Exploring Application of Big Data in Elections – From Data to Action

Amandeep Kaur[1], Dr. Gagandeep Jagdev[2]

[1]*Research Scholar (Ph.D), Dept. of Comp. Appl., Guru Kashi University, Talwandi Sabo (PB)*

[2]*Dept. of Comp. Science, Punjabi University Guru Kashi College, Damdama Sahib (PB)*

*****Corresponding Author:** *Dr. Gagandeep Jagdev*, *Dept. of Comp. Science, Punjabi University Guru Kashi College, Damdama Sahib (PB), India.*

Abstract: *"The numbers don't lie". It is a known and famous saying. But even if this is true most of the times, the numbers can be slippery, cryptic, and, sometimes, two-faced. As big data is multiplying itself and infiltrating more and more parts of our business, the necessity for understanding the meaning of numbers is getting more acute. There is a need to fundamentally rethink how the data analysis can create value. 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. To handle such huge data, there is a need to deploy infrastructure differently to support distributed processing requirements. This paper discusses the working of evolving technologies like Hadoop framework and MapReduce technology responsible for transforming big data into meaningful insights. The research paper concentrates on the utilization of big data to utmost potential in winning elections held at any level. The primary aim of the paper is to elaborate the different issues and challenges involved in handling big data and how to attract electorates.*

Keywords: *Apache Hadoop Framework; Big Data; Elections; Map Reduce.*

Abbreviations: *HDFS*

1. INTRODUCTION

Billions of Internet users and machine-to-machine connections are causing a tsunami of data growth. Big data [1, 2] is the data that exceeds the processing capacity of conventional database systems. 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. Big data has found its applications in every sector and the elections is an exception anymore. Today political parties concentrate and focus deeply on social media relevant dynamics like comments, likes, dislikes, followers, popularity and much more. Gone are the days when grey- haired politicians used to fight elections on the basis of religion, caste, and ethnicity. Today the only importance is of numbers and statistics irrespective of the fact that whether the election is ata local or national level. When we consider election scenario, the political parties must avoid broadcast approach where the same message is delivered to the whole nation. The preference should be given to a targeted approach where data canbe utilized to directly converse with the individuals and masses on the critical issues that matter most to them. There has been 240 percent increase in politicians making use of social media sites like Facebook and Twitter world over. The electoral success of Narendra Modi led BharatiyaJanta Party (BJP) in India is also credited with the effective use of big data. The BJP applied the big data sets to get aware of the demand for masses. It assisted them in shaping core issues and media campaign in 2014 elections in India. By conducting data analysis, BJP managed to raise funds, edit advertisements and engage voters from different regions.

2. PROMINENT SOURCES OF BIG DATA

There are many sources responsible for the generation of big data as mentioned under and shown in Fig. 1.

A. Social Networking

The data mining competences for business are growing continuously with the growth of social networks like Twitter, Facebook, and Pinterest. These act as a wealth of information when it comes to

defining the inspiration of the customers. Because of its open-source nature, social networking data is continuously available for all marketers. Some examples of this type of data are follower/following accounts, interaction with posts (likes, comments, pins, etc.), the geography of social media user, and sentiment analysis of posts and tweets.

B. Review Networks

The businesses are provided an unparalleled look at online review sites when viewed by their customers. These acts as a windfall for companies who are interested to gather insights from data that is easy to segment. For example, TripAdvisor and Google Reviews provide sentiment data after analyzing the tone and language of entries.

C. Government Data Sources

At first listening about government data sources at the big data party seems boring, but it plays a big role in the number of analyses. Healthcare, legal information, population, weather and disaster forecast; these kinds of data are authoritative when looking to apply big data research across the board. Many countries are accepting the free data mindset and making it publically available on the internet.

D. Blogesphere

In the context of big data, blogs are a useful source of data because of their targeted audiences. Whether a company is engaged in selling toys or providing consulting services to other businesses, there is most likely a blog there with a strong audience of users constantly interacting with posts. The comments section of blogs is another source of data capturing opinions of interested visitors.

E. Media Outlets

Today's user communication driven media outlet websites are alike to blogs in being actual data sources due to their generally targeted audiences and active visitor feedback. Using various data mining techniques, marketers can collect information on the wide extent of subject matter carried by newspaper and news magazine sites. The comment sections of these sites are overflowing with insight into the views and opinions of the general public that could be relevant to marketers.
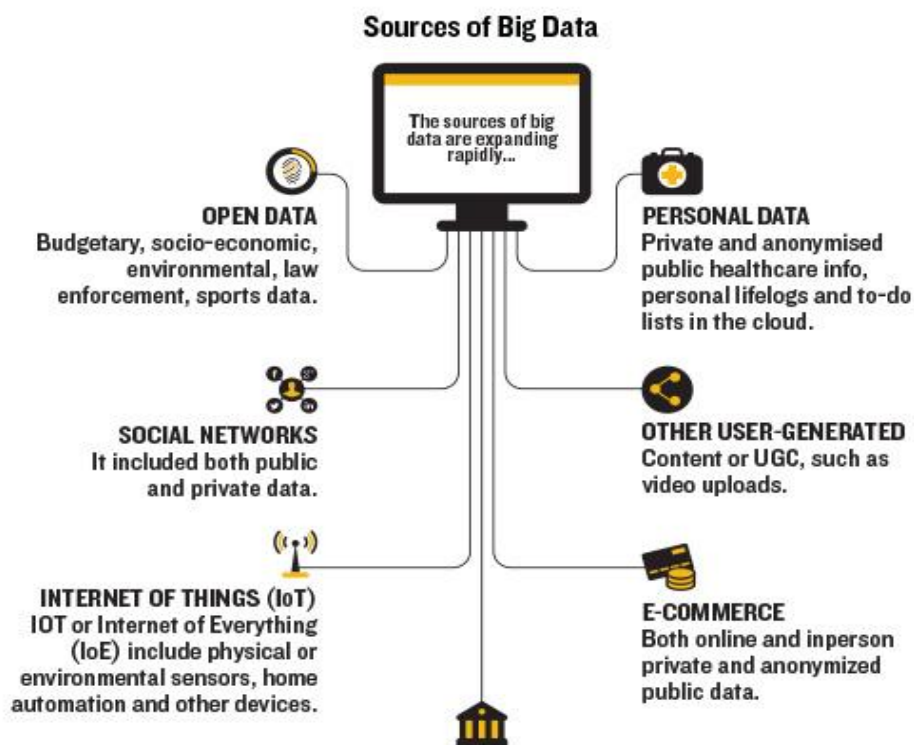

Fig1. *The figure depicts the different sources responsible for generation of big data*

The amount of data generated by some prominent agencies and organizations every 24 hours is shown in Fig. 2.
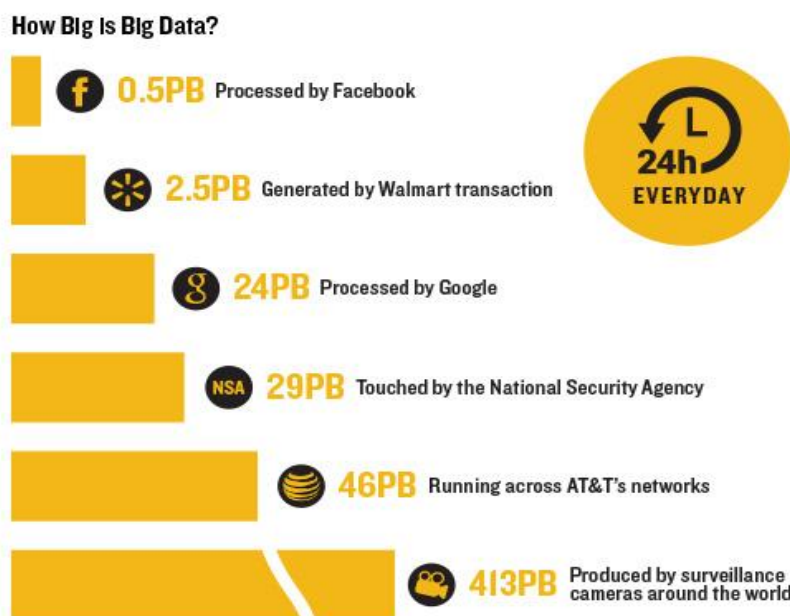
Fig2. The figure shows the amount of data generated by few famous agencies, social networking sites, and organizations every 24 hour

3. BIG DATA FROM ELECTIONS PERSPECTIVE

The data related to Indian elections are huge and it is increasing day by day. This data needs to be handled in such a manner that it could be utilized to its best and used for the analyzing the different scenarios which frequently comes up in elections. It may be said that massive data is all set to play a maximum vital function in any countrywide elections to be conducted in future. Political parties need to ponder on the use of technology, a brilliant deal more than another count. The research paper is focused to remove the confusion in the minds of electorates who to vote for. A record related to elections varies in size and ranges from Petabytes to Yottabytes. The giving out an analysis of the big amount of information related to elections or producing the high priced information is the tough assignment. Inside the large data is the latest generation that could be beneficial for the general public and political parties. So it's far essential that special issues and challenges related to this technology have to convey out into the light. However, with the first-rate possibility comes top-notch demanding situations and there's a host of upcoming obstacles which we have to address.

4. APACHE HADOOP FRAMEWORK AND MAPREDUCE ALGORITHM

Big Data, because of its volume, veracity, variety, and velocity can't be handled by traditional database tool. So, a special tool is needed to handle Big Data. Hadoop finds its application in several sectors which comprises of retail, sports, medical science, business, education and of course now in elections.Hadoop is a framework for running applications on large clusters of commodity hardware which produces huge data and to process it. Hadoop is a Java-based programming framework that supports the processing of large data sets in a parallel computing environment.

Hadoop is undoubtedly the preferred choice for such a requirement due to its key characteristics of reliability, flexibility and a scalability. Hadoop provides the ability to store huge data on Hadoop Distributed File System (HDFS).There are many other solutions available for analyzing this huge data such as MapReduce, Pig, and Hive. Due to advancements in data analysis technologies to analyze the big data, there are different thoughts about which and when Hadoop data analysis technology should be used efficiently.A well-executed big data analysis provides the possibility to uncover hidden markets, finds customer demands and cost reduction opportunities.It can be helpful in improving efficiency in telecommunication and medical treatments, and big data analytics is also helpful in social media campaigns and related digital marketing.

The important characteristics of Hadoop framework are detailed as under.

- **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 stop 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 the 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.

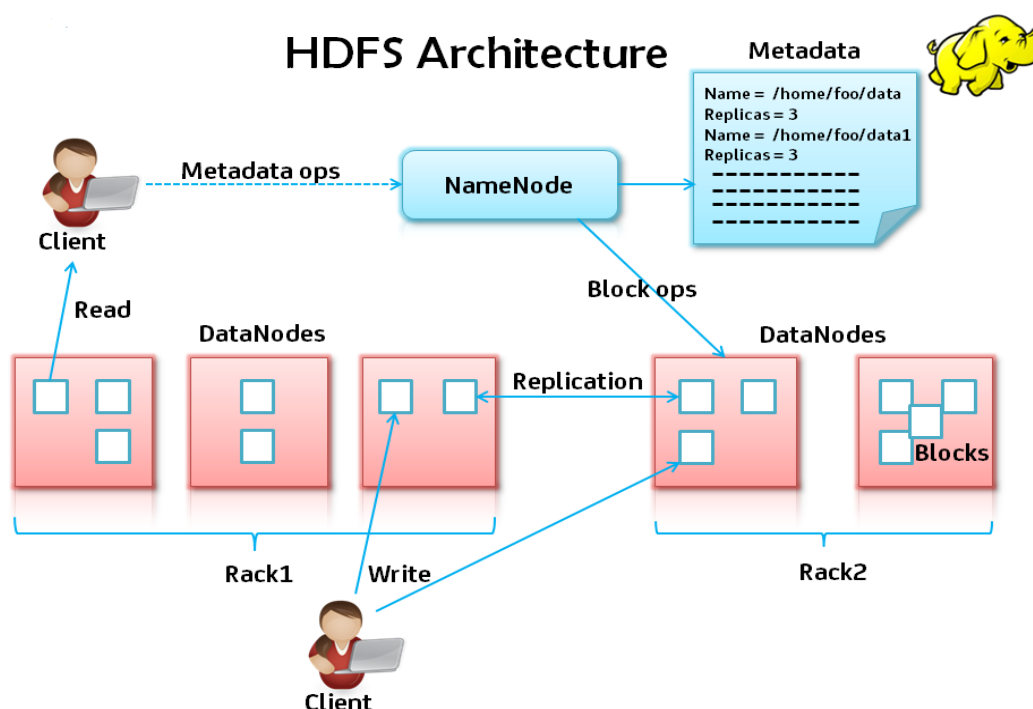Fig. 3 shows the Hadoop Distributed File System (HDFS).

Fig3. *The figure depicts the Hadoop Distributed File Structure*

Big data had found its applications in almost every sector and an election is no more an exemption. It had been predicted that technology today is much more influential in planning strategies for contesting elections than any other thing. The US president Barack Obama had set up a platform describing the use of big data in contesting and winning elections. In India political parties are now adopting the same concept. It had been brought to the forefront that majority of electorates in India do not cast their own vote. On having deep insight into this matter it had been confirmed that either these people are confused or unable to decide who to vote for or they simply don't care to vote. Political parties are now busy targeting such electorates. The research work would guide such electorates to decide who to vote for based on their own judgment and not under the influence of any external factor. The structured database of different political leaders of Punjab state has being constructed who contested past elections held in Punjab. Next year in 2017, these leaders are again all set to contest the elections. In the structured database, fifteen different attributes have been considered mentioned below and provide details about a political leader. One could pass appropriate query or write a suitable script to know what he/she wants to know about any political party or political leader.

Name, Age, Education, Sex, Party, Partytype, Votes_in_favour, perofvotesinfavour, Criminalcases, Assets, Liabilities, Status, Winningchances, Constituency_name, Year.

The glimpses of snapshots of structured database are shown in Fig. 4as under.

NAME	AGE	EDUCATION	SEX	PARTY	PARTYTYPE	VOTESINFAVOUR	PEROFVOTESI	CRIMINAL	ASSETS	LIABILITIES	STATUS	WINNINGCH	CONSTITUENCY	YEAR
Harsimrat Kaur Badal	47	10th	Female	SAD	State	514727	46.09	0	100 Crore	41 crore	Crorepati	Bright	BATHINDA	2014
Manpreet Singh S/O Gurdas Singh	52	Graduate	Male	INC	National	495332	44.35	0	42 Crore	4 Crore	Crorepati	Bright	BATHINDA	2014
Jasraj Singh Longia	40	Graduate	Male	AAP	National	87901	7.87	0	31 Lac	Nil	Lakhpati	Average	BATHINDA	2014
Kuldeep Singh	32	10th	Male	BSP	National	13732	1.22	0	80 Lac	Nil	Lakhpati	Poor	BATHINDA	2014
Ashish	25	10th	Male	IND	Independent	6626	0.59	3	30 Thousand	Nil		Poor	BATHINDA	2014
Bhagwant Singh Samaon	36	8th	Male	CPI	National	5984	0.53	0	2 Lac	15 Thousand	Lakhpati	Poor	BATHINDA	2014
Satish Arora	51	12th	Male	IND	Independent	5936	0.52	0	1 Crore	Nil	Crorepati	Poor	BATHINDA	2014
Manpreet Singh S/O Gurdev Singh	33	10th	Male	IND	Independent	4618	0.413	0	7 Lac	Nil	Lakhpati	Poor	BATHINDA	2014
Shaminder Singh	34	12th	Male	JKKNPP	National	4610	0.412	0	33 Lac	2 Lac	Lakhpati	Poor	BATHINDA	2014
Geeta Rani	50	5th	Female	ABSR	National	4380	0.39	0	Nil	Nil		Poor	BATHINDA	2014
Sanjeev Kumar Thapar	34	10th	Male	IND	Independent	2833	0.253	0	26 Lac	2 Lac	Lakhpati	Poor	BATHINDA	2014
Makhan Lal	43	8th	Male	BSP(A)	National	2587	0.231	0	91 Thousand	Nil		Poor	BATHINDA	2014
Vijay Kumar	44	10th	Male	IND	Independent	2258	0.202	2	6 Lac	Nil	Lakhpati	Poor	BATHINDA	2014
Swaran Singh	45	Graduate	Male	IND	Independent	2077	0.185	0	1 Crore	Nil	Crorepati	Poor	BATHINDA	2014
Rajinder Singh	34	10th	Male	SAD(A)	State	1960	0.175	0	26 Lac	Nil	Lakhpati	Poor	BATHINDA	2014
Navneet	42	12th	Male	SHS	National	1893	0.169	0	57 Thousand	5 Lac		Poor	BATHINDA	2014
Jagdeep Singh Gehri	39	10th	Male	IND	Independent	1665	0.149	0	1 Lac	Nil	Lakhpati	Poor	BATHINDA	2014
Bhupesh Kumar	36	Graduate	Male	IND	Independent	1487	0.133	0	6 Lac	Nil	Lakhpati	Poor	BATHINDA	2014
Gurdeep Singh	44	Illiterate	Male	IKL	National	1404	0.125	0	5 Lac	Nil	Lakhpati	Poor	BATHINDA	2014
Gurmeet Singh Ranghreta	44	12th	Male	PLP	National	1386	0.124	1	35 Lac	1 Lac	Lakhpati	Poor	BATHINDA	2014
Sushil Kumar Jindal	31	10th	Male	IND	Independent	1330	0.119	0	1 Lac	Nil	Lakhpati	Poor	BATHINDA	2014
Jagdish Rai Sharma	57	Post Graduate	Male	NBDP	National	1248	0.111	1	95 Lac	3 Lacs	Lakhpati	Poor	BATHINDA	2014
Sukhwinder Singh	33	12th	Male	BMP	National	1086	0.097	0	2 Lac	Nil	Lakhpati	Poor	BATHINDA	2014
Dyal Chand	52	8th	Male	IND	Independent	955	0.085	0	58 Thousand	Nil		Poor	BATHINDA	2014
Suresh Kumar Goyal	70	NA	Male	JSP	National	944	0.084	0	15 Lac	1 Lac	Lakhpati	Poor	BATHINDA	2014

Fig4. *The figure shows the glimpse of constructed database*

The huge database can be explored to extract useful information by writing appropriate scripts and queries using Apache Hadoop framework. The backbone of the working of the Hadoop framework is the MapReduce algorithm. MapReduce operates in three different stages; mapping, shuffling and reducing. Fig. 5 shows the working of MapReduce algorithm.
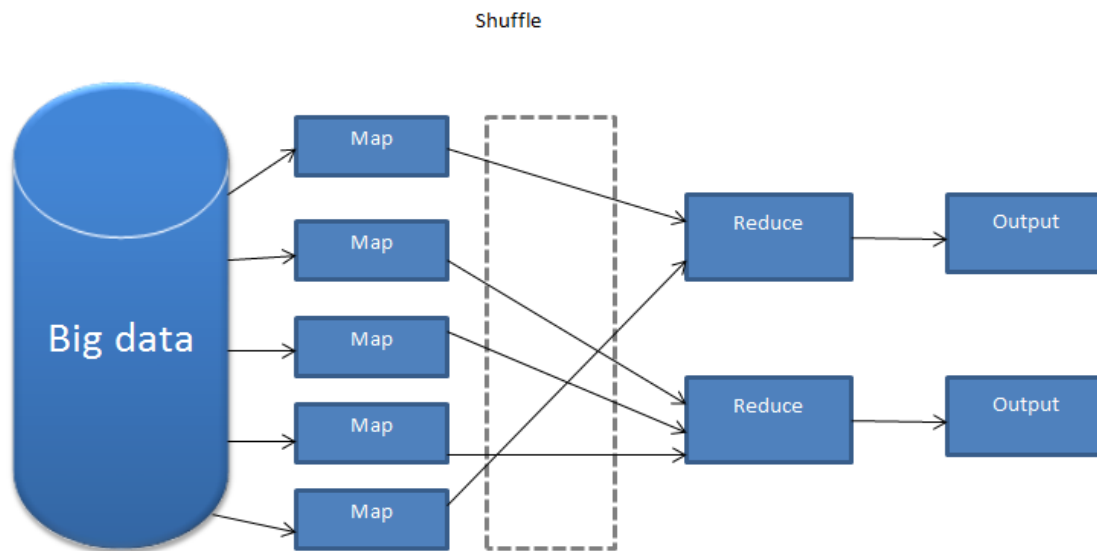

Fig5. *The figure shows the working of MapReduce technique*

The below-mentioned examples illustrate the extraction of useful information from the database in accordance with the fulfillment of the objectives of the research work.

Query 1

To construct the constituency wise sub-tables from the main table "election_data". Fig. 6 shows the query written in Query Editor of Apache Hive to form tables of two constituencies titled "Bathinda" and "Amritsar". The attributes included in both the tables are the *name, age, party, votesinfavour, constituency, and year.*
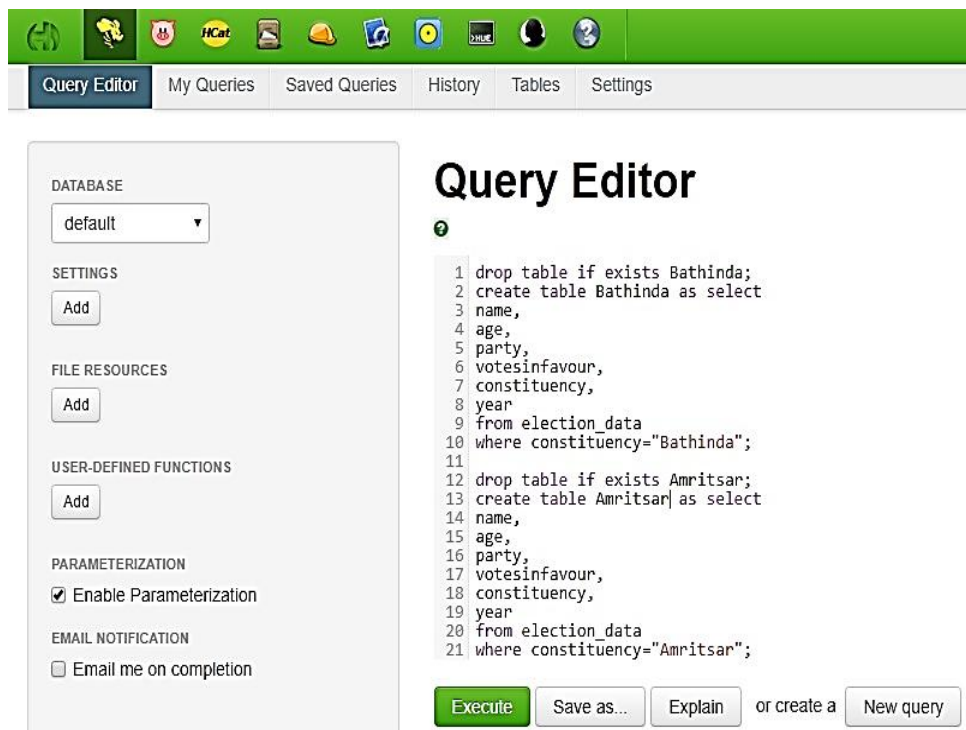
Fig6. *The figure displays the query written in Query Editor of Apache Hive*

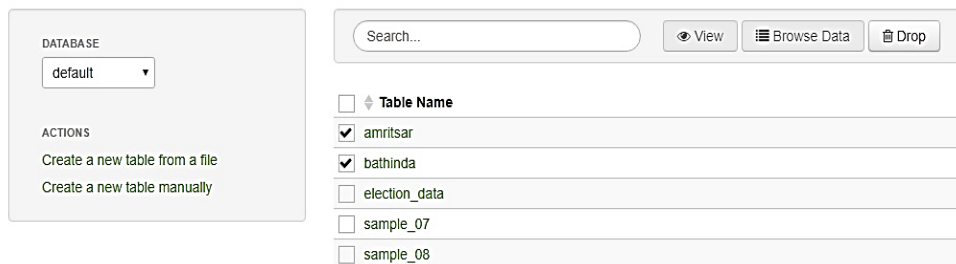Fig. 7 displays the tables "Bathinda" and "Amritsar" formed in the table list of HCatalog.

Fig7. *The figure displays the tables "Amritsar" and "Bathinda" in the table list.*

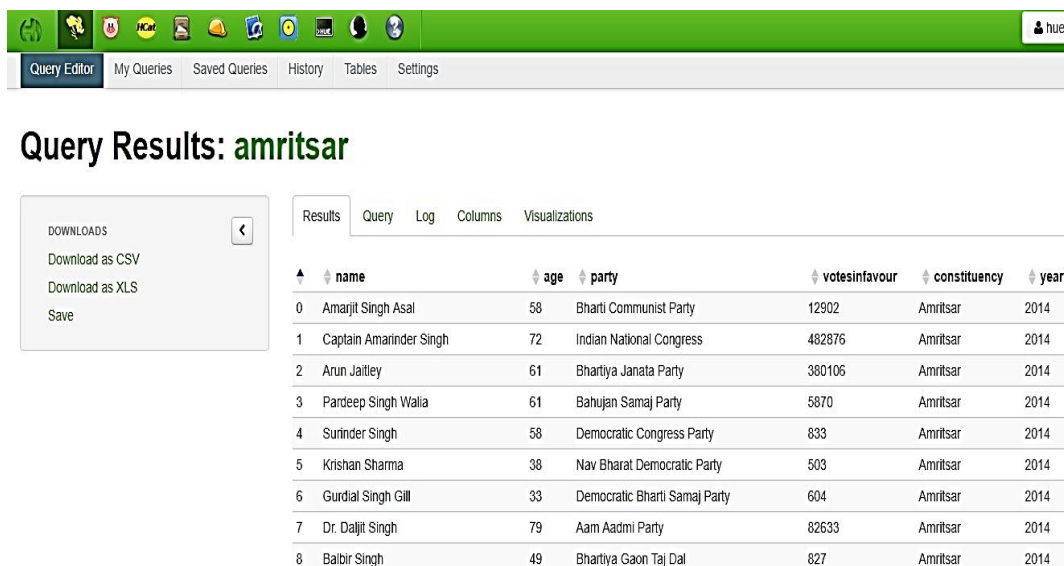The contents of the table "Amritsar" are shown below in the Fig. 8.

Fig8. *The figure displays the contents of table "Amritsar" in tabular form*

Fig. 9 shows the graphical result of the table *"Amritsar"* in which the attribute *"name"* constitutes X-axis and the attribute *"age"* constitutes Y-axis. The bar depicts the age of candidates against their names.
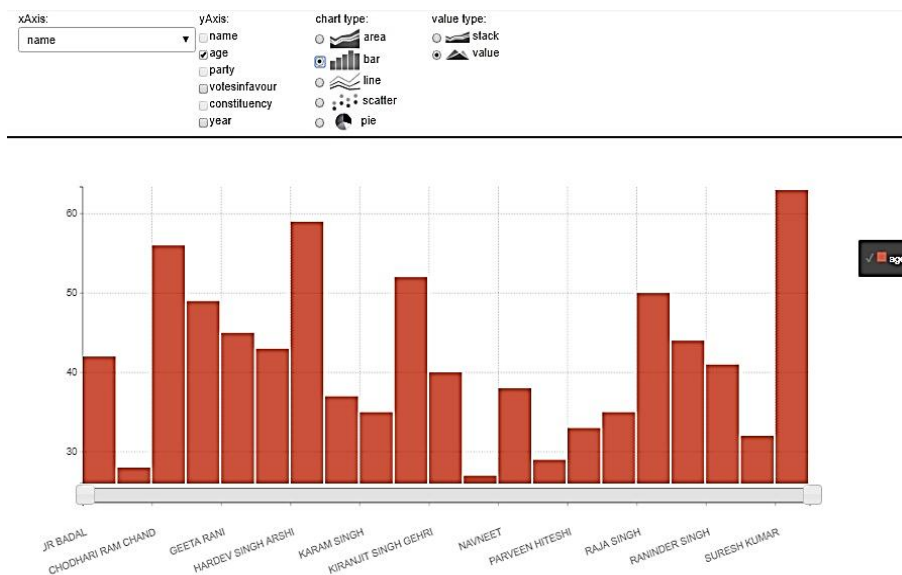

Fig9. The figure depicts the *"name"* of the candidate against his/her age in the table *"Amritsar"*

The Fig. 10 shows the graphical result of the table *"Amritsar"* in which the attribute *"name"* constitutes X-axis and the attribute *"votesinfavour"* constitutes Y-axis. The area depicts the name of candidates against the attribute *"votesinfavour"*.
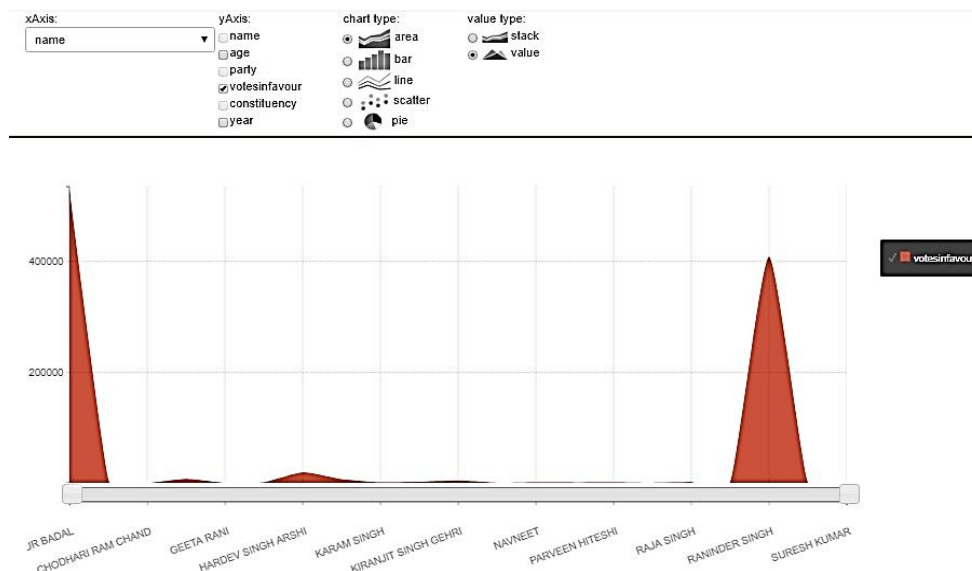

Fig10. *The figure depicts the "name" of the candidate against the attribute "votesinfavour" in the table "Amritsar"*

In a similar manner, the database can be mined to obtain fruitful results via writing appropriate queries and scripts.

5. CONCLUSION

India is the largest democracy in the world and it is prime duty of every citizen to vote who is eligible to do so. The significance of the research conducted in this research paper is directly and primarily concerned with the electorates of India who are all set to cast their votes in the upcoming elections in 2019. It was observed that in 2014 Lok Sabha elections, 814 million electorates were supposed to vote but it was found that out these people 164 million electorates did not vote or either opted for NOTA. The implication of the study is to provide guidance to the people who did not vote because either they did not care to vote or the appropriate message which they were waiting to hear from any political party did not reach their ears. The research work conducted can assist common electorates to decide

about the best option to vote for among several candidates belonging to different parties and independent ones. The paper elaborated the working of Apache Hadoop Framework and Map Reduce technology with elections perspective.

REFERENCES

[1] Dean, Jeffery, and Ghemawat Sanjay. 2004. "MapReduce: Simplified Data Processing on Large Clusters." Google.

[2] Katal, A., Wazid, M., &Goudar, R. H. (2013). Big Data: Issues, Challenges, Tools and Good Practices. IEEE, 404-409.

[3] Gagandeep Jagdev et al., "Comparing Conventional Data Mining Algorithms with Hadoop based Map-Reduce Algorithm considering elections perspective", *International Journal of Innovative Research in Science and Engineering (IJIRSE)*, ISSN: 2454-9665 (O), ISSN: 2455-0663(P), Volume – 3, Issue – 3, March 2017.

[4] Gagandeep Jagdevet al., "Association of Big Data with Map-Reduce Technology Augments for Economic Growth in Retail", *International Journal of Engineering Technology Science and Research (IJETSR),* ISSN: 2394 - 3386, Volume 4, Issue 2, February 2017.

[5] N. Marz and J. Warren. Big Data: Principles and best practices of scalable realtime data systems. Manning Publications, 2013.

[6] R. Smolan and J. Erwitt. The Human Face of Big Data.Sterling Publishing Company Incorporated, 2012.

[7] Gagandeep Jagdevet al., "Analyzing Maneuver of Hadoop Framework and MapR Algorithm Proficient in supervising Big Data", *International Journal of Advanced Technology in Engineering and Science (IJATES)*, ISSN – 2348-7550,Volume – 05, Issue – 05, May 2017.

[8] Gagandeep Jagdev et al., "Implementation and Applications of Big Data in Health Care Industry", *International Journal of Scientific and Technical Advancements (IJSTA)*, ISSN: 2454-1532, Volume 1, Issue 3, pp: 29 – 34.

[9] Undefined by Data: A Survey of Big Data Definitions, by Jonathan Stuart Ward and Adam Barker, School of Computer Science, University of St. Andrews, UK, 2013. Available at: http://arxiv. org/pdf/ 1309.5821v1.pdf.

[10] Gagandeep Jagdev et al., "Big Data proposes an innovative concept for contesting elections in Indian subcontinent", *International Journal of Scientific and Technical Advancements (IJSTA),*Volume 1, Issue 3, pp. 23-28, 2015, ISSN No. 2454-1532.

[11] Gagandeep Jagdev et al., "Analyzing and Scripting Indian Election Strategies using Big Data via Apache Hadoop Framework", *IEEE Xplore*, DOI: 10.1109/WECON.2016.7993431, INSPEC Accession Number: 17061464, 27 July 2017.

[12] Gagandeep Jagdev et al., "Excavating Big Data associated to Indian election scenario via Apache Hadoop",*International Journal of Advanced Research in Computer Science*. 2016 November, 7 (6), pp. 117-123.

[13] Gagandeep Jagdev et al., "Big Data Alliance with MapR Technology to Govern Indian Elections Panorama", Proceedings of IDES Fifth International Conference IETE, Kurukshetra, India, 2016, pp. 55-61.

[14] Gagandeep Jagdev et al., "A Comparative study of conventional data mining algorithms against Map-Reduce algorithm",*International Journal of Advanced Research in Science and Engineering*. 2017 May. 6 (5), pp. 325-335.

AUTHOR'S BIOGRAPHY

Dr. Gagandeep Jagdev, is a faculty member in Dept. of Computer Science, Punjabi University Guru Kashi College, Damdama Sahib (PB). His total teaching experience is above 11 years and has above 118 international and national publications in reputed journals and conferences to his credit. He is also a member of editorial board of several international peer-reviewed journals and has been active Technical Program Committee member of several international and national conferences conducted by renowned universities and academic institutions. His field of expertise is Big Data, ANN, Biometrics, RFID, Cloud Computing, Cryptography, and VANETS.