International Journal of Scientific and Technical Advancements

ISSN: 2454-1532

Big Data Proposes an Innovative Concept for Contesting Elections in Indian Subcontinent

Gagandeep Jagdev[1], Bhalwinder Singh[2], Mahabli Mann[3]

[1]Dept. of Computer Science, Punjabi University Guru Kashi College, Damdama Sahib, Bathinda (PB), India

[2]Research Scholar, M. Phil (Comp. Sc.), Punjabi University, Patiala (PB), India

[3]Research Scholar, M. Phil. (Comp. App.), Guru Kashi University, Talwandi Sabo (PB), India

Email address: [1]gagans137@yahoo.co.in, [2]bhalwindergori@gmail.com, [3]mahablisingh12@gmail.com

***Abstract*—**Big data refers to the data sets that are too big to be handled using the existing database management tools and are emerging in many important applications, such as Internet search, business informatics, social networks, social media, genomics, and meteorology. In simple words it can be said that any data which challenges the currently existing techniques for handling data can be referred as big data. Big data presents a grand challenge for database and data analytics research. The central theme of my research work is to explore the use of big data concepts in the process of conducting elections in such a manner that it would help the political parties in canvassing and targeting voters and on the other hand it would help electorates to appoint an efficient representative from their constituency. Now elections would not be fought on the basis of caste and religion, but on the basis of numbers, quants and stats. Political parties have to concentrate on the use of technology much more than other matters. The data related to elections is huge, especially when these elections are conducted at national level in country like India which has a large and diverse population. In this research paper, we would discuss about the Hadoop framework, techniques capable of handling big data, challenges faced by big data and also highlight about the beginning of the new era in the Indian politics which makes use of big data and how BJP (Bhartiya Janta Party) made use of big data in Lok Sabha Elections 2014 and gained total majority in Lok Sabha. Election Commission of India (ECI) also made use of techniques associated with big data in mining the huge data of 118 million people of India for conducting effective and healthy elections.

Keywords— Big data; electoral; ECI (Election commission of India); map reduce; MPP (Massive parallel processing).

I. INTRODUCTION

Billions of Internet users and machine-to-machine connections are causing a tsunami of data growth. Big data [1], [2] is data that exceeds the processing capacity of conventional database systems. Big data is too big, it moves too fast, and doesn't fit the structures of our existing database architectures. Big Data is the ocean of information we swim in every day. Vast zeta bytes of data flowing from our computers, mobile devices, and machine sensors form the source of big data. 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. This data is "big data". Utilizing big data requires transforming information infrastructure into a more flexible, distributed, and open environment [3], [6].

Big data analytics is one of the great new frontiers of IT. Data is exploding so fast and the promise of deeper insights is so compelling that IT managers are highly motivated to turn big data into an asset they can manage and exploit for their organizations. 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.

II. HADOOP AND ITS ARCHITECTURE

Hadoop 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. Any of these fragments can run on any node in the cluster. The components involved in Hadoop ecosystem are Hadoop kernel, Map Reduce, Hadoop distributed file system and many related projects like Apache hive, HBase and Zookeeper. The use of Hadoop framework is done by major players like Yahoo, IBM and Google. Ideal operating systems for Hadoop are Windows and Linux, but it can also work with BSD and OS X. The technology involved was developed by Google during their earlier days in order to index all valuable textural and structural information collected by them. All this was done to provide meaningful result to the user. Later this innovation of google was integrated into Nutch which was an open source project and Hadoop was spun off from it. Yahoo was the most prominent in developing Hadoop for enterprise applications. Hadoop platform is preferred to solve problems where data is enormous and does not fit appropriately in tables. It is used to run analytics that involves extensive computation. There are numerous sectors where Hadoop finds it application. It is used in finance to perform exact portfolio evaluation and risk analysis. In online retail, it helps in providing better answers to the customers to increase the probability of their buying the things.

Hadoop runs on large number of machines that do not share any memory space. It is similar to buying a whole bunch of commodity servers, slap them in a rack and run Hadoop software on each one of them. Suppose one decides to load all the data of their organization in Hadoop, the software busts that data into pieces and spread them across different servers.

Gagandeep Jagdev, Bhalwinder Singh and Mahabli Mann, "Big data proposes an innovative concept for contesting elections in Indian subcontinent", *Journal of Scientific and Technical Advancements*, Volume 1, Issue 3, pp. 23-28, 2015.

ISSN: 2454-1532

Hadoop keeps track of where the whole data resides. Also mentionable is that if data on server goes offline, it can be replicated automatically from a known good copy. In centralized database systems one big disk that is connected to four, eight or sixteen processors. But in comparison, each of these servers has two, four or eight CPUs in Hadoop. You can run your indexing job by sending your code to each of the dozens of servers in your cluster, and each server operates on its own little piece of the data. The result is then delivered back collectively. This is what is known as map - reduce. Map reduce technique involves mapping the operations out to all these servers and thereafter these results are reduced back into a single result set [10].

As Map Reduce is an algorithm, it can be written in any programming language. Hadoop map reduce works in three stages:

- *First Stage*: *Mapping*: In this stage, a list of elements is provided to a 'mapper' function to get it transferred into pairs. The mapper function does not modify the input data, but simply returns a new output list.
- *Intermediate stages*: *Shuffling and Sorting*: After the mapping stage, the program exchanges the intermediate outputs from the mapping stage to different 'reducers'. This process is called shuffling.
- *Final Stage*: *Reducing*: In the final reducing stage, an instance of a user-provided code is called for each key in the partition assigned to a reducer. In particular, we have one output file per executed reduce task.

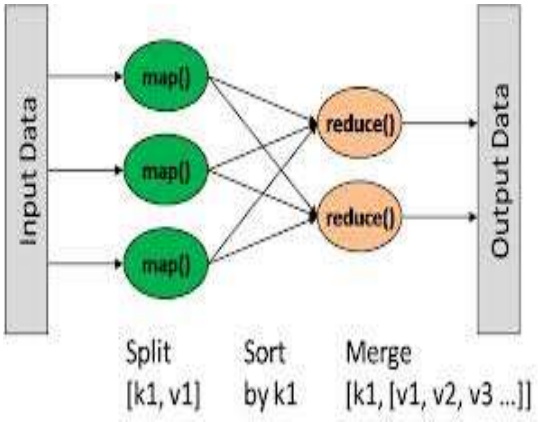

Fig.1. Working of Map Reduce.

III. Features of HDFS and Map Reduce

A. Hadoop Distributed File System

HDFS is a robust and fault tolerant distributed file system aimed to turn a cluster of industry standard servers into a massively scalable pool of storage. Developed specifically for large-scale data processing workloads where scalability, flexibility and throughput are critical, HDFS accepts data in any format regardless of schema, optimizes for high bandwidth streaming, and scales to proven deployments of 100PB and beyond.

- *Scalability* – Servers can be added later on to increase capacity
- *Availability* - Serve mission-critical workflows and applications
- *Fault Tolerance* – It refers to automatic recovery from failures. If a workload is running on a system and some parts of the system stops performing their task, in such a situation the other parts of the system should configure themselves to share the work of the failed parts. This means that the service does not fail even in the face of some component failures.
- *Flexible Access* – Multiple and open frameworks for serialization and file system mounts
- *Load Balancing* – It refers to placing of data intelligently for maximum efficiency and utilization. Many systems related to big data takes un-curated data. It means there are always data points that are extreme outliers and introduces hotspots in the system. The workload in such systems is not uniform. Some small parts are major hotspots and bear high load as compared to rest of the system. Such distribution of load is to be taken care of.
- *Tunable Replication* - Multiple copies of each file provide data protection and computational performance
- *Security* –Enhanced security for users and groups

IV. Phases Involved in Big Data

Big data processing involves five different phases [13], [5].

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. In scientific experiments, considerable detail regarding specific experimental conditions and procedures may be required to be able to interpret the results correctly, and it is important that such metadata be recorded with observational data.

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.

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.

Gagandeep Jagdev, Bhalwinder Singh and Mahabli Mann, "Big data proposes an innovative concept for contesting elections in Indian subcontinent", *Journal of Scientific and Technical Advancements*, Volume 1, Issue 3, pp. 23-28, 2015.

ISSN: 2454-1532

Query Processing, Data Modeling, and Analysis Methods for Querying and Mining – There is no doubt in the fact that big data is noisy, dynamic, diverse, inter-related and untrustworthy. But even noisy Big Data could be more valuable than tiny samples because general statistics obtained from frequent patterns and correlation analysis usually overpower individual fluctuations and often disclose more reliable hidden patterns and knowledge. Further, interconnected Big Data forms large heterogeneous information networks, with which information redundancy can be explored to compensate for missing data, to crosscheck conflicting cases, to validate trustworthy relationships, to disclose inherent clusters, and to uncover hidden relationships and models. Mining requires integrated, cleaned, trustworthy, and efficiently accessible data, declarative query and mining interfaces, scalable mining algorithms, and big-data computing environments. At the same time, data mining itself can also be used to help improve the quality and trustworthiness of the data, understand its semantics, and provide intelligent querying functions. A problem with current Big Data analysis is the lack of coordination between database systems, which host the data and provide SQL querying, with analytics packages that perform various forms of non-SQL processing, such as data mining and statistical analyses.

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.

V. Challenges Faced by Big Data

Heterogeneity and Incompleteness

A great amount of heterogeneity is tolerated when humans consume information. But in case of machine analysis algorithms homogeneous data is accepted. The first step in data analysis is that data should be carefully structured. Many data analysis systems require greater structure. However, less structured design is more effective for many. However the computer system works most efficiently if they are capable of storing multiple items that are identical in size and structure. Even after cleaning data and correcting errors, some errors are likely to reside behind in the data. Performing this task with perfection is a big challenge.

Scale

The first thing on hearing the name big data what comes into the mind is size. Managing enormous amount of data has been a challenging task since decades. In the past these challenges were met by processors getting faster. But today data volume is scaling much faster as compared to computer resources and CPU speeds. During the past five years, the processor technology has under gone a dramatic shift. Earlier processors used to double there clock frequency every 18-24 months, but now processors are being built with increasing numbers of cores. The second major shift is concerned with the move been made towards cloud computing which aggregate multiple dissimilar workloads with different performance goals into very large clusters. This level of sharing of resources on expensive and large clusters requires new ways of determining how to run and execute data processing jobs so that we can meet the goals of each workload cost-effectively, and to deal with system failures, which occur more frequently as we operate on larger and larger clusters.

Privacy/Security

First hurdle that is faced by big data is that of privacy and security issues. The issues related to privacy that are prominent in big data applications are

- What types of confidential information can be shared and with whom
- Whether cyber communication can be transmitted without anyone else viewing such communication

Data Silos

One of the biggest challenges many enterprises face is trying to collect insights from big data that's trapped in the data silos that exist across business units and organizational functions. Departments like HR, Marketing, sales, etc. within an organization cares more for their own position rather than the organization as a whole. These data vaults prevent decision makers from gaining unified views of customer and operational information. Moreover, they confound the efforts of enterprises to leverage the full range of available data that can reveal insights into emerging customer trends or market shifts that can be acted on quickly to optimize business performance. Data silos create a number of barriers that impede decision making and organizational performance. Enterprises are generating impressive business and operational gains with the use of big data analytics. But a data silo is a big hurdle in achieving a 360-degree view of customers as well as a comprehensive picture of business operations. It has been analyzed that the organizations using big data analytics are seeing an average 26% improvement in performance and it is been expected to see a full 41% improvement over the next three years. So for a better performance organizations needs to pool their data to find the answers to and get a complete view of their data [4].

Data Scientists

It has been usually observed that in an enterprise, there are typically ten times more IT employees as compared to analysts or data scientists. The process of analysis starts with a line of business request. IT employees collect data from various databases and transfer it to data scientists. Thereafter, data scientists are deployed who spend months or even years querying the data. It has been found that people having expertise in statistics and computer science are extremely scarce. The demand and interest in data scientists are skyrocketing. It is a need of time to amplify the impact of data scientists, and allow more people to become data scientists.

Communication

Finally, the biggest hurdle to fully realize the potential of data science is the lack of communication between data scientists and business users. The gap between a data scientist and a business user is so broad that even communicating insights pose a problem. Anything that does not make intuitive sense is often regarded with uncertainty, or not fully understood, by business users, which can lead to missed opportunities. So it is required that both data scientists and business users align themselves and work more closely together, and build trust to solve business problems.

VI. Purpose of Using Big Data in Elections in India

Elections in India till now comprised heat, dust, dirt, drama, traditional wisdom, opinion polls, speeches, processions, door-to-door visits, sweat and toil. Two irreversible trends that were registered in 2014 parliamentary elections were very large young voter base and use of technology to its best. 2014 Lok Sabha elections in India were smartly led by digital/social media technologies. Obama's 2008 presidential campaign ushered use of Social Media and 2012 brought Big Data Analytics to forefront. World's largest democracy went one step further

Gagandeep Jagdev, Bhalwinder Singh and Mahabli Mann, "Big data proposes an innovative concept for contesting elections in Indian subcontinent", *Journal of Scientific and Technical Advancements*, Volume 1, Issue 3, pp. 23-28, 2015.

ISSN: 2454-1532

by integrating Social Media and Big Data Analytics for the first time [7]. The massive exercise was the largest ever on the earth involving 81.4 crore voters compared to USA's 19.36 crore voter, Indonesia's 17.1, Brazil's 13.58 and UK's 4.55 crore. Complexities involved in 2014 Lok Sabha elections were:

- 543 Parliamentary and 4120 assembly constituencies
- 9.3 lakh polling booths
- Voter Rolls in PDF in 12 languages
- 9 lakh PDFs, amounting to 2.5 crore pages to be deciphered
- Diverse range of Voter Names and Information

The real challenge was extraction of voter info from 2.5 crore PDF pages and transliteration of the same into English to fuse with other sources. Technology was a big hurdle. A special infrastructure was built to handle this project which included 64 nodes Hadoop, PostgreSQL and servers that process master file containing over 8 Terabytes of data. Besides this testing and validation was another big task to be performed. Several heuristic algorithms were developed for people classification based on name, geography etc., which help in identification of religion, caste and even ethnicity.

Data from multiple sources which included Census, Economic and Social surveys were mapped to polling booths. Because of this complex nature, no big IT company ever ventured into this. The Indian context is heterogeneous, diverse, non-uniform, hard to collect and complex as compared to USA where high quality data already exists. Finding and reaching a voter is easy and less expensive in US, but in India it's a huge challenge. For instance, Delhi and Andhra Pradesh have good quality electoral data, while UP comes last.

The last time election process was carried out in India, people saw the largest democracy in the world pull in almost 600 million of its residents into the voting process which ushered in the new government. These residents are diverse in every way possible which includes beliefs, sentiments, faith, language and motivations. The selections made by people are also based on a multitude of factors which may be direct or indirect. Among direct factors falls policies for that region, local polarizations, past records and indirect includes factors like geography, television penetration, mobile penetration, financial stability, climate, readership, media, etc. A large part of the voting population (about 30 percent) still feel confused about who to vote for and are guided by social, familial or political influencers. And also there are people who just do not vote – either because they forget, or don't care, or live in regions which are really hard to reach. For a political party to win, they need a combination of new voters, influenced voters and those that form the core groups who always vote for them. Political parties construct different strategies to target each category. Imagine this, what if a political party is able to identify those people who are most likely to vote? Why would they then spend a fortune to do general canvassing with the people who are not likely to vote? Shouldn't they rather analyze social circles and find out key influencers and mavens and try to convince them to side with them. They could spend all their effort to combine the "want to's" with the "have to's". How can they do this?

An individual can be judged by examining his or her interest in reading the type of newspaper or magazine. If you are in the west and you subscribe to "Samna" or if you are in the north and subscribe to "Ajit", it does display your appetite towards a certain political party. Similarly if you are a woman and subscribe to newspaper or magazines related to politics or national and international affairs, you may be more likely to vote than a person reading good Housekeeping magazines. Political parties could then analyze the views you express on social media as well as identify your social circles. If they find you influential enough, they could reach out to you over a variety of channels.

Further imagine what if the parties could find out what a voter really wants to listen to and then target them with specific key words and messages. We watch the television daily and if one smart party tells us something really nice about themselves during an advertisement break, we are bound to listen to it. Once we see the message we could tweet something good or bad about it and if the party finds public response against them, the party can change its messaging. Further to this, they could try finding the right spokesperson (like a movie star or a sportsperson, etc.) which may be selected from local movies, media and teams, who is best received by a target segment and dynamically keep changing them based on the kind of chatter it generates. Voters can also draw out comparative scorecards of various party candidates and how they have done in various constituencies based on news feed, policy decisions, project completions, etc., all automatically updated based on inputs from various channels. Such score cards may help people decide objectively on which candidate is actually better.

All this is about mining patterns of interest and once found, they can be applied to a large set of data to predict outcomes. What do this all sounds like? To us it sounds like a massive BIG data problem waiting to be solved. 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. The patterns found could then be applied to past election data and voter turnouts in each constituency and then various prediction models could be built around them which could simulate various scenarios. 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], [9].

VII. Challenges Faced by Big Data in Administrating Elections Procedure

Many statistical organizations already started to investigate the possibility of using the Big Data as a source to complement and support the election procedure. The use of Big Data in administrating elections process presents many challenges, falling into the following categories:

- Legislative, i.e. with respect to the access and use of data
- Privacy, i.e. managing public trust and acceptance of data reuse and its link to other sources
- Financial, i.e. potential costs of sourcing data vs. benefits
- Management, e.g. policies and directives about the management and protection of the data
- Methodological, i.e. data quality and suitability of statistical methods used in election process.

Technological, i.e. issues related to information technology.

VIII. Conclusion

Big data analytics is going to be main stream with increased adoption among every industry and form a virtuous cycle with more people wanting access to even bigger data.

Gagandeep Jagdev, Bhalwinder Singh and Mahabli Mann, "Big data proposes an innovative concept for contesting elections in Indian subcontinent", *Journal of Scientific and Technical Advancements*, Volume 1, Issue 3, pp. 23-28, 2015.

ISSN: 2454-1532

However, often the requirements for big data analysis are really not well understood by the developers and business owners, thus creating an undesirable product. For organizations to not waste precious time and money and manpower over these issues there is a need to develop expertise and process of creating small scale prototypes quickly and test them to demonstrate its correctness, matching with business goals.

Since Big data is an emerging technology and is at its youth, so it needs to attract organizations and youth with diverse new skill sets. The skills should extend from technical to research, analytical, interpretive and creative ones. These skills need to be developed in individuals hence requires training programs to be held by the organizations. Also universities should introduce curriculum on big data to produce skilled employees and data scientists in this expertise.

Finally, it can be concluded that big data is all set to play a major role in any national elections to be conducted in future. Sophisticated campaigns develop and use voter databases that contain a range of detailed information on individual citizens. As a result, campaign data analysts occupy an increasingly important role in politics. They develop predictive models that produce individual-level scores that predict citizen's likelihoods of performing certain political behaviors, supporting candidates and issues, and responding to targeted interventions. The improved capability to target individual voters offers campaigns an opportunity to concentrate their resources where they will be most effective.

Political parties have to concentrate on the use of technology much more than other matters. Appropriate use of big data guarantees the big win of the political parties. The use of big data by BJP has already been proved by the huge victory they obtained in 2014 Lok Sabha elections. But free access to electronic electoral databases may have an adverse effect and would endanger the very goal it seeks to achieve because the electronic database may pose threat to privacy of the voters and also lead to security breach. It may be argued that the ECI is mandated by the law to publish the electoral database and hence, it is beyond the operation of the IT Act. But Section 81 of the IT Act has an overriding effect on any law inconsistent, therewith. The appropriate Government should take necessary steps under the IT Act and notify electoral databases as a protected system.

REFERENCES

[1] www.sas.com/en_us/insights/big-data/what-is-big-data.html

[2] www.mongodb.com.big-data-explained

[3] www.forbes.com/sites/lisaarthur/2013/08/15/what-is-big-data/

[4] www.ibmbigdatahub.com/tag/587

[5] www.ibm.com/software/data/bigdata/what-is-big-data.html

[6] http://www.business2community.com/digital-marketing/4-vs-big-data-digital-marketing-0914845

[7] www.firstpost.com/tag/election-commission-of-india

[8] http://blog.expressanalytics.net/index.php/uncategorized/role-of-social-media-data-analytics-in-loksabha-generalelections-2014-in-india/

[9] http://www.informationweek.in/informationweek/news-analysis/277217/help-predicting-results-elections-india

[10] Ensuring "Big Data" Security with Identity and Access Management., Aveksa Inc., Waltham, MA: Aveksa, 2013.

Gagandeep Jagdev, Bhalwinder Singh and Mahabli Mann, "Big data proposes an innovative concept for contesting elections in Indian subcontinent", *Journal of Scientific and Technical Advancements*, Volume 1, Issue 3, pp. 23-28, 2015.