DIGICENS

SHUBHAM MURUDKAR[1], GAURAV PAGAR[2], ROHAN HATODE[3]

[1]*(INFORMATION TECHNOLOGY, PADMABHUSHAN VASANTDADA PATIL PRATISHTHAN'S COLLEGE OF ENGINEERING/ MUMBAI UNIVERSITY, INDIA)*

[2]*(INFORMATION TECHNOLOGY, PADMABHUSHAN VASANTDADA PATIL PRATISHTHAN'S COLLEGE OF ENGINEERING/ MUMBAI UNIVERSITY, INDIA)*

[3]*(INFORMATION TECHNOLOGY, PADMABHUSHAN VASANTDADA PATIL PRATISHTHAN'S COLLEGE OF ENGINEERING/ MUMBAI, UNIVERSITY, INDIA)*

Abstract: Now-a-days the government employees has to come to house and gathered information or details of the particular family on a book or some information sheets and after collecting details they'll pass it to particular BMC office and then proceed on. To resolve such kind of problem we want to create another method to work where it help us to directly pass all those information to the main office instead of passing to local BMC office. So DIGICENS is the solution for this as it is a digitized way of calculating the census or population with the help of modernized technology instead of doing it manually on sheets or papers.

Keywords: Android, Census, Hadoop, Java, Map Reduce, Sqoop.

I. INTRODUCTION

In Digital Census is being proposed with a motive of evaluating hidden patterns from the census data of our country in order to reveal the unemployed ratio, criminal ratio of a particular state. The system involves the concept of big data analysis and visualizing the insights of the data.

The system would evaluate the big data sets of census of India, to generate insights from the data. The aim of the system is to provide valuable insights from the data sets and provide the data to appropriate organizations to increase the employment in a state or country. The system can be used to reveal the buried patterns of information from the crime data set, to generate the crime rate of country. The system might also work for a social cause by evaluating the ratio of people who are non-workers due to their physical disability to do work, this eminent information could inspire various NGO’s to conduct social activities by providing employment to the disabled people.

The system would use Hadoop as its big data framework. Though the analysis can be done using the traditional RDBMS as well, but since the data is big using RDBMS the total time of analysis would be 10 times more than Hadoop. The analysis tool which the system would be using is Hive.

In survey on the existing system and proposed system is discussed forcensus.This discuss the aim of the project, Development of Efficient system to count census. The chapter also highlighted the problem definition of the project along with the solutions to the functional problems in the existing system.

Registrar General and Census Commissioner of India for the 2011 Indian Census. Census data was collected in 16 languages and the training manual was prepared in 18 languages. In 2011, India and Bangladesh also conducted their first-ever joint census of areas along their border.The census was conducted in two phases. The first, the house-listing phase, began on 1 April 2010 and involved collection of data about all the buildings and census houses.Information for the National Population Register was also collected in the first phase. The second, the population enumeration phase, was conducted from 9 – 28 February 2011 all over the country. The eradication of epidemics, the availability of more effective medicines for the treatment of various types of diseases and the improvement in the standard of living were the main reasons for the high decade growth of population in India.

There were 2 Phases taken into consideration:

House Listing
Population Enumeration

Every user was issued a 12-digit unique identification.

The Other Question that arises is why we chosen Android Platform and why Android Is The Most Popular Mobile Operating System in the World Android, The World's Most Popular Mobile Platform, Android powers hundreds of millions of mobile devices in more than 190 countries around the world, Global Partnerships And Large Installed Base, For developers, Android innovation lets you build powerful, differentiated applications that use the latest mobile technologies, Powerful Development Framework , Easily optimize a single binary for phones, tablets, and other devices, Open Marketplace For Distributing Your Apps Google Play is the premier marketplace for selling and distributing Android apps 1.5 billion downloads a month and growing. Get your apps in front of millions of users at Google's scale, You Can Have Your Pick of Phone at Any Price Point.

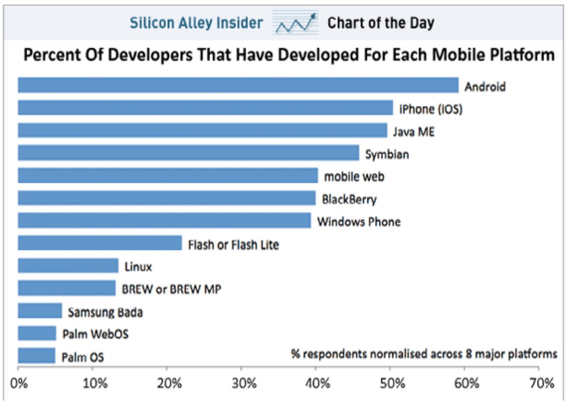


Figure 1.Chart of the day app developers mobile platform

HADOOP

As the system involves census data analysis, the amount of raw census data that would be evaluated for insights will be very large. Thus, the involvement of very large data for analysis urges to the use of Hadoop framework for data analysis.

Though the data analysis can be done with traditional RDBMS system as well but since the data is very large, the time required for analysis would be very long. The use of Hadoop framework for the data analysis incurs many benefits:

- Flexibility: Hadoop Framework is very much convenient for big data analysis since it can compute any format of data, be it (structured, un-structured, multi-structured) data.
- Scalability: Linear processing of the big data. As the system consist of petabytes of data, Hadoop's involvement for data analysis would provide great scalability for the proposed system.
- Reliability: As the Hadoop's file system HDFS involves redundant storage of data it ultimately increases the reliability of the data availability. The HDFS will store 3 copies of the data to increase the system reliability
- Economical: Since the Hadoop framework is completely free it eradicates the software cost.

Thus, the involvement of Hadoop for the proposed system increases the overall efficiency of the system. It provides very prompt results for any type of query.

The proposed system exercises large amount of data thus, the usage of Hadoop is the first contender of the big data analysis.

II. METHODS AND MATERIAL

As the Big Data comes into picture, Hadoop the most preferred method of the big data analysis comes into existence because of its key characteristics of being reliable, flexible, economical, and a scalable solution. The Hadoop provides a very large disk space for storing any amount of data. Using a Hadoop cluster the files are segregated very swiftly, and it is the responsibility of the name node of Hadoop to make the data available. The Hadoop file system provides very less disk time , thus proving to be time efficient. The Hadoop framework is built on google's MapReduce programming structure and is flexible with various data analysis tools.

The data in the HDFS can be analyzedirrespective ofit's structure.

The technology to be used for analysis depends on following factors:

- Structure of the data set. (unstructured, structured or multi-structured)
- Awareness with the technology. For eg. If you are a SQL master using Hive would be very convenient
- Real time processing. Since Hadoop is a batch processing system, using it in a system which requires real time processing is big conundrum. Thus, for real time processing with Hadoop appropriate technology must be used.

The overall process of big data analysis using Hadoop can be termed as an ETL process. ETL refers to (**E**xtract, **T**ransform, **L**oad)

- The extract step of the ETL involves retrieval of the datasets from various data sources. The extract process is the initial step in the ETL process. The extraction of the data involves retrieval of imminent data from sources to carry on the analysis part. For eg, stream data is extracted from solar panels for the predictive analysis of battery level.
- The second step of the ETL process involves transformation of extracted data to a particular format for the analysis. Though the HDFS of Hadoop can store any type of data, it is sometimes difficult to analyze data of different formats. For eg,the stream data received from solar panels are transformed using Java Yuart
- Load the data into HDFS is the third step of ETL. The loading of data simple involves the command as Hadoop fs –copy FromLocal souce destination paragraphs must be indented. All paragraphs must be justified, i.e. both left-justified and right-justified.

We all know Hadoop is a framework which deals with Big Data but unlike any other framework it's not a simple framework, it has its own family for processing different thing which is tied up in one umbrella called as Hadoop Ecosystem. Before jumping directly to members of ecosystem let's have a understanding of classification of data.

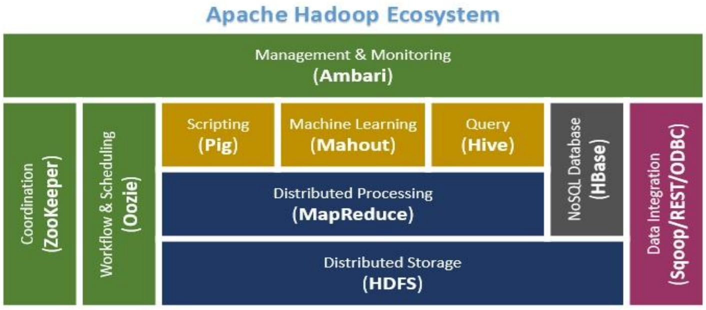

Apache Hadoop Ecosystem
Management & Monitoring (Ambari)
Coordination (ZooKeeper)
Workflow & Scheduling (Oozie)
Scripting (Pig)
Machine Learning (Mahout)
Query (Hive)
Distributed Processing (MapReduce)
Distributed Storage (HDFS)
NoSQL Database (HBase)
Data Integration (Sqoop/REST/ODBC)

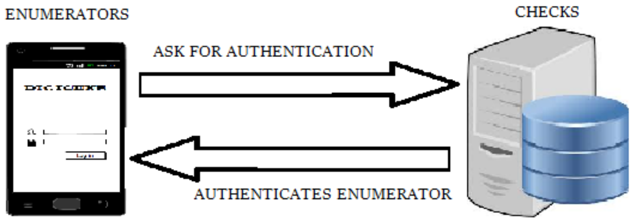
Figure 2. Hadoop Eco-System

Data is mainly categorized in 3 types under Big Data platform.

- **Structured Data**-- Data which has proper structure and which can be easily stored in tabular form in any relational databases like Mysql, Oracle etc is known as structured data.Example- Employee data .
- **Semi-Structured Data**-- Data which has some structure but cannot be saved in a tabular form in relational databases is known as semi structured data. Example-XML data, email messages etc.
- **Unstructured Data**-- Data which is not having any structure and cannot be saved in tabular form of relational databases is known as unstructured data. Example- Video files, Audio files, Text file etc.

III. RESULTS AND DISCUSSION

In Digicens first the data is collected by the enumerator from the user with the help of the android app,which considers of all the information of the user which will be needed for the census data and which are needed by the enumerator for the completion of the form.Once the enumerator has collected the data from the user the enumerator submits the form which includes all the data of the user.When the data is submitted the data is stored in the database and then the data can be processed and functions or conditions can be applied on the data which is done with the help of Hadoop.Hadoop is used to manage the data ,Hadoop smartly manages the data by creating clusters of data which can be easily accessed later on.While retrieving the data an interface has been created using Java which is like a Java application.The data can be easily retrieved by the person who needs to examine census data by using this application.It is easy to access and the person can retrieve the data by applying some filters and can get the desired output within fraction of seconds as Hadoop processes the data very rapidly.



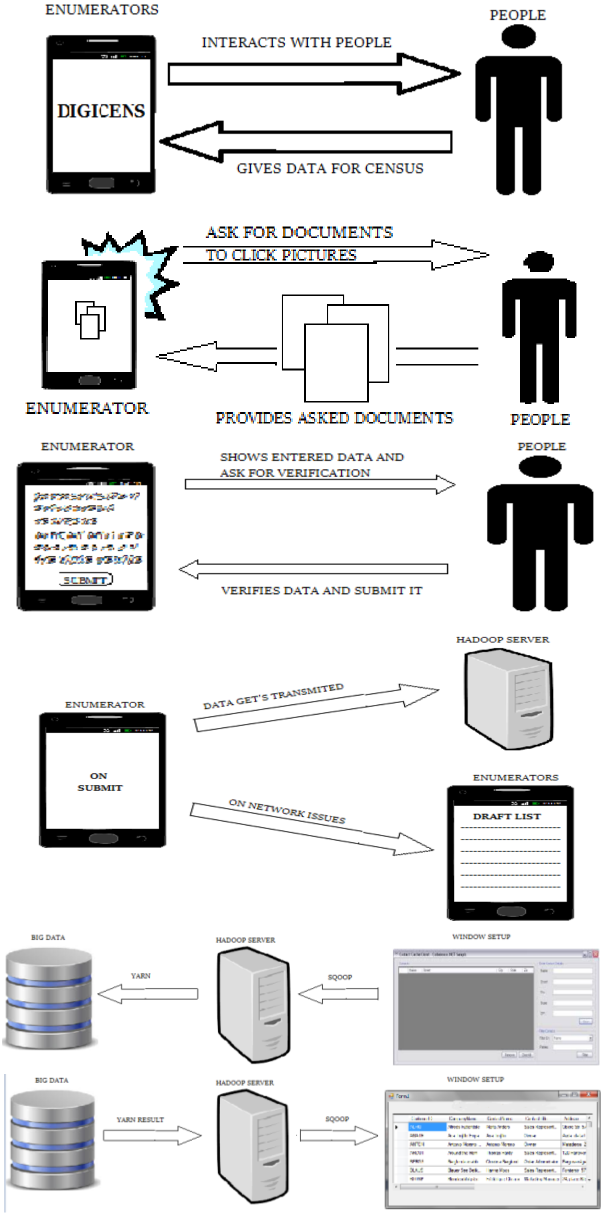


Figure 3.Proposed Working

IV. CONCLUSION

DigiCensus is a software which is been basically designed for the betterment of any locality by providing impeccable insights by big data analysis. By using Hadoop as the data analysis framework and Hive as analysis tool, we generate the imminent insights, which can be used by various organizations, entrepreneurs & various firms. Thus, the system provides a helping hand for any particular organization for establishment of their project by providing them the number of non-workers of a particular area, which is very difficult to find manually.

REFERENCES

[1]. Apache Hadoop http://hadoop.apache.org/.

[2]. https://www.mssqltips.com/sqlservertip/3262/big-data-basics--part-6--related-apache-projects-in-hadoop-ecosystem/

[3]. http://www.census.gov/data/data-tools.html

[4]. http://developer.android.com

[5]. https://www3.ntu.edu.sg/home/ehchua/programming/java/j4a_gui.html