Marupaka Nagaraju *et al*, International Journal of Computer Science and Mobile Applications,
Vol.6 Issue. 2, February- 2018, pg. 116-130 **ISSN: 2321-8363**
Impact Factor: 5.515

Theoretical Understanding the Concepts of Big Data and its Approaches

Marupaka Nagaraju[1], Dr. Dhanalaxmi Vadlakonda[2]
[1]*Research Scholar, Department of Computer Science, University College of Science, Osmania University, Hyderabad, Telangana,India*
E-mail: marupaka101@gmail.com

[2]*Associate Professor, Department of Mathematics, University College of Technology(Autonomous), Osmania University, Hyderabad,Telangana,India*
E-mail: v_laxmi_sudha@yahoo.co.in

Abstract
In the present scenario, the massive growth in the scale of data has been observed as a key factor of the big data. Now-a-days data is considered to be the biggest assets. As the enormous amount of data is making its space inside the world there is a new evolution of data. Big Datasets are endemic, but are often notoriously difficult to analyze because of their size, heterogeneity and quality. Data sets with sizes beyond the capability of usually used software tools to capture, manage, and process data within a tolerable elapsed time, are included in Big Data. One who has maximum relevant data is considered to be rich in the information industry. But only the collection of data is not enough, it needs to be analyzed. This huge amount of data which is termed as big data cannot be analyzed by traditional tools and techniques; rather it requires more advanced techniques which can make data retrieval, management and storage much faster are required. The rapid evolution and adoption of big data by industry has leapfrogged the discourse to popular outlets, the statistical methods in practice were devised to infer from sample data. The heterogeneity, noise and the massive size of structured big data calls for developing computationally efficient algorithms that may avoid big data pitfalls, such as spurious correlation. Addressing big data is a challenging and time-demanding task that requires a large computational infrastructure to ensure successful data processing and analysis. In this paper, we connote the basic concepts of big data environment.

Keywords: Big data, Data mining, MapReduce, Servitization, Data transformation, heterogeneity and 5V.

1. Introduction
Big data is data sets that are so voluminous and complex that traditional data processing application software are inadequate to deal with them. A consensual definition that states that "Big Data represents the Information assets characterized by such a high volume, velocity and variety to require specific Technology and Analytical Methods for its transformation into Value". Big data challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating and information privacy. There are five dimensions to big data known as Volume, Variety, Velocity, Veracity and Value.

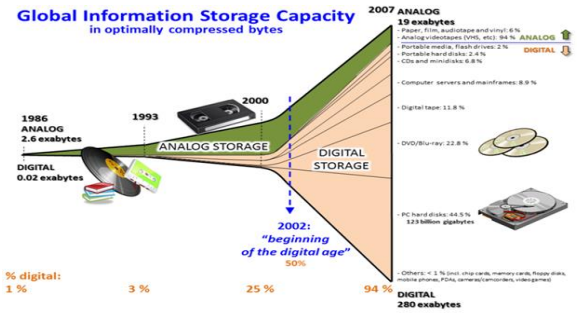

Figure 1: Growth of and digitization of global information-storage capacity

Analysis of data sets can find new correlations to "spot business trends, prevent diseases, and combat crimes and so on."[4]. Scientists, business executives, practitioners of medicine, advertising and governments alike regularly meet difficulties with large data-sets in areas including Internet search, fintech, urban informatics, and business informatics. Data sets grow rapidly - in part because they are increasingly gathered by cheap and numerous information-sensing Internet of things devices such as mobile devices, aerial (remote sensing), software logs, cameras, microphones, radio-frequency identification (RFID) readers and wireless sensor networks.[4][5]. Relational database management systems and desktop statistics- and visualization-packages often have difficulty handling big data. What counts as "big data" varies depending on the capabilities of the users and their tools, and expanding capabilities make big data a moving target. [6]. Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time.[7] . Big Data philosophy encompasses unstructured, semi-structured and structured data; however the main focus is on unstructured data. [8]. Big data "size" is a constantly moving target, as of 2012 ranging from a few dozen terabytes to many petabytes of data.[9] Big data requires a set of techniques and technologies with new forms of integration to reveal insights from datasets that are diverse, complex, and of a massive scale.[10]. Big Data also suffer of the aforementioned negative factors. Big data pre-processing constitutes a challenging task, as the previous existent approaches cannot be directly applied as the size of the data sets or data streams make them unfeasible. In this overview we gather the most recent proposals in data pre-processing for Big Data, providing a snapshot of the current state-of-the-art. Besides, we discuss the main challenges on developments in data pre-processing for big data frameworks, as well as technologies and new learning paradigms where they could be successfully applied.

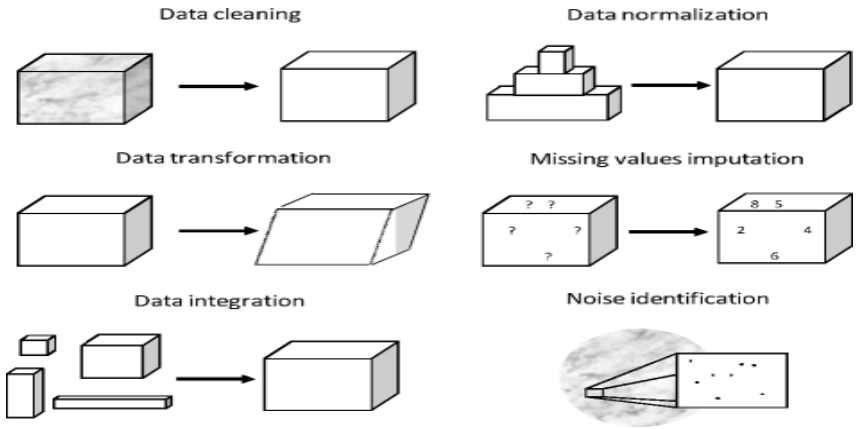

Figure 2: Data processing tasks

1.1 Characteristics (V's of big data)

Big data can be described by the following characteristics (5V) [12]

- **Volume**

Volume refers to the magnitude of data. Big data sizes are reported in multiple terabytes and petabytes. The size of the data determines the value and potential insight- and whether it can actually be considered big data or not.

- **Variety**

Variety refers to the structural heterogeneity in a dataset. The type and nature of the data. This helps people who analyze it to effectively use the resulting insight.

- **Velocity**

Velocity refers to the rate at which data are generated and the speed at which it should be analyzed and acted upon. In this context, the speed at which the data is generated and processed to meet the demands and challenges that lie in the path of growth and development.

- **Variability**

Inconsistency of the data set can hamper processes to handle and manage it. Variability is different from variety. A coffee shop may offer 6 different blends of coffee, but if you get the same blend every day and it tastes different every day, that is variability. The same is true of data; if the meaning is constantly changing it can have a huge impact on your data homogenization.

- **Veracity(Complexity)**

The data quality of captured data can vary greatly, affecting the accurate analysis [12], which represents the unreliability inherent in some sources of data. Variability refers to the variation in the data flow rates.

Figure 3: V's of big data [27]

1.2 Applications of big data

- Banking and Securities
- Communications, Media and Entertainment
- Healthcare Providers
- Education
- Manufacturing and Natural Resources
- Government
- Insurance
- Retail and Whole sale trade
- Industrial and Transportation
- Energy and Utilities

Figure 4: Applications of big data

The remainder of this paper organized as follows:

Chapter 1 explained introduction about Big data concepts, Chapter 2 explained Related work. Chapter 3 explained Methodology, Chapter 4 gives conclusion. In Chapter 5 acknowledgment and at last, references have been given which were used for preparing this paper.

2. Related Work

Big data repositories have existed in many forms, often built by corporations with a special need. Commercial vendors historically offered parallel database management systems for big data beginning in the 1990s. For many years, WinterCorp published a largest database report. [14]. The **DBC/1012** Data Base Computer was introduced by Teradata Corporation in 1984, as a back-end data base management system for mainframe computers.[14].The DBC/1012 was designed to manage databases up to one terabyte (1,000,000,000,000 characters) in size; "1012" in the name refers to "10 to the power of 12".[16].

Major components included:

- Mainframe-resident software to manage users and transfer data
- Interface processor (IFP) - the hardware connection between the mainframe and the DBC/1012
- Ynet - a custom-built system interconnect that supported broadcast and sorting
- Access module processor (AMP) - the unit of parallelism: includes microprocessor, disk drive, file system, and database software
- System console and printer
- TEQUEL (TEradata QUEry Language) - an extension of SQL

The DBC/1012 was designed to scale up to 1024 Ynet interconnected processor-disk units. Rows of a relation (table) were distributed by hashing on the primary database index. The DBC/1012 preceded the advent of redundant array of independent disks (RAID) technology, so data protection was provided by the "fallback" feature, which kept a logical copy of rows of a relation on different AMPs. The collection of AMPs that provided this protection for each other was called a cluster. A cluster could have from 2 to 16 AMPs. Teradata installed the first petabyte class RDBMS based system in 2007. As of 2017, there are a few dozen petabyte class Teradata relational databases installed, the largest of which exceeds 50 PB. Systems up until 2008 were 100% structured relational data. Since then, Teradata has added unstructured data types including XML, JSON, and Avro. In 2000, Seisint Inc. (now LexisNexis Group) developed a C++-based distributed file-sharing framework for data storage and query. The system stores and distributes structured, semi-structured, and unstructured data across multiple servers. Users can build queries in a C++ dialect called ECL [17]. ECL uses an "apply schema on read" method to infer the structure of stored data when it is queried, instead of when it is stored. In 2004, Google published a paper on a process called MapReduce algorithm that uses a similar architecture [18]. The MapReduce concept provides a parallel processing model, and an associated implementation was released to process huge amounts of data. With MapReduce, queries are split and distributed across parallel nodes and processed in parallel (the Map step). The results are then gathered and delivered (the Reduce step). Apache Spark [19] was developed in 2012 in response to limitations in the

MapReduce paradigm, as it adds the ability to set up many operations (not just map followed by reduce).2012 studies showed that a multiple-layer architecture is one option to address the issues that big data presents. Distributed parallel [20] architecture distributes data across multiple servers; these parallel execution environments can dramatically improve data processing speeds. This type of architecture inserts data into a parallel DBMS, which implements the use of MapReduce and Hadoop frameworks. This type of framework looks to make the processing power transparent to the end user by using a front-end application server.[21].Big data analytics for manufacturing applications is marketed as a 5C architecture (connection, conversion, cyber, cognition, and configuration) [22]. The data lake allows an organization to shift its focus from centralized control to a shared model to respond to the changing dynamics of information management. This enables quick segregation of data into the data lake, thereby reducing the overhead time.

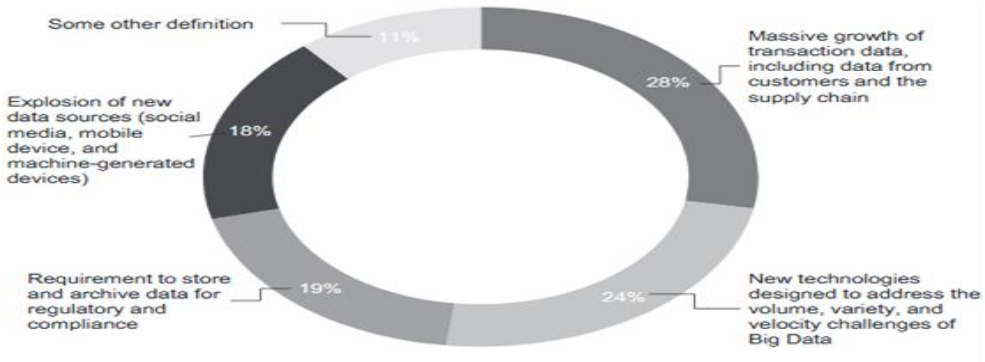

Figure 5: Definition of big data based on online survey

3. Methodology

3.1 Manufacturing Servitization and innovation

Many advanced countries whose economic base is the manufacturing industry have made efforts to transform their economy and reinvigorate the industry. They suffer threats from emerging markets and the global manufacturing supply chain. Therefore, manufacturing firms not only seek manufacturing technique innovation, but are also beginning to focus on induction and impetus of service. This way, the fuzzy boundary of the manufacturing industry and service industry drive will stimulate the development of manufacturing servitization. Servitization is defined as the strategic innovation of an organization's capabilities and processes to shift from selling products, to selling an integrated product and service offering that delivers value in use, i.e. a Product-Service System .Servitization was proposed by Vandermerve and Rada in 1988 [23].Baines defined manufacturing servitization as innovation of organizational capabilities and processes, from product sales to integrated product services [24].

3.2 Industrial big data environment

Recently, big data becomes a buzzword on everyone's tongue. It has been in data mining since human-generated content has been a boost to the social network. Lots of research organizations and companies have devoted themselves to this new research topic, and most of them focus on social or commercial mining. This includes sales prediction, user relationship mining and clustering, recommendation systems, opinion mining, etc. Using appropriate sensor installations, various signals such as vibration, pressure, etc. can be extracted. In addition, historical data can be harvested for further data mining. Communication protocols, such as MTConnect and OPC, can help users record controller signals. [25].The actual processing of big data into useful information is then the key of sustainable innovation within an Industry 4.0 factory.

3.3 Self-aware and self-maintenance machines for industrial big data environment

For a mechanical system, self-awareness means being able to assess the current or past condition of a machine, and react to the assessment output. Such health assessment can be performed by using a data-driven algorithm to analyze data/information collected from the given machine and its ambient environment. The condition of

the real-time machine can be fed back to the machine controller for adaptive control and machine managers for in-time maintenance. However, for most industrial applications, especially for a fleet of machines, self-awareness of machines is still far from being realized. Current diagnosis or prognosis algorithms are usually for a specific machine or application, and are not adaptive or flexible enough to handle more complicated information.

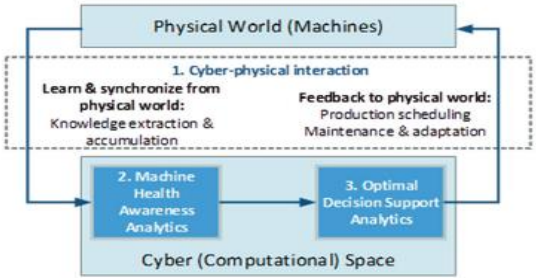

Figure 6: Cyber-physical system framework for self-aware and self-maintanance machines

3.3.1 Machine health awareness analytics with self-learning knowledge base

Machine health awareness analytics are designed to fulfill such a task. Using adaptive learning and data mining algorithms, a knowledge base representing machine performance and degradation mechanisms can be automatically populated. The knowledge base will be able to grow with new data to eventually enhance its fidelity and capability of representing complex working conditions that happen to real-world machines. With data samples and associated information collected from machines, both horizontal (machine to machine) and vertical (time to time) comparison will be performed using specifically designed algorithms for knowledge extraction. Whenever health information of a particular machine is required, the knowledge base will provide necessary information for health assessment and prediction algorithms. Because of the comprehensiveness of the knowledge base, PHM algorithms can be more flexible on handling unprecedented events, and more accurate on PHM result generation.

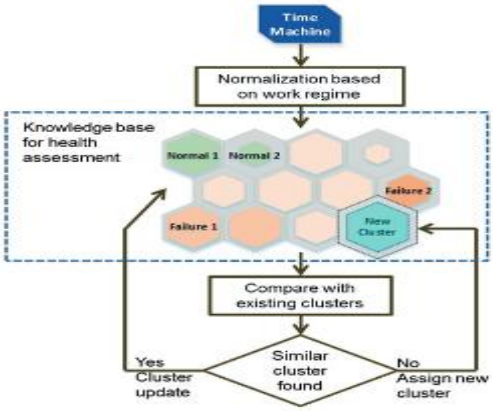

Figure 7: Adaptive-learning for machine clustering

Besides these, there are many techniques being used to analyze datasets.

3.3.2 A/B testing: a technique in which a control gropis compared with a variety of test groups in order to determine what changes will improve a given objective variable. This is also known as split testing or bucket testing. Big data enables huge numbers of tests to be executed and analyzed.

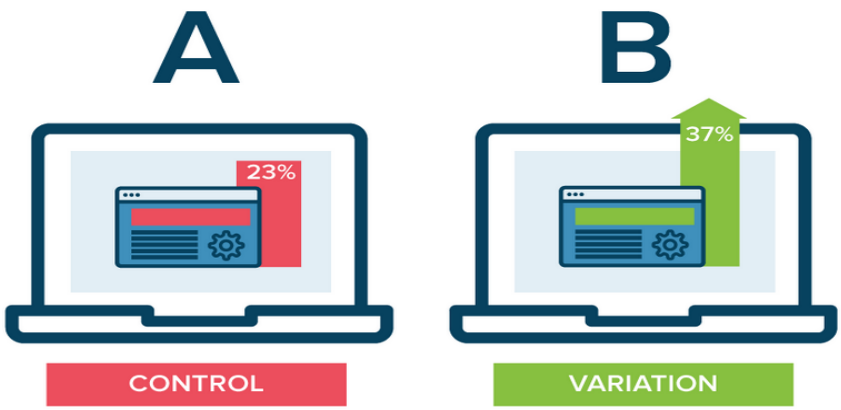

Figure 8: A\B testing of big data

3.3.3 Association rule learning: a set of techniques for discovering interesting relationships (association rules) among variables in large databases. This is another major challenge for Big Data pre-processing and it will concern different learning paradigms, besides classification and regression, such as:

- Unsupervised learning: Clustering and rule association mining have been addressed in Big Data. Developments on real-time applications can be also found in the literature. It is well-known that the success of these problems depends heavily on the quality of data, being the data cleaning, transformation and discretization the techniques with the most important role for this.
- Semi-supervised learning: A significant growth of applications and solutions on this paradigm is expected in the near future. Due to the fact of generating and storing more and more data, the labeling of examples cannot be done for all and the predictive or descriptive task will be supported by a subset of labeled examples. Data preprocessing, especially at the instance level, would be useful to improve the quality of this kind of data.
- Data streams and real-time processing: processing large data offering real-time responses are one of the most popular and demanding paradigms in business. Currently, there are some specific approaches in Big Data streams and even software development. Data preprocessing techniques, such as noise editing, should be able to tackle Big Data scenarios in upcoming applications.
- Non-standard supervised problems: there are some other popular supervised paradigms in which Big Data solutions will be necessary soon. This is the case of ordinal classification/regression, multi-label classification or multi-instance learning. All the possible data preprocessing approaches will also be required to enable and improve these solutions.

3.4 Big data analytics and Frameworks

Big data are worthless in a vacuum. The overall process of extracting insights from big data can be broken down into five stages. These five stages from the two main sub-processes called data management and analytics. Big data analytics can be viewed as a sub-process in the overall process of 'insight extraction' from big data. In the following section, we briefly review big data analytical techniques for structured and unstructured.

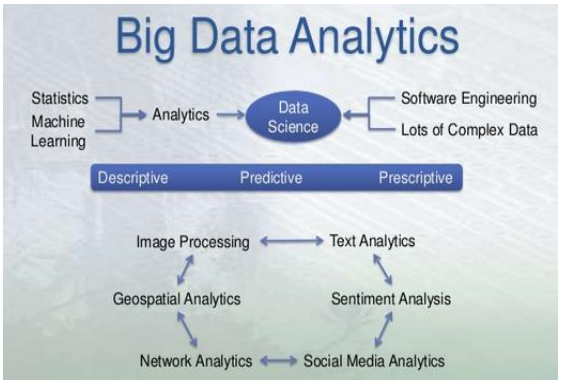

Figure 9: Big data analytics

3.4.1 Text Analytics

Text analytics (text mining) refers to techniques that extract information from textual data. Social network feeds, emails, blogs, online forums, survey responses, corporate documents, news, and call center logs are examples of textual data held by organizations. Information extraction (IE) techniques extract structured data from unstructured text. For example, IE algorithms can extract structured information such as drug name, dosage, and frequency from medical prescriptions. Two sub-tasks in IE are Entity Recognition (ER) and Relation Extraction (RE) (Jiang,2012). ER finds names in text and classifies them into predefined categories such as per- son, date, location, and organization. RE finds and extracts semantic relationships between entities (e.g., persons, organizations, drugs, genes, etc.) in the text. For example, given the sentence "Steve Jobs co-founded Apple Inc. in 1976", an RE system can extract relations such as Founder of [Steve Jobs, Apple Inc.] or Founded In [Apple Inc., 1976]. Text summarization techniques automatically produce a succinct summary of a single or multiple documents. Summarization follows two approaches: the extractive approach and the abstractive approach. In extractive summarization, a summary is created from the original text units (usually sentences). And extractive summarization techniques do not require an 'understanding' of the text. In order to parse the original text and generates the summary, abstractive summarization incorporates advanced NLP techniques. Semantic analysis technique analyses opinion text. It is also called as opinion text.

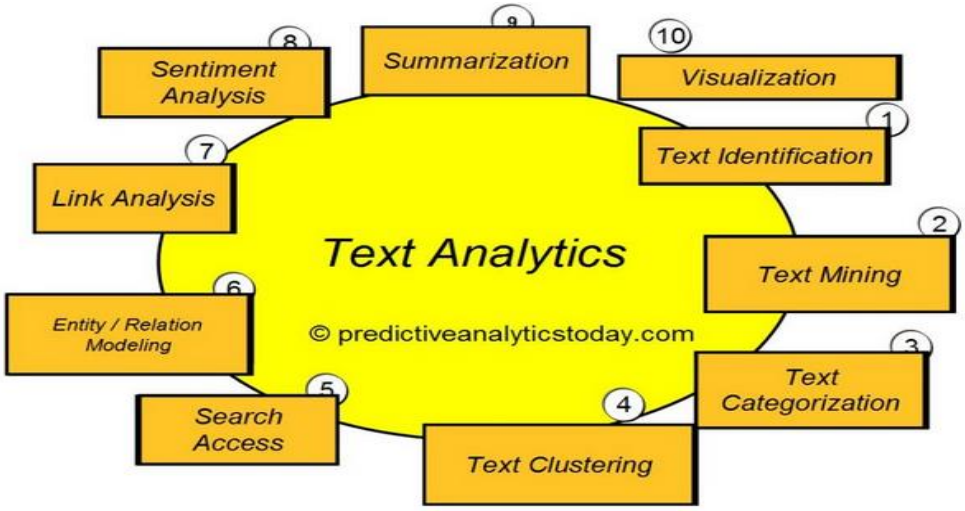

Figure 10: Text Analytics of big data

3.4.2 Audio Analytics

Audio analytics analyze and extract information from unstructured audio data. It is also called speech analytics. Speech analytics follows two common technological approaches: the transcript-based approach (widely known as large-vocabulary continuous speech recognition LVCSR and the phonetic-based approach. LVCSR system follows a two-phase process indexing and searching. Phonetic-based systems work with sounds or phonemes. Phonemes are the perceptually distinct units of sound in a specified language that distinguishes one word from another.

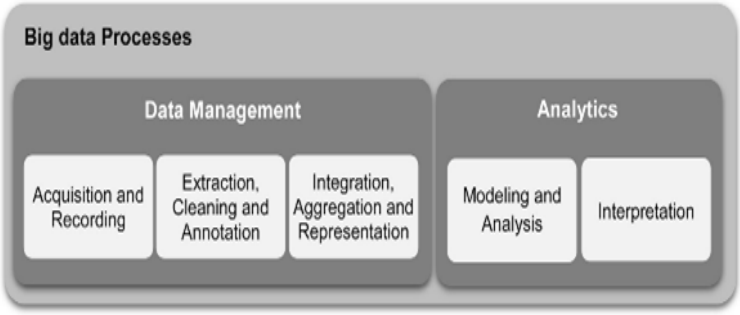

Figure 11: Processes of extracting insights from big data

3.4.3 Video Analytics

Video analytics also known video content analysis (VCA). It involves a variety of techniques to monitor, analyze, and extract meaningful information from video streams. The primary application of video analytics in recent years has been in automated security and surveillance systems. Big data technologies turn this challenge into opportunity. Obviating the need for cost-intensive and risk-prone manual processing. Big data technologies can be leveraged to automatically sift through and draw intelligence from thousands of hours of video. The primary application of video analytics in recent years has been in automated security and surveillance systems. For instance, CCTV generated data. Another potential application of video analytics in retail lies in the study of buying behaviour of groups. Among family members who shop together, only one interacts with the store at the cash register, causing the traditional systems to miss the data on buying patterns of other members. Automatic video indexing and retrieval constitutes and another domain of analytics applications. Audio analytics and text analytics techniques can be applied to index a video based on the associated soundtracks and transcripts.

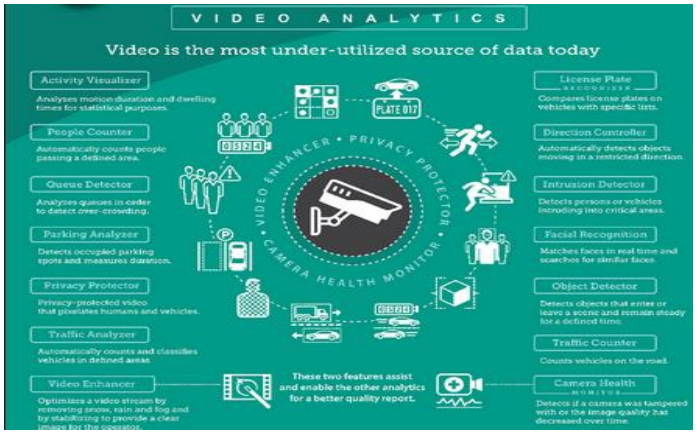

Figure 12: Video analytics of big data

3.4.4 Social Media Analytics

Social Media analytics refer to the analysis of structured and instructed data from social media channels. Social media is a broad term encompassing a variety of online platforms that allows users to create and exchange content. We can categorize social media into social networks, blogs, micro blogs, social news, social bookmarking, media sharing and wikis.

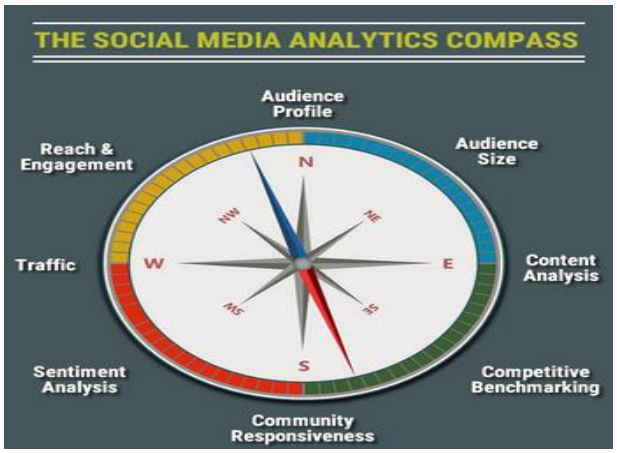

Figure 13: Social Media analytics compass of big data

3.4.5 Predicative analytics

Predicative analytics comprise a variety of techniques that predict future outcomes based on historical and current data. At its core, predicative analytics seek to uncover patterns and capture relationships in data. Based on the underlying methodology, these techniques categorized into regression and machine learning techniques. And another classification is based on the type of outcome variables. These techniques are primarily based on statistical methods. The goal of this paper is to describe and reflect on big data. Although major innovations in analytical techniques for big data have not yet taken place.

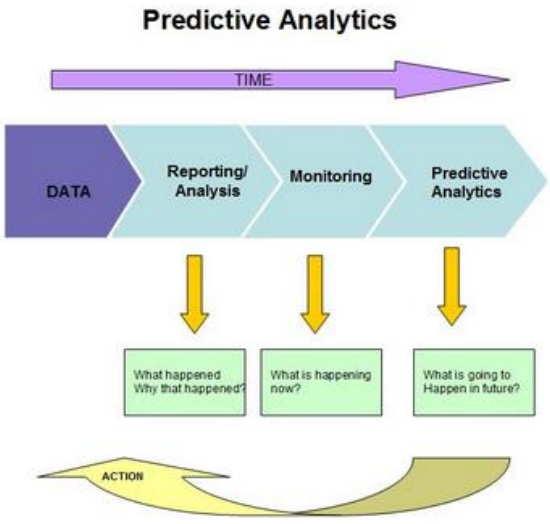

Figure 14: Predicative analysis of big data

3.5 Frameworks

3.5.1 Apache Hadoop

Hadoop is one of the most popular framework for Big Data analytics. It is a collection of Hadoop and MapReduce ecosystem tools like Pig, Flume, Hive, HDFS etc. A number of frameworks are available nowadays. Hadoop is mostly useful because it is very simple to use. Other frameworks such as Spark can also use many Hadoop tools such as YARN-The resource management layer.

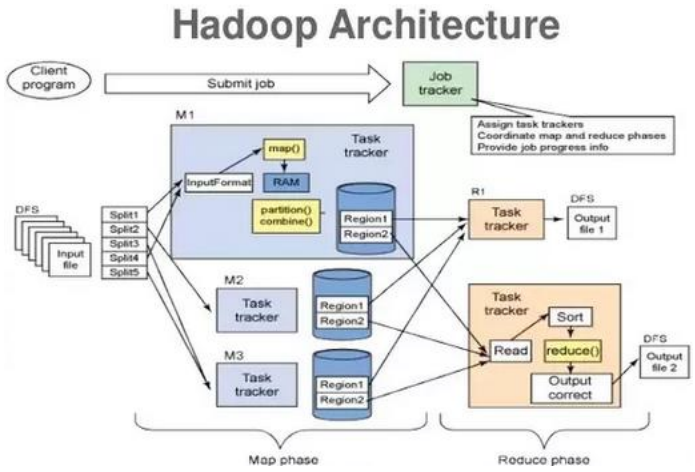

Figure 15: Hadoop Architecture of big data

3.5.2 MapReduce

MapReduce is the programming model most commonly used with Hadoop It works with mappers and reducers. Mappers deal with collection of data and analyzing it and producing Intermediate data, which is then passed to reducers. Reducers deals with aggregation of the results and give proper output.

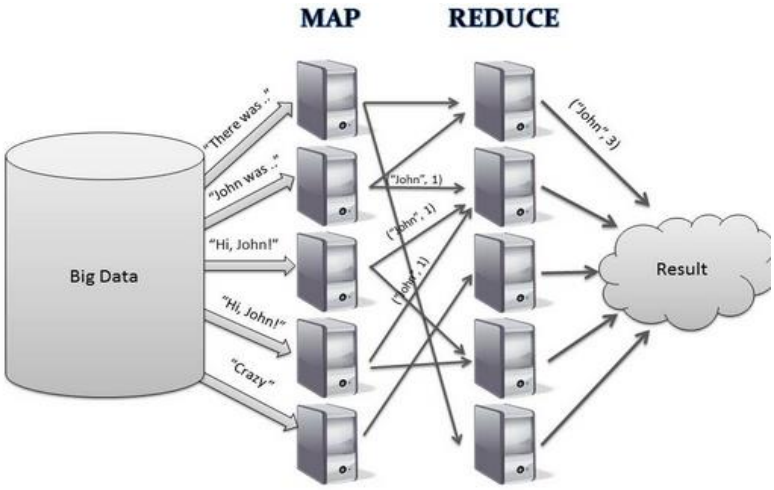

Figure 15: MapReduce of big data

3.5.3 Apache Pig and Hive

These are the two wrappers, which provide easy way instead of dealing with Map Reduce. Apache Pig is an SQL like environment, which helps in analyzing of Data. It is lying on top of Hadoop, which allows higher-level languages to use Hadoop's MapReduce library. While Hive is a technology which turns Hadoop into a data warehouse with a help to use SQL Queries.

Apache Pig Vs Hive

- Both Apache Pig and Hive are used to create MapReduce jobs. And in some cases, Hive operates on HDFS in a similar way Apache Pig does.

Apache Pig	Hive
Apache Pig uses a language called **Pig Latin**. It was originally created at **Yahoo**.	Hive uses a language called **HiveQL**. It was originally created at **Facebook**.
Pig Latin is a data flow language.	HiveQL is a query processing language.
Pig Latin is a procedural language and it fits in pipeline paradigm.	HiveQL is a declarative language.
Apache Pig can handle structured, unstructured, and semi-structured data.	Hive is mostly for structured data.

Figure 16: Apache Pig Vs Hive

3.5.4 Spark

Spark and Hadoop are often considered same but in reality, it is not so. Spark can be used in Hadoop ecosystem in place of MapReduce and the tools for both can be used to perform certain actions. Spark performs in-memory processing and allows pipelined construction for data flow unlike Hadoop and MapReduce.

Figure 16: Spark of big data

3.5.5 Apache Flink

It is a Streaming data flow engine, which helps us to perform distributed operation on stream of data. Flink consist of many API such as stream API, Static API and SQL like query API. It also has its own Machine Learning and Graph Libraries. It works with Stream flow in real time also, which is not possible with Hadoop or Spark.

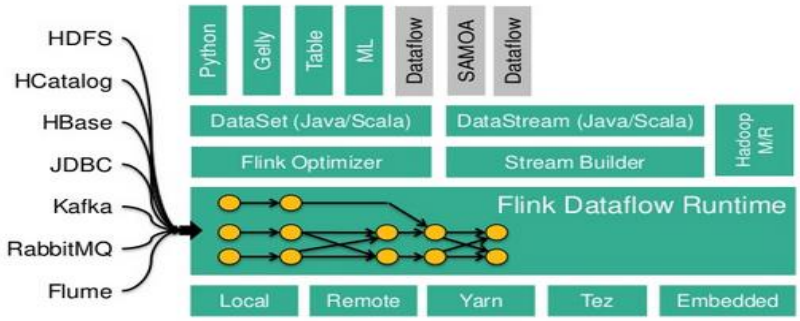

Figure 17: Apache Flink of big data

3.5.6 Apache Strom

It is a Distributed Computational System, which designs its problem as DAG (directed acyclic graph). It can be used with any programming languages with all its application. Some of its application is Real-Time analysis, Distributed Machine Learning etc. It can run on top of YARN and thus can work with Hadoop ecosystem. It is a stream Processing Engine unlike Spark.

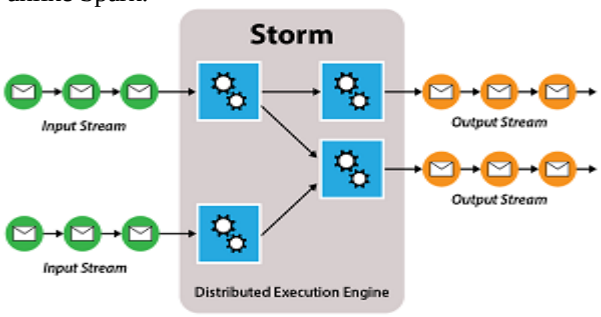

Figure 18: Storm of big data

3.5.7 Apache Flume

It is a distributed and reliable system for collecting large amounts of log data to a centralized data space. There are three main components of Flume: sinks, sources and channels. Sink is mostly a distributed file system like HDFS. Sources do the task of listening and consuming the data. Channels are the mechanism by which Flume transfers data from sources to sinks.

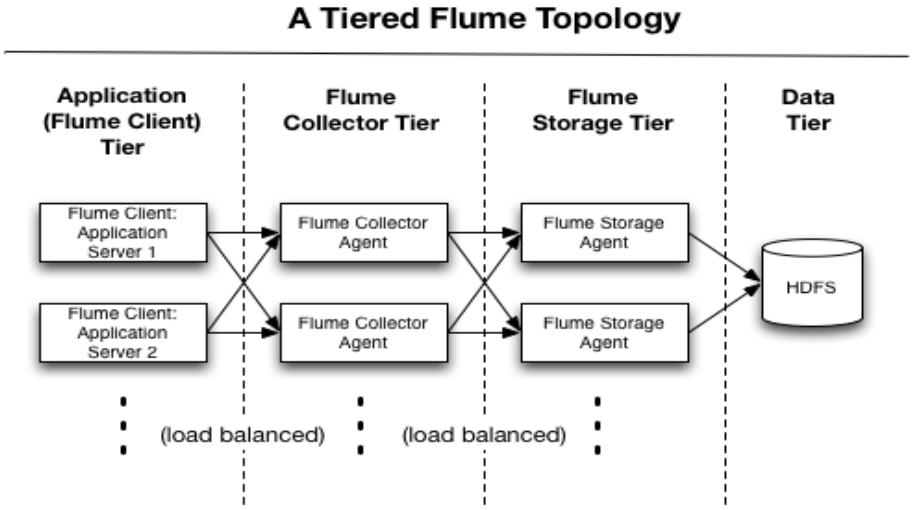

Figure 18: Apache Flume of big data

3.5.8 Apache Samza

It is another distributed stream processing framework. It is built on YARN for cluster Resource management to work with Hadoop.

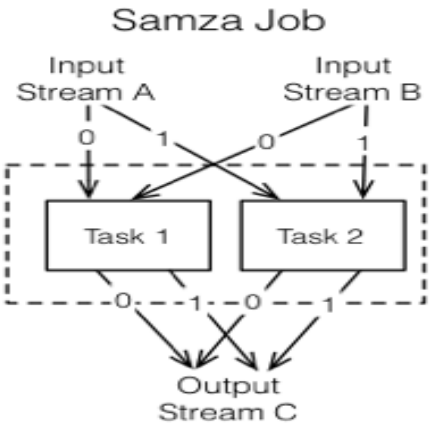

Figure 19: Apache Samza big data

4. Conclusion

Data management and distribution in Big Data environment is critical for achieving self-aware and self-learning machines. The importance of leveraging additional flexibility and capabilities offered by cloud computing is inevitable, but adapting prognostics and health management algorithms to efficiently implement current data management technologies requires further research and development [22] .In the future, significant challenges and topics must be addressed by the industry and academia, especially those related to the use of new platforms such as Apache Spark/Flink, the enhancement of scaling capabilities of existing techniques and the approach to new big data learning paradigms. Researchers, practitioners, and data scientists should collaborate to guarantee the long-term success of big data pre-processing and to collectively explore new domains.

5. Acknowledgement

The authors would like to thank all, who helped to their in this research work.

References

[1] Snijders, C.; Matzat, U.; Reips, U.-D. (2012). "'Big Data': Big gaps of knowledge in the field of Internet". International Journal of Internet Science. **7**: 1–5.
[2] The Economist. 25 February 2010. Retrieved 9 December 2012.
[3] Hellerstein, Joe (9 November 2008). "Parallel Programming in the Age of Big Data". Gigaom Blog.
[4] Segaran, Toby; Hammerbacher, Jeff (2009). Beautiful Data: The Stories Behind Elegant Data Solutions.
[5] Jacobs, A. (6 July 2009). "The Pathologies of Big Data"
[6] Snijders, C.; Matzat, U.; Reips, U.-D. (2012). "'Big Data': Big gaps of knowledge in the field of Internet". International Journal of Internet Science. **7**: 1–5.
[7] Dedić, N.; Stanier, C. (2017). "Towards Differentiating Business Intelligence, Big Data, Data Analytics and Knowledge Discovery". **285**. Berlin ; Heidelberg: Springer International Publishing. ISSN 1865-1356
[8] Everts, Sarah (2016). "Information Overload" Distillations. **2** (2): 26–33. Retrieved 17 February 2017.
[9] Ibrahim; Targio Hashem, Abaker; Yaqoob, Ibrar; Badrul Anuar, Nor; Mokhtar, Salimah; Gani, Abdullah; Ullah Khan, Samee (2015). "big data" on cloud computing: Review and open research issues". Information Systems. **47**: 98–115. doi:10.1016/j.is.2014.07.006
[10] Hilbert, Martin. "Big Data for Development: A Review of Promises and Challenges. Development Policy Review". martinhilbert.net. Retrieved 7 October 2015.
[11] https://spotlessdata.com/blog/big-datas-fourth-v
[12] http://www.eweek.com/database/survey-biggest-databases-approach-30-terabytes

[13] https://en.wikipedia.org/wiki/DBC_1012.
[14] Arthur Trew, Greg Wilson, eds. (December 6, 2012). Past, Present, Parallel: A Survey of Available Parallel Computer Systems.
[15] https://en.wikipedia.org/wiki/ECL_(data-centric_programming_language)
[16] https://en.wikipedia.org/wiki/Google
[17] https://en.wikipedia.org/wiki/Apache_Spark
[18] https://en.wikipedia.org/wiki/List_of_file_systems#Distributed_parallel_file_systems
[19] Boja, C; Pocovnicu, A; Bătăgan, L. (2012). "Distributed Parallel Architecture for Big Data". Informatica Economica. **16** (2): 116–127.
[20] http://www.imscenter.net/cyber-physical-platform
[21] https://ac.els-cdn.com/S2212827114000857/1-s2.0-S2212827114000857-main.pdf?_tid=ed99b2d4-1605-11e8-9efe-00000aab0f27&acdnat=1519107715_6b7b5c92aad65124473526d1a1ea68c2
[22] Vandermerwe, S., & Rada, J. (1989). Servitization of business: adding value by adding services. European Management Journal, 6(4), 314-324.
[23] Baines, T. S., Lightfoot, H. W., Benedettini, O., & Kay, J. M. (2009). The servitization of manufacturing: a review of literature and reflection on future challenges. Journal of Manufacturing Technology Management, 20(5), 547-567.
[24] Lee, B. E., Michaloski, J., Proctor, F., Venkatesh, S., & Bengtsson, N. (2010, January). MTConnect-based Kaizen for machine tool processes. In ASME 2010 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference (pp. 1183-1190). American Society of Mechanical Engineers.
[25] Sabia,Sheetal Kalra "Applicat ions of big Data: Current Status and Future Scope" Intern ational Journal on Advanced Computer Theory and Engineering (IJACTE).

Authors:

Dr. Dhanalaxmi Vadlakonda(v_laxmi_sudha@yahoo.co.in) pursed Bachelor of Science in Mathematics, Master of Science in Mathematics, Master of Philosophy in mathematics, Master of Education in Mathematics all degrees from Osmania University and Master of Technology from JNTU,Hyderabad. And also she achieved Doctoral degree in Mathematics from Osmania University, Hyderabad, Telangana, India.Currently she is working as an Associate Professor, Department of Mathematics, University College of Technology, Osmania University, Hyderabad, Telangana, India. She is supervising many Ph. D.'s. She has excellent teaching track record with 27 years teaching experience.

Mr. Marupaka Nagaraju (marupaka101@gmail.com) pursed Bachelor of Science in Computer Science, Master of Science in Computer Science from Osmania University, Hyderabad, Telangana, India. Presently he is pursuing Ph. D in computer science from Osmania University. His main research work focuses on Big data.

