Data Mining and Data Pre-processing for Big Data

Ashish R. Jagdale, Kavita V. Sonawane, Shamsuddin S. Khan

Abstract— Big Data is a term which is used to describe massive amount of data generating from digital sources or the internet usually characterized by 3 V's i.e. Volume, Velocity and Variety. From the past few years data is exponentially growing due to the use of connected devices such as smart phone's, tablets, laptops and desktop computer. Moreover E-commerce which is also known as online market, internet services and social networking sites are generating tremendous user data in the form of documents, emails and web pages. This generated data volume is so vast and overwhelming which makes complex to process and analyze using traditional software systems consuming more time. This paper presents a pre-processing algorithm to extract real time user accessed data from windows operating system environment and an approach from Apache's Hadoop Distributed File System (HDFS) framework using Map Reduce functionality to mine and analyze this large dataset. The ability to mine and analyze Big Data gives organization richer and deeper insights into business patterns and trends. The performance metrics of the proposed system can be evaluated on the basis of execution time, data heterogeneity, scalability, flexibility and mining algorithm used.

Index Terms— Apache Hadoop, Big Data, Data Analysis, Datasets, Data Pre-processing, Data Mining, HDFS, MapReduce

1. INTRODUCTION

A dramatic increase in our ability to collect data from various sources such as connected devices, sensors, log records or click-stream in web exploring and other applications have been witnessed that too in different formats whether structured or unstructured. This has outpaced our capability to store, understand, process and analyze this large datasets. Considering the internet data, the web pages indexed by Google were around one million in 1998 and reached to one billion in 2000 and now have already exceeded in trillions. The reason for this expansion is mainly due to evolution of social networking websites such as Facebook, Twitter, LinkedIn, MySpace, etc. allowing users to create content freely thereby expanding the volume of data over the internet. From this heap of Big Data, information must be discovered and converted to knowledge to help improve our decision making and make this world a better place.

Nowadays, the quantity of data that is created every two days is estimated to be 5 Exabyte's. This amount of data is similar to the amount of data created from the dawn of time up until 2003. Moreover, it was estimated that 2007 was the first year in which it was not possible to store all the data that we are producing. This massive amount of data opens new challenging discovery tasks [1].

The term 'Big Data' appeared for first time in 1998 in a Silicon Graphics (SGI) slide deck by John Mashey with the title of "Big Data and the NextWave of InfraStress" [7]. Each day Google has more than 1 billion queries per day, Twitter has more than 250 million tweets per day, Facebook has more than 800 million updates per day, and YouTube has more than

4 billion views per day. The data produced nowadays is estimated in the order of Zettabytes, and it is growing around 40% every year. A new large source of data is going to be generated from mobile devices, and big companies such as Google, Apple, Facebook, Yahoo, Twitter is starting to look carefully to this data to find useful patterns to improve user experience. Alex 'Sandy' Pentland in his 'Human Dynamics Laboratory' at MIT, is doing research in finding patterns in mobile data about what users do, and not in what people say they do [2]. We need new algorithms and new tools to deal with all of this data.

Doug Laney [3] was the first one in talking about 3 V's in Big Data management:

Volume: There is more data than ever before; its size continues increasing, but not the percent of data that our tools can process.

Variety: There are many different types of data, as text, sensor data, audio, video, graph and more.

Velocity: Data is arriving continuously as streams of data, and we are interested in obtaining useful information from it in real time.

Nowadays, there are 2 more V's:

Variability: There are changes in the structure of the data and how users want to interpret that data

Value: Business value that gives organization a compelling advantage, due to the ability of making decisions based in answering questions that were previously considered beyond reach [4].

Data mining is the process of finding useful information and deriving patterns by using certain data mining algorithms. It uses the Knowledge Discovery in Database (KDD) process which involves data cleaning, data integration, data selection and data transformation. It is the pre-processing step which is done prior to data mining. Developing the organizational skill to mine and process big data to perform predictive and prescriptive analytics will be a key driver of performance in the future, enabling to make better decisions, increase business velocity, accelerate the pace of innovation, discover and tap new markets. Advances in parallel computing now make it possible to handle big data, to the point where it is now becoming standard practice to capture and store information well before its value is completely understood, and tackle many business problems that have been previously too large to handle [5].

Data Mining Challenges with Big Data [6]:

Heterogeneity and Incompleteness: When humans consume information, a great deal of heterogeneity is comfortably tolerated. In fact, the nuance and richness of natural language can provide valuable depth.

Scalability: Of course, the first thing anyone thinks of with Big Data is its size. After all, the word "big" is there in the very name. Managing large and rapidly increasing volumes of data has been a challenging issue for many decades.

Timeliness: The larger the data set to be processed, the longer it will take to analyze. The design of a system that effectively deals with size is likely also to result in a system that can process a given size of data set faster.

Privacy: The privacy of data is another huge concern, and one that increases in the context of Big Data.

The structure of this paper is organized as follows: In section 2, the paper introduces related work by different authors. Section 3 includes the proposed work and overview of the entire system. Section 4 includes experimental setup & results followed by performance analysis in section 5 and finally Conclusion.

2. LITERATURE SURVEY

In [8] authors present a HACE (Heterogeneous, Autonomous, Complex and Evolving) theorem that characterizes the features of the Big Data revolution, and proposes a Big Data processing model, from the data mining perspective. This data-driven model involves demand-driven aggregation of information sources, mining and analysis, user interest modeling, and security and privacy considerations. They analyze the challenging issues in the data-driven model and also in the Big Data revolution.

In [9] authors present the KEOPS data mining methodology centered on domain knowledge integration. In this paper, the authors focuses first on the pre-processing steps of business understanding and data understanding in order to build an ontology driven information system (ODIS). Then they show how the knowledge base is used for the post-processing step of model interpretation. Detailed the role of the ontology and define a part-way interestingness measure that integrates both objective and subjective criteria in order to evaluate model relevance according to expert knowledge.

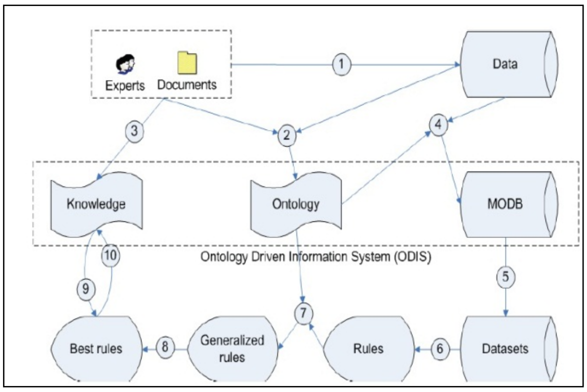


Fig 1. KEOPS Methodology

In [10] authors present insight about Big Data mining infrastructures and the experience of doing analytics at Twitter. Two major topics are discussed here. First, schemas play an important role in helping data scientists understand petabyte-scale data stores, but they are insufficient to provide an overall "big picture" of the data available to generate insights. Second, a major challenge in building data analytics platforms stems from the heterogeneity of the various components that must be integrated together into production workflows- refer to this as "plumbing". The goal of this paper is to share experiences at Twitter for academic researchers to provide a broader context for data mining in production environments, pointing out opportunities for future work.

3. Proposed Work

The proposed system as shown in the block diagram below is executed under the Apache Hadoop framework environment which collects user accessed data from windows operating system, pre-processes the data generating dataset and then finally data mining algorithm followed by analysis.

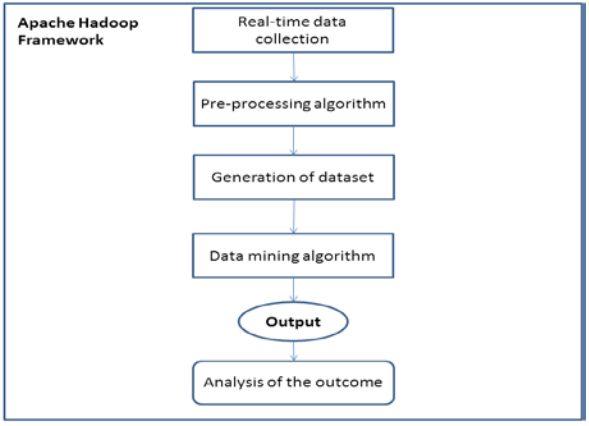


Fig 2. Block diagram of proposed system

The important phases as part of the proposed system can be divided into two major category:

A] Data Pre-processing

1. Real-time data collection

Here the data from individual user machine which is generated by accessing different files and folders is collected with the help of pre-processing algorithm to extract the relevant information. In our experimental setup we have used data collected from the windows operating system machine where a user is performing certain activity like accessing certain files and folders

2. Pre-processing algorithm

From the data collected, pre-processing algorithm transforms the data to a specific format. This means extracting, cleaning and loading of appropriate data to a text file takes place. This file is also known as a log file. We have used java algorithm for pre-processing to extract and transform the data to given log file format.

3. Generation of dataset

Dataset means a collection of data. Here relevant data are grouped together to form a dataset. In this experimental setup, the dataset includes the timestamp, type & name of file/directory accessed. This dataset which is basically a text file is given as an input to the data mining algorithm.

B] Data mining and Analysis

1. Data mining algorithm

Data mining algorithm is applied to the generated dataset or the text/log file to track the number of files/directory accessed in different time period. This is one of the most core steps in the process. Data mining algorithm written in java using Apache Hadoop HDFS and MapReduce functions are used for mining and analyzing.

2. Analysis of the outcome

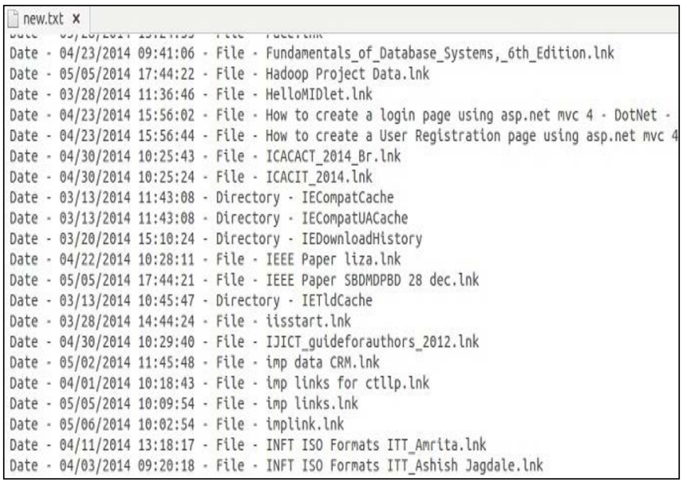
Output is analyzed by tracking files/directory accessed by month/year and giving different input criteria. Also same output can be mapped in Microsoft Power Business Intelligence (BI) which is a powerful BI tool to represent the output in the form of charts and graphs for visualization and better insights.

4. EXPERIMENTAL SETUP & RESULTS

Desktop machine with Linux operating system and Apache Hadoop packages as well as java installed.
For evaluating the output of the proposed system, we have used semi-structured data which is basically a text file containing thousands of records.

Different output results obtained after performing the proposed steps are as shown below:

1. **Input**: Pre-processing java algorithm

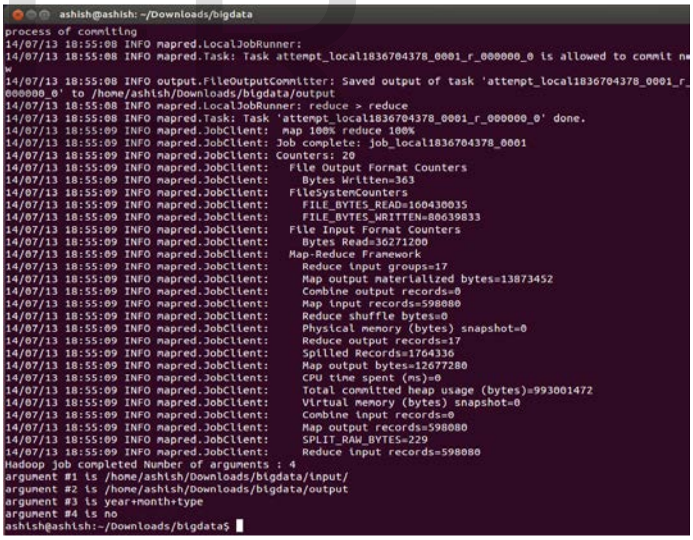


```
Date - 04/23/2014 09:41:06 - File - Fundamentals_of_Database_Systems,_6th_Edition.lnk
Date - 05/05/2014 17:44:22 - File - Hadoop Project Data.lnk
Date - 03/28/2014 11:36:46 - File - HelloMIDlet.lnk
Date - 04/23/2014 15:56:02 - File - How to create a login page using asp.net mvc 4 - DotNet -
Date - 04/23/2014 15:56:44 - File - How to create a User Registration page using asp.net mvc 4
Date - 04/30/2014 10:25:43 - File - ICACACT_2014_Br.lnk
Date - 04/30/2014 10:25:24 - File - ICACIT_2014.lnk
Date - 03/13/2014 11:43:08 - Directory - IECompatCache
Date - 03/13/2014 11:43:08 - Directory - IECompatUACache
Date - 03/20/2014 15:10:24 - Directory - IEDownloadHistory
Date - 04/22/2014 10:28:11 - File - IEEE Paper liza.lnk
Date - 05/05/2014 17:44:21 - File - IEEE Paper SBDMDPBD 28 dec.lnk
Date - 03/13/2014 10:45:47 - Directory - IETldCache
Date - 03/28/2014 14:44:24 - File - iisstart.lnk
Date - 04/30/2014 10:29:40 - File - IJICT_guideforauthors_2012.lnk
Date - 05/02/2014 11:45:48 - File - imp data CRM.lnk
Date - 04/01/2014 10:18:43 - File - imp links for ctllp.lnk
Date - 05/05/2014 10:09:54 - File - imp links.lnk
Date - 05/06/2014 10:02:54 - File - implink.lnk
Date - 04/11/2014 13:18:17 - File - INFT ISO Formats ITT_Amrita.lnk
Date - 04/03/2014 09:20:18 - File - INFT ISO Formats ITT_Ashish Jagdale.lnk
```

Fig 3. Output of Java Pre-processing algorithm

Output: It's a log file with timestamp, type and name of the file/directory

2. **Input**: Log file given to Hadoop MapReduce data mining algorithm

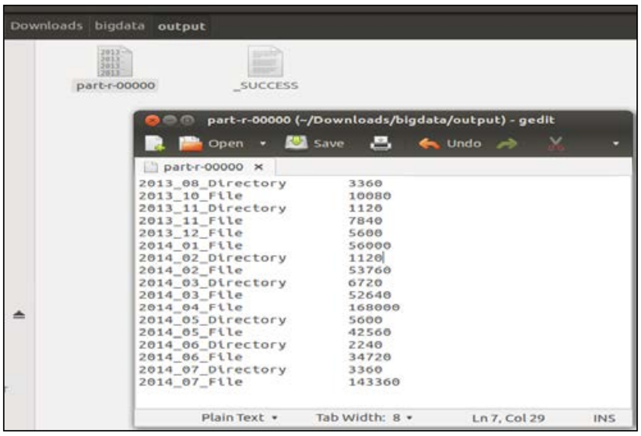


```
process of commiting
14/07/13 18:55:08 INFO mapred.LocalJobRunner:
14/07/13 18:55:08 INFO mapred.Task: Task attempt_local1836704378_0001_r_000000_0 is allowed to commit ne
w
14/07/13 18:55:08 INFO output.FileOutputCommitter: Saved output of task 'attempt_local1836704378_0001_r_
000000_0' to /home/ashish/Downloads/bigdata/output
14/07/13 18:55:08 INFO mapred.LocalJobRunner: reduce > reduce
14/07/13 18:55:08 INFO mapred.Task: Task 'attempt_local1836704378_0001_r_000000_0' done.
14/07/13 18:55:09 INFO mapred.JobClient:  map 100% reduce 100%
14/07/13 18:55:09 INFO mapred.JobClient: Job complete: job_local1836704378_0001
14/07/13 18:55:09 INFO mapred.JobClient: Counters: 20
14/07/13 18:55:09 INFO mapred.JobClient:   File Output Format Counters
14/07/13 18:55:09 INFO mapred.JobClient:     Bytes Written=363
14/07/13 18:55:09 INFO mapred.JobClient:   FileSystemCounters
14/07/13 18:55:09 INFO mapred.JobClient:     FILE_BYTES_READ=160430035
14/07/13 18:55:09 INFO mapred.JobClient:     FILE_BYTES_WRITTEN=80639833
14/07/13 18:55:09 INFO mapred.JobClient:   File Input Format Counters
14/07/13 18:55:09 INFO mapred.JobClient:     Bytes Read=36271200
14/07/13 18:55:09 INFO mapred.JobClient:   Map-Reduce Framework
14/07/13 18:55:09 INFO mapred.JobClient:     Reduce input groups=17
14/07/13 18:55:09 INFO mapred.JobClient:     Map output materialized bytes=13873452
14/07/13 18:55:09 INFO mapred.JobClient:     Combine output records=0
14/07/13 18:55:09 INFO mapred.JobClient:     Map input records=598080
14/07/13 18:55:09 INFO mapred.JobClient:     Reduce shuffle bytes=0
14/07/13 18:55:09 INFO mapred.JobClient:     Physical memory (bytes) snapshot=0
14/07/13 18:55:09 INFO mapred.JobClient:     Reduce output records=17
14/07/13 18:55:09 INFO mapred.JobClient:     Spilled Records=1764336
14/07/13 18:55:09 INFO mapred.JobClient:     Map output bytes=12677280
14/07/13 18:55:09 INFO mapred.JobClient:     CPU time spent (ms)=0
14/07/13 18:55:09 INFO mapred.JobClient:     Total committed heap usage (bytes)=993001472
14/07/13 18:55:09 INFO mapred.JobClient:     Virtual memory (bytes) snapshot=0
14/07/13 18:55:09 INFO mapred.JobClient:     Combine input records=0
14/07/13 18:55:09 INFO mapred.JobClient:     Map output records=598080
14/07/13 18:55:09 INFO mapred.JobClient:     SPLIT_RAW_BYTES=229
14/07/13 18:55:09 INFO mapred.JobClient:     Reduce input records=598080
Hadoop job completed Number of arguments : 4
argument #1 is /home/ashish/Downloads/bigdata/input/
argument #2 is /home/ashish/Downloads/bigdata/output
argument #3 is year+month+type
argument #4 is no
ashish@ashish:~/Downloads/bigdata$
```

Fig 4. MapReduce processing

Output: We can observe that different MapReduce Jobs run simultaneously to get the output within few seconds.

3. **Input:** Applying analytics criteria to data mining algorithm

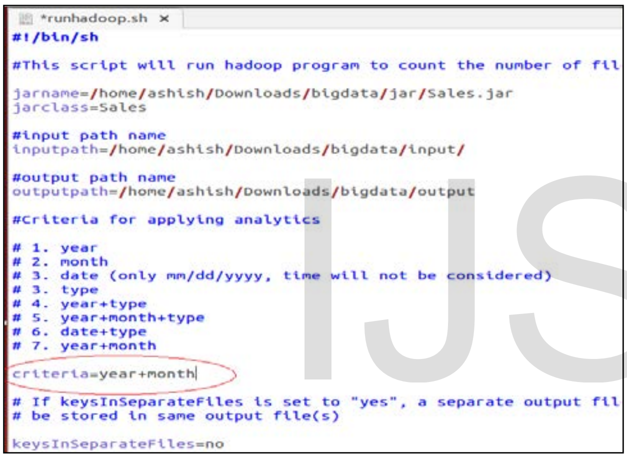


```
2013_08_Directory	3360
2013_10_File		10080
2013_11_Directory	1120
2013_11_File		7840
2013_12_File		5600
2014_01_File		56000
2014_02_Directory	1120
2014_02_File		53760
2014_03_Directory	6720
2014_03_File		52640
2014_04_File		168000
2014_05_Directory	5600
2014_05_File		42560
2014_06_Directory	2240
2014_06_File		34720
2014_07_Directory	3360
2014_07_File		143360
```

Fig 5. Output of Hadoop MapReduce mining algorithm

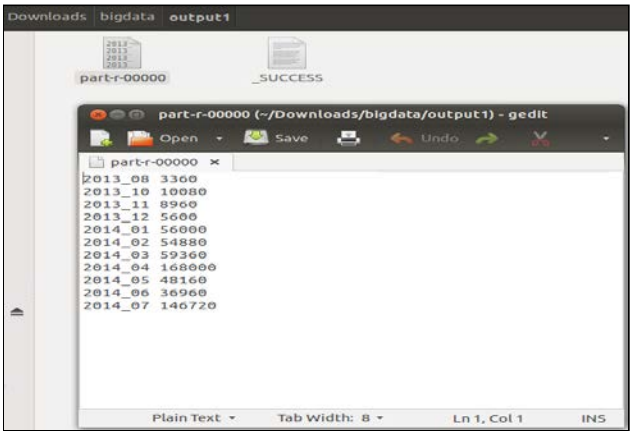
Output: Number of files and directories accessed by year and month

4. **Input**: Changing the analytics criteria



```
#!/bin/sh

#This script will run hadoop program to count the number of fil

jarname=/home/ashish/Downloads/bigdata/jar/Sales.jar
jarclass=Sales

#input path name
inputpath=/home/ashish/Downloads/bigdata/input/

#output path name
outputpath=/home/ashish/Downloads/bigdata/output

#Criteria for applying analytics

# 1. year
# 2. month
# 3. date (only mm/dd/yyyy, time will not be considered)
# 3. type
# 4. year+type
# 5. year+month+type
# 6. date+type
# 7. year+month

criteria=year+month

# If keysInSeparateFiles is set to "yes", a separate output fil
# be stored in same output file(s)

keysInSeparateFiles=no
```

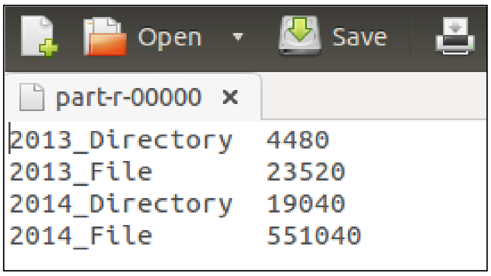
Fig 6. Changing to year + month as analytics criteria



```
2013_08 3360
2013_10 10080
2013_11 8960
2013_12 5600
2014_01 56000
2014_02 54880
2014_03 59360
2014_04 168000
2014_05 48160
2014_06 36960
2014_07 146720
```

Fig 7. Output after changing the criteria

Similarly, we can get different results on changing the condition for analysis



```
2013_Directory	4480
2013_File	23520
2014_Directory	19040
2014_File	551040
```

Fig 8. year + type as analytics criteria

5. **Input**: Applying BI tool for visualization

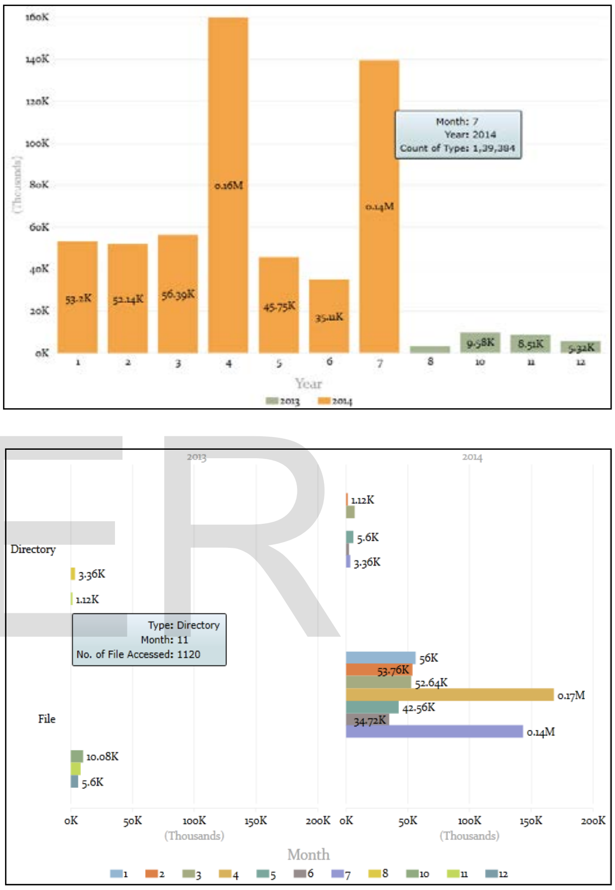


Fig 9. Visualization of output using Power BI

Output: For further analysis, we used Microsoft Power BI tool which consists of power query, power pivot & power view to represent data in different form of visualization.

5. PERFORMANCE ANALYSIS

Performance analysis of the proposed system can be evaluated on the basis of different factors such as the execution time of the algorithm, data scalability and flexibility and data heterogeneity which are discussed below.

Execution Time:

Time taken for performing data mining operation and getting the desired output quickly and efficiently will be computed based on the efficiency and time complexity of the algorithm used. If traditional data mining algorithms are applied to large datasets containing billions of records, processing the algorithm consumes more time as compared to algorithm which uses Hadoop Map Reduce functions. Also using traditional systems it becomes difficult to handle such large datasets.

Scalability & Flexibility:
Even on scaling data from thousands to billions, it does not scale down the performance as seen in our system. Hadoop processes terabytes of data without any issues which makes it most scalable. Also Hadoop can handle workloads pretty quickly due to high bandwidth which makes it quite flexible.

Data heterogeneity:
Different types of data coming from different sources and of different types i.e. structured, semi structured and unstructured. Ideally system should accept any kind of heterogeneous data, preprocess it and produce the desired information for the user within minimum possible duration. The given system will check with respect to different types of heterogeneous data and its performance will be evaluated.

6. Conclusion

Mining and analyzing Big Data have shown to be a challenging yet very compelling task. We have seen that Apache Hadoop HDFS and Map Reduce play an important role in the processing, handling and analyzing of large datasets also known as Big Data. Using this methodology, tracking user activities and later analyzing it helps in gaining better insights, understanding and quick decision making. Hadoop plays a vital role in the process in terms of execution time, scalability, flexibility and cost. Also it enhances the overall performance of the process.

Here we have used a single node to demonstrate this concept. Similar approach can be applied to nodes distributed geographically in different locations by parallel running of MapReduce programming model. For our experimental purpose, we have tested the algorithm for approximately 5, 80,000 records which was a semi-structured dataset i.e. data embedded in text file in relational format. Hadoop runs more efficiently and quickly for billions of record containing Exabyte's of data having variety of dataset i.e. structured, semi-structured or unstructured.

Our ability to handle many Exabyte's of data across different application areas in the near future will be crucially dependent on the existence of a rich variety of datasets, techniques and software frameworks. It is clear that stream data mining offers many challenges and equally many opportunities as the quantity of data generated in real time continues to grow.

ACKNOWLEDGEMENT

The authors take this opportunity to thank all the researchers from this domain of Big Data and Data Mining for their immense knowledge and kind support throughout the work. Also would like to thank our institute for their resources and constant inspiration. Lastly thanks to our family and friends for encouraging us to make this a success.

REFERENCES

[1] A. Bifet, "Mining Big Data in Real Time", Informatica 37, pp.15-20, 2013.

[2] A. Petland. Reinventing society in the wake of big data. Edge.org, http://www.edge.org/conversation/reinventing-society-in-the-wake-of-big-data, 2012

[3] D. Laney. 3-D Data Management: Controlling Data Volume, Velocity and Variety. META Group Research Note, February 6, 2001.

[4] W. Fan, A Bifet, "Mining big data: current status, and forecast to the future", ACM SIGKDD Explorations Newsletter, Vol. 14, pp.1-5, 2013.

[5] Mining Big Data in the Enterprise for Better Business Intelligence, Intel White Paper, July 2012

[6] R Anand, J David, "Mining of massive datasets", Cambridge University Press, 2012 E Bertino et al. "Challenges and Opportunities with Big Data", 2011.

[7] F. Diebold. On the Origin(s) and Development of the Term "Big Data". Pier working paper archive, Penn Institute for Economic Research, Department of Economics, University of Pennsylvania, 2012.

[8] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding, "Data Mining with Big Data", (In Press) IEEE Transactions on Knowledge and Data Engineering, 2013.

[9] L. Brisson and M. Collard, "How to Semantically Enhance a Data Mining Process?" Enterprise Information Systems, Springer Berlin Heidelberg, Vol. 19, pp. 103–116, April 2010.

[10] J. Lin, D. Ryaboy, "Scaling big data mining infrastructure: the twitter experience", ACM SIGKDD Explorations Newsletter, Vol 14, pp.6-19, 2013.

ABOUT THE AUTHORS

Mr. Ashish R. Jagdale has received Bachelor's Degree in Information Technology from Mumbai University in 2012 and is currently pursuing Master's in Computer Engineering from Mumbai University. This is the first paper of his research work and area of interest includes learning and working on new and latest technologies like Big Data, Power Business Intelligence, Data Mining, Database's and Web Technologies.

Email: ashishjagdale@yahoo.com

Ms. Kavita V. Sonawane has received M.E (Computer Engineering) degree from Mumbai University in 2008, currently Pursuing Ph.D. from Mukesh Patel School of Technology, Management and Engineering, SVKM's NMIMS University, Vile-Parle (w), Mumbai, INDIA. She has more than 12 years of experience in teaching. Currently working as a Assistant professor in Department of Computer Engineering at St. Francis Institute of Technology Mumbai. Her area of interest is Image Processing, Data structures and Computer Architecture. She has 28 papers in National/ International conferences / Journals to her credit. She is the member of ISTE.

Email: kavitavinaysonawane@gmail.com

Mr. Shamsuddin S. Khan is currently Assistant Professor at St. Francis Institute of Technology, Mumbai in Computer Engineering department. His areas of interests include artificial intelligence, neural network, database systems, data mining and distributed computing and have published several research papers.

Email: Shams21980@gmail.com