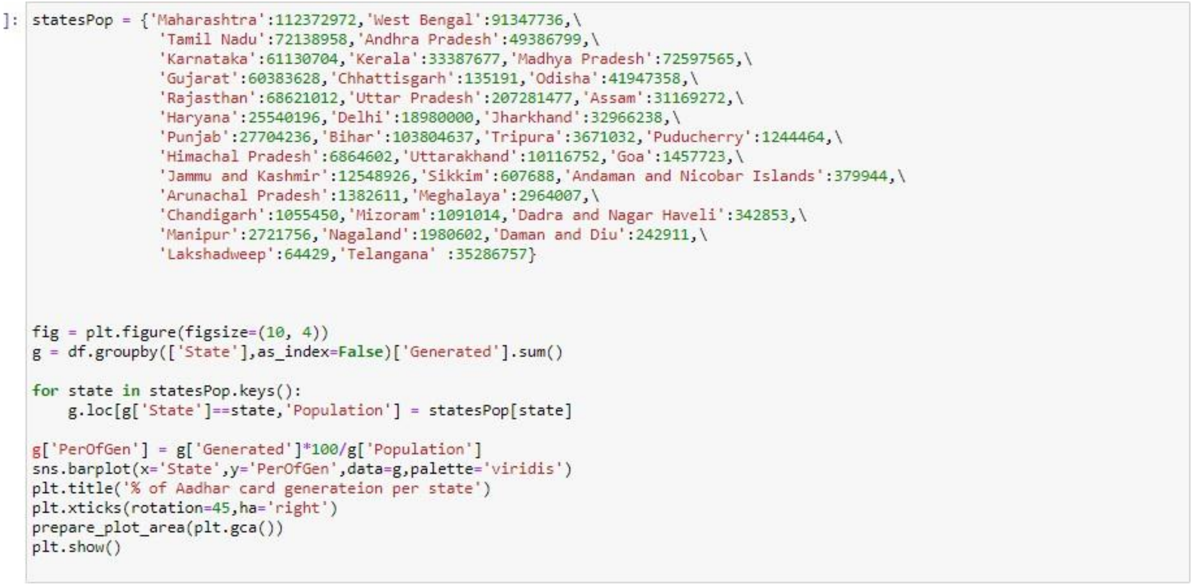
```
perM = np.around(df[df['Gender']== 1]['Gender'].sum()/df['Gender'].count()*100,2)
perF = np.around(df[df['Gender']== 0]['Gender'].count()/df['Gender'].count()*100,2)
perT = np.around(df[df['Gender']== 2]['Gender'].count()/df['Gender'].count()*100,2)
print("Percentage man :" , perM )
print("Percentage woman :" , perF )
print("Percentage trans :" , perT )
```



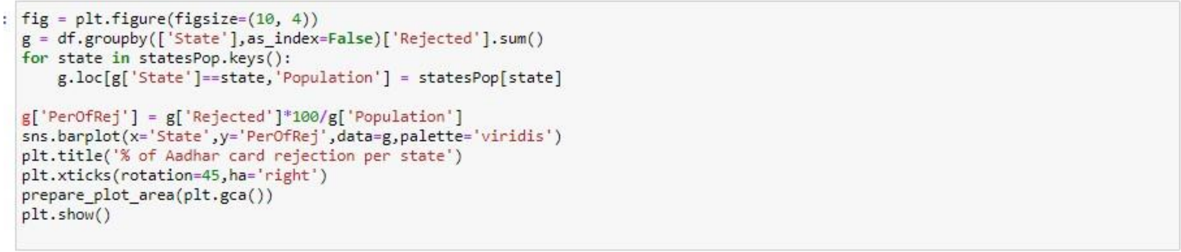
```
fig,axes = plt.subplots(figsize=(10, 4), nrows=1, ncols=2)
plt.sca(axes[0])
g = df.groupby('Gender')['Generated'].count()
g.plot(kind = 'bar',color = colrcode[0],alpha = 0.75,edgecolor = [1,1,1])
plt.title('Aadhaar applicants by gender')
plt.xticks([0,1,2],['F','M','T'])
prepare_plot_area(plt.gca())
plt.sca(axes[1])
plt.bar(['F','M'],[perF,perM],color = colrcode[0],alpha = 0.75,edgecolor = [1,1,1])
plt.xticks([1.5,2.5],['F','M'])
plt.title('% Aadhaar generated by gender')
plt.ylabel('% Aadhaar generated')
prepare_plot_area(plt.gca())
plt.show()
```



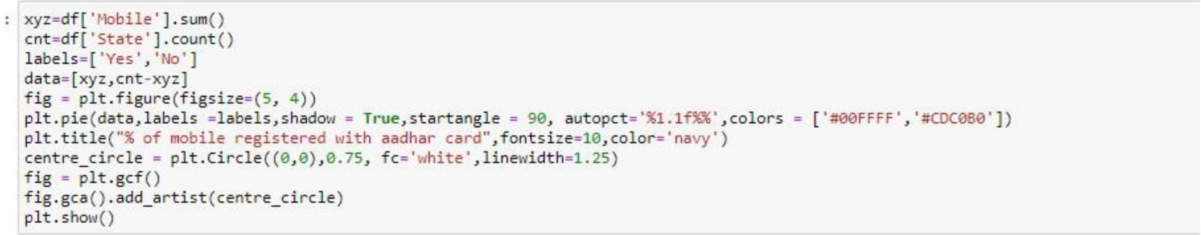
```
d = df.groupby('State')['Generated'].sum()
c = df.groupby('State')['Generated'].count()
perc_gen_per_state = c/d*100
perc_total =  d/d.sum()*100
fig,axes = plt.subplots(figsize = (12,12),nrows = 1,ncols =2)
plt.sca(axes[0])
perc_total.plot(kind = 'barh')
plt.ylabel('% of overall applicants')
plt.title('% of overall applicants per state')
prepare_plot_area(plt.gca())

plt.sca(axes[1])
perc_gen_per_state.plot(kind = 'barh',color = colrcode[0],edgecolor = [1,1,1],alpha=  0.75)
plt.ylabel('% of approvals per enrolled')
plt.title('% of applications that were approved')
prepare_plot_area(plt.gca())
plt.show()
```



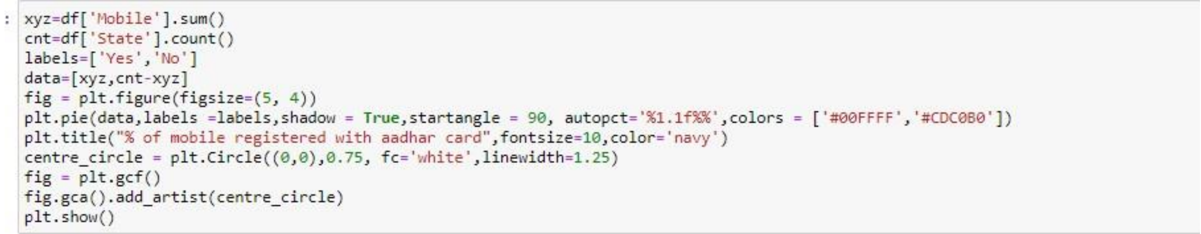
```
statesPop = {'Maharashtra':112372972,'West Bengal':91347736,\
                 'Tamil Nadu':72138958,'Andhra Pradesh':49386799,\
                 'Karnataka':61130704,'Kerala':33387677,'Madhya Pradesh':72597565,\
                 'Gujarat':60383628,'Chhattisgarh':135191,'Odisha':41947358,\
                 'Rajasthan':68621012,'Uttar Pradesh':207281477,'Assam':31169272,\
                 'Haryana':25540196,'Delhi':18980000,'Jharkhand':32966238,\
                 'Punjab':27704236,'Bihar':103804637,'Tripura':3671032,'Puducherry':1244464,\
                 'Himachal Pradesh':6864602,'Uttarakhand':10116752,'Goa':1457723,\
                 'Jammu and Kashmir':12548926,'Sikkim':607688,'Andaman and Nicobar Islands':379944,\
                 'Arunachal Pradesh':1382611,'Meghalaya':2964007,\
                 'Chandigarh':1055450,'Mizoram':1091014,'Dadra and Nagar Haveli':342853,\
                 'Manipur':2721756,'Nagaland':1980602,'Daman and Diu':242911,\
                 'Lakshadweep':64429,'Telangana' :35286757}



fig = plt.figure(figsize=(10, 4))
g = df.groupby(['State'],as_index=False)['Generated'].sum()

for state in statesPop.keys():
    g.loc[g['State']==state,'Population'] = statesPop[state]

g['PerOfGen'] = g['Generated']*100/g['Population']
sns.barplot(x='State',y='PerOfGen',data=g,palette='viridis')
plt.title('% of Aadhar card generateion per state')
plt.xticks(rotation=45,ha='right')
prepare_plot_area(plt.gca())
plt.show()
```



```
: fig = plt.figure(figsize=(10, 4))
  g = df.groupby(['State'],as_index=False)['Rejected'].sum()
  for state in statesPop.keys():
      g.loc[g['State']==state,'Population'] = statesPop[state]

  g['PerOfRej'] = g['Rejected']*100/g['Population']
  sns.barplot(x='State',y='PerOfRej',data=g,palette='viridis')
  plt.title('% of Aadhar card rejection per state')
  plt.xticks(rotation=45,ha='right')
  prepare_plot_area(plt.gca())
  plt.show()
```



```
: xyz=df['Mobile'].sum()
  cnt=df['State'].count()
  labels=['Yes','No']
  data=[xyz,cnt-xyz]
  fig = plt.figure(figsize=(5, 4))
  plt.pie(data,labels =labels,shadow = True,startangle = 90, autopct='%1.1f%%',colors = ['#00FFFF','#CDC0B0'])
  plt.title("% of mobile registered with aadhar card",fontsize=10,color='navy')
  centre_circle = plt.Circle((0,0),0.75, fc='white',linewidth=1.25)
  fig = plt.gcf()
  fig.gca().add_artist(centre_circle)
  plt.show()
```

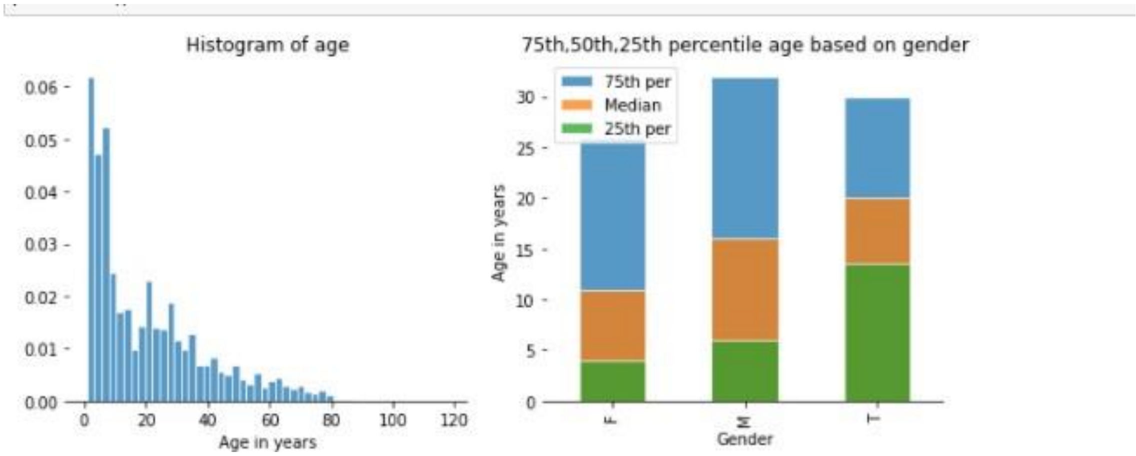


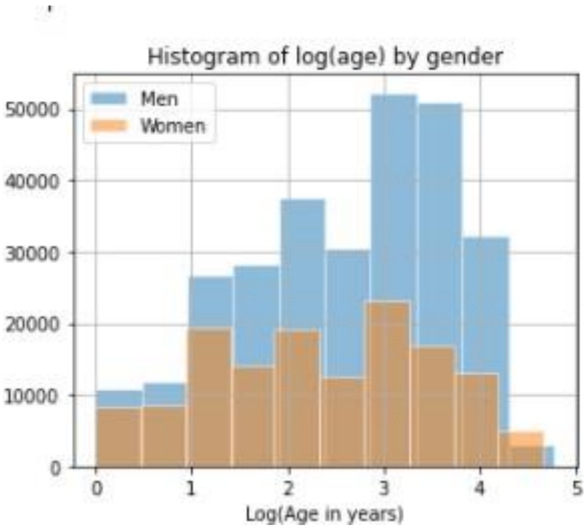
```
: xyz=df['Mobile'].sum()
  cnt=df['State'].count()
  labels=['Yes','No']
  data=[xyz,cnt-xyz]
  fig = plt.figure(figsize=(5, 4))
  plt.pie(data,labels =labels,shadow = True,startangle = 90, autopct='%1.1f%%',colors = ['#00FFFF','#CDC0B0'])
  plt.title("% of mobile registered with aadhar card",fontsize=10,color='navy')
  centre_circle = plt.Circle((0,0),0.75, fc='white',linewidth=1.25)
  fig = plt.gcf()
  fig.gca().add_artist(centre_circle)
  plt.show()
```

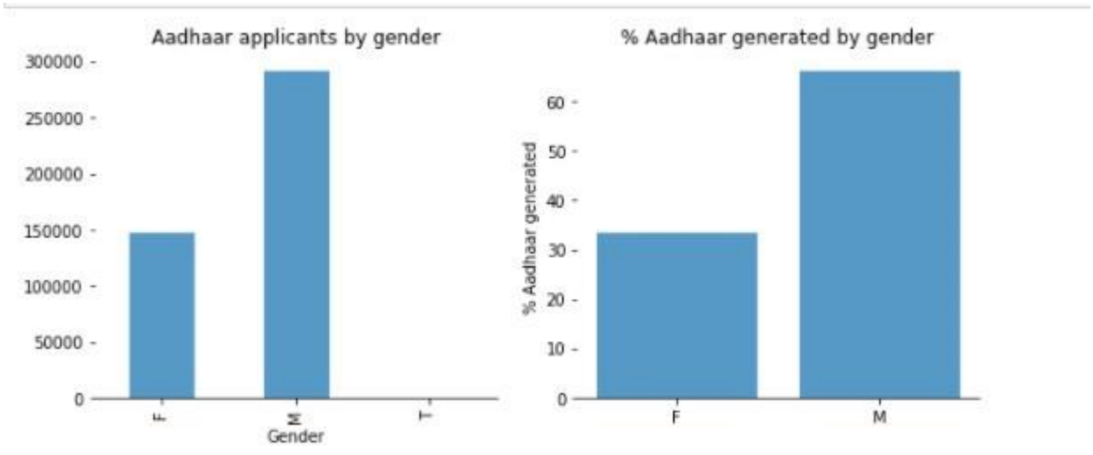
4. SYSTEM TESTING AND RESULT ANALYSIS

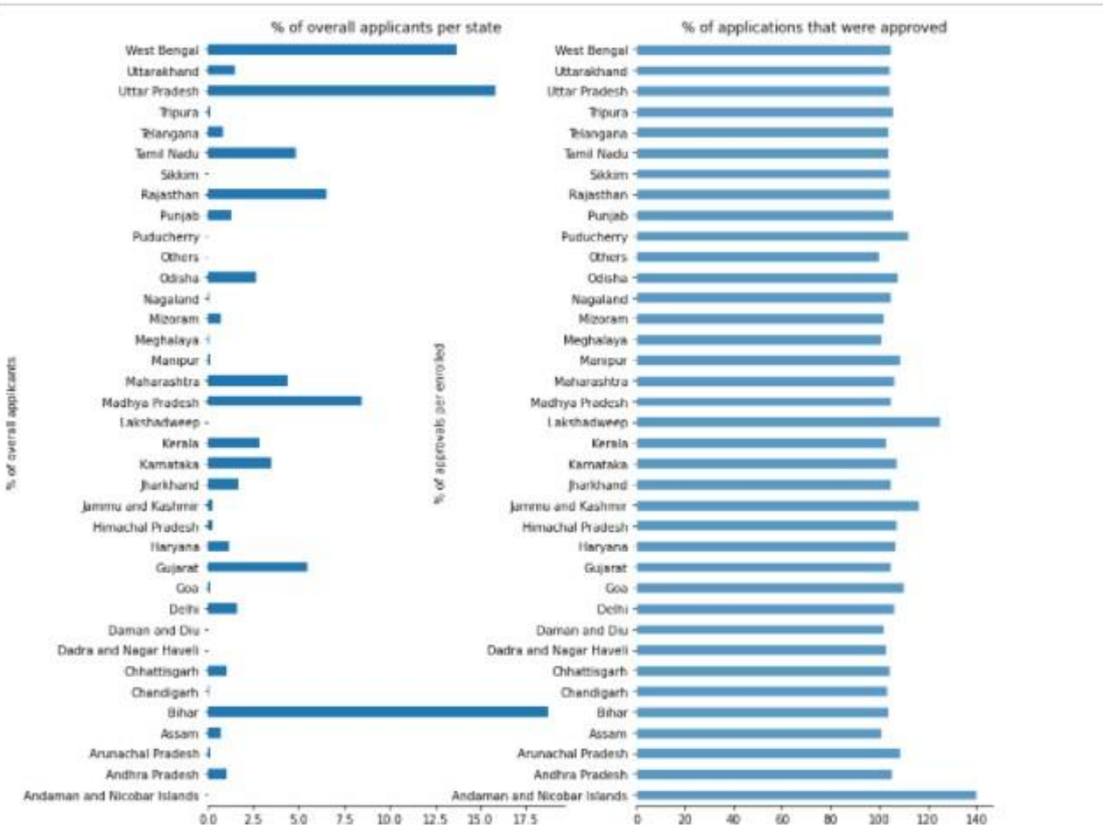
1. Identify the total number of cards approved by gender wise
2. Identify the total number of cards approved by state wise
3. Identify the total number of cards approved by age wise
4. Identify the number of cards rejected by government (State wise)
5. Identify the number of cards approved by government (State wise)
6. Identify the total number of cards linked with mobile

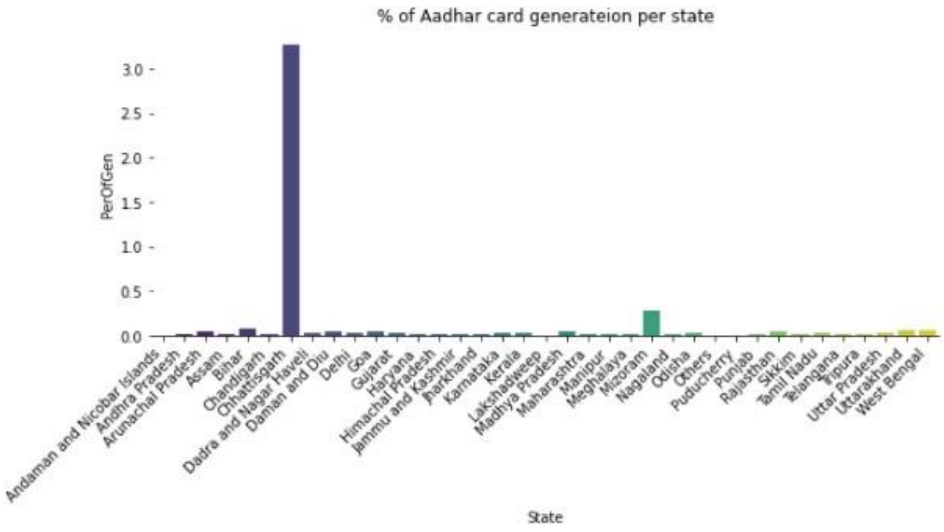
Output of Projects:

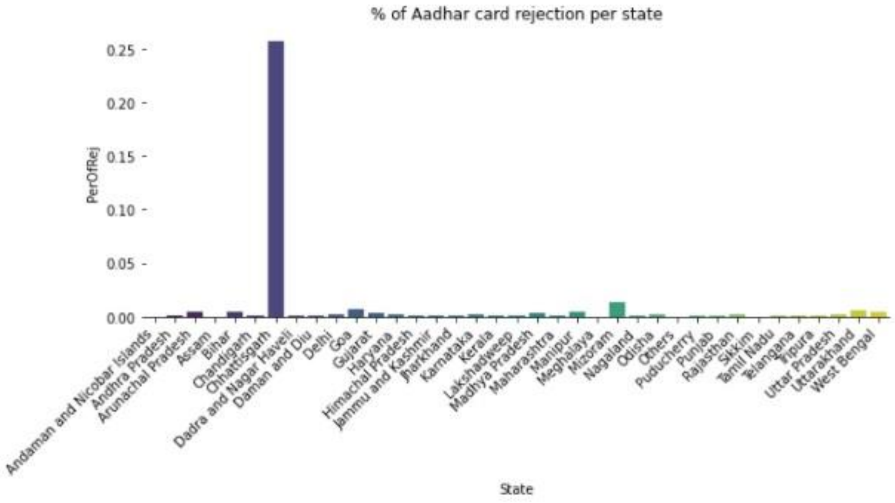


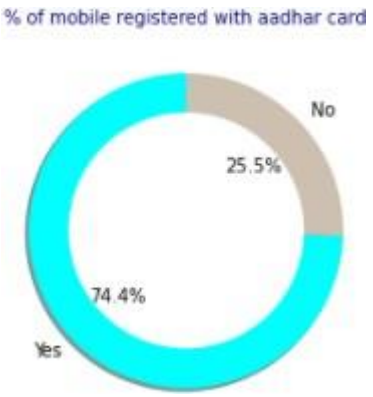


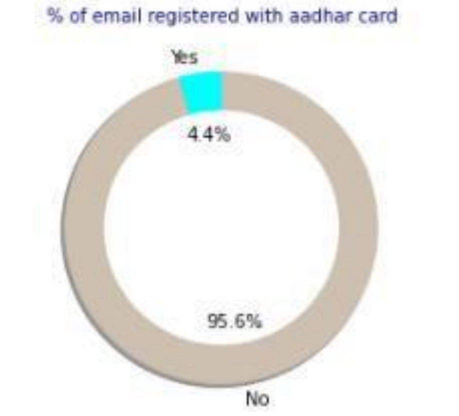

Aadhaar applicants by gender
300000
250000
200000
150000
100000
50000
0
F
M
T
Gender
% Aadhaar generated by gender
% Aadhaar generated
60
50
40
30
20
10
0
F
M


% of overall applicants per state
% of overall applicants
West Bengal
Uttarakhand
Uttar Pradesh
Tripura
Telangana
Tamil Nadu
Sikkim
Rajasthan
Punjab
Puducherry
Others
Odisha
Nagaland
Mizoram
Meghalaya
Manipur
Maharashtra
Madhya Pradesh
Lakshadweep
Kerala
Karnataka
Jharkhand
Jammu and Kashmir
Himachal Pradesh
Haryana
Gujarat
Goa
Delhi
Daman and Diu
Dadra and Nagar Haveli
Chhattisgarh
Chandigarh
Bihar
Assam
Arunachal Pradesh
Andhra Pradesh
Andaman and Nicobar Islands
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
% of applications that were approved
% of approvals per enrolled
0
20
40
60
80
100
120
140


% of Aadhar card generateion per state


% of Aadhar card rejection per state


% of mobile registered with aadhar card


% of email registered with aadhar card

CONCLUSION

In this project we analyzed the aadhaar data set against different queries using hadoop environment for processing and storing the massive data while pig latin language for analyzing and extracting meaningful information. From the first it could be concluded that maximum aadhaar cards are accepted from the states like Andhra Pradesh and Uttar Pradesh while least from Goa. Therefore government has to make people more aware regarding the benefits of aadhaar card so that maximum people can enroll themselves in this policy and can be benefited. While from the second query we could conclude that considerable amount of aadhaar card are rejected therefore government has to take some actions to train there manpower so that correct information of the common masses can be recorded. From third query we interpreted that there is considerably low female population as compared to male population in each state. Therefore government has to take some strict measures against female infanticide and feticides to save girl child. Government also has to come up with certain number of policies, technologies and facilitations for senior citizens so that they also have meaningful opportunities.

REFERENCES

1. Doug Howe, Maria Costanzo, Petra Fey, Takashi Gojobori, Linda Hannick, Winston Hide, David P. Hill, Renate Kania, Mary Schaeffer, Susan St Pierre, Simon Twigger, Owen White and Seung Yon Rhee, “Big data: The future of biocuration”, Nature, international weekly journal of science 455, 47-50,4 September 2008.

2. Clifford Lynch,” Big data: How do your data grow? “,Nature , international weekly journal of science ,455, 28-29

3. Adam Jacobs,”The pathologies of big data”, Communications of the ACM - A Blind Person's Interaction with Technology ,Volume 52 Issue 8, August 2009

4. Min Chen, Shiwen Mao and Yunhao Liu, “Big Data: A Survey”, Springer-Mobile Networks and Applications, April 2014, Volume 19, Issue 2, pp 171-209

5. Yeonhee Lee and Youngseok Lee, “Toward scalable internet traffic measurement and analysis with Hadoop”, ACM SIGCOMM Computer Communication, Volume 43 Issue 1, January 2013