**American Journal of Information Science and Computer Engineering**
Vol. 1, No. 2, 2015, pp. 38-43
http://www.aiscience.org/journal/ajisce

# BIG-Data Challenges: A Review on Existing Solutions

**Sheikh Muhammad Saqib*, Hamid Masood Khan, Khalid Mahmood, Tariq Naeem**

Institute of Computing and information Technology, Gomal University, Dera Ismail Khan, Pakistan

## Abstract

Big Data is a new term in today technology era. Data generation rate is turning as well as exceeding from Peta byte to Exa byte. Companies take major decisions emphasizing their Big Data. To take decisions, there is need of management of such data. Although Map Reduce and Hadoop is playing a vital role in management but there are still some challenges to be managed. Normally 3Vs (Volume, Velocity and Variety) has been focused. Besides these 3Vs authors have explored many other challenges related to Big Data and as a result, they provided the description and existing solutions related to each issue. Results would be very fruit full for all those working on Big Data.

## Keywords

Big Data, Big Data Challenges, Volume of data, Velocity of data, Variety of Data, Map Reduce and Hadoop

Received: April 17, 2015 / Accepted: April 30, 2015 / Published online: June 8, 2015

## 1. Introduction

Big Data means a dataset with huge amount of data with high volume, velocity and variety data which requires efficient processing for decision making [12].

In existing work data was generated by internal employs, but nowadays employees, partners and customers are considered as a source of data. Machines are also considered as a source for data generation i.e. through smart phone lot of data is generating on daily basis [17].

Before making decision from Big Data, there are lot of challenges mostly 3Vs which should need some attentions. Different researches [12] [13] [14] have done lot of work to handle such challenges. Such challenges can be resolves by using Map Reduce as pre-processing. There is also lot of work [2] [3] in Map Reduce.

Each new technology have brought many advantages, major advantages of Big-Data is cost reduction, substantial improvements in the time required to perform a computing task, or new product and service offerings [15]. Before utilizing information from Big-Data, challenges should be covered.

Here we explored all the challenges related to Big Data and found their solutions, and also explored all those challenges which required further research. We examined fourteen challenges and found out how each challenge was paid attention by the researchers previously with respect to its solution. In proposed solution we suggested that apart from two challenges, three challenges require serious attention for solution and remaining eight challenges have already been provided solutions but require attention for further improvement.

## 2. Big Data Challenges and Solutions

### 2.1. Heterogeneity

If data is not in natural language or in heterogeneity format

*Corresponding author
E-mail address: saqibsheikh4@hotmail.com (S. M. Saqib)

then it may not provide valuable depth, because machine algorithms are suited for homogeneous data. So for data analysis data must be structured carefully [1]. Data can be structured by using Map Reduce techniques, because all keys generated by the Map Reduce must fit into main memory [10].

## 2.2. Inconsistency and Incompleteness

Since big data comprises of lot of information coming from different sources (each source has different nature of reliability), so error and missing values is a challenge for managing them. These errors and incompleteness must be managed in data analysis [1]. However, there is also challenges include with data analysis such as efficient representation, access, and analysis of unstructured or semi-structured data in the further researches. To remove noise and correct inconsistencies, different types of data preprocessing techniques can be applied such as data cleaning, data integration, data transformation and date reduction [6].

## 2.3. Scale

Size of big data is an important challenge, although there are many researches to handle this issue such as handle big data with processor speed but some time increasing volume of data is faster than processor speed [1][6]. In Big Data applications, the state-of-the-art techniques and technologies cannot ideally solve the real problems, especially for real-time analysis. As Big Data requires a more storage and medium, if Hard Disk Drives (HDDs) are used for such purpose, then HDDs is slower than data processing engines, this challenges can be handled by using Solid State Drives (SSDs) and Pulse-code modulation (PCM) technologies [6]. A sheer volume of data requires very high speed [4] because data grows as the degree of granularity increases. To store such data there is also memory issues. Faster growth of data and memory issues can be solved by using grid computing approach. Big Data is comprised of large no of inputs, outputs and attributes, these are lead to the complexity related to running time [9], and to handle such types of issue distributed frameworks with parallelized machines are preferred.

## 2.4. Velocity of Data

In Bigdata, data is generated with very high speed, and this speedy data requires processing in timely manner. This problem in learning algorithm can be solved by using online learning approaches [9].

## 2.5. Timeliness

How data without outlier can be filtered at real time for storage purpose [1]. This issue can be handled through Index structure of traffic management system. If data is not analyzed quickly and there is not proper framework for users then data for decision making will not be fruitful. To address data quality, there are some proactive methods that can address data quality and timely [4].

## 2.6. Privacy and Data Ownership

For data analysis, data from all relevant side is requiring, but in some situations, where, there are strict laws governing what data can be revealed in different contexts. Another issue is that many online services today require us to share private information (think of Facebook applications), but beyond record-level access control we do not understand what it means to share data, how the shared data can be linked, and how to give users fine-grained control over this sharing in an intuitive [1], so proper preventive measures are taken to protect such sensitive data, to ensure its safety [7]. Keeping track of a particular individual's data throughout big data analytics contexts is merely an organizational requirement that can e.g. be met by means of log files. Linkage of disjoint datasets can often be performed without relying on linkage via user's identities, but based on other types of data. In the same direction, many types of data can be preprocessed with proper anonymization or pseudonymization prior to sharing, such that linkage of datasets remains feasible, but linkage to an individual's identity becomes hard [8].

## 2.7. The Human Perspective

For data analysis, there should be considerable volume of data that must be understandable for a human [1]. Understanding right data is an important issue that can be handled by using visualization as a part of data analysis [4]. Companies have difficulty identifying the right data and determining how to best use it. Building data-related business cases can be used for identifying useful data [5].

## 2.8. Visualization

Due to large and high dimension of Big Data, visualization of data is very difficult. There are some tools for visualization but mostly have poor performance and response time. New framework for visualization is highly necessary [6].

There is no communication path between data points, so companies cannot aggregate and manage the data across the enterprise. Recently [5] using smart grid such as real-time grid management has provided such solution.

## 2.9. Displaying Meaningful Results

As in Big Data, there are extremely large amounts of information or a variety of categories of information so representation of such information is a big challenge. This issue can be resolved by making data cluster into a higher-level view where smaller groups of data become visible [4].

Data structure, class and type as well as integrated technologies can be reflected through efficient data representation. This can be achieved by efficient operations on different datasets [7].

### 2.10. Dealing with Outliers

Although outlier data has small amount of percentage in normal data, but in massive amount of data, visibility of outlier is very difficult. Possible solution for such issue is to create a chart for outlier [4].

### 2.11. Finding Talent for Big Data

Company requires a talent for interpretation of data to find meaningful business insights [5].

### 2.12. IT Architecture for Big Data

Creation of right IT architecture for data is also a big challenge that should have to adopt the technology landscape; it required a lot of research to manage such architecture [5].

### 2.13. Big Data Functions

Leveraging Big Data often means working across functions like IT, engineering, finance and procurement. These issues require a path from scratch for collaboration over functions and businesses [5].

### 2.14. Security

Data protection is a major issue related to security. If it is solved from all dimensions (Data will not be displayed to others, Data cannot be changed, Data cannot be deleted) then companies could take full advantage of their data [5].

## 3. Results and Conclusion

Due to [16] three big advantage (Cost reduction, Faster, better decision making, new products and services) of Big Data, we cannot ignore the survival of Big Data in current IT-Market. These advantages can be achieved after resolving the issues related to different challenges. In following table authors maps the all Big-Data related challenges and their solutions, which are done by different researchers. In Table-1, there is a list of almost challenges of Big Data. These challenges have been defined by different authors in different ways while they have not provided the solutions.

Here we studied that with respect to solution, some of the challenges have received full attention from different researchers such as inconsistency and incompleteness, displaying meaningful results, scale, timeliness, privacy & data ownership and security while some of them have not been paid proper attention such as velocity, the human perspective, visualization and dealing with outliers. Besides, remaining challenges such as finding talent for Big Data, IT architecture for Big Data, Big Data functions, which have just given a description but not been provided a solution yet, could be the future work.

**Table 1.** Big Data Challenges and Solutions

| Challenge | Description | Solution |
|---|---|---|
| Heterogeneity | Structured Data Required | MapReduce [10] |
| Inconsistency and incompleteness | Identified Errors and missing values, efficient representation | Managed in data analysis [1]. Data cleaning, data integration, data transformation and date reduction [6] |
| Scale | Size of Bid Data | Handle with processor speed [1], handled by using Solid State Drives (SSDs) and Pulse-code modulation (PCM) technologies [6]. Grid computing approach [4], parallelized machines[9] |
| Velocity of Data | Coming Data with high speed | Using online learning approaches [9]. |
| Timeliness | Data without outliers | Using traffic management system [1]. Using proactive methods [4] |
| Privacy and data ownership | Data cannot be leakage | Using strict laws governing [1]. Using Log Files [8] |
| The human perspective | Data understandable for human | Using visualization as a part of data analysis [4]. |
| Visualization | Data Appearance | New framework is needed (existing have poor performance) [6]. Using smart grid such as real-time grid management [5] |
| Displaying meaningful results | Representation of Big Data | Using Data Cluster [4]. Using different Dataset techniques [7]. |
| Dealing with outliers | Identification of Errors | Create Chart of outliers [4] |
| Finding Talent for Big Data | How data can be interpreted | Required research for such Talent [5] |
| IT Architecture for Big Data | Architecture for Big Data | Required research for such architecture [5] |
| Big Data Functions | Working of Big Data across functions | Required research for such functions [5] |
| Security | Data Protection | Required research for such functions [5]. But can handle Using strict laws governing [1]. Using Log Files [8] |

Now we are going to analyse all challenges' solutions with respect to weight. If a challenge has no solution, we will assign 0.9. If existing solution requires more attention then we will assign 0.8. If existing solution requires less attention we will assign 0.6. If existing solution is enough then we will assign 0.4. Here instead of taking 0.0 we are taking 0.4, means in future it may need attention towards solution. For a particular challenge, if one author paid much attention while other paid little attention towards its solution, in such case we will take the average value as shown in Table-2.

**Table 2.** Solution of Challenges with respect to Weight

| Challenge | Solution | Assigning Weight | Final Weight |
|---|---|---|---|
| Heterogeneity | MapReduce [10] | There is lot of work about different dimensions of MAPREDUCE. (Less attention) | 0.6 |
| Inconsistency and incompleteness | Managed in data analysis [1]. Data cleaning, data integration, data transformation and date reduction [6] | Solution Exists but requires attention on how to clean, integrate and reduce data.(More attention) | 0.8 |
| Scale | Handle with processor speed [1], handled by using Solid State Drives (SSDs) and Pulse-code modulation (PCM) technologies [6]. Grid computing approach [4], parallelized machines[9] | SSDs available (0.4), Grid Computing approach available (0.4), parallelized machines requires less attention. Work can be done in MapReduce (0.6) | (0.4+0.4+0.6)/3= 0.45 |
| Velocity of Data | Using online learning approaches [9]. Speed is also the big challenge. | OLA available, but requires algorithm to handle speed. (0.8), also challenge (0.9) | (0.8+0.9)/2=0.85 |
| Timeliness | Using traffic management system [1]. Using proactive methods [4] | Proactive method and UTMS also requires a searching on efficient algorithm. | 0.8 |
| Privacy and data ownership | Using strict laws governing [1]. Using Log Files [8] | Log Files are working in different web based profiles. | 0.4 |
| The human perspective | Using visualization as a part of data analysis [4]. | Data analysis also needs some technique. (0.8) | 0.8 |
| Visualization | New framework is needed (existing have poor performance) [6]. Using smart grid such as real-time grid management [5] | New framework required (0.9) and some attention on smart grid (0.6) | (0.9+0.6)/2=0.75 |
| Displaying meaningful results | Using Data Cluster [4]. Using different Dataset techniques [7]. | Data clustering requires algorithm (0.6), Dataset techniques are available (0.4) | (0.6+0.4)/2=0.5 |
| Dealing with outliers | Create Chart of outliers [4] | Requires attention on how to create a chart (0.6) | 0.6 |
| Finding Talent for Big Data | Required research for such Talent [5] | Wait for attention | 0.9 |
| IT Architecture for Big Data | Required research for such architecture [5] | Wait for attention | 0.9 |
| Big Data Functions | Required research for such functions [5] | Wait for attention | 0.9 |
| Security | Required research for such functions [5]. But can handle Using strict laws governing [1]. Using Log Files [8] | Wait for attention (0.9) but can handle with laws or log files (0.4). | (0.9+0.4)/2=0.65 |

Table-3 contains all BigData challenges in sorted order (ascending) with respect to final weight.

**Table 3.** Sorted Challenges w.r.t. Weight

| Challenges | Weight w.r.t. Attention |
|---|---|
| Privacy and data ownership | 0.4 |
| Scale | 0.45 |
| Displaying meaningful results | 0.5 |
| Heterogeneity | 0.6 |
| Dealing with outliers | 0.6 |
| Security | 0.65 |
| Visualization | 0.75 |
| Inconsistency and incompleteness | 0.8 |
| Timeliness | 0.8 |

| Challenges | Weight w.r.t. Attention |
|---|---|
| The human perspective | 0.8 |
| Velocity of Data | 0.85 |
| Finding Talent for Big Data | 0.9 |
| IT Architecture for Big Data | 0.9 |
| Big Data Functions | 0.9 |

Hence Fig-1 depict that BigData functions, Architecture, Finding Talent (having maximum value) require more attention while Timeline, Security, visualization, heterogeneity etc. have some solutions but also require further attention. Privacy and scaling (having value less than 0.5) needs little attention on solution because existing solution can handle prevailing challenges.

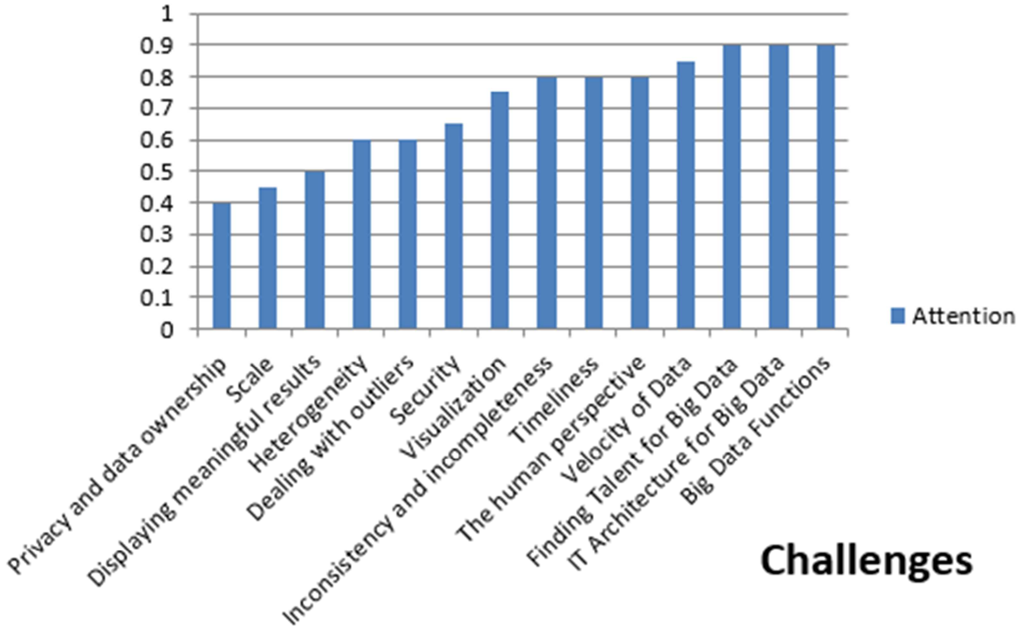

**Fig. 1.** Progress of Existing Attentions on BigData Challenges

## References

[1] Big Data and Its Technical Challenges. By H. V. Jagadish, Johannes Gehrke, Alexandros Labrinidis, Yannis Papakonstantinou, Jignesh M. Patel, Raghu Ramakrishnan, Cyrus Shahabi Communications of the ACM, Vol. 57 No. 7, Pages 86-94.

[2] Elif Dede, Zacharia Fadika, Madhusudhan Govindaraju, Lavanya Ramakrishnanb. Benchmarking MapReduce implementations under different application scenarios. Future Generation Computer Systems 36 (2014) 389–399. ELSEVIER.

[3] CHEN Kai, WAN Wen-qiang, LI Yun. Differentially private feature selection under MapReduce framework. The Journal of China Universities of Posts and Telecommunications. October 2013, 20(5): 85–90. ELSEVIER.

[4] Five big data challenges And how to overcome them with visual analytics.

[5] Six Challenges of Big Data Mar 26, 2014, The Wall Street Journal http://blogs.wsj.com/experts/2014/03/26/six-challenges-of-big-data/.

[6] Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. C.L. Philip Chen ⇑, Chun-Yang Zhang, Information Sciences 275 (2014) 314–347, ELSEVIER.

[7] Big Data: A Survey. Min Chen · Shiwen Mao · Yunhao Liu, Mobile Netw Appl (2014) 19:171–209.

[8] Meiko Jensen, Challenges of Privacy Protection in Big Data Analytics, 2013 IEEE International Congress on Big Data.

[9] XUE-WEN CHEN, Big Data Deep Learning: Challenges and Perspectives, Digital Object Identifier 10.1109/ACCESS.2014.2325029.

[10] Tharso Ferreira*, Antonio Espinosa, Juan Carlos Moure, Porfidio Hern´andez, An Optimization for MapReduce Frameworks in Multi-core Architectures, International Conference on Computational Science, ICCS 2013. ELSEVIER.

[11] Big Data, http://en.wikipedia.org/wiki/Big_data.

[12] Sergey V. Kovalchuk, Artem V. Zakharchuk, Jiaqi Liao1, Sergey V. Ivanov1, Alexander V. Boukhanovsky. A Technology for BigData Analysis Task Description using Domain-Specific Languages. Volume 29, 2014, Pages 488–498 ICCS 2014. 14th International Conference on Computational Science.

[13] Tatiana Gavrilova, Margarita Gladkova. Big Data Structuring: The Role of Visual Models and Ontologies. Information Technology and Quantitative Management (ITQM 2014). Procedia Computer Science 31 (2014) 336 – 343.

[14] Andrew Rau-Chaplin, Zhimin Yao, and Norbert Zeh .Efficient Data Structures for Risk Modelling in Portfolios of Catastrophic Risk Using MapReduce. Volume 29, 2014, Pages 1557–1568 ICCS 2014. 14th International Conference on Computational Science.

[15] Thomas H. Davenport Jill Dyché. Big Data in Big Companies. 2013.

[16] Tom Davenport, IIA Director of Research and faculty leader. Three big benefits of big data analytics. 2014. https://www.sas.com/en_za/news/sascom/2014q3/Big-data-davenport.html.

[17] Diya Soubra on July 5, 2012 at 5:11am. The 3Vs that define Big Data. http://www.datasciencecentral.com/forum/topics/the-3vs-that-define-big-data.