Enhancement of Governance through Data Analytics: Mapreduced mining Census data for policy making

[1]Prof. Smita Khot, [2]Pranav Ghildiyal, [3]Prasanna Dubey, [4]Mitali Patil, [5]Pooja Matere
[1,2,3,4,5]Department of Computer Engineering,
[1,2,3,4,5] Dr. D. Y. Patil Institute of Technology, Pune, India

Abstract: The advancements of technology have increased awareness about what government is doing among the citizens of the country. Any action taken by government makes way for both positive and negative reactions among people. While positive reactions are something to cheer about for the ruling people, negative ones can be signs of threats. In order to become more citizen friendly, the government can analyze the citizen's data that it possess to decide the track for policy making, in a way that it brings as much as positive reviews from people and as less as possible of negative ones. Mining/Analytics provides the best possible platform for this.

Index Terms **– Census Data, Map Reduce, Data Analytics, Knowledge Discovery**

I. INTRODUCTION

The data about the citizens of a country is known as the Census Data. The process of this collection i.e., census is done usually every 10 years. This census process allows the government to have not only count of population of the country but also gives possession of data, carrying huge amount of knowledge, the only need is to discover this knowledge. We implement this Analytics mechanism using the MapReduce algorithm. MapReduce is a very famous algorithm for analyzing data. The MapReduce algorithm can be applied for different variety of fields present in the data, giving variety of outputs, but still this analytics would be half complete without proper representation of this. We implement this presentation part using Python's matplotlib library package. The MapReduce can be triggered by making dynamically choice of fields from the python GUI application, and the result of this operation is processed in order to be given to matplotlib package's function in order to be graphically represented in the form of bar graphs, plots, pie charts etc.

II.PROBLEM STATEMENT

To help the government increase the positive review by making use of analyzing of the census data that it has, the target of this system is to do this task for the government in doing this analysis and hence, it turn using the result of this system to improve the policy structure thereby serving the required purpose.

III. LITERATURE REVIEW

Census can provide the fundamental population data of the whole nation. The census data is rich with hidden knowledge which can be used to assess the current condition of the nation and its citizens. Most of the national policies are constituted straightly based on the population status, hence, this huge amount of data can be used to develop effective policies for the citizens of the country, to improve their life and hence it turn speed up the nations progress i.e., to provide services for country's social and economic development. According to them classification, is one of the important Data Mining techniques, when it comes to Census or Government Data Mining [1].

Data mining can extract implicit, previously unknown, and potentially useful information from the data. They say that Census is a significant investigation of national condition and national power, strongly believe that Data Mining in census data has very high learning value. When it comes to Mining Big Data, it might not be always possible to mine detailed data, hence though details is lost by generalization, but the generalized data may be more meaningful and easier to interpret. Hence by generalizing the Big Data, concept hierarchies can be made, which can be mined to extract useful information [2].

They highlight the impact of big data on current scene of the society. They say that the term big data occurs more frequently now than ever before. A large number of fields and subjects, ranging from everyday life to traditional research fields involve big data problems. Big data incorporates endless amount of information. In many industries, it is growing, providing a means to improve and streamline business. Big data has changed the world in terms of predicting customer behavior. The actual challenge of big data is not in collecting it, but in managing it as well as making sense of it. While working on big data, it is crucial to determine whether the benefits outweigh the costs of storage and maintenance. They also review recent research in data types, storage models, privacy, data security, analysis methods, and applications related to network big data [3].

Big Data concerns with large-volume, complex, growing data sets with multiple, autonomous sources. They believe that big

data can prove out to be very useful, only if we can harness the data, to extract the hidden knowledge that it contains. They propose a theorem by the name of HACE theorem (Big Data starts with large-volume, **H**eterogeneous, **A**utonomous sources with distributed and decentralized control, and seeks to explore **C**omplex and **E**volving relationships among data) [4].

According to them, the growing popularity and development of data mining technologies bring serious threat to the security of individual's sensitive information. An emerging topic in data mining, known as privacy preserving data mining (PPDM), has been extensively studied in recent years. The basic idea of PPDM is to modify the data in such a way so as to perform data mining algorithms effectively without compromising the security of sensitive information contained in the data. Current studies of PPDM mainly focus on how to reduce the privacy risk brought by data mining operations, while in fact, unwanted disclosure of sensitive information may also happen in the process of data collecting, data publishing, and information (i.e., the data mining results) delivering. They identify four different types of users involved in data mining applications, namely, data provider, data collector, data miner, and decision maker. And discuss privacy concerns and the methods that can be adopted to protect sensitive information for each type of user [5].

Taking reference of Educational Data Mining, they tell us that though Data Mining of Educational data provides new insights for a better education system. However, sharing or analysis of educational data introduces privacy risks for the data subjects, mostly students, this holds true, that too in even greater proportion, when it comes to census data and mining of census data. Many initiatives and regulations protect personal data privacy in domains such as health, commerce, communications and most regulations do not enforce absolute confidentiality which would cause more harm than good but rather protect individually identifiable data that can be traced back to an individual with or without external knowledge. This gives rise to a wide range of studies primarily focusing on de-identifying private data with as little harm to its information content as possible, in an attempt to preserve both the privacy and usefulness of the data, should be considered with respect to various scenarios. Research on data privacy has formally defined and enforced privacy primarily in two scenarios: (1) Sharing data with third parties without violating the privacy of those individuals whose (potentially) sensitive information is in the data. This is often called privacy-preserving data publishing. (2) Mining data without abusing the individually identifiable and sensitive information within. This is often called privacy-preserving data mining or disclosure control [6].

Present us with Incremental Association rule mining approach. They tell that lot of study has been done in the context of preserving privacy of raw data and sensitive data, but has focused on one time mining. As per them, No work till now, has been published on incrementally mining association rules with privacy protection when the data upon which mining occurs is quantitative and subject to change. They study this problem against the backdrop of supply chain management. They, taking reference of this supply chain management, present uses of Discrete Wavelet Trans- form (DWT) to mask the original data in such a way that a majority of the original association rules are preserved [7].

According to them, it might not be always possible to store Big Data at a single place. Hence the data storage has to be distributed. Privacy being another crucial factor along with security restricts sharing or centralization of data. Privacy-preserving data mining has emerged as an effective method to solve this problem. Though distributed solutions have been proposed that can preserve privacy while still enabling data mining. However, while perturbation based solutions do not provide stringent privacy, cryptographic solutions are too inefficient and infeasible to enable truly large scale analytics to face the era of big data. Previous work on random decision trees (RDT) show that it is possible to generate equivalent and accurate models with much smaller cost, which can be exploited the fact that RDTs can naturally fit into a parallel and fully distributed architecture, and develop protocols to implement privacy-preserving RDTs that enable general and efficient distributed privacy-preserving knowledge discovery [8].

Propose Pincer-Search as an efficient Algorithm for Discovering the Maximum Frequent Set, from data. Knowledge Discovery of data can be defined as the non trivial process of identifying valid, potentially useful, and ultimately understandable patterns in data. Mining of frequent Set can discover trends hidden in the data. They tell that typical algorithm for finding the frequent set, operate in bottom-up, breadth-first fashion. The computation starts from frequent 1-itemsets at the bottom and then extends one level up in every pass until all maximal (length) frequent item sets are discovered. The efficiency of such algorithms decreases when any of the maximal frequent set becomes longer. In data mining applications the maximum frequent item sets could be long. To solve this problem, they present a novel Pincer-Search algorithm, which searches for Maximum Frequent Set from both bottom-up and top-down directions. They say that it performs well even when the maximal frequent item sets are long [9].

Big Data is characterized by high volumes of data, volumes huge enough that it becomes difficult to manage it by using existing data management concepts and tools. Today every organization is trying to move towards big data and investing towards it, the main reason that why organizations do big data processing or why organizations are getting attracted to habit called big data, is the enormous potential it carries with itself. It gives organization the power to analyze huge amounts of data they possess, they can posses, they want to possess, and use it for competitive edge in business arena, against their competitors. Map reduce is one of the go to tool when it comes to big data processing Reasons behind which is the popularity that MapReduce enjoys due to its unique features which includes simplicity and communicative manners of its programming model as MapReduce has mainly two functions map( ) and reduce( ) even though a large number of data analytical tasks can be expressed as a set of MapReduce jobs.[10]

IV. MATHEMATICAL MODEL

Let S be the system.

S= {}
Identify I as input.
I= {a, b, c}
where
a -> Census Data
b -> Current Job and Salary
c -> Education Details
S = {I}
Identify P as processes
P = {p}
where
p-> Map reduce as per user request
S = {I, P}
Identify O as output.
O= {o}
where
o -> Pictorial Data Representation as Bar Graphs, Pie Charts etc as applicable
S = {I, P, O}
Identify A as case of success.
A= {s}
where
s -> Requested information displayed pictorially Bar Graphs, Pie Charts.
S = {I, P, O, A}
Identify F as case of failure.
F= {k, l}
where
k -> Improper data preprocessing
l -> Data connection error.
S = {I, P, O, A, F}

V. SYSTEM OVERVIEW

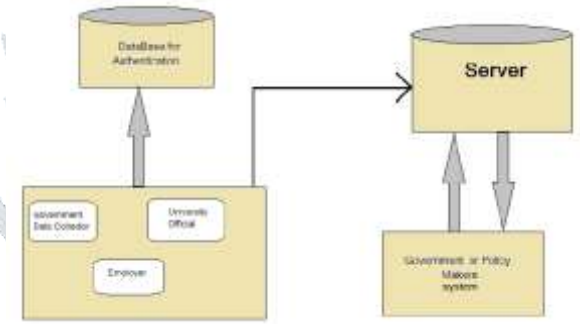

Figure 1: System Overview

Figure 1 shows the overview of the system. The data is stored in the database (MongoDB) which comes majorly from census officials, who enter data for each and every individual citizen of the country, the other two mini sources are as applicable and as and when required type of sources, as the university (official) is responsible for entry of educational data only for those who are passing out from their institution as and when, similarly, the employer (official) only look after the required data's entry only for their own employees.. The data stored in the database will be utilized, by the government through a python GUI application, which through the use of matplotlib package of python will display the results of query graphically. This graphical representation can be used to understand the scene of the citizens of the country, the state in which the country is and hence decisions related to policies can be taken based upon it.

VI. RESULT AND DISCUSSIONS

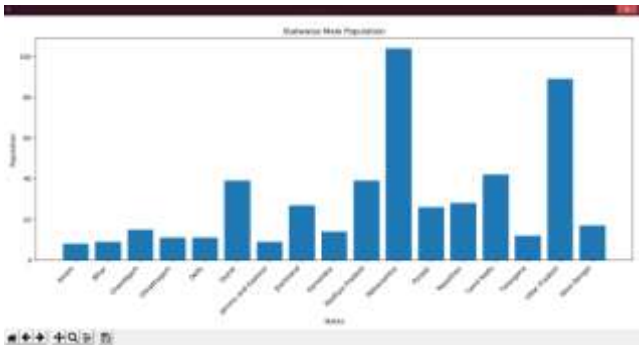

Figure 2: Result - I

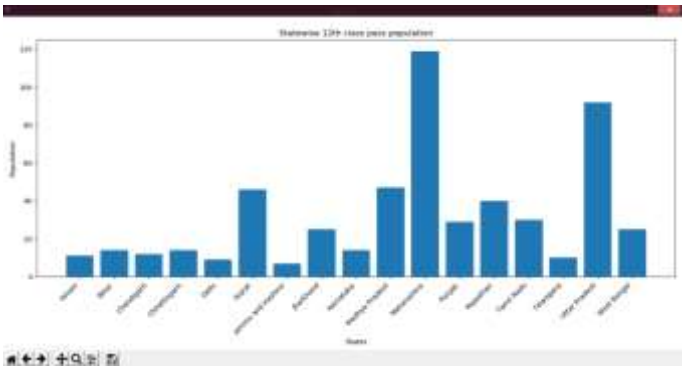

Figure 3: Result – II

Figure 2 and Figure 3 show two sample results generated by the system, while Figure 2 shows the State wise male population distribution Figure 3 shows the State wise distribution for 12th pass people based on the sample data considered for testing the system. Like these two results the system is capable of generating more such results covering various areas like population distribution (General overall, gender wise and so on), Education distribution, Occupation distribution, caste-wise population, education distribution, religion-wise population education distribution and many more such query terms. These and more such results provide government with multi-dimensional view of the status report of the citizen-wise scenario of the country, which in turn will help them understand what needs to be done, for the citizens, improvements possible, areas with scope of improvements, strength areas of the country, hence helping it take more effective actions.

VII. ADVANTAGES

The advantage of this system is that it will help government make use of the data it possess, in understanding the state of people and make use of it in improving existing as well as designing new policies for the benefit of the citizens of the country.

VIII. APPLICATION

The system can be used by the Government for understanding the current state of the people for better citizen centric policy making.

IX. CONCLUSION

With progressing technology and increasing awareness, the governments ruling the country has pressure of performing well, making good beneficial policies for the citizens. Hence, the government can make use of the census data available with it, for the same. An effective and efficient way for which, is the system described in this paper.

REFERENCES

[1] Bing Sheng and Sun Ghengxin, Data Mining in Census Data with CART, 3rd International conference on Advanced Computer Theory and Engineering (ICACTE), pp. V3-260- V-264, 2013

[2] Sheng Bin and Gengxin Sun, The preprocessing in census with Concept Hierarchy, 2nd International Conference on Computer Engineering and Technology, Volume 1, pp V1-535 V1-538, 2010

[3] Zhihan Lv, Houbing Song, Pablo Basanta-Val, Anthony Steed, Minho Jo, Next Generation of Big Data Analytics : State of Art, Challenges, and Future Research Topics, IEEE Transactions on Industrial Informatics, Volume 13, Issue 4, pp 1891-1899, 2014

[4] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding, Data Mining with Big Data, IEEE Transactions on knowledge and data Engineering, Volume 26, Issue 1, pp 97-107, January 2014

[5] Lei Xu, Chunxiao Jhang, Jian Wang, Han Yung, and Yong Ren, Information Security in Big Data: Privacy and Data Mining, IEEE Access, 2014

[6] Mehmet Emre Gursoy, Ali Inan, Mehmet Ercan Nergiz and Yucel Saygin, Privacy Preserving Learning Analytics : Challenges and Techniques, IEEE Transactions on Learning technologies, Volume 10, Issue 1, pp 68-81, 2017

[7] Madhu V Ahluwalia, Aryya Gangopadhyay, Zhlyuan Chen and Yelena Yesha, Target Based Privacy Preserving and Incremental Association Rule Mining, IEEE Transactions on Services Computing, Volume 10, Issue 4, pp 633-645, 2017.

[8] Jaideep Vaidya and Wei Fan, A random decision tree framework for privacy preserving Data Mining, IEEE transactions on dependable and secure computing, Volume 11, No. 5, pp 399-411, Sept/Oct 2014

[9] Dao-l Lin and Zvi M. Kedem, Pincer-Search: An efficient Algorithm for discovering the maximum frequent set, IEEE Transactions on Knowledge and data engineering, Volume 14, No 3, pp 553-566, MAY/JUNE 2002

10] Shweta Pandey, Dr.Vrinda Tokekar, Prominence of MapReduce in BIG DATA Processing, 4th International Conference on Communication Systems and Network Technologies (ICCSNT), pp.555-560, 2014.