ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering
(IJERCSE)
Vol 4, Issue 4, April 2017

Challenges and Issues in Big Data Era- A Survey

[1]Shony KM [2]Navaneeth Krishnan K, [3]Daniel F Netto
ER&DCI-Institute of Technology,C-DAC Campus,Trivandrum,Kerala.

***Abstract:--* Big Data analytics have significant promises in this era of information explosion. Cloud computing, ecommerce , social networking etc. are enhancing importance of big data. However, there are a number of challenges that must be overcome to realize its true potential. Big Data analytic systems use cluster computing infrastructures that are more reliable and available as compare to traditional security analytics technologies. Big data absolutely has the potential to change the way governments, organizations, and academic institutions conduct business and make discoveries, and it is likely to change how everyone lives their day-to-day lives. This will constitute its future scope.**

***Keywords*— Big data, Cloud computing, Challenges and Issues, Data analytics, Anonymity, Big data Security.**

I. INTRODUCTION

Digital world is overwhelmed with huge amount of data generated by number of users worldwide. These data are of diverse in nature, come from various sources and in many forms. Big Data refers to the dynamic, large and disparate volumes of data. Volume, Velocity and Variety are the three different properties of big data. [1]Increase in any one of these property will increase the complexity of analysis. New technologies have emerged to address exploding volumes of complex data with the increase in challenges and Issues. It is important to ensure the various kinds of security needs when considering Big data challenges and issues. The ultimate aim of a big data initiative is the output. This paper is discussing about different challenges and issues that is to be considered at the time of big data analytics.[2]

II. BIG DATA ANALYTICS: CHALLENGES AND ISSUES

Organizations deal with numerous challenges. When working with Big Data we need to understand the need of technology and need of user. Meeting challenges presented by Big Data will be difficult; volume of data increasing every day, velocity of its generation is increasing faster than ever; variety of data is also expanding. Current tools, technologies, analysis approaches, architecture may unable to cop up with complexity of data presented, so challenges and issues are also increases with the increase in needs and complexity.

A. *Specific Challenges Related To Big Data Analytics*

Here we discuss the challenges which are specific to Big Data Analytics.

- ***Data Storage and Retrieval***

Currently available technologies are able to handle data entry and data storage. It is yet difficult to handle semi or unstructured data for transaction processing [2].

- ***Quality vs. Quantity***

While dealing with huge amount of data, sometimes it is difficult to decide which data is inappropriate and how to select most appropriate data? How to ensure authenticity of the data? , How to estimate the value of data? etc [2].

- ***Data Growth and Expansion***

As the organizations increase their services, their data is also expected to grow [2]. Few organizations consider data expansion also whereas most of them won't consider data expansion which poses severe issue on the go.

- ***Speed and scale***

As the volume of data grows, it is difficult to attain deeper perception of data within the given time period, which is more important than processing the complete set of data.

- ***Structured and Unstructured Data***

Difference between structured data (data stored in well-defined tables) and unstructured data (images, videos, text) will affect end to end processing of data. Newer non-relational technologies needed to be invented that will provide some flexibility in data representation and processing.

- ***Data Ownership***

Magnanimous amount of data resides in the servers of social media service providers. These data is stored on behalf of their users and are not really owned by social media service providers. This is one of the ongoing and biggest challenge in area of social media [5].

B. *Technology And Application Challenges*

Most of the technology required for big-data computing is developing at an acceptable rate due to market forces and technological evolution [8] For example, due to the ongoing progress of magnetic storage technology and the large economies of scale provided by both personal computers and large data centers disk drive capacity is increasing and prices are dropping. Whereas the other aspects that pose challenge in big data analytics and require more focused attention, such as:

- ***High-Speed Networking***

Transferring magnanimous amount of data over even on a typical high-speed internet connection requires roughly over a day. This limitation in bandwidth increases the challenge of making efficient use of the cluster based storage recourses and computation. Then difference between the amount of data that is practical to store, vs. the amount that is practical to communicate will continue. And so we need better technology for the same.[4]

- ***Cluster Computing Programming***

Another major challenge is programming large-scale, distributed computer systems for processing very large data sets in reasonable amounts of time. The software must distribute the data and computation. MapReduce programming framework introduced by Google is a method to organize and program such systems. More powerful and general techniques must be developed to fully realize the power of big-data computing across multiple domains.[8]

- ***Extending The Reach Of Cloud Computing***

Much time and expense is involved in getting data in and out of a cloud facility with limited bandwidth. In an ideal world, the cloud systems should be geographically dispersed to reduce problems due to earthquakes and other catastrophes. But, this requires much greater levels of interoperability and data mobility. On the administrative side, organizations must adjust to a new costing model.[5]

- ***Machine Learning And Other Data Analysis Techniques***

Machine learning is still in its early stages of development. Many algorithms are still unable to manage beyond a few million elements or cannot tolerate the statistical noise and gaps found in real practical scenario. The automated or semi-automated analysis of enormous volumes of data lies at the heart of big-data computing for all application domains.

- ***Widespread Deployment***

Although many organizations are collecting large amounts of data, only a handful are making full use of the insights that this data can provide. Until recently, the main innovators in this domain have been companies with Internet-enabled businesses, such as search engines, online retailers, and social networking sites. Most of developing organizations and Technologists are becoming familiar with the capabilities and tools.

- ***Security and Privacy***

Data sets consist of huge amount of sensitive data, and the tools to extract and make use of this information give rise to many possibilities for unauthorized access and use. Privacy preservation is one of the most challenging issue. For example, people are monitored by video cameras in many locations – ATMs, convenience stores, airport security lines, and urban intersections. In addition, cloud facilities become a platform for malicious agents in a low cost manner, e.g., to launch a botnet or to apply massive parallelism to break a cryptosystem. So it is necessary to create safeguards to prevent abuse while developing this technology. [9]

C. *Obstacles In Big Data Implementation*

The obstacles that limit the implementation of big data by any industry are aplenty. Few are listed below.

- ***Data Democratization***

The present business paradigm has brought several small and medium sized organizations who are trying to harness Big Data. When incorporated with approachable analytics capabilities organizations focus on the ability to reduce the time for getting the essence of entire voluminous data

- ***Providing Encryption***

Massive amounts of data being generated, ensuring that the data doesn't fall at risk is essential. Such unsecured data may put organizations or the general human race at risk. , mass awareness of this is need to be initiated and smaller organizations should ensure that the data is in a safe mode. Data encryption can help a lot to solve this issue but the feasibility of encrypting this much amount of data is a challenge.

D. Big Data Security Concerns

- ***Lack of Designed Security***

Big data and many of the big data platforms weren't designed to address different security concerns. That means many platforms not enhanced with security features like encryption, policies, compliance, risk management. If organizations want to ensure their data is secured they have to develop their own features.

- ***Big Data Becomes the Obvious Target of Cyber Attack***

In the cyber space, database integrated by big data is easier to become the target of hackers [14]. First of all, the huge amount of integrated data makes the hacker who successfully attacks the database obtains more data and de-creases the cost of the attack tourist industry, aircraft industry, e-commerce and IT service etc.

- ***Big Data Challenges Existing Save and Security Measures***

Big data brings new security issues for save measures due to its variety. Most data is semi-structured or unstructured. For example, design data, customer in-formation and operating data are always put together, and this might cause some troubles for the management of enterprises such as business security. Updating speed of security protection measures cannot catch up with the growth of mass or large data, therefore there must be some loopholes needing help in the protection of big data[11]

- ***Big Data Is Used As an Attacking Measure***

Hackers use big data technology to launch an attack to companies when these companies use big data to obtain information for commercial value to prepare for the attack and this can make the attack more accurate and result in big loss. In addition, there exist opportunities for hackers to attack based on big data. Hackers use big data to launch botnet exploit, which may control millions of computers and launch attacks at the same time. In addition, big data is used as the carrier of the virus.

E. Anonymity Concern

Detailed information about the customer identities, behaviors, motivations, and other sensitive facts can be collected from big data this is an issue from customer side. Some companies respond to these concerns with data masking policies, much more. Data consumers can be anyone from high level executives to customer's antibusiness users. More diverse data simply means more work is needed to protect it.

- ***Data Breaches***

Cyber crime will get easier if criminals get more amount of data, and companies that collect and store it are big targets. Attacks like APT collect terabytes of data to collapse the entire targets. Data breaches are now far common and will likely not go away anytime seen.[3]

- ***Security Spending Still Low***

When considering about developing companies, amount spending on their security is low. Awareness and carefulness can also improve security without extra expense. Still necessary security measures have to taken.

- ***Big Data Skills Gap***

Avoiding different vulnerabilities and security issues would be possible even with limited resources if the right people for the job were working on it, but many businesses doing without properly trained people. It has to find the right people with the right skills to handle the work, which only increases the severity of the problems.

- ***Data Brokers***

Company may sell some data to third parties or share data for doing some part of the process by collaborating with someone else, It will increase risks could increase.

F. System Challenges

Designing and deploying a big data analytics system is not a common & straight forward task. For the proper working of new hardware and software platforms demand new infrastructure and models to address the wide range of challenges of big data. Following are some among them,

- ***Data Representation***

Many datasets are heterogeneous in type, structure, semantics, accessibility, granularity and organization. An effective data presentation should be designed to reflect the structure integration technique should be designed to work across different datasets.

- ***Data Compression and Redundancy Reduction***

Typically, there is a large number of redundant data in raw datasets. Efficient methods have to be implemented for data compression and redundancy reduction without scarifying potential value.

- ***Data Life-Cycle Management***

As data sets grow and the real time requirement becomes more strict, analysis of the entire dataset becoming difficult. Approximate results can provide to solve this problem. The accuracy of the result and the groups omitted from the output will be the notion of approximation.

- ***Social Media***

Social media possesses unique properties, such as statistical redundancy, vastness and the availability of user feedback. Various extraction techniques have been successfully used to identify references from social media to locations specific product names, or people on websites. By connecting interfiled data with social media, applications can achieve high levels of precision and distinct points of view.[8]

- ***Deep Analytics***

When go for deep analytics we have to consider the potential pillars of privacy and security mechanisms like access control, privacy-aware data mining and analysis, security storage and management.

- ***Energy Management***

The energy consumption of large scale computing systems has attracted greater concern from economic and environmental perspectives. Transmission, storage, and processing will inevitably consume more energy, as data volume and analytics demand increases. [6]

- ***Scalability***

A big data analytics system must be able to support very large datasets from present to future. All the components in big data systems must be capable of scaling to address the ever-growing size of complex datasets.

- ***Collaboration***

Big data analytics is research field that requires specialists from multiple professional fields to harvest output. A comprehensive big data cyber infrastructure is necessary to allow communities of engineers and scientists to access the diverse data apply their respective expertise and cooperate to accomplish the goals of analysis and to share their findings improves the efficiency ahead.[10]

G. Issues And Challenges Due To Data Provenance

In big data research, privacy and security of big data play a major role. Provenance of big data is relevant as well. Data provenance concerns with the problem of detecting from where the data is originated and the propagation process of data within the entire system. In other words, data provenance consists in the lineage and derivation of data and data objects, and it puts its conceptual roots in extensively studies performed in the past in the contexts of arts, literary works, manuscripts, sculptures, and so forth. Ownership of data which refers to the issue of defining and providing information about the rightful owner of data assets[7].

- ***Information Sharing***

Information sharing introduces relevant research challenges as well as technological drawbacks. Credential and private information may leak.

- ***Minimum Computational Overhead Requirement***

Data provenance techniques may be intensive and resource consuming. This requires hard ware and implementation techniques that introduce a minimum computational overhead, in order to avoid impacting on the performance of the target system.

- ***Query Optimization Issues***

Data provenance techniques need to access and query data in order to determine their provenance. This will make severe drawbacks when these techniques run over big data.

- ***Transformation Issues***

During data provenance tasks, data sources need to be transformed among different data formats. Tracing provenance must be introduced accordingly, in order to track all the different transformations occurred.

- ***When Computing Provenance?***

There exist two alternatives for computing provenance. One predicates to compute provenance only when the same provenance is required (lazy provenance model). The other one argues to computer provenance every time data are transformed (eagerly provenance model). Both models have pros and cons. They also imply different computational overheads. This one is still an open problem to be considered in future efforts.

- ***Heterogeneity of Data Source Models***

Data provenance techniques usually run over heterogeneous data sources, so that they need to cope with heterogeneous data models. Heterogeneity of data sources is a big challenge for such techniques, as data sources expose different formats and characteristics.

- ***User Annotation Support for Provenance***

Big data contains different user annotation, In order to enhance the effectiveness of this process corresponding domain experts are assigned to annotate data. As a consequence, data provenance tools need to introduce.[7]

III. CONCLUSION

Big data is leading the information revolution with the development of information technology. People are still facing the challenges that restricted them from enjoying the convenience from the maximum benefits of big data. If the security of information could not be guaranteed miss use of big data also become increase. Thus, information security is of great importance in the age of big data.

REFERENCES

[1] Keith C.C. Chan, "Big data analytics for drug discovery," IEEE International Conference on Bioinformatics and Biomedicine, 2013. The Hong Kong Polytechnic University.

[2] ParthChandarana and M. Vijayalakshmi, "Big Data Analytics Frameworks," International Conference on Circuits, Systems, Communication and Information Technology Applications (CSCITA), 2014

[3] Han Hu, Yonggang Wen2Tat-Seng Chua1 and Xuelong Li, "Toward Scalable Systems for Big Data Analytics,"

[4] KapilBakshi,"Big Data Analytics Approach for Network Core and Edge Applications,"

[5] Qiufen Xia, Weifa Liang and ZichuanXu, "Data Locality-Aware Query Evaluation for Big DataAnalytics in Distributed Clouds," Second International Conference on Advanced Cloud and Big Data, 2014.

[6] Jiankun Huand Athanasios V. Vasilakos, "Energy Big Data Analytics and Security: Challenges and Opportunities," DOI 10.1109/TSG.2016.2563461.

[7] Alfredo Cuzzocrea,"Provenance Research Issues and Challenges in the Big Data Era," IEEE 39th Annual International Computers, Software & Applications Conference,2015.

[8] Kyounghyun Park, Minh Chau Nguyen, Heesun Won, "Web-based Collaborative Big Data Analytics on Big Data as a Service Platform,"

[9] Gennady Smorodin and Olga Kolesnichenko, "Big Data as the Big Game Changer Big Data-driven world needs Big Data-driven ideology"

[10] Randal E. Bryant, Randy H. Katz, Edward D. Lazowska, "Big-Data Computing: Creating revolutionary break throughs in commerce, science, and society,"

[11] YogeshDhinwa, "Applying Data Analytics on Vulnerability Data," SANS Institute, Info Sec Reading Room December 21, 2015.