A Novel Approach to Big Data Management

Praveena Chaturvedi
Kanya Gurukul Campus,
Gurukula Kangri Vishwavidyalaya, Haridwar
Email: praveena_c1@rediffmail.co

***Abstract*: In the digital and computing world, information is generated and collected at a rate that rapidly exceeds the boundary range. The increase in the amount of data sources also increases the amount of the data acquired. Therefore storing and processing data become difficult and classical approaches remain incapable to do such transactions. By means of Big Data large amount of data with a wide range can be stored, managed and processed. In this paper we have devised a general Data Life Cycle Model that uses the technologies and terminologies of Big Data. The stages in the proposed model include acquisition, analysis, and storage.**

Keywords: Big data, big data management, acquisition, analysis, storage.

I. INTRODUCTION

The term 'Big Data' came in light for the very first time in 1998 in a Silicon Graphics (SGI) slide area by John Mashey with the title of "Big Data and the Next Wave of Infra Stress" [1]. The derivation of the term 'Big Data' is due to the fact that we are generating a huge amount of data every day. This data is known as Big Data [2]. Big Data is considered by the following three aspects:

(a) The data are abundant,

(b) The data cannot be considered into regular relational databases, and

(c) Data are generated, captured, and processed very quickly.

Conventionally, the data is stored in a highly structured way to exploit its informational contents. However, in current state the data volumes are determined by both unstructured and semi structured data. Therefore, end-to-end processing can be obstructed by the conversion between structured data in relational systems of database management and unstructured data for analytics.

Various studies in the literature show that big data has 3, 4 or 5 characteristics; 3 of whom are common at all (3 Vs): Volume, Velocity and Variety. Others are Veracity and Value (3, 4, 5, 6, and 7).

These 5 main characteristics are explained as follows:

- **Volume:**

  It denotes the size of the big data set. It is the most significant characteristic of big data.

- **Variety:**

  Various data come to the companies from numerous resources (internal or external). This data may be structured or unstructured.

- **Velocity:**

  The production rate of big data is called its velocity and it is very high. The heavy increase in data means that the data should be analyzed more swiftly. The faster the data increases, the faster the need for the data increases; therefore the process shows increase as well.

- **Veracity:**

  It is the correctness of the data. The data should be obtained from legal resources and it should be secure. Only authorized people should have the access authorization.

- **Value:**

  An outcome should be generated after all of the processes and the result should augment the process.

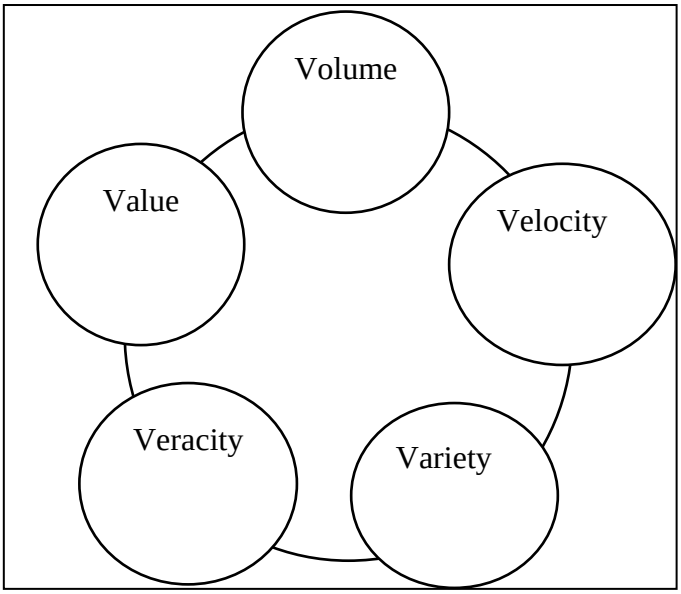

Figure 1: 5 Characteristics of Big Data (Elragal, 2014)

II. LIFE CYCLE MODEL OF DATA

The real importance of Big Data is when it is able to drive decision making for an organization. To enable such evidence-based decision making, organizations require efficient processes to turn high volumes data into meaningful understandings. The overall method of extracting visions from big data can be broken down into various four stages, shown in Fig. 2. This process is known as Data Life Cycle (DLC). The following section presents a general data life cycle that uses the technology and terminology of Big Data. The proposed data life cycle consists of the following stages: Acquisition (sub-phases are Extraction, Aggregation), Analysis, and Storage. The life cycle is called the AAS cycle for data processing.

(a) Acquisition

Data collection or generation is generally the first stage of any data life cycle. In process of data collection, special techniques are utilized to acquire raw data from a specific environment. The typical sources of the data are: User Generated Contents, Transactional Data, Scientific Data, and Web Data etc. [8]. Data generation is closely associated with the daily lives of people. These data are also similarly of low density and high value. The data acquisition phase is divided into two sub-phases: Extraction and Aggregation.

In number of cases, big data do not have proper relational structures. They often comprise of information of assorted types such as texts, images, tags, metadata, provenance data etc. It causes that big data cannot be simply abstracted using single data model. Countless strategies can be applied for dissimilar types of big data. For big unstructured data, they cannot be excellently understood and efficiently processed when in raw format. As such, information extraction procedures have been broadly functional to extract important and manageable structured data from the raw unstructured ones [9]. Finally, the big unstructured data are processed through solutions on the extracted structured data. The extracted data are summaries and sketches of the original unstructured data. There must be information loss after the transformation, reduction and aggregation. Solutions on how to efficiently reduce and transform big data are therefore very essential.

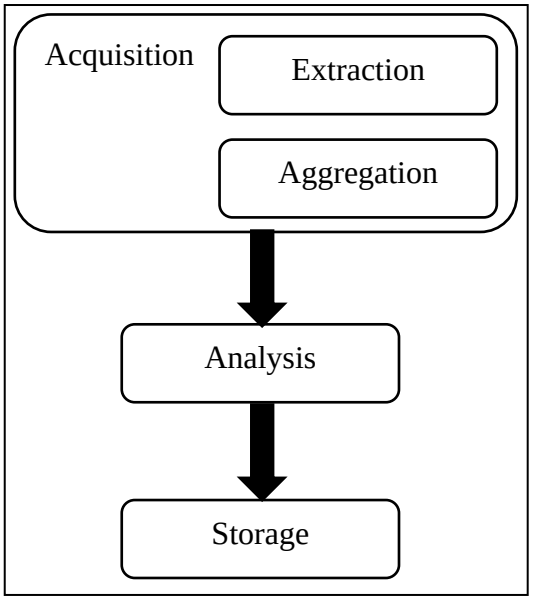

Figure 2: Data Life Cycle Model

(b) Analysis

Once we have the good way to analyze and mine the big data, it can bring us the big value. However, due to its noisy, dynamic, heterogeneous, inter-related and untrustworthy properties, the analysis and mining of the big data is very challenging [8].

Data analysis enables an organization to handle ample information that can affect the business. However, data analysis is challenging for various applications because of the complexity of the data that must be analyzed and the scalability of the underlying algorithms that support such processes [10]. Data analysis has two main objectives: to understand the relationships among features and to develop effective methods of data mining that can accurately predict future observations [11]. Various devices currently generate increasing amounts of data. Accordingly, the speed of the access and mining of both structured and unstructured data has increased over time [12]. Thus, techniques that can analyze such large amounts of data are necessary. Available analytical techniques include data mining, visualization, statistical analysis, and machine learning. For instance, data mining can automatically discover useful patterns in a large dataset.

(c) Storage

Data and its resources are composed and studied for storing, sharing, and publishing to benefit audiences, the public, tribal administrations, academicians, scholars, scientific partners, federal agencies, and other stakeholders (e.g., industries, communities, and the media)[13]. Large and widespread datasets must be stored and accomplished with reliability, availability, and easy accessibility. Storage infrastructures must suggest trustworthy space and a strong access interface that can not only analyze huge amounts of data, but also store, manage, and determine data with relational DBMS structures.

III. CONCLUSION

Big Data is going to endure growing during the next years, and each organization will have to manage much more volume of data every year. This data is going to be more diverse, larger, and faster. The objective of this paper is to describe, review, and reflect on big data. The paper first defined what is meant by big data to merger the divergent discourse on big data. To enhance the productivity of data management, we have developed a common Data Life Cycle Model that uses the technologies and terminologies of Big Data. The steps in this life cycle include acquisition, analysis, and storage. All these stages (collectively) transform raw data to published data.

References

[1] F. Diebold. On the Origin(s) and Development of the Term "Big Data". Pier working paper archive, Penn Institute for Economic Research, Department of Economics, University of Pennsylvania, 2012.
[2] D. Che, M. Safran, and Z. Peng, “From Big Data to Big Data Mining: challenges, issues, and opportunities,” in Database Systems for Advanced Applications, pp. 1–15, Springer, Berlin, Germany, 2013.
[3] Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of “big data” on cloud computing: Review and open research issues. Information Systems, 47, 98-115.
[4] Elragal, A. (2014). ERP and Big Data: The Inept Couple. Procedia Technology, 16, 242-249.
[5] Fadiya, S. O., Saydam, S., & Zira, V. V. (2014). Advancing big data for humanitarian needs. Procedia Engineering, 78, 88-95.
[6] Yang, C., Zhang, X., Zhong, C., Liu, C., Pei, J., Ramamohanarao, K., & Chen, J. (2014). A spatiotemporal compression based approach for efficient big data processing on Cloud. Journal of Computer and System Sciences, 80(8), 1563-1583.

[7] López, V., del Río, S., Benítez, J. M., & Herrera, F. (2015). Cost-sensitive linguistic fuzzy rule based classification systems under the Map Reduce framework for imbalanced big data. Fuzzy Sets and Systems, 258, 5-38.
[8] Jinchuan CHEN, Yueguo CHEN, Xiaoyong DU, Cuiping LI, Jiaheng LU , Suyun ZHAO, Xuan ZHOU, "Big data challenge: a data management perspective", Front. Comput. Sci., 2013, 7(2): 157–164, DOI 10.1007/s11704-013-3903-7
[9] Doan A, Naughton J F, Baid A, Chai X, Chen F, Chen T, Chu E, DeRose P, Gao B J, Gokhale C, Huang J, Shen W, Vuong B Q. The case for a structured approach to managing unstructured data. In: Proceedings of the 4th Biennial Conference on Innovative Data Systems Research. 2009.
[10] A. Labrinidis and H. Jagadish, "Challenges and opportunities with big data," Proceedings of the VLDB Endowment, vol. 5, no. 12, pp. 2032–2033, 2012.
[11] J. Fan and H. Liu, "Statistical analysis of big data on pharmacogenomics," Advanced Drug Delivery Reviews, vol. 65,no. 7, pp. 987–1000, 2013.
[12] D. E. O'Leary, "Artificial intelligence and big data," IEEE Intelligent Systems, vol. 28, no. 2, pp. 96–99, 2013.
[13] Nawsher Khan, Ibrar Yaqoob, Ibrahim Abaker Targio Hashem, Zakira Inayat, Waleed KamaleldinMahmoud Ali, Muhammad Alam, Muhammad Shiraz, and Abdullah Gani, "Big Data: Survey, Technologies, Opportunities, and Challenges", Hindawi Publishing Corporation.