Big Data: Myth, Reality and Parametric Relationship

*Abdul Alim[1] and Diwakar Shukla[2]
[1,2]*Department of Computer Science and Applications*
Dr. Harisingh Gour Vishwavidyalaya, Sagar (MP)
[1]abdulaleem1990@gmail.com , [2]diwakarshukla@rediffmail.com

.

Abstract

Data becomes big data when the volume of digitalized data has increased rapidly from various sources such as insurance company, medical, education sector, and Social sites. The Social sites are generating a huge amount of datasets at a time.the data has been increased from several years back to till now and approx ever day 2.5 exabyte of data is generated. These datasets are very complex so that the RDBMS (Relational Database Management System)package often have difficulty to handling big data. Big data techniques and technology has the capability to process these complex datasets. Basically, big data expended into 3Vs, Volume, Velocity, and Variety. It must be processed with advanced tools like Hadoop, MapReduce, and Spark. This paper has explored myth and reality of big data with various types of techniques and challenge. Also, we have suggested a relationship between those parameters.

Keywords: *3Vs, 9Vs, Hadoop, MapReduce, big data analytics, cloud computing.*

1. Introduction

We are living within a society where the people constantly interact to each other due to the rapid development of Social computing and Social media services, which has mediated by information technology and take place in the digital realm. Through social online services like Facebook, Twitter, Youtube, Instagram, an average internet user consume and shares large amounts of digital data every day. For example communication text, videos, self-representation, sharing of news and other content. The content may be in structured, unstructured or semi-structured, yet semantically rich data has been argued to constitute 95% of all big data [1].

The big data term has picked up since 2011, this was the year that Gartner introduced big data and extreme information processing and management in its hype cycle. The popularity of this term, there is much debate about the definition of big data. The big data technological advances in data storage, processing and analysis huge amount of data. For example the rapidly decreasing cost of storage and CPU power in a recent year, flexibility and cost-effectiveness of data centers and develop a new framework such as Hadoop.big data analytics can improve information security and situational awareness. It can be employed to analyze financial transactions, log files and network traffic to identify anomalies and suspicious activities [2].

In 2011, Gartner has published a report which in hindsight is often referred to the first description of big data. He has defined the term through information technology challenges described by three Vs: Volume, Velocity, and Variety which is showing pictorial form in *Figure* 1[3]. The report suggested that data creation will continue to grow at a rate of 40 to 60% in a year. A Google analytics tool aggregation search queries reveal that big data queries have grown tenfold in a matter of some 2.5 years. Google was indexing for more than 3.5 billion search queries performed every day. Facebook is handling about a billion content information queries every day [4]. The telecom companies are capturing lots of data volume such as consumers are making more calls and connecting more and more Internet and Smartphone users in the UK tend to do 220

*Corresponding Author

tasks a day [5]. The growth in data can be attributed to a number of technological and economic factors. According to Mckinsey Global Institute report, 2011 on big data, an economic and business research arm of Mckinsey and company highlighted big data analytics as a key-driven in the next wave of economic innovation [6].

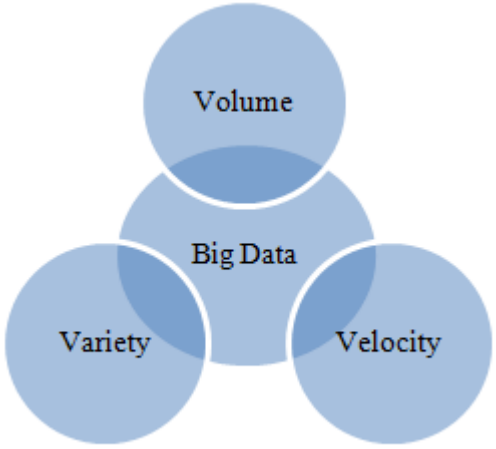

Figure 1. 3Vs in Big Data

2. Big Data Storage

NoSQL, it is familiar group of non-relational database management system which is design for large-scale data storage and for massively parallel data processing across a large number of commodity servers. It can support multiple activities including exploratory and predictive analytics. The *Table* 1 exploring characteristics of NoSQL databases.

Table 1. Characteristics of NoSQL

Strong Consistency	All clients see the same versions of data even on updates to the datasets.
High Availability	The clients can always find at least one copy of the requested data even if some of the machines in a cluster is down.
Partition - Tolerance	The system keeps its characteristics even when being deployed on different servers, transparent to the client.

Table 2. Categories of NoSQL

Category	Description
Key-Value	It is store items as alphanumeric identifiers (key) and associated values in a simple standalone table (hash table). The value may be a simple text string or more complex list and sets.
Document Databases	It is managed and store documents which ended in standard data exchange format such as XML.
Column-Oriented	A column-oriented data structure that accommodates multiple attributes per key. Basically, it is used to distribute data storage and large-scale batch-oriented data processing.
Graph Databases	Graph databases replace the relational table with structured relational graphs of interconnected key-value pairings.

Furthermore, *Table* 2 is explaining classification of NoSQL databases that are in basic four categories.

The various organizations that collect lots of unstructured data which are increasingly turning to no-relational databases [7]. The Bigtable is a distributed storage system for structured data and the emergence of a plurality of derivative versions such as HBase, Cassandra, Hyper table etc. [8]. Cassandra developed within Facebook and it is open source on 2008 Google code and accepted as Apache Incubator project on 2009. It is fully distributed and built based on the Amazon's Dynamo. The Cassandra cluster can run on different commodity servers and also multiple data centers [9]. Bigtable database shares many implementation strategies with databases and also provides a different interface [10].

3. Big Data Mining

Big data mining is a process to extract the meaningful data from huge amounts of data by using various tools and techniques. The popular techniques for big data mining are given below-

Table 3. Big Bata Mining Techniques

Techniques	Descriptions
Regression	It is used predicting values of a dependent variable by estimating the relationship between variables using statistical analysis.
Nearest Neighbor	The values are predicted based on the predicted values of the records that are nearest to the record that needs to be predicted.
Clustering	Clustering values grouping of records that are similar to identifying the distance between them in an n-dimensional space and n is the number of variables.
Classification	Identification of the category or class to which a value belongs to on the basis of previously categorized value.

There are several open source tools are available for processing complex data like big data. *Table* 4 explains some top-level tool which helps in big data mining.

Table 4. Big Data Mining Tools

Open source tool	Description
Mongo DB	It is a cross-platform document-oriented database management system.
Hadoop	It is a framework that allows distributed processing of big data sets across clusters of networked computers using simple programming models.
MapReduce	It is a programming model and framework based on Hadoop which enables processing huge amount of data in parallel on large clusters of compute node.
Orange	It is a python based tool for processing and mining big data. It has drag and drops functionalities with a variety of add-ons that's why easy to use interface.
Weka	It is a Java-based tool for processing a large amount of data. It has a varied selection of algorithms that can be used in mining big data.

The above *Table* 4 highlighted some top-level big data tools which are open source and it processes the huge data [11].

4. Big Data Analytics

The goal of big data analytics is to extract useful value and support decision making. Big data analytics basically faces three types of challenges like storage, management, and processing. Integrating heterogeneous information from various sources provides a holistic view of the domain [12]. There is lots of data available in various sectors. The size of volumes are constantly increasing and currently ranging from terabyte to many petabytes of data in the single dataset. In big data, analytics is where advanced analytic techniques are applied to datasets and extract meaningful information [13]. Big data analytics can be used in various fields like e-commerce and marketing intelligence, e-Government, science and technology, smart health and wellbeing and security and public safety [14]. The following *Figure* 2 shows some types of analytics-

Figure 2. Types of Big Data Analytics (Source: 15)

5. Cloud Computing and Big Data

Cloud computing is a fastest growing technology which has established in IT industries and businesses. It has provided reliable software, hardware, and Internet as a service delivered over the Internet and remote data center.

Table 5. Cloud Computing Services

Platform as a Service (PaaS)	The delivered services like Google's Apps engine, salesforce.com, force platform and Microsoft Azure refers to different resources operating on a cloud to provide platform computing for end user
Software as a Service (SaaS)	Services such as Google Docs, Gmail, salesfore.com, and online payroll, refers to applications operating on a remote cloud infrastructure offered by the cloud provider.
Infrastructure as a Service (IaaS)	Flexi scale and Amazon's Ec2 refers to hardware equipment operating on a cloud provided by service providers and used by end users upon demand.

The cloud computing basically providing three types of services. *Table* 5 shows the services of cloud computing.

Big data hs utilized distributed storage technology based on cloud computing rather than local storage attached to a computer device. The following *Figure* 3 has shown the relationship between cloud computing and big data.

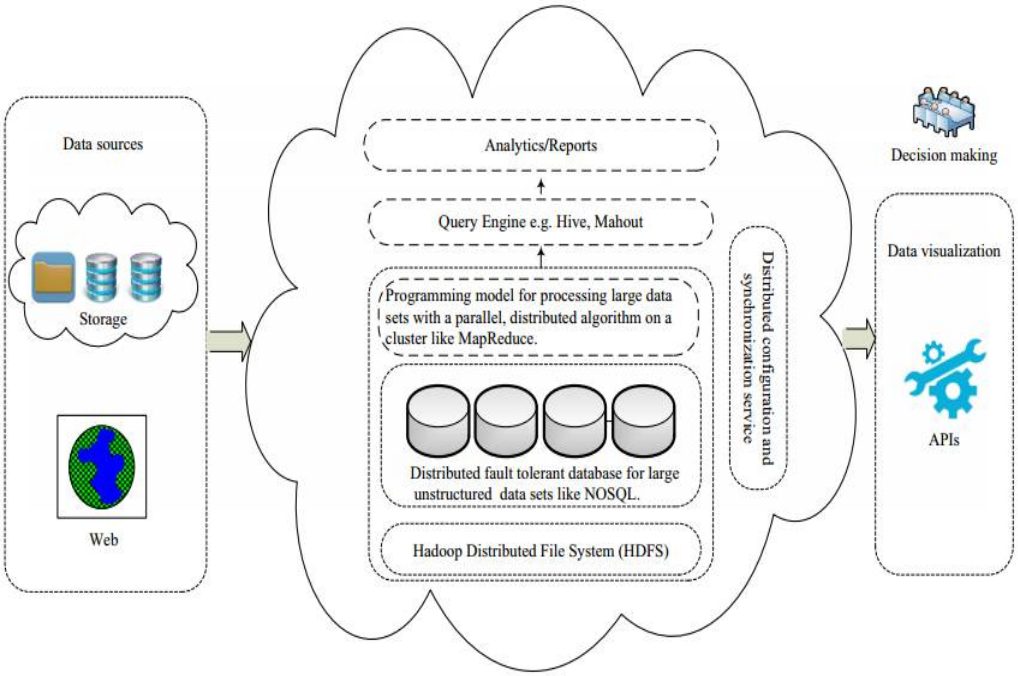

Figure 3. Relationship between Cloud Computing and Big Data (Source: [16])

In this *Figure* 3 has given large data sources from cloud and web are stored in a distributed fault-tolerant database and processed through a programming model for huge datasets with a parallel distributed algorithm in a cluster [16].

6. Big Data Parameters

There are basically three parameters in big data that are Volume, Velocity, and Variety but after this parameters, some authors have collected six another parameter or characteristics then now 3Vs become 9Vs.

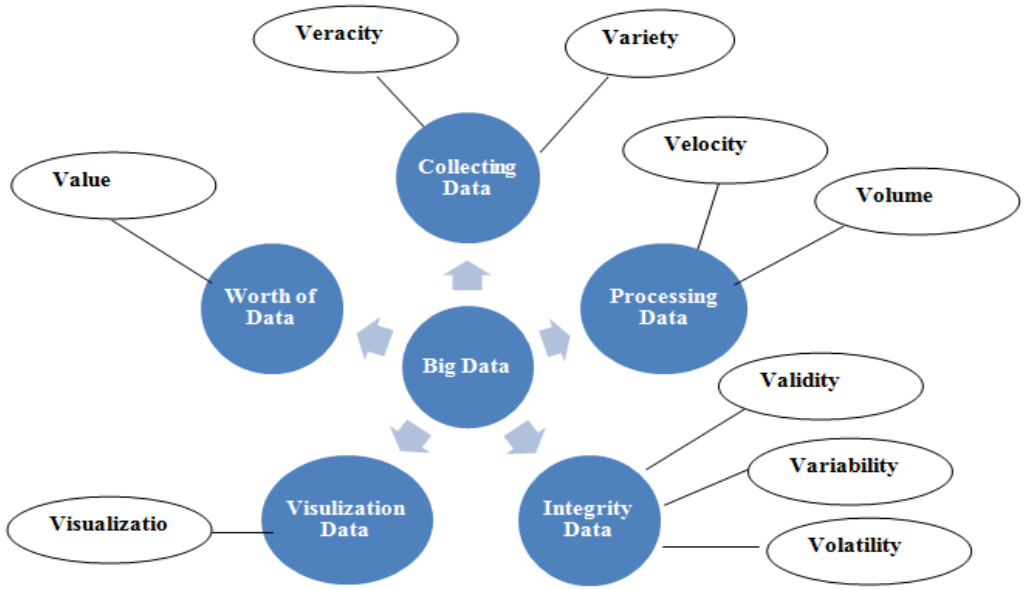

Figure 4. Five Categories of Big Data with their Characteristics (Source: [17])

The *Table* 6 explains 9Vs of big data such as Veracity, Variety, Velocity, Volume, Validity, Variability, Volatility, Visualization, and Value. The big data parameters can be categorized into five categories. The categories are Collecting Data, Processing Data, Integrity Data, Visualization Data and Worth of Data. Figure 4 shows five categories of big data with 9Vs.Table 6 is describing the importance of 9Vs.

Table 6. Big Data Mining Techniques

Parameter Name	Description
Veracity	It is referred to the biases, noise, and abnormality in data.
Variety	Various types of data like structured, semi-structured and unstructured.
Velocity	It is referred to, how fast data is to be produced and processed to meet the demand.
Volume	Volume is a size of data such as terabyte, petabyte, Exabyte, zettabyte etc.
Validity	The data is correct and accurate for the intended.
Variability	Along with the velocity, the data flows may be highly inconsistent with the data. The need to be found by anomaly and outlier detection methods in order for any meaningful analytics to occur[18]
Volatility	Recall the retention policy of structured data that the implement every day in the businesses.
Visualization	It makes all that huge amount of data comprehensible and easy to understand and read.
Value	It has a low-value density as a result of extracting value from massive data

The above *Table* 6 has been clarified 9Vs of big data were grouped in a cluster to get five categories [17].

7. Suggested Relationship among Parameters

In the above table, we have discussed the various type of parameters which are involved in big data. According to this, we have suggested the relationship among those parameters V_1= Volume, V_2= Velocity, V_3=Variety, V_4=Veracity, V_5=Validity, V_6=Variability, V_7=Volatility, V_8=Visualization, and V_9=Value.

[A]: $V_1 = f(V_3)$, V_1 varies according to whatever value V_3 have been taken.

We say that $V_1 \propto V_3$

So $V_1 = K_1V_3$, where K_1, is a constant

The logical part is that variety (V_3) of data leads to more volume (V_1) due to the creation of several backup tables and their additional storage for each variety. Therefore constant k1 calculating it's a challenging problem.

[B]: $V_2 = f(V_3)$, V_2 varies according to whatever value V_3 takes on.

We say that $V_2 \propto V_3$

So $V_2 = K_2 V_3$, where K_2, is a constant

V_2 (Velocity) is fully depended on V_3(Variety), we cannot imagine that within 1 minute how much data generated by the various portal in the form of text, image, video, comments etc. so it's difficult to calculate k_2. Higher in the data generation, more is the expected velocity.

[C]: $V_4 = f(V_1, V_3)$, V_4 varies according to whatever value V_3 takes.

We say that $V_3 \propto (V_1, V_3)$

So $V_4 = K_3 V_1.V_3$, where K_3 is a constant

Here the logic is that V_4 will increase when V_1 and V_3 make any changes because V_4 is is the veracity which is biased. Noise and it should be increased when volume and variety have been increased.

[D]: $V_5 = f(V_1, V_3)$, V_5 varies according to whatever value V_1, V_3 takes on.

We say that $V_5 \propto (V_1, V_3)$

So $V_5 = K_4 V_1.V_2$, where K_4, is a constant

To find correct and accurate data from the huge amount of data so its need to data should be in the structure for but here when V_1 and V_3 have increased then find the value of K_5 is very difficult.

[E]: $V6 = f(V_2)$, V_6 varies according to whatever value V_2 takes on;

We say that $V_6 \propto V_2$

So $V_6 = K_5 V_2$, where K_5, is a constant

Here to calculate K_6 is very difficult because V_2(Velocity) is changing continuously and may be in highly consistent with the data.

[F]: $V7 = f(V_3)$, V_7 varies according to whatever value V_3 takes on;

We say that $V_7 \propto V_3$

So $V_7 = K_6 V_3$, where K_1, is a constant

V_7(Volatility) need to changeable policy and further, it can implement within the business but when there are a variety of data available that's why if V_3(variety) has increased then K_7 calculating the challenging task.

[G]: $V_8 = f(V_1, V_2)$, V_8 varies according to whatever value V_1, V_2 takes on;

We say that $V_8 \propto (V_1, V_3)$

So $V_8 = K_7 (V_1, V_3)$, where K_7, is a constant

It depends on volume and variety because V_8(Visualization) of the data will be difficult when V_1 and V_3 are increasing continuously then it's a challenging task to find the visualization data.

[H]: $V_9 = f(V_1, V_3)$, V_9 varies according to whatever value V_1, V_3 takes on;

We say that $V_9 \propto (V_1, V_3)$

So $V_9 = K_8(V_1, V_3)$, where K_8, is a constant

It also depends on V_1 and V_3 because for finding V_9(Value) to valuable information from the massive amount of data.

8. Big Data Challenges

The key set of challenges faced in today's tight market is the need to find and analyze the required data at the least speed possible[19]. Jae-Gil Lee and Minseo Kang have discussed some global challenges which are facing by the humanities such as Sustainable Development and Climate Change, Clean Water, Pollution and Resources, Democratization, Global Foresight, and Decisionmaking, Global Convergence of Information Technology, Rich-Poor gap, Health issues, Education, Peace and Conflict, Status of Woman, Transactional Organized Crime, Energy, Science and Technology and Global Ethics[20] and also some another challenge within big data which are shown in *Table* 7.

Table 7. Challenges in Big Data (Source: [21])

Challenges	Description
Data Representation	Heterogeneity in type, structure, semantics, organizations, granularity, and accessibility
Redundancy reduction and data comparison	There is high-level redundancy in datasets, most data generated by the sensor networks are highly redundant
Data lifecycle management	A data importance principle related to the analytical value should be developed to decide which data shall be stored and which data shall be discarded
Analytical mechanism	Traditional RDBMSs are strictly designed with a lack of scalability and expandability which could not meet the performance requirements
Data confidentiality	Big data service providers at present could not effectively maintain and analyze such huge datasets because of their limited capacity. They must rely on professionals or tools to analyze such data which increase the potential safety risk.
Energy management	Processing, storage, and transmission of big data will consume more and more electronic energy
Expandability and scalability	The analytical system of big data must support present and future datasets
Cooperation	Big data network architecture must be established to helps scientist and engineers in various fields access different kinds of data and fully utilize their expertise

9. Conclusion

In this paper, we have discussed the current state of the big data which are involving in various sectors, the big data parameters or characteristics and their importance. This paper explores the storage of big data in databases like NoSQL and its categories. It has been discussed data mining techniques which are used to extract useful information from complex data. We have also explored the relation between big data and cloud computing, global challenges which is facing humanity in daily life. Through this paper, one can address the reality of big data and its possible relationship between the parameters.

10. References

[1] Olshannikova, Ekaterina, Olsson, Thomas, Huhtamaki, Jukka, and Karkkainen, "Conceptualizing Big Social Data", Journal of Big Data, vol. 4, no. 3, **(2017)**, pp. 1-19.

[2]. Cárdenas, Alvaro A., 2013, "Big Data analytics for security intelligence" Cloud Security Alliance, available at www.cloudsecurity alliance.org/ research/big-data.

[3] Altena, Allard J. Van, Moerland, Perry D., Zwinderman, Aeilko H. and Olabarriaga, Silvia D., " Understanding big data themes from scientific biomedical literature through topic modeling", Journal of Big Data, vol. 3, no. 23, **(2016)**, pp. 1-21.

[4] Bughin, Jacques, "Big Data, Big Bang?", Journal of Big Data, vol. 3, no. 2, **(2016)**, pp. 1-14.

[5] Bughin, Jacques, " Reaping the benefits of big data in telecom", Journal of Big Data, vol. 3, no. 14, **(2016)**, pp. 1-17.

[6] O'Donovan, P., Leahy, K., Bruton, K. and O'Sullivan, D.T.J., " An industrial big data pipeline for data-driven analytics maintenance applications in large-scale smart manufacturing facilities", Journal of Big Data, vol. 2, no. 25, **(2015)**, pp. 1-26.

[7] Moniruzzaman, A. B. M., and Hossain, Syed Akhter, " NoSQL Database: New Era of Databases for Big data Analytics - Classification, Characteristics, and Comparison", International Journal of Database, Theory and Application, vol. 6, no. 4., **(2013)**, pp. 1-13.

[8] Liu, Na, and Zhou, Jianfei, "The Research and Application of a Big Data Storage Model", International Journal of Database Theory and Application, vol.8, no.4, **(2015)**, pp.319-330.

[9] Ben Brahim, Mohamed, Drira, Wassim, Filali, Fethi and Hamdi, Noureddine, " Spatial data extension for Cassandra NoSQL database", Journal of Big Data, vol. 3, no. 11, **(2016)**, pp. 1-16.

[10] Chang, Fay, Dean, Jeffrey, Ghemawat, Sanjay, Hsieh, Wilson, Wallach, Deborah A., Burrows, Mike, Chandra, Tushar, Fikes, Andrew and Gruber, Robert E., "Bigtable: A Distributed Storage System for Structured Data", vol. 26, no. 2, **(2008)**, pp. 1-14.

[11] Sin, Katrina and Muthu Loganathan, "Application of Big Data in education Data Mining and Learning Analytics- A Literature Review", ICTACT Journal on Soft Computing, vol. 5, no. 4, **(2015)**, pp. 1035-1049.

[12] Fan, Shaokun, Lau Raymond Y.K. and Zhao J. Leon, "Demystifying big data analytics for business intelligence through the lens of marketing mix", Journal of big data research, vol. 2, no. 1, **(2015)**, pp. 28-32.

[13] Elgendy, Nada and Elragal, Ahmed, "Big Data Analytics: A Literature Review Paper", Springer International Publishing Switzerland, Proceeding of the 14th Industrial Conference st. Petersburg, Rasia, **(2014)** July 16-20, pp. 214-227.

[14] Chen, Hsinchun, Chiang, Roger H. L. and Storey, Veda C., "Business Intelligence and analytics: from Big Data to Big Impact", MIS Quarterly, vol. 36, no. 4, **(2012)**, pp. 1165-1188.

[15] Kaiser, Stephen, Armour, Frank, Espinosa, J. Alberto and Money, William, "Big Data: Issues and Challenges Moving Forward", Proceeding of the 46th Hawaii International Conference on System Sciences, Grand Wailea, Maui, Hawaii, **(2013)** January 7-10.

[16] TargioHashem, IbrahimAbaker, Yaqoob, Ibrar, Anuar, Nor Badrul, Mokhtar, Salimah, Gani, Abdullah, and Khan, Samee Ullah, "The rise of big data on cloud computing: Review and open research issues", Journal of Information System, vol. 42, **(2015)**, pp. 98-115.

[17] Owais, Sushil Sami and Hussein, Nada Seal, "Extract Five Categories CPIVW from the 9V's Characteristics of the Big Data", International Journal of Advanced Computer Science and Applications, vol.7 no. 3, **(2016)**, pp. 254-258.

[18] Firican, George, "The 10 Vs of Big Data", **(2017)**, pp. 1-6, available on https://tdwi.org/articles/2017 /02/0 8/10vsofbigdata.aspx accessed on 22.01.2018.

[19] Mukherjee, Samiddha, and Shaw, Ravi, "Big Data – Concepts, Applications, Challenges and Future Scope", International Journal of Advanced Research in Computer and Communication Engineering, vol. 5, no. 2, **(2016)**, pp. 66-74.

[20] Lee, Jae-Gil and Kang Minseo, "Geospatial Big Data: Challenges and Opportunities", Journal of Big Data Research, vol. 2, no. 2, **(2015)**, pp. 74-81.

[21] Chen, Min, and Mao, Shiwen, Liu, Yunhao, "Big Data: A survey", Mobile Network and Applications, Springer Link, vol. 19, no. 2, **(2014)**, pp. 171-209.