Big Data Analytics, Integration with Cloud Computing – An Overview and Security Challenges

Rekha Sunny T, Kshema Suku

Assistant Professor, SCMS, Cochin, Kerala, India
Scholar, SCMS, Cochin, Kerala, India

***Abstract:** The wide popularity of Internet and the tremendous increase in data processed through the increasingly popular applications such as social networking and semantic web require effective methods to be introduced for the management and analysis of the large-scale data. Big data analytics arose as a technology that enables us to store and process big and varied volumes of data, providing both business and scientific areas a way to correlate data, find patterns and predict new trends. At the same time, the emergence of cloud computing and its steady growth resulted in providing a reliable, fault-tolerant and scalable environment to harbour the big data. In this paper we present an overview of both these technologies, the scope and benefits of integrating big data and cloud frameworks and the challenges involved particularly that associates with security of data.*

***Keywords:** Big Data Analytics, Cloud Computing, Security Issues, Confidentiality*

1. Introduction

The tremendous increase of computational power has resulted in the production of an overwhelming flow of data and the need for techniques to efficiently manage and utilise this data. Consequently the term big data and in particular big data analytics, arose as a technology that stores and processes big and varied volumes of data, providing both business and scientific areas a way to correlate data, find patterns and predict new trends. Significant resources were allocated to support these data intensive operations which lead to high storage and data processing costs. The current technologies such as grid and cloud computing have all intended to access large amounts of computing power by aggregating resources and offering a single system view. Among these technologies, cloud computing is becoming a powerful architecture to perform large-scale and complex computing, and has revolutionized the way that computing infrastructure is abstracted and used. Cloud provides a reliable, fault tolerant, available and scalable environment so that big data systems can perform (Hashem et al., 2014). Hence leveraging these two technologies, can provide businesses with a competitive advantage, and science with ways to aggregate and summarize data from various experiments.

Big data is defined as "new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling the high velocity capture, discovery and or analysis" [1]. The elements of big data can be categorized in to velocity, volume, veracity, value and variety. Cloud computing is an information technology paradigm that enables ubiquitous access to shared pools of configurable system resources and high-level services that can be rapidly provisioned with minimal management efforts. Cloud computing relies on sharing of resources to achieve coherence and economies of scale. Three most popular cloud paradigms includes: Infrastructure as Service (IaaS) , Platform as a Service(PaaS) and Software as a Service(SaaS) [5].Cloud computing have all intended to access large amount of computing power by aggregating resources and offering a single system view. The cloud computing can be used as a solution for tacking big data, such as large scale, multi-media and high dimensional data sets. Cloud computing is associated with new paradigm for the provision of cloud computing infrastructure and big data processing method for all kinds of resources. Big data is responsible for storing and processing of data and cloud provides a reliable, available and scalable environment. Cloud computing delivers all these through hardware virtualization and enables big data to be available, scalable and fault tolerant [6]. Security issues in the cloud are a major concern for the business and cloud providers today, because the attackers are relentless and they keep inventing new ways to crack the security.

In this paper, both these technologies are discussed in detail along with the areas to improve or have yet to be addressed in both technologies. The rest of this paper is organized as follows: Section 2 provides the literature review comprising an overview of big data, cloud computing and the characteristics; Section 3 discusses big data classification and section4 explains big data in cloud computing. In section5, the security challenges associated with the integrated environment are discussed and section6 concludes the paper.

2. Literature Review

2.1 Big data analytics

Big data is the term which describes massive data set which is having large, varied, complex or unstructured data with the difficulties of storing, analyzing and also visualizing processes or the results. The process of searching massive amount of data to reveal hidden patterns and the secret correlations named as big data analytics. Big data analytics refers to the strategy for analyzing large volume of big data and is considered as a subcomponent to the analytical process of big data. A maturity-scape model that examines the technology process using five-stages such as intent, data, technology, people and process is shown in figure1.

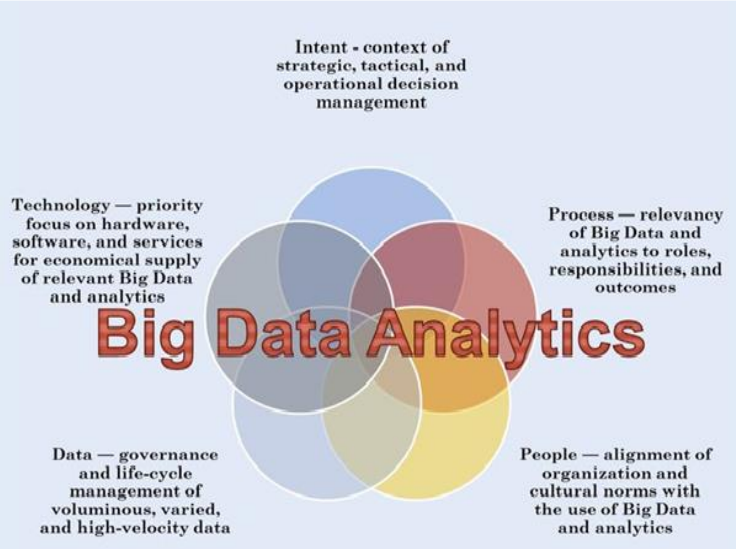
Intent - context of strategic, tactical, and operational decision management
Technology — priority focus on hardware, software, and services for economical supply of relevant Big Data and analytics
Process — relevancy of Big Data and analytics to roles, responsibilities, and outcomes
Big Data Analytics
Data — governance and life-cycle management of voluminous, varied, and high-velocity data
People — alignment of organization and cultural norms with the use of Big Data and analytics

Figure 1: Big data analytics

2.2 Cloud computing

Cloud computing is one of the most significant aspect in modern ICT and service for enterprise applications which has become a powerful architecture to perform large scale and complex computing process. The main advantages of cloud computing include visualized resources, parallel processing, security and data service integration with scalable data storage. Cloud computing not only minimize the cost but also restrict for automation and computerization by individuals and enterprises. Some of the first adopters of big data in cloud computing are users who deploy hadoop clusters with high scalability and elastic computing environments provided by the vendors such as IBM, Microsoft Azure and Amazon AWS. Virtualization is the one of the basic technologies which is applicable for the implementation of cloud computing. Storage process, analysis, access and managing distributed computing components in bigdata environment are achieved by virtualization process. Virtualization is the process of resource sharing and isolation of hardware to increase computer resource utilization, efficiency and also scalability [1].Big data analytics deals with the high level of capacity and often demand for the resources such as data and domain experts and analytic skills. Hence, we need to identify the business model- especially data and models that reside on the providers. Clouds enables big data analytics to explore means to allocate and utilize these special resources in proper manner [2].

2.3 Characteristics of big data

The big data is dimensioned in to five categories such as volume, variety, value, velocity and veracity [3] [figure 2].

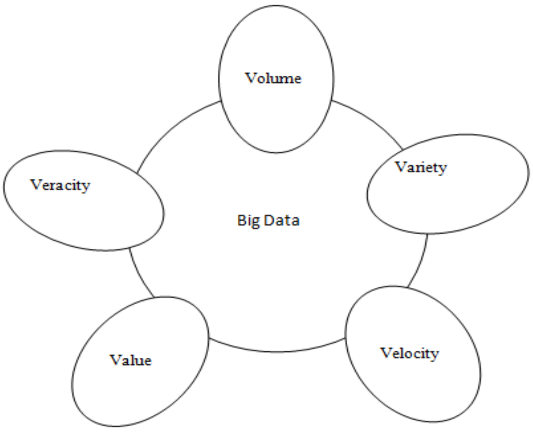

Figure 2: BD5

1. **Volume-** Volume refers to the amount of all type of data which is generated from different sources and continue to expand. The advantage for gathering of large amount of data include creation of hidden information and patterns through data or information analysis. Collection of longitudinal data requires considerable effort and underlying investment. Mobile data produces interesting results similar to that of examination of human behavior patterns and visualization techniques for complex data.
2. **Variety-** Variety refers to the different types of data collected from sensors, smartphones or social networks etc. These data type include video, images, text, audio and data logs which is either structured or unstructured format. Most of the data which is generated from mobile applications are unstructured format. For example, text messages, online games, blogs and social media where generated different types of unstructured data through mobile devices and also from different sensors.

3. **Velocity-** Velocity refers to the speed of data transfer. The content of data constantly changes because of absorption of complementary data collections, introduction of previously archived data and streamed data arrived from multiple sources [1].
4. **Veracity-** Veracity refers to the biases, noise and abnormality in data, is the data which is being stored and also being mined meaningful to the problem being analysed [7]. The data quality of the captured data can be varied gently and it can affect the accurate analysis of the data or the process[8].
5. **Value-** Value is the most important aspect of big data and refers to the discovering process of huge hidden values from large datasets with various types and rapid generation [1].

3. Big Data Classification

Big data classified in to different categories for better understanding of their characteristics. Figure 3 shows the numerous categories of big data. The classification is important because of large scale data in cloud. The classification which is done depending up on the five aspects, they are (i) data sources,(ii) content format, (iii) data store,(iv) Data staging and (v) data processing .Each categories having their own characteristics and complexities.

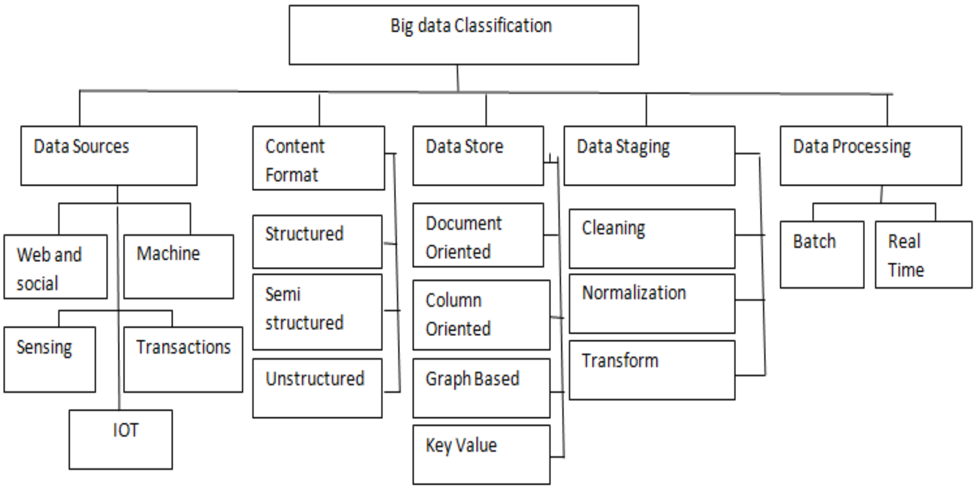

Figure 3: Big data classification

The data source which include internet data, sensing and all storing of transnational information which is ranging from unstructured to highly structured and thus they are stored in different formats. As result of wide varieties of data sources the captured data differ in size with respect to its redundancy, consistency and noise etc.[1]

4. Cloud Computing usages in Big Data

Integration of cloud computing with big data can provide businesses with a competitive advantage, and science with ways to aggregate and summarize data from various experiments. Big data provide users ability to use computing process to distributed queries and across multiple datasets and return resultant set in timely manner. Cloud computing on the other hand, delivers all these through hardware virtualization and enables big data to be available, scalable and fault tolerant and Hadoop can be used for the same. Usage of cloud computing is detailed in Figure4.

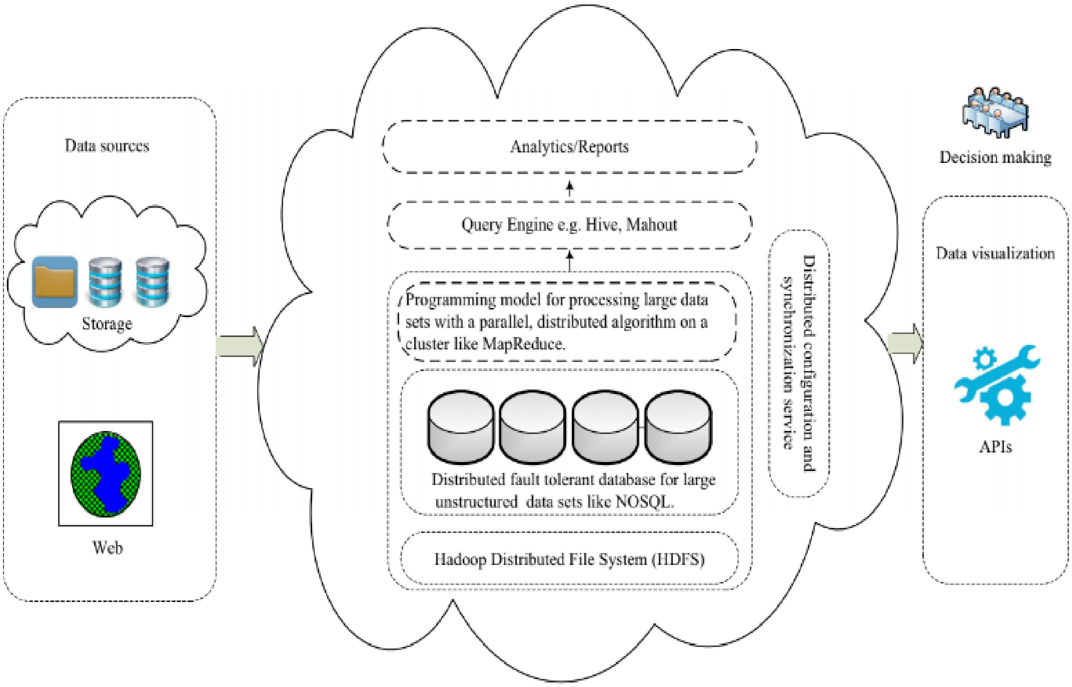
Data sources
Storage
Web
Analytics/Reports
Query Engine e.g. Hive, Mahout
Programming model for processing large data sets with a parallel, distributed algorithm on a cluster like MapReduce.
Distributed fault tolerant database for large unstructured data sets like NOSQL.
Hadoop Distributed File System (HDFS)
Distributed configuration and synchronization service
Decision making
Data visualization
APIs

Figure 4: Cloud computing usages in big data

Large amount of data sources from the cloud and the web are stored in the form of distributed fault-tolerant database and processed by the programming model for the large datasets and with a parallel cluster algorithm [1]. The cloud computing not only provide facilities but also help for the computational and processing of big data and also serves as service models [2]. Cloud computing is related by a new pattern and is used for the provision of cloud computing infrastructure and big data processing. Several cloud based technologies have to be cope with this new environment because it is dealing with big data for concurrent processing and has become increasingly complicated. Map reduce is the good example for the big data processing in cloud computing environment and they allow for the processing of large amount of data sets stored in parallel to the clusters[1]. Big data that are implemented in data mining and they deals with the big data mining. Big data processing framework rely on the cluster computer with high performance on the computer platform. Where the data mining task which is deployed by running programming tool parallel, Map reduce or Enterprises Control Language (ECL) on large number of computer nodes. The role of the software component is to make sure that the single data mining task (find best match of query) from database with billions of samples and they are split in to sub tasks (small tasks) and each running on one or multiple computer nodes[3].

5. Security Challenges

The challenges associated with big data analytics such as scalability, availability, data storage and data staging can be resolved to a certain extent by Hadoop and MapReduce. Several security issues such as privacy, integrity, confidentiality and availability of data that exist in big data must need to be assessed at regular interval to protect it from threats. Effective cryptographic techniques have to be used to encapsulate sensitive amount of data in cloud computing environment. The following are some of the challenges faced by the big data in cloud security.

- In most of the distribution system, single level of protection is not recommended.
- There is having security issues on the demand with the non- relational databases (NoSQL).
- Automated data transfer requires additional security measures, which are not available yet.
- If the system receives large amount of information or data, it need to be validated to ensure if it remains in trustworthy and also having accuration, but this operation which doesn't always occur [4].
- Some of the organization doesn't have institute access control to increase the level of confidentiality with the company.
- Because of large or massive amount of data in the cloud, its origins are not consistently monitored and tracked.
- The access control encryption and the security connections, can be dated and also inaccessible to the IT specialists who relay on it.
- In most cases, unethical IT specialists practicing mining the information can gather the personal data, without asking user permissions or notify them [4].

Different criteria taken in consideration, are as listed below:

- Confidentiality- The data which sent from the sender which should be read by the receiver or the receiving user only.
- Integrity-The message which is sent by the sender to the receiver which should be received at the same time without any delay.
- Security-which specifies with the security level of algorithm.
- Adversary types- Which determine whether the algorithm is malicious or not.
- Time complexity-It is the amount of time which is taken by an algorithm to run, the [4] functions of the length with the string representing the input.
- Performance scrutinizing-Which determine the performance of the algorithm while testing.

6. Conclusion

Integrating big data and cloud frameworks provides a reliable, fault-tolerant and scalable environment to harbour the big data and researches are going on to effectively implement the same. Several challenges are been faced with among which providing perfect security for big data in cloud computing is found to be the most critical. With this regard, a detailed study of cloud computing usages in big data and the security issues is done in this paper.

References

[1] Ibrahim Abaker Targio Hashem, Ibrar Yaqoob , Nor Badrul Anuar ,Salimah Mokhtar , Abdullah Gani, SameeUllah Khan, *The rise of "big data" on cloud computing: Review and open research*, Elsevier Ltd, 1-18.

[2] Marcos D. Assuncao, Rodrigo N Calheiros , Silvia Bianchi, Marco A.S. Netto , Rajkumar Buyya, *Big Data computing and clouds: Trends and future directions*, Elsevier Ltd, 1-13.

[3] Changqing Ji, Yu Li, WenmingQiu, UchechukwuAwada , Keqiu Li, *Big Data Processing in Cloud Computing Environments*, International Symposium on Pervasive Systems Algorithms and Networks , 1-7.

[4] SreenivasaB.L, Manish Kumar, Mohammed Nueed Shaikh and Dr. S Sathyanarayanana , *A Study On Encryption Decryption Algorithm For Big-data Analytics In Cloud* , International Journal of Latest Trends in Engineering and Technology, 1-7.

[5] J.K. Laurila, D. Gatica-Perez, I. Aad, J. Blom, O. Bornet, T.-M.-T. Do, O. Dousse, J. Eberle, M. Miettinen, *The mobile data challenge: Big data for mobile computing research*, Workshop on the Nokia Mobile Data Challenge, in: Proceedings of the Conjunction with the 10th International Conference on Pervasive Computing, 2012, pp. 1–8.

[6] F. Chang, J. Dean, S. Ghemawat, W.C. Hsieh, D.A. Wallach, M. Burrows, T. Chandra, A. Fikes, R.E. Gruber, *Bigtable: a distributed storage system for structured data*, ACM Trans. Comput. Syst. (TOCS) 26 (2008) 4

[7] P. Noordhuis, M. Heijkoop, A. Lazovik, *Mining twitter in the cloud: A case study, Cloud Computing (CLOUD)*, 2010, Proceedings of IEEE 3rd International Conference on, IEEE, Miami, FL, 2010, pp. 107–114.

[8] C. Zhang, H. De Sterck, A. Aboulnaga, H. Djambazian, R. Sladek, *Case study of scientific data processing on a cloud using hadoop*, High Performance Computing Systems and Applications, Springer, 2010, 400–415.

[9] F. Faghri, S. Bazarbayev, M. Overholt, R. Farivar, R.H. Campbell, W.H. Sanders, *Failure scenario as a service (FsaaS) for Hadoop clusters*, Proceedings of the Workshop on Secure and Dependable Middleware for Cloud Monitoring and Management, ACM, 2012, p. 5.