A Literature Inspection on Big Data Analytics

Dr.E.Laxmi Lydia,
Associate Professor,
Department of Computer Science and Engineering,
Vignan's Institute Of Information Technology,
Visakhapatnam,
Andhra Pradesh,
India.

M. Vijay Laxmi,
M.Tech 2nd year,
Department of Computer Science and Engineering,
Vignan's Institute Of Information Technology,
Visakhapatnam,
Andhra Pradesh,
India.

Dr. M.Ben Swarup,
Professor,
Department of Computer Science and Engineering,
Vignan's Institute Of Information Technology,
Visakhapatnam,
Andhra Pradesh,
India.

ABSTRACT

Many business cases exploiting big data have been realized in recent years: Twitter, LinkedIn, and Face book are case of organizations in the person to person communication area for big data. Likewise, implementation architectures of the utilization cases have been distributed. Notwithstanding, theoretical work coordinating the methodologies into one rational reference architecture has been constrained. Late technological advancements have lead knowledge of data from unmistakable areas in the course of recent decades. The term big data caught the importance of this developing pattern. Notwithstanding its sheer volume, big data additionally exhibits other extraordinary attributes as contrasted and conventional data. For example big data is a normally unstructured data and require all the more constant examination. This improvement calls for new system data procurement, transmission, stockpiling, and vast scale data handling components. In this paper we exhibit a literature Inspection for big data analytics platform. To start with we exhibit brief history of big data, Introduction to big data and hadoop, Map reduce, hadoop distributed file system (HDFS) In the second stage took after by procedures and technologies for dissecting big data and points of interest and contrast between big data and hadoop in the third stage we give security utilizing cloud computing, need of security in big data and big data system architecture (data era, procurement, stockpiling and analytics). These four modules form a big data esteem chain. Big data applications are an incredible advantage to associations, business, organizations and numerous vast scale and little scale ventures. We additionally talk about different conceivable answers for the issues in cloud computing, security and Hadoop. Cloud computing security is creating at a fast pace which incorporates PC security, system security, information security, and data protection. Cloud computing assumes an exceptionally key part in ensuring data, applications and the related framework with the assistance of policies, technologies, controls, and big data devices. In addition, cloud computing, big data and its applications, points of interest are prone to speak to the most encouraging new outskirts in science.

Keywords: Big data analytics, data generation, data acquisition, data storage, data analytics, cloud computing, Map reduce, Hadoop distributed file system (HDFS) and Hadoop.

1. INTRODUCTION

Lately, big data has rapidly developed into a hotspot that pulls in incredible consideration from academic ,industry, and even governments around the globe [1–2]. Nature and Science have distributed exceptional issues committed to talk about the opportunities and challenges brought by big data [3, 4]. Big data has infiltrated into each zone of today's industry and business works and has turned into an imperative variable in production. Numerous big data use cases have been acknowledged, which make extra esteem for companies, end users and third parties. As of now, constant data is assembled from a huge number of end users by means of prominent long range interpersonal communication services. For instance, LinkedIn [5] gathers data from users, and offers services such as "Individuals you may know", aptitude endorsements or news encourage upgrades to end users based on examination of the data. Another case is Netflix, which uses big data for providing recommendations and ranking related services to customers [6]. Twitter uses gathered data for ongoing inquiry proposal and spelling corrections of their search algorithm [7]. Examination of gathered data likewise increases understanding of consumers, which is a critical asset for the big data companies. Esteem from data can likewise be extracted with different applications such as monitoring of network traffic [8] or enhancing fabricating procedure of digital presentations [9].

Apache Hadoop is a software framework that supports data-intensive distributed applications under a free license, which has been used by many big technology companies, such as Amazon, Face book, Yahoo and IBM. Hadoop [10] is well known for Map reduces and it's distributed file system (HDFS). Map reduce idea is described in a Google paper [11], to be minimize the task of Map reduce is another processing of divide and concur. Hadoop [12] is focused at problems that need examination of all the available data. For instance, text analysis and image processing generally need that every single record be read and often interpreted in the context of similar records .A wide variety of technologies and heterogeneous architectures have been applied in the implementation of the big data use cases. The publications have for the most part concentrated on depicting architectures of individual commitments by vast big data companies such as Face book [13] or LinkedIn [5]. Then again, architectural work joining the individual reports into one lucid reference design has been constrained, in spite of the fact that the principal commitments have been made [14-17]. Innovation autonomous reference design and categorization of related execution technologies and services would be important for research and development of big data systems.

The contribution of this research paper is reference architecture for big data systems, and grouping of related technologies and products/services. In the first place, big data research, reference architectures, and use cases are reviewed from literature. Hence, the configuration of reference architecture for big data systems is displayed, which has been built inductively in light of analysis of the exhibited use cases. At long last, a characterization is accommodated the motivation behind making a general picture of big data research, related technologies, products, and services.

Google has presented Map Reduce [18] framework for processing large amounts of data on commodity hardware. Apache's Hadoop distributed file system (HDFS) is evolving as a superior software component for cloud computing combined along with integrated parts such as Map Reduce. Hadoop, which is an open-source implementation of Google Map Reduce, including a distributed file system, provides to the application programmer the abstraction of the map and the reduce. With Hadoop it is easier for organizations to get a grip on the large volumes of data being generated every day, but at the same time can also create problems pertain to security, data access, monitoring, high availability and business continuity

We should have a framework that can scale to handle an huge number of sites and likewise have the capacity to process substantial and monstrous measures of data. Nonetheless, best in class frameworks using HDFS and Map Reduce are not exactly enough/adequate as a result of the way that they don't give required efforts to establish safety to ensure delicate data. In addition, Hadoop framework is utilized to take care of issues and oversee data helpfully by utilizing diverse strategies, for example, consolidating the k-means with data mining innovation [19].

The structure of the paper is as follows: section1 Discovers a brief history of big data and followed by introduction to big data and hadoop , map reduce and hadoop distributed file systems big data techniques and technologies for analyzing big data and advantages and differences between big data and hadoop. section 2 describes the introduction to cloud computing and issues and challenges in cloud computing and survey of big data, a reference architecture for big data systems and examples of companies in the social networking domain followed by a big data system value chain(data generation, data acquisition, data storage and data analytics).section 3 describes big data system challenges, security measures in big data ,phases in processing data analysis, need of security in big data, review of big data technologies, section 4 describes Hadoop framework and applications followed by improvements.

2. A BRIEF HISTORY OF BIG DATA

2.1 INTRODUCTION TO BIG DATA

What is big data? So far, there is no universally accepted definition. In Wikipedia, big data is defined as "an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications" [20]. We move to understanding the historical backdrop of big data, i.e., how it developed into its present stage. Considering the evolution and complexity of big data frameworks, past depictions depend on an uneven perspective point, for example, sequence [21] or milepost innovations. In this survey, the historical backdrop of big data is introduced as far as the data size of interest.

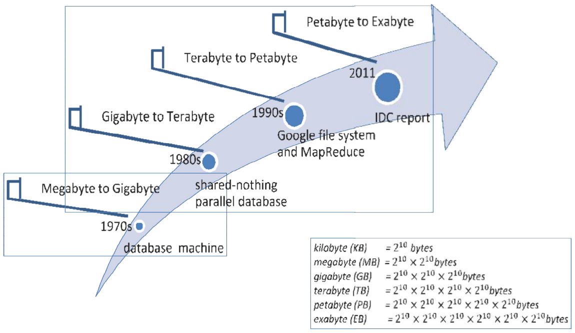
Petabyte to Exabyte
Terabyte to Petabyte
2011
IDC report
Gigabyte to Terabyte
1990s
Google file system and MapReduce
1980s
shared-nothing parallel database
Megabyte to Gigabyte
1970s
database machine
kilobyte (KB) = 2^{10} bytes
megabyte (MB) = $2^{10} \times 2^{10}$bytes
gigabyte (GB) = $2^{10} \times 2^{10} \times 2^{10}$bytes
terabyte (TB) = $2^{10} \times 2^{10} \times 2^{10} \times 2^{10}$bytes
petabyte (PB) = $2^{10} \times 2^{10} \times 2^{10} \times 2^{10} \times 2^{10}$bytes
exabyte (EB) = $2^{10} \times 2^{10} \times 2^{10} \times 2^{10} \times 2^{10} \times 2^{10}$bytes

Figure 1: Demonstrates a brief history of big data with major milestones.

Figure 1demonstrates a brief history of big data with major milestones. It can be roughly split into four stages according to the data size growth of order, including Megabyte to Gigabyte, Gigabyte to Terabyte, Terabyte to Petabyte, and Petabyte to Exabyte. Under this framework, the history of big data is tied tightly to the capability of efficiently storing and managing larger and larger datasets, with size limitations expanding by orders of magnitude. Specifically, for each capability improvement, new database technologies were developed. Thus, the history of big data can be roughly split into the following stages:

- Megabyte to Gigabyte: In the 1970s and 1980s, his historical business data introduced the earliest ''big data'' challenge in moving from megabyte to gigabyte sizes. The urgent need at that time was to house that data and run relational queries for business analyses and reporting. Research efforts were made to give birth to the database machine that featured integrated hardware and software to solve problems. The underlying philosophy was that such integration would provide better performance at lower cost. After a period of time, it became clear that hardware-specialized database machines could not keep pace with the progress of general-purpose computers. Thus, the descendant database systems are soft- ware systems that impose few constraints on hardware and can run on general-purpose computers.

- Gigabyte to Terabyte: In the late 1980s, the popularization of digital technology caused data volumes to expand to several gigabytes or even a terabyte, which is beyond the storage and/or processing capabilities of a single large computer system. Data parallelization was proposed to extend storage capabilities and to improve performance by distributing data and related tasks, such as building indexes and evaluating queries, into disparate hardware. Based on this idea, several types of parallel databases were built, including shared-memory databases, shared-disk databases, and shared- nothing databases, all as induced by the underlying hardware architecture. Of the three types of databases, the shared-nothing architecture, built on a networked cluster of individual machines - each with its own processor, memory and disk [22] - has witnessed great success. Even in the past few years, we have witnessed the blooming of commercialized products of this type, such as Teradata [23], Netezza [24], Aster Data [25], Greenplum [26], and Vertica [27]. These systems exploit a relational data model and declarative relational query languages, and they pioneered the use of divide-and- conquer parallelism to partition data for storage.
- Terabyte to Petabyte: During the late 1990s, when the database community was admiring its ‘‘finished’’ work on the parallel database, the rapid development of Web 1.0 led the whole world into the Internet era, along with massive semi-structured or unstructured web- pages holding terabytes or petabytes (PBs) of data. Unfortunately, although parallel databases handle structured data well, they provide little support for unstructured data. Additionally, systems capabilities were limited to less than several terabytes. To address the challenge of web-scale data management and analysis, Google created Google File System (GFS) [28] and Map Reduce [29] programming model. GFS and Map Reduce enable automatic data parallelization and the distribution of large-scale computation applications to large clusters of commodity servers. A system running GFS and Map Reduce can scale up and out and is there- fore able to process unlimited data. In the mid-2000s, user-generated content, various sensors, and other ubiquitous data sources produced an overwhelming flow of mixed-structure data, which called for a paradigm shift in computing architecture and large-scale data processing mechanisms. NoSQL databases, which are scheme-free, fast, highly scalable, and reliable, began to emerge to handle these data. In Jan. 2007, Jim Gray, a database software pioneer, called the shift the ‘‘fourth paradigm’’ [30]. He also argued that the only way to cope with this paradigm was to develop a new generation of computing tools to manage, visualize and analyze.

2.2 BIG DATA

Big Data is the word used to describe massive volumes of structured and unstructured data that are so large that it is very difficult to process this data using traditional databases and software technologies. The term “Big Data [31]” is companies who had to query loosely structured very large distributed data. Figure 2 and 3 describes about it.

The three main terms that signify Big Data have the following properties:

- Volume: Many factors contribute towards increasing Volume streaming data and data collected from sensors etc.,
- Variety: Today data comes in all types of formats Emails, video, audio, transactions etc.,
- Velocity: This means how fast the data is being produced and how fast the data needs to be processed to meet the demand. The other two dimensions that need to consider with respect to Big Data are Variability and Complexity [31].
- Variability: Along with the Velocity, the data can be highly inconsistent with periodic peeks.
- Complexity: Complexity of the data also needs to be considered when the data is coming from multiple sources. The data must be linked, matched, cleansed and transformed into required formats before actual processing.

Figure 2: Big Data

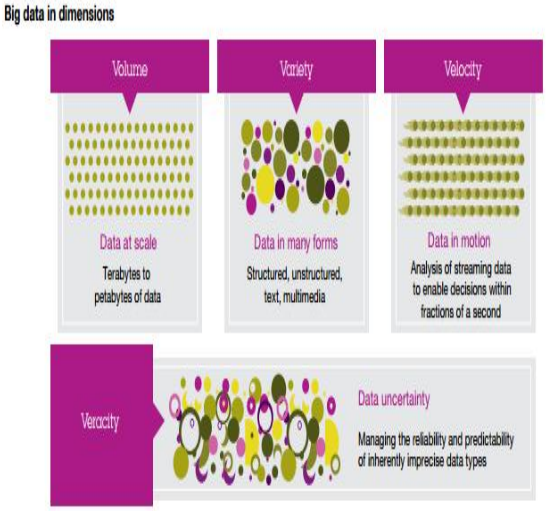
Big data in dimensions
Volume
Variety
Velocity
Data at scale
Terabytes to petabytes of data
Data in many forms
Structured, unstructured, text, multimedia
Data in motion
Analysis of streaming data to enable decisions within fractions of a second
Veracity
Data uncertainty
Managing the reliability and predictability of inherently imprecise data types

Figure 3: Big data in Dimensions

2.4 HADOOP

Hadoop, which is a free, Java-based programming framework, supports the processing of large Sets of data in a distributed computing environment. Hadoop uses a technique called Map Reduce to carry out this exhaustive analysis quickly. HDFS gives the distributed computing storage provides and support. They are the two main subprojects for Hadoop platform. It is a part of the Apache project sponsored by the Apache Software Foundation. Hadoop cluster uses a Master/Slave structure [32]. Using Hadoop, large data sets can be processed across a cluster of servers and applications can be run on systems with thousands of nodes involving thousands offer a bytes. Distributed file system in Hadoop helps in rapid data transfer rates and allows the system to continue its normal operation even in the case of some node failures. This approach lowers the risk of an entire system failure, even in the case of a significant number of node failures. Hadoop enables a computing solution that is scalable, cost effective, and flexible and fault tolerant. Hadoop Framework is used by popular companies like Google, Yahoo, Amazon and IBM etc., to support their applications involving huge amounts of data. Hadoop has two main sub projects – Map Reduce and Hadoop Distributed File System (HDFS).

2.4.1 HADOOP FUNDAMENTALS

Teaches you the basics of Apache Hadoop and the concept of Big Data. This Hadoop course is entirely free, and so are the materials and software provided. This is the third version of our most popular Hadoop course. Since Version 2 was published, several more detailed courses covering topics such as Map Reduce, Hive, HBase, Pig, Oozie, and Zookeeper have been added. We recommend you start here and then dig deeper into the specific Hadoop technology you wish to learn more about.

2.4.2 LEARN HADOOP

This Hadoop course is designed to give you a basic understanding of key Big Data technologies. In this Hadoop tutorial, we first begin with describing what Big Data is and the need for Hadoop to be able to process that data in a timely manner. This is followed by describing the Hadoop architecture and how to work with the Hadoop Distributed File System (HDFS) both from the command line and using the Big Insights Console that is supplied with Info Sphere Big Insights. This Hadoop course was recently tested and updated for Big Insights Quick Start 4.0 (IBM's edition of Hadoop). Figure 4 describes about the ecosystem of the Hadoop.

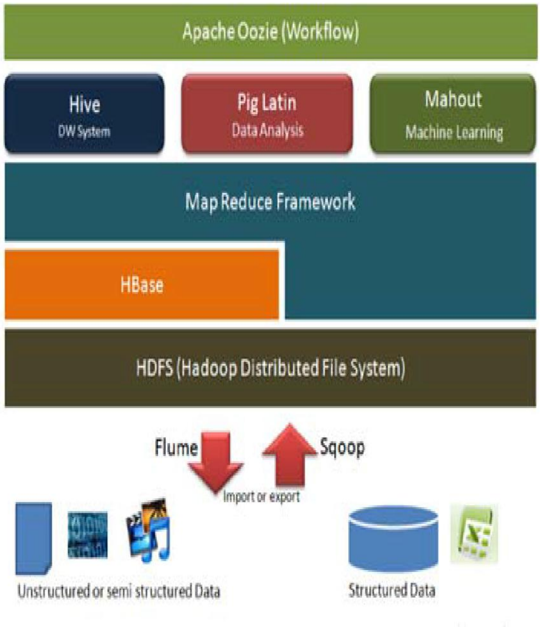

Figure 4: Hadoop Ecosystem

2.5 MAP REDUCE

The Map Reduce frame work [33] consists of a single master Job Tracker and one slave Task Tracker per cluster node. The master is responsible for scheduling the jobs' component tasks in the slaves, monitoring them, and re-executing any failed tasks. The slaves executed the tasks as directed by the master. As mentioned, Map Reduce applications are based on a master-slave model [34].

Hadoop Map Reduce is a framework used to write applications that process large amounts of data in parallel on clusters of commodity hardware resources in a reliable, fault-tolerant manner.

A Map Reduce job first divides the data into individual chunks which are processed by Map jobs in parallel. The outputs of the maps sorted by the framework are then input to the reduce tasks. Generally the input and the output of the job are both stored in a file-system. Scheduling, Monitoring and re-executing failed tasks are taken care by the framework.

2.6 Hadoop Distributed File System (HDFS)

Hadoop Distributed File System (HDFS)[35] is the primary storage system used by Hadoop applications. HDFS [36] is a file system that spans all the nodes in a Hadoop cluster for data storage. It links together file systems on local nodes to make it into one large file system. HDFS improves reliability by replicating data across multiple sources to overcome node failures. Figure 5 demonstrates about the working of HDFS.

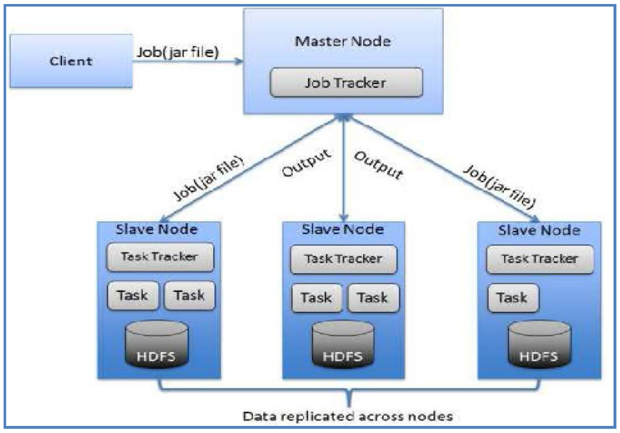

Figure 5: HDFS

3. BIG DATA TECHNIQUES AND TECHNOLOGIES

Big data techniques and technologies A wide variety of techniques and technologies has been developed and adapted to aggregate, manipulate, analyze, and visualize big data. These techniques and technologies draw from several fields including statistics, computer science, applied mathematics, and economics. This means that an organization that intends to derive value from big data has to adopt a flexible, multidisciplinary approach. Some techniques and technologies were developed in a world with access to far smaller volumes and variety in data, but have been successfully adapted so that they are applicable to very large sets of more diverse data. Others have been developed more recently, specifically to capture value from big data. Some were developed by academics and others by companies, especially those with online business models predicated on analyzing big data.

This report concentrates on documenting the potential value that leveraging big data can create. It is not a detailed instruction manual on how to capture value, a task that requires highly specific customization to an organization's context, strategy, and capabilities. However, we wanted to note some of the main techniques and technologies that can be applied to harness big data to clarify the way some of the levers for the use of big data that we describe might work. These are not comprehensive lists—the story of big data is still being written; new methods and tools continue to be developed to solve new problems. To help interested readers find a particular technique or technology easily, we have arranged these lists alphabetically. Where we have used bold typefaces, we are illustrating the multiple interconnections between techniques and technologies. We also provide a brief selection of illustrative examples of visualization, a key tool for understanding very large-scale data and complex analyses in order to make better decisions.

3.1 TECHNIQUES FOR ANALYZING BIG DATA

There are many techniques that draw on disciplines such as statistics and computer science (particularly machine learning) that can be used to analyze datasets. In this section, we provide a list of some categories of techniques applicable across a range of industries. This list is by no means exhaustive. Indeed, researchers continue to develop new techniques and improve on existing ones, particularly in response to the need to analyze new combinations of data. We note that not all of these techniques strictly require the use of big data—some of them can be applied effectively to smaller datasets (e.g., A/B testing, regression analysis). However, all of the techniques we list here can be applied to big data and, in general, larger and more diverse datasets can be used to generate more numerous and insightful results than smaller, less diverse ones.

3.1.1 A/B testing: A technique in which a control group is compared with a variety of test groups in order to determine what treatments (i.e., changes) will improve a given objective variable, e.g., marketing response rate. This technique is also known as split testing or bucket testing. An example application is determining what copy text, layouts, images, or colors will improve conversion rates on an e-commerce Web site. Big data enables huge numbers of tests to be executed and analyzed, ensuring that groups are of sufficient size to detect meaningful (i.e., statistically significant) differences between the control 28 and treatment groups (see statistics). When more than one variable is simultaneously manipulated in the treatment, the multivariate generalization of this technique, which applies statistical modeling, is often called "A/B/N" testing.

3.1.2 Association rule learning: A set of techniques for discovering interesting relationships, i.e., "association rules," among variables in large databases.27 These techniques consist of a variety of algorithms to generate and test possible rules. One application is market basket analysis, in which a retailer can determine which products are frequently bought together and use this information for marketing (a commonly cited example is the discovery that many supermarket shoppers who buy diapers also tend to buy beer). Used for data mining.

3.1.3 Classification: A set of techniques to identify the categories in which new data points belong, based on a training set containing data points that have already been categorized. One application is the prediction of segment-specific customer behavior (e.g., buying decisions, churn rate, consumption rate) where there is a clear hypothesis or objective outcome. These techniques are often described as supervised learning because of the existence of a training set; they stand in contrast to cluster analysis, a type of unsupervised learning. Used for data mining.

3.1.4 Cluster analysis: A statistical method for classifying objects that splits a diverse group into smaller groups of similar objects, whose characteristics of similarity are not known in advance. An example of cluster analysis is segmenting consumers into self-similar groups for targeted marketing. This is a type of unsupervised learning because training data are not used. This technique is in contrast to classification, a type of supervised learning. Used for data mining.

3.1.5 Crowd sourcing: A technique for collecting data submitted by a large group of people or Community (i.e., the "crowd") through an open call, usually through networked media such as the Web.28 This is a type of mass collaboration and an instance of using Web 2.0.29.

3.1.6 Data fusion and data integration: A set of techniques that integrate and analyze data from multiple sources in order to develop insights in ways that are more efficient and potentially more accurate than if they were developed by analyzing a single source of data. Signal processing techniques can be used to implement some types of data fusion. One example of an application is sensor data from the Internet of Things being combined to develop an integrated perspective on the performance of a complex distributed system such as an oil refinery. Data from social media, analyzed by natural language processing, can be combined with real-time sales data, in order to determine what effect a marketing campaign is having on customer sentiment and purchasing behavior.

3.1.7 Data mining: A set of techniques to extract patterns from large datasets by combining methods from statistics and machine learning with database management. These techniques include association rule learning, cluster analysis, classification, and regression. Applications include mining customer data to determine segments most likely to respond to an offer, mining human resources data to identify characteristics of most successful employees, or market basket analysis to model the purchase behavior of customers.

3.1.8 Sentiment analysis: Application of natural language processing and other analytic techniques to identify and extract subjective information from source text material. Key aspects of these analyses include identifying the feature, aspect, or product about which a sentiment is being expressed, and determining the type, "polarity" (i.e., positive, negative, or neutral) and the degree and strength of the sentiment. Examples of applications include companies applying sentiment analysis to analyze social media (e.g., blogs, micro blogs, and social networks) to determine how different customer segments and stakeholders are reacting to their products and actions.

3.1.9 Signal processing: A set of techniques from electrical engineering and applied mathematics originally developed to analyze discrete and continuous signals, i.e., representations of analog physical quantities (even if represented digitally) such as radio signals, sounds, and images. This category includes techniques from signal detection theory, which quantifies the ability to discern between signal and noise. Sample applications include modeling for time series analysis or implementing data fusion to determine a more precise reading by combining data from a set of less precise data sources (i.e., extracting the signal from the noise).

3.1.10 Spatial analysis: A set of techniques, some applied from statistics, which analyze the topological, geometric, or geographic properties encoded in a data set. Often the data for spatial analysis come from geographic information systems (GIS) that capture data including location information, e.g., addresses or latitude/longitude coordinates. Examples of applications include the incorporation of spatial data into spatial regressions (e.g., how is consumer willingness to purchase a product correlated with location?) or simulations (e.g., how would a manufacturing supply chain network perform with sites in different locations?).

3.1.11Statistics: The science of the collection, organization, and interpretation of data, including the design of surveys and experiments. Statistical techniques are often used to make judgments about what relationships between variables could have occurred by chance (the "null hypothesis"), and what relationships between variables likely result from some kind of underlying causal relationship (i.e., that are "statistically significant"). Statistical techniques are also used to reduce the likelihood of Type I errors ("false positives") and Type II errors ("false negatives"). An example of an application is A/B testing to determine what types of marketing material will most increase revenue.

4. BIG DATA TECHNOLOGIES

There is a growing number of technologies used to aggregate, manipulate, manage, and analyze big data. We have detailed some of the more prominent technologies but this list is not exhaustive, especially as more technologies continue to be developed to support big data techniques, some of which we have listed.

4.1 Big Table: Proprietary distributed database system built on the Google File System. Inspiration for HBase.

4.2 Business intelligence (BI): A type of application software designed to report, analyze, and present data. BI tools are often used to read data that have been previously stored in a data warehouse or data mart. BI tools can also be used to create standard reports that are generated on a periodic basis, or to display information on real-time management dashboards, i.e., integrated displays of metrics that measure the performance of a system.

4.3 Cassandra: An open source (free) database management system designed to handle huge amounts of data on a distributed system. This system was originally developed at Face book and is now managed as a project of the Apache Software foundation.

4.4. Cloud Computing: A computing paradigm in which highly scalable computing resources, often configured as a distributed system, are provided as a service through a network.

4.5 Data mart: Subset of a data warehouse, used to provide data to users usually through business intelligence tools.

4.6 Data warehouse: Specialized database optimized for reporting, often used for storing large amounts of structured data. Data is uploaded using ETL (extract, transform, and load) tools from operational data stores, and reports are often generated using business intelligence tools.

4.7 Distributed system: Multiple computers, communicating through a network, used to solve a common computational problem. The problem is divided into multiple tasks, each of which is solved by one or more computers working in parallel. Benefits of distributed systems include higher performance at a lower cost (i.e., because a cluster of lower-end computers can be less expensive than a single higher-end computer), higher reliability (i.e., because of a lack of a single point of failure), and more scalability (i.e., because increasing the power of a distributed system can be accomplished by simply adding more nodes rather than completely replacing a central computer).

4.8 Dynamo: Proprietary distributed data storage system developed by Amazon.

4.9 Extract, transform, and load (ETL): Software tools used to extract data from outside sources, transform them to fit operational needs, and load them into a database or data warehouse.

4.10 Google File Systems: Proprietary distributed file system developed by Google; part of the inspiration for Hadoop. [37].

4.11 Hadoop: An open source (free) software framework for processing huge datasets on certain kinds of problems on a distributed system. Its development was inspired by Google's Map Reduce and Google File System. It was originally developed at Yahoo! and is now managed as a project of the Apache Software Foundation.

4.12 HBase: An open source (free), distributed, non-relational database modeled on Google's Big Table. It was originally developed by Power set and is now managed as a project of the Apache Software foundation as part of the Hadoop.

4.13 Map Reduce: A software framework introduced by Google for processing huge datasets on certain kinds of problems on a distributed system.[38], Also implemented in Hadoop.

4.14 Mash up: An application that uses and combines data presentation or functionality from two or more sources to create new services. These applications are often made available on the Web, and frequently use data accessed through open application programming interfaces or from open data sources.

4.15 Metadata: Data that describes the content and context of data files, e.g., means of creation, purpose, time and date of creation, and author. [37].

4.16 Non-relational data base: A database that does not store data in tables (rows and columns). (In contrast to relational database).

4.17 Relational database: A database made up of a collection of tables (relations), i.e., data is stored in rows and columns. Relational database management systems (RDBMS) store a type of structured data. SQL is the most widely used language for managing relational databases (see item below).

4.18 Semi-structured data: Data that do not conform to fixed fields but contain tags and other markers to separate data elements. Examples of semi-structured data include XML or HTML-tagged text. Contrast with structured data and unstructured data.

4.19 SQL: Originally an acronym for structured query language, SQL is a computer language designed for managing data in relational databases. This technique includes the ability to insert, query, update, and delete data, as well as manage data schema (database structures) and control access to data in the database.

4.20 Stream processing: Technologies designed to process large real-time streams of event data. Stream processing enables applications such as algorithmic trading in financial services, RFID event applications, fraud detection, process monitoring, and location-based services in telecommunications. Also known as event stream processing.

4.21 Structured data: Data that reside in fixed fields. Examples of structured data include relational databases or data in spreadsheets. Contrast with semi-structured data and unstructured data.

4.22 Unstructured data: Data that do not reside in fixed fields. Examples include freeform text (e.g., books, articles, body of e-mail messages), untagged audio, image and video data. Contrast with structured data and semi-structured data.

4.23 Visualization: Technologies used for creating images, diagrams, or animations to communicate a message that are often used to synthesize the results of big data analyses (see the next section for examples). VISUALIZATION presenting information in such a way that people can consume it effectively.

5. BIG DATA ADVANTAGES

In Big data, the software packages provide a rich set of tools and options where an individual could map the entire data landscape across the company, thus allowing the individual to analyze the threats he/she faces internally. This is considered as one of the main advantages as big data keeps the data safe. With this an individual can be able to detect the potentially sensitive information that is not protected in an appropriate manner and makes sure it is stored according to the regulatory requirements. There are some common characteristics of big data, such as

- Big data integrates both structured and unstructured data.
- Addresses speed and scalability, mobility and security, flexibility and stability.
- In big data the realization time to information is critical to extract value from various data Sources, including mobile devices, radio frequency identification, the web and a growing.

All the organizations and business would benefit from speed, capacity, and scalability of cloud storage. Moreover, end users can visualize the data and companies can find new business opportunities. Another notable advantage with big-data is, data analytics, which allow the individual to personalize the content or look and feel of the website in real time so that it suits the each customer entering the website. If big data are combined with predictive analytics, it produces a challenge for many industries. The combination results in the exploration of these four areas:

- Calculate the risks on large portfolios
- Detect, prevent, and re-audit financial fraud
- Improve delinquent collections
- Execute high value marketing campaigns

6. CONCLUSION

The era of big data is upon us, bringing with it an urgent need for advanced data acquisition, management, and analysis mechanisms. In this paper, we have presented the concept of big data and highlighted the big data value chain, which covers the entire big data lifecycle. The big data value chain consists of four phases: data generation, data acquisition, data storage, and data analysis. Moreover, from the system perspective, we have provided a literature survey on numerous approaches and mechanisms in different big data phases. In the big data generation phase, we have listed several potentially rich big data sources and discussed the data attributes. In the big data acquisition phase, typical data collection technologies were investigated, followed by big data transmission and big data pre-processing methods. In the big data storage phase, numerous cloud-based NoSQL stores were introduced, and several key features were compared to assist in big data design decisions. Because programming models are coupled with data storage approaches and play an important role in big data analytics, we have provided several pioneering and representative computation models. In the data analytics phase, we have investigated various data analytics methods organized by data characteristics. Finally, we introduced the main stay of the big data movement, Hadoop, HDFS and MapReduce and providing security to big data using cloud computing and issues and challenges in cloud and different phases in data analysis and techniques and technologies and big data benchmarks.

REFERENCES

[1]V. MayerSchonberger, K. Cukier, Big Data: A Revolution That Will Transform How We Live, Work, and Think, Houghton Mifflin Harcourt, 2013.

[2]A. Cuzzocrea, Privacy and security of big data: current challenges and future research perspectives, in: Proceedings of the First International Workshop on Privacy and Security of Big Data, PSBD '14, 2014.

[3]Big data, Nature 455(7209) (2008) 1–136

[4]Dealing with data, Science 331(6018) (2011) 639–806.

[5]R.Sumbaly, J. Kreps, S. Shah, The “Big Data” Ecosystem at LinkedIn, in: 2013 ACM SIGMOD International Conference on Management of Data, New York, New York, USA, 22–27 June, 2013.

[6]X. Amatriain, Big & Personal: data and models behind Netflix recommendations, in: The 2nd International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications, Chicago, Illinois, USA, 11 August, 2013.

[7]G. Mishne, Fast data in the era of big data: Twitter's real-time related query suggestion architecture, in: The 2013 ACM SIGMOD International Conference on Management of Data, New York, New York, USA, 22–27 June, 2013.
[8]D. Simon celli, M. Dusi, F. Gringoli, S. Niccolini, Stream-monitoring with Block-Mon: convergence of network measurements and data analytics platforms, ACM SIGCOMM Commun. Rev. 43 (2013) 29–35.

[9]C. Zeng, et al., FIU-miner: a fast, integrated, and user-friendly system for data mining in distributed environment, in: 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, Illinois, USA, 11–14 August, 2013.

[10] V. Mayer-Schonberger, K. Cukier, Big Data: A Revolution That Will Transform How We Live, Work, and Think, Houghton Mifflin Harcourt, 2013.

[11]J.Dean,S. Ghemawa. MapReduce : SimplifiedDataP rocessingonLargeCluster. OSDI'04, Sixth Symposium on Operating System Design and Implementation, SanFrancisco, CA, December,2004.

[12] G. Li, X.Cheng, Research status and scintific science thinking of big data, Bull. China. Acad. Sci. 27(6)(2012) 647–657.

[13]A. Thusoo, et al., Data warehousing and analytics infrastructure at Facebook, in: 2010 ACM SIGMOD International Conference on Management of Data, Indianapolis, Indiana, USA, 6–11 June, 2010.

[14]M. Meier, Towards a big data reference architecture, Master's thesis, Eindhoven University of Technology, October 2013.

[15]R. Schmidt, M. Möhring, Strategic alignment of cloud-based architectures for big data, in: 17th IEEE International Enterprise Distributed Object Computing Conference Workshops, Vancouver, Canada, 9–13 September, 2013.

[16]Y. Demchenko, C. Ngo, P. Membrey, Architecture framework and components for the Big Data Ecosystem, SNE Technical Report, University of Amsterdam, September 12, 2013.

[17]C.E.Cuesta,M.A.MartinezPrieto, J.D. Fernandez, Towards an architecture for managing big semantic data in real-time, in: 7th European Conference on Soft-ware Architecture, Montpellier, France, 1–5 July, 2013.

[18] Ren, Yulong, and Wen Tang. "A SERVICE INTEGRITY ASSURANCE FRAMEWORKFOR CLOUDCOMPUTING BASED ON MAPREDUCE."Proceedings of IEEE CCIS2012. Hangzhou: 2012, pp 240 –244, Oct. 30 2012-Nov. 1 2012.

[19] Hao, Chen, and Ying Qiao. "Research of Cloud Computing based on the Hadoop platform.".Chengdu, China: 2011, pp. 181 – 184, 21-23 Oct 2011.

[20]Bigdata,http://en.wikipedia.org/wiki/Big_data, 2014.

[21]V.R.Brokar,M.J.Carey,andC.Li,”Bigdataplatforms:Wh at'snext?”XRDS,Crossroads,ACM, vol.19,no.1,pp.44-49,2012.

[22]D.Dewitt and J.Gray,” parallel data base systems: The future of high data base systems,”.Commun.ACM,vol.35,no.6,pp.85-98,1992.

[23](2014).Teradata.Teradata,Dayton,OH,USA[online].Available:http://www.teradata.com/

[24](2013).Netezza.Netezza,Marlborough,MA,USA.[online].Available:http://www-01.ibm.com/software/data/netezza

[25](2013).AsterData.ADATA.Beijing,China[online].Available:http://www.asterdata.com/

[26](2013).Greenplum.Greenplum,SanMateo,CA,USA[online].

[27](2013).Vertica[online].Available:http://www.vertica.com/.

[28]S.Ghemawat, H.Gobioff, and S.-T.Leung,"The google file system," in proc.19th ACM symp.operating syst.principles.2003,pp.29-13.

[29]J.DeanandS.Ghemawat,"Mapreduce:Simplified data processing on large clusters,"Commun,ACM,vol.51,no.1,pp,107-113,2008.

[30] T.Hey,S.Tansley and k.Tolle, the fourth paradigm:data-intensive scientific discovery. Cambridge,MA,USA:Microsoft Res.,2009.

[31] A, Katal, Wazid M, and Goudar R.H. "Big data: Issues, challenges, tools and Good practices.". Noida:2013, pp. 404 – 409, 8-10 Aug. 2013.

[32] Lu, Huang, Ting-tin Hu, and Hai-shan Chen. "Research on Hadoop Cloud Computing Model and its Applications.". Hangzhou, China: 2012, pp. 59 – 63, 21-24 Oct. 2012.

[33] Wie, Jiang , Ravi V.T, and Agrawal G. "A Map-Reduce System with an Alternate API for Multi-coreEnvironments.". Melbourne, VIC: 2010, pp. 84-93, 17-20 May. 2010.

[34] C. O'Neil, R. Schutt, Doing Data Science: Straight Talk from the Frontline, O'Reilly Media, Inc.,2013.

[35] Dealing with data, Science 331(6018) (2011) 639–806.

[36] K, Chitharanjan, and Kala Karun A. "A review on hadoop — HDFS infrastructure extensions.". JeJuIsland: 2013, pp. 132-137, 11-12 Apr. 2013.

[37] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung, "The Google file system," 19th ACM Symposium on Operating Systems Principles, Lake George, NY, October 2003 (labs.google. com/papers/gfs.html).