Proc. of Int. Conf. on Emerging Trends in Engineering & Technology, IETET

Big Data Alliance with MapR Technology to Govern Indian Election Panorama

Gagandeep Jagdev[1] and Amandeep Kaur[2]
[1]Dept. of Computer Science, Punjabi University Guru Kashi College, Damdama Sahib (PB).
Email: drgagan137@pbi.ac.in
[2]Research Scholar (Ph.D.), Dept. of Computer Applications, Guru Kashi University
Talwandi Sabo, Punjab

***Abstract*—Data is not a new term in the field of computer science, but Big Data is essentially a new word. When data grows beyond the capacity of currently existing database tools, it begins to be referred as Big Data. Big Data posses a grand challenge for both data analytics and database. The central theme of this research paper is concerned with handling huge amount of data that is concerned with different formats of elections that are been contested in India. We have initially limited my database only to Punjab state. It is no more a hidden fact that in future elections will be fought on the basis of statistics and figures and not on the basis of caste and religion. We have created a structured database which includes thirteen different attributes providing information related to different candidates who have contested MP elections in different districts of Punjab state. These attributes are candidates name, political party to which they belong, there assets, there liabilities, criminal cases registered on them, number of votes obtained, percentage of votes obtained, chances of winning next elections etc. In this research paper we will discuss Apache Hadoop framework which makes use of Map-Reduce technology. The objective behind this research paper is to assist common electorates of Punjab state to take best decision on the basis of previous track record of politician or political party and decide who to vote for to get better governance.**

***Index Terms*— Big Data, Big Data analytics, elections, Hadoop framework, Map-Reduce.**

I. INTRODUCTION

Internet is the major source which has resulted in the tsunami of data in the past few years. Big data is too big, it moves too fast, and doesn't fit the structures of our existing database architectures. It is like an ocean of data in which we people swim in every day with an effort to come on the surface, but every day the level of data increases tremendously. Gone are the days when memory was used to be measured in Gigabytes or Terabytes or Petabytes, today it is measured in exabytes, zettabytes or yottabytes. With Big Data solutions, organizations can dive into all data and gain valuable insights that were previously unimaginable. The term "big data" can be pretty nebulous, in the same way that the term "cloud" covers diverse technologies. Utilizing big data requires transforming information infrastructure into a more flexible, distributed, and open environment [1, 2]. Big data promises deeper insights that data scientists are highly involved in exploring this

Grenze ID: 02.IETET.2016.5.10

data in such a manner that organizations are benefited to its best with total customer satisfaction. Big data analytics is one of the great new frontiers of IT. Emerging technologies such as the Hadoop framework and MapReduce offer new and exciting ways to process and transform big data—defined as complex, unstructured, or large amounts of data—into meaningful insights, but also require IT to deploy infrastructure differently to support the distributed processing requirements and real-time demands of big data analytics [3, 4].

II. Issues Related With Big Data Characteristics

Heterogeneity, scale, timeliness, complexity, and privacy problems with Big Data impede progress at all phases of the pipeline that can create value from data. The problems start right away during data acquisition, when the data tsunami requires us to make decisions, currently in an ad hoc manner, about what data to keep and what to discard, and how to store what we keep reliably with the right metadata. Much data today is not natively in structured format; for example, tweets and blogs are weakly structured pieces of text, while images and video are structured for storage and display, but not for semantic content and search: transforming such content into a structured format for later analysis is a major challenge. The value of data explodes when it can be linked with other data, thus data integration is a major creator of value. Since most data is directly generated in digital format today, we have the opportunity and the challenge both to influence the creation to facilitate later linkage and to automatically link previously created data. Data analysis, organization, retrieval, and modeling are other foundational challenges. Data analysis is a clear bottleneck in many applications, both due to lack of scalability of the underlying algorithms and due to the complexity of the data that needs to be analyzed. Finally, presentation of the results and its interpretation by non-technical domain experts is crucial to extracting actionable knowledge [6, 7, 12].

III. Five Phases Of Big Data

Big data processing involves five different phases [2, 6, 7].

A. Data Acquisition and Recording

Big data definitely have some source of origin. It is not created from a vacuum. Different scientific experiments being carried out in the world today produces petabytes of data per day. Much of this data is of no use and has to be filtered out. The first challenge faced is to set filtering parameters as such that useful data doesn't gets discarded. For example, suppose one sensor reading differs substantially from the rest: it is likely to be due to the sensor being faulty, but how can we be sure that it is not an artifact that deserves attention? We need research in the science of data reduction that can intelligently process this raw data to a size that its users can handle while not missing the needle in the haystack. The second challenge encountered is related to automatically generating right metadata to illustrate what data is recorded, how it is recorded and measured.

B. Information Extraction and Cleaning

It is mention able here that information collected is not in an analysis ready format. For example, consider the collection of electronic health records in a hospital, comprising transcribed dictations from several physicians, structured data from sensors and measurements, and image data such as x-rays. The data in this format cannot be effectively analyzed. An information extraction process should be applied to such data to pull out the required information from the sources under consideration and present it in a structured format suitable for analysis. This is really a big challenge. This data may include images and videos and such extraction is highly application dependent.

C. Data Integration, Aggregation, and Representation

It is not enough to merely collect record and throw the data into a repository. If we have large data sets in repository, then it will be almost impossible for the user to find the desired data when required. But with sufficient amount of metadata there is some hope but still challenges persists due to differences in experimental details and in data record structure. Data challenging is much more than simply locating, identifying, understanding and citing data. All this process needs to occur in a complete automated manner for an effective large scale analysis. Suitable database design is most important. There are many different ways in which data can be stored. Certain designs will be better than others for certain purposes and possibly

may carry drawbacks for other purposes. Therefore it can be concluded that database design is an art and needs to be carefully executed by trained professionals.

D. Query Processing, Data Modeling, and Analysis Methods for querying and mining

There is no doubt in the fact that big data is diverse, imprecise and unstructured. Even then big data is of much value as compared to small individual observations as general statistics obtained from large sample are more precise. When it comes to mining, it requires clean and efficiently accessible data. Provision should be there for declarative query and mining interfaces. Efficient mining algorithms and computing environments is another important requirement.

E. Interpretation

The analysis of big data remains of no value if users are not able to understand the analysis concept. Decision maker is provided with the result of analysis and is expected to interpret these results. This interpretation requires efforts. It involves deeply examining all the assumptions made and retracing the analysis. There are several sources of errors like system may carry bugs and conclusions may be based on error prone data. No responsible user will yield authority to computer system for all this. Instead one will try to understand and verify the results produced by computer system. All this should be made easy by computer system and this is a big challenge with big data due to its complexity.

IV. Elections Scenario In India

One method for predicting the results of upcoming elections is via exit poll. The most valuable information regarding campaigns and their affect on general public is provided by citizens themselves. Data analysts develop models based on this information and perform predictions regarding winning and losing chances of any political party and any political leader. If such results are properly harnessed, they could gain sizeable gains. Elections in India have always comprised issues based on caste, religion, sentiments, traditional wisdom, opinion polls and rallies. But 2014 Lok Sabha elections witnessed the use of technology to its very best by political parties. All this idea was actually borrowed by the way Barack Obama contested his elections in America and raise to power in 2008 and 2012.

In an extraordinary attempt to engage digitally literate electorates of India, Google and some other social platforms started a forceful digital information campaign. Google India launched one such hub related to elections where electorates can search for political candidates, political parties, and election platforms and voting related information in their regions. They even launched one site on the counting date which updated about live status of results on the day of counting. It was revealed that Narendra Modi consistently topped the search trends when compared to other candidates. In India elections have been always influenced by matters related to caste, religion, minorities, majorities, sentiments, election rallies etc. But it was for the first time during 2014 Lok Sabha election that the presence of technology and its importance in contesting election was felt. The concept was very smartly copied by BJP from the way President Barack Obama used technology and came to power in 2008 and 2012. The year 2014in India was the largest display of democracy undergoing elections on the planet earth. As much as 543 Parliamentary and 4120 assembly constituencies were set up. Election Commission of India brought up 9 lakh 30 thousand polling booths all over the country with objective of conducting fair elections. India being a very diverse country has many languages which officially vary from state to state. For e.g. in Punjab, Punjabi is an official language, but in neighboring state of Haryana the official language is Hindi and down south it changes to Malayalam, Telugu etc. So in all voter rolls were prepared in 12 different languages constructing 9 lakh pdf files and when their hard copy were created, it consumed approximately 2.5 crore pages. The biggest hurdle was to decipher these 2.5 crore pages into English to fuse them appropriately with other sources.

The citizens of India are unlike in every way possible which comprises religious faith, beliefs, languages and sentiments. The candidates selected by the electorates are based on many factors which can be broadly classified into direct factors and indirect factors. Direct factors include region specific policies; work done for the region in past and other local polarizations. On the other hand, indirect factors include factors like geography, financial stability, penetration of media via television, mobile etc and climate. It is a fact that large proportion on Indian electorates (about 30 percent) is still undergoing through the confusion of who to vote for. Such people are often targeted by familial, social or political influencers. This 30 percent also include those electorates who just don't vote because either they forget or they just don't care to vote and

have a strong belief that nothing can be changed via their single vote. The political parties target electorates by dividing them into three categories: new voters; influenced voters and core group voters which involve voters who always vote for one particular party. Each group is targeted in a different manner after constructing appropriate strategies. If political party can find out those electorates who are definitely going to vote, then there is no need for the party to waste time in influencing voters residing in regions who are not interested in voting process. This would help political parties to do canvassing in selected regions rather than general canvassing. This would also assist them in targeting key influencers and side with them. Political parties should lay stress on combining the "want to's" with the "have to's". Now question is how this can be accomplished by political parties?

When an individual is to be judged, it is good to know about his/her reading interest. For e.g. Suppose an individual lives in Punjab and have subscribed "Ajit" newspaper or lives in the west and have subscribed "Samna", this clearly displays individuals enthusiasm towards a certain political party. In same manner, if one is a woman and has subscribed magazines related to national and international affairs, she is more likely to vote than woman subscribing magazines related to housekeeping. Political parties actively analyze the public views on social media and if they find any person influential having healthy social circle, such person can be targeted by them [12]. Further if political parties can find out what a voter of a specific region wants to listen, the party can target them with specific messages and key words. It is usual practice performed by parties to advertise themselves on social media, radio and television. People watch television daily and if one finds something really nice in advertisement of a political party, he/she is bound to listen it. Once the advertisement becomes common among public, they start tweeting and writing comments on social media websites. It is from here that political party can come to know the reaction among the public regarding that particular advertisement. If the reaction is on a positive side, this would encourage the political party and they would further work on it. But on the other hand, if the majority of the public feedback is on negative side, the parties can rollback the advertisement by giving any suitable excuse. Political parties also keep an eye on selecting the best spokesperson for canvassing for the party. This spokesperson should be such who is best received by the target segment. This spokesperson can be any popular personality from local movies, media and sports. But it selection of spokesperson generates a negative wave among the voters, party can replace him/her after analyzing the chatter it generates in the society. For e.g. recent announcement of Salman Khan as India's ambassador at Rio Olympics 2016, has been condemned by many reputed sports person of India. Similarly, voters can draw out comparative scorecard of different party candidates and what development have they done in past, all automatically updated via different social channels. Such scorecard can assist voters in deciding which candidate is actually better.

All this is related to mining patterns of interest. It sounds like a huge Big Data problem waiting to be handled. It could turn into a massive data gathering where unique databases will be integrated and explored to find appropriate patterns and correlations. These patterns can be applied on previous election data and voters turnout to come up with appropriate prediction models. It has been explored that approximately 160 million Indian voters who are confused and can't decide who to vote for, can be targeted via different means of media. These people are waiting for a right message to reach their ears which is hidden somewhere waiting to be uncovered. So, it can be concluded that big data analytics could act as a key to reveal the winning mantra which could assure a political party their major win [8, 11].

V. Implementation Via Apache Hadoop Framework And Map-Reduce Technology

Hadoop [9, 11, 12] is a java based framework that is efficient for processing large data sets in a distributed computing environment. Hadoop is sponsored by Apache Software Foundation. The creator of Hadoop was Doug Cutting and he named the framework after his child's stuffed toy elephant. Applications are made run on systems with thousands of nodes making use of thousands of terabytes via Hadoop. Distributed file system in Hadoop facilitates fast data transfer among nodes and allows continuous operations of the system even if node failure occurs. This concept lowers the risk of disastrous system failure even if multiple nodes become inoperative. The inspiration behind working of Hadoop is Google's Map reduce which is a software framework in which application under consideration is broken down into number of small parts [5, 6]. Hadoop is a framework which comprised of several components mentioned as under.

- HDFS – HDFS are distributed cages where all animals live i.e. where data resides in a distributed format.
- Apache HBase – It is a smart and large database.

- Zookeeper- Zookeeper is the person responsible for managing animals play.
- Pig – Pig allows playing with data from HDFS cages.
- Hive- Hive allows data analysts play with HDFS and makes use of SQL.
- HCatalog helps to upload the database file and automatically create table for the user.

The Apache Hadoop software [10] library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. The working of Map Reduce technology is shown in Fig. 1.

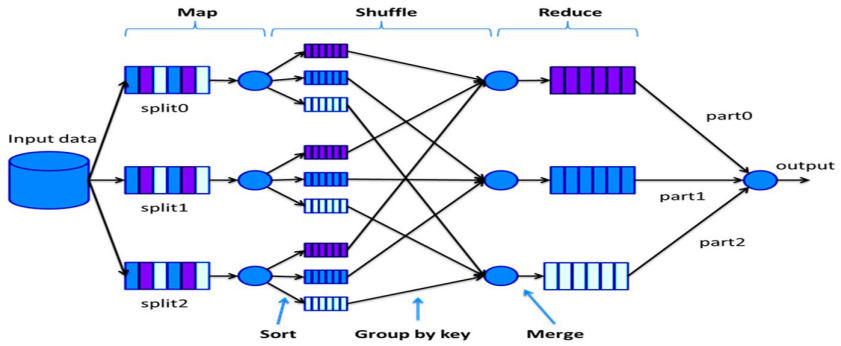

Fig. 1 Working of Map Reduce Technology

This algorithm allows splitting of a single computation task to multiple nodes or computers for distributed processing. As a single task can be broken down into multiple subparts, each handled by a separate node, the number of nodes determines the processing power of the system. There are various commercial and open-source technologies that implement the MapReduce algorithm [5] as a part of their internal architecture. A popular implementation of MapReduce is the Apache Hadoop, which is used for data processing in a distributed computing environment. As MapReduce is an algorithm, it can be written in any programming language.

The initial part of the algorithm is used to split and 'map' the sub tasks to computing nodes. The 'reduce' part takes the results of individual computations and combines them to get the final result. In the MapReduce algorithm, the mapping function reads the input data and generates a set of intermediate records for the computation. These intermediate records generated by the map function take the form of a (key, data) pair. As a part of mapping function, these records are distributed to different computing nodes using a hashing function. Individual nodes then perform the computing operation and return the results to the reduce function. The reduce function collects the individual results of the computation to generate a final output [14, 15].

Let's consider an example to understand the concept of working of Map-Reduce technique. Suppose we have four different files having names and salaries of employees of the company. Each file further consists of four different records. Our aim here is to find out the employee with the maximum salary in the entire database under consideration.

File A:	File B:	File C:	File D:
Surendra, 26000	Gagan, 50000	Surendra, 26000	Surendra, 26000
Kulwant, 25000	Kulwant, 25000	Kumar, 45000	Kulwant, 25000
Sukhdeep, 15000	Sukhdeep, 15000	Sukhdeep, 15000	Manpreet, 45000
Ramandeep, 10000	Ramandeep, 10000	Ramandeep, 1000	Ramandeep, 10000

The Map phase – In map phase, the job is to enclose each record in form of key-value pair as <k, v> : <emp name, salary>.

File A:	File B:	File C:	File D:
<Surendra, 26000>	<Gagan, 50000>	<Surendra, 26000>	<Surendra, 26000>
<Kulwant, 25000>	<Kulwant, 25000>	<Kumar, 45000>	<Kulwant, 25000>
<Sukhdeep, 15000>	<Sukhdeep, 15000>	<Sukhdeep, 15000>	<Manpreet, 45000>
<Ramandeep, 10000>	<Ramandeep, 10000>	<Ramandeep, 1000>	<Ramandeep, 10000>

The combiner phase accepts the output of map phase as input. Here code can be written to find out maximum salaried person from each file.

```
<k: employee name, v: salary>
```

```
Max= the salary of a first employee. Treated as max salary
if(v(second employee).salary > Max)
{
   Max = v(salary);
}
else
{
   Continue checking;
}
```

The expected result is as follows –

A - <Sukhdeep, 26000> B- <Gagan, 50000> C- <Kumar, 45000> D-<Manpreet, 45000>

Reducer phase – The same code is reused to find the maximum salaried person among four < k, v> pairs generated from four different files and the final result is obtained.

<Gagan, 50000>

The self-created database has been created by the authors to monitor the election scenario of past and present in Punjab state. The glimpse of the database has been shown in Fig. 2. One can write different queries, scripts and codes to get appropriate results from this database when uploaded on Apache Hadoop Framework. Map-Reduce technology works in the background and comes up with required results which can be obtained in the form of tables and visualizations as shown in Fig. 3.

	NAME	AGE	EDUCATION	SEX	PARTY	PARTY TYPE	VOTES IN FAVOUR	% OF VOTES IN FAVOUR	CRIMINAL CASE	ASSETS	LIABILITIES	STATUS	WINNING CHANCES
2	Harsimrat Kaur Badal	47	10th	Female	SAD	State	514727	46.09	0	100 Crore	41 crore	Crorepati	Bright
3	Manpreet Singh S/O Gurdas Singh	52	Graduate	Male	INC	National	495332	44.35	0	42 Crore	4 Crore	Crorepati	Bright
4	Jasraj Singh Longia	40	Graduate	Male	AAP	National	87901	7.87	0	31 Lac	Nil	Lakhpati	Average
5	Kuldeep Singh	32	10th	Male	BSP	National	13732	1.22	0	80 Lac	Nil	Lakhpati	Poor
6	Ashish	25	10th	Male	IND	Independent	6626	0.59	3	30 Thousand	Nil		Poor
7	Bhagwant Singh Samaon	36	8th	Male	CPI	National	5984	0.53	0	2 Lac	15 Thousand	Lakhpati	Poor
8	Satish Arora	51	12th	Male	IND	Independent	5936	0.52	0	1 Crore	Nil	Crorepati	Poor
9	Manpreet Singh S/O Gurdev Singh	33	10th	Male	IND	Independent	4618	0.413	0	7 Lac	Nil	Lakhpati	Poor
10	Shaminder Singh	34	12th	Male	JKKNPP	National	4610	0.412	0	33 Lac	2 Lac	Lakhpati	Poor
11	Geeta Rani	50	5th	Female	ABSR	National	4380	0.39	0	Nil	Nil		Poor
12	Sanjeev Kumar Thapar	34	10th	Male	IND	Independent	2833	0.253	0	26 Lac	2 Lac	Lakhpati	Poor
13	Makhan Lal	43	8th	Male	BSP(A)	National	2587	0.231	0	91 Thousand	Nil		Poor
14	Vijay Kumar	44	10th	Male	IND	Independent	2258	0.202	2	6 Lac	Nil	Lakhpati	Poor
15	Swaran Singh	45	Graduate	Male	IND	Independent	2077	0.185	0	1 Crore	Nil	Crorepati	Poor
16	Rajinder Singh	34	10th	Male	SAD(A)	State	1960	0.175	0	26 Lac	Nil	Lakhpati	Poor
17	Navneet	42	12th	Male	SHS	National	1893	0.169	0	57 Thousand	5 Lac		Poor
18	Jagdeep Singh Gehri	39	10th	Male	IND	Independent	1665	0.149	0	1 Lac	Nil	Lakhpati	Poor
19	Bhupesh Kumar	36	Graduate	Male	IND	Independent	1487	0.133	0	6 Lac	Nil	Lakhpati	Poor
20	Gurdeep Singh	44	Illiterate	Male	IKL	National	1404	0.125	0	5 Lac	Nil	Lakhpati	Poor
21	Gurmeet Singh Ranghreta	44	12th	Male	PLP	National	1386	0.124	1	35 Lac+	1 Lac	Lakhpati	Poor
22	Sushil Kumar Jindal	31	10th	Male	IND	Independent	1330	0.119	0	1 Lac	Nil	Lakhpati	Poor
23	Jagdish Rai Sharma	57		Male	NBDP	National	1248	0.111	1				Poor
24	Sukhwinder Singh	33	12th	Male	BMP	National	1086	0.097	0	2 Lac	Nil	Lakhpati	Poor
25	Dyal Chand	52	8th	Male	IND	Independent	955	0.085	0	58 Thousand	Nil		Poor

Fig. 2 Structured database of different political leaders and parties who have been contesting elections in Punjab

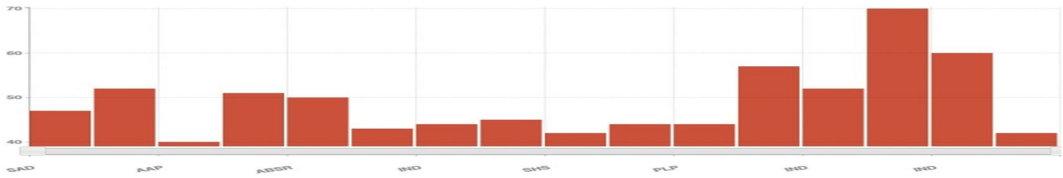

Fig. 3 Results obtained from Hadoop framework in form of bar graphs

VI. Conclusion

The next elections may be path breaker in the way it's fought. It could turn into a massive data gathering work out where unique databases (for e.g. voter registration, social media, subscription data, transaction profile, mobile records, television viewership and channel bouquet, work profile, location, etc.) will be integrated together and analyzed with eagerness to find correlations and patterns. It has been analyzed that about 160 million of those who are not sure about who to vote could be targeted through mobile phones and about a 100 million through television. These people are waiting to hear the right message to make that choice of which party to vote for and may be the right message is hidden somewhere waiting to be

uncovered. So, it can be concluded that big data analytics could act as a key to reveal the winning mantra which could get a political party their major win [8].

REFERENCES

[1] Laney, Doug. 2012. "3D Data Management: Controlling Data Volume, Velocity and Variety."
[2] Information Week. 2012. "Big Data Widens Analytic Talent Gap." Information Week April.
[3] Heudecker, Nick. 2013. "Hype Cycle for Big Data." Gartner G00252431
[4] Edala, Seshu. 2012. "Big Data Analytics: Not Just for Big Business Anymore." Forbes.
[5] Dean, Jeffery, and Ghemawat Sanjay. 2004. "MapReduce: Simplified Data Processing on Large Clusters." Google.
[6] Kaisler, S., Armour, F., Espinosa, J. A., & Money, W. (2013). Big Data: Issues and Challenges Moving Forward. International Conference on System Sciences (pp. 995-1004). Hawaii:IEEE Computer Soceity.
[7] Katal, A., Wazid, M., & Goudar, R. H. (2013). Big Data: Issues, Challenges, Tools and Good Practices. IEEE, 404-409.
[8] Gagandeep Jagdev et. al.," Scrutinizing Elections Strategies by Political Parties via Mining Big Data for Ensuring Big Win in Indian Subcontinent", 4th Edition of International Conference on Wireless Networks and Embedded Systems.
[9] http://hadoopilluminated.com/hadoop_illuminated/Intro_To_Hadoop.html#d1575e686
[10] http://hadoop.apache.org/docs/r1.2.1/ hdfs_design.html
[11] http://searchdatamanagement.techtarget.com/definition/MPP-database-massively-parallel-processing-database
[12] http://www.slideshare.net/rupenmomaya/big-data-insights-challenges
[13] http://www.salient.com/docs/books/SALIENT_MPP.pdf
[14] Dr. Gagandeep Jagdev et. al., "Big Data commence a new Trend for Political Parties to Contest Elections in Indian Subcontinent" at National Conference FPIIT-2015 at D.A.V. College, Abohar, Punjab.
[15] Dr. Gagandeep Jagdev et. al., "Big Data proposes an innovative concept for contesting elections in Indian subcontinent", IJSTA Volume 1, Issue 3, pp. 23-28, 2015, ISSN No. 2454-1532.