International Journal of Application or Innovation in Engineering & Management (IJAIEM)
Web Site: www.ijaiem.org Email: editor@ijaiem.org
Volume 5, Issue 10, October 2016 **ISSN 2319 - 4847**

Big Data Analytics: A Survey

Gavara Sai Sri Lakshmi Alekhya[1], Dr. E. Laxmi Lydia[2],Dr. Narasimham Challa[3]

[1]M.Tech Research Scholar, Department of Computer Science and Engineering, Vignan's Institute of Information Technology, Visakhapatnam, Andhra Pradesh, India.

[2]Associate Professor, Department of Computer Science and Engineering, Vignan's Institute Of Information Technology, Visakhapatnam, Andhra Pradesh, India.

[3] Professor, Department of Computer Science and Engineering, Vignan's Institute Of Information Technology, Visakhapatnam, Andhra Pradesh, India.

ABSTRACT

In this technical and enthusiastic world, every day we come across new aspects that drag our attention to know something about it. So our enthusiasm portrays to do research on it and make a file of documentary of each and every aspect from ancient history to present and ma extend in future. Later the maintenance and reference of these documentaries used to be difficult to future research. So to solve this, Big Data came into existence. Big Data acts as a datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze. It is a data whose scale, diversity and complexity require new architecture, techniques, algorithms and analytics to manage it and extract value and hidden knowledge from it. For Big Data to scalably implement this, Big Data has different technologies that include MPP database, the Apache Hadoop framework, the Internet and archival storage systems. Of these Hadoop plays a prominent role. Hadoop is an open source software tool used to optimize massive volume of a data for distributed processing of large datasets across large clusters of servers. This Hadoop runs based on HDFS and Map Reduce. After surveying many highly knowledge papers of many professionals, I got a brief idea of Big Data and Apache Hadoop. In this paper I shared brief concepts of Big Data and its analytics, its various definitions survey, benefits, challenges. I also gave a gist of concepts on Hadoop technology and its ecosystem components

.

Keywords: Big Data Introduction, Characteristics, Introduction of Big Data Analytics, Big Data Definitions Survey, Benefits, Impacts on Lives, Challenges, Hadoop Ecosystem and its Components.

1. INTRODUCTION:

Big Data is an omnipresent term in many parts of industry, scientific and academia. As there is a chance for frequent utilization of the same words in different contexts it leads or poses a threat towards the structured evolution of its meaning. Therefore it is mandatory to invest time and put effort in the acceptance of a standard definition of Big Data that would pave the way to its systemic evolution and reduce the confusion related to its usage. In order to describe Big Data we have decided to start from an "as is" analysis of the contexts where the term most frequently appears. Given its remarkable success and its hectic evolution, Big Data possesses multiple and diverse shades of meaning, all of which have the right to exist. By analyzing the most significant occurrences of this term in both academic and business literature, Big Data refers to four key themes: Information, Technologies, Methods and Impact. It's reasonably asserted that the vast majority of references to Big Data encompass one of the four themes listed above. It's believed that having such a definition will enable a more conscious usage of the term Big Data and a more coherent development of research on this subject.

Figure 1: Big Data and Analytics

Nowadays as the information technology spreading widely, most of the data were created and edited under digital as well as exchanged on internet. According to Lyman and Varian [1] estimations, the newly updated or created data stored in digital media devices have been already more than 92 % in 2002, and also the size of these new data was more than five Exabyte's. It is fact that analyzing the problems of huge scale data or finding things from the existing data is more difficult than creating a set of data. When compared to 1930's even though we are having well advanced computers, it is strain to analyze large scale data. So big data can give a relief to some extent as it analyzes the huge data present previously to present data in the required format we want.

Figure 2: Big Data view

Big data involves more than simply the ability to handle large volumes of data. Examples of some firms like Google, eBay, LinkedIn and Facebook were the first organizations to embrace it, and were built from the beginning around big data [1].

2. V'S IN BIG DATA:

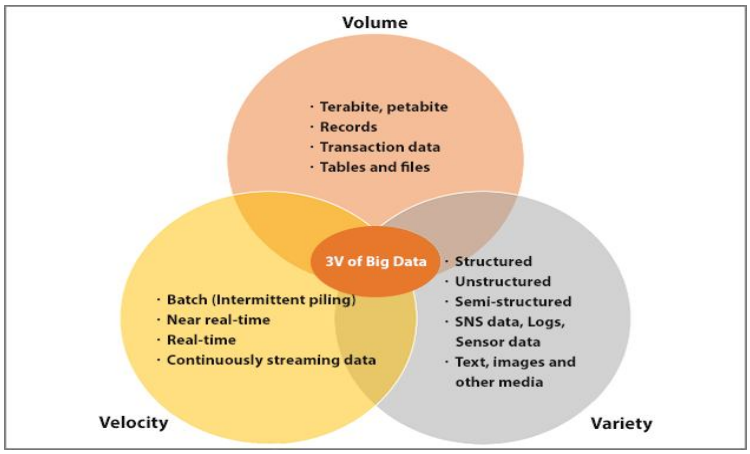


Figure 3: 3V's of BigData

Volume, Velocity, Variety
Doug Laney calls 3V's in his article *3-D Data Management: Controlling Data Volume, Velocity and Variety*, published in 2001, represents key elements for the characteristics of Big Data systems [2].

2.1. Volume:
The first main characteristic of Big Data is 'Volume', which refers to the quality of data that is being manipulated and analyzed in order to obtain the desired results. Some quantified data by counting their records, transactions, tables or file, but some found it more useful to quantify big data in terms of time. One of the examples is, in the U.S. some prefer to keep data available for legal analysis for seven years which is statute of limitations[1].This attribute can represent a challenge as big data can manipulate and analyze a big volume of data which requires a lot of resources that will eventually materialize in displaying the requested results. If we consider an example a computer system, it is limited by latest technology such as regarding the speed of processing operations which is constant while size of the data can be processed unlimited. We need to develop infrastructure to achieve higher processing speeds as more computer power is

needed but at higher costs. It is tedious process to compress huge volumes of data and then analyzing it, which will ultimately prove more ineffective. It takes time to compress the data and also same time it takes to decompress it to analyze. So ultimately displaying of results is delayed. One of the methods of mining through huge amount of data is with OLAP solutions (Online Analytical Processing) (Fig.4. Data warehouse -> OLAP). In this OLAP solution it has tools and multidimensional databases that allow users to easily navigate and extract data from different points of view. Hence, it identifies relations between elements in the database so it can be reached in a more intuitive way. Hence there is an example of how OLAP systems are rearranging the data imported from a data warehouse is given below. Therefore to obtain results, various OLAP tools are used in order for the data to be mined and analyzed.

Figure 4: Data warehouse -> OLAP

2.2. Variety:

The second characteristic is the variety of data. The word itself says that data come from a variety of sources like logs, streams, social media, text data, and semi-structured data from B2B process. This type of data that is stored, analyzed and used is represented here which consists of location coordinates, video files, data sent from browsers, simulations etc. in this variety, the key challenge is how to sort all this data so it can be "readable" by all users who access it and does not create ambiguous results. There are two key variable mechanics of sorting at the beginning. They are: the system that transmits data and the system that receives it and interpret so that can be later displayed (Fig. 5. Send-Receive).

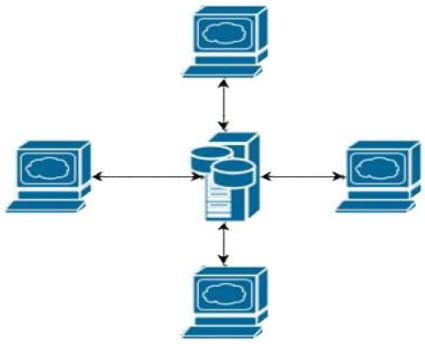


Figure 5: Send-Receive

From these two key aspects, there are some issues that they might not be compatible regarding the content of the data transferred between them. For example, if a browser sends a data that consists of user's location, favorite search terms and so on to Big Data. On other hand, if the Big Data system receives all this information unsorted, so it's difficult to understand whether this user is from "London" or from "orange". To avoid this confusion created in Big Data solutions, all systems that send data should be standardized so that, the send data can be in a logical array that, afterwards, it can be easily analyzed and displayed in a proper manner[3].

2.3. Velocity:

The third characteristic of big data is the velocity that refers to the low-latency, real-time speed at which analytics need to be applied. "Velocity" can be described as the speed that data travels from point A i.e., it can be an end user interface or a server, to point B, which can have the same characteristics as point A is described. With this feature it has a key

issue as well due to high requests that end users have for streamed data over numerous devices (laptops, mobile phones, tablets etc.). This is a challenge for most of the companies which they can't keep up. As we know that data transfers are done at less than the capacity of the systems. Therefore, transfer rates are limited but requests are unlimited, so streaming data in real-time or close to real-time are a big challenge. Hence there is a solution for this point that is to shrink the data that is being sent.

A good example is Twitter where interaction on Twitter consists of text, which can be easily compressed at high rates. But, where as in the case of "Volume" challenge, this operation is still time killing i.e., time-consuming and there will still be delay in sending-receiving data. The solution to this right now is only to invest in infrastructure [3].

Other V's derived From Big Data:
After Laney's "3V's" another two "V's" where added as key aspects of Big Data systems.

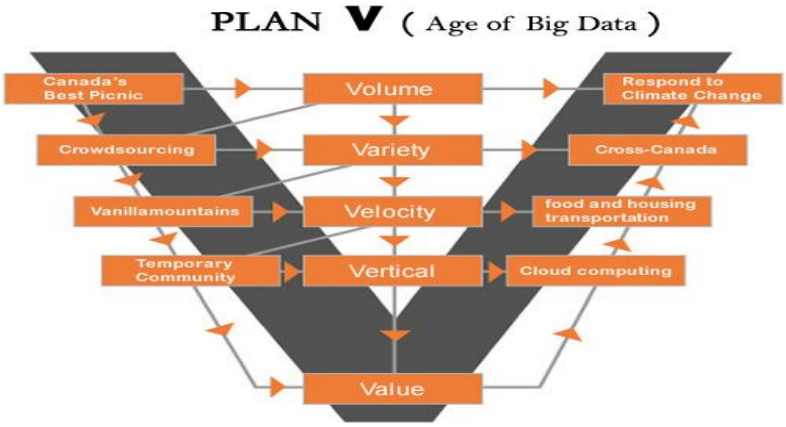


Figure 6: Other V's.

2.4. Value:
The fourth characteristic is "Value" which all measures about the quality of data that is stored and the further use of it. It's simply about the usefulness of data in making decisions. Large quantity of data is being stored from mobile phones call records to TCP/IP logs where the question arises that all together can have any commercial value. If the outcome can't offer insights for a good development and can't be properly managed then there is no use in storing large amount of data. Hence users can deduct important results from the filtered data obtained and can also rank it according to the dimensions they require to find the business trends according to which they can change their strategies. Even data science is exploratory and useful in getting to know the data, but "analytic science" encompasses the predictive power of big data

2.5. Veracity:
The fifth characteristic of Big Data is "Veracity" portrays that the possible consistency of data is good enough for Big Data. It measures the richness of the data representation– text, images video, audio, etc. Data is not only produced from single category but also includes the traditional data and the semi structured data from various resources like web Pages,Web Log Files, social media sites, e-mail, documents.

For example, if A is sending an email to B, B will have the exact content that A sent it, and then it's said to be reliable service and trusted by people. Is not an issue in Big Data, if there is a loss regarding the data stored from one geo-location, because there a hundreds more that can cover that information. Software and Current technologies try to overcome the challenges that "V's" raises. One of the solutions can be Apache Hadoop, which is open source software that its main goal is to handle large amounts of data in areas on able time. Hadoop divides data across a multiple systems infrastructure in order to be processed. Also, Hadoop can be easily found and accessed as it creates a map of the content that is scattered. [3]

3. BIG DATA ANALYTICS [4]:

Nowadays, the data need to be analyzed are not only large, but they are composed of various data formats and types, and even including streaming data [5]. The statistical and data analysis approaches can be changed through the unique features of big data that are "massive, high dimensional, heterogeneous, complex, unstructured, incomplete, noisy, and erroneous," [6]. Through big data makes it possible for us to collect more data to find more useful information, but the

truth is that more data do not necessarily mean more useful information. It may contain more ambiguous or abnormal data. For example, a user may have multiple accounts, or an account may be used by multiple users, which may degrade the accuracy of the mining results [7]. Therefore, various new issues for data analytics has come up, such as privacy, security, storage, fault tolerance, and quality of data etc [8].The big data may be created by handheld device, social network, internet of things, multimedia, and many other new applications that all have the characteristics of volume, velocity, and variety. The following are some perspectives through which the whole data analytics has to be re-examined from:

— From the **volume** perspective, the very first thing we face is the flood of input data because it may paralyze the data analytics. For the wireless sensor network data analysis, Baraniuk[9] stated out that the bottleneck of big data analytics will be shifted from sensor to processing, communications, storage of sensing data, as shown in Fig. 7 which is different from traditional data analytics. This is because sensors can gather much more data, but when uploading such large data to upper layer system, it may create bottlenecks everywhere.

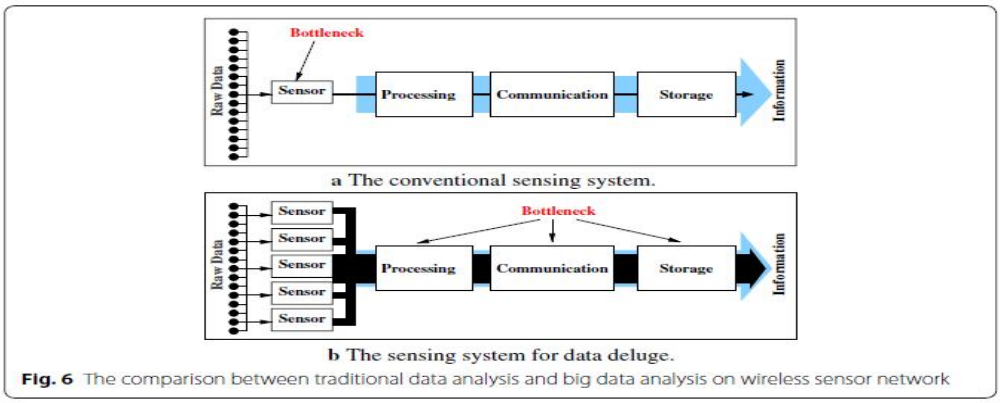


Fig. 6 The comparison between traditional data analysis and big data analysis on wireless sensor network

Figure 7: The comparison between traditional data analysis and big data analysis on wireless sensor network.

From the **velocity** perspective, it is difficult to handle the problems of real-time or streaming data bring up the large quantity of data coming into the data analytics within a short duration but the device and system may not be able to handle these input data. This situation is similar to that of the network flow analysis for which we cannot mirror and analyze everything we can gather.

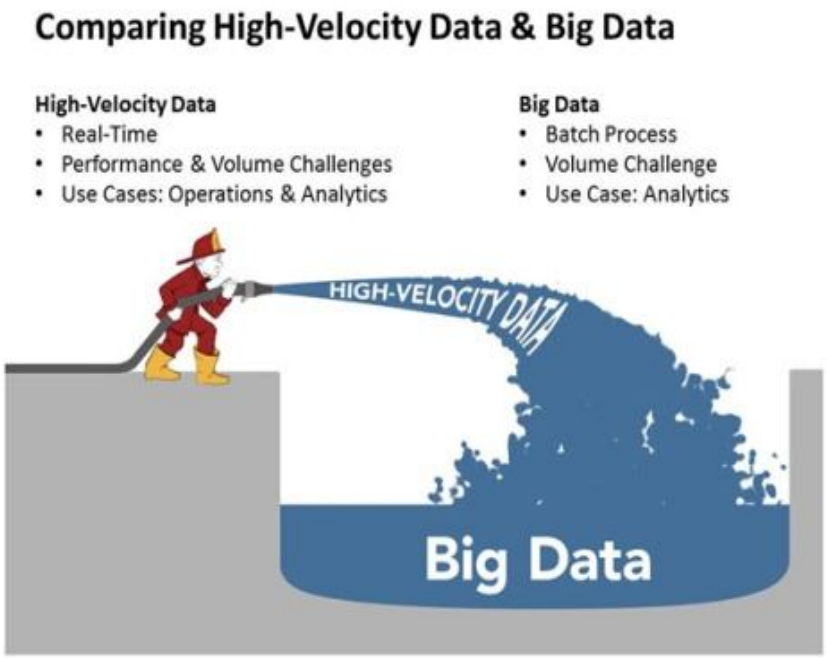


Figure 8: Comparing High-velocity data and Big data

From the **variety** perspective, it makes an issue for the input operators of data analytics as the incoming data may use different types or have incomplete data that includes how to handle them.

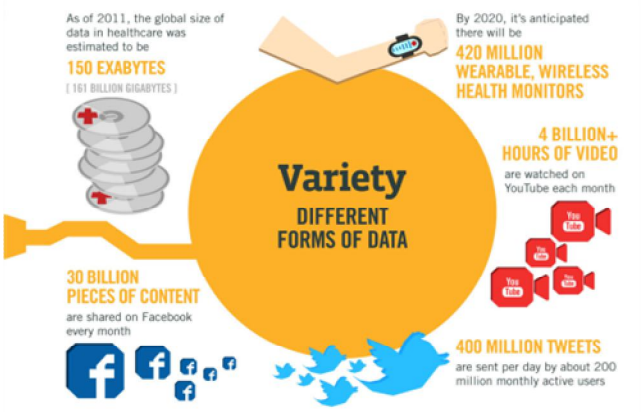


Figure 9: Different forms of data

Four Types of Analytics:

New analytics application include Video and audio application [52] that are needed to process streaming Big data. Another example, Sound monitor to predict earthquakes and satellite images to recognize cloud patterns are which have to be analyzed. The term "analytics" has four types [53]: Quantitative Research and Development, Operational Analytics, Data Scientists and Business Intelligence and Discovery. By combining big data and analytics together we will discover most significant results in business value.

4. A DEFINITION FOR BIG DATA:

"Big Data" lacks a consistent definition that researchers face a first issue in exploring Big Data in financial audits. One website lists 32 definitions and another website had seven more. [10] McKinsey (2011) states that: "Big data refers to datasets where the size is beyond the ability of typical database software tools to capture, store, manage, and analyze." This is similar to the Wikipedia definition, which presumably represents a consensus perspective: "Big data is the term for a collection of large and complex data sets that it becomes difficult to process using on-hand database management tools or traditional data processing applications. There are some challenges of big data that include capture, data curation, storage, search, sharing, transfer, analysis, and visualization. As compared to separate smaller sets with the same total amount of data, the trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, allowing correlations to be found to "spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway traffic conditions."[11]

A convincing definition of a concept is a designer of its scientific development. As Ronda-Pupo and Guerras-Martin [12] suggest, the level of consensus shown by a scientific community on a definition of a concept can be used as a measure of progress of a discipline. Big Data has no universally accepted formal statement denoting its meaning existed instead evolved so quickly and disorderly. Many attempts are made on definition for Big Data, more or less popular in terms of utilization and citation. However, no proposals have prevented authors of Big Data-related works to extend, renovate or even ignore previous definitions and propose new ones. Although Big Data is still a relatively emerging concept that enables the proper development of the discipline among cognoscenti and practitioners which it certainly deserves an accepted vocabulary of reference. By considering the existing definitions and analyzing their commonalities we will propose a consensual definition of Big Data. Consensus in this case comes from the acknowledgement of centrality of some recurring attributes associated to Big Data, and from the assumption made by scholars and practitioners today that they define the essence of what Big Data means. I expect that such a definition would be less prone to attack from previous definitions' authors and users as it is based on the most central aspects associated until now to Big Data. A thorough consensus analysis based on Cohen's K coefficient [13] and co-word analysis, as in [12], goes beyond the scope of this work and is held for future study. Even late technological advancements have lead knowledge of data from unmistakable areas in the course of present decades. The term big data caught the importance of this developing pattern and notwithstanding its sheer volume, big data additionally exhibits other extraordinary attributes as contrasted and conventional data[50].

5. SURVEY OF EXISTING DEFINITIONS OF BIG DATA:

Implicitly Big Data has been often described through success stories or anecdotes, characteristics, technological features, emerging trends or its impact to society, organizations and business processes. It is found that Big Data is used when referring to a variety of different entities including – but not limited to - social phenomenon, information assets, data sets, analytical techniques, storage technologies, processes and infrastructures. I have surveyed multiple definitions that have been proposed previously and characterized into 3 groups.

- ✓ Big Data definitions focuses on enlisting its characteristics
- ✓ Big Data definitions emphasizes the technological needs behind the processing of large amounts of Data
- ✓ Big Data definitions highlights the impact of Big Data advancement on society

I also listed some of them herethat areadapted from different articles.

- Demand cost-effective, innovative forms of information processing for enhanced insight and decision making are assets by High volume, velocity and variety information.[14]
- Volume, Velocity, Variety and Value are the four characteristics for defining big data.[15]
- Complex, unstructured, or huge amounts of data.[17]
- Cardinality, Continuity and Complexity are three other data characteristics can be defined.[18]
- In today's digitized marketplace it is an advantage for organizations to gain competitiveness as big data is a combination of Volume, Variety, Velocity and Veracity that creates an ample opportunity.[16]
- A scalable architecture for efficient storage, manipulation, and analysis is required for extensive datasets, primarily in the characteristics of volume, velocity and/or variety.[20]
- A series of techniques included for the storage and analysis of large and or complex data sets, but not limited to: NoSQL, Map Reduce and machine learning.[23]
- The process of applying the latest machine learning, serious computing power and artificial intelligence, to seriously massive and often highly complex sets of information.[19]
- The processing capacity of conventional database systems was exceeded by that data.[21]
- Processing and handling of that data that cannot be done in a straightforward manner.[22]
- There are too big datasets to fit on a screen.[24]
- The size of datasets is beyond the ability of typical database software tools to capture, store, manage, and analyze.[27]
- The data sets and analytical techniques in applications are so large and complex such that they require advanced and unique data storage, management, analysis, and visualization technologies.[28]
- The interplay of Technology, Analysis and Mythology are rested by a cultural, technological, and scholarly phenomenon's.[25]
- 'More data', 'Messier (incomplete) data', 'Correlation overtakes causality' are the three key shifts phenomenon that brings in the way we analyze information that transform how we understand and organize society.[26]

5.1. Consensual Definition:

By considering both the existing definitions of Big Data and at the main research topics associated to it, we can affirm that the concept of Big Data can be expressed by:

- Characteristics of Information involved describing the 'Volume', 'Velocity' and 'Variety';
- To clarify the unique requirements strictly Specific 'Technology' and 'Analytical Methods' are needed to make use of such Information;
- Transformation into insights and consequent creation of economic 'Value', as the principal way Big Data is impacting companies and society.

From this it is know that the "object" to which Big Data should refer to in its definition is 'Information assets', as this entity is clearly identifiable and is not dependent on the field of application.
Therefore, the proposed formal definition can be:

"Big Data is characterized as a Huge Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value by its Information assets."

When referring directly to the specific technology and methods mentioned in the main definition above definition of Big Data is compatible with the existence of terms like "Big Data Technology" and "Big Data Methods" that should be used.[29]

6. BENEFITS, COSTS, AND EXTERNALITIES OF BIG DATA:

Crafting policy for big data requires that various costs, benefits and externalities be considered. Big data obviously has a number of private benefits and positive externalities. There are also social and economic costs and negative externalities.

6.1. Social and Economic benefits and positive externalities:

Data can help intensify economic efficiency, develop access to social services, strengthen security, personalize services and make increased availability of relevant information and innovative platforms for communications (Kang, 1998; Smolan&Erwitt, 2012). For example, drivers on road are provided with real time information about road congestions through mapping apps, which would allow them to select efficient routes.

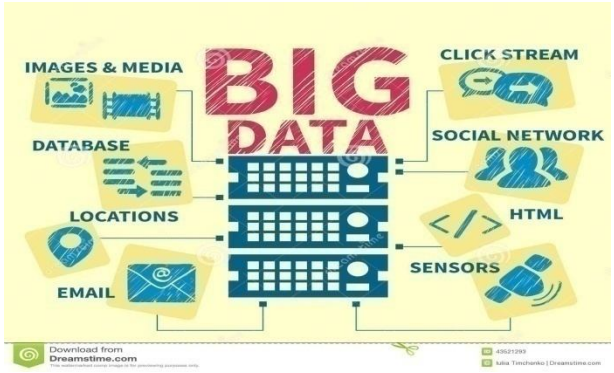


Figure 10: Benefits of Big Data

Organizations are becoming more efficient by improving operations, facilitating innovation and adaptability and optimizing resources allocations through big data. For instance, BMW Car Company detects potential issues and vulnerabilities quickly and eliminates them before new models are launched by combining and analyzing data from test drives of prototypes, workshop reports and other sources. The time period is reduced to analyze some type of data from several months to few days using Big Data and analytics technology. The number of workshop visits and the time required to repair are reduced by timely discovery of the patterns and anomalies in the products and analysis of maintenance and repair data that allowed the company to issue repair instructions on a timely basis (IBM, 2014). In the same way, the big data analytics allowed the yogurt company, Dannon to forecast the demand of its retailer customers more accurately, which led to higher consumer satisfaction, more customer following, less wastes, and a higher profitability (IBM, 2013).

The big data helps Scientists in research that can improve human well-being. Even in implementation of drug therapies big data provides huge volumes of information and patient data that helped to detect drug interactions and design (healthworkscollective.com, 2014 and Smolan and Erwitt, 2012). Data regarding individual patients available through state and federal health information exchanges can contribute to effective drug regulation and reduce direct costs of medical expenditures and indirect costs associated with lower productivity (Abbott, 2013).

The performance of Government agencies services are improved by Big data (Lane et al., 2014). For example, big data helps law enforcement agencies to deploy resources more efficiently, respond quickly and increase presence in crime prone areas (Kang, 1998). The spread of communicable diseases can be fought with the help of big data. For instance, that mining data from Twitter and online news reports could have given the country ′s health officials an accurate indication of the disease ′s spread with a lead time of two weeks which is showed in retrospective analysis of the 2010 cholera outbreak in Haiti showed (Chunara, Andrews, & Brownstein, 2012).

From a variety of sources, Firms have access to a large amount of transactional data. This data is described by Burrows and Savage (2014, p. 3) as a “crucial part of the informational infrastructures of contemporary capitalism”. This type of data can be used to tailor pricing and product offerings, which improve consumer welfare and increase firm’s profits.

6.2. Social and Economic costs and potential negative externalities:

The creepy factor is that it is too intrusive and invasive to personal privacy with big data’s revelation information. To make predictions of a sensitive nature such as sexual orientation and financial status it is possible with non-personal data (Daniels, 2013). If we consider an example of Facebook, researchers have demonstrated that Facebook Likes can

be used to accurately predict highly sensitive personal attributes such as sexual orientation, ethnicity, religious/political views, personality traits, intelligence, degree of happiness, addictive substance consumption, parental separation, age, and gender (Kosinski, Stillwell, &Graepe, 2013). Big data also challenges the Fair Information Practices (FIPs), which are an established set of principles for addressing privacy concerns on which latest privacy or security laws are based (Rubinstein, 2013, Table 2). Big data helps firms arise with better advertising/promotional programs and persuasion attempts, which sometimes could be predatory. Few analysts suggest that firms could determine the probability that someone has suffered from a serious illness and use that information to market unnecessary insurance policies (Drum, 2013). Others state that a medical insurance company can increase premiums in areas with a high incidence of certain diseases (King, 2014). Similarly, predictions like life expectancy based on individual's consumption patterns by some life insurers reportedly and use of that information to offer rates, coverage and other services.[30]

7. HOW BIG DATA CHANGES OUR LIVES: IMPACT

Through anecdotes and success stories of methods and technology implementations depicted, Big Data is impacting our society and our companies to some extent. We can represent a valuable contribution to the creation of knowledge on the subject when these stories are accompanied by proposals of new principles and methodological improvements. The pervasive nature of the current information production and availability leads to many applications spanning or dividing in numerous scientific and industry sectors that can be very distant from each other. Sometimes, in distant domains the same techniques and data have been applied to solve problems. For instance, correlation analysis was leveraged to use logs of Google searches to forecast influenza epidemics [33] as well as unemployment [34] and inflation [35]. There are many existing applications on Big Data and expected to grow: hence, their systematic description constitutes a promising development area for those willing to contribute in the scientific progress in this field. Society can also be adversely impacted by big data. In fact, there are multiple concerns arising from the quick advancement of Big Data [32], of these first being privacy. It is not always true that consequences of using the data will not impact a single individual in an invasive and/or unexpected way when large data sets would normally proceed from actions done by a multitude of individuals. The identifiability of the individual person can be avoided through a thorough anonymization of the data set, even though it is hard to be fully guaranteed as the reverse process of deanonymization can be potentially attempted [36]. On top of freedom in the present, the predictability of future actions, made possible by the analysis of behavioral patterns, poses also the ethical issue of protecting free will in the future.

Figure 11: Impact of Big Data on lives

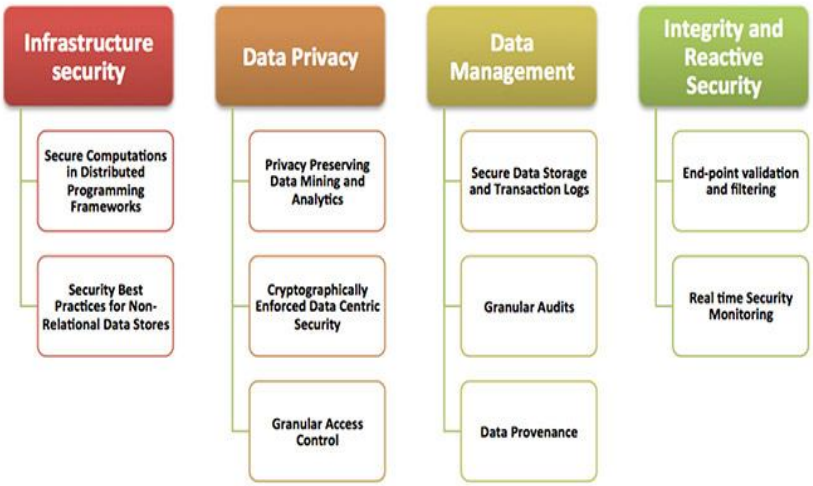
There are other issues to be considered which are related to the accessibility of information: the exclusive control over data sources can become an abuse of dominant position and restrict competition by posing unfair entrance barriers to the marketplace. For example, as Manovich notices [37], "*large social data i.e., transactional data are accessed by only social media companies*" and they have full control over who can access what information. A new digital division can be created when there exists a split between information-rich and data-lacking companies [32] that can slow down innovation in the sector. Within antitrust regulations specific policies will have to be promoted and data is likely to become a new dimension to consider. Not only society but also companies are heavily impacted by the rise of Big Data: the call to arms for acquiring vital skills and technology to be competitive in a data-driven market implies a full realm of business processes and a serious reconsideration of the firm organization [38]. In today's business world the transformation of data into competitive advantage [31] is what makes "Big Data" such an impactful revolution.

8. FIVE MAIN CHALLENGES IN BIG DATA:

The big data wends number of challenges on enterprise, information technology (IT) practitioners and business sponsors that must be addressed before any big data program can be successful. Five of those challenges are[d]:

1. **Uncertainty of the Data Management Landscape** – Various rivals may rise with increasing competitive technologies. So the first challenge is to make the best choices without introducing additional unknowns and risk to big data adoption.
2. **The Big Data Talent Gap** – Implementation helps in the excitement around big data applications to imply that there is a broad community of experts available to help in. So, the second challenge clears this by possessing the talent gap.
3. **Getting Data into the Big Data Platform** –The third challenge that can overwhelm is the unprepared data practitioner, making data accessibility and integration which the scale and variety of data to be absorbed into a big data environment.
4. **Synchronization Across the Data Sources** – The fourth challenge is to incorporate into an analytical platform, the potential for time lags to impact data currency and consistency from diverse sources of data sets.

5. **Getting Useful Information out of the Big Data Platform** – The fifth challenge making a big data syndication lastly, using big data for different purposes ranging from storage augmentation to enabling high-performance analytics is impeded if the information cannot be adequately provisioned back within the other components of the enterprise information architecture.



:

Figure 12: Challenges of Big Data

Apart from this there are other main common challenges [40] that required are:

A. Privacy and Security It is the most important challenges with Big data which is sensitive that includes the personal information, information regarding the people, Social stratification where a literate person would be taking advantages of the Big data predictive analysis, law enforcement.

B. Data Access and Sharing of Information: Sharing data between companies is awkward because of the need to get an edge in business. Sharing of data about their clients and operations threatens the culture of secrecy and competitiveness. The data in the companies' information systems is to be used to make accurate decisions in time it becomes necessary that it should be available in accurate, complete and timely manner.

C. Analytical Challenges: This includesthe main challenging that deals with data volume, data needed to be stored, data needed to be analyzed, finding data points that are important and est advantageous use of data etc., for the making correct decisions.

D. Human Resources and Manpower: Big data is an emerging technology attracting organizations and youth with diverse skills that are not only limited to technical but also should extend to research, analytical, interpretive and

creative ones. Skills are need to be developed in individuals hence requires training programs to be held by the organizations.

E. Technical Challenges: Thisincludes issues regarding Fault Tolerance, Scalability, Quality of Dataand Heterogeneous Data.

9. HADOOP (Highly Archived Distributed Object Oriented Programming):

As we know the term Big data describes the large volume of data that contains data in the form of both structured and un-structured datasets which are very large and complex so that it becomes difficult to process using traditional data processing applications. With this Big data it is difficult to work with using most relational database management systems and desktop statistics and visualization packages. Instead requires "massively parallel software running on tens, hundreds, or even thousands of servers". So to solve this, Hadoop technology is used by big data application to handle the massive data through Hdfs, Map Reduce, Pig, Apache Hive, Hbase and Spark. These technologies handle huge amount of data in KB, MB, GB, TB, PB, EB, ZB, YB and BB.

9.1 Introduction to Hadoop:

Hadoop is the core platform for structuring Big Data, and solves the problem of making it useful for analytics purposes. It is an Open source Framework technology written in Java, helps to store, access and gain large resources from big data in a distributed fashion at less cost, high degree of fault tolerance and high scalability.
Hadoop was initially created by Goug Cutting and Mike Cafarella in 2005 for supporting a distributed search Engine Project. Later it was founded by Apache. Hadoop [54] handles large data from different system like Images,videos, Audios, Folders, Files, Software, Sensor Records, Communication data, Structured Query, unstructured data, Email& conversations, and anything which we can't think in any format. There are various components involved in Hadoop like Avro, Chukwa, Flume, HBase, Hive, Lucene, Oozie, Pig, Sqoop and Zookeeper. The Hadoop Package provides Documentation, source code, location awareness, Work scheduling.

9.2 Characteristics of Hadoop:

a. **Scalable**– when required new nodes can be added without changing the data formats, like as the way which data is loaded, way the jobs or application are written.
b. **Cost effective**– Hadoop brings massively parallel computing to commodity servers whose output is a sizeable decrease in cost which in turn makes it affordable to model all the data.
c. **Flexible**– Hadoop can absorb any type of data, structured or not, from number of sources. Deeper analysis can be done on the Data from multiple sources that can be joined and aggregated in arbitrary ways. Here no restricted data formats considered.
d. **Fault tolerant**– During loss of node, the system redirects work to another location of the data and continues processing without missing a beat.

9.3 Key benefits of Hadoop:

1. Designed to large files
2. Designed to be parallelism
3. Flexible development platform
4. HDFS runs to existing file system

9.4 Hadoop Master/Slave Architecture:

A Hadoop cluster contains one Master node and Many Slave nodes. The master node consists of Data node, Name node, Job Tracker and Task Tracker. Whereas the slave node acts as both a TaskTracker and Data node which holds compute only and data only worker node. The Job Tracker here manages the job scheduling.

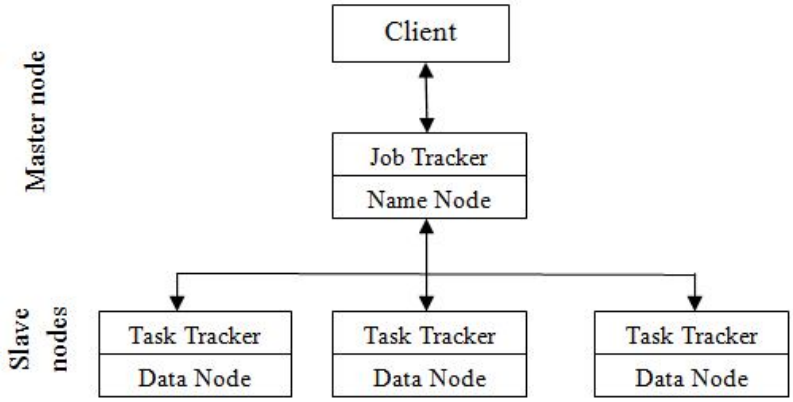


Figure 13: Hadoop Master/Slave Architecture

9.5 Hadoop Ecosystem:

Hadoop ecosystem consists of 2 main components called:

- HDFS (Hadoop Distributed File System) for storage.
- MapReduce for processing

In addition to these it has other components like:

- Hive
- Pig Latin
- Mahout
- Apache Oozie
- HBase
- Flume
- Sqoop

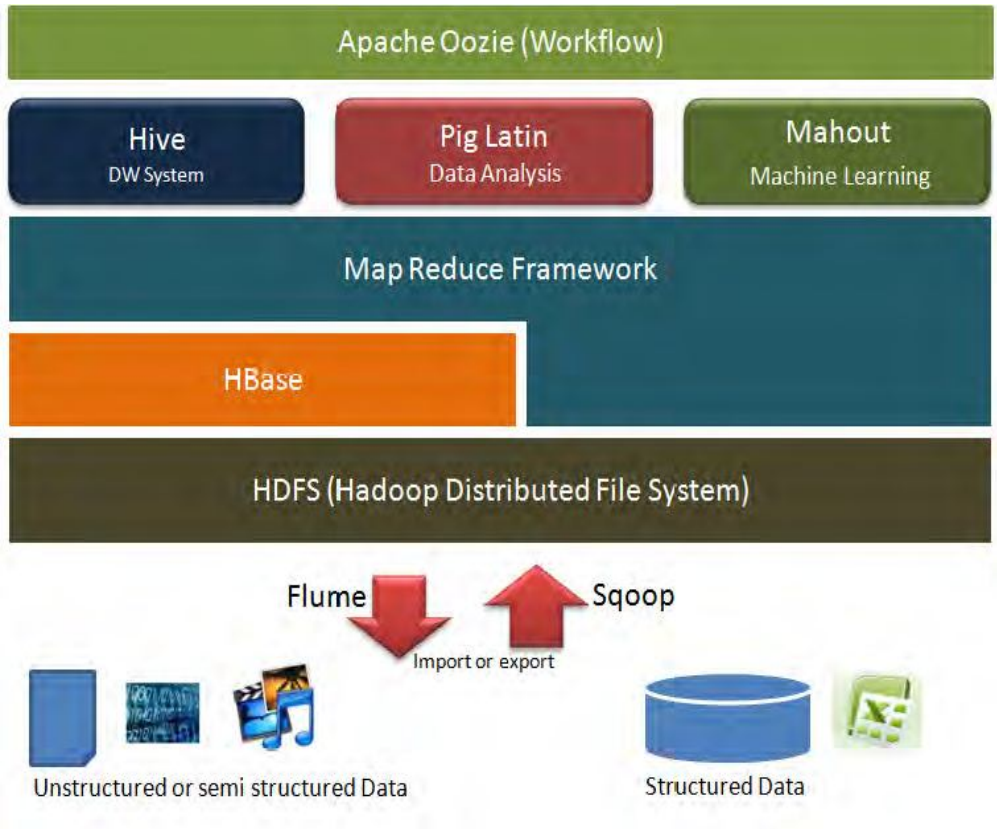


Figure 14: Hadoop Ecosystem

9.6.1 HDFS (Hadoop Distributed File System):

HDFS is one of the primary components of Hadoop cluster. It is the Java portable file system which is more scalable, reliable, distributed in the Hadoop framework environment. In it, each group has a single Name node and group of Data nodes. The operation like "Write Once, Read Many Times" is performed in using Commodity Hardware that provides redundant storage of large amounts of data with low latency. A data node has numerous blocks of same size aside from last piece which have distinctive size. Each block in HDFS has a size of 64 MB or numerous of 64 MB. The communication between these nodes occurs through Remote Procedure calls. There are atleast 3 duplicates of each datanode available. Datanodes correspond with one another torebalance information or duplicate information or to keep high replication of information Name node stores metadata like the name, replications, f ile attributes, locations of each block address and the quick access of metadata is stored in Random Access Memory by Metadata. It reduces the data loss and prevents corruption of the file system. Namenode only monitors the number of blocks in data node and if any block lost or failed in the replication of a datanode, the name node creates another duplicate of the same block. Each block in the data node is maintained with timestamp to identify the current and present status. If any failure occurs in the node, it need not be repair immediately it can be repaired periodically. HDFS [41] can allow more than 1000 nodes by a single Operator. Each block is replicated across many data nodes where original data node is mentioned as rack 1 and replicated node as rack 2 in Hadoop framework. This never supports Data [44] Cache [42][43] due to Large set of data.

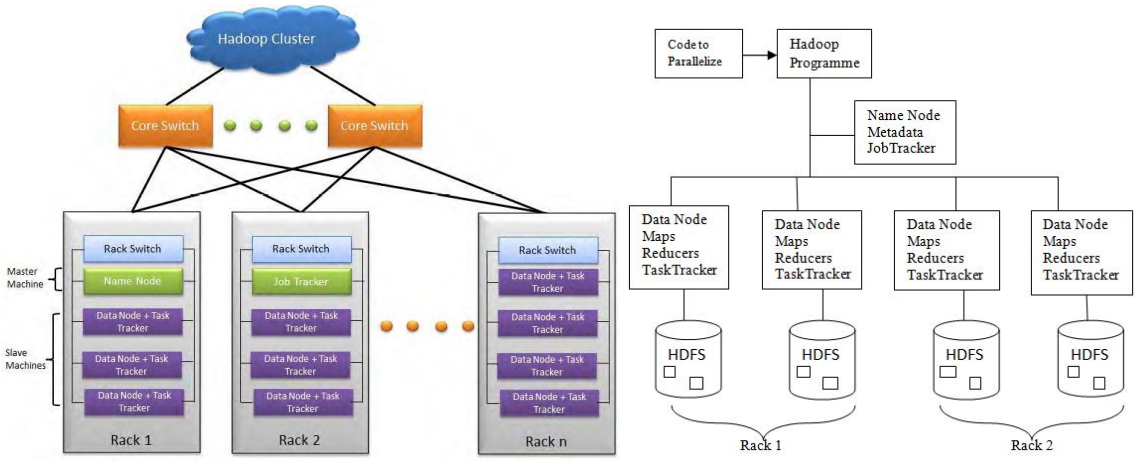


Figure 15: HDFS System

HDFS is designed like Master-slave architecture. The Master (NameNode) manages the file system operations like opening, closing, renaming files and directories and determines the mapping of blocks to DataNodes along with regulating access to files by clients. Slaves (DataNodes) are responsible for serving read and write request from the clients along with performing block creation, deletion, and replication upon instruction from the Master node(NameNode). When a client makes a request of a Hadoop cluster, this request is handled by the JobTracker. The JobTracker, working with the NameNode, will distributes work so closely to the data on which it will work. Here the NameNode is the master of the file system, that provides metadata services for data distribution and replication. The JobTracker schedules map and reduce tasks into available slots at one or more TaskTrackers. The TaskTracker(the slave portions of the distributed file system) working with the DataNode to execute map and reduce tasks on data from the DataNode. After the completion of the map and reduce tasks, the TaskTracker notifies the JobTracker, who identifies that when all tasks are completed and eventually notifies the client of job completion.

Master {Jobtracker} acts as interaction between users and the map/reduce framework. When a map/reduce job is submitted, Jobtracker puts it in a queue of pending jobs and executes them by first-come/first-served basis and then manages the assignment of map and reduce tasks to the tasktrackers. Slaves {tasktracker} execute tasks given by instruction from the Master {Jobtracker} and also handle data motion between the maps and reduce phases.

9.6.2 MapReduce:

Hadoop Map Reduce is a software system framework for for effortlessly composing applications that process large amounts of data (multi-terabyte data-sets) in-parallel on substantial groups (a huge number of hubs) of item in an exceedingly reliable, fault-tolerant manner.A Map Reduce job splits the input data-set into freelance chunks in a fully parallel manner that is processed by the map tasks. The framework sorts the input to the reduce tasks, which are outputs of the maps. Here,both the input and the output of the work are keep in an exceedingly file-system. The framework takes into consider of scheduling tasks, monitoring them and re-executes the failing tasks.The Map Reduce framework consists of a single master Job tracker and a single slave Task tracker per cluster-node. The master is

responsible for scheduling the jobs' element tasks, monitoring them and re-executing the unsuccessful tasks on the slaves. The slaves execute the tasks as directed by the master i.e., slaves follow masters order.

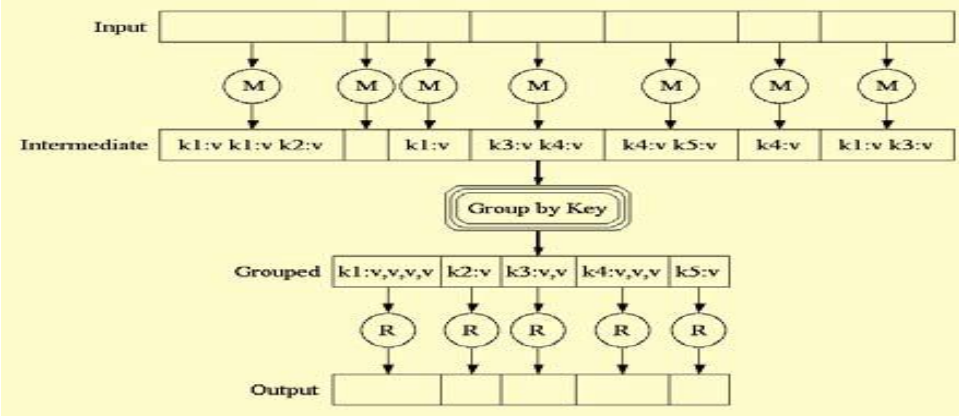


Figure 16: Map Reduce operations

Inputs and Outputs:
The Map Reduce framework operates completely on <key, value> pairs, that it is conceivable of different varieties where the framework views the input to the job as a group of <key, value> pairs and produces a setof<key, value>pairs as the output of the work. The key and value pairs have to be serializable by the framework and hence got to implement the Writable interface. Therefore, the key classes have to implement the Writable Comparable interface to facilitate sorting by the framework. Here is an example of operation of Input and Output types of a Map reduce job (shown in figure 16): (input) <k1, v1> -> map -><k2, v2> -> combine -><k2, v2> -> reduce -><k3, v3>(output).

Yet unbending structure design Unique MapReduce executes jobs in astraightforward.MapReduce changes step ("map"), a synchronization step("shuffle"), along with a stage to join results from every one of thenodes in a cluster ("reduce"). Accordingly to defeat theinflexible structure of guide and diminish Dr.E.Laxmi Lydia[51] proposed as of late presented Apache Spark – both of which give a handling model to breaking down enormous information which is primary contender for "successor to MapReduce" today is Apache Spark.

9.6.3 Pig:
Pig was initially developed at Yahoo Research around 2006 but moved into the Apache Software Foundation in2007 to allow individuals using Apache Hadoop tofocus a lot of on analyzing massive data setsand pay less time having to put inwriting mapper and reducer programs.The Pigprogramming language is meant to handleany reasonably data—hence the name!
Pig consists of a two components, first is the language called as Pig Latin and secondly an execution environment where Pig Latin programs areexecuted. [46].

PigLatin:
Piglatin is a dataflow language used by Pig. It is a typeof language where you program by connecting things together. Even Pig can handle complex data structure, eventhose who have levels of nesting. Pig has two types of execution environment: local and distributed environment. When distributed environment cannot be deployed Local environment is used for testing. Pig Latin program is collection of statements that can be an operation or command.
Let's us consider at the programming languageitself so you can see how it's significantly easierthan having to write mapper and reducerprograms.
1. The first step in Pig programis to LOAD the data you wish to control fromHDFS.
2. Next we run the data through a group oftransformations (which, under the covers, aretranslated into a set of mapper and reducertasks).
3. Finally, we DUMP the data to the screen or we STORE the results in a file somewhere.

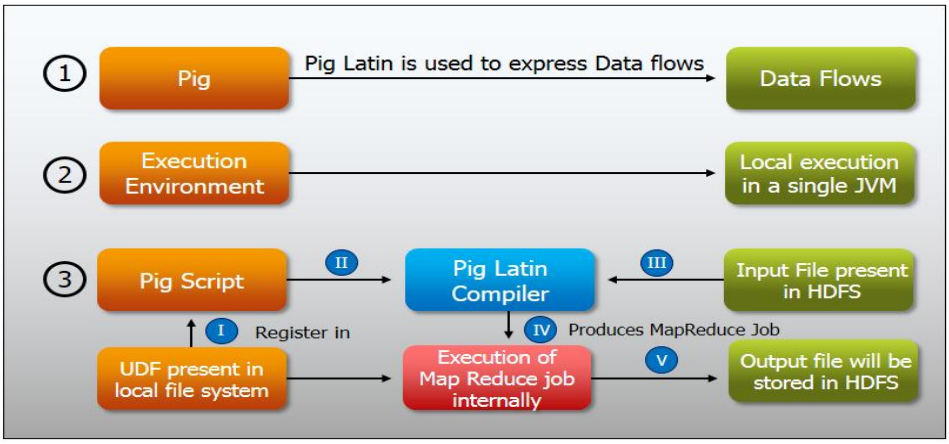


Figure 17: Apache Pig Operations

9.6.4 Hive:

Apache Hive is a data warehouse system for Apache Hadoop [45].Hive is a technology which is developed byFacebook that turns Hadoop into a datawarehouse which complete with an extension of sql for querying.Hive is used as HiveQL which is a declarative language. In piglatin, dataflow is described but inHive results must be described. Hive by itself find out a dataflow to get those results. Hive must have a schema that is more than one.
Hive must be configured in three different ways before use. They are:

- By editing a file hive-site.xml,
- By hiveconf option in Hive command shell
- By using set command.

9.6.5 Oozie:

Oozie is a java based web-application that runs in a java servlet that uses the database to store definition of Workflow that is a collection of actions. Hadoop jobs are managed y this.

9.6.6 HBases:
Hbase is non-relational columnar distributed column oriented database where as HDFS is file system. It is built and run on top of HDFS system. It is a management system that is open-source, versioned, and distributed based on the Big Table of Google. It is written in Java. It is serving as the input and output for the Map Reduce. For instance, read and write operations involve all rows but only a small subset of all columns.

9.6.7 Sqoop:
Sqoop is a tool used to transfer the data from relational database environments like oracle, mysql and postgresql into hadoop environment. It is a command-line interface platform is used for transferring data between relational databases and Hadoop.

9.6.8 Mahout:
Mahout is a library for machine-learning and data mining which is divided into four main groups: collective filtering, categorization, clustering, and mining of parallel frequent patterns. The Mahout library belongs to the subset that can be executed in a distributed mode and executed by Map Reduce.

9.6.9 Flume:
Flume is an open source programming which is made by cloud era to go about as an organization for gathering and moving enormousmeasure of data around a Hadoop bundle as data is conveyed or in no time. Crucial use case of flume is togather log records from all machines in cluster to continue on them in a united store. For instance, in HDFS, we have to make data streams by building up chains of sensible center points and partner them to source andsink.

10. CONCLUSION:

It impossible to handle huge amounts of secure and research oriented data at instant amount of time. But big data as made it possible by it features and technology that as solved many problems related to storage, managing, analyzing

etc. this made the work of every developer easier with ample advantages. For suppose if we consider the computation time, there is no doubt at all that parallel computing is one of the important future trends to make the data analytics work for big data, and on this account the technologies of cloud computing, Hadoop, and map-reduce will play the important roles for the big data analytics. Therefore the scheduling method is another future trend to handle the computation resources of the cloud based platform and to finish the task of data analysis as fast as possible. By using efficient methods to reduce the computation time of input, comparison, samplingand a variety of reduction methods will play an important role in big data analytics. As these methods usually do not consider parallel computing environment,how to make them work on parallel computing environment will be a future researchtrend. Thus I conclude that the Big data and Hadoop had made a revolutionary achievement in part of everyone's world and will continue to be.

Reference:

[1] Bogdan NEDELCU, University of Economic Studies, Bucharest, Romania "About Big Data and its Challenges and Benefits in Manufacturing",
[2] TWDI Research –"Big Data Analytics".
[3] Alexandru Adrian TOLE Romanian – American University, Bucharest, Romania " Big Data Challenges"
[4] "How much information 2003" Tech. Rep, 2004. [Online]. Available: http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/printable_report.pdf., Lyman P, Varian H.
[5] "Big data analytics". TDWI: Tech., Russom P,Rep ; 2011.
[6] Ma C, Zhang HH, Wang X." Machine learning for big data analytics in plants." Trends Plant Sci. 2014;19(12):798–808.
[7] "Critical questions for big data". Boyd D, Crawford K., Inform Commun Soc. 2012;15(5):662–79.
[8] Katal A, Wazid M, Goudar R." Big data: issues, challenges, tools and good practices". In: Proceedings of the International Conference on Contemporary Computing, 2013. pp 404–409.
[9] Baraniuk RG." More is less: signal processing and the data deluge". Science. 2011;331(6018):717–9.
[10] "Definitions-of-big-data-you-should-know-about".html respectively http://www.opentracker.net/article/definitions-big-data; http://timoelliott.com/blog/2013/07/7-. Accessed 9/3/2015 9:53:42 AM.
[11] http://en.wikipedia.org/wiki/Big_data.
[12] L. Á. Guerras and Martin G. A. Ronda-Pupo, Strateg. Manag. J. 33, 162 (2012).
[13] J. Cohen, Educ. Psychol. Meas. 20, 37 (1960).
[14] M. A. Beyer and D. Laney, "The Importance of "Big Data": A Definition", Gartner report (2012), pp. 1–9.
[15] "Big Data for the Enterprise", J. Dijcks, Oracle report (2012).
[16] M. Schroeck, R. Shockley, .et.al.,"Analytics: The Real-World Use of Big Data", IBM report (2012), pp. 1–20.
[17] Intel, "Big Data Analytics. Intel's IT Manager Survey on How Organizations Are Using Big Data", Intel report (2012).
[18] S. Suthaharan, ACM SIGMETRICS Perform. Eval. Rev. 41, 70 (2014).
[19] Microsoft, (2013), available at https://www.microsoft.com/en-us/news/features/2013/feb13/02-11bigdata.aspx.
[20] NIST Big Data Public Working Group,"*Big Data Interoperability Framework: Definitions" (draft)* (2014).
[21] E. Dumbill, "Big Data", (2013).
[22] D. Fisher, R. DeLine, M. Czerwinski, and S. Drucker, Interactions (2012).
[23] J. Ward and A. Barker, preprint arXiv:1309.5821 (2013).
[24] B. Shneiderman, in "*Proc. 2008 ACM SIGMOD Int. Conf. Manag. Data* (2008)", pp. 3–12.
[25] D. Boyd and K. Crawford, Information, Commun. Soc. 15, 662 (2012).
[26] V. Mayer-Schönberger and K. Cukier, *Big Data: "A Revolution That Will Transform How We Live"*(2013).
[27] J. Manyika, M. Chui, B. Brown, and J. Bughin, "*Big Data: The next Frontier for Innovation, Competition, and Productivity"*(2011).
[28] H. Chen, R. Chiang, and V. Storey, MIS Q. 36, 1165 (2012).
[29] Andrea De Mauro, Marco Greco, Michele Grimaldi"What is Big Data? A Consensual Definition and a Review ofKey Research Topics",
[30] By: NirKshetriKshetri, N. (2014). "Big data's impact on privacy, security and consumer welfare". *Telecommunications Policy*, *38(11)*, 1134-1145. doi: 10.1016/j.telpol.2014.10.002, Elsevier: http://dx.doi.org/10.1016/j.telpol.2014.10.002
[31] A. McAfee, E. Brynjolfsson, T. H. Davenport, D. Patil, D. Barton, Harv. Bus. Rev. 90, 61 (2012).
[32] D. Boyd and K. Crawford, Information, Commun. Soc. 15, 662 (2012).
[33] J. Ginsberg, et.al., Nature 457, 1012(2009).
[34] N. Askitas and K. F. Zimmermann, Appl. Econ. Q. 55, 107 (2009).
[35] G. Guzman, J. Econ. Soc. Meas. 36, 119 (2011).

[36] A. Narayanan and V. Shmatikov, in *Proc. - IEEE Symp. Secur. Priv.* (2008), pp. 111–125.
[37] L. Manovich, Debates Digit. Humanit. 1 (2011).
[38] T. Pearson and R. Wegener, Big Data: The Organizational Challenge, Bain & Company report (2013).
[39] ”Addressing Five Emerging Challenges Of Big Data” David Loshin, President of Knowledge Integrity, Inc.
[40] ”A Survey on Challenges and Advantages in Big Data” by LENKA VENKATA SATYANARAYANA ,Dept. of CSE, IJCST Vol. 6, Issue 2, April - June 2015
[41] “Hadoop and HDFS:Storage for Next Generation Data Management”, Cloudera, Inc, 2014.
[42] P. Victer Paul, et.al., "Efficient service cache management in mobile P2Pnetworks", Future Generation Computer Systems, Elsevier, Volume 29, Issue 6, August 2013, Pages 1505–1521. ISSN: 0167-739X.
[43] "An Effective Model for QoS Assessment in Data Caching in MANET Environments", by N. Saravanan, R. Baskaran, M. Shanmugam, M.S. SaleemBasha and P. Victer Paul, International Journal of Wireless and Mobile Computing, Inderscience, Vol.6, No.5, 2013, pp.515-527; ISSN: 1741-1092.
[44] R. Baskaran, P. Victer Paul and P. Dhavachelvan, "Ant Colony Optimization for Data Cache Technique in MANET", International Conference on Advances in Computing (ICADC 2012), “Advances in Intelligent and Soft Computing" series, Volume 174, Springer, June 2012, pp 873-878, ISBN: 978-81-322-0739-9.
[45] Apache Hadoop. Available at http://hadoop.apache.org
[46] AshishThusoo, et.al., Murthy “Hive – A Petabyte Scale Data Warehouse Using Hadoop” By Facebook Data Infrastructure Team
[47] SURVEY PAPER ON BIG DATA ANALYTICS USING HADOOP TECHNOLOGIES” by Vikas Goya (Assistant Professor), Deepak Soni (Student), CSE Department, MIMIT, Malout, India
[48] Dr. E. Laxmi Lydia, Associate Professor, Department of Computer Science and Engineering, et.al.,”Analysis of Big data through HadoopEcosystem Components like Flume,MapReduce, Pig and Hive”Dr. E. Laxmi Lydia et al. / International Journal of Computer Science Engineering (IJCSE)
[49] ”A Survey of Big Data Processing in Perspective of Hadoop and Mapreduce” by D.Usha A and AslinJenil A.P.SA, A Hindustan University, Chennai, Accepted 05 March 2014, Available online 01 April 2014, Vol.4, No.2 (April 2014).
[50] ”A Literature Inspection on Big Data Analytics” by Dr.E.Laxmi Lydia (Associate Professor), Dr. M.BenSwarup(Professor) and M. Vijay Laxmi (student), International Journal of Innovative Research in Engineering & Management (IJIREM), ISSN: 2350-0557, Volume-3, Issue-5, September-2016.
[51] Dr.A.Krishna Mohan,Dr.E.Laxmi Lydia, “Implementing K-Means for Achievement Study between Apache Spark and Map Reduce” International Journal of Innovative Research in Engineering & Management (IJIREM), ISSN: 2350-0557, Volume-3, Issue-5, September-2016.
[52] “Big Data: Challenges, Opportunities and Cloud Based Solutions” by Hamid Bagheri ,Abdusalam Abdullah Shaltooki, International Journal of Electrical and Computer Engineering (IJECE), Vol. 5, No. 2, April 2015, pp. 340~343, ISSN: 2088-8708
[53] NeilRaden, Big Data Analytics Architecture, 2012, Hired Brains, Inc.
[54] An Oracle White Paper, “Leveraging Massively Parallel Processing in an Oracle Environment for Big Data”, November 2010.