# Challenges in Handling Big Data

**M.Sadiq Ali Khan[1] , Huma Jamshed[2]**

[1]Assitant Professor Computer Science Department, University of Karachi, Karachi Pakistan

[2]Lecturer, Computing Department, Shaheed Zulfiqar Ali Bhutto Instiute of Science and Technology (SZABIST), 90 & 100 Clifton, Karachi, Pakistan

## ABSTRACT

***With the influx of enormous information accumulating, all over data storage and warehousing facilities, "BIG DATA" became a reality. This multifaceted and colossal amount of data, pouring through various sources, brings versatile complexities to all aspects of management, such as, managerial systems, applications development and network management. These structural modules are all interrelated, complimenting each other, and the system will defunct if not completely collapse if either will seize to function in its true capacity. However, extra ordinary considerations become necessary towards network modules, primarily invoking network performance, network structure, network security and in turn privacy, especially with amplifying demand for real time applications. Variation in input devices, from hand held gadgets to sophisticated industrial equipment and now further elongating the parameters to household equipment, transferring data through various platforms of internet communication topology, come with pertinent network security issues. Infrastructures leading towards big data processing facility need exceptional security measures complimenting the system and associating applications.***

**Keywords:** Big Data; 4Vs, Traditional Data, Technologies, NoSQL, Privacy; Security

## 1. INTRODUCTION

With the rise in Internet popularity and speedy development of emerging applications such as, web and social network analysis, variety of data to be processed observe drastic increase periodically. The management and analysis of data in a large-scale infrastructure poses a significant challenge. For enhanced association between business accomplices, an appropriate IT infrastructure is need for utilization of big data [1].

Traditional security and data processing mechanism are inadequate for huge data networks. Big Data can be defined as; collection of massive amounts of digital information by different organizations for which special data processing applications are required for data analysis. The size of data sets is huge that it is beyond the ability of current technology and technique to manage and process the data within an acceptable elapsed time. The data move so fast that it cannot fit the organization of traditional database architecture. The Big Data usually have low value for usage before handing it out. The data usually have diverse structure, which signifies as NoSQL (not only SQL) [1]. For instance, the variety is due to sources such as transactional records, business documents, images, recording, web pages, and social media updates etc. [2]. Cloud and grid computing architecture are accessing large amount of computing power accumulating resources and offering single system view. Intend of these technologies is to resolve and undertake big data issues [3]. Publication in Journal of Science 2008 defines big data as representation of progress of human cognitive processes with data size beyond the processing capability of current technology [5].

Douglas defined Big Data in his research work as high velocity, high volume, high variety information that require new forms of handing out to enable improved decision making, discovery and process optimization. Big data is a transformative, persistent avalanche that is not disappearing but it just keeps accelerating [6]. Organizations are rapidly investing on implementations of big data programs for strategic changes in organization business model to gain a competitive advantage and expand their global presence. The size and complexity of Big Data has affected the area of network and it has become emerging hot topic in network security fields. The data collected from heterogeneous network sources triggers security and privacy issues. Big data sets are generally understood by their four characteristics ; the volumes of data up till peta-byte; Varity data types, including logs, audio files, text files, pictures videos, geographical location information, fast processing speed velocity and data Veracity. Big Data bring significant security privacy and transfer risks that are real, magnified and will continue to grow. The future trends towards analytics and use of big data leads to erosion of privacy and user's access rights. This paper primarily focuses on challenges in handling big data networks.

## 2. BACKGROUND

With the speedy development of web technology and emerging applications, massive information is accumulating all over data storage and warehousing facilities making "BIG DATA" a reality. For example, social network website accumulates around millions of data per month and uses artificial-intelligence strategies for business decision making. Database systems, which are involved in knowledge discovery, still face problems of efficient access of the data. The query formulations need properly structuring for accessing tera (peta) bytes of data. In year 2012, Big Data Research and Development Initiative", came into existence as national policy for big data research by American government. Considerations were to invest on resources, which were cable of intensive data operation, which resulted in high data processing cost and demanded high storage systems.

With the realization of the web and internet technologies, more IT organizations have, expanding needs to accumulate and investigate the data collected from various sources such as web crawlers, search logs and variety of web services. The data is usually unstructured and non-relational thus poses challenges for present data processing technologies [5][13].Google is providing distributed storage system "Big-table" for managing structured data that is above petabytes. Big-table is a simple model that supports dynamic control over data design and format [4]. On cloud or grid architecture, cost of data communication is major concern as compared to data processing. Network parameters such as bandwidth and latency face different challenges while communicating client and server [11] [12].

Data security is another important factor of computer networks as on cloud infrastructure the security mechanism is generally weak and data can be easily tampered, which is one of the major concerns of technologist [10]. With the rapid rises in information technology, organizations are trying to fetch mean full information from the data accumulated from different sources on daily basis. Data gathered from different sources such as mobile devices, radio frequency identification, sensors, streaming servers etc, which is in raw form, diverse, noisy and variant. Until now, scientists are not able to combine the fundamental features of big data. According to many organizations, handling of big data is difficult using convenient technology and theory. It boosts challenges for data management, security, privacy and analysis, and for the entire IT industry. Usually of big data, analytics involves processing personal data, which demands data protection by ensuring the processing of personal data is fair and is important for decision-making when it is affecting individual.

**Table 1**: Comparisons of Conventional and Big Data Analysis

| | **Conventional Analysis** | **Big Data Analysis** |
|---|---|---|
| Size | Mega bytes/ giga bytes | Terabytes/ petabytes |
| Technology | Relational Databases | Hadoop Platforms |
| Processing | Batch or Statistical | Online or Real-time |
| Operational | Individual Organization | Everyone/related organizations |

## 3. BIG DATA CHARACTERISTICS

The terminology "Big Data" indicates large and composite data sets made up of a diverse of structured and unstructured data, which are complex to manage by conventional techniques [7]. Focusing at the fast-growing volume of digital data, the McKinsey Global Institute estimates data volume to increase at rate of 40% per year. According to the institution by 2018, there will be 50 to 60 percent gap between the availability and the essential demand of deep analytic talent and data managers, respectively in United States alone [14]. Big Data represented by four V's, which are velocity, variety, volume and veracity [8] [9].

Velocity specifies the speed with which data is being generated and how fast it must be processed, volume refers data quantity and size whereas variety indicated the diversity of data types. Veracity means the uncertainty, inconsistency, and ambiguity in data, which is difficult to understand and perceive.

Technologist suggests that, of the 'four Vs', variety is the most important feature of big data. This observation put forwards that, if a company is investigating its own patron database, it may not raise issues related to analytics or data protection. However, when it combines external data sourced with its own database, then it is doing something, which is challenging and can raise privacy and security related issues.

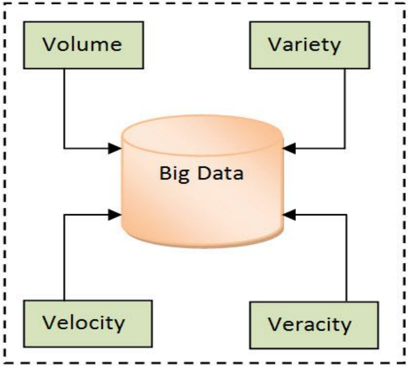

**Figure 1** The fours V's of Big Data

## 4. BIG DATA NETWORK CHALLENGES

Big Data is complex and immense amount of data, pouring through various sources, brings versatile complexities which significant increase security, privacy and transfer risks that are existent and will persist to escalate. The issue of security and privacy are magnified by Big Data 4V's characteristics, for instance a cloud infrastructure with diverse data sources and data formats, the inter cloud migration fails the traditional security mechanism and results in data which is inadequate [15]. Concerns required by the organizations who are handling big data to identify these risks and adopt techniques to avoid data breach. Companies must be practical in considering any information security risks posed by big data. Organizations should apply proper risk assessment strategies and introduce risk management policies. In short, it should adopt methods to improve information security. For example, Law enforcement organization has turned to Data Surveillance to fight against crime, where as social networks accumulate massive amount of data for creating larger, searchable digital footprints. Major big data specific management, privacy and security issues and challenges are;

### 4.1 Distributed Platform Issue

Distributed systems apply parallel programming framework for processing data. One of the examples of such system is map reduction algorithm. Risks associated with map reduction systems are secure mappers and data in the presence of un-trusted mapper. There can be an intentional or unintentional leakage of personal information by mapper.

### 4.2 Data Storage and Managements

Existing data management systems are not able to satisfy the requirements of big data, and size of data is increasing day by day as compared to storage capacity which is much less than that of data, thus a uprising renovation of information framework is urgently needed. Proposed solution is to design and implement hierarchical storage architecture. Such system, perform excellent in processing homogeneous data but gives poor performance for heterogeneous data. Therefore, data organization and classification is one big problem in big data management. Other issue related to management includes bottleneck problems in master slave configurations.

### 4.3 Data in Multi-tier

Data structured in multi-tiers, requires continuous movement of data between different layers. Data transaction logs are also stored in this multitier system. Auto-tiering solutions are introduced as data size is growing exponentially which does not keep track of where data is stored thus introduces a new challenge of secure data storage.

### 4.4 Big Data Governance

Big data implementation may lead to discovery of earlier undisclosed sensitive information through amalgamation of different data sets. Organizations attempting to implement big data initiatives requires strong governance regime, in order to prevent ethical dilemma. Therefore, a strong ethical code of conduct along with training is required in handling big data.

### 4.5 Access Control and Secure Communication

Sensitive information is being stored in unencrypted form in cloud infrastructure. The main problem in encrypting data in large-scale sharable network is that it disallow user in performing fine-grained data sharing activities. This issue reduced by implementing public key encypto-system that is un-encrypt less sensitive information where as sensitive data should have proper access control mechanism.

### 4.6 Security and Privacy

The concept of cloud computing and big data networks have reduced over all IT cost. By using third-party tools and services have raised security and privacy issue in handling data. Challenges are hoist as data rate is growing exponentially. Dynamic data monitoring and protection is a big challenge of current era as sensitive information exposed to big data users. Therefore, privacy policy implementation needs attentions and legal regulatory bodies' intervention needed.

### 4.7 Data Reliability

Reliability big data is a serious challenge on a network. From one perspective, valuable information extracted from unstructured data using techniques such as data cleaning; on the other hand, it is essential to attain secure data access and privacy protection. These two characteristics are key requirements of big data reliability.

## 5. RESULTS AND DISCUSSIONS

With the advancement and digitalization of business and society, immense pressure is developing on management and utilization of resources available. The data accumulated at exponential rate on daily basis is so compound and diverse that it poses numerous questions regarding it safety and usage. The experts have concluded that the existing processing techniques and technology is insufficient to process such a complex mass of data. Researcher suggests techniques such as parallel processing (MAP Reduce) to handle and deal with this unstructured data. For enhanced storage and management, good choice is use of distributed file system and NoSQL databases. As far as privacy in big data contexts encloses a huge amount of challenges and concerns. Most of them are based on organizational and legislation matter.

## 6. CONCLUSION

This paper presents some of issues in handling big data networks. Big data is a hot topic of discussion of current era. Numbers of fields including marketing, scientific research and government agencies are using big data analytics. The transformation of data into information and knowledge, depend on the technical capabilities of Big Data. Numerous literatures reviews, highlights big data technologies challenges and address these issues under development stage. Extra ordinary considerations become necessary towards network traffic is an affluent foundation of information for security monitoring.

## REFERENCES

[1] S. Robak, B. Franczyk, M. Robak, “Research Problems Associated with Big Data Utilization in Logistics and Supply Chains Design and Management,” Federated Conference on Computer Science and Information Systems, pp. 245-249, 2014. Available: https://fedcsis.org/proceedings/2014/pliks/472.pdf. [Accessed: June, 2015].

[2] S. Robak, B. Franczyk, M. Robak, “Applying Big Data and Linked Data Concepts in Supply Chains Management”, Proceedings of the Federated Conference on Computer Science and Information Systems FedCSIS. IEEE Conference Publications, pp 1203-1209, 2013. Available: http://annals-csis.org/BA554816-A125-40D8-95F7-FCFD63642184/FinalDownload/DownloadId-8F7286899B1525BC29FAE9084933/BA554816-A125-40D8-95F7-FCFD63642184/proceedings/2013/pliks/269.pdf [Accessed: June, 2015].

[3] J.Changqing , L.Keqiu , A. Uchechukwu , Q. Wenming , L. Yu, "Big Data Processing in Cloud Computing Environments", International Symposium on Pervasive Systems, Algorithms and Networks , pp 17 - 23, 2012

[4] F. Chang,J. Dean, S. Ghemawat ,W.Hsieh, D. Wallach, "Bigtable: A distributed structured data storage system", 7th OSDI, pp 305–314. 2006

[5] Nature 455 (7209), Big data: science in the petabyte era, 2008, Available: http://www.nature.com/nature/journal/v455/n7209/edsumm/e080904-01.html [Accessed: June, 2015]

[6] M.A Beyer, D. Laney, " The importance of ‘big data’: A definition", Gartner, 2008

[7] M.Cafarella, A. Halevy, "Web data management" , Proceedings of the 2011 ACM SIGMOD International Conference on Management of data , pp 1199–1200, 2011

[8] D. deRoo, T. Deutsch, C. Eaton, G. Lapis, P. Zikopoulos,” Understanding big data- Analytics for enterprise class Hadoop and streaming data”, McGraw-Hill, 2012

[9] S. Suthaharan, "Big Data Classification: Problem and challenges in Network Intrusion Prediction with Machine Learning", ACM SIGMETRICS Performance Evaluation Review archive Volume 41 Issue 4 ,pp 70-73, 2014

[10] C. Barton, I. Muttik, "Cloud security technologies", Information security technical report, pp 1-6.2009

[11] S. Carlin, K. Curran, "Cloud Computing Technologies", International Journal of Cloud Computing and Services Science (IJ-Closer) , pp 59-65, 2012

[12] C. Chen, R. Johnson, W. Shen, B. Ross, P. Wong, "Top ten challenges in Extreme Scale Visual Analytics", Computer Graphics and Applications, IEEE, pp 63-67, 2012

[13] N. Yuhanna, " Today's challenge in government: What to do with unstructured information and why doing nothing isn't an option", Forrester Research. 2012

[14] McKinsey Global Institute, " Big Data: The Next Frontier for Innovation, Competition and Productivity",2012

[15] Murthy, Praveen K., "Top ten challenges in Big Data security and privacy", Test Conference (ITC) IEEE International, 2014

## AUTHOR

**M.Sadiq Ali Khan** received his BS and MS degree in Computer Engineering in 1998 and 2003 respectively. He received his Ph.D degree in computer science in 2011. He is currently an Assistant Professor in Computer Science Department University of Karachi. He is also member of various higher academic boards of different universities. His research interest includes Network communication, network security, artificial intelligence, mobile communication system and cloud computing.

**Huma Jamshed** received her B.E and M.E degree in Computer Engineering from NED University of Engineering & Technology in year 2008 and 2011 respectively. She is PhD Student in University of Karachi and doing her PhD research under the supervision of M. Sadiq Ali Khan. She is Lecturer in Computer Science Department at Shaheed Zulfiqar Ali Bhutto Institute of Science and Technology (SZABIST). Her research area is networking and she is focusing on issues related to data management security and privacy issues in cloud and other distributed network environment.