ISSN No. 0976-5697

Volume 7, No. 6(Special Issue), November 2016

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

# Excavating Big Data associated to Indian Elections Scenario via Apache Hadoop

Dr.Gagandeep Jagdev[#1] Amandeep Kaur[#2] Amandeep Kaur[#3]
[#1]Dept. of Computer Science, Punjabi University Guru Kashi College, Damdama Sahib (PB).
[#2]M.C.A., Guru Kashi University, Talwandi Sabo (PB).
[#3]Web & Graphic Designer, Eternal University, Baru Sahib, Himachal Pradesh.
[#1]drgagan137@pbi.ac.in

***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. It has been only in 2013 to 2015 that we humans have created 90 percent of data existing on the planet earth since existence of humans on this planet. The huge technological up gradation in social network, in retail industry, in health sector, in engineering disciplines, in the field of wireless sensors, in stock market, in public and private sector, all has collectively amassed enormous data. This data is very huge in volume, it gets created at very high speed, it may be structured, unstructured, semi-structured or may be in text, audio or video format and most important that it is not totally precise and can be messy or misleading. The central theme of our research work is concerned with handling huge amount of data that is concerned with different formats of elections that are been contested in India. The framework used in this research work is Apache Hadoop. Apache Hadoop framework makes use of Map-Reduce technology which operates in three steps: mapping, shuffling and reduction. Map-Reduce is the same technique which Facebook use for handling its section of "People you may know". Research paper also discusses the working of Map-Reduce technology with competent examples.

***Keywords*** *- 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 daywith 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 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, 16].

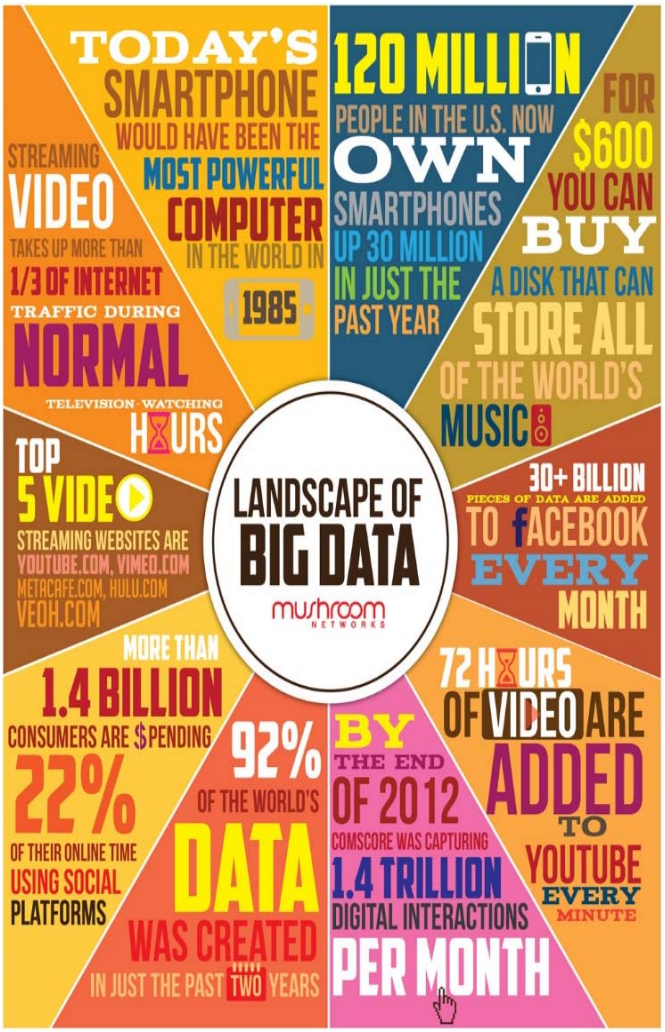


Fig. 1 – Sources and statistics of Big Data

978-93-85670-72-5 © 2016 (RTCSIT)

CONFERENCE PAPER
International Conference on
Recent Trends in Computer Science & Information Technology (RTCSIT-2016)
21st August 2016
Guru Nanak College Budhlada, Punjab India

Some surprising facts related to big data are as under.

- Over 90% of all the data in the world was created in the past 2 years.
- Around 100 hours of video are uploaded to YouTube every minute and it would take you around 15 years to watch every video uploaded by users in one day.
- If you burned all of the data created in just one day onto DVDs, you could stack them on top of each other and reach the moon – twice.
- 571 new websites spring into existence every minute of every day.
- The number of Bits of information stored in the digital universe is more than the number of stars in the physical universe.
- The total amount of data being captured and stored by industry doubles every 1.2 years.
- Every minute we send 204 million emails, generate 1.8 million Facebook likes, send 278 thousand Tweets, and up-load 200,000 photos to Facebook.
- Google alone processes on average over 40 thousand search queries per second.

## II. Issues Related With Big Data Characteristics

- Data Volume – It refers to the enormous amount of data that is been created each second, each minute and each hour of the day. 571 websites are created in a single minute. Total of 625000 GB of data is transferred from one end to another in single internet minute, may be terms of mails, pictures, posts etc. If we burn the amount of data present on planet earth today on DVDs and pile them in the form of a stack one upon another, the pile will be such huge that one can climb it and touch the moon, come back to earth and again repeat this process once.
- Data Velocity – Data is being created at such high velocity that companies are finding it difficult to cope up with such high speed. They have to establish their infrastructure in such a manner that it is capable of handling such generated data Social media, E-Commerce has rapidly increased the speed and richness of data used for different business transactions.
- Data Variety - All the data being generated is totally diverse consisting of raw, structured, semi structured and even unstructured data which is difficult to be handled by the existing traditional analytic systems. Mismatched data formats and data structures highlight major challenges that can be responsible for analytic collapse.
- Data Value –There is a big breach between the business leaders and the IT professionals. Business leaders are primarily concerned with adding values to their business and to maximize their profit. On the other hand, IT leaders are busy dealing with workings of the storage and processing.
- Data Complexity – The biggest complexity faced while running big data using relational databases is that they require parallel software running on hundreds of servers and data scientists have to match and transform data across systems coming from various sources.
- Data Veracity - Veracity refers to the preciseness of data or how much faith one can have on data. The data on internet is not always accurate or precise. For example, if some male pretends himself as a female on his Facebook profile, there is no authenticity check in such cases. Similarly twitter makes use of abbreviations and hash tags, but big data enables us to work with even this type of imprecise data [1, 6, 7, 12].

## III. Five Phases Of Big Data

Big data processing involves five different phases [2, 6, 14, 15].

***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 [14, 15].

***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 [15].

***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

978-93-85670-72-5 © 2016 (RTCSIT)

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 [14, 15].

***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.

***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 [14, 15].

## IV. Electioneering - Using Big Data In Elections 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.

For conducting 2014 Lok Sabha elections, 543 Parliamentary constituencies and 4120 assembly constituencies were set up. All over India total of 9 lakh 30 thousand polling booths were set up for conducting fair elections. Voter rolls were prepared in 12 different languages and total of 9 lakh pdf files which amounted to 2.5 crore pages were deciphered. The real challenge was extraction of voter info from these 2.5 crore PDF pages and transliteration of the same into English to fuse with other sources. Technology was a big hurdle.

Behavior scores use past behavior and demographic information to calculate explicit probabilities that citizens will engage in particular forms of political activity. The primary outcomes campaigns are concerned with include voter turnout and donations, but other outcomes such as volunteering and rally attendance are also of interest.

Support scores predict the political preferences of citizens. In the ideal world of campaign advisers, campaigns would contact all citizens and ask them about their candidate and issue preferences. However, in the real world of budget constraints, campaigns contact a subset of citizens and use their responses as data to develop models that predict the preferences of the rest of the citizens who are registered to vote. These support scores typically range from 0 – 100 and generally are interpreted to mean "if you sample 100 citizens with a score of X, X percent would prefer the candidate/issue". A support score of "0" means that no one in a sample of 100 citizens would support the candidate/issue, "100" means that everyone in the sample would support the candidate/issue, and "50" means that half of the sample would support the candidate/issue. Support scores only predict the preferences at the aggregate-level, not the individual-level. That is, people with support scores of 50 are not necessarily undecided or ambivalent about the candidate/issue and, in fact, may have strong preferences. But when citizens have support scores of 50, it means that it is difficult to predict their political preferences.

Responsiveness scores predict how citizens will respond to campaign outreach. While there are theoretical rationales as to who might be most responsive to blandishments to vote and attempts at persuasion, in general, predicting which individuals will be most and least responsive to particular direct communications in a given electoral context is difficult. Campaigns can use fully randomized field experiments to measure the response to a campaign tactic. The results of these experiments can then be analyzed to detect and model heterogeneous treatment effects (i.e., predictive scores) that guide targeting decisions. Some of the results of these experiments can only be used to inform decisions in future elections (e.g., the results of most voter turnout experiments necessarily come after Election Day), but others can be conducted during the election cycle to improve efficiency in real time [8, 14, 16].

## V. Apache Hadoop Platform And Map-Reduce Technology

Hadoop [9, 10, 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

978-93-85670-72-5 © 2016 (RTCSIT)

CONFERENCE PAPER
International Conference on
Recent Trends in Computer Science & Information Technology (RTCSIT-2016)
21st August 2016
Guru Nanak College Budhlada, Punjab India

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 six components shown in Fig. 2[4].

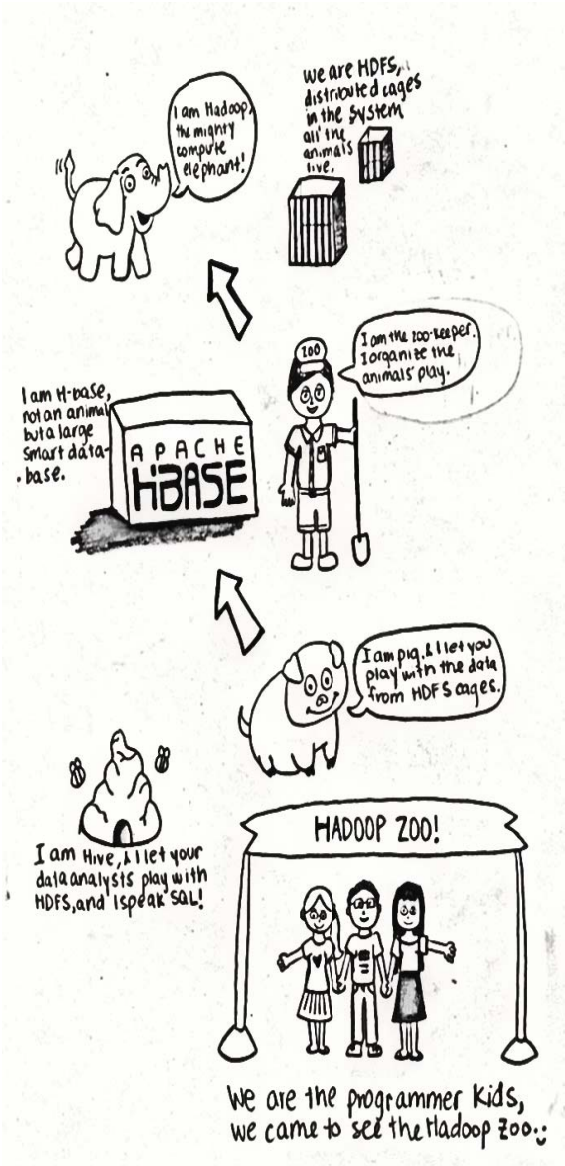


Fig. 2 - Hadoop Zoo

- 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. 3.

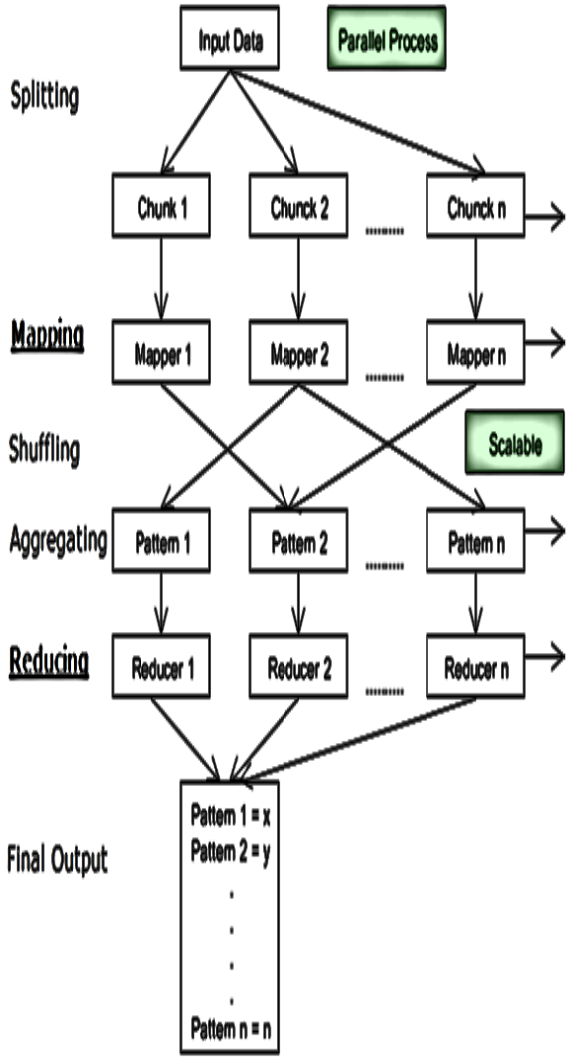


Fig. 3.Flowchart - Working of Map Reduce Technology

Algorithm for Map-Reduce

- The input data can be divided into n number of chunks depending upon the amount of data and processing capacity of individual unit.
- Next, it is passed to the mapper functions. All the chunks are processed simultaneously at the same time, which embraces the parallel processing of data.
- After that, shuffling happens which leads to aggregation of similar patterns.
- Finally, reducers combine them all to get a consolidated output as per the logic.

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].

To understand the concept, consider the example of how Facebook manages its section “People you may know”. Consider five persons – A, B, C, D and E.

Assume the persons are stored as Person → [List of Friends], our friends list is then:

A -> B C D

B -> A C D E

C -> A B D E

D -> A B C E

E -> B C D

Each line will be an argument to a mapper. For every friend in the list of friends, the mapper will output a key-value pair. The key will be a friend along with the person. The value will be the list of friends. The key will be sorted so that the friends are in order, causing all pairs of friends to go to the same reducer. This is hard to explain with text, so let's just do it and see if you can see the pattern. After all the mappers are done running, you'll have a list like this:

For map (A→B C D):

(A B) →B C D

(A C) → B C D

(A D) → B C D

For map (B →A C D E) : (Note that A comes before B in the key)

(A B) → A C D E

(B C) → A C D E

(B D) → A C D E

978-93-85670-72-5 © 2016 (RTCSIT)

(B E) → A C D E

For map(C → A B D E):

(A C) → A B D E

(B C) → A B D E

(C D) → A B D E

(C E) → A B D E

For map (D → A B C E):

(A D) → A B C E

(B D) → A B C E

(C D) → A B C E

(D E) → A B C E

And finally for map (E -> B C D):

(B E) → B C D

(C E) → B C D

(D E) → B C D

Before we send these key-value pairs to the reducers, we group them by their keys and get:

(A B) → (A C D E) (B C D)

(A C) → (A B D E) (B C D)

(A D) → (A B C E) (B C D)

(B C) → (A B D E) (A C D E)

(B D) → (A B C E) (A C D E)

(B E) → (A C D E) (B C D)

(C D) → (A B C E) (A B D E)

(C E) → (A B D E) (B C D)

(D E) → (A B C E) (B C D)

Each line will be passed as an argument to a reducer. The reduce function will simply intersect the lists of values and output the same key with the result of the intersection. For example, reduce ((A B) -> (A C D E) (B C D)) will output (A B): (C D) and means that friends A and B have C and D as common friends.

The result after reduction is:

(A B) -> (C D)

(A C) -> (B D)

(A D) -> (B C)

(B C) -> (A D E)

(B D) -> (A C E)

(B E) -> (C D)

(C D) -> (A B E)

(C E) -> (B D)

(D E) -> (B C)

Now when D visits B's profile, we can quickly look up (B D) and see that they have three friends in common, (A C E).

This is how Facebook analyses millions of user accounts created on it and finds out that to which what people should be shown in there "people you may know section".

## 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].

It can be concluded that big data is all set to play a major role in any national elections to be conducted in future. 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.

## 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.


978-93-85670-72-5 © 2016 (RTCSIT)

[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.

[16] Dr.Gagandeep Jagdev et. al.,"Big Data sets novel drift for contesting elections in India",Wireless Communication", at National Conference RTEST-2016,ISBN: 978-93-84935-82-5.