www.ijecs.in

*International Journal Of Engineering And Computer Science ISSN:2319-7242*
*Volume – 5 Issue -03 March, 2016 Page No. 15947-15961*

# A review: Issues and Challenges in Big Data from Analytic and Storage perspectives

*[1]Zaharaddeen Karami Lawal, [2]Rufai Yusuf Zakari, [3]Mansur Zakariyya Shuaibu, [4]Alhassan Bala*

(deenklawal13@gmail.com)

(rufaig6@gmail.com)

(mansurzshuaibu@gmail.com)

(alumgarko200@gmail.com)

[1,2,3,4]Dept. CSE, Jodhpur National University, Jodhpur, India

## ABSTRACT

There is an up-and-coming topic in the field of computer science and technology that is getting a lot of publications these days and that is Big Data. The term **Big data** is referring for a collection of large and complex data sets which is very hard to process by database management tools or data processing applications. The volumes are in the range of Exabyte and above. Most of companies and government agencies are dealing with it in this current speedy moving technology environment. Stand -alone applications as well as today's newer web-based processes are generating large amount of data. This raise some issues that has to be considered when dealing with big data. Some of the issues are storage, management and processing. This data also creates new challenges which make the framework based developers to come up with solutions to the problems. They introduced different frameworks for the storage, management and processing of big data that are scalable and portable with fault tolerance capabilities. Newer technologies for the analysis of such large amount data that is complex to handle are also introduced. These challenges include Privacy and Security, Scale, and Heterogeneity to mention a few. This paper reviews the Issues and Challenges of Big Data from Data Analytic and Storage Perspectives.

We also briefly discussed about the frameworks and databases that can be used to tackle the challenges that is facing Big Data world.

Keywords: Big Data, Analytic, Fault tolerant, Framework, Heterogeneity, Privacy, Stand-alone

## 1.0 INTRODUCTION

The term 'Big Data' appeared for first time in 1998 in a Silicon Graphics (SGI) slide deck by John Mashey with the title of "Big Data and the Next Wave of Infra Stress". It is the term for data sets so large and complicated that it becomes difficult to process using traditional data management tools or processing applications [1].

The concept of big data has been endemic within computer science since the earliest days of computing. "Big Data" originally meant the volume of data that could not be processed (efficiently) by traditional database methods and tools. Each time a new storage medium was invented, the amount of data accessible exploded because it could be easily accessed. Even if we have the storage capacity to store our data, but the available data analytics tool can't handle a data it's still a Big Data. This has lead to a growing interest in the development of tools capable in the automatic extraction of knowledge from data [2]. Data are collected and analyzed to create information suitable for making decisions. Hence data provide a rich resource for knowledge discovery and decision support. A database is an organized collection of data so that it can easily be accessed, managed, and updated [3].

"We define "Big Data" as the amount of data just beyond technology's capability to store, manage and process efficiently. These imitations are only discovered by a robust analysis of the data itself, explicit processing needs, and the capabilities of the tools (hardware, software, and methods) used to analyze it" [4].

Enormous amount of data are generated every minute. A recent study estimated that every minute, Google receives over 4 million queries, e-mail users send over 200 million messages, YouTube users upload 72 hours of video, Facebook users share over 2 million pieces of content, and Twitter users generate 277,000 tweets [5] [6]. With the amount of data growing exponentially, improved analysis is required to extract information that best matches user interests. Big data refers to rapidly growing datasets with sizes beyond the capability of traditional data base tools to store, manage and analyse them. Big data is a heterogeneous collection of both structured and unstructured data. Increase of storage capacities, Increase of processing power and availability of data are the main reason for the appearance and growth of big data. Big data refers to the use of large data sets to handle the collection or reporting of data that serves businesses or other recipients in decision making [5].

**Figure 1 Big Data**

### 1.1 Types of Data

There are two types of data. These are:

1. **Structured Data**

Structured Data are numbers and words that can be easily categorized and analyzed. These data are generated by things like network sensors embedded in electronic devices, smart phones, and global positioning system (GPS) devices. Structured data also include things like sales figures, account balances, and transaction data [1].

**2. Unstructured Data**

Unstructured Data include more complex information, such as customer reviews from commercial websites, photos and other multimedia, and comments on social networking sites. These data cannot easily be separated into categories or analyzed numerically. The explosive growth of the Internet in recent years means that the variety and amount of big data continue to grow. Much of that growth comes from unstructured data [1]. These types of data include: Log Data, Images, Audios, Videos, etc.

DOI: 10.18535/ijecs/v5i3.12

### 1.2 Characteristics of Big Data

One view, espoused by Gartner's Doug Laney describes Big Data as having three dimensions: volume, variety, and velocity. Thus, IDC defined it: "*Big data technologies describe a new generation of technologies and architectures designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis.*" [4] [7]
Although IBM defines Big Data with four characteristics by adding veracity to the above 3 characteristics. Other Big Data Researchers defined it with 5 Vs by adding Value.
In this paper we will define Big Data based on 5 Vs.

1. **Volume**
   The first characteristic of Big Data, which is "Volume", refers to the quantity of data that is being manipulated and analyzed in order to obtain the desired results. It represents a challenge because in order to manipulate and analyze a big volume of data requires a lot of resources that will eventually materialize in displaying the requested results [8]. For instance, a computer system is limited by current technology regarding the speed of processing operations. The size of the data that is being processed can be unlimited, but the speed of processing operations is constant.
2. **Velocity**
   Data velocity measures the speed of data creation, streaming, and aggregation. Ecommerce has rapidly increased the speed and richness of data used for different business transactions (for example, web-site clicks). Data Variety: Data variety is a measure of the richness of the data representation – text, images video, audio, etc [1].
3. **Variety**
   Data variety is a measure of the richness of the data representation – text, images video, audio, etc. From an analytic perspective, it is probably the biggest obstacle to effectively using large volumes of data. Incompatible data formats, non-aligned data structures, and inconsistent data semantics represents significant challenges that can lead to analytic sprawl [4].
4. **Veracity**
   The quality of the data being captured can vary greatly. Accuracy of analysis depends on the veracity of the source data.
5. **Value**
   Data value measures the usefulness of data in making decisions. It has been noted that "the purpose of computing is insight, not numbers". Data science is exploratory and useful in getting to know the data, but "analytic science" encompasses the predictive power of big data [4].

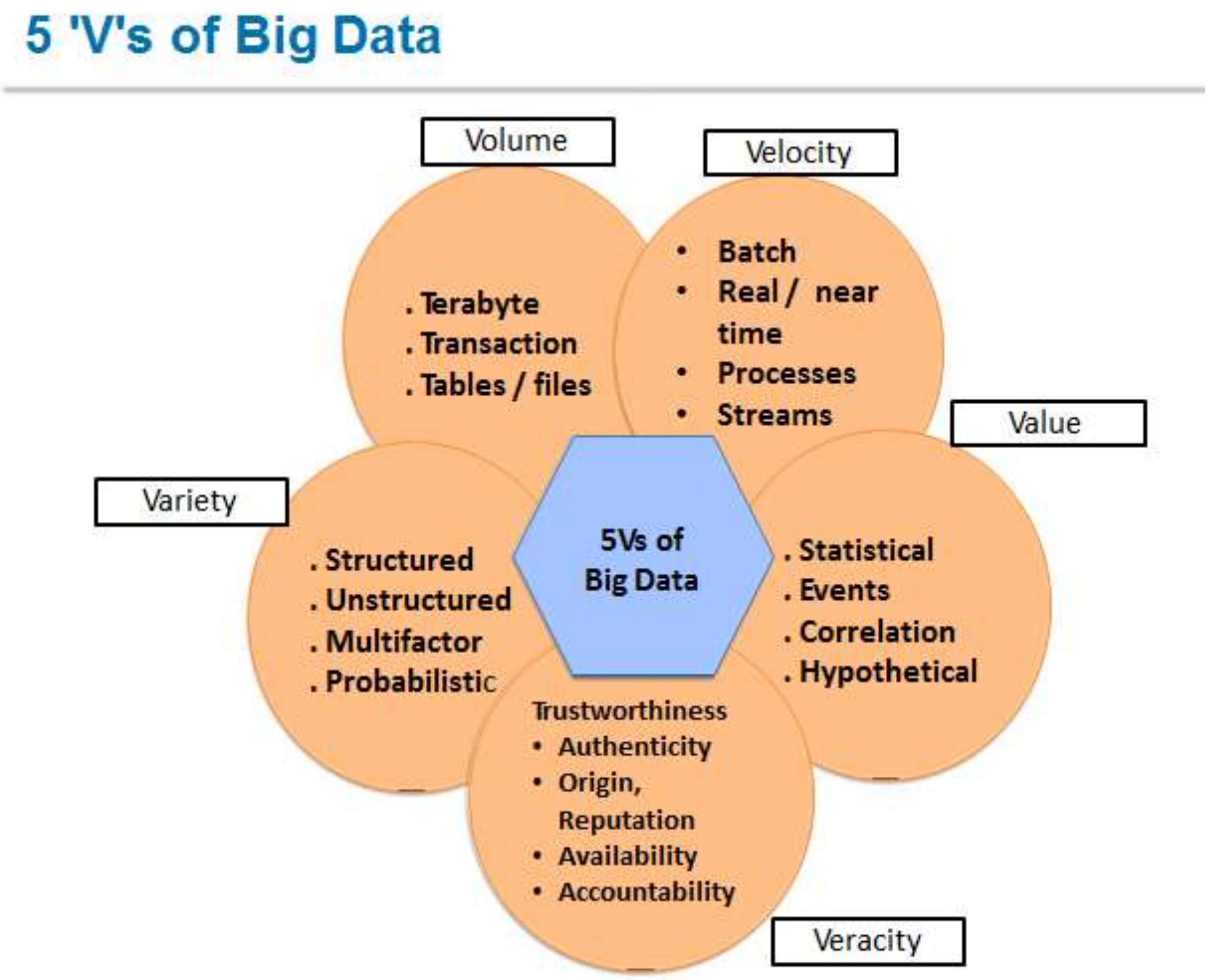

**Figure 2 5Vs of Big Data**

### 1.3 Big Data Architecture

Big data can be stored, acquired, processed, and analyzed in many ways. Every big data source has different characteristics, including the frequency, volume, velocity, type, and veracity of the data. When big data is processed and stored, additional dimensions come into play, such as governance, security, and policies. Choosing an architecture and building an appropriate big data solution is challenging because so many factors have to be considered.

This "Big data architecture and patterns" series presents a structured and pattern-based approach to simplify the task of defining an overall big data architecture. Because it is important to assess whether a business scenario is a big data problem, we include pointers to help determine which business problems are good candidates for big data solutions.

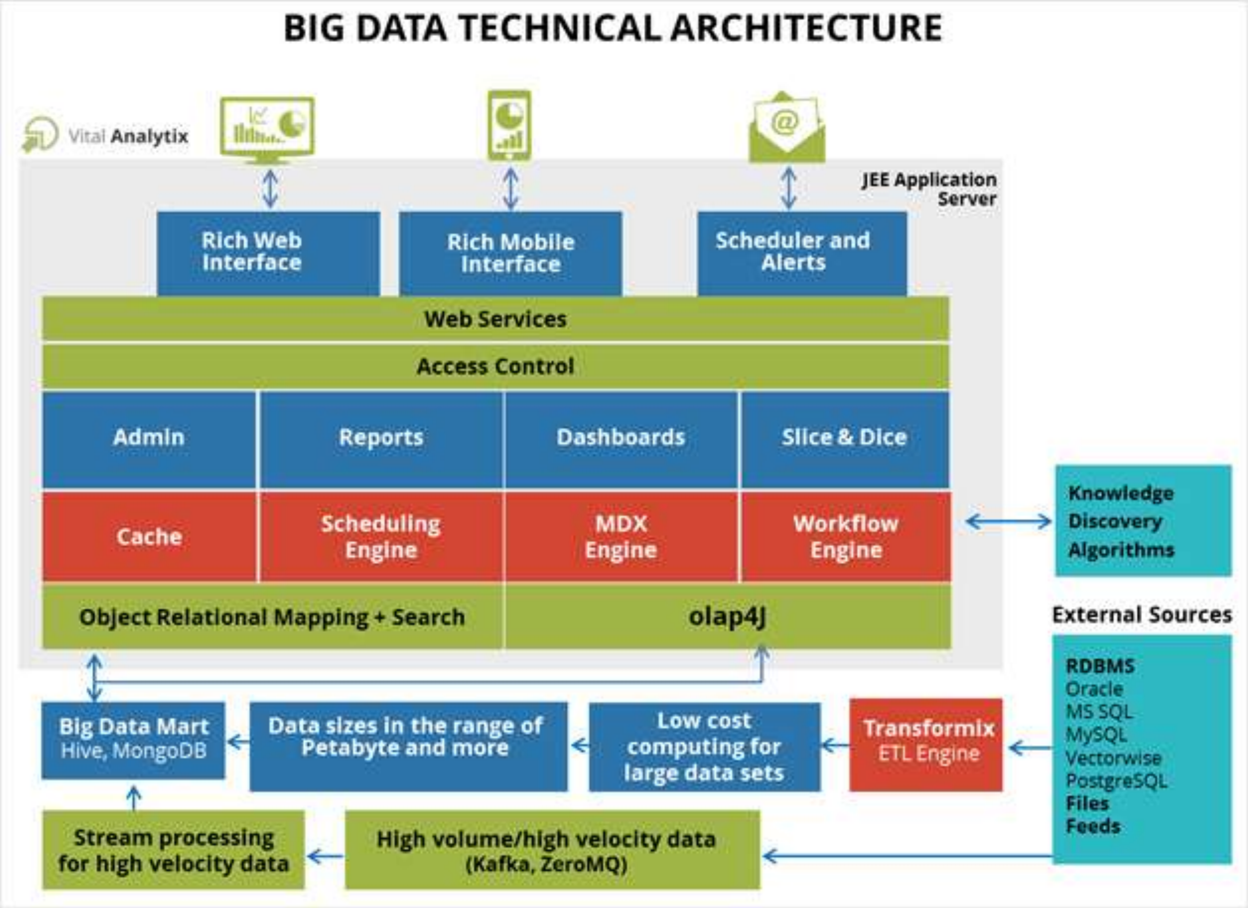

**Figure 3 Big Data Architecture**

### 1.4 Big Data Opportunities

Big data creates tremendous opportunity for the world economy both in the field of national security and also in areas ranging from marketing and credit risk analysis to medical research and urban planning. The extraordinary benefits of big data are lessened by concerns over privacy and data protection.

Big data is obviously giving us with thrilling opportunities in data mining research.

- **Manageability -** when data can grow in a single file system namespace the manageability of the system increases significantly and a single data administrator can now manage a petabyte or more of storage versus 50 or 100 terabytes on a scale up system [9].
- **Eradication of stovepipes -** since these systems scale linearly and do not have the bottlenecks that scale up systems create, all data is kept in a single file system in a single grid eliminating the stovepipes introduced by the multiple arrays and files systems required [9].
- **Just in time scalability -** as my storage needs grow I can add an appropriate number of nodes to meet my needs at the time I need them [9]. With scale up arrays I would have to guess at the final size my data may grow while using that array which often led to the purchase of large data servers with only a few disks behind them initially so I would not hit bottleneck in the data server as I added disks [9].
- **Increased utilization rates -** since the data servers in these scales out systems can address the entire pool of storage there is no stranded capacity [10].

### 2.0 Issues in Big Data

There are many issues that needs to be address in Big Data. Such issues they are topic of research on their own, because each one can be taken and treated as a single issue entity. The Big Data Issues that needs to be address are:-

1. **Storage and Transport Issues**
   The quantity of data has exploded each time we have invented a new storage medium. What is different about the most recent explosion – due largely to social media – is that there has been no

new storage medium. Moreover, data is being created by everyone and everything (e.g., devices, etc) – not just, as heretofore, by professionals such as scientist, journalists, writers, etc [4].

Current disk technology limits are about 4 terabytes per disk. So, 1 exabyte would require 25,000 disks. Even if an exabyte of data could be processed on a single computer system, it would be unable to directly attach the requisite number of disks. Access to that data would overwhelm current communication networks. Assuming that a 1 gigabyte per second network has an effective sustainable transfer rate of 80%, the sustainable bandwidth is about 100 megabytes. Thus, transferring an exabyte would take about 2800 hours, if we assume that a sustained transfer could be maintained. It would take longer to transmit the data from a collection or storage point to a processing point than it would to actually process it! Two solutions manifest themselves. First, process the data "in place" and transmit only the resulting information. In other words, "bring the code to the data", vs. the traditional method of "bring the data to the code." Second, perform triage on the data and transmit only that data which is critical to downstream analysis. In either case, integrity and provenance metadata should be transmitted along with the actual data [4].

2. **Security and Privacy Issues**

As big data expands the sources of data it can use, the trust worthiness of each data source needs to be verified and techniques should be explored in order to identify maliciously inserted data.

Information security is becoming a big data analytics problem where massive amount of data will be correlated, analyzed and mined for meaningful patterns [3].

Any security control used for big data must meet the following requirements:

- It must not compromise the basic functionality of the cluster.
- It should scale in the same manner as the cluster.
- It should not compromise essential big data characteristics.
- It should address a security threat to big data environments or data stored within the cluster.

Unauthorized release of information, unauthorized modification of information and denial of resources are the three categories of security violation. The following are some of the security threats:

An unauthorized user may access files and could execute arbitrary code or carry out further attacks.

- An unauthorized user may eavesdrop/sniff to data packets being sent to client.
- An unauthorized client may read/write a data block of a file.
- An unauthorized client may gain access privileges and may submit a job to a queue or delete or change priority of the job [3].

3. **Management Issues**

Management will, perhaps, be the most difficult problem to address with big data. This problem first surfaced a decade ago in the UK eScience initiatives where data was distributed geographically and "owned" and "managed" by multiple entities. Resolving issues of access, metadata, utilization, updating, governance, and reference (in publications) have proven to be major stumbling blocks [4]. Unlike the collection of data by manual methods, where rigorous protocols are often followed in order to ensure accuracy and validity, digital data collection is much more relaxed. The richness of digital data representation prohibits a bespoke methodology for data collection. Data qualification often focuses more on missing data or outliers than trying to validate every item. Data is often very fine-grained such as clickstream or metering data. Given the volume, it is impractical to validate every data item: new approaches to data qualification and validation are needed.

The sources of this data are differs – both temporally and spatially, by format, and by method of collection. Individuals contribute digital data in mediums comfortable to them: documents, drawings, pictures, sound and video recordings, models, software behaviors, user interface designs, etc – with or without adequate metadata describing what, when, where, who, why and how it was collected and its provenance. Yet, all this data is readily available for inspection and analysis.

Going forward, data and information provenance will become a critical issue. JASON has noted that [11] "there is no universally accepted way to store raw data, … reduced data, and … the code and parameter choices that produced the data." Further, they note: "We are unaware of any robust, open source, platform independent solution to this problem." "As far as we know, this remains true today. To summarize, there is no perfect big data management solution yet. This represents an important gap in the research literature on big data that needs to be filled" [4].

4. **Processing Issues**

   Assume that an exabyte of data needs to be processed in its entirety. For simplicity, assume the data is chunked into blocks of 8 words, so 1 exabyte = 1K petabytes. Assuming a processor expends 100 instructions on one block at 5 gigahertz, the time required for end-to-end processing would be 20 nanoseconds. To process 1K petabytes would require a total end-to-end processing time of roughly 635 years. Thus, effective processing of exabytes of data will require extensive parallel processing and new analytics algorithms in order to provide timely and actionable information [4].

### 2.1 Big Data challenges

1. **Network level**

   The challenges that can be categorized under a network level deal with network protocols and network security, such as distributed nodes, distributed data, Internode communication [1].

2. **Scale and complexity**

   The size of big data is easily recognized as an obvious challenge. Big data is pushing scalability in storage, with increases in data density on disks to match. The current Redundant Array of Independent Disks (RAID) approach that is in widespread use does not provide the level of performance and data durability that enterprises dealing with escalating volumes of data require. For example, committing data from memory to disk can increase overhead and cause processing delays if multiple disks are involved in each commit process. Moreover, as the scale of data increases, the mean time between failures (MTBF) falls. For example, a system with a billion cores has an MTBF of one hour. The failure of a particular cluster node affects the overall calculation work of the large infrastructure that is required to process big data transactions.

   Of course, managing large and rapidly increasing volumes of data has been a challenging issue for many decades. In the past, this challenge was mitigated by processors getting faster, following Moore's law, to provide us with the resources needed to cope with increasing volumes of data [12],But there is a fundamental shift underway now: data volume is scaling faster than compute resources, and CPU speeds are static. Traditional software tools are not enough for managing the increasing volumes of data. Data analysis, Traditional software tools are not enough for managing the increasing volumes of data. Data analysis, organization, retrieval and modeling are also challenges due to scalability and complexity of data that needs to be analysed [3].

1. **Timeliness**

   As the size of the data sets to be processed increases, it will take more time to analyse. In some situations, results of the analysis is required immediately. For instance, if a fraudulent credit card transaction is suspected, it should ideally be flagged before the transaction is completed by preventing the transaction from taking place at all. Obviously a full analysis of a user's purchase history is not likely to be feasible in real time. So we need to develop partial results in advance so that a small amount of incremental computation with new data can be used to arrive at a quick determination [3]. The larger the data set to be processed, the longer it will take to analyze. The design of a system that effectively deals with size is likely also to result in a system that can process a given size of data set faster. However, it is not just this speed that is usually meant when one speaks of Velocity in the context of Big Data [12].

2. **Heterogeneity and Incompleteness**

   When humans consume information, a great deal of heterogeneity is comfortably tolerated. In fact, the nuance and richness of natural language can provide valuable depth. However, machine analysis algorithms expect homogeneous data, and cannot understand nuance. In consequence, data must be carefully structured as a first step in (or prior to) data analysis. Consider, for example, a patient who has multiple medical procedures at a hospital. We could create one record per medical procedure or laboratory test, one record for the entire hospital stay, or one record for all lifetime hospital interactions of this patient. With anything other than the first design, the number of medical procedures and lab tests per record would be different for each patient. The three design choices listed have successively less structure and, conversely, successively greater variety. Greater structure is likely to be required by many (traditional) data analysis systems. However, the less structured design is likely to be more effective for many purposes – for example questions relating to disease progression over time will require an expensive join operation with the first two designs, but can be avoided with the latter. However, computer systems work most efficiently if they can store multiple items that are all identical in size and structure. Efficient representation, access, and analysis of semi-structured data require further work [12].

3. **Analytics Architecture**
   It is not clear yet how an optimal architecture of an analytics system should be to deal with historic data and with real-time data at the same time. An interesting proposal is the Lambda architecture of Nathan Marz. The Lambda Architecture solves the problem of computing arbitrary functions on arbitrary data in real time by decomposing the problem into three layers: the batch layer, the serving layer, and the speed layer. It combines in the same system Hadoop for the batch layer, and Storm for the speed layer [1].
4. **Data privacy. Data security.**
   This has many implications and it concerns individuals and companies as well. Individuals have the right, according to International Telecommunications Union, to control the information that may be disclosed regarding them. Information posted by users on their online profiles is likely to be used in creating a “users profile” so that can be further used by companies to develop their marketing strategies and to extend their services. Individual’s privacy is still a delicate problem that can only be solved with drastic solutions. Allowing persons to choose whether they post or not information about them is a more secure way to achieve privacy, but will also cause software to “malfunction”. For example in a social network, if a person is allowed to choose whether he/she wants to complete the fields regarding personal information and, in the same time, allow them to choose if the social network can store information about their IP address, location etc., this could be a possible threat to everyone else that is using the same social network [8]For companies the privacy issue is more related to the sensitive data that they work with. Whether is financial data, clients list, perspective projects, all represents valuable data that may or may not be disclosed. Companies have multiple choices regarding where to store their information. They can either store it on cloud systems, “in-house” systems or a hybrid solution. By storing data on cloud systems is more convenient for companies in terms of cost. Also, a cloud system is not only characterized by storage space, but as well for the speed of processing requested operations. The data security still remains a contentious issue in this case.

5. **Distributed Mining**
   Many data mining techniques are not trivial to paralyze. To have distributed versions of some methods, a lot of research is needed with practical and theoretical analysis to provide new methods [1].
6. **Sharing data**
   Sharing the information proves to be one of the most valuable characteristics of development. Information about almost anything can be found by simply doing a “Google search”. Every person and company has at their disposal large amount of information that can use it to serve their purposes. Everything is available only if everyone shares it. Regarding persons, there is a difference between what is personal and what can be made public. The issue of what is personal and what is public mostly resides in the point of view of the services that they use [13].
7. **Right Time, Real Time, Online Analytics**
   The traditional data miner, in practice, has generally been involved in batch-oriented model building, using machine learning and statistical algorithms. From our massive store of historical data we use

algorithms such as logistic regression, decision tree induction, random forests, neural networks, and support vector machines. Once we have built our model(s) we then proceed to deploy the models. In the business context we migrate our models into production [14]. The models then run on new transactions as they arrive, perhaps scoring each transaction and then deciding on a treatment for that transaction—that is, based on the score, how should the transaction be processed by our systems?

The process is typical of how data mining is delivered in many large organizations today, including government, financial institutions, insurance companies, health providers, marketing, and so on. In the Australian Taxation Office, for example, every day a suite of data mining models risk score every transaction (tax return) received. The Australian Immigration Department [15], as another example, has developed a risk model that assesses all passengers when they check-in for their international flight to Australia (known as Advance Passenger Processing) using data mining models. Such examples abound through industry and government.

Today’s agile context now requires more than this. The larger and older organizations, world wide, have tended to be much less agile in their ability to respond to the rapid changes delivered through the Internet and our data rich world. Organizations no longer have the luxury of spending a few months building models for scoring transactions in batch mode. We need to be able to assess each transaction as the transaction happens, and to dynamically learn as the model interacts with the massive volumes of transactions that we are faced with as they occur. We need to build models in real time to respond in real time and that learn and change their behavior in real-time. Research in incremental learning is certainly not new. Incremental learning [14] [16], [17], just-intime or any-time learning [18], and data stream mining [19] have all addressed similar issues in different ways over the past decades. There is now increasing opportunity to capitalize on the approach. The question continues as to how we can maintain and improve our knowledge store over time, and work to forget old, possibly incorrect, knowledge?

The development of dynamic, agile learners working in real-time—that is, as they interact with the real world—remain quite a challenge and will remains a central challenge for our big data world [14].

8. **Failure handling**

Devising 100% reliable systems on the go is not an easy task. Systems can be devised in such a way that the probability of failure must fall within the permitted threshold. Fault tolerance is a technical challenge in big data. When a process started it may involve with numerous network nodes and the whole computation process becomes cumbersome. Retaining check points and fixing the threshold level for process restart in case of failure, are greater concerns [20].

9. **Data quality**

Huge amount of data pertaining to a problem is undoubtedly a big asset for both Business as well as IT leaders [21]. For predictive analysis or for better decision making amount of relevant data helps a lot. But the quality of such data is based on the source through which they are derived. Though big data stores large relevant data, the accuracy of data is completely dependent on the source domains. Hence, there is a question of how far the data can be trusted and it definitely requires appropriate trust agent filters [20].

10. **Inconsistencies**

The Big data research area is embedded with multi-dimensional technical and scientific spaces. The objectives of big data analytics differ with the stakeholders. As this big data analysis is the next frontier for innovation and advancement of technology, one should not underestimate the impact of it on the society. Soon the big data cloud is going to cover all domains and sectors like manufacturing, special data science, life and physical sciences, communication, finance and banking etc., As big data comprises all domains, definitely inconsistencies arise. Inconsistencies in data level, information level and knowledge level have to be addressed [20]. There are four types of inconsistencies like temporal, textual, spatial and functional dependency. These inconsistencies are well addressed by Zhang in [20] [22].

DOI: 10.18535/ijecs/v5i3.12

## 11. Extreme Data Distribution—Privacy and Ownership

Having considered two intermediate challenges around big data, we now consider a game-changing challenge. The future holds for us the prospect of individuals regaining control of their data from the hands of the now common centralized massive stores. "We expect to see this as an orderly evolution of our understanding of what is best for society as we progress and govern our civil society in this age of big data collection and surveillance and of its consequent serious risk to privacy. Data ownership has become a challenging issue in our data rich world" [14]. Data collectors in the corporate and government spheres are learning to efficiently collect increasingly larger holdings of big data. However, with the help of civil libertarians, philosophers and whistle blowers, society is gradually realizing the need for better governance over the collection and use of data. Recent events like Wikileaks [23] and Edward Snowden [24] help to raise the level of discussion that is providing insight into the dangers of aggregating data centrally—with little regard about who owns the data.

Data ownership presents a critical and ongoing challenge, particularly in the social media arena. While petabytes of social media data reside on the servers of Facebook, MySpace, and Twitter, it is not really owned by them (although they may contend so because of residency) [4]. Certainly, the "owners" of the pages or accounts believe they own the data. This dichotomy will have to be resolved in court. Kaisler, Money and Cohen [25] addressed this issue with respect to cloud computing as well as other legal aspects that we will not delve into here.

## 12. Infrastructure faults

Storing and analyzing large volumes of data that is crucial for a company to work requires a vast and complex hardware infrastructure. If more and complex data is stored, more hardware systems will be needed. A hardware system can only be reliable over a certain period of time. Intensive use and, rarely, production faults will most certainly result in a system malfunction. Companies can't afford to lose data that they gathered in the past years, neither to lose their clients. For avoiding such catastrophic events they use a backup system that does the simple operation of storing all data. By doing this, companies obtain continuity, even if they are drawn back temporary. The challenge is to maintain the level of services that they provide when, for example, a server malfunction occurs right when a client is uploading files on it. To achieve continuity, hardware systems are backed by software solutions that respond in order to maintain fluency by redirecting traffic to another system. When a fault occurs, usually a user is not affected and he/she continues work without even noticing that something has happened [8].

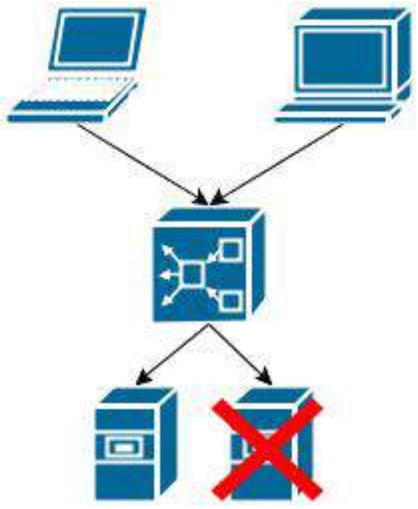

**Figure 4 System Failure**

The flow of data must not be interrupted in order to obtain accurate information. For example, Google is sending one search request to multiple servers, rather than sending it to only one. By doing this, the response time is shortened and also there is no inconsistency in the data that users sends – receives.

System failure affects the process of storing data and is making more difficult to work with. There can be created a permanent connection between the device, that is sending data. Hence this problem can be solve by creating a loop, the “sender” will make sure that the “receiver” has no gaps regarding the data that should be stored. This loop should work as long as the system that is receiving data tells the system that sends it to stop because the data that is stored is identical to the one sent. So, is a simple comparison process that can prevent loosing data.

This process can also slow down the whole process. To avoid this from happening, for any content that is transmitted, the sender must generate a “key”. This key is then transferred to the receiver to compare it with the key that it generated regarding the data that was received. If both keys are identical than the “send-receive” process was successfully completed. For better understanding, this solution is similar with the MD5 Hash that is generated over a compressed content. But, in this case, the keys are compared automatically [8].

## 2.2 Other Issues

There are some issues in big data that are not strange but they are not getting attention which are not supposed to get from the Data Analytics point of view. Some of these issues are:

- **Data trust**: While data is rapidly and increasingly available, it is also important to consider the data source and if the data can be trusted. More data is not necessarily correct data, and more data is not necessarily valuable data [14]. A keen filter for the data is a key.
- **Distributed existence**: Owners of different parts of the data might warrant different access rights. We must aim to leverage data sources without access to the whole data, and exploit them without transporting the data. We will need to pay attention to the fact that different sources may come with different label quality, there may be serious noise in the data due to crowdsourcing, and the *i.i.d.* assumption may not hold across the sources [14].
- **Diverse demands**: People may have diverse demands whereas the high cost of big data processing may disable construction of a separate model for each demand. Can we build one model to satisfy the various demands? We also need to note that, with big data, it is possible to find supporting evidence to any argument we want; then, how to judge/evaluate our “findings”? [14].
- **Rapid model**: As the world continues to “speed up”, decisions need to be made more quickly because fraudsters can more quickly find new methods in an agile environment, model building must become more agile and real-time [14].

## 3.0 Tackling Big Data Challenges

The volume of and variety/heterogeneity of data with the velocity/speed its generated, makes difficult for the current computing infrastructure to manage Big Data. Unlike traditional data management, Big Data uses distributed file system(DFS) architecture to store data.

- **Apache Hadoop** is used which is scalable, flexible and fault tolerant framework to solve the problem of storage, distributed processing and analysis on top of the entire population of very large data sets on computer clusters built from commodity hardware. which traditional data mining and warehousing fails to address by using Hadoop Distributed File System **(HDFS)** for data storage and MapReduce for analysis. It’s an Open-source software framework written in Java for distributed storage but its inherently a batch-oriented system which has limitations with real time operations. Large scale data processing is a difficult task. Managing hundreds or thousands of processors and managing parallelization and distributed environments makes it more difficult. Map Reduce provides solution to the mentioned issues since it supports distributed and parallel I/O scheduling. It is fault tolerant and supports scalability and it has inbuilt processes for status and monitoring of heterogeneous and large datasets as in Big Data [26].

- **Apache Spark** was also introduced to handle the problem of real time operations which Hadoop was unable to support. It's an Open source cluster computing framework for fast large scale data processing. Advanced Directed Acyclic Graph (DAG) execution engine that supports cyclic data flow and in-memory computing, allowing programs to run up to 100 times faster than Hadoop MapReduce. Arranges data into 'Resilient Distributed Data sets' which can be recovered following failure. Does not have its own distributed storage system. But it has some drawbacks. Firstly, it requires higher amount of RAM, hence expensive. If data size exceeds memory, processing becomes slower than batch processing MapReduce.
- **Presto**
  Presto is an efficient Big Data system developed by data engineers at the popular social networking site, Facebook. An open source distributed SQL query engine for running interactive analytical queries against data sources of all sizes ranging from gigabytes to petabytes.
- **NoSQL Databases**
  All of the above frameworks uses NoSQL as their databases. An important aspect of NoSQL databases is that they have no predefined schema, records can have different fields as necessary, this may be referred to as a dynamic schema. Many NoSQL databases also support replication which is the option of having replicas of a server, this provides reliability as in the case one goes offline the replica would become the primary server [27].

## Conclusion

We have tried to discover the issues and challenges that Big Data is facing from data storage and analytics perspectives. Some of the challenges that we have mentioned can easily be overcome. These practical challenges are common across a large variety of application domains, and consequently not cost-effective to address in the context of one domain alone. Therefore, there is need to support and encourage fundamental research towards addressing these technical challenges if we are to attain the assured benefits of Big Data.

## References

[1] K. S. Dr. Jangala., M. Sravanthi, K. Preethi and M. Anusha, "Recent Issues and Challenges on Big Data in Cloud computing," *IJCST , ,* vol. Vol. 6, no. Issue 2, April - June 2015.

[2] J. D. M. and K. Balakrishnan, "Prediction of Key Symptoms of Learning Disabilities in School-Age Children using Rough Sets," *Int. J. of Computer and Electrical Engineering, Hong,* vol. 3(1), pp. pp163-169, 2011.

[3] K. U. Jaseena and J. M. David, "ISSUES, CHALLENGES, AND SOLUTIONS: BIG DATA MINING," *Computer Science & Information Technology (CS & IT),* no. pp. 131–140, 2014.

[4] S. Kaisler, F. Armour and J. A. Espinosa, "Big Data: Issues and Challenges Moving Forward," *Hawaii International Conference on System Sciences,* no. 46th, 2013.

[5] J. . K. U. and J. . M. David, "ISSUES, CHALLENGES, AND SOLUTIONS: BIG DATA MINING," *Computer Science & Information Technology (CS & IT).*

[6] "http://www.domo.com/blog/2014/04/data-never-sleeps-2-0/," [Online].

[7] G. J. and . E. Reinsel, "“Extracting Value from Chaos”, IDC's Digital Universe Study, sponsored by EMC," 2011.

[8] A. A. TOLE, "Big Data Challenges," *Database Systems Journal,* vol. vol. IV, no. no. 3, pp. 31-40, 2013.

[9] K. Upadhyay, "A Review: Opportunities and Challenges for Big Data Storage.," *International Conference on Cloud, Big Data and Trust ,* vol. RGPV, pp. 13-15, 2013, Nov.

[10] . J. C. M. Strategist,, " "Big Data and the New Storage," *National Storage OnX Enterprise Computing OnX Enterprise Solution Design.Build. Manage. Accelerate,,* vol. A WHITE PAPER , pp. pp.4, 6, November 2011.

[11] JASON, ""Data Analysis Challenges"," *Mitre Corporation, McLean, VA, JSR-08-142,* 2008.

[12] "Challenges and Opportunities with Big Data," *A community white paper developed by leading researchers across the United States.*

[13] "Frontiers in Massive Data Analysis.," National Research Council, Washington D.C: The National AcademiesPress, 2013.

[14] Z.-H. Zhou, N. V. Chawla, Y. Jin and G. J. Williams, "Big Data Opportunities and Challenges: Discussions from Data Analytics Perspectives," *IEEE COMPUTATIONAL INTELLIGENCE MAGAZINE .*

[15] "Australian Department of Immigration, "Fact sheet 70 - managing the border," Internet, 2013. [Online].," [Online]. Available: Available: http://www.immi.gov.au/media/factsheets/70border.htm.

[16] Q. Da, Y. Yu, Z. -H. and Zhou, ""Learning with augmented class by exploiting unlabeled data,"," in *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, Quebec City, Canada., 2014.

[17] J. J. Schlimmer and . D. Fisher, ""A case study of incremental of incremental concept induction,"," in *5th National Conference on Artificial Intelligence, pp. 496–501.*, Philadelphia, PA, 1986.

[18] "P. Smyth, "Recent advances in knowledge discovery and data mining (invited talk),"," in *14th National Conference on Artificial Intelligence,Providence*, RI, USA,, 1997.

[19] M. M. Gaber, A. Zaslavsky and S. Krishnaswamy, ""Mining data streams: A review,"," *ACM SIGMOD Record.,* Vols. vol. 34, no. 2, p. pp. 18–26, 2005.

[20] S. J. Samuel, K. RVP, K. Sashidhar and C. R. Bharathi, "A SURVEY ON BIG DATA AND ITS RESEARCH CHALLENGES," *ARPN Journal of Engineering and Applied Sciences,* Vols. VOL. 10, NO. 8, MAY 2015.

[21] A. Katal, M. Wazid and R. H. Goudar, "Big data: Issues, challenges, tools and Good practices," in *Proceedings of Sixth International Conference on Contemporary Computing (IC3), (Page: 404-409 ),* 2013.

[22] D. Zhang, "Inconsistencies in big data Informatics," in *Proceedings of the 12th IEEE International Conference on Cognitive(Page: 61-67)*, New York City, NY, 2013.

[23] "Wikileaks," [Online]. Available: http://en.wikipedia.org/wiki/Wikileaks.

[24] "Wikipedia," [Online]. Available: http://en.wikipedia.org/wiki/Edward Snowden.

[25] K. S., W. Money and S. . J. Cohen., """A Decision Framework for Cloud Computing"," in *45th Hawaii*

*International Conference on System Sciences,*, Grand Wailea, Maui, HI,, Jan 4-7, 2012.

[26] P. S. Duggal and S. Paul, "Big Data Analysis: Challenges and Solutions," in *International Conference on Cloud, Big Data and Trust 2013,* , RGPV, Nov 13-15, 2013.

[27] "MongoDB," MongoDB Manual, [Online]. Available: http://docs.mongodb.org/manual/. Accessed.

[28] "Pattern-Based Strategy: Getting Value from Big Data.," *Gartner Group press release. . Available at http://www.gartner.com/it/page.jsp?id=1731916,* July 2011.

[29] D. S. Kiran, M. Sravanthi, K. Preethi and M. Anusha, ""Recent Issues and Challenges on Big Data in Cloud Computing,"," *International Journal of Computer Science And Technology,IJCST,* vol. Vol. 6, April - June 2015..

[30] "“Fact sheet 70 - managing the border,” Internet, 2013. [Online].Available: http://www.immi.gov.au/media/factsheets/70border.htm," *Australian Department of Immigration.*

[31] A. Hazra, C. Holban and Y. e. Li, "Performance and Capacity Implications for Big Data," *IBM (International Technical Support Organization),* vol. Red Paper (ibm.com/redbooks), January 2014.

Zaharaddeen Karami Lawal: Graduated in B.Sc. Computer Science from Kano University of Science and Technology, Wudil. He served as a part time lecturer at School of Technology, Kano State Polytechnic, Kano State Nigeria. Before he moves to India for his Master’s degree (MSc.). At present he is a student of Jodhpur National university. His areas of interest are Big Data Technology, Computer Programming and Security.

Rufai Yusuf Zakari: Hails from Kano state Nigeria, a graduate from Bayero University Kano, Nigeria. Currently pursing M.Sc. Computer science at Jodhpur National University, India His field of interest includes Cloud Computing and Big Data Analytic.

Mansur Zakariyya Shuaibu: Graduated from Kano University of Science and Technology, Wudil with B.Sc. in Computer Science. Currently pursing M.Sc. Computer science at Jodhpur National University, India. He served as a part time lecturer at School of Technology, Kano State Polytechnic, Kano State Nigeria. His areas of interest are Big Data Technology, Digital Forensics and Networking.

Alhassan Bala: Hails from Kano state Nigeria, a graduate from Bayero University Kano, Nigeria. Currently pursing M.Sc. Computer science at Jodhpur National University, India His field of interest includes Data mining, Cloud Computing, Big Data Analytic and .net programing.