Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 8, Issue, 04, pp.29317-29320, April, 2016

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

ISSN: 0975-833X

RESEARCH ARTICLE

CRITICAL ANALYSIS OF EXISTING BIG DATA ANALYTICS FRAME WORKS

***Belesti Melesse Asress and Dr. Tibebe Beshah**

Ethiopia

ARTICLE INFO

Article History:
Received 23rd January, 2016
Received in revised form
16th February, 2016
Accepted 09th March, 2016
Published online 26th April, 2016

Key words:

Big Data analytics,
Big Data analytics frameworks,
Unstructured data analysis,
Big Data architecture.

ABSTRACT

Recent technological advancements have led to a flood of data from various domains over the past few decades. Big Data incorporates large volume of structured, semi-structured, and unstructured data, which is beyond the processing capabilities of traditional databases. In addition to its huge volume, Big Data is commonly unstructured and requires more real-time analysis. On the other hand, the processing and analysis of Big Data plays a central role in decision making, forecasting, business analysis, product development, customer experience, and loyalty. Hence, organizations dealing with Big Data and analytics need to manage the challenges and opportunities related to datasets they have. The IT industry has responded by providing Big Data tools and technologies as well as approaches. However, many of the existing approaches and technologies experience noted limitations. In this paper, attempt has been made to examine the distinctive features of Big Data along the lines of the 3Vs (variety, volume, and velocity) using literature review and provide an understanding of the Big Data processing approaches. Furthermore, Various Big Data analytics frameworks that deal with Big Data analysis workloads were also investigated and analyzed against set of criteria. Finally, analysis and discussions of existing Big Data analytics frameworks along with a way forward approach is presented.

Citation: Belesti Melesse Asress and Dr. Tibebe Beshah, 2016. "Critical analysis of existing Big Data analytics frame works", *International Journal of Current Research*, 8, (04), 29317-29320.

INTRODUCTION

The emerging Big Data paradigm, owing to its broader impact, has profoundly transformed our society and will continue to attract diverse attentions from both technological experts and the public in general. It is obvious that we are living in a data flood era, evidenced by the overwhelming volume of data from a variety of sources and its growing rate of generation (Hu *et al*., 2014). Big Data plays a great role in multiple domains such as science, research, engineering, medicine, healthcare, finance, business, and ultimately society itself. It can be used for analyzing and predicting business trends, profit, and loss, and identifying real-time road traffic conditions, healthcare, Weather forecasting, and so on (Casado and Younas, 2014). Big Data encloses large volume of complex structured, semi-structured, and unstructured data, which is beyond the processing powers of conventional databases (Casado and Younas, 2014). Thus, any of these huge data sets produced in the data deluge may be considered Big Data. It is clear that they are too big, they move too fast, and they do not fit, generally, to the relational model strictures (Akerkar, 2013). Big Data is commonly unstructured and require more real-time analysis. This development calls for new system architectures for data acquisition, transmission, storage, and large-scale data processing techniques (Hu *et al*., 2014). As far back as 2001, Doug Laney's (Industry Analyst) expression of the now conventional definition of Big Data as the 3Vs of Big Data like volume, velocity and variety is described in (Kumar *et al*., 2014). Various works also defined the concepts.

- Volume: Volume refers to the size of the data to be processed. Volume of Big Data goes far beyond the accepted limits of megabytes or gigabytes and reaches the terabytes or even petabytes (Casado and Younas, 2014).
- Velocity: Velocity refers to the speed at which the data is being produced or the frequency with which it is delivered (Zadrozny and Kodali, 2013). The stress forced by high-velocity data streams will likely require a more real-time approach (Nugent *et al*., 2013).
- Variety: variety refers to the data form, i.e., structured, semi-structured, and unstructured (Hu *et al*., 2014). The attribute of variety takes account of the fact that Big Data is fuelled by diverse sources such as data warehouses, document stores, logs, click-streams, social media, sensors, mobile phones, and many others. These data vary in structure, volume, and format and due to their

Corresponding author: Belesti Melesse Asress
Ethiopia.

heterogeneity create serious challenges to storage, integration, and analytics (Akerkar, 2013).

Therefore it is easy to learn that Big Data refers to datasets which for one of many reasons (Volume, Velocity, or Variety as shown in Figure 1) do not fit a conventional relational database (Kumar *et al*., 2014). Unstructured data refers to information that either does not have a pre-defined data model or does not fit well into relational tables. Unstructured data is the most rapidly growing type of data, some samples could be imagery, sensors, telemetry, video, documents, log files, and email data files (Bakshi, 2012).

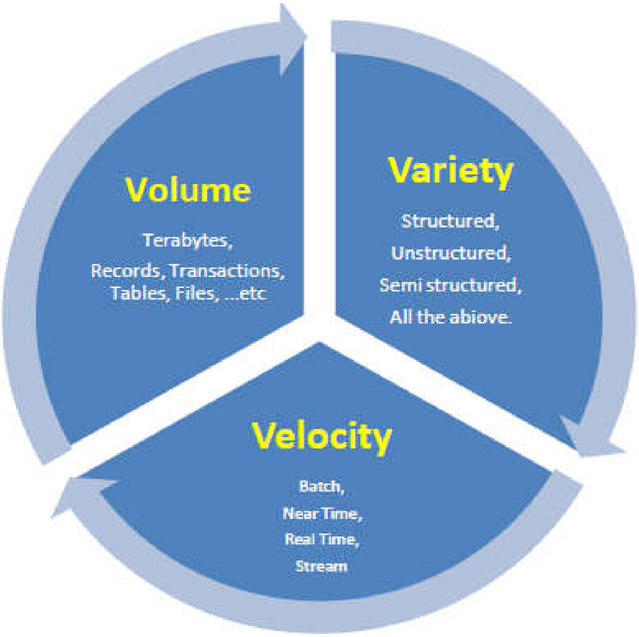

Figure 1. Characteristics of Big Data (Al-Barhamtoshy and Eassa, 2014)

While Big Data can produce extremely useful information, it also presents challenges with respect to how much data to store, how much this will cost, whether the data will be secure, and how long it must be stored (Michael and Miller, 2013). Big Data problems were mostly related to the presence of unstructured data, that is, information that either do not have a default schema or that do not fit to relational tables; it is therefore necessary to turn to analysis mechanisms for unstructured data, to address these problems (Zadrozny and Kodali,, 2013). Growth rate of data collected is challenging. This growth rate is very fast exceeding design capability to manage data effectively and also get relevant meaning for decision making. Consequently, many organizations are challenged to manage the Big Data they have. User’s requirements, technologies that are needed, and the design of the proposed system are challenges required to work with Big Data discovery (Al-Barhamtosh and Eassa, 2014). Every day, 2.5 quintillion bytes of data are created (Kumar *et al*., 2014). These data are generated from digital pictures, videos, posts to social media sites, intelligent sensors, purchase transaction records, cell phone GPS signals, to list some. There is no doubt that Big Data and particularly what we do with it has the potential to become a driving force for breakthrough and value creation (Akerkar, 2013). Big Data has also given organizations a new way to analyze and visualize their data effectively. Therefore, it is worth having to conduct a research on Big Data analytics.

MATERIALS AND METHODS

Overview

New discoveries don't materialize out of nowhere; they build upon the findings of previous experiments and investigations. A literature review shows how the investigation a researcher is conducting fits with what has gone before and puts it into context (Reading, 2015). In this section, literatures on Big Data architecture, Big Data storage, and unstructured data analytics are reviewed. The first phase of the review was to identify relevant literatures. Different searching techniques were used to obtain those literatures related to Big Data architecture, Big Data storage, and unstructured data analytics from electronic journals, thesis papers, databases, internet, conference proceedings, and books. The keywords used for searching were unstructured data analysis, Big Data concepts, Big Data frameworks, Big Data applications, Big Data architecture, unstructured data storage, Big Data solutions, Big Data challenges, Big Data Analytics, and Big Data opportunities. In the second phase, researches on Big Data analytics framework that were published since 2011 were selected for review. Then the selected literatures were thoroughly analyzed and investigated to understand the state of the art for the body of knowledge under review and to identify gaps that would be filled by future researchers. We investigated and analyzed the selected Big Data analytics frameworks that deal with Big Data analysis workloads against set of criteria. The criteria set to investigate and analyze those literatures are:

- capability to store and process unstructured data (Variety)
- capability to store and process huge volume of data (Volume)
- capability to handle data generated at high speed (Velocity)
- Capability to handle data having a combination of all the three attributes (3Vs)
- Requirement of expert knowledge in using the proposed Big Data analysis solution

Big Data analytics technologies

Different types of frameworks are required to run different types of analytics. A variety of workloads present in large-scale data processing enterprise. In order to attain a business goal, it needs to see a blend of workloads deployed: batch-oriented processing, OLTP (Online Transaction Processing), stream processing, and interactive ad-hoc query (Chandarana and Vijayalakshmi, 2014).

- Apache Hadoop: Apache Hadoop is a framework that allows handling distributed processing of Big Data across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each of them offering local computation and storage (Kumar *et al*., 2014). Hadoop is designed to scan large data set to produce results through a distributed and highly scalable batch processing systems.

Table 1. Summary of Results

Author(s)	Title	capability to store and process unstructured data (Variety)	capability to store and process huge volume of data (Volume)	capability to handle data generated at high speed (Velocity)	Capability to handle data having all the three attributes (3Vs)	Requirement of expert knowledge in using the proposed Big Data analysis solution
Das and Kumar (2013)	Big Data Analytics: A Framework for Unstructured Data Analysis	+	-	+	-	+
Boja, Pocovnicu, and Batagan (2012)	Distributed Parallel Architecture for "Big Data"	+	+	+	+	+
Kim, Kim, Park, Seo, and Lee (2013)	RUBA: Real-time Unstructured Big Data Analysis Framework	+	-	+	-	+
Al-Barhamtoshy and Eassa (2014)	A Data Analytic Framework for Unstructured Text	+	+	-	-	+
Herodotou, Lim, Luo, Borisov, Dong, Cetin, Babu (2011)	Starfish: A Selftuning System for Big Data Analytics	+	+	-	-	-
Ha, Back, and Ahn (2015)	MapReduce Functions to Analyze Sentiment Information from Social Big Data	+	+	-	-	+

Project Storm: Storm is an open source low latency processing stream processing system designed to integrate with existing queuing and bandwidth systems (Barlow, 2013). In order to carry out rigorous real-time analysis, Storm (Hadoop for real-time) has been developed. Storm is distributed real-time computation system developed and released as open source by Twitter (Casado and Younas, 2014).

- Apache Drill: Apache Drill is a distributed system for interactive ad-hoc analysis of large-scale datasets. Designed to manage up to petabytes of data spread across thousands of servers, the aim of Drill is to respond to ad-hoc queries in a low latency manner. Many times it occurs that human sits in front of business application and need to execute ad-hoc queries as per business needs. Query should not take more than few seconds to execute even at scale; sometimes users do not know which query to fire in advance; also users need to react to changing circumstances. Apache drill will provide the solution for all these issues (Chandarana and Vijayalakshmi, 2014).

Apache Hadoop is suited for workload where time is not critical feature whereas Project storm is well suited for data stream analysis in which analysis made is real time and Apache drill is best for interactive and ad-hoc analysis (Chandarana and Vijayalakshmi, 2014).

Big Data analytics literatures

In this section six attempts of proposing Big Data analytics formwork were presented and discussed. Das and Kumar (2013) developed a framework for analyzing unstructured data (Das and Kumar, 2013). Their proposed approach consists of acquiring unstructured data from public tweets of Twitter, storing the data in NOSQL database like HBase, and retrieving and analyzing the data. However, there are many limitations in their research. Data acquired from Twitter were stored in HBase after sentiment analysis. Therefore their system is unable to store unstructured data, and unable to process data in parallel during analysis. Furthermore, since HBase data requires java coding, business analysts who are supposed to use this solution need to have java knowledge and skills. Processing large datasets obtained from diverse sources is a challenging task as it requires tremendous storing and processing capabilities. Distributing the data across multiple processing units and parallel processing unit produces linear improved processing speeds. Boja, Pocovnicu, and Batagan (2012) investigated the problem of storing, processing and retrieving meaningful insight from petabytes of data. They developed a Distributed Parallel Architecture for Big Data that can be used to solve the analysis problem. When distributing the data is critical that each processing unit is assigned the same number of records and that all the related data sets reside on the same processing unit.

Using a multi-layer architecture to acquire, transform, load and analyze the data, ensures that each layer can use the best of breed for its specific task (Boja, 2012). But the authors focused only on the volume aspect of Big Data while Big Data has three attributes. Kim and Kim (2013) attempted to discover facts on Real-time Unstructured Big Data Analysis Framework (RUBA). In their research, they proposed a novel framework for real-time unstructured Big Data analysis, such as a movie, sound, text and image data. The proposed framework offers functions of a real-time analysis and dynamic modification for unstructured Big Data analysis. To find a specific data from real-time data stream, the continuous query processing was studied. In the existing query processing, data is stored in the database firstly and queries are executed whenever user requests. Therefore, they couldn't only analyze the Big Data but also adjust the analysis strategies in real-time (Kim *et al*., 2013). However, the proposed framework was not tested if it could address the volume attribute of Big Data. Al-Barhamtoshy and Eassa (2014) presented a systematic flow of unstructured data in organizations, and proposed an unstructured data framework for managing Big Data analysis (Al-Barhamtoshy and Eassa, 2014). They described unstructured data, collected data and stored data issues and challenges. Some of the major issues were identified, and many of technical points were forwarded; such as, Big Data contents, samples and challenges. Their study focused on developing unstructured data analysis, defining and designing methodologies, taken into account language processing, and machine translation. But the proposed framework ignored the velocity characteristic of Big Data.

Timely and cost-effective analytics on Big Data has come out as a key ingredient for success in many businesses, scientific and engineering disciplines, and government endeavors. The Hadoop software stack is a popular choice for Big Data analytics, but most practitioners of Big Data analytics lack the expertise to tune the system to get good performance in Hadoop. Herodotou, Lim, Luo, Borisov, Dong, Cetin, and Babu (2011) introduced Starfish, a self-tuning system for Big Data analytics (Herodotou *et al*., 2011). Starfish was developed on Hadoop while adapting to user needs and system workloads to bring good performance automatically, without any need for users to understand and operate the many tuning knobs in Hadoop. This approach enables Starfish to handle the significant interactions arising among choices made at different levels even though Starfish can't manage real time data analysis.

Ha, Back, and Ahn (2015) proposed a method to obtain sentiment information from various types of unstructured social media text data from social networks by using a parallel Hadoop Distributed File System (HDFS) to save social multimedia data and using Map Reduce functions for sentiment analysis (Ilkyu Ha and Byoungchul Ahn, 2015). The proposed method has successfully performed data collecting and data loading and maintained stable load balancing of memory and CPU resources during data processing by the HDFS system. The proposed method successfully processes data loading according to the increase in the number of data items. The system load is distributed to each node by parallel processing. When the proposed sentiment analysis functions have processed the data effectively, the system load is not concentrated on a single node but is evenly distributed among all nodes. However, the proposed architecture doesn't work for real time sentiment analysis of social media data via the Map Reduce functions.

RESULTS AND DISCUSSION

Here, various Big Data analytics frameworks which deal with Big Data analytics workloads are analyzed against set of criteria and summarized in Table 1 below. As a result, only 16.7% of the Big Data analytics frameworks are able to address all the three attributes of Big Data (Volume, Velocity, and Veracity). Regarding requirement of expert knowledge in using the proposed Big Data analysis solution, only 16.7% of the proposed frameworks fulfilled the criteria. However, none of the proposed frameworks or architectures satisfied both criteria, the three attributes of Big Data and requirement of expert knowledge in using the proposed solution. Therefore, this limitation is shared by all existing approaches.

Conclusion and Recommendation

In this research, a critical and in depth review of prior studies has been conducted to get understanding of the main Big Data processing approaches and to analyze Big Data analytics frameworks that deal with Big Data analysis workloads. Hence, we have identified that all existing approaches have one limitation in common. In other words, existing frameworks or architectures either fail to address all the three attributes of Big Data together or they require expert knowledge to work with. However, most practitioners of Big Data analytics lack the expertise to use the system. Therefore, a new approach is required to deal with Big Data analytics problems.

REFERENCES

Akerkar, R. 2013. Big Data computing, CRC Press.

Al-Barhamtoshy, H.M. and Eassa, F.E. 2014. A Data Analytic Framework for Unstructured Text. *Life Science Journal*, 11(10).

Bakshi, K. 2012. Considerations for Big Data: Architecture and approach. in Aerospace Conference, 2012 *IEEE*.

Barlow, M. 2013. Real-time Big Data analytics: emerging architecture. "O'Reilly Media, Inc.".

Boja, C., Pocovnicu, A. and Batagan, L. 2012. Distributed Parallel Architecture for" Big Data. *Informatica Economica,* 16(2): p. 116-127.

Casado, R. and M. Younas, 2014. Emerging trends and technologies in Big Data processing. Concurrency and Computation: Practice and Experience.

Chandarana, P. and Vijayalakshmi, M. 2014. Big Data analytics frameworks. in Circuits, Systems, Communication and Information Technology Applications (CSCITA), 2014 International Conference on. *IEEE*.

Das, T. and Kumar, P.M. 2013. Big Data analytics: A framework for unstructured data analysis. *International Journal of Engineering Science & Technology,* 5(1): p. 153.

Herodotou, H., *et al*. 2011. Starfish: A Self-tuning System for Big Data Analytics. in CIDR.

Hu, H., *et al*., 2014. Towards Scalable Systems for Big Data Analytics: A Technology Tutorial. Access, *IEEE*.

Ilkyu Ha, B.B. and Byoungchul Ahn, 2015. MapReduce Functions to Analyze Sentiment Information from Social Big Data. *International Journal of Distributed Sensor Networks,* 2015.

Kim, J., *et al*. 2013. RUBA: Real-time unstructured Big Data analysis framework. in ICT Convergence (ICTC), 2013 International Conference on. *IEEE*.

Kumar, R., *et al*. 2014. Apache Hadoop, NoSQL and NewSQL Solutions of Big Data. *International Journal of Advance Foundation and Research in Science & Engineering (IJAFRSE),* 1(6): p. 28-36.

Michael, K. and Miller, K.W. 2013. Big Data: New Opportunities and New Challenges (Guest editors' introduction). *Computer*, 46(6): p. 22-24.

Nugent, A., Halper, F. and Kaufman, M. 2013. Big Data for dummies, John Wiley & Sons.

Reading, U.O. 2015. Study Advice. (cited 2015; Available from: http://www.reading.ac.uk/internal/studyadvice/StudyResources/Essays/sta-startinglitreview.aspx.

Zadrozny, P. and Kodali, R. 2013. Big Data and splunk, in Big Data Analytics Using Splunk, *Springer*. p. 1-7.

****