International Journal of Computer Engineering in Research Trends
Multidisciplinary, Open Access, Peer-Reviewed and fully refereed
Research Paper Volume-5, Issue-8, 2018 Regular Edition E-ISSN: 2349-7084

Literature survey on Big Data Analytics

Mallikarjuna Reddy Beram*

*Lead I, UST Global, Bengaluru, Karnataka 560066, India,
Email: berammkr@gmail.com

Available online at: http://www.ijcert.org

Received:25/07/2018, Revised: 17/08/2018, Accepted: 22/08/2018, Published: 28/08/2018

Abstract: In this latest era of computer's, a enormous amount of data is available to managerial. Big data doesn't only refer to datasets that are big, but also high in volume, velocity, value, veracity and variety, which is tough to handle using old-fashioned tools and methods. Due to rapid growth of such data, some ways are necessary to found to get important knowledge and values from these data sets. Also, decision makers need to gain some valuable vision from such big and endlessly changing data, ranging from daily transactions to customer connections and data of social network. Such vision can be given using Big Data Analytics, which is the application of Advanced Analytics Technique on big data. This paper aims to literature of some of the analytics methods and tools which can be applied to big data, as well as the charge provided by the applications of big data analytics in different decision domain. We have discussed the different processing techniques for big data and processing steps as well.

Keywords: Big Data Analytics, Advanced Analytics, Transactions Data, Data processing.

---

1. Introduction

Big data analytics is the often complex process of examining big data to uncover information -- such as hidden patterns, correlations, market trends and customer preferences and that can help organizations make informed business decisions. On a broad scale, data analytics technologies and techniques give organizations a way to analyze data sets and gather new information. Business intelligence (BI) queries answer basic questions about business operations and performance. Big data analytics is a form of advanced analytics, which involve complex applications with elements such as predictive models, statistical algorithms and what-if analysis powered by analytics systems. Organizations can use big data analytics systems and software to make data-driven decisions that can improve business-related outcomes. The benefits may include more effective marketing, new revenue opportunities, customer personalization and improved operational efficiency. With an effective strategy, these benefits can provide competitive advantages over rivals.

Data analysts, data scientists, predictive modelers, statisticians and other analytics professionals collect, process, clean and analyze growing volumes of structured transaction data as well as other forms of data not used by conventional BI and analytics programs.

1.1 Characterstics Of Big Data

The characteristics of the big data depend on the three factors which include Data Velocity, Data Volume and Data Variety. Big Data is not just about the size of data but also includes data variety and data velocity. These are the five V's of the big data. [1]

Volume: Big data denotes its massive character, i.e. a huge amount of information involved. Data is evergrowing day by day of all types ever Kilo Byte, Mega Byte, Peta Byte, Yotta Byte, Zetta Byte, Tera Byte of information. The data results into massive files. Excessive volume of information is main issues of storage. This main issue is resolved by reducing storage value. Data volumes are expected to grow more than 50 times by 2020.

Variety: Data sources (even in the same field or in distinct) are extremely heterogeneous. The files comes in various formats and of any type, it may be unstructured or structured such as text, audio, log files, videos and more.

 https://doi.org/10.22362/ijcert/2018/v5/i8/v5i801

The varieties are endless, and the data enters the network without having been quantified or qualified in any way.[2]

Velocity: The data comes at high speed. Sometimes one minute is too late so big data is time sensitive. Most organizations data velocity is main challenge. The credit card transactions and social media messages done in millisecond and data generated by this putting in to databases.

Value: Which addresses the requirement for valuation of enterprise data? It is a most important V in big data. Value is main buzz for big data because it is important for IT infrastructure system, businesses to store large amount of values in database.

Veracity: The increase in the range of values typical of a large data set. When we dealing with high volume, velocity and variety of data, the all of data are not going 100% correct, there will be dirty data. Big data and analytics technologies work with these types of data.

2. Steps Of The Big Data Analytics Process

Step1: Data professionals collect data from a variety of different sources. Often, it is a mix of semi structured and unstructured data. While each organization will use different data streams, some common sources include: internet clickstream data; web server logs; cloud applications; mobile applications; social media content; text from customer emails and survey responses; mobile phone records; and machine data captured by sensors connected to the internet of things (IoT).

Step2: Data is prepared and processed. After data is collected and stored in a data warehouse or data lake, data professionals must organize, configure and partition the data properly for analytical queries. Thorough data preparation and processing makes for higher performance from analytical queries.

Step3: Data is cleansed to improve its quality. Data professionals scrub the data using scripting tools or data quality software. They look for any errors or inconsistencies, such as duplications or formatting mistakes, and organize and tidy up the data.

Step4: The collected, processed and cleaned data is analyzed with analytics software. This includes tools for: data mining, which sifts through data sets in search of patterns and relationships, predictive analytics, which builds models to forecast customer behavior and other future actions, scenarios and trends, machine learning, which taps various algorithms to analyze large data sets deep learning, which is a more advanced offshoot of machine learning text mining and statistical analysis software artificial intelligence (AI) mainstream business intelligence software data visualization tools.

3.Methods For Big Data Processing

With the evolution of technology and therefore the increased multitudes of knowledge flowing in and out of organizations daily, there has become a necessity for faster and more efficient ways of analyzing such data. Having piles of knowledge available is not any longer enough to create efficient decisions at the proper time. Such data sets cannot be easily analyzed with traditional data management and analysis techniques and infrastructures. Therefore, there arises a necessity for brand spanking new tools and methods specialized for giant data analytics, in addition because the required architectures for storing and managing such data. Accordingly, the emergence of huge data has an effect on everything from the information itself and its collection, to the processing, to the final extracted decisions. Consequently, [8] proposed the large – Data, Analytics, and Decisions (B-DAD) framework which includes the large data analytics tools and methods into the choice making process [8]. The framework maps the various big data storage, management, and recessing tools, analytics tools and methods, and visualization and evaluation tools to the various phases of the choice making process. Hence, the changes related to big data analytics are reflected in three main areas: big data storage and architecture, data and analytics processing, and, finally, the large data analyses which may be applied for knowledge discovery and informed deciding. Each area is further discussed during this section. However, since big data continues to be evolving as a crucial field of research, and new findings and tools are constantly developing, this section isn't exhaustive of all the chances, and focuses on providing a general idea, instead of an inventory of all potential opportunities and technologies.

Big data analytics is the process of using analysis algorithms running on powerful supporting platforms to uncover potentials concealed in big data, such as hidden patterns or unknown correlations. According to the processing time requirement, big data analytics can be categorized into two alternative paradigms.

 https://doi.org/10.22362/ijcert/2018/v5/i8/v5i801

3.1 Batch Processing: In the batch-processing paradigm, data are first stored and then analyzed. Map Reduce has become the dominant batch-processing model. The core idea of Map Reduce is that data are first divided into small chunks. Next, these chunks are processed in parallel and in a distributed manner to generate intermediate

results. The final result is derived by aggregating all the intermediate results.[12 This model schedules computation resources close to data location, which avoids the communication overhead of data transmission. The Map Reduce model is simple and widely applied in bioinformatics, web mining, and machine learning

3.2 Streaming Processing: The start point for the streaming processing paradigm is the assumption that the potential value of data depends on data freshness. Thus, the streaming processing paradigm analyzes data as soon as possible to derive its results. In this paradigm, data arrives in a stream. In its continuous arrival, because the stream is fast and carries enormous volume, only a small portion of the stream is stored in limited memory.

One or few passes over the stream are made to find approximation results. Streaming processing theory and technology have been studied for decades. Representative open source systems include Spark, Storm. [9] The streaming processing paradigm is used for online applications, commonly at the second, or even millisecond, level.

4. Big Data Analytical Tools

There are varieties of tools available in the current market. A few of the eminent tools are briefly described below:

• **Apache Spark:** Apache Spark's "key point of this open source Big Data tool is it fills the gaps of Apache Hadoop concerning data processing" (Verma, 2018). Data processing is done much faster than the traditional disk processing. It facilitates distributed task transmission and scheduling. This software was originally developed at the University of California Apache Spark (2018).

• **Hadoop:** This is one of the most prominent tools used for data analytics, which can process data in large scales. It is 100% open source and be run on a cloud infrastructure as well. It is a collection of open-source software, which solves massive data problems using a network of computers. Hadoop was released in 2006 Apache Hadoop (2018).

• **Mongo DB:** is a cross platform object-oriented database program, which is free and open source. The development of Mongo DB software began in 2007 by the 10gen software company now known as Mongo DB (2018). According to Zakir (2015), this software is also considered as one of the most popular NoSQL databases.

• **Qlik:** Qlik (2018) was previously known as QlikTech and was founded in Sweden in 1993 as a software company in business intelligence. It is a user-friendly tool, which allows for instant report generation.

• **Rapid Miner:** was developed by Rapid Miner company, and it was previously known as YALE (Yet another Learning Environment). Rapid Miner is a data science software platform that is used for business and commercial applications supporting machine-learning process as presented by Rapid Miner (2018).

• **SAP:** is enterprise software with the domain of providing business intelligence solutions and collaborative planning, supported by predictive analysis and machine learning technology. Its key features include data visualization, reporting and analysis, mobile data analytics and interactive role-based dashboards as stated by SAP Analytics Cloud (2017)

• **SAS:** is a collection of software that mines data, makes necessary changes, manages, and retrieves data from different sources whereby statistical analysis is performed on this data. Advanced options are provided to the users as well as graphical user interface for non-technical users as mentioned by SAS Software (2018).

• **Tableau:** is a software that has capabilities to produce data visualization products with its focus on business intelligence. Tableau Software (2018) also has mapping functionality.

Big Data Analytics Tools in Use

Based on the responses received, the following tools were listed as being used currently by businesses within the capital city of the Small Island State.

• **ArcGIS:** is an information system that deals with geographical and spatial data.

• **Clarity:** is an integrated modular management information system for business that enables data entry, analysis, and report generation.

• **Cognos:** is a data analytical tool for business intelligence and performance management.

• **Crystal Reports:** SAP crystal reports is business intelligence software, which is suitable for small and medium sized businesses. Some of the features of this software include writing customized reports from multiple data sources, visualizing data in dashboards and scorecards, displaying key performance indicators, and

 https://doi.org/10.22362/ijcert/2018/v5/i8/v5i801

other metrics in relation to project or department performance as per

SAP Crystal Reports Software (2018).

• DNS Analytics allows for the collection and analysis of DNS (Domain Name Servers) traffic on a particular computer network, assisting in identifying threats and malware. It allows visual representation of data in form of bar graphs and tables enabling the user to monitor the domains data logs as per DNS Made Easy (2017).

• **Excel:** Microsoft Excel allows for manipulation of numbers using formulas and functions. This software is used by nearly all businesses in the Small Island State for electronic record keeping.

• **Google Analytics** is a fermium web analytics service offered by Google that tracks and reports website traffic, as described by Google Analytics (2018).

• **Heat map:** does Heat map (2018) mention a graphical representation of data that uses a system of color-coding to represent different values as. This is used to track user behavior on the web, such as identifying the number of clicks or scrolls on a website.

• **Oracle analytics:** has the entire core "capabilities and languages on a powerful in-database architecture" which includes data mining algorithms implemented in the database as presented by Big Data Analytics Advanced Analytics in Oracle Database (2013).

• **SAP:** as mentioned in the earlier section, is enterprise software with the domain of providing business intelligence solutions and collaborative planning, supported by predictive analysis and machine learning technology. Its key features include data visualization, reporting and analysis, mobile data analytics and interactive role- based dashboards as presented by SAP Analytics Cloud (2017).

• **SAS:** as mentioned in the earlier section, is a collection of software that mines data, makes necessary changes, manages, and retrieves data from different sources whereby statistical analysis is performed on this data. Advanced options are provided to the users as well as graphical user interface for non-technical users as per SAS Software (2018).

• **SQL Server:** is a product of Microsoft, a relational database management system that has the primary function of storing and retrieving data when being requested by other software applications as presented by Microsoft SQL Server (2018).

• **TABLEAU:** as mentioned in the earlier section, is software that has capabilities to produce data visualization products with its focus on business intelligence. It also has a mapping functionality as per Tableau Software (2018).

• **Webmaster:** reveals the way Google views a website online and if there are any problems with the site, then webmaster uses its tools to fix such problems as presented by Difference between Google Analytics and Google Webmaster Tools (2018).

5. Benefits Of Big Data Analytics

With the use of analytical tools, the companies / businesses stated the following benefits:

• Using Big Data tools provides initiative, and it is great for modern day analysis.

• Use of tools such as tableau gives visual analysis. This will help in easy interpretation of the result, which can be easily understood by all stakeholders. It further makes it easier to understand subtle trends with in the datasets with means of visual aids.

• Increases the efficiency in business and enables better decisions to be made for the future growth and direction of the business.

• Using Big Data analytics assists companies to gain competitive advantage and outdo their competitors. The competitor companies or any new company joining the industry to derive schemes to maintain value and enhance their business processes with inventive ideas can use accurately analyzed data.

• Supports the businesses in their objectives for new developments and progression opportunities by consolidating and analyzing industry data. These businesses have sufficient data about the items and administrations, purchasers and providers, customer inclinations that can be evaluated. It also adds to profitability of the business. As stated by Farah, et al. (2017) allows for personalization of customer's needs, wants, and contribution to the profit of the organization.

• Enhances and improves businesses processes. Stock is easily adjusted by the vendors based on the predictive analysis and models based on the search trends from the web or data from social sites. Predictive analytics empowers users to have more confidence in areas of forecasting

• It provides a clear understanding of the data in the system. It provides a clearly understanding and representation of the available data which allows detailed decision making

 https://doi.org/10.22362/ijcert/2018/v5/i8/v5i801

• Provides a better insight about users and their actions

• Provides an easier method of capturing data, which can later be used in marketing plan

• Huge amounts of data are analyzed in a short amount of time, hence its time saving

• Makes it easier to collaborate, sharing analytics with clear facts and figures included with visual aids

• Real-Term and Long-Term Application usage reporting which allows monitoring of customer usage and helps in troubleshooting internet issues.

• Security attacks and vulnerabilities are identified based on the user traffic and host behavior such as spamming, Denial of Service attacks, Port-Scans

• Allows organizing and displaying data so that it adds meaning to it.

• Better interpretation of data to stakeholders and people who do not have experience with complex datasets to make efficient decisions.

6. Conclusion And Futurework

Big data is having the impact on the technological and the business world in the modern era. As data is growing, similarly technology also is advancing with new technologies. Business faces the challenges with huge amount of data and main aim is to the competitive advantage and business intelligence. Business can use the right tools and technologies for mine the relative patterns and uncover hidden insights. The objective of this paper is mainly determined the big data concepts and current analytical tools to be used and processing techniques to process the big data.

We would like to recommend future researchers to expand the scope of their research and include companies and businesses from different geographical locations. Whenever numbers of participants are increasing and will get more precise facts when collecting and analyzing the data.

References

[1] M.H.Padgavankar, Dr.S.R.Gupta, Big Data Storage and Challenges, M.H.Padgavankar, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 2218-2223.

[2]. Sabia, Sheetal Kalra, Applications of big Data: Current Status and Future Scope, International Journal on Advanced Computer Theory and Engineering (IJACTE), , Volume -3, Issue -5, 2014, ISSN 2319-2526.

[3] Apache Hadoop. What Is Apache Hadoop?, 2014. http://hadoop.apache.org/, accessed April 2014.

[4] H S. Bhosale1, Prof. D. P. Gadekar2, A Review Paper on Big Data and Hadoop, International Journal ofScientific and Research Publications, 4(10),2014.

[5] C.Jin, R.Liu, Z.Chen, Alok Choudhary, A Scalable Hierarchical Clustering Algorithm Using Spark, IEEE,

[6] Christos Doulkeridis, Kjetil , A Survey of Large-Scale Analytical Query Processing in MapReduce, TheVLDB Journal manuscript No.5.

[7]. Lekha R.Nair, DR. Sujala,D.Shetty, streaming Twitter Data Analysis Using Spark For Effective Job Search, Journal of Theoretical and Applied Information Technology ,. Vol.80. No. 2 2005 – 2015.

[8] Satish Gopalani,Rohan Arora, Comparing Apache Spark and Map Reduce with Performance Analysis using K-Means, International Journal of Computer Applications Volume 113 – No. 1, March 2015. (0975 –8887)

[9] D. Rajasekar, C. Dhanamani, S. K. Sandhya, A Survey on Big Data Concepts and Tools

[10] Suresh Lakavath, Ramlal Naik L, A Big Data Hadoop Architecture for Online Analysis, International Journal of Computer Science and Information Technology & Security (IJCSITS), Vol. 4, No.6, December 2014, ISSN: 2249-9555.

[11] M. Dhavapriya, N. Yasodha, Big Data Analytics: Challenges and Solutions Using Hadoop, Map Reduce and Big Table, International Journal of Computer Science Trends and Technology (IJCST) – Volume 4 Issue 1, Jan - Feb 2016

[12] H.HU1, Y. WEN 2 , TAT-SENG CHUA1, AND XUELONG LI 3, Toward Scalable Systems for Big Data Analytics, A Technology Tutorial, IEEE, 2 ,655-687, 2014.

[13] Ambika P R, Dr. K.N. Narasimha Murthy, Sowmya Naik PT, Aparna J S, Big Data: Towards Next Generation Analytics, International Journal of Innovative Research in Computer and Communication Engineering. Vol.3, Special Issue 5, May 2015.50 | P a g e

[14]. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauley, M. Franklin, S. Shenker, and I. Stoica. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. Technical Report UCB/EECS-2011-82, EECS Department, University of California, Berkeley, 2011

[15] Reynold Xin, Joshua Rosen, Matei, Zaharia, Michael J. Franklin, Scott Shenker, Ion Stoica. Shark: SQL and Rich Analytics at Scale. SIGMOD 2013. June 2013.

