Big Data on Cloud Computing, Challenges and Opportunities – A Conceptual Model

A S. T. Nishadi

Faculty of Graduate Studies, Masters in Information Systems Management, University of Colombo, Sri Lanka

Abstract: ***The rapid growth of the Internet, Internet of Things(IoT) and cloud computing have been resulting to the explosive growth of big data in almost most of the industries and academia. However, analyzing the large volumes of data which generated from various sources requires a lot of efforts at multiple levels to extract knowledge for decision making. Cloud computing is one such powerful technology to perform massive scales of data and complex computing. The paper will introduce the concepts and definitions of big data and cloud computing, challenges of big data and possible cloud solutions to address the challenges. Further, this will suggest a conceptual model which expressing three segments of challenges including data, processing and organizational and also provide ideal cloud solutions in order to handle the proposed challenges.***

Keywords: big data, big data challenges, cloud computing

1. Introduction

The dramatic increase of volumes of data generated by mobile devices, Global Positioning Systems(GPS), computer logs, social media, sensor devices and all the other digital monitoring systems has been producing both structured and unstructured formats of data. Therefore, managing the large volume of data is a significant challenge. The term big data is referred as huge and complex, structured and unstructured data which is difficult to manage using traditional technologies such as Database Management Systems (DBMS) [1]. Big Data is not simply a data rather it has become a vast computer technology which involve various tools, techniques and frameworks. Big data, and in particular big data analytics, are viewed by both business and scientific areas as a way to correlate data, find patterns and predict new trends.

Cloud computing is one of the most significant shifts in modern technology for enterprise applications and has become a powerful architecture to perform large-scale and complex computing. There are many advantages provided by cloud computing including virtualized resources, parallel processing, security and data service integration with scalable data storage [2].

Cloud computing is the perfect vehicle for hosting big data. Big data needs for multiple servers to hand parallel processing. Moreover, cloud computing already uses remote multiple servers and allow resource allocation. In addition to that, using cloud computing provides faster provisioning to big data as provisioning servers in the cloud is very fast and feasible. Hadoop and Map reduce are used in cloud environment which support to processing large sets of data in a distributed computing environment.

The aim of the study is to introduce a comprehensive analysis of big data and cloud computing with providing definitions and characteristics. In addition to that, it will further identify the implementation of big data in cloud environment, the tools and technologies provided by cloud computing in order to handle issues of big data.

The section II and III of the paper will present the concepts and definitions of big data, cloud computing and big data analysis. Further, in section IV of the study will explain the challenges of big data based on three identified segments which elaborating data, processing and organization. Finally, the section V and VI the report will suggest the conceptual model which expressing three segments of challenges and also ideal cloud solutions in order to handle the proposed challenges.

2. Characteristics of Big Data

The growth of digital data which generated from various sources and fast transition though the digital technologies has led to big data. In general, big data refers to the collection of large volumes and complex data which is difficult to store, process and analysis using traditional database technologies. The first 3Vs model of big data which referred as volume, velocity, and variety introduced by Laney [3]. Volume refers to the large amount of data that are being generated everyday whereas velocity refers to the rate of growth and how fast the data are gathered for being analysis. Variety provides information about the types of data such as structured, unstructured, semi-structured etc. The prime objective of big data analysis is to process data of high volume, velocity, variety, and veracity using various traditional and computational intelligent techniques. Further, the forth V which refers to veracity that includes availability and accountability [4].However, the 5Vs model which added another crucial factor of big data which indicated volume of data, variety, velocity, veracity (alternatively guarantee of the data quality or validity) and value which denotes the added value for companies [1]. Moreover, the 6V model which indicates volume, variety, velocity, veracity, variability and value further defines the modern view of big data [5].

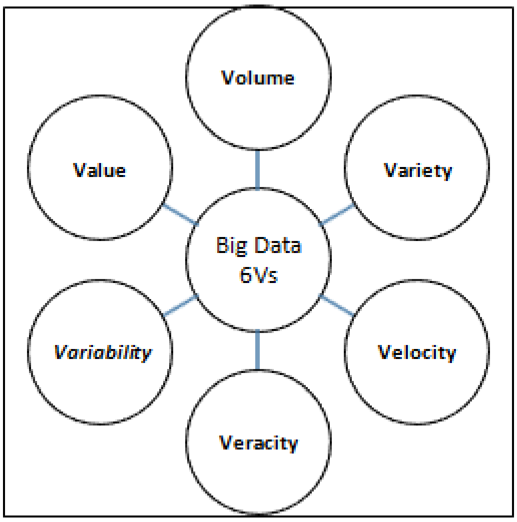

Figure 1: 6V model of Big Data

- **Volume** – the huge amount of data that is produced each day companies (i.e. the generation of data is large and complex hence no longer be saved or analysed using conventional methods)
- **Variety**– diversification of data types and data sources available (for example nearly 80% of data in the world today unstructured therefore at first glance does not indicate any relationships)
- **Velocity-** refers to the speed with which the data is generated, analysed and reprocessed.
- **Veracity-** guarantee of the data quality (alternatively
- **Validity** is the authenticity and credibility of the data*)*
- **Variability** - variability regards about consistency of the data over time
- **Value -** denotes the added value for companies. (Many companies have recently established their own data platforms and invested a lot of money in infrastructure. Therefore, this is a measure of evaluating the business value generating from their investments.)

3. Cloud Computing

Cloud computing has been considered as a one of the emerged technology for hosting and delivering services over the internet. Therefore, the adoption of cloud computing is maturing in the global context. The National Institute of Standards and Technology defines cloud computing as a type of model which enables convenient, on-demand network access to a shared pool of configurable computing resources and also can be rapidly provisioned and released with minimal management effort or service provider interaction [6]. According to the European Network and Information Security Agency defines the concept cloud as, 'On-demand service model for IT provision, often based on virtualization and distributed computing technologies' [7]. However, the first definition of cloud computing status as a computing pedagogy where the boundaries of computing will be determined rationale rather than technical. The characteristics of includes a disruptive shift of the computer stack to online services, allowing on-demand access to software applications, development and deployment environments, and computing infrastructure on a pay-per-usage basis [8].Therefore, cloud is referred as a technical framework, which support on-demand network access and shared pool of resources which cater for modern business requirements.

4. Challenges of Big Data

Big data has many challenges which ranging from the designing of big data to acquiring the new knowledge. Among these challenges, some of the issues caused by the characteristics of big data, some, by its current analysis models and methods, and others are generated due to the limitations of current data processing systems. In this section, it will briefly describe three basic types of challenges of big data namely data, processing and organizational.

A. Data challenges

Data challenges are related to the characteristics of data. Different researchers have indicated different aspects of data. The 3Vs of data (volume, velocity and variety) [3], 4V aspects (volume, velocity, variety and veracity) [4]and 5V model (volume, variety, velocity, veracity and value) [1]. The 6V areas volume, velocity, variety, veracity, variability and value [5].

B. Processing challenges

The issues which encountered while processing the data from data capturing to interpreting is indicate as process challenges. As many of the forms of big data sets are usually non-relational, unstructured or semi-structured, thus processing these types of data generate a significant challenge than managing big data [9].

Big data processing life cycle consists of four phases; collection phase, data storage phase, data processing and analysis, and knowledge creation [10].

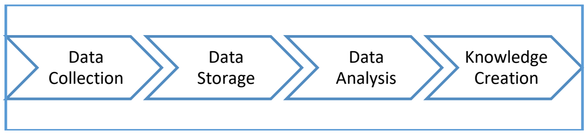

Figure 2: Big data life cycle

Data Collection: This is related to acquiring or collecting data from different sources; structured, semi-structured, and unstructured. Millions of terabytes of big data are generated daily due to the expansion of modern devises in many business sectors. However, the real challenge in here filtering required for the data that is what is useful to capture and what is useless to discard [11].

Another major issue in data collection is lack of provenance. The Provenance information of data explains the creation process and origin of data by recording which transformations are responsible in creating a certain piece of data and from which data items a given data item is derived [12]. Since the recording the origin and movement of data in the processing pipeline will help to indicate the next processing steps. However, if there is an error happened during one stage, it will affect to all the subsequent analysis stages. The third major issue of data collection is automatic generation of metadata. Metadata provides the information about the meaning of data, terminology, concepts, relationships of data, provide information about the source of the data(provenance) [13].Therefore, the efficient analytical algorithms are required to understand the provenance of data and process the vast streaming data and to reduce data before storing [14].

Data Storage: In this phase, the collected data is stored and prepared for being used in the next phase (data analytics phase).Basically, the challenges in this stage are related to extracting and cleaning data from collection of large unstructured data. The main challenge in this phase is right information extraction. Due to the strident, vibrant, diverse, inter-related and unreliable features, the mining, cleansing and analysis proves to be very challenging [15].Moreover, the collected data is mostly not in the proper formats which required for processing. For example, there are mismatching data in health records such as medical reports, prescripts, reading and captured data by sensors and monitoring machines and image data (X-Rays). These mismatching generate due to various reasons such as patients hide some risky information, application errors etc. Therefore, the second challenge is detection of error models. Thus, there is a need to develop proper extraction methods that mines data from unstructured data sources.

Data Analysis: The useful knowledge using data mining methods such as clustering, classification, and association rules generate in this phase. The main challenge in this phase of big data is heterogeneity. The term heterogeneous refers to multiple dimensions, multiple sources, structured, semi-structured and unstructured data and huge amount of data [16]. Big data mostly aggregates various online sources of data such as tweets, microblogging, Facebook data [17]. Therefore, big data needs mechanisms of maintaining and aggregating heterogeneous data.

Knowledge Creation: The final stage of big data processing, new information and valued knowledge are derived by decision makers. Further, this is relatively similar to visualizing data and making understandable patterns for users that is the data analysis and modelling thus results are presented tothe decision makers to interpret the findings for extracting sense and knowledge [18]. Major challenge of big data is the shortage of people with analytical skills to interpret big data [19]. In addition to that, technical issues of data presentations such as application issues, wrong data modelling and data errors. As much of big data activate and reside on online resources, defining the internet computing technological solutions is a challenge which allow access, aggregate, analyze, and interpret big data [20].

C. Organizational challenges

Organizational challenges related to the big data are considered as challenges of data accessing, managing and governing data. Therefore, the main four challenges of organizational has identified as security, privacy, governance and ethical issues.

Security: Big data will not achieve the required level of trust if it does not provide high level of security. Cloud Security Alliance defines four different aspects of big data security i.e. infrastructure security, data privacy, data management and integrity and reactive security [21]. Big data analysis massive amount of data which is correlated, analyzed and mined for meaningful patterns. Therefore, preserving sensitive data is a huge risk for organizations. Security of big data can be enhanced by using techniques such as authentication, authorization and encryption. The major challenge is to develop a multi-level security, privacy preserved data model for big data [22].

Privacy: Extracting data using different analytical tools generates privacy issues. Not only for organization wise but also individuals must to put access controls for information. Organizations may need to protect their competitiveness by not sharing sensitive data about their clients and users, or data about their own operations. Privacy is undoubtedly an issue as systems getting increased with huge quantities of personal information. Therefore, it is highly required to have personally identifiable regulated framework in order to make sure the confidentiality of big data.

Governance: Data governance refers to practices and organizational policies which describe managing data. The practices composed with structural practices (roles and responsibilities of key IT and non – IT decision makers regarding the data ownership), operational practices (data migration, data retention, access rights, cost allocation and backup and recovery) and relational practices (knowledge sharing, value analysis, education, training and strategic planning) [23].

Ethical Issues: The ethical implications associated with big data analytics is just emerging. Big data analytics relies on algorithmic decision making process which data captured and combined from different sources and aims to predict individual behaviors of current or past [24]. Most probably, the algorithmic decision making process focus on identifying patterns of relationships which meaningful and represent the cause and effect in a social phenomenon [25]. Ethical issues might arise in each phase of big data analysis; therefore, recognizing big data is not only as a technology, but rather it combines with several other parties such as industries, producers, distributes and customers. Thus, ethical issues are highly influencing in big data analysis.

5. Conceptual Framework of Big Data Challenges

The proposed conceptual framework illustrates the relationship between three major variables of big data namely data, processing and organizational. Further, the data challenges sub divided into six independent (IVs) such as volume, velocity, variety, veracity, variability and value. Processing challenges composed with four phases including data collection, data storage, data analysis and knowledge creation. Three challenges of data collection have indicated as data filtering, lack of provenance and metadata generation. In addition to that, two major challenges which derived in the data storage are valid extraction and error detection. The third stage of processing challenge is heterogeneity. Knowledge and skills and technical issues are two main proposed challenges in the final stage.

Further, four major challenges identified as organizational challenges are indicated as privacy, security, governance and ethical issues.

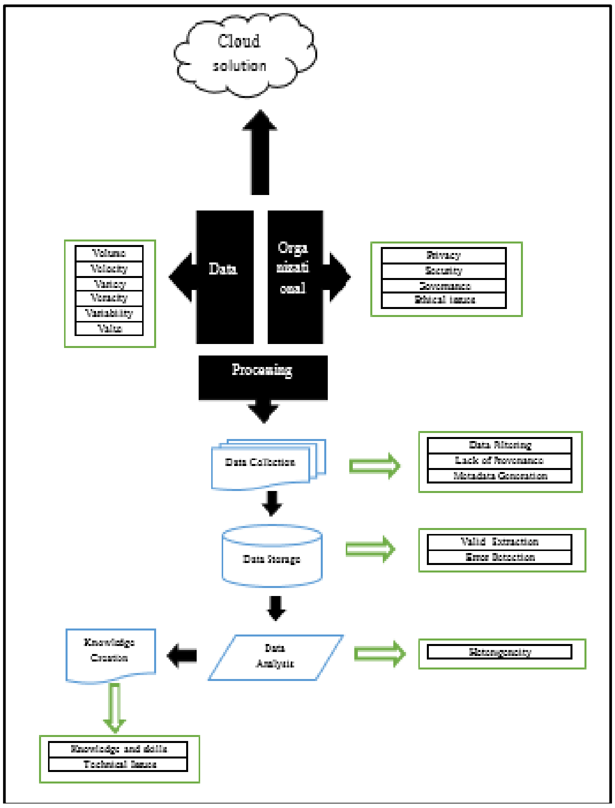

Figure 3: Conceptual framework of big data challenges

Integrating big data with cloud computing provides many options for the big data challenges.

6. Proposed Cloud Solutions

Integrating cloud with big data provides many advantages. Due to massive growth of data, big data need for multiple servers in parallel. (big data demand for high velocity and high variability). Meanwhile, cloud computing already uses multiple servers and allow for parallel resource allocation. Therefore, this is a great solution for build big data on cloud multi servers as cloud provide dynamic resource allocation (support for better efficiency of analysis). Cloud systems are mainly based on remote multi servers which makes it feasible to handle massive amount of data simultaneously.
The integration between cloud computing and big data would result in cost reduction (the feature of pay-as-you-go model in cloud). In addition to that, using cloud computing provides faster provisioning to big data as provisioning servers in the cloud is so easy and feasible. Therefore, the processing requirements of big data can be scaled due to feature of scalability provided by cloud environment. Further, big data needs skilled workers to handle the complexity. However, cloud computing complements big data and provides convenient, on-demand and shared computing environment with minimal management effort and reduced overhead. In addition to that, the integration between both, makes big data resources more controllable, monitored and report-ed. Moreover, this integration enables reducing the complexity and improving the productivity.

There are many issues arise in the context of data processing in the big data. Hadoop and MapReduce in provides ideal solutions for these issues. Hadoop, an open source MapReduce implementation, which allows for the creation of clusters that use the Hadoop Distributed File System (HDFS) to partition and replicate data sets to nodes.

Hadoop: Hadoop is a java-based programming framework which supports processing of large volumes of data in a distributed computing. This is a part of the Apache project. Hadoop mainly use for processing a cluster of servers and application which consist with terabytes of data. This supports with rapid transfer rates even in the case of some node failures.

Hadoop Map Reduce: This also a framework which used to write applications with large amounts of data in parallel clusters. Hadoop Map Reduce is a framework used to write applications that process large amounts of data in parallel on clusters of commodity hardware resources in a reliable, fault-tolerant manner. A Map Reduce job first divides the data into individual chunks which are processed by Map jobs in parallel. The outputs of the maps sorted by the framework are then input to the reduce tasks. Generally, the input and the output of the jobs are both stored ina file-system.

Hadoop Distributed File System (HDFS): This is a file system which used to store all the nodes in a Hadoop cluster and also help to improve the reliability, support security, privacy issues of big data by replicating data among the multiple servers.

NoSQL: Another trend provided by cloud computing is NoSQL data bases for storing and retrieval of data (support for heterogeneity – structured, unstructured and semi-structured sources)

7. Conclusion

Big data is one of the most critical technological trends in the world. In general, large volumes of complex data (big data) difficult to store, process and analysis using traditional database management systems. Initially, the study covers definitions of big data and cloud computing. Then, it has explained the existing challenges of big data in three segments such as data, processing and organizational. However, integration of big data with cloud computing seems to be a perfect match for hosting big data. The major advantage with the cloud computing and big data integration is the data storage and processing power availability, cloud has access to a large pool of resources and various forms of infrastructures. In addition to that, the study proposes conceptual framework for big data challenges and finally provide cloud based solutions in order to handle generated big data issues.

References

[1] Marr, B.2015. Big Data: Using SMART Big Data. Analytics and Metrics to Make Better Decisions and Improve Performance
[2] Hashem, I., Yaqoob, I., Anuar, N., Mokhtar, S., Gani, A. and Ullah Khan, S. (2018). The rise of big data on cloud computing: Review and open research issues.
[3] Laney, D. (2001), Application Delivery Strategies, Retried from https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data- Management-Controlling-Data-Volume-Velocity-and-Variety.pdf on 1st May 2018
[4] M. K.Kakhani, S. Kakhani and S. R.Biradar, Research issues in big data analytics, International Journal of Application or Innovation in Engineering & Management, 2(8) (2015), pp.228-232.
[5] Ristevski, B., & Chen, M. (2018). Big Data Analytics in Medicine and Healthcare. Journal of Integrative Bioinformatics, 0(0). doi: 10.1515/jib-2017-0030
[6] NIST. (2011). US Government Cloud Computing Technology Roadmap Volume II Release (Draft) Useful Information for Cloud Adopters. U.S. Department of Commerce. Retrieved from http://www.nist.gov/itl/cloud/upload/SP_500_293_volumeII.pdf on 5th September 2016.
[7] ENISA. (2015). Security Framework for Governmental Clouds. European Union Agency for Network and Information Security (ENISA), Retrieved from https://www.enisa.europa.eu/.../cloud...cloud-security/security...clouds/ on 5th August 2015
[8] Wardley, S. (2009). Maturity models for the cloud. Retrieved from http://blog.gardeviance.org/2008/12/maturitymodels-for-cloud.html on 25th March 2018
[9] Kaisler, S., Armour, F., Espinosa, J. A., & Money, W. (2013). Big data: Issues and challenges moving forward.46th Hawaii International Conference on System Sciences (HICSS) (pp. 995–1004).
[10] Alshboul, Y., Nepali, R.K.,Wang, Y.(2015),Big Data LifeCycle: Threats and Security Model, Retrieved from https://pdfs.semanticscholar.org/9ef4/fa7c505b92a5a0a9621fb5646b9d70739d27.pdf on 1st May 2018
[11] Zhang, X., Hu, Y., Xie, K., Zhang, W., Su, L., & Liu, M. (2015b). An evolutionary trend reversion model for stock trading rule discovery. Knowledge-Based Systems, 79,27–35.
[12] Galvic, B. (2015), Big Data Provenance: Challenges and Implications for Benchmarking, Retrieved from https://pdfs.semanticscholar.org/9ad4/3adc16c975deb5dfe2d5d1fec815906d3fc9.pdf on 1st May 2018
[13] Bilalli, B., Abello, A., Banet, T.D, Wrembel, R (2015), Towards Intelligent Data Analysis: The Metadata Challenge, Retrieved from http://www.essi.upc.edu/~aabello/publications/16.IoTBD.Besim.pdfon 1st May 2018
[14] Zhang, X., Hu, Y., Xie, K., Zhang, W., Su, L., & Liu, M. (2015). An evolutionary trend reversion model for Knowledge-Based Systems, doi: 10.1016/j.knosys.2014.08.010
[15] Chen, J., Chen, Y., Du, X., Li, C., Lu, J., Zhao, S., & Zhou, X. (2013). Big data challenge: a datamanagement perspective. Frontiers of Computer Science, 7(2), 157–164.
[16] Liu, A.Y., Wang, Q., & Qiang, H. (2015), Research on IT Architecture of Heterogeneous Big Data, Journal of Applied Science and Engineering, Vol. 18, No. 2, pp. 135142 (2015) DOI: 10.6180/jase.2015.18.2.05
[17] Edwards, R., & Fenwick, T. (2015). Digital analytics in professional work and learning. Studies in Continuing Education (pp. 1–15).
[18] Simonet, A., Fedak, G., & Ripeanu, M. (2015). Active Data: A programming model to man-age data life cycle across heterogeneous systems and infrastructures. Future Generation Computer Systems, 53,25–42
[19] Phillips-Wren, G., & Hoskisson, A. (2015). An analytical journey towards big data. Journal of Decision Systems, 24(1), 87–102
[20] Bhimani, A., & Willcocks, L. (2014). Digitization, Big Data and the transformation of ac-counting information. Accounting and Business Research, 44(4), 469–490
[21] Cloud Security Alliance (CSA) (2013), Expanded Top Ten Big Data Security and Privacy, Retrieved from https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Expanded_ Top_Ten_Big_Data_Security_and_Privacy_Challenges.pdf on 2nd May 2018
[22] D.P. Achariya, P.K. Ahamed(2016), A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 2, Retrieved from https://thesai.org/Downloads/Volume7No2/Paper_67-A_Survey_on_Big_Data_Analytics_Challenges.pdf on 1st May 2018
[23] Morgan Kaufmann, B., 2013. Chapter 5 – data governance for big data analytics: considerations for data policies and processes, in: D. Loshin (Ed.), big data Analytics., pp.pp. 39–48.
[24] Newell, S., & Marabelli, M. (2015). Strategic opportunities (and challenges) of algorithmic decision-making: A call for action on the long-term societal effects of 'datification'. The Journal of Strategic Information Systems, 24(1), 3-14. doi: 10.1016/j.jsis.2015.02.001
[25] Clarke, R. (2015). Big data, big risks. Information Systems Journal, 26(1), 77-90. doi: 10.1111/isj.12088