Analysis and Classification for Big Data Computing Techniques: A Retrospective Study

Shriya kumari[1],Dr. K Srujan Raju[2] , P.Satyavathi[3]

[1]Department of CSE, CMRTC, JNTU, Hyderabad
Email: shriya.thakur@gmail.com
[2]Department of CSE, CMRTC, JNTU, Hyderabad
Email: drksrujanraju@gmail.com
[3]Department of CSE, CMRTC, JNTU, Hyderabad
Email: satya.potlapalli@gmail.com

***Abstract*—*Big data, defined as a massive dataset based of the five V's volume, velocity, variety, veracity and value[1]. The world is connecting and producing a huge amount of data from applications like social media, e commerce, internet, sensors, health care, informatics, net banking etc. There are few major challenges of these datasets: data collection, data analysis, data enrichment, data distribution, data security and privacy. These datasets play a vital role in any of the application areas. There are various tools and techniques which are developed to manage Big data. There are mainly two processing techniques to process these datasets: Batch processing and Non batch processing. Different application areas use different tools and techniques depending upon the efficiency and capability of managing different forms of data produced by them. In this paper we have suggested some of the efficient tools and techniques for managing the data produced by a particular application area by analyzing and comparing all the possible techniques to manage Big data.*

***Keywords*— Big data, Datasets, Hadoop, MapReduce, computing techniques.**

I. INTRODUCTION

Big data is a collection of different forms of data such as Structured, Unstructured and Semi structured datasets. Big data is generated by all the applications around us like social media and internet which is transmitted by mobile devices and different sensors. Big data is defined by gartner as “ A high volume , high velocity and/or high variety information assets that demand cost effective, innovative forms of information processing that enable enhanced insight, decision making and process automation.” The data is growing exponentially and has reached from bits to bytes and now bytes to zetabytes in last 2 to 3 years. It is being produced by several application areas such as Social Networking, Natural Language Processing, Digital Forensics, Bio Informatics, Sensor Networks, GIS, Medical Sciences, Weather Forecasting, Human Behavior Monitoring, Cloud Control System, Multimedia, Net Banking etc. As big data is being produced in different forms it is imperative to use different tools and techniques to process these datasets. The two major processing techniques often used to manage Big Data is: Batch processing and Non Batch Processing.

II. BIG DATA TOOLS AND TECHNIQUES

Big data is processed using two main processing techniques:

i) Batch Processing
ii) Non Batch Processing

2.1 Batch Processing: The data is accumulated over a certain period of time and then it is processed to produce desired output. MapReduce framework is used in batch processing. MapReduce is a programming model used for distributed computing. It is done in 2 computation stages: Map and Reduce. Another framework which is often used in batch processing is Hadoop which is an open source framework used in distributed environment. It uses two components: HDFS and MapReduce to store and process Big data[6] .

2.2 **Non Batch Processing:** In non batch processing, the input of data is continuous and processed in small period of time. There are two categories of non batch processing:

2.2.1 Real time Big data processing: It is done by two processes such as

A. In memory computing : It is a technique which is used to minimize the computation time of MapReduce. It is used to efficiently minimize the execution time of jobs. The frameworks used in this technique are Apache Spark, GridGain, XAP etc.

B. Real time queries over Big data: It is an optimized technique used for real time input queries to make them respond in less than a second. Cloudera Impala, Apache Drill are some of the techniques used for the above purpose.

2.2.2 Stream processing: Stream processing uses two prominently used frameworks such as Storm and S4[7].

There are some more tools used in Big data scenarios such as:

- NoSQL : DatabasesMongoDB, CouchDB, Cassandra, Redis, BigTable, Hbase, Hypertable, Voldemort, Riak, ZooKeeper
- MapReduce : Hadoop, Hive, Pig, Cascading, Cascalog, mrjob, Caffeine, S4, MapR, Acunu, Flume, Kafka, Azkaban, Oozie, Greenplum
- Storage : S3, Hadoop Distributed File System
- Servers : EC2, Google App Engine, Elastic, Beanstalk, Heroku
- Processing : R, Yahoo! Pipes, Mechanical Turk, Solr/Lucene, ElasticSearch, Datameer, BigSheets, Tinkerpop.

III. BIG DATA APPLICATION AREAS

Big data is being generated in most of the application areas. Let us have a brief idea about a) how the huge amount of data is produced , b) what are the challenges each application area faces and c) how it is overcoming the defined challenge.

The application areas to be discussed are as follows:

3.1 Natural Language Processing: Natural Language Processing is the study of designing machines or programs that can understand verbal and written communications. Extracting meaningful information from large volume of unstructured human language is a Big Data problem. The unstructured data like voice call, emails, text messages etc. is increasing exponentially and need to be analyzed accurately, which would lead to more insights and better predictive models. Hadoop framework [8] is able to solve the NLP problems but Hadoop is a Batch processing technique, so it requires complete input set of each NLP module. It is very difficult or not practical to get complete input set of each NLP module at the starting stage of execution. Apache SPARK which is built on top of Hadoop and is also an extension of Hadoop. It is used to overcome this problem by executing interactive and streamed queries [9].

3.2 Human Behavior Modeling: Nowadays, human life is data centric. Emotions and sentiments, relationships and interactions, speech, offline and back-office activities, culture etc. generates huge set of data. By analyzing Big Data of human behavior, it can lead to a detailed insight and precise predictive models [10]. Big Data technology Apache Hadoop is used to process these huge set of data. For better result Stream based processing Storm can be used.

3.3 GIS : It is a powerful system, which is designed for making better decision about location. It includes storing, manipulating, managing, collecting, selecting and sorting of geographical data [11]. Big Data will enable a number of transformative societal applications. Societal applications in the context of understanding climate change, next-generation routing services and disaster response. Apache Hadoop is able to process this large amount of data using MapReduce and HDFS. Apache Spark, which has less “Reduce” step, will need to be evaluated with iterative GIS Workloads [12].

3.4 **Weather Forecasting:** Human has tried to get a better understanding of weather and forecast. Nowadays, satellite sensors and other resources are used by weather forecasting system to help general people for accurate predictions of weather. The volume of environmental data is increasing exponentially. So, there are needs of efficient Big Data techniques to manage, store and process this data.The main advantage of Big Data is that it compares separate datasets to obtain associated observations. It enables better risk management to improve performance in the organizations. Companies work with Big Data and predictive analysis parallely to focus on real time forecast using the growing data[17]. Apache Hadoop MapReduce framework is used to analyze huge dataset of weather forecasting [18].

3.5 Multimedia: In today’s digital world social networking applications are being used everywhere. They have multiplied fast and their uses have grown exponentially. Multimedia has become one of the largest application area that produces enormous amount of data at faster rate. The multimedia data include audio, video, texts, images, graphic objects, animation sequence etc. The multimedia resources have grown so fast that it has brought the need for Big Data processing technique.Multimedia Big Data is in memory processing that is, the data is processed in the memory instead of on hard disks, significantly reducing the processing latency. So, Apache SPARK is used to process Big Data produced by multimedia [19].

3.6 Sensor Networks: In order to monitor various purposes like vital signs, gait patterns, balance and fall, daily activities etc., or sensing, communicating and processing various physiological parameters, Body Sensor Networks are used. Billions of data stream comes from large scale sensor networks challenges the traditional approach of data management related content capturing techniques. The benefits to analyze Big Data generated by body sensor networks are; in making energy-aware network protocol and energy gathering techniques. It is also beneficial in data compression, on- node processing, power transceiver etc [20].Hadoop framework is able to solve the body sensor network problems but for interactive queries Apache SPARK [21] (an extension of Hadoop) is used to overcome this problem by executing interactive and streamed queries.

3.7 Bio Informatics: Bio-informatics [13] is the study of understanding the molecular mechanism of life on earth by analyzing Genomic information. Genomic information includes genomic sequencing and expressed gene sequencing. Sequencing mechanism has improved these days, which leads the volume of sequence of data being produced to exceed the capabilities of traditional method of database model. Big Data possess a great impact on the bio-informatics field and a researcher in this field faces many difficulties in using biological Big Data to extract valuable information from the data easily thereby enhancing further advancement in the decision-making process related to diverse biological process, diseases and disorders. A tool Hadoop MapReduce platform, such as BioPig [14] and Crossbow [15] has been developed for sequence analysis.

3.8 Digital Forensics: Digital Forensic is a branch of Applied Science which deals with the identification, collection, organization, preservation and presentation of evidence data which is permissible in court room [1]. Recently Network forensics has been evolved from digital forensics, which deals with collection of evidences from Internet or local intranets [2]. Digital or Network Forensics helps the security and forensic investigators to analyze evidences collected from Internet. This type of forensic analysis also deals with the cloud and other distributed environments. The process of Digital Forensic comprises of following main sub-processes [3]:

- Identification
- Collection
- Organization
- Preservation
- Presentation

Big data is facing a lot of challenges in digital forensics. Traditional digital forensic tools are not proficient to handle big data in order to identify and analyze the evidences effortlessly [4].The challenges and opportunities of big data in several sub processes of digital forensics are as follows:

a) Identification: The challenge is to find accurate evidence from big data due to its huge volume and high velocity.

b) Collection: The major challenge is the collection of erroneous or worthless data.

c) Organization: Organizing big data evidences is a biggest challenge faced by forensic investigators. Its nearly impossible to review and organize evidences manually.

d) Preservation: Due to unavailability of appropriate tool forensic investigators often face the problem for preserving the evidences to maintain their security and integrity for the future use.

e) Presentation: The biggest challenge faced by investigators in the presentation process is to prepare the final report on the basis of available evidences. Limited knowledge of judges on big data also possess a lot of difficulty in explaining the entire process.

Big data tools and techniques can be used to correlate distinct criminal data set, for identification of Cyber Criminals, to identify mental state of a criminal, for identification of Phishing Email and to send alerts on accessing fake Social media accounts.

3.9 Health Care: Data is growing faster than medical science can consume it. The unstructured data generated by medical science is huge (around 80% of the total relevant medical data). From genetic to genomic, internal imaging to motion picture, treatment to life course assessment etc. are all Big Data. There is a need of Big Data technology to capture all the information of every patient. Big Data technique can be used to evaluate data generated from routine care of entire patient. By statistically analyzing Big Data effectively, it will be beneficial for medical science to easily diagnose and effectively cure the ailments [16].The easy and efficient analysis of Big Data benefits such as: (i) detection of diseases can be done earlier; (ii) identification of health care deception can be done more quickly. Open source platforms Hadoop, MapReduce is used for Big Data analytics in Medical Science [11].

3.10 **Banking Sector:** Banking sector generates a huge amount of structured to unstructured data. Traditional data tools are not efficient enough to store, process and analyze such data. Banking is a financial service organization whose goal is to aquire new customers and retain the existing ones. Therefore other than storage and retrieval there are several challenges[22] in banking where big data analytics is being used such as:

a) Sentiment analytics: It monitors what customers say to increase marketing success, identify key customers to boost word-of-mouth marketing and examine customer feedback to improve products and services.

b) Customer 360: It identifies the customer profile, understands the product engagement of the customer and detect when a customer is about to leave.

c) Customer Segmentation : It designs targeted marketing programs, creates loyalty programs based on card usage habits, optimize pricing strategy and build relationships with valuable customers.

d) Next best offer: It enhances loyalty through targeted offers, increases product propensity and product bundling to uplift revenue.

e) Channel Journey: It provides more relevant content in the preferred channel, recognizes multi-channel behaviors that lead to sales and measures marketing effectiveness across channels. Hadoop is being used to handle the above usecases. Hive, cloudera, spark is also being used along with hadoop framework.

IV. COMPARATIVE ANALYSIS OF BIG DATA TOOLS AND TECHNIQUES IN VARIOUS APPLICATION AREAS

Table1 is a comparative analysis of Big data tools and processing techniques in various application areas. According to the challenges occurred in the application areas mentioned in above section and the usability of the big data tools and techniques we have put Y against the tool if it is used in the given application area to overcome the mentioned challenges.

Table 1: Comparative analysis of Big data tools and processing techniques.

S No.	Tools & Techniques	Banking Sector	Health Care	Natural Language Processing	Human Behavior Modeling	GIS	Digital Forensics	Sensor Networks	Bio-Informatics	Weather forecasting	Multimedia
1.	Hadoop	Y	Y	Y	Y	Y	Y	Y	Y	Y	
2.	Spark	Y		Y		Y	Y	Y			Y
3.	Pig	Y	Y		Y		Y		Y		
4.	Hive	Y	Y	Y			Y	Y		Y	
5.	Storm				Y			Y			Y

Hadoop : Hadoop is an open-source software framework that allows distributed processing of large data sets across clusters of computers using simple programming models. Hadoop is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Hadoop is used in most of the application areas due to its linear scalability, flexibility, low cost, high computing power and fault tolerance features[23]. It uses distributed computing model for processing big data. Hadoop in banking sector is used for sentiment analytics, predictive analytics etc., to understand the customer, acquire new customers and retain the existing ones. For example : Let us use MapReduce algorithm to predict the pattern of usage of credit card by a particular customer. Depending upon the maximum number of usage of a credit card in a particular sector such as food, shopping , movies, travel etc, the next offers can be planned for the customer. As a result of this predictive analysis the bank as well as the customer will be benefited. Map can be used to sort the various sectors where the card has been used by the customer and then Reduce can analyze the top 2 sectors where the card has been used the most to understand the pattern in order to predict efficiently. Once the top 2 sectors are reduced the next personalized offer can be planned on the basis on Analysis by the bank.

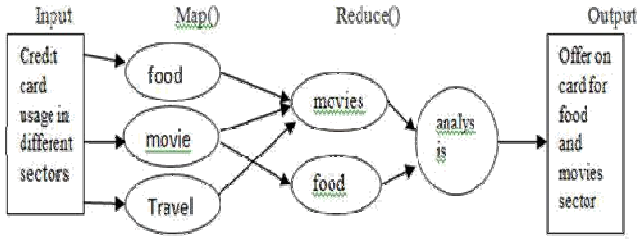
Input
Map()
Reduce()
Output
Credit card usage in different sectors
food
movie
Travel
mcvies
food
analys is
Offer on card for food and movies sector

Fig.1: Example for using MapReduce algorithm in banking Sector

In health care sector Healthcare analytics is done using Hadoop framework. It helps in the detection of diseases at earlier stage, in maintaining the privacy of the patient's information and in the quick identification of any health care fraud.

Apache Spark: Apache spark is a cluster computing technology designed for faster computation. Spark not only supports 'Map' and 'reduce'. It also supports SQL queries, Streaming data,

Machine learning (ML), and Graph algorithms. Spark helps to run an application in Hadoop cluster, up to 100 times faster in memory, and 10 times faster when running on disk. It supports multiple languages.[24]

Apache spark can be used in many application areas such as banking, GIS, NLP, Digital forensics, sensor networks, multimedia etc. as it supports real time and streaming datasets with a high velocity processing.

Hive: Apache Hive is an open source Hadoop application for data warehousing. It offers a simple way to apply structure to large amounts of unstructured data, and then perform batch SQL-like queries on that data. Queries are written using a SQL-like language called HiveQL, which Hive translates into MapReduce jobs that are executed on the Hadoop cluster. More complex queries are supported through User Defined Functions (UDF) can be written in Java and referenced by a HiveQL query.

Apache Hive is a data warehouse infrastructure built on top of Hadoop for providing data summarization, query, and analysis.Hive gives an SQL-like interface to query data stored in various databases and file systems that integrate with Hadoop. The traditional SQL queries must be implemented in the MapReduce Java API to execute SQL applications and queries over a distributed data.

Hive can be used in application areas such as banking Sector, Health Care, Natural Language Processing, Digital Forensics, Sensor Networks and Weather forecasting.

Apache Pig: Apache Pig is an abstraction over MapReduce. It is a tool/platform which is used to analyze larger sets of data representing them as data flows. Pig is generally used with Hadoop; we can perform all the data manipulation operations in Hadoop using Pig. Apache Pig is generally used by data scientists for performing tasks involving ad-hoc processing and quick prototyping. Apache Pig is used :

- To process huge data sources such as web logs.
- To perform data processing for search platforms.
- To process time sensitive data loads.
- To mask the sensitive information.

Apache Pig can be used in application areas such as banking sector, health care, digital forensics, human behavior modeling, bio informatics etc.

Storm: Apache Storm is a distributed real-time big data-processing system. Storm is designed to process vast amount of data in a fault-tolerant and horizontal scalable method. It is a streaming data framework that has the capability of highest ingestion rates. Though Storm is stateless, it manages distributed environment and cluster state via Apache ZooKeeper. Storm is easy to setup, operate and it guarantees that every message will be processed through the topology at least once.

Apache storm is used in the applications areas such as human behavior modeling, sensor networks, multimedia etc. For example: Under sensor networks Storm is being used as mobile call log analyzer. The input data is mobile calls and its duration and the output is the group of calls between the same caller and receiver and its total number of calls.

V. CONCLUSION

Big data is being generated in several application areas in structured to unstructured form. It possesses a lot of challenges for the applications to acquire and analyze this huge amount of data.

Different tools and processing techniques are being used. In this paper we have compared some of the application areas on the basis of usage of the Big data tools to find the optimum technique to handle big data. As per the comparative analysis Hadoop framework is being used by most of the applications. Hadoop is a framework which uses batch processing technique and it has an ability to scale itself in accordance with scalable data. In addition to Hadoop, Apache Spark is being used for real time Non batch processing ie for streamed data.

REFERENCES

[1] B. Marr, "why only one of the 5 vs of big data really matters".

[2] Casey Eoghan, Digital evidence and computer crime: Forensics Science, Computers and the Internet. Academic Press 2011.

[3] Mukkamala, Srinivas, and Andrew H. Sung."Identifying significant features for network forensic analysis using artificial intelligent techniques." International Journal of digital evidence 1.4 (2003): 1-17.

[4] E. Casay, “Digital Evidences and Computer Crime”, Elsevier Inc., 2011.

[5] Raghupathi W,Raghupathi V:Big data Analytics in health Care:Promise and Potential.Health inform Sci syst,2014,2(1):3-10.1186/2047-2501-2-3.

[6] L. Neumeyer, B. Robbins, A. Nair, and A. Kesari, “S4: Distributed Stream Computing Platform,” in Proceedings of IEEE International Conference on Data Mining Workshops (ICDMW), 2010, pp.170–177.

[7] https://www. atkearney. com documents/10192/698536/FG-Big-Data-and-the-Creative-Destruction f –Todays Business-Models-1.png/0c519468-de82-45cb-afde-5d7b431b00b5?t=1358296806758.

[8] Rabkin and R. H. Katz, “How Hadoop Clusters Break,”IEEE Software, vol. 30,pp. 88-94, 2013.

[9] Rodrigo Agerri,Xabier Artola, Zuhaitz Beloki, German Rigau, Aitor Soroa “Big Data for Natural Language Processing:A streaming approach“ in Knowledge-Based Systems Volume 79, May 2015, Pages 36–42.

[10] http://www.forbes.com/sites/martinzwilling/2015/03/24/what-canbig-data-ever-tell-us-about human behavior/#1580346f1bed

[11] M. R. Evans, D. Oliver, K. Yang, X.Zhou, S. Shekhar, “Enabling Spatial Big Data via CyberGIS: Challenges and Opportunities,” Ed. S. Wang, M. F.Goodchild, CyberGIS: Fostering a New Wave of Geospatial Innovation and Discovery. Springer, 2014

[12] https://en.wikipedia.org/wiki/Spatial_database

[13] Hirak Kashyap, Hasin Afzal Ahmed, Nazrul Hoque, Swarup Roy and Dhruba Kumar Bhattacharyya “Big Data Analytics in Bio-informatics: A Machine Learning Perspective” in Journal of LATEX CLASS FILES, vol. 13, no. 9,September 2014

[14] H. Nordberg, K. Bhatia, K. Wang and Z.Wang, “BioPig: a Hadoop-based analytic toolkit for large-scale sequence data,” Bioinformatics, vol. 29, no. 23, pp. 3014–3019, 2013

[15] B. Langmead, M. C. Schatz, J. Lin, M. Pop, and S. L. Salzberg, “Searching for SNPs with cloud computing,”

[16] Genome Biol, vol. 10, no. 11, p. R134, 2009 http://www-01.ibm.com/common/ssi/ cgi-bin/ssialias?htmlfid=IML14338USEN&appname=skmwww

[17] Hossein Hassani and Emmanuel Sirimal Silva“Forecasting with Big Data: A Review “in Annals of Data Science March 2015, Volume 2, Issue 1, pp 5-19

[18] Veershetty Dagade, Mahesh Lagali,Supriya Avadhani and Priya Kalekar, Big Data Weather Analytics Using Hadoop,International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE) ISSN: 0976-1353, Volume 14 Issue 2 –APRIL 2015

[19] H. Hu, Y. Wen, T.-S. Chua, and X. Li, ‘‘Towards scalable systems for Big Data analytics: A technology tutorial,’’ IEEE Access, vol. 2, pp. 652–687, 2014

[20] Carmen C. Y. Poon et al, “Body Sensor Networks: In the Era of Big Data and Beyond” in IEEE Reviews in Biomedical Engineering, VOL. 8, 2015.

[21] T. Jiang, Q. Zhang, R. Hou, L. Chai, S. A. McKee, Z. Jia, and N. Sun,“Understanding the behavior of in-memory computing workloads,” in Workload Characterization (IISWC),IEEE International Symposium on,2014,pp. 22–30

[22] Yook-Pei Shee, Hillie Richter, Svenn-Petter Maehle, David Crompton, “Big data in banking for marketers How to derive value from big data”.Innovation Lab, EVRY.

[23] http://www.sas.com/en_us/insights/big-data/hadoop.html

[24] https://www.tutorialspoint.com/apache_spark/apache_spark_i ntroduction.htm