ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 3, Issue 3 , March 2016

Digitizing Indian Census Data for Analytics, Using Big Data Technology

Remya Panicker

Asst. Professor, MET's Institute of Engineering BKC, Nasik, Maharashtra. India

ABSTRACT: The paper presents a method of projecting census data of a nation as Big Data for Analytics purpose. Census Data provides the demographics and details of the geographical, social and economic status of a country. This data are recorded once in a decade and is maintained by government. In India also we have a census data recording since 1872. We can imagine if we combine these data we will get a treasure of information which will be beneficial for the government, policy maker, planning commission to mend and implement beneficial policies and laws for citizens of the country. The paper explores the opportunities that will be created if we digitalized this historical data and use it for analytic purpose. It also discovers some suitable technologies that can be suited to develop such tool.

KEYWORDS: Digital India, Census Data Analysis Tool, Big Data, Census data Analysis, Analytics.

I. INTRODUCTION

The paper presents a conceptual model census data analysis tool (CDAT)[4] which can be used by policy maker and govt. officials to design and frame policies that can be well suited for the masses. This paper describes history of census data collection, how they are actually digitized, how census data can be accumulated , problems of accumulating census data and a short description of creation of a repository for these data, proposed infrastructure for the model and opportunities that can be created by implementing this model.

The Indian Census has a tradition of accumulating nation's data and considered as the best in the world. India is the largest democracies in the world. So Indian census data has a remarkable significance. Census 2011 is the fifteenth census, and seventh after independence. Census data collection has started from 1872. Participation in the Census by the people of India is indeed a true reflection of the national spirit of unity in diversity [1]. The Indian Census is the largest source of a variety of statistical information on different characteristics of the people of India. From more than 130 years, is used as a statistical tool for measuring the country's growth and development. From 1872 when the first census was conducted in India by the British but it was conducted in some parts of the country. Researchers for various fields like demography, economics, anthropology, sociology, statistics and many other disciplines uses the Indian Census as a source of data for analysis and fact finding. The rich diversity of the people of India i.e the data about the population, male female ratio, literacy, income, personal and social status, etc. is collected by the decennial census which has become one of the tools to understand and study India

The Census is conducted by Office of the Registrar General and Census Commissioner, India under Ministry of Home Affairs, Government of India. In India Census was officially set up in 1951 which was on adhoc basis. The Census Act was enacted in 1948 to provide for the scheme of conducting population census with duties and responsibilities of census officers. The systematic collection of statistics of the country was set up in 1949. It was decided that Ministry of Home Affairs under Registrar General and ex-Officio Census Commissioner, India will be responsible for this. This organization was made responsible for generating data on population statistics including Vital Statistics and Census. Later, this office was also entrusted with the responsibility of implementation of Registration of Births and Deaths Act, 1969 in the country [2].

We have tons of data collected in the due course from 1872 to 2011, which will give us untouched and unexplored facts and trends of our nation. Since because of the size of this data we can refer it as a Big Data. This paper presents a terminology that these data that is collected can be analyzed by using sophisticated analysis tools that are available today.

II. WHAT IS BIG DATA TECHNOLOGY?

Big data refers to huge data sets characterized by larger *volumes* (by orders of magnitude) and greater *variety* and complexity, generated at a higher *velocity*. These three key characteristics are sometimes described as the three Vs of big data. Big Data refers to data which has larger sizes more than billion zettabytes. Big Data is about turning

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 3, Issue 3 , March 2016

imperfect, complex, often unstructured data into implementable information. The data can be text, images, RFID codes, satellite images, search engine hits, tweets, facebook tweets etc. Since these data are in wide variety of formats and sizes. It is referred as Big Data.

Big Data Analytics refers to discovering knowledge from the information. Big Data analytics is referred as a collection of tools and methodologies that is used to transform large quantities of raw data into "data about data" – i.e. a useful data which can be used for fining out some knowledge.

Indian census data is also large sized data which will contain all historical data of 130 years along with other survey data and social networking sites' data. So we can refer it as Big Data. The proposed tool will gather all those data and build a ware house which will contain Big Data repository which can be used for Big Data Analytics. The Central store will also contain social networking data since now most of the citizens of the country are active member of internet, so most of the present status, problem and condition of society can also be predicted from these online data. Online data along with census data and various government survey data is gathered under a Big Data infrastructure for analytics. Big Data technology is an accurate solution in building CDAT.

III.ACCUMULATING CENSUS DATA FOR ANALYTICS

Before building the tool the major challenge will be to gather all the data of past censuses. All the census data is under the government control. Now government has shared most of the data to the public for research and analysis purpose. Let see how these data are collected and which technologies are used to process these data. Use of Digital Technologies in census data collection and processing started from 1961, before this data were collected and processed manually. The major and most important part in building up this model is to gather those data and store it. Since we are using census data which is the secure and sensitive data of a nation. Before the accumulation is done we will review how those data are stored and processed and what the challenges in collecting and storing it.

IV.TECHNOLOGIES USED IN CENSUS DATA PROCESSING

The census data was recorded manually before 1961, after 1961 data was captured electronically. Following are the details how technology was used:

A. In 1961 census[13]:

In 1961 "Unit record" machines were used. The Hand Punching machines (inserting one card at a time) using 80 columns (Hollirith) punch cards were used for converting data into machine readable form. The processing was done on the sample (data) selected from the entire data. The data schedules were coded in different parts of India for data entry. Punched cards were duplicated by reproducers. Verifiers and sorting machine was used for data processing. "Serial Rolling Total Tabulator cum printer (SRTT)" was used for printing

B. In 1971 census[13]:

The Key-punching (electrical cum mechanical) machines were used for data entry. The machine used a stack of 80 column (Hollirith) punch cards. An IBM 1401 computer along with IBM card reader and printer was used at the Headquarter at New Delhi for processing. The large size spools of magnetic tape were used for data processing and for keeping backup storage.

C. In 1981 Census[13]:

GCS,ECIL & ICT provided Key to Disk machines for converting information into machine readable form. HP1000, CD-Cyber 730 & NEC-1000 computer systems at NIC, New Delhi and Regional Computer Centre (RCC) was used for processing the data. All the required software (for data validation, editing, processing and tabulation) was developed by the officers of Data Processing Division, ORGI, Headquarters.

D In 1991 census[13] :

During 1991 Census, a major revolution took place in the data processing activities in ORGI. Medha- 930 main frame system at DP Division. For the data entry Unix supported Dump terminals were setup. These dump terminal were connected to servers (data centres) . Data transfer and exchange between them was done using magnetic tapes. Camera ready copies of entire tabulation in Hindi and English was produced for the first time in the history.

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 3, Issue 3 , March 2016

E. In 2001 census[13]:

"Automatic Form Processing Technology using "Intelligent Character Recognition (ICR)" technology was used in this census. The OMR/OCR/ICR technologies were evaluated and ICR was found to be acceptable being the viable solution for census data processing activities. The data centres were setup using 25 High speed heavy duty duplex scanners (Kodak) for scanning data sheets. 45 NT servers were used to set up Servers, 1060 PIII PCs were used for computing terminals. The backup was achieved by using ZIP SLR & DLT drives. First time in the history of census one billion records were scanned. Computer Assisted Coding (CAC) was used to scan certain columns which also avoid manual coding.

F. In Census 2011[13]:

The servers used were HP ProLiant DL380G6 Quad Core Base Servers with ,HP Storage Works EVA6400, Capacity – 10/100 TB .Kodak High Speed Scanners were used for storage. The data centers were set up using Nos Microsoft Windows 2008R2 servers, SQL Server -2008, Windows 7 Professional Clients, High speed heavy duty duplex scanners (Kodak) and backup HP Storage Works EVA6400, Capacity – 10/100 TB. The Intelligent Character Recognition (ICR) used eFlow4.5 software. CAC was used for scanning and data file creation automatically.

V. SOURCES FOR COLLECTING CENSUS DATA

Older Data were collected manually. But government had released those data in various formats on internet from official websites like, http://www.indiastat.com, http://www.censusindia.gov.in, http://**data**.gov.in *www.archives.gov*, http://india.gov.in, *etc.* Government has also started an open government data initiative in which it has made census data public which can be downloaded by any one for analysis. A workstation has been set up by government where we can access census data for research purpose. This is at JNU New Convention Centre, JNU Campus, Delhi. Here census tables from 1991 to 2011 Censuses are made open to researchers. Researchers have to visit this center in order to use those data. Prior to digitization, these records can be accessed via microfilm publication, National Archives Microfilm Publication (NAMP) M595, Indian Census Rolls, 1885-1940 (692 rolls) contains census rolls that were usually submitted each year by agents or superintendents in charge of Indian reservations, to the Commissioner of Indian Affairs, as required by an act of July 4, 1884 (23 Stat. 98)., which is done by NARA(**National Archives and Records Administration**) is an independent agency of the United States government charged with preserving and documenting government and historical records and with increasing public access to those documents, which comprise the National Archives.

VI. CHALLENGES FACED IN ACCUMULATING CENSUS DATA

A. DIFFERENT DATA STORAGE PLATFORMS AND FORMATS

As from the above history we can see that data is recorded and stored using different technique and formats. The problem here will be bringing all those data under one platform and same schemas. Building schemas will also come up with a new problem of variety of data. Though census data of last 30 years have been digitized all the data base schemas are in different structures. The data of last three decades are digitized and made available public on web. These data are in the form of books, reports, maps, tables, maps, audio clips, PowerPoint presentations etc. The forms are in various computer file formats, viz. JPG, PDF, BMP, MP3, CSV, PPT, WAV etc. Most of data is in pdf formats which contain tabular information, graphs, histogram, pie charts etc.

The data after the independence period can be obtained in any format for the above mentioned sources. The records before independence period 1872-1951 are available in microfiche forms of 4458 reels. The information from microfiche can be digitized using optical scanner and can be converted in charge-coupled device (CCD) array. International Institute for Population Sciences (IIPS) had undertook a project to create a digital library for the census data from 1872 to 1951. Information from this library is also in PDF formats.

The challenge here is to convert all the data set into a uniform repository based on some predefined data structure.

B. VARIATIONS IN INFORMATION COLLECTED

In every census data collected for the country varies accordingly, like name, age, marital status, religion etc. this parameters changes in every census. The table below shows the variation in the attributes of the data that were collected during census. So when repository is created at that time we have to normalize such attributes. In every census the

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 3, Issue 3 , March 2016

question asked varies. The number of questions canvassed during Population Enumeration in Censuses from 1951 to 2011 is listed below:

Sr. No	Census Year	No. of Question Asked
1	1951 Census	14 Questions
2	1961 Census	13 Questions
3	1971 Census	17 Questions
4	1981 Census	16 Questions
5	1991 Census	21 Questions
6	2001 Census	23 Questions
7	2011 Census	29 Questions

Table 1: No. of Questions Canvassed During Census [14]

The challenge over here will be while extraction and creation of centralized schemas we will have some incomplete data which were not monitored during earlier censuses. Schema creation will be very difficult task if uniform repository has to be created. Relevant and Irrelevant fields may appear while the data collections.

C. LIMITATION OF ACCESS TO ENTIRE CENSUS INFORMATION (RAW DATA).

The data that are made public is in aggregated form, raw data i.e. each and every record of each individual is not revealed in those data. We get only the summary, statistical graphs, charts and maps for the census data. Raw data are very crucial when fact have to be mined. The raw data contains each and every record of every citizen of the nation. Most of the data available is not sufficient enough as they are summarized data. For example we can get the per capita income of a particular village but we will not get the information about the per capita income of a particular person for the data exposed so far. As the confidentiality of such data is responsibility of government. Government Permission and consent has to taken to get an access to those data. It is obvious since it is very sensitive data and had to be secured.

VII. TACKLING THE CHALLENGES

The data that is made public is in aggregated form, raw data i.e. each and every record of each individual is not revealed in those data. We get only the summary, statistical graphs, charts and tables as a source of information. These data has to be converted into processable data or flat files. One of the easiest ways can be converting all data into flat files suitable for analysis. We deploy pdf to word convertor for doing this. Another mechanism is using Image convertor. Since Big Data Infrastructure is used for this we can accommodate data in various different formats. For the variation in the data collected we can apply some data preprocessing activities like Data Cleaning to fill out incomplete and inconsistent data. Clustering, Missing Value elimination techniques like Nullification, Assigning a Constant values, Range Approximation , Co-relation analysis etc can be employed for normalization of Variant data. For overcoming the problem of restricted data access we can take special permission for government agencies for granting access to those data for research activities under various government policies for development of innovative technologies.

VIII. INFRASTRUCTURE FOR DATA ACCUMULATION

Open sources technologies can be incorporated to build a census data analysis tool. Fig. 1 shows the block diagram of the proposed model. Hadoop Infrastructure will be best suited to build up such application. All the modules in Hadoop are designed with a fundamental assumption that hardware failures (of individual machines or racks of machines) are common and thus should be automatically handled in software by the framework. Apache Hadoop's MapReduce[5] and HDFS components originally derived respectively from Google's MapReduce and Google File System (GFS). Map Reduce Frameworks are often based upon Hadoop and Hadoop-like technologies. They work by providing parallel processing capabilities that move subsets of the data to distributed servers. The primary use is processing massive amounts of data in a scalable manner. To build a data warehouse we can use a HBASE[12] to populate the warehouse. Online Analytical Storage can be used for some analytical processing. Populating database will require NoSQL technologies. Hadoop uses inexpensive servers; it contains name node which contains metadata and data note which will contain the data. The data is stored and managed using HDFS, which provides a facility where

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 3, Issue 3 , March 2016

data is stored in files, files are divided into uniform blocks and stored across clusters nodes. HBASE provides storage for Hadoop Distributed Computing Environment. Data is logically organized into tables, rows and columns in HBASE. We can store the census data as clusters on data nodes and whose links will be maintained by name node. The clustering of the data collected will be maintained on distributed environment, using HDFS. The clusters will also contain the data collected from the social networking sites and internet which will be refined using MAPREDUCE[5], NoSQL and HIVE will be used for retrieving data from the store. HBASE can also used to store the data where rows and columns data will have its significance.

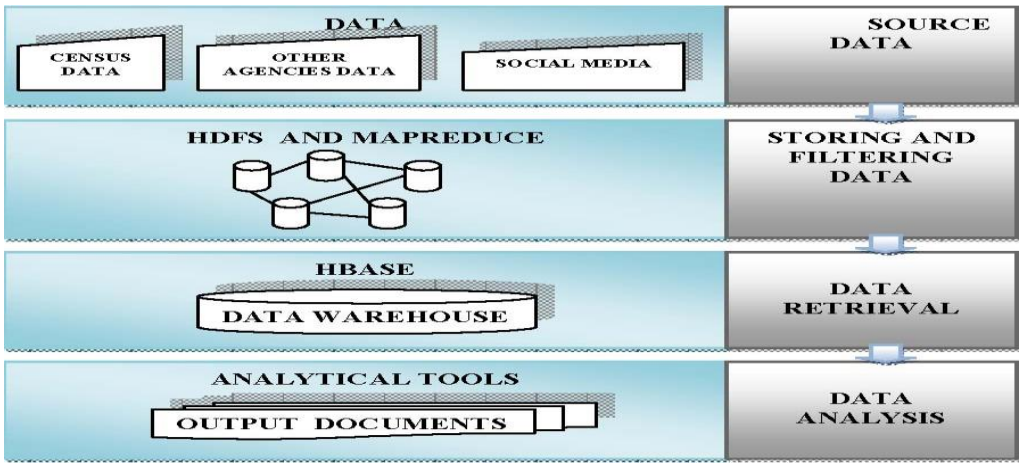

Fig 1: CDAT Model [4]

IX. CONCLUSION

If the entire census data along with the day by the day data generated on the internet are collected then we will a vast store of knowledge. This treasure of knowledge will help planning commission, policy makers, law makers, scheme drafters etc to have an insight of the actual need of the country, some trends, present status, and prevalent problems. We can also employ this in healthcare, disaster management, space research etc. For e.g. if government wants to implement some law related female foeticide, we can get the trend of lower sex ratio in particular parts of the country, so laws can be drafted according to social and cultural background of those areas. Satellite Images of the areas where some natural disaster had occurred can give some parameters that may occurs before such disaster. So when we mine such historical data we can predict some natural calamities. We can find out the need of the nation sate wise, region wise, social class wise, culture wise, religion wise, gender wise, age wise etc which will be very use full in mending policies and laws.

REFERENCES

[1] “Census of India 2011” source: Drop in Article http://www.censusindia.gov.in
[2] Web Source: http://www.censusindia.gov.in/2011- Common/aboutus.html
[3] “Big Data Infographic and Gartner 2012 Top 10 Strategic Tech Trends.” Business Analytics 3.0 (blog) (November 11, 2011). .
[4] Remya Panicker, “Value Creation in Government Management Models and Practices Through Census Data Analysis using Big Data In India” Conf. IRCTET-14 Chandwad .
[5] Jeffrey Dean and Sanjay Ghemawat ,”MapReduce: Simplified Data Processing on Large Clusters” *GoogleResearch Paper*
[6] Paul, M.J. and M. Dredze. *You Are What You Tweet: Analyzing Twitter for Public Health*. Rep. Center for Language and Speech Processing at Johns Hopkins University, 2011.
[7] Zhuo Feng,Pritam Gundecha,Huan Liu,"Recovering Information Recipients in Social Media via Provenance"
[8] Revolution analytics white paper "Advanced ‘Big Data’ Analytics with R and Hadoop"
[9] White Paper: “Using Cloudera to Improve data Processing” *Cloudera, Inc 1-888-789-1488 or 1-650-362-0488*
[10] White Paper:” Oracle NoSQL Database” An Oracle White Paper September 2011
[11] Revolution analytics white paper "Advanced ‘Big Data’ Analytics with R and Hadoop
[12] Shoji Nishimura,Sudipto Das, Divyakant Agrawa, Amr El Abbadi,”HBase: A Scalable Multi-dimensional Data Infrastructure for Location Aware Services”
[13]Web source: “ http://censusindia.gov.in/eval/data_processing_division.pdf taken on 26/08/2015.
[14] Web Soruce:http://censusindia.gov.in/Ad_Campaign/drop_in_articles/05-History_of_Census_in_India.pdf taken on 26/08/2015