Application of Big Data Analytics in Cloud Computing via Machine Learning

Somya Goyal[a,1], Anubha Parashar[b], Anita Shrotriya[c],
[a] *Vaish College of Engineering, Rohtak, India*, [bc] *Manipal University Jaipur, India.*

Abstract. Analysis of big data can be done in many ways. To define trends in Big Data we need to concentrate on the biggest challenges faced by this technology and various strategies have been developed in order to process such large data efficiently. We usually describe this by three factors, popularly known as 3Vs i.e. Volume, Velocity and Variety, which describes most of the features of data. The goal of this chapter is to provide traditional meaning and definitions of big data and how this technology is evolved in order to meet today's requirements and challenges. In this chapter we will describe all the attributes of big data i.e. popularly known as 9Vs. Business Intelligence contains these nine attributes on the basis of statistical models or hypothesis in order to provide better predictions and outcomes for any research or results. Machine Learning provides the platform where the big data analysis can be done using cloud computing. With the help of this chapter academicians, business people and researchers can easily find opportunities as a solution for their required purpose.

Keywords. Big data, Machine learning, Cloud Computing, Statistics Analysis, 3Vs, 4Vs, 5Vs, 6Vs, 7Vs, 9Vs.

1. Introduction

Big Data is highly renowned and popular now-a-days, No common perspective can be established regarding what it is meant for. As per the field specialists, the ETL procedure to Extract, Transform and Load for huge databases is key to the term Big Data. Another widely accepted description about Big Data is founded above three pillar-like data-attributes. These are also called 3V's of Big Data namely volume, velocity and variety. But, this is not the fair description for the capabilities of Big Data precisely. To define actually the term Big Data, a review of periodic timeline in this arena is desirable and then the journey of evolution of today's current technology can be accurately predicted, shown in figure 1.

[1] Somya Goyal, Computer Science and Engineering, Vaish College of Engineering, Rohtak, India; E-mail: somyagoyal1988@gmail.com.

Figure 1. Big data analysis.

Looking at the history, this term itself appears not so clear and precise; rather it is apparently an ambiguous term. Because it is comprised of two words "Big + Data" and the direct implication of it is about something is big [1].The relevance with the key data-object is not implied by the term. Like, what data is big? Or, how much the data is big? The range of answers to such questions is really very wide and diverse. In-fact, the size of "data" is continuously evolving. Internet can be considered [2] as a source of answering the question "How much the data is big?" The answer can be "the volume is Tera or Zetta (in bytes). According to the research contributed by Cisco after studying the periodic growth rate of data, 2015 was the time when we entered the Zetta-World [2]. Table 1 shows the impact of data-size in our daily life by considering the size (on the average) of typical files, we use in routine life.

Table 1. Size of Typical Data

S.No.	File Type	Average Size
1	Still Image	254 – 812 KB
2	Electronic Text	1 - 5 MB
3	Audio	3.5 - 5.8 MB
4	Video	100 - 120 GB

The overall objective is to discuss the timeline of Big Data evolution and reveal that it is not about only 3V's (namely Volume, Velocity and Variety), but in reality squared of 3V's resulting 9 data aspects on which the foundation of Big Data has been laid. The journey from the history to the present scenario in the field of Analytics of Big Data is driven about by all the 9V's not only the initially defined 3V's, shown in figure 2. Through this chapter, we try to answer some core questions precisely regarding the term Big Data. It covers what makes Big Data essential? What is the kind of problem being dealt by Big Data? What disbeliefs are made about Big Data? What are the common problems that are being wrongly considered as Big Data Analytics?

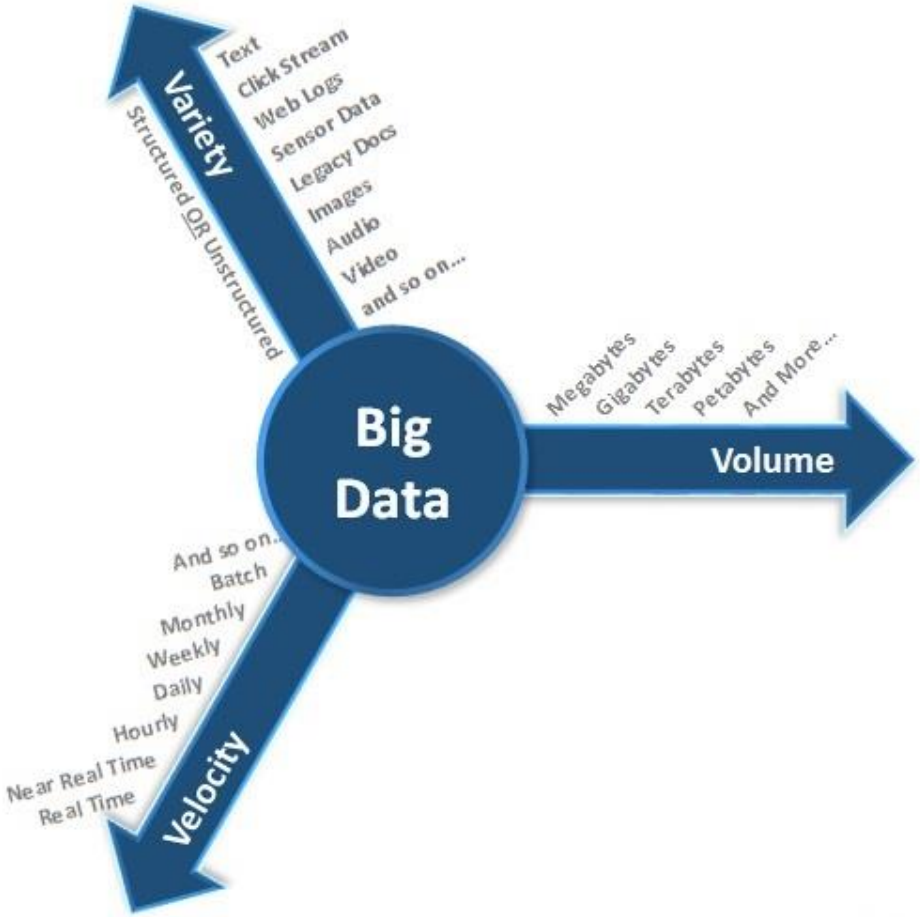

Figure 2. 3Vs of Big Data.

This chapter addresses all the aforementioned issues by making the analytical study of evolution of Big Data from the historic point of view. The latest technologies which are being used by current market professionals for big data analytics are also covered in this chapter. The chapter is structured into nine sub-sections as:

1) Evolution of the term Big Data

2) Meaning of 3V's, 4V's and 6V's as expanding dimensions
3) Definition of Big Data
4) Big Data in Machine Learning
5) Big Data in Cloud Computing
6) Hadoop, HDFS, Map-Reduce, Spark and Flink
7) ML combining CC to form BDA
8) Conclusion

2. Evolution of the term Big Data

The question "what makes Big Data essential?" can be answered only after considering the entire timeline of evolution of Big Data from the initial birth stage through mediocre stages to the present stage. This would give a clear vision leading to the justified definition for the term Big Data.

1.1. *The Birth of Big Data*

Various reports are made by several researchers on the evolution of BD and its ongoing growth. An historic study [3] about BD has been carried out by Gil Press highlighting 1944 as the birth year for BD. The basis of his report was Rider's [4] contribution. He reported the journey Big Data evolution starting from its birth in 1944 up to 68th birthday in 2012. This entire lifespan of BD is beautifully illustrated through 32 events in relation to Big Data. He demonstrated through his research that a very thin boundary between the big sized growing data and the actual meaning of the term Big Data has been violated somewhere. Gil Press extended this study for one more life year of BD that was 2013. His report covers a huge number of events related to Big Data along with Data Science. Yes, this was the next term complementing Big Data, Data Science, shown in figure 3.

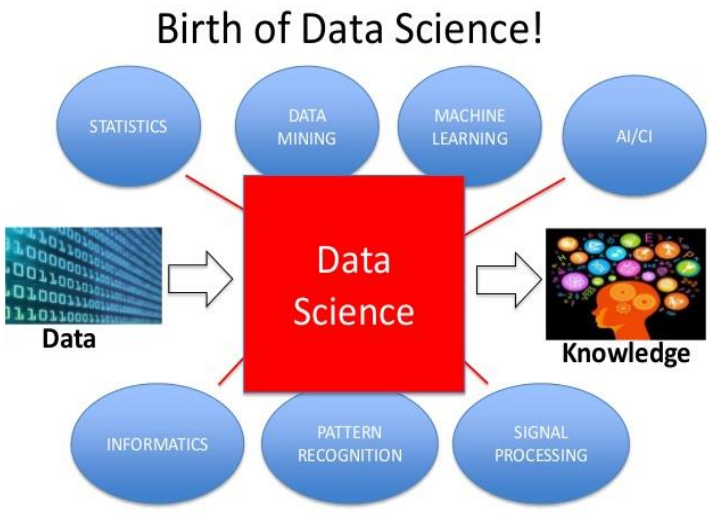

Figure 3. Big Data came into existence.

Another crucial case study was published by Frank [5] who reported 1880 as the birth year of Big Data. His work was inspired by the problem raised in 19th century based on some statistical grounds. It was about a survey to be made over North-American citizens counting to almost 50 million. Rather, he considered that statistical problem as basis for his work, but even till today Big Data and Statistics are totally different. Of course, BD can Problems may be overlapping statistical elements. Winshuttle [6] argued for the birth of Big Data in 19th century.

The point of argument was about the declaration of the very huge and highly complicated data which could be processed using traditional methods as Big Data. His work was dominated by ERP (Enterprise Resource Planning) and motivated by Cloud based implementation. An important data-forecasting report was made by Winshuttle about high growth of data in 2020. Under his review, he covered 220 years of total life span of BD. He remarkably illustrated SAP events along with HANA like products.

In Bernard's review [7], the largest study of Big Data historic development was traced with the birth of Big Data in 18,000 BCE. Marr [7] emphasized that the focus of study should be shifted to the pillar concepts. The fundamental key concepts on which Big Data foundation laid were the diverse techniques and methods to data / information capture, its storage, analysis and retrieval. According to Francis, E. Larson coined the term "Big Data" [9], also presented a write-ups for Harper's magazine and Washington post.

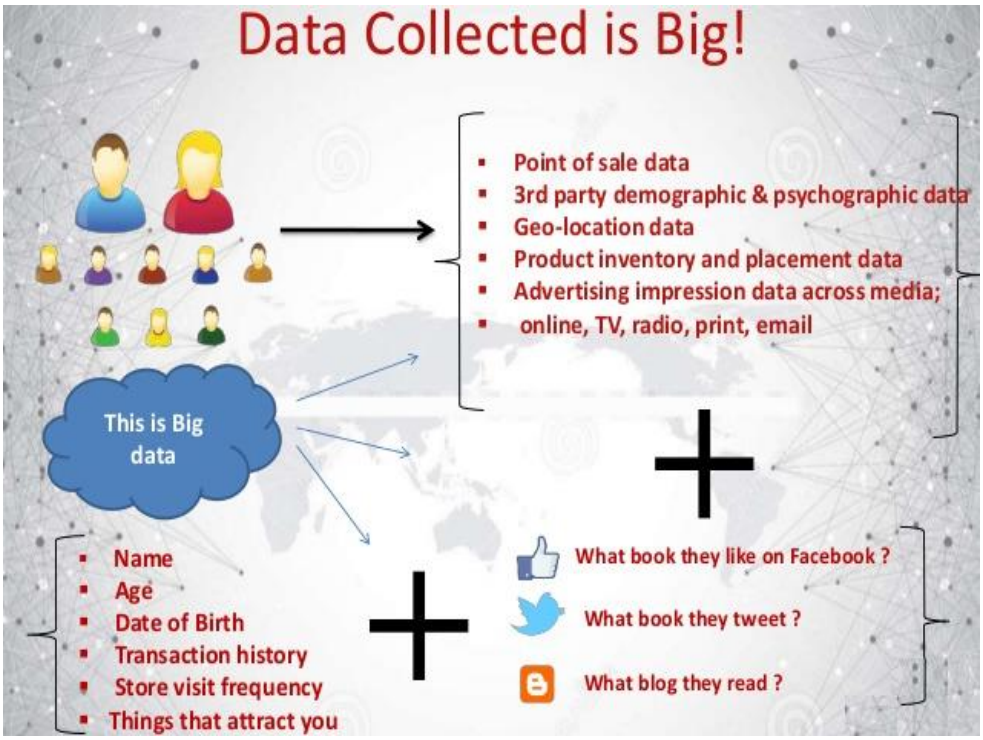

Figure 4. How big is Big Data?

The review made by Steve Lohr lit the fire of argument [10] among the interested reviewers. He produced a strong statement that belittled the others' work regarding the overall objective of the case studies which were being carried out by multiple contributors. He stated that just the adoption of the term or the birth of initial phrase "Big + Data" is not the intention of review, because Big Data is such a general term that the initial implication of this term is not worthy, rather early usage of this term leading to the current meaning is target. And, of course it is not about only the big amount of data. The meaning lies in the early use of this term, connotation to the present meaning. The present implication of the term "Big Data" is not about the large volume of the data only shown in figure 4, it's about the handling of the multi-dimensional data in multiple variant approaches. It was a turning point in the life span of Big Data history and many technologies were evolved shown in figure 5.

Figure 5. Evolved Technologies

From this turning edge, the review was steered in a new and appropriate direction. Cox and Ellsworth [8] made a remarkable contribution in this direction because a more precise interpretation to the current scenario of Big Data was made by them. They proposed the statement "the problem of Big Data is when the main memory or even the local disk gets exhausted with the data". This is not the precise description of current meaning of today's Big Data, but can be considered best connotation.

Visualizing .org [11] reviewed the lifespan of Big-Data in the newly steering direction. It produced a timeline about the usage along with the implementation of BD. It illustrated the events in relation to the Big Data (BD). Those events were uploaded by various IT Corporation and websites like Google, Twitter, and Facebook etc. The influence caused by Hadoop in the evolution of BD was emphasized by Visualizing .org. The evolution timeline produced is shown in fig. 6.

The existence of conflicting and perplexing reviews lead to several debates. Some reviewers were in favor of big data implications and others were against of it. The agenda was to find out the interpretation of current term "Big Data" from the very early use of it. Along with the precise description, benefits/shortcomings of this technology and its usage were also the topics of interest.

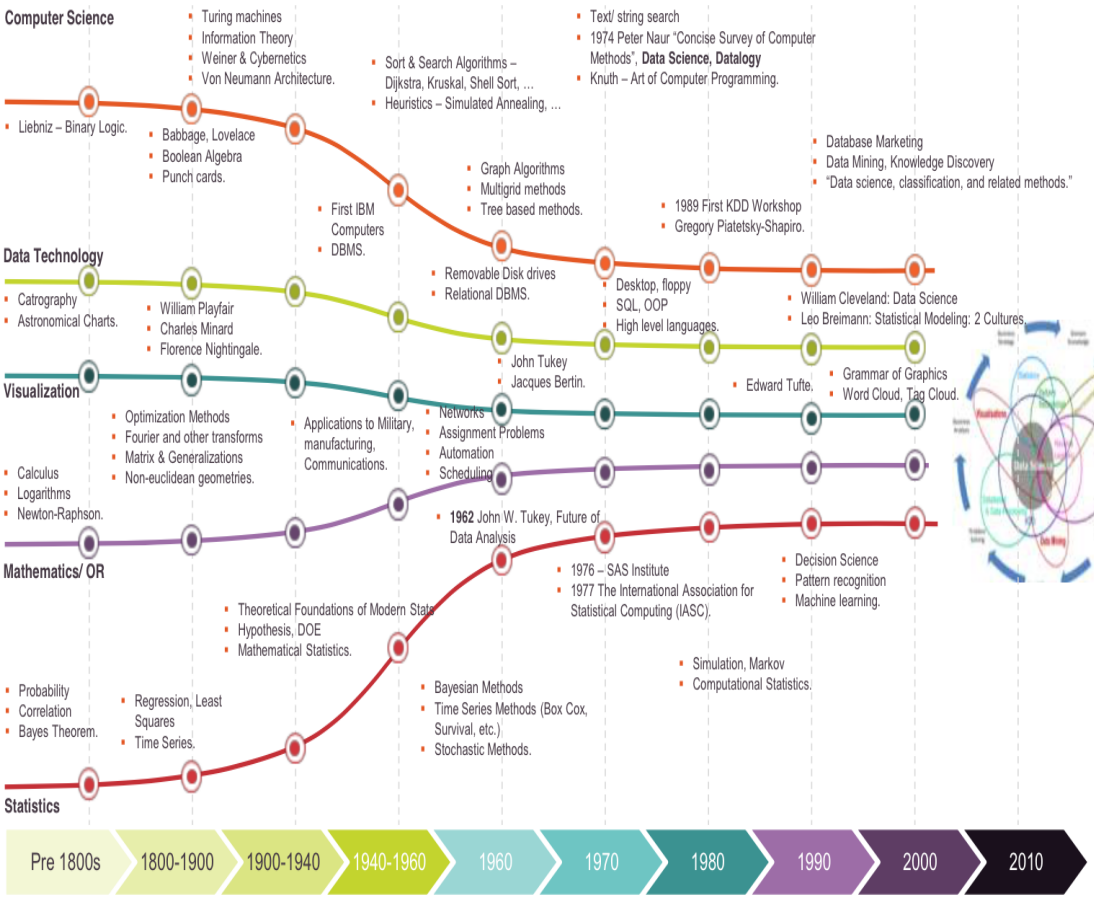

Figure 6. Evolution timeline of Big Data

1.2. *Conflicting reviews on Big Data implication*

1.2.1. In favor

Among the several debates made upon the implication issues of Big Data, many reviewers talked in favor and honored the Big Data Analytics with the title of "rock-stars" [20]. NRCNA declared it as "Frontiers in Massive Data Analysis" [21]. James Manyika, et al. deputed Big-Data as "the next frontiers in regard of real world data to innovate, compete and to be productive" [22]. The real world data is produces by either the man-made machines or by the man itself. Ultimately, the daily life activities are the source of the generation of such data. Rob Kitchin reported the Big Data as an epistemic and causing a fast shift in the entire science paradigm [23]. To somewhat extent, Viktor M. Schonberger and K. Cukier, [24] presented Big Data as a revolutionary to the pattern of our thoughts, our behaviors, working styles and entire living standards.

The more data collected in quantity causes betterment in the overall BD quality because quantity is directly proportional to the quality in the case of BDA for ML, MetaData and Parallelism. Various reviewers advocates Big Data being a new inspiration

for economic innovations. They were at the inference that the term Big Data can itself reveal its importance without intervention, so ultimately we should not interrupt the self-descriptor big data about its importance. Montjoye et al's and Yves-Alexandre de Montjoy et al. also raised their voice in the same direction [25] . They took an example of transaction via credit card and revealed that the probability of the re-identification of a person using just a few spatio-temporal data points is almost 90% using BDA.

They reached to the conclusion that "big amount of data about the behavior of man can be fruitful when processed with BDA to find applications in architecture, engineering, research and medical field etc.

1.2.2. In Against

Some reviewers viewed the entire concept of Big Data as a hypocritical subject. They argued it is overstated and exaggerated. According to Thomas C. Redman, data cannot speak for itself [12]. The size of a Dataset does not make any difference about it. It seems a false belief. L. Gomes described a situation where a large number of monkeys are typing, and out of the blue, one writes Shakespeare [13]. Interpretation made was if one monkey typed Shakespeare then it should not be conferred that monkeys are smart enough to be Shakespeare. Dobelli [14] also disfavor the Big Data implication from the psychological view-point. He appealed to think in better way not like this one. Drenik [15] referred Big Data as "Madness of Crowds" because his perception about Big Data was just a delusion and an overstated term fascinating the crowd. This title was elaborated by Charles Mackay [16] via his book titled "Extraordinary Popular Delusions and the Madness of Crowd". They connected the entire positive wave about the Big Data with psychology of crowd. They considered it as an emotion of a large count of people who were highly moved and fascinated by the term Big Data. Further, that was just a trend or fashion of these extraordinary people. This trend loses its passion as soon as something new and more attractive appears. The explanation was purely dealing with human thinking.

Mackay pointed that the entire crowd just put their eyes at Big Data for a while until their focus got shifted to something new and more fascinating. Drenik stated that "the real meaning was duped by the illusions and it was too short-spanned". CIO magazine published behind the scene story in 2013. The remarked quotes were really hard. [34].

Danah Boyd et al [17] joined "in doubt of Big Data" team regarding the volume. They served the point that the large sized data are not all time suitable data from the perceptions of society and out the ball in Boyd's [18] court. He made assertions in "The End of Theory" that the entire theoretical or methodological descriptions are genuine for only statistics of today and how big the data is just irrelevant in numerous situations of research problem where the small dataset is found to be best suited one. They made a suggestion that the quantity is not much significant over the quality. Mill [35] criticized over the logical point that there is no use of gathering abundant datasets without any methodology or theory at all [32]. This so called big data which is big only in its volume cannot be used without any theory.

Another criticism made by David Lazer et al. [19] is found to be very descriptive. They referred GFT predictions as the most suitable comparison study and highlighted two issues. Issues were hubris of Big Data and algorithm dynamics which were major source of mistakes[39].

The hubris issue was about some reviewers who believed that it can be the suitable replacement of KDD and mining processes. The algorithm dynamics was subjected to the changing approach of algorithms which surely influence the consumers of the commercial services. It was based on the improvements done by Google. It causes the data collection and processing would be carried out algorithms resulted out of desires. Lazer revealed the hidden weak points of BD. BDA causes traps for research in social sciences domain especially. All of these reviewers made the conclusion that still there are miles to go to the point where big data could supersede the existing theories and methodologies.

So, the different inferences, different points of view, vivid thinking patterns of reviewers because of wide variety of Big Data implementation and interpretations resulted into this apples and oranges kind of situation. After looking at both the perspectives, a clear notion is to define the term Big Data precisely, it is essential to rectify the conflicting matters among these remarks.

3. Meaning of 3Vs, 4Vs and 6Vs as expanding dimensions

3.1. *Method to Define Big Data*

The "big" of term "big data" cannot be accurately measured. The question, "big is how much big?" cannot be answered correctly. The data from past was big? Or the data of today is big? May be it will seem too short in front of next generation data size, then will the name of big data would be redefined? Ultimately, we are at the same stuck point that is "how to define the Big Data precisely?" To resolve the debate issues and to accurately give the definition Thomas [26] and Irvig [27] made a major contribution. They provided the methodology to definition depicted in Figure 7.

This approach for definition suggests to keenly analyze the historical development of the big data definition. The very first effort is to find out the lexical meaning of the term. Then next is to add the semantic meaning by expanding the dimension of V's. It is about supplementing the term with some additional attributes to comply the meaning of it. Then, the final step of bringing clarity and removal of skepticism in the defining of Big Data can be made successfully completed.

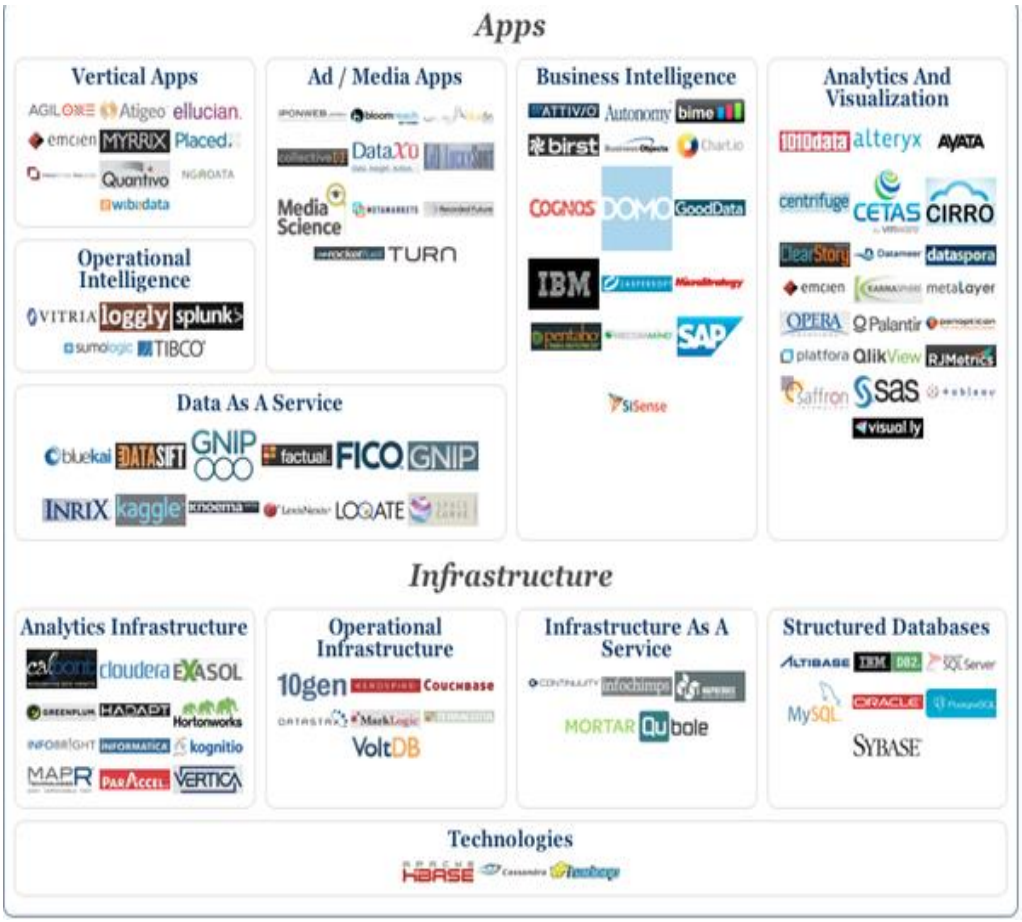

Figure 7. Methodology to define

3.2. The Definitions from History

3.2.1. Definition by Gartner of 3V's:

When we look at the history then, a variety of attributes defining Big Data were introduced. The renowned ones are 3V's shown in figure 8. The other name is Gartner's interpretation after his name. These attributes acquired a special place in all the citations and literature. The seed of 3V's was implanted in 2001 when Douglas Laney [28] published his white paper. Gartner expanded that in 2004. The electronic surge in commercial transactions and processes resulted in the data-growth along some specific dimensions, this was pointed by Douglas. Those were three dimensions calling Volume, Velocity and Variety. The time has witnessed the wide acceptance of 3V's as definition of Big Data. Volume refers to the in-stream of the data and the volume of data gathered. Velocity denotes the movement data which is caused by interactions. Variety means the range of formats resulting in non-compatible data and inconsistencies for data structures.

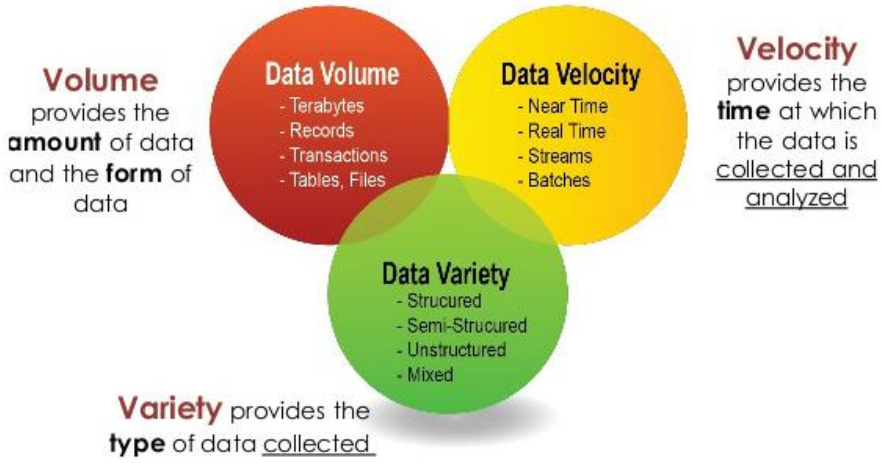

Figure 8. Definition by Gartner of 3V's

From the historic perspective, Gartner's definition of 3V's is widely accepted and these attributes are the common attributes of Big Data term.

3.2.2. Definition by IBM of 4V's

After that, the era of expansion started for big data term, when IBM introduced another attribute of "Veracity" as a cherry on the cake built by Douglas's 3Vs notion. This newly added 'V' caused the definition as 4V's [29] [30] shown in figure 9. Now, Volume refers to the scale of data. Velocity means the analysis of streamed data. Variety denotes variety of data formats. Veracity stands for the data uncertainty.

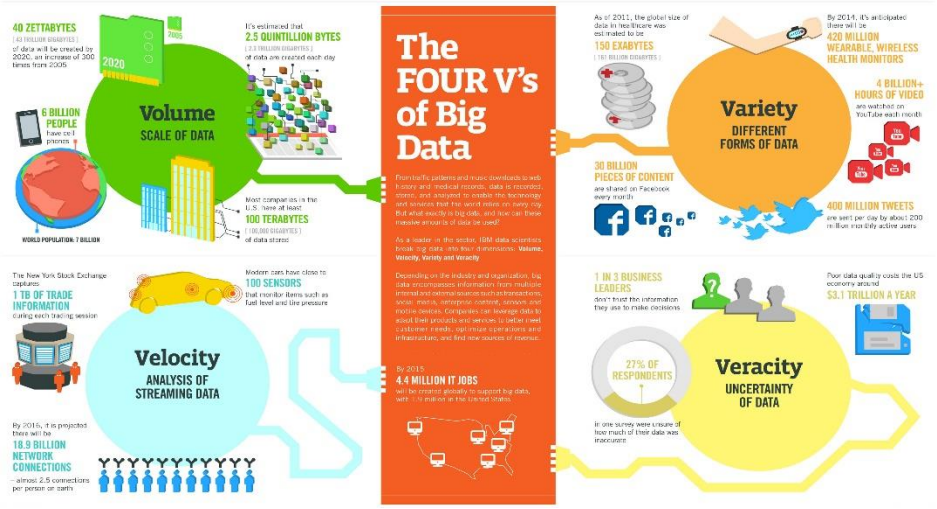

Figure 9. Definition by IBM of 4V's

The motivation behind this new attribute as a 4th 'V' in the definition was explained well by Paul C. Zikopoulos et al. [31] which was directly from the problems faced by clients about the quality and source of data when Big Data came into the implementation. They also considered all other definitions and attributes of Big Data proposed by market analysts.

3.2.3. Definition by Microsoft of 6V's

Now, Microsoft entered into the play. With the introduction of such a commercial name in the race of defining the Big Data term, it surged the maximum commercial value and importance. Microsoft introduced exactly 3 more V's and doubled the dimensions of Big Data. New 3 V's were Variability, Veracity and Visibility. Volume, Velocity and Variety still represents the same attributes as dictated by IBM. Veracity tells the trustworthiness of sources of data. Variability denotes the complexity in data by counting the number of variables in data sets. Visibility is about the knowledge of full frame of scene to get the exactly correct information from the Data show in figure 10.

Figure 10. Definition by Microsoft of 6V's

3.2.4. Expanding dimensions of Big Data from 3V's to more Vs

In 2013, another definition of Big Data was also released by Yuri Demchenko [33] named 5V's. He expanded IBM's definition by adding one more 'V' representing "Value" dimension to Big Data depicted in Figure 3. Douglas Laney introduced 3V's in 2001, since then various additional attributes with initial letter published in the literature. The Eleven V's can be added to complete the definition of Big Data [41] shown in figure 11.

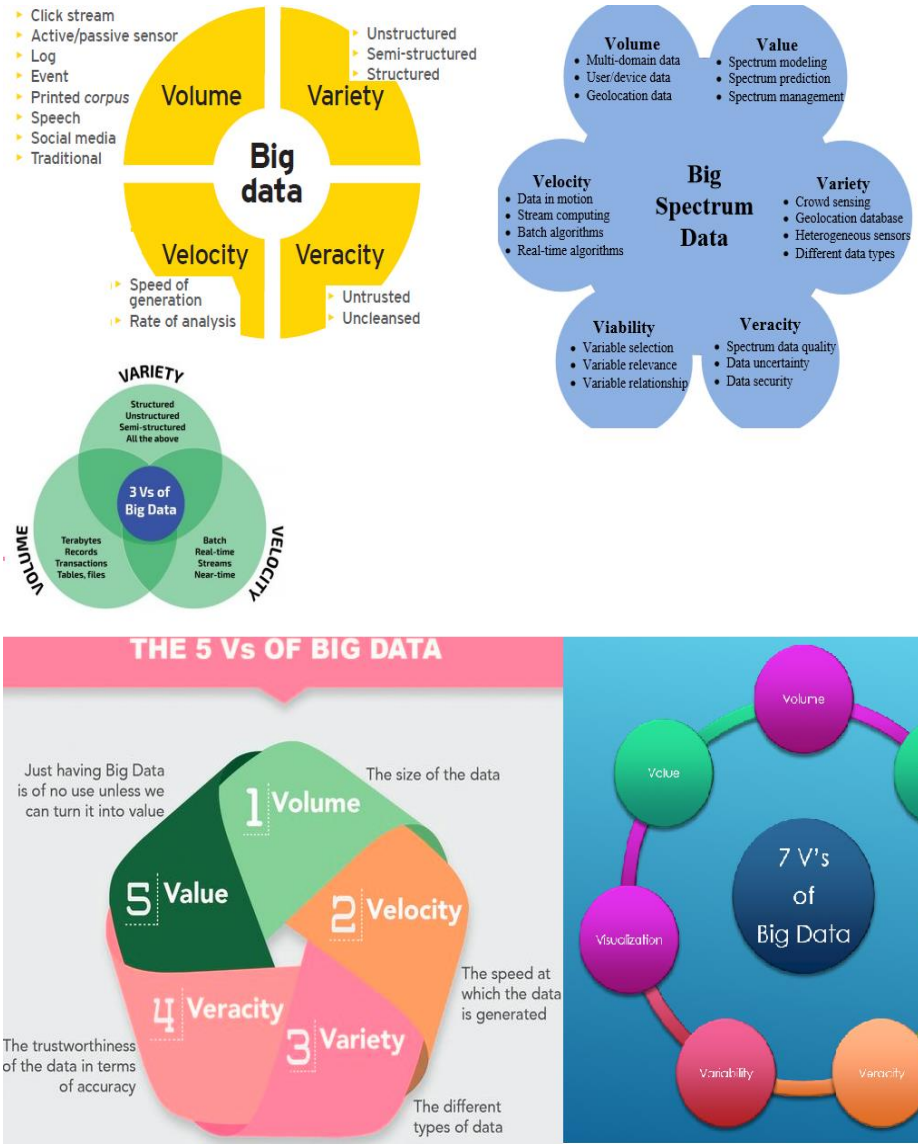

Figure 11. Expanding dimensions of Big Data from 3V's to more Vs

After having a glimpse over the variety of definitions, we can detect a common theme behind all these definitions. That theme is about to explore the dimensions of Big Data so that it can explain the necessity of Big Data Analytics. But, none of these clarify the BD essence. The core meaning of BDA can be stated only after the clear understanding of the meaning of the term "data".

What is Data? Data is every raw fact, figure or value. Everything is data actually. Every activity generates data. Entire life, world and living beings are data. The amount of data in our surroundings is non-measurable. In-fact, it is infinite. In case, there is no restriction of technology and storage, then no end of the data exists. The meaning of data is self-explanatory behind the question "why to gather the data?". The motive to capture the

data does not lie in neither of the proposed definitions. It lies in the usage of data to provide the intelligent knowledge to solve multiple research problems and also to find business solutions. We capture big amount of data to extract knowledge from it to perform intelligent decisions. So, Big Data Analytics cannot be purely data driven. Harper calls Big Data, Hard because a large volume of data cannot be analyzed only from data driven approach [34].

3.2. *Definitions of Big Data*

Summary of Seven Definitions of Big Data was published by Timo Elliott [41] as shown in Table 2. He took into consideration more than 33 Big Data definitions proposed by various researchers [42].

Table 2. Summary of Definitions

S .No.	Definition Term	Definition Description
1	3V's of Big Data	Douglas Laney defined 3 V's of Big Data- Volume, Velocity and Variety or 3Vs which wer widely accepted from 2001. Various researchers expanded the dimensions from 3V's to 4V's, 5V's, 6V's, at max 11V's.
2	Big Data – A Technology	Map-Reduce, Bulk Synchronous Parallel - Hama, Resilient Distributed Datasets - Spark, and Flink brought a new meaning to Big Data and interpreted it as a Technology.
3	Big Data-An Application	Big Data solved heterogeneous kinds of application based problems. Then, Barry Devlin [43] proposed a definition for Big Data as an application for processing the all kinds of data generated by man, machine or any activity. Shaun Connolly [44] looked for hindsight of data focusing the analysis of transactions and interaction.
4	Big Data- A Signal	It is also application based definition of Big Data with the only difference that its core emphasis was on timing instead of data-type. It is foresight of data.
5	Big Data-An Opportunity	Matt Aslett [45] viewed Big Data as an opportunity for analyzing the data which was ignored at first. After reassessing the gathered data, new potential in this novel opportunity was detected.
6	Big Data- A Metaphor	It refers Big Data as the process of human thinking [46]. It took the concept to the new level by extending it to the human brain.
7	Big Data – Old Stuff, New Name	It was just a new definition for the old concepts. Like, new packaging for the already existing data mining techniques.

A variety of aspects have been addressed by all these definitions. But, individually each one is found to be limited to only one approach for BD. Apparently, a big definition for Big Data is desirable to comprehend all the aspects notified by these individual definitions. But that would become very perplexing one. Karl Popper introduced a rational idea for the restructure of definition. His notion was based on analysis of reasons behind the entire paradigm and processes. That analysis must be done explicitly and in a way that is easy to understand.

3.3. *Driving Force for the Definitions*

The overall objective of Big Data Analytics is to gain 3D view of the data. These 3-dimensions are depth, foresight and hindsight of the data. Depth is about the very detailed and deep knowingness for the data. Foresight is to predict the possibility of data for the future. Hindsight refers to the analysis of patterns from data of history. No definition is found to be comprehensive enough to cover all these aspects alone. Hence, to address all the possible issues and comprehend the data aspects, redefinition is required. Now, to give a complete, unambiguous, precise and comprehensive definition for Big Data, we shall consider all data attributes from all possible aspects.

4. Definition of Big Data

The definition of Big Data from today's implication relies on three aspects which were not addressed altogether by any of the previous definitions. The sole purpose of capturing a huge amount of data and then analyzing it, is to obtain answers to the business and research problems. The data must be analyzed in such a manner that it serves as a source of knowledge which can further empower the intelligent decision making capability. To satiate all the desires behind collecting "Big" data amount, we essentially have to discern three more aspects and the inter-relationships among these attributes. New aspects are from Data Domain, Business Intelligent Domain and Statistical Domain. Now, let's have a look on these three domains.

4.1. Data Domain-Searching for patterns

Data Domain is vibrant with 3V's namely Volume, Velocity and Variety, introduced by Laney, who successfully captured the most important characteristic of Big Data. That attribute was "Volume". The entire evolution timeline of Big Data is flooded with the hype of the growth in volume of data. Other 2V's are not enough dominating as the "volume". So, the amount of data is significant, but it is not the only one factor for BD. Data Domain is about the volume of data to find out the patterns existing among the captured data.

4.1.1. Business Intelligent Domain-Making predictions

Big Data in BI (Business Intelligent) Domain is based on three another V's which are Visibility, Verdict and Visibility. This is the most crucial domain for Big Data. It is the primary agenda to be accomplished by the BDA. The capturing and analysis of a very large volume of data generated by human or machines are done to obtain Business intelligence to support decision-making. There is no point of BDA without BI domain. The major component of BDA is Business Intelligent Domain. To make intelligent decisions using the knowledge obtained from BDA these three V's are essential. Visibility covers 3-D view of Data including insight, foresight and hindsight of data. It involves the vision to find the solution of BI problem. It extends the insight not only for the data, but also for the metadata (information about the data). Verdict is the most sensitive "V" as the entire BI solution

is about making the best decisions and most optimum choices at an instant. The decision is to be made on the basis of the definition of the problem, the scope of the problem and the set of possible choices for the problem. Along this, the restricted available resources make it more challenging part. To capture, store, extract, transform and load the data for verdictive decisions involves a lot of "what-if" clauses. This adds extra cost to the entire process heavily. Value is the measurement of how valuable the data is for the BI. It has nothing to do with the value of data-set but it is about the information it carries with it. It implies the usefulness of the data in long run. Figure 12. Depicts these three V's of BI Domain of BDA.

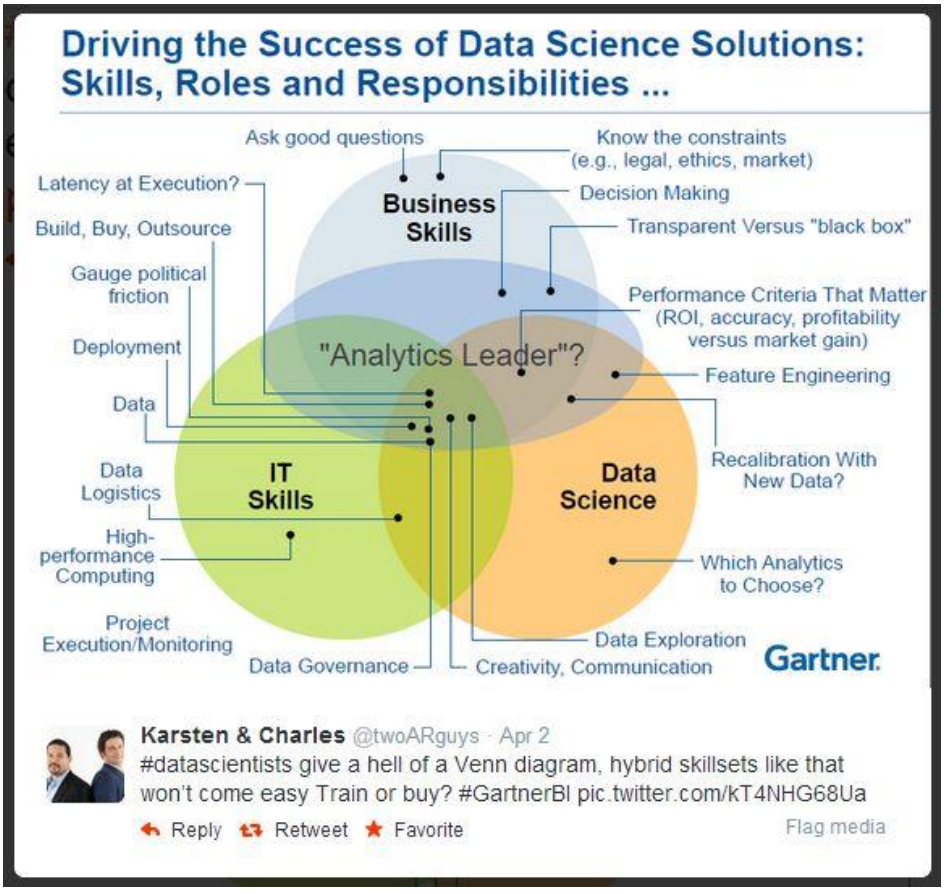

Figure 12. Domains of Big Data Analytics

BI domain revolutionized by the arrival of trendy platforms and frameworks in the market for BDA. It covers Map-Reduce, Hadoop and others. The steering direction set the BDA to answer the core questions of BI. These five basic questions are upshot of the Business Intelligent Domain. Questions are about the storage, accession and processing of data to find best decision-maker's knowledge. First, how to store the massive amount of data on the restricted resources. Second, how to access captured large volume of data quickly and efficiently. Third, how to transform the heterogeneous data-formats to a common data format. Fourth, how to process the data with the high scalability, flexibility and fault-tolerance. Fifth, how to extract the knowledge hidden in such massive data to make intelligent decisions. The extraction must be cost-effective and interactive.

3V's of BI are closely interlinked with one another. "Verdict" is the most complicated and crucial "V". "Visibility" is to get the deep insight of data in real-time. "Value" is considerable when we have to capture the data. "Value" can be judged only after getting insight of data in all three dimensions which is discerned by "Visibility". "Verdict" is possible only after capturing the relevant and worthy data and dependent to the "Value" which in turn depends upon "Visibility". So, the relationship among these three V's is Visibility → Value → Verdict. Without "Visibility", there is no meaning of other 2V's.

4.1.2. Statistical Domain-Making assumptions

Statistical Domain is defined with three more additional V's namely Validity, Veracity and Variability of Data. It suggests how to make assumptions in "what-if" form to precisely judge the degree of reliability on the gathered datasets. These assumptions build a set of hypothesis. The entire set is required to be fully accurate, precise and correct. Because if the assumptions made are wrong or the data is unreliable or the statistics went wrong then the results of Big Data Analytics would be incorrect. The dirty data sets may cause failure of the entire business solution. Such a failure was well reported by Literary Digest Magazine [36] in 1936. A prediction was made for the US elections for the president. The data was a survey report from 2.4 million sources of response. That data was later found to be dirty which caused a big disaster in the prediction of winning candidate for the company conducting the polling. Statistical domain deals with the accuracy, reliability and validity of datasets via its 3V's. Validity denotes the verification test for the soundness of the data-quality logically. It emphasizes validity of the whole inference system being based on the statistics. It explains precisely the acquisition of data in a correct and fair fashion. Veracity is the degree of clarity in the data. Uncertainty and vagueness can be removed from data by this aspect of statistical domain. It is responsible for the high degree of faith and reliability in data sets. Variability implies the variations in the data complexity. Bruce Ratner [37] brought a new belief about this statistical domain factor that a data can be nominated as BD only if it has about 50 or more different variables. Then, such a complex dataset is processed via BDA on statistical grounds that the needs of business solutions can be satisfied.

"Veracity" is the key factor in this domain. It brings the Big Data Analytics closer to the reality while building statistical model. A fitting to a curve and the implication of Veracity is analogous. Lesser the constraints, the larger errors in the regression. The more constraints, the more problems of over-fitting.

4.1.3. Definition of Big Data-9V's & its Venn Diagrams

The three major domains-Data Domain, BI Domain and Statistical Domain are explored fully now. Each domain is defined with 3V's. In total, we have 3 X 3 V's that counts to 9V's of Big Data. These 9V's are the dimensions of Big Data definition. The entire set of V's covers all essential perspectives. This set of 9V's is leverage to the definition of Big Data comprehending all the possible aspects for the precise and complete definition. The Venn diagrams of 9V's reflect our definition of Big Data in Figure 13.

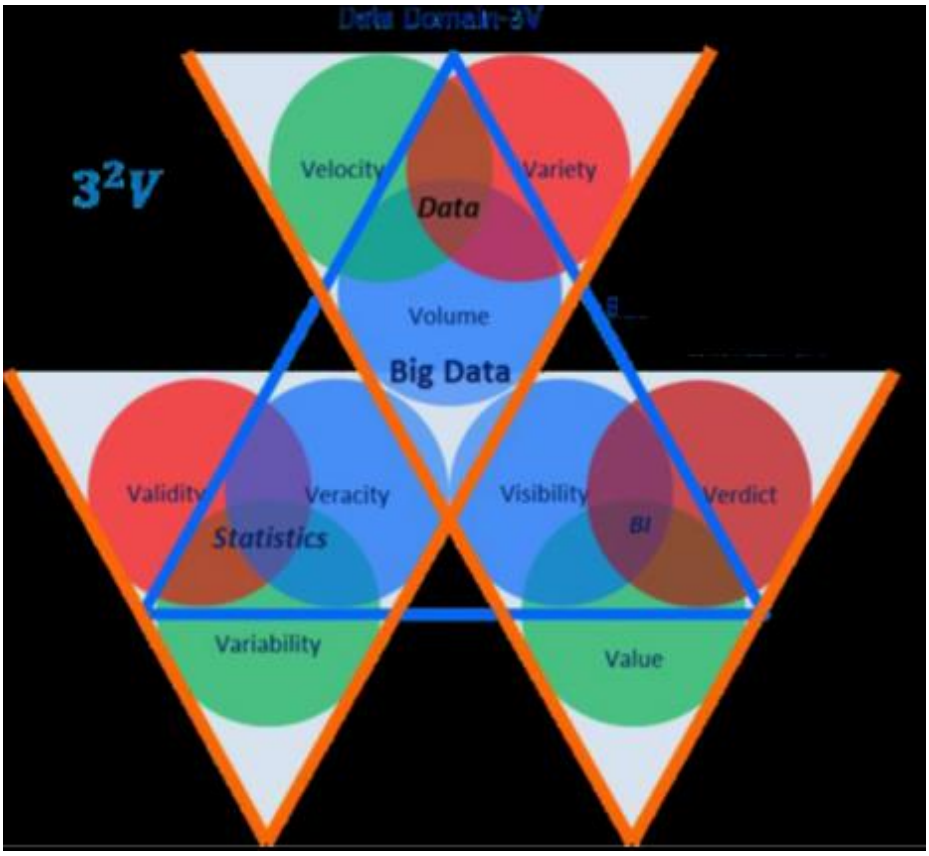

Figure 13. Venn Diagrams for 9V's

It can be clearly seen that each of the 3 domains is represented with a V-shape. Each "V' has 3 attribute within it. All 3 X 3 = 9 or 32 aspects are constituents of Big Data definition. A triangle shape can be drawn as shown in figure to cover all these 9V's with the area covered representing the degree of participation of each of 32 or 9 aspects.

This triangle covers Veracity, Visibility and Volume completely under its area. Rest of the V's are partially (exactly half) are shaded under this triangle. This triangle is the comprehensive definition of Big Data. It reveals the semantic meaning of Big Data term. From historical definitions, Douglas Laney served the Big Data term with 3V's which was the lexical meaning of it. But, this definition explains it's semantic and syntactic both the meanings. It covers three major domains with 9V's to explain all possible aspects of Bog Data for today's realm. So, here is definition for Big Data with 9V's out of which the more concentrated aspects are Veracity, Visibility and Volume. The "volume" is all time justified Aspect. "Visibility" refers to the insight of the data, hence its importance is basic to all other aspects. "Veracity" is justified with the essence of "high certainty of data" to achieve best decision-maker's knowledge and to make the mission of Big Data possible. So, these 3 are heart of the definition and their learning via ML is heartbeat to the BDA.

5. Big Data in Machine Learning

5.1. Big Data Analytics

The term "Big Data" is now defined at both Syntactic and semantic levels with the advent of 9V's. Now, The term "Big Data Analytics" adds the practical meaning to the BD. It supplies the pragmatic meaning to Big Data. 9V's Venn Diagrams for the definition of Big Data can be extended to the Venn Diagrams of the pragmatic meaning of Big Data, by applying computational point of view.

The pragmatic meaning of Big Data comprised of 3 technological paradigms including Machine Learning. Samuel [38] refers the very first definition of ML about being the field which deals with making the machines capable to learn from the data. This learning is self-learning by finding patterns among the data being fed. NO human intervention is expected for this learning. From the historic perspective, ML has always been the point of attraction for professionals along with the "Big Data". To precisely define the meaning of ML, numerous terms were introduced like data sciences, pattern recognition, data mining, business intelligence etc. There are more than 32 key terms and descriptions of different aspects of ML. All these different orientations represent some certain aspect about ML. The major aspects are Data, Information, Knowledge and Intelligence. Now the classification of 32 terms defined for ML can be made from the 4 aspects as in Table 3.

Table 3. Terms defining ML

Data	Information	Knowledge	Intelligence
Data Mining	Information Analytics	Real time Analytics	Business analysis
Data Science	Information visualization	Predictive analytics	Business Intelligence
Data Warehouse	Information System Management	Machine Learning	Artificial Intelligence
Learning from Data	Text Analytics	Knowledge Base System	Decision Support System
Data Smart	Text Mining	Pattern Recognition	Actionable Intelligence
Data Analytics	Web Analytics	Statistical Application	Business Forecasting
Making Sense of Data	Web Semantic Analysis	Knowledge Discovery	Business Strategy
Data Ingestion	Web Searching	Expert Systems	Business Transformation

5.2. Machine Learning (ML)

ML is about to automate the process of learning for machine by finding patterns. To surpass the human ability of decision-making or problem solving from easy to complicated levels is sole objective of ML. It has come up into the existence directly from Artificial Intelligence, which was focused to imitate humans by making robots. Then, over the years it got extension from

robots to solving the problems using machines. The problem-set has been generalized to be solved by the machines. In past, Intelligence was achieved by supplying a rule base, over which the machine could be programmed to solve problems. Ultimately, algorithms were fed to transform inputs into the outputs. It was called GOFAI-Good Old Fashion for AI. Another name given was "Expert-System".

Some problems cannot be solved just by supplying a flat algorithm like "pattern Recognition" with handwriting as input. The transformation of handwritten content to the fine typed letter is not easy. The alternate led the baseline for ML that is to learn from the data. The learning is based on hit and trial mechanism and also "The wisdom of crowds" plays important role [40]. It is about making a trial, then to match the outcome with the desired output, later the error are fed back to adjust the direction of trial. It is hit and trial complemented with wisdom. This powerful aggregation leads to learning from data by machines as illustrated in Figure 14.

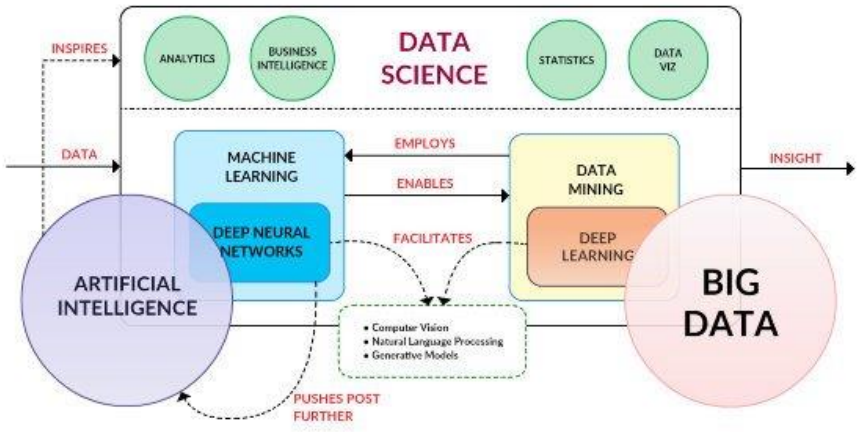

Figure 14. Machine Learning

History supplies multiple definitions of ML from the contributions from different authors. They pointed different characteristics out of ML [47]. Some addressed it as an application, other associated it with being a utility or process. Arthur Samuel's perspective of ML is that the automated learning is ML. Each component of ML was described by Tom [50] in detail. Murphy [51] and Bishop [52] focused on the pattern recognition process as ML. Nisan and Schocken [53] stressed the ability of ML to transform abstracted thoughts into physical actions. More than 30 definitions are available in the history of ML, the summary of all these can be centrifuged to find major components of ML [48-49]. These traits are from Training the machines to automatically learn from the data through the Extraction of meaningful patterns and knowledge hidden in the datasets and to make predictions for the unknown inputs from the previously supplied data, ultimately to serve the purpose of problem solving on the basis of intelligence they learnt from the data automatically. All these specified elements can be combined to give a comprehensive definition of ML. Ml is the field grown as a result of the overlapping of popular fields of computers and statistics with a common target to train the machines to learn automatically for pattern-recognition and intelligent decision-making [54]. Peter says in simple words that the transformation of data into information is all about ML [55]. In nut-shell, ML is to provide human competent machines to make decisions and to solve complicated problems.

ML supports the implementation techniques for BDA. Without ML, the continuously growing enormous amounts of data cannot be processed. ML is the brain of BDA, and rest of the BDA components are just sub-ordinates to the ML paradigm. S. Wadkar et al [56] defined 4 architectures from computational aspect to extract the knowledge from the massive data effectively. These are:

- Massively parallel processing (MPP) database system.
    - For example, EMC's Greenplum and IBM's Netezza
- In-memory database systems.
    - For example, Oracle Exalytics, SAP's HANA and Spark
- Map-Reduce processing model and platforms
    - For example, Hadoop and Google File System (GFS)
- Bulk Synchronous Parallel (BSP) systems
    - For example, Apache HAMA and Giraph

After these four models, a new model has been launched to perform BDA in the most cost effective way that is Cloud Computing (CC). It has become a recommended and popular solution especially for small and media businesses (SMEs) [74].

5.3. Big Data Analytics and Cloud Computing

Another important part of BDA is Cloud Computing. It provides computing infrastructure, data, and application services on subscription [33]. The BDA tends to provide computing ability for the data from/to the Internet including the indexing and searching of web pages. This dataset is of massive volume, so BDA looks for an implementation which cost effective and provides high degree of fault-tolerance. CC is implemented under 3S-4D-5C definition. It is comprised of 3 service models, 4 deployment models and 5 Characteristics [23].

1. 3S- 3 service models namely SaaS, PaaS and IaaS
2. 4D- 4 deployment models are Private, Public, Community and Hybrid Cloud.
3. 5C- 5 Characteristics namely on-Demand, Broad network Access, Resource Pool, Rapid elasticity and measured service

The cloud possess the capabilities for being an easy and fast accessible infrastructure. It is suitable for small companies and medium scale companies for BDA implementation [73].

CC supports the scaling of the large data down effectively. It seems just opposite to BD term which all about the scaling of data out. That's not the case for all situations. Sometimes data size is moderate, in-face fluctuates to ultimate increase of the overall data volume. The needs of BDA processing unit vary as per situ. We utilize the elastic property of Cloud to implement BDA in a cost effective manner. It is most suitable for batch, micro-batch, interactive, real time, and near real time BDA processing.

High flexibility for computation is provided by CC infrastructure as per the requirement of BDA. Amazon sets an example for this statement by providing spots instances at the regular rate, which can be extended to increase capacity to complete the processing in shorter interval if batch mode is desired [69].

Hadoop is one the most popular platform that is open and implemented on cloud infrastructure. Its overall implementation is inspired by Google Map-Reduce and Google File System (GFS).

6. Hadoop, HDFS, Map-Reduce, Spark and Flink

Hadoop is reopresented in Figure 14 which is the most popular platform for BDA. It has become the priority for analysts and professionals for BDA implementation to make intelligent decisions. Michael Cafarella stated that

Nutch which is predecessor to Hadoop is The National Public Radio (NPR) of search engines [63]. The reasons for the development of Hadoop are numerous. First, it is programmed in Java and provides open source platform. Second, It is linearly scalable, reliable and accepts hardware failure. Third, it provides a fault tolerant system. Fourth, it is a practical platform to store and process greater than 10s of TB data. Fifth, it is best fit for diversified data sources. Next, it leverages commodity type of hardware. And, it is "schema on read" or has "data agility" character shown in figure 15.

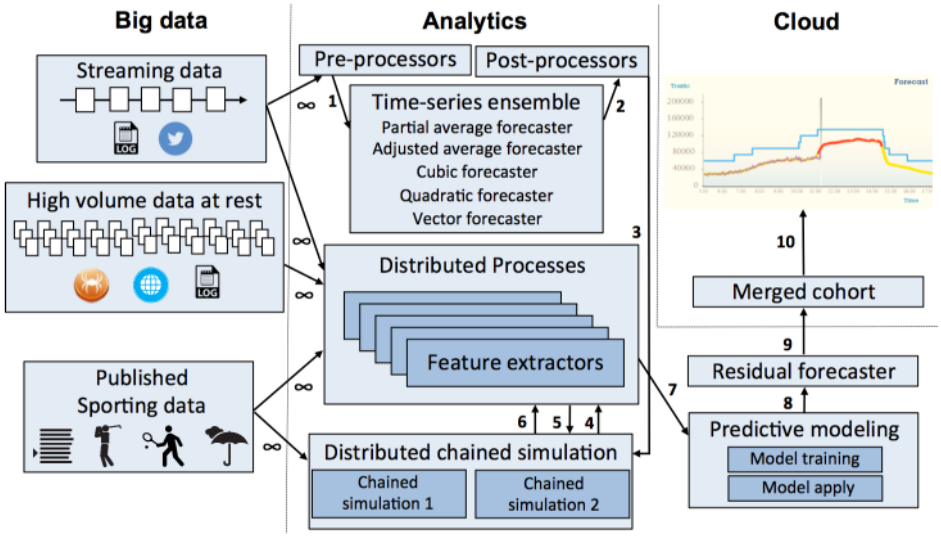

Figure 15. Overview of Hadoop Framework or Technology Stack and Ecosystem

The driving force behind Hadoop is the continuously increasing data and the hardware cost for computations. The idea of Hadoop platform is to leverage the commodity hardware for processing the workload at a large scale. It is achieved via only mainframe computers which are very expensive. Hadoop allows to scale-out the computational capability rather the scale-up. Often, both these terms are interchangeably [57] used but the standard definition qualifies "scale-up" as the sense of quality improvement while "scale-out" as adding the same unit horizontally.

The wide acceptance of Hadoop platform is caused by the attractive features it provides. Hadoop is a computational platform available openly free of any cost. It allows the use of commodity hardware for the storage and

processing of large amounts of data via clusters. The basic working principle of Hadoop architecture relies on the statement. Its architecture consists of three major components displayed in Figure 16. Components are:
1. HDFS - Hadoop Distributed File System
2. Map for distribute function
3. Reduce for parallel processing function.

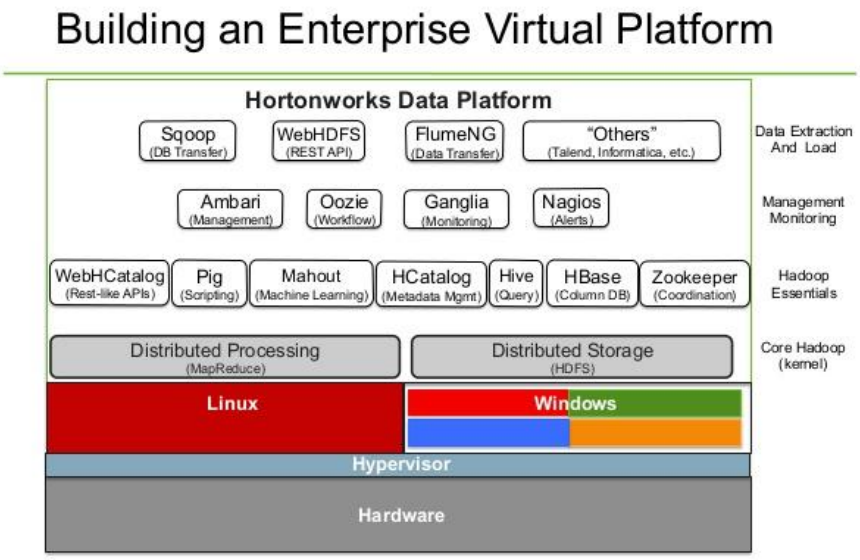

Figure 16. Hadoop Kernel

Every coin has two faces, so the Hadoop has. This generic platform has one drawback is that it can process the dataset only in batch-mode. It is because of the basic design principle of Hadoop. It is originally designed to process to run queries and perform data read operations on very high volume data [58]. Early released versions of Hadoop were incapable to stream the data and does not support interactions during processing. Core Attributes and corresponding features of Hadoop are summarized in Table 5.

Table 5. Core Attributes of Hadoop

Attributes	Characteristics of Hadoop
Initiators	Doug Cutting and Michael J Cafarella
Predecessor	Nutch
Subsequent Version	YARN or Hadoop 2.0
Hadoop Written Language	Java
Philosophy of computation	Divide and Conquer for large datasets
Principle of Computational Processing	Bring computer to data rather than bring data to computer
System	A distributed programming framework
Main Characteristics	Accessible, Robust, Scalable, Simple and Fault tolerance
Storage -Hadoop Distributed File System (HDFS)	Self-healing Distributed and shared storage element
Initial Computational Program - Map-Reduce	Distributed, aggregated and collaborated parallel processing
Map-Reduce Library written language	C++ code
Process Type	Batch
Hardware Type	Heterogeneous commodity hardware
Software licence	Open Source
Initial Applications	Information Retrieval (IR) and searching index and Web Crawler
Solution Type	Software solution not hardware solution
Scalability Solution	Scale-out not Scale-up
Typical Size of Data Set	From few GBs to few TBs
Capable Size of Data Set	From Tens of TBs to Few PBs
Simple Coherency Model	Write-once and Read many
Default Replication Factor	3
A typical size of data block for HDFS	64MB
Permission Model	Relaxing POSIX [2] model
Main Application Modules	Mahout, Hive, Pig, HBase, Sqoop, Flume, Chukwa, Pentaho …
Typical Vendors	MapR, Cloudera, Hortonworks, IBM, Teradata, Intel, AWS, Pivotal Software and Microsoft

In 2002, ASF - Apache Software Foundation conceived the notion of Nutch project which later succeeded as Hadoop captured in Figure 17. Initially, the platform was open source. It was based the Map-Reduce Model [60] and DFS. This implementation was proposed by Google originally. Google allowed

Apache to incorporate its Map-Reduce model in their Hadoop. This license was open source, free of course and its distribution was without any patents.

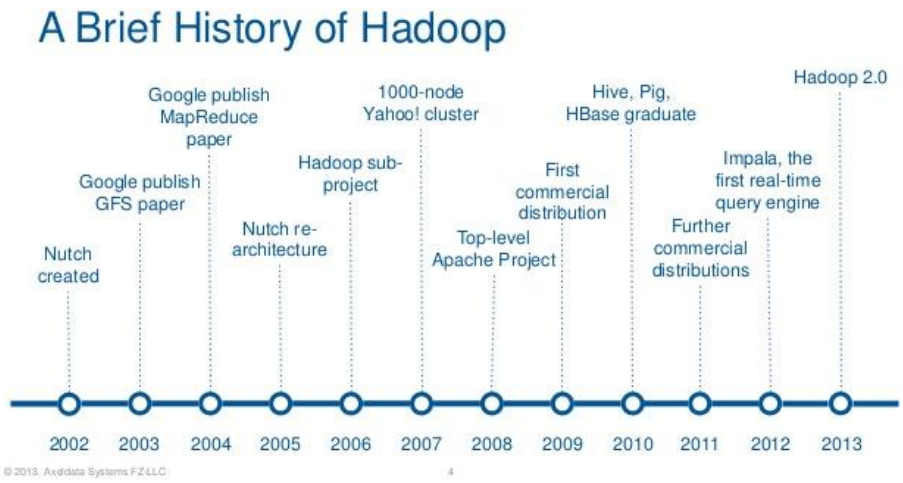

Figure 17. Brief history of Hadoop

6.1. File System of Google and Hadoop

The author set Google File System (GFS) architecture and developed Hadoop Distributed File System (HDFS) in the Hadoop project also it based on four important characteristics for GFS:

- System Principles
- System architecture
- System assumptions and
- System Interfaces

In the GFS system the conventional system design creed that a failure was not allowed and the computation system used must be reliable. In contrast, GFS expects the certain number of system failures with specified redundancy or replicating factor and automatic recovery [62]. GFS is capable of handling billions objects and I/O in comparison with the traditional file standard, also it should be revisited. Moreover, most of files will be altered by affixing rather than overwriting. Finally, the GFS flexibility is increased by balancing the benefits between GFS applications and file system API [61].

The master server maintains 6 types of the GFS's metadata, which are:

1) Namespace,

2) Access control information,

3) Mapping from files to chunks (data),

4) Current locations of chunks or data,

5) System activities: (chunk lease management, garbage collection of orphaned chunks and chunk migration between chunk servers),

6) Master communication of each chunk server in heart beat messages.

GFS was designed with five basic conventions according to its particular application requirements:

1. GFS will anticipate any commodity hardware outages caused by both software and hardware faults. This means that an individual node may be unreliable. This assumption is similar to one of its system design principles.

2. GFS accepts a modest number of large files. The quantities of "modest" is few million files. A typical file size is 100 MB/per file. The system also accepts smaller file but will not optimize them.

3. Typical workload size for stream reading would be from hundred KBs to 1MB with small random reads for few KBs in batch mode

4. GFS has its will defined sematic for multi-clients with minimal synchronization overhead

5. A constant high file storage network bandwidth is more important than low latency.

In contrast to other file systems, GFS does not adopt a standard API POSIX permission model such as Andrew File System (AFS) or Server less File System (xFS) or Swift, rather than relax its rules to support the usual operations to create, delete, open, close and write. According to the assignment processing assumptions, GFS is a file storage system and having two basic structure: Logs and String Table (SSTable).

According to these workload processing assumptions, GFS is actually a file storage system or framework that has two basic data structure: logs (metadata) and Sort String Table (SSTable). The main object of having GFS is to implement Google's data-intensive applications. Initially, it was designed to handle the issues of web crawler and file indexing system under the pressure of accelerating data growing. The aim that Google published these influential papers [59] was to show how they scale out the file storage system for large distributed data-intensive applications. Doug Cutting and Mike Cafarella leveraged the Google's GFS idea to develop their file system – Nutch or Nutch Distribute File System (NDFS) for web crawling application, namely Apache Lucene. NDFS was the predecessor of HDFS (see Figures 13 and 15). Although HDFS is based on GFS concept and has many similar properties and assumptions as GFS, it is different with GFS in many ways, especially in term of scalability, data mutability, communication protocol, replication strategy, and security.

6.2. Map-Reduce

Map-Reduce is a core component of the Apache Hadoop software framework [75]. Hadoop enables resilient, distributed processing of massive unstructured data sets across commodity computer clusters, in which each node of the cluster includes its own storage. Map-Reduce serves two essential functions: It parcels out work to various nodes within the cluster or map, and it organizes and reduces the results from each node into a cohesive answer to a query. Map-Reduce is composed of several components, including:

- JobTracker -- the master node that manages all jobs and resources in a cluster.
- TaskTrackers -- agents deployed to each machine in the cluster to run the map and reduce tasks.
- JobHistoryServer -- a component that tracks completed jobs, and is typically deployed as a separate function or with JobTracker.

To distribute input data and collate results, Map-Reduce operates in parallel across massive cluster sizes. Because cluster size doesn't affect a processing job's final results, jobs can be split across almost any number of servers. Therefore, Map-Reduce and the overall Hadoop framework simplify software development. Map-Reduce is available in several languages, including C, C++, Java, Ruby, Perl and Python. Programmers can use

Map-Reduce libraries to create tasks without dealing with communication or coordination between nodes.

Map-Reduce is also fault-tolerant, with each node periodically reporting its status to a master node. If a node doesn't respond as expected, the master node re-assigns that piece of the job to other available nodes in the cluster. This creates resiliency and makes it practical for Map-Reduce to run on inexpensive commodity servers.

From a programming viewpoint, Map-Reduce has other two meanings that "Mapping" is splitting for distribution and "Reducing" is shuffling + sorting in parallel. A major advantage is its capability of shared-nothing data processing, which means all mappers can process its data independently. The characteristic of shared-nothing enable Map-Reduce to run a simple program cross thousands or even millions of unreliable and homogeneous machines in parallel and complete a task in very short time. Theoretically speaking, it allows any programmer to access almost unlimited commodity type of computing resources instantly (theoretically) or within an acceptable time frame (practically) e.g. cloud infrastructure. Several Cloud computing platforms have implemented their own Map-Reduce processing model such as CouchDB, Cloud Map-Reduce and Aneka

Map-Reduce in action: For example, users can list and count the number of times every word appears in a novel as a single server application, but that is time consuming. By contrast, users can split the task among 26 people, so each takes a page, writes a word on a separate sheet of paper and takes a new page when they're finished. This is the map aspect of Map-Reduce. And if a person leaves, another person takes his place. This exemplifies Map-Reduce's fault-tolerant element [76-81].

Spark was developed by UC Berkeley RAD Lab (now called as AMP Lab). The main funder is Matei Zaharia et al. To adopt Resilient Distributed Datasets (RDDs) in memory computation (micro-batch) technique it extended Hadoop to a general purpose framework. In a simple term, it aims to replace MapReduce model with a better solution. It highlights the interactive quries and computational efficiency of iterative and recursive algorithms of data mining. It assumed that for certain type of assignment such as performing iterative algorithm it would be 10-20X faster than MapReduce. Although it tries to replace MapReduce, it did not unrestraint HDFS. It influences Hadoop's file storage system. It is an open source project under Apache Software Foundation (ASF) like many other Hadoop related projects. Based on large cluster Generally, Spark is a fast and general- purpose computation platform. In contrast to MapReduce Spark includes SQL, interactive query, data stream, graph, and machine learning analytic functions into its computation platform that is basically designed for web crawler, indexing system and limited machine learning.

6.3 Flink and other data process engines

There are several data processing engines such as Microsoft Dryad, Storm, Tez, Flink and CIEL apart from Spark that are proficient of supporting MapReduce like processing requirements. The objective to sustenance more computational functions, like standard queries, stream analysis, machine learning, graphic analysis and interactive or ad hoc queries efficiently. The effort made by these platforms is to generalize Hadoop to be able to support a wide variety of BDA workloads. Stephan Ewen et. al. [70], Kostas Tzoumas [71] and Marton Balassi [72] claimed that Flink is the next generation or the 4 th generation data processing engine in contrast with others, although every data processing engine has its individual distinct article. Flink data engine is really general

purpose framework for Big Data Analytics (BDA). They privilege that Flink is accomplished of outperforming Spark by 2.5 times [84].

A probable reason for Ewen to privilege that Flink is enhanced than Spark is that it is centered on Lambda architecture and able to process arbitrary Big Data workloads in real time. To build a data processing engine or system to deal with a subsection of data the Lambda architecture is used.it included the number of layers with properties these layers are line of code to implement the total seven steps.

The idea for launching these three layers, according to Nathan Marz, is to meet the characteristic requirements of all types of Big Data workloads. They are:

- Robustness and fault tolerance
- Low latency reads and updates
- Scalability
- Generalization
- Extensibility
- Ad hoc queries
- Minimal maintenance
- Debuggability

Hadoop and Elephant DB both can handle batch and serving layers and scalability is the main requirement for that. In MapReduce and Pregel, the users must follow the pre-defined programming model (e.g., map-reduce model and vertex-centric model), whereas in xtensibility, the users can design their customized programming model [67]. The batch layer permits users to figure out another view of batch. It can also be surmised in that the batch layer can also perform ad hoc queries as the master data set are in one location. Nominal maintenance is allowed as Hadoop is robust and a serving layer database only gets a batch view per few hours for data process both input as a batch layer and output as a serving layer are recorded for intermediate steps. Therefore, if the process has any glitch, the debug analysis is fairly easier [68]. The speed layer is the top element of Lambda architecture [82-83]. The speed layer is to use for arbitrary computing function in real time, which is also fill the gap of new data for both batch and serving layer [64].

The speed layer checks the latest data rather than batch layer covers all the data in one batch and and it does in incremental manner. The big data requirements meets with capability of speed layer for low latency and updates. In contrast to MapReduce (batch only), the Lambda architecture can meet all requirements of Big Data query whether it is batch or real time [65].

7. Machine Learning combining Cloud Computing to form Big Data Analysis

With the help of machine learning and cloud computing big data can be easily processed as machine learning has come with some challenging learning methods shown in figure 20 [66].

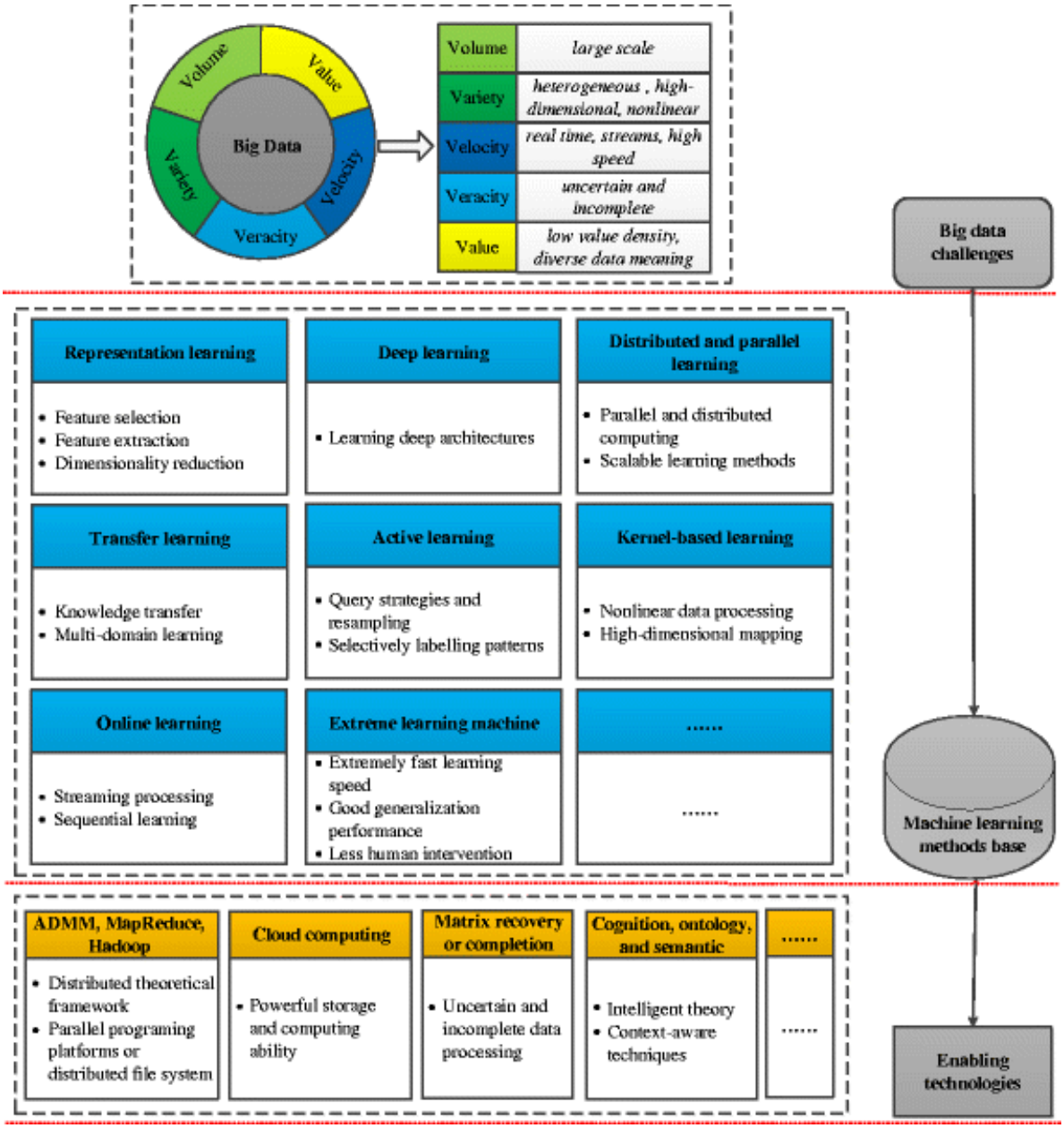

Figure 21. Machine Learning and Cloud Computing to calculate large data

8. Conclusion

Many important issues have been discussed in order to introduce the concept of big data and the concept of three Vs. We just don't limit ourselves up to the concept of 3Vs but

also incorporated 9Vs concept in order to get all the features of big data analysis. Here we discussed many important topics like Hadoop, BDA, and ML shown in figure 21.

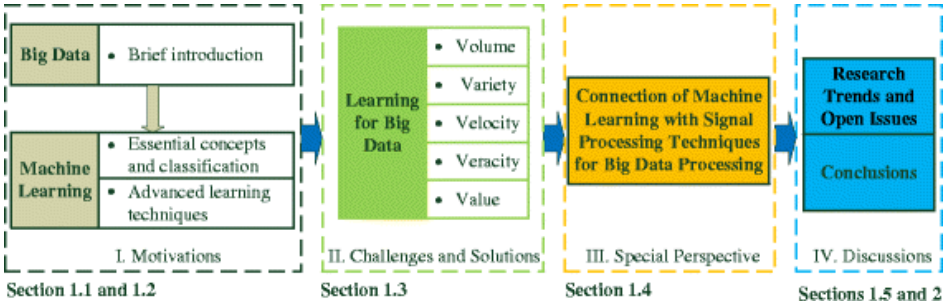

Figure 21. Discussed technology.

Thus we can say now if machine learning is applied on larger dataset in cloud computing paradigm then big data analysis can be done in very efficient manner.

References

[1] A.N. Author, *Book Title*, Publisher Name, Publisher Location, 1995.
[2] A.N. Author, Article title, *Journal Title* **66** (1993), 856–890.
[3] Timothy Paul Smith, How Big is Big and How Small is Small, The size of everything and why, Oxford University Press, 2013,pp 14-29.
[4] Gil Press, "A Very Short History Of Big Data", Forbes Tech Magazine, May 9, 2013. URL:
[5] Fremont Rider, "The Scholar and the Future of the Research Library. A Problem and Its Solution", New York, Hadham
[6] Press, 1944.
[7] Frank Ohlhorst, Big Data Analytics, Turning Big Data into Big Money, John Wiley & Sons, Inc., 2013, pp 2. pp 171
[8] http://www.winshuttle.com/big-data-timeline/
[9] http://au.linkedin.com/pulse/brief-history-big-data-everyone-should-read-bernard-marr
[10] Michael Cox and David Ellsworth, "Application-Controlled Demand Paging for Out-of-Core Visualization", Proceedings of Visualization '97, Phoenix AZ, October 1997, pp 1-12.
[11] 9. Francis X. Diebold, A Personal Perspective on the Origin(s) and Development of 'Big Data': The Phenomenon, the Term, and the Discipline, Second Version, SSRN, PIER working paper No 13-003, http://dx.doi.org/10.2139/ssrn.2202843.
[12] Thomas C. Redman, Data Doesn't Speak for Itself, Harvard Business Review, April 29, 2014.
[13] Lee Gomes, Machine-Learning Maestro Michael Jordan on the Delusions of Big Data and Other Huge Engineering Efforts, IEEE Spectrum, Oct 20, 2014.
[14] Gary Drenik, "Big Data and the Madness of Crowds", Forbes Tech Magazine, USA, Jun 17, 2014.
[15] http://www.forbes.com/sites/prospernow/2014/06/17/big-data-and-the-madness-of-crowds/
[16] Charles Mackay, Extraordinary Popular Delusions and the Madness of Crowds, Harriman House Ltd, 2003.
[17] Donah Boyd et al, Critical Questions for Big Data, Information, Communication & Society Vol. 15, No. 5, June 2012, pp. 662–679.
[18] Chris Anderson, "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete", Wired Magazine,
[19] June 23, 2008. http://www.wired.com/2008/06/pb-theory/
[20] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani, The Parable of Google Flu: Traps in Big Data Analysis,
[21] Science 14 March 2014: Vol. 343 no. 6176 pp 1203-1205
[22] IEEE Computer Society, Rock Stars of Big Data Analytics Presentations, October 21, 2014, San Jose, California.
[23] http://www.computer.org/web/rock-stars/big-data-analytics/presentations

[24] National Research Council of the National Academies, Frontiers in Massive Data Analysis, the National Academy of Sciences, 2013.
[25] James Manyika, et al. Big Data: The next frontier for innovation, competition, and productivity, McKinsey Global Institute, 2011, pp 1-13.
[26] Rob Kitchin, Big Data, new epistemologies and paradigm shifts, Big Data & Society, SAGE, April–June 2014, pp 1–12.
[27] Viktor Mayer Schonberger and Kenneth Cukier, Big Data, A Revolution that will transform how we live, work and think, Houghton Mifflin Harcourt Publishing Company, 2013.
[28] Yves-Alexandre de Montjoy et al. Unique in the shopping mall: On the re-identifiability of credit card metadata, American Association for the Advancement of Science, Science 347, no. 6221 (January 29, 2015): 536–539.
[29] Thomas A. Tweed, Crossing and Dwelling: A theory of Religion, Harvard University Press, 2008.
[30] Irving M. Copi, Carl Cohen, Kenneth McMahon, Introduction to Logic, 14th Edition, Person Education, 2014, pp 83-90.
[31] Douglas Laney, 3D Data Management: Controlling Data Volume, Velocity and Variety, Application Delivery Strategies, Meta Group, 6 Feb 2001, pp 1-4.
[32] http://www-01.ibm.com/software/data/bigdata/
[33] http://www.ibmbigdatahub.com/infographic/four-vs-big-data
[34] Paul C. Zikopoulos et al., Harness the Power of Big Data, The IBM Big Data Platform, McGraw-Hill, 2013.
[35] www.microsoft.com/bigdata
[36] Yuri Demchenko, Defining architecture components of the Big Data Ecosystem, IEEE Collaboration Technologies and System (CTS), 2014, pp 104-112.
[37] Rohan Pearce, "Big data is BS: Obama campaign CTO", CIO Magazine, May 28, 2013.
[38] http://www.cio.com.au/article/462961/big_data_bs_obama_campaign_cto/
[39] Antis Loizides, Mill's A System of Logic Critical Appraisals, Routledge, 2014, pp 11, 192-213
[40] http://www.qualtrics.com/blog/the-1936-election-a-polling-catastrophe/
[41] Bruce Ratner, Statistical Modelling and Analysis for Database marketing Effective Techniques for Mining Big Data, CRC Press, 2003.
[42] Arthur Samuel, Some Studies in Machine Learning Using the Game of Checkers, IBM Journal of Research and Development July 1959, pp 211-229.
[43] Yaser S. Abu-Mostafa et al. Learning from data, a short course, AMLBook.com, 2012.
[44] James Surowiecki, The Wisdom of Crowds, why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations, Anchor Books, 2004, pp 66-83.
[45] http://timoelliott.com/blog/2013/07/7-definitions-of-big-data-you-should-know-about.html
[46] http://www.opentracker.net/article/definitions-big-data
[47] http://books.google.com.au/books/about/Data_Warehouse.html?id=hdRQAAAAMAAJ
[48] http://hortonworks.com/blog/implementing-the-blueprint-for-enterprise-hadoop/
[49] https://451research.com/biography?eid=333
[50] http://thehumanfaceofbigdata.com/
[51] Daniel Larose and Chantal Larose, Discovering Knowledge in Data: An Introduction to Data Mining, John Wiley & Sons, 2014.
[52] Jiawei Han, Micheline Kamber and Jian Pei, Data Mining Concepts and Techniques 3rd Edition, Elsevier Inc., 2012, pp 23
[53] Ian H. Witten and Eibe Frank, Data Mining Practical Machine Learning Tools and Techniques, 3rd Edition, Elsevier Inc. 2011.
[54] Tom M. Mitchell, Machine Learning, McGraw-Hill Science, 1997, pp 2
[55] Kevin P. Murphy, Machine Learning, A Probabilistic Perspective, Kevin P. Murphy, The MIT Press, 2012.
[56] Christopher M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.
[57] Noam Nisan and Shimon Schocken, The Element of Computing Systems Building a Modern Computer from First Principles, MIT press 2005, pp 57-58.
[58] Xin Liu et al, Computational Trust Models and Machine Learning, CRC Press, 2015.
[59] Peter Harrington, Machine learning in Action, Manning Publications, 2012.
[60] Sameer Wadkar, Madhu Siddalingaiah, Pro Apache Hadoop, 2nd Edition, Apress, 2014.
[61] Danil Zburivsky, Hadoop Cluster Deployment, Packt Publishing, 2013.
[62] K.G. Srinivasa and Anil Kumar Muppalla, Guide to High Performance Distributed Computing: Case Studies with Hadoop, Scalding and Spark, Springer, 2015.
[63] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung, "The Google File System", SOSP'03, October 19–22, 2003, pp 1-15

[64] Jeffrey Dean and Sanjay Ghemawat, MapReduce: Simplified Data Processing on Large Clusters, Communications of the ACM, January 2008, pp 107-113.
[65] Gregory Mone, Beyond Hadoop, The leading open source system for processing Big Data continues to evolve, but new approaches with added features are on the rise, Communications of the ACM, Jan 2013.
[66] Jimmy Lin, Chris Dyer, Data-Intensive Text Processing with Map Reduce, Morgan & Claypool, 2010.
[67] Erik Hatcher and Otis Gospodnetic, Lucene in Action, A guide to the Java search engine, Manning Publication Co, 2005.
[68] Dr. Zakir Laliwala and Abdulbasit Shaikh, Web Crawling and Data Mining with Apache Nutch, Packt Publishing, 2013.
[69] Trey Grainger and Timothy Potter, Solr in Action, Forward by Yonik Seeley, Manning Publications Co., 2014.
[70] Alfredo Serafini, Apache Solr Beginner's Guide, Configure your own search engine experience with real-world data with this practical guide to Apache Solr, Packt Publishing, 2013.
[71] http://mahout.apache.org/
[72] Matei Zaharia et al. Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing, Proceedings of the 9th USENIX Symposium on Networked Systems Design and Implementation, April 25-27, 2012.
[73] Holden Karau, Andy Konwinski, Patrick Wendell, and Matei Zaharia, Learning Spark, O'Reilly Media Inc. 2015.
[74] Stephan Ewen, Sebastian Schelter, Kostas Tzoumas, Daniel Warneke, Volker Markl, Iterative Parallel processing with Stratosphere An Inside Look, Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, SIGMOD 13, pp 1053-1056.
[75] Kostas Tzoumas, Apache Flink Next Generation Analysis, http://www.slideshare.net/FlinkForward/k-tzoumas-s-ewen-flinkforward-keynote
[76] Marton Balassi, Gyula Fora, The Flink Big Data Analytics Platform,
[77] https://apacheconeu2014.sched.org/overview/type/big+data#.Vj_-9r8nK1I
[78] Nathan Marz, Big Data Principles and Best Practices of Scalable Real-Time Data Systems, Manning Publications Co. 2015.
[79] Rajkumar Buyya, Christian Vecchiola, and Thamarai Selvi, Mastering Cloud Computing, Morgan Kaufmann, USA, May 2013.
[80] CouchDB, https://en.wikipedia.org/wiki/CouchDB
[81] Caesar Wu, Rajkumar Buyya, Cloud Data Centers and Cost Modeling, Morgan Kaufmann, 2015.
[82] C. J. Date, SQL and Relational Theory How to Write Accurate SQL Code, 3rd Edition, O'Reilly Media, Inc., 2015, pp 523
[83] Dan Sullivan, NoSQL For Mere Mortals, Pearson Education, Inc., 2015.
[84] Joe Celko's Complete Guide To NoSQL, What every SQL professional need to know about Nonrelational databases, Elsevier Inc. 2014.