International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 05 | May 2019 www.irjet.net p-ISSN: 2395-0072

A Comparative Study on Big Data Analytics Approaches and Tools

Sapna[1], Umesh Goel[2], Pankaj Sharma[3]

[1]*Research Scholar, Dept. of Computer Science & Engineering, DITMR Faridabad, India*
[2]*Assistant Professor, Dept. of Computer Science & Engineering, DITMR Faridabad, India*
[3]*Assistant Professor, Dept. of Computer Science & Engineering, DITMR Faridabad, India*

Abstract - *In this paper we have done a comparative study of various methodologies (technologies/tools) based on the certain parameters to check optimal solution which on is best for the concern organization. To Compute the data it's need tool and technique. Having data bigger consequently requires different approaches, techniques, tools & architectures to manage the data in a better way. Big data technologies provide more accurate analysis which help in decision making. To manage and process huge volume of structured semi-structured and unstructured data you would require an infrastructure that can secure, privacy and protect the data. The aim of this paper is to identify different Big Data strategies a company may implement and provide a set of organizational contingency factors that influence strategy choice. In order to do so, we reviewed existing literature in the fields of Big Data analytics tools and techniques in choosing a suitable Big Data approach. We find that while every strategy can be beneficial under certain corporate circumstances.*

The paper also evaluates the difference in the challenges faced by a small organization as compared to a medium or large scale operation and therefore the differences in their approach and treatment of BIG DATA. A number of application examples of implementation of BIG DATA across industries varying in strategy, product and processes have been presented.

1. INTRODUCTION

Big Data technologies are transforming the way data is used to be analyzed. One reason is the massive amount of data that is being generated from different sources such as social networks, sensors, search engines, banks, telecommunication and web, handling this massive amount of data take us in the era of Big Data.

Data is everywhere, from social sciences to physical science, business and commercial world, for example, digitizing the past fifty year's newspapers will results the massive amount of data, in astronomy storing billions of astronomical objects, in biology storing genes, proteins and small molecules results in massive amounts of data. In business world such as handling millions of call data records in telecommunication, handling millions of transactions in banking and handling millions of transactions for multinational grocery store results in large data sets. Analyzing these large datasets and getting out meaningful information from it is a challenging in itself. By intelligently using the information in and around them, organizations are able to improve their decision-making and better realize their objectives [1], [2]. Some authors even claim that organizations may lose competitiveness by not systematically analyzing the available information [3]. However, to obtain the desired insights, data need to be sourced, stored, and analyzed [4], [5]. During the past years, accessing and processing the collected, voluminous, and heterogeneous amounts of data has become increasingly time consuming and complex [6].

1.1 Big Data

Big Data can be described in 3 V's such as variety, volume and velocity [3].

Variety : Data has different variations, for example semi-structured or unstructured, such as data, generated from web sites, social networks, emails, sensors and web logs is unstructured. Structured data refers to as data generated in result of conversion from call data record to tabular format in order to calculate the monetary value out of it or banks transactions data or data generated from the airline ticketing system are different varieties in the data.

Volume: Volume refers to the amount of data or size of the data set. Nowadays figures are in Tera and Peta bytes. For instance Airbus can generate half of terabytes of data in one flight [4].

Velocity: Velocity refers to the speed of data generation which is very fast nowadays. For example weather sensors are kept on generating data as new updates comes, Twitter is generating data at 9100 tweets per second and on Facebook users is sending 3 million messages to each other every 20 minutes [5].

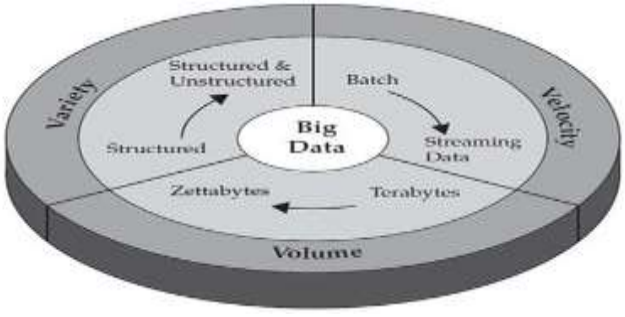

Fig-1: Characterizes Big Data by its volume, Velocity and veriety or V[3]

1.2 Big Data Technologies

Apache Hadoop : Hadoop was initially inspired by papers published by Google outlining its approach to handling an avalanche of data, and has since become the de facto standard for storing, processing and analyzing hundreds of terabytes, and even petabytes of data. Apache Hadoop is 100% open source, and pioneered a fundamentally new way of storing and processing data. Instead of relying on expensive, proprietary hardware and different systems to store and process data, Hadoop enables distributed parallel processing of huge amounts of data across inexpensive, industry-standard servers that both store and process the data, and can scale without limits. With Hadoop, no data is too big. And in today's hyper-connected world where more and more data is being created every day, Hadoop's breakthrough advantages mean that businesses and organizations can now find value in data that was recently considered useless. Hadoop can handle all types of data from disparate systems: structured, unstructured, log files, pictures, audio files, communications records, email – just about anything you can think of, regardless of its native format. Even when different types of data have been stored in unrelated systems, you can dump it all into your Hadoop cluster with no prior need for a schema. In other words, you don't need to know how you intend to query your data before you store it; Hadoop lets you decide later and over time can reveal questions you never even thought to ask.

Cloudera: Cloudera founded in 2008 and was the first enterprise distributors of Apache Hadoop and other Big Data technologies. The contributions of Cloudera to Big Data world is the abstraction over different Big Data technologies such as Hadoop, Hive, HCatalog [15] etc., which provides easiness to users to use these technologies without going into technical details. Cloudera solutions are ready to install on any commodity hardware which hides the technical details of compiling and configuring the Big Data technologies and provides the system management such as configuration, deployment, security management, diagnostics, operational reports generation etc. Cloudera is at forefront of providing back-end solutions for Big Data exploration and analysis but does not provide any tool or framework which actually can be used on top of these technologies especially for non-expert users [16].

Cassandra: For a number of years, modern businesses have been looking for an alternative to the legacy relational model for storing and accessing data. Information management requirements have undergone subtle to radical changes in these organizations, such that the practice of forcing the proverbial square peg into the round hole is no longer a viable solution. The rising interest of using NoSQL databases in the enterprise has demonstrated that developers and CTOs alike have recognized the fact that many progressive companies are now using NoSQL to manage both operational and analytical data. Businesses that were early adopters of NoSQL already have progressed to the point where NoSQL is powering a number of their key production systems, with both new development and RDBMS replacement projects being run in parallel. The release of Cassandra 1.0 in October 2011 is a significant milestone, both in the history of Cassandra itself and in the NoSQL data management movement in general. It is atypical for companies to run pre-1.0 software in production, yet because of a need for the features Cassandra offers, that is exactly what many businesses have been doing. With the release of 1.0, companies with software policies that negate the use of non-GA software can now also begin to enjoy the benefits Cassandra provides. Apache Cassandra is one of the pillars behind Facebook's massive success, as it allows processing structured data sets distributed across huge number of nodes across the globe. It works well under heavy workloads due to its architecture without single points of failure and boasts unique capabilities no other NoSQL or relational DB. The Apache Cassandra database is the right choice when you need scalability and high availability without compromising performance.

2. LITERATURE REVIEW

The characteristics of Big Data were first described in 2001, when Laney [4] identified three key attributes of large data amounts: high variety, volume, and velocity. To date, these attributes have become the defining characteristics of Big Data. However, contemporary authors and business specialists enlarged these defining characteristics with further aspects such as dedicated storage, management, and analysis techniques [8], [9], [10]. Further amendments to the definition include the addition of a fourth V, veracity, by IBM [11], emphasizing the aspect of data quality. Taking these different extensions of the original definition into account, we define Big Data as a phenomenon characterized by an ongoing increase in volume, variety, velocity, and veracity of data that requires advanced techniques and technologies to capture, store, distribute, manage, and analyze these data. The economic potential of Big Data is as diverse as the data itself and the key driver for organizations to adopt Big Data analytics. There are four Big Data categories organizations can leverage: external structured data such as Global Positioning System (GPS) or credit history data, internal structured data such as Customer Relationship Management (CRM) or inventory data, external unstructured data such as Facebook or Twitter posts, as well as internal unstructured data such as text documents and sensor data [12]. All four categories have specific characteristics that certain Big Data strategies may address better than others. Application fields of Big Data cover a wide range of industries and businesses. Suggestions range from health care (reduction of costs resulting from over-and under-treatment) with a 300 billion annual potential, to the public sector and e–government (more efficiently

collection of taxes and service quality improvement from education to unemployment offices) with a 250 billion annual potential, over e-commerce, marketing and merchandising (better understanding of consumers with respect to product and price preferences) with a potential of 60% increase in operating margins [8], [9]. These fields' potentials are unlocked by the application of different Big Data techniques such as crowd sourcing, data fusion and data integration, natural language processing, network analysis, predictive modeling, simulation, and visualization. Thus, the application possibilities and economic potentials of Big Data technologies are enormous and executives should assess whether and how they could make use of these potentials.

Mark Beyer, Douglas Laney Published on 21 June 2012 "Big data" warrants innovative processing solutions for a variety of new and existing data to provide real business benefits. But processing large volumes or wide varieties of data remains merely a technological solution unless it is tied to business goals and objectives.

Ming Ke, Yuxin Shi published on September 17, 2014 "Big Data, Big Change: In the Financial Management" In recent years, "Big Data" has attracted increasing attention. It has already proved its importance and value in several areas, such as aerospace research, biomedicine, and so on. In "Big Data" era, financial work which is dominated by transaction, business record, business accounting and predictions may spring to life. This paper makes an analysis about what change that "Big Data" brings to Accounting Data Processing, Comprehensive Budget Management, and Management Accounting through affecting the idea, function, mode, and method of financial management. Then the paper states the challenges that "Big Data" brings to enterprise aiming to illustrate that only through fostering strengths and circumventing weaknesses can an enterprise remain invincible in "Big Data" era.

Aicha Ben Salem, Faouzi Boufares, Sebastiao Correia published on April 2014 **"**Semantic Recognition of a Data Structure in Big-Data" In fact, good governance data allows improved interactions between employees of one or more organizations. Data quality represents a great challenge because the cost of non-quality can be very high. Therefore the use of data quality becomes an absolute necessity within an organization. To improve the data quality in a Big-Data source, our purpose is to add semantics to data and help user to recognize the Big-Data schema. The originality of this approach lies in the semantic aspect it offers. It detects issues in data and proposes a data schema by applying a se-mantic data profiling

Pwint Phyu Khine, Wang Zhao Shun published on March 13, 2017 "Big Data for Organizations: A Review" Big data challenges current information technologies (IT landscape) while promising a more competitive and efficient contributions to business organizations. What big data can contribute to is what organizations have been wanted for a long time ago. This paper presents the nature of big data and how organizations can advance their systems with big data technologies. By improving the efficiency and effectiveness of organizations, people can benefit the can take advantages of a more convenient life contributed by Information Technology".

Big data is a very wide and multi-disciplinary field which requires the collaboration from different research areas and organizations from various sources. Big data may change the traditional ETL process into Extract-Load-Transform (ELT) process as big data give more advantages in moving algorithms near where the data exist. Like other information systems, the success of big data projects depend on organizational resistance to change.

3. PROBLEM DISCUSSION

The volume of data is growing rapidly and on a global scale. Traditional data management tools cannot analyze the big datasets produced today. Therefore, Big Data is expected to transform the modern Tools and technique.

Failing to adapt to a business environment where Big Data is utilized will give way for competitors who do adapt. Embracing Big Data can give organizations competitive advantage and growth. Although Big Data represents the "ultimate opportunity", it also represents the "ultimate challenge". Examples of challenges are lack of skills and adequate technology. Big Data analytics can improve their performance, a statement verified by Brown et al. (2011) who argues that organizations need to address considerable challenges to be able to seize the potential with Big Data. Such statements show that there is a gap between the positive prospects of Big Data and the actual knowledge and use of Big Data in organizations today.

From the above perspective study will also address the issue of how the perceived gap can be decreased and select appropriate tools and technique for the organization. This part should be seen as an explanatory extension of the result from the above paragraphs.

3.1 Problem Formulation

Firstly, we have chosen to study Big Data in a wider perspective by not looking at specific countries or sectors. The surveys used in this study are global and spans across several sectors. Secondly, we have avoided the technology solutions debate as that is out of scope for our purpose and research question. Thirdly, regarding change management theories, we have chosen to look at research about change readiness and preparation as well as change related to innovation. This is due to the fact that Big Data is still at the

early stages of implementation. Change management theories about the mature stages of change are therefore not relevant to this study. Finally, we have chosen to study the supply side in more detail than the demand side based on our starting point which is the diffusion of innovation theory. The theory focuses on the emitters of an innovation rather than the receivers. Also, the time frame does not allow for detailed studies of both sides.

4. BIG DATA ANALYSIS TOOLS

Big data tools: Jaspersoft BI Suite: The Jaspersoft package is one of the open source leaders for producing reports from database columns. The software is well-polished and already installed in many businesses turning SQL tables into PDFs that everyone can scrutinize at meetings. This is a well-developed corner of the software world, and Jaspersoft is expanding by making it easier to use these sophisticated reports with newer sources of data. Jaspersoft isn't offering particularly new ways to look at the data, just more sophisticated ways to access data stored in new locations.

Big data tools: Pentaho Business Analytics: Pentaho is another software platform that began as a report generating engine; it is, like JasperSoft, branching into big data by making it easier to absorb information from the new sources. You can hook up Pentaho's tool to many of the most popular NoSQL databases such as MongoDB and Cassandra. Once the databases are connected, you can drag and drop the columns into views and reports as if the information came from SQL databases. Pentaho was actually made as a report generating engine, which later took its form as a software program. Similar to JasperSoft, Pentaho gathers information from new sources by branching into big data. It can be integrated with most of the popular NoSQL databases like MongoDB and Cassandra.

Big data tools: Karmasphere Studio and Analyst: Many of the big data tools did not begin life as reporting tools. Karmasphere Studio, for instance, is a set of plug-ins built on top of Eclipse. It's a specialized IDE that makes it easier to create and run Hadoop jobs. Karmasphere also distributes a tool called Karmasphere Analyst, which is designed to simplify the process of plowing through all of the data in a Hadoop cluster. It comes with many useful building blocks for programming a good Hadoop job, like subroutines for uncompressing Zipped log files. Then it strings them together and parameterizes the Hive calls to produce a table of output for perusing.

Big data tools: Talend Open Studio : Talend also offers an Eclipse-based IDE for stringing together data processing jobs with Hadoop. Its tools are designed to help with data integration, data quality, and data management, all with subroutines tuned to these jobs. Talend Studio allows you to build up your jobs by dragging and dropping little icons onto a canvas. If you want to get an RSS feed, Talend's component will fetch the RSS and add proxying if necessary. There are dozens of components for gathering information and dozens more for doing things like a "fuzzy match." Then you can output the results.

Big data tools: Skytree Server: Not all of the tools are designed to make it easier to string together code with visual mechanisms. Skytree offers a bundle that performs many of the more sophisticated machine-learning algorithms. All it takes is typing the right command into a command line. Skytree is more focused on the guts than the shiny GUI. Skytree Server is optimized to run a number of classic machine-learning algorithms on your data using an implementation the company claims can be 10,000 times faster than other packages. It can search through your data looking for clusters of mathematically similar items, then invert this to identify outliers that may be problems, opportunities, or both. The algorithms can be more precise than humans, and they can search through vast quantities of data looking for the entries that are a bit out of the ordinary. This may be fraud or a particularly good customer who will spend and spend.

Big data tools: Tableau Desktop and Server: The Tableau is an American Software Company with its headquarters located in Seattle. It manufactures a variety of interactive data visualization products built in business intelligence [7]. Tableau software implemented Hadoop years ago and it uses Hive for query structuring. Then it attempts with great effort to cache information in the memory in order to make the tool more interactive. Among the other tools that are developed to create reports offline, Tableau needs to give an interactive mechanism so that the user can dig the data as much as possible. Caching assists work with few latency of a Hadoop cluster. Tableau Desktop is a visualization tool that makes it easy to look at your data in new ways, then slice it up and look at it in a different way. You can even mix the data with other data and examine it in yet another light. The tool is optimized to give you all the columns for the data and let you mix them before stuffing it into one of the dozens of graphical templates provided.

Big data tools: Splunk: Splunk is a bit different from the other options. It's not exactly a report-generating tool or a collection of AI routines, although it accomplishes much of that along the way. It creates an index of our data as if our data were a book or a block of text. Yes, databases also build indices, but Splunk's approach is much closer to a text search process. Splunk's mission is to make machine data accessible across an organization by identifying data patterns.

4. RESULT AND DISCUSSION

We have just suggested different tools along with their dependencies. Now It’s totally the choice and ability of

user to choose the suitable Data analytics tools as per requirement of the industry/Organization. Financial and economical constraints may lead to choose open source based tools.

Table-1: Data Analytics Tools for user analysis

Big Data Tools	Mode	Data Types	Database Support	Operating System
JASPER SOFT	Commercial and Open Source	Structured and Unstructured data	Mongo DB, Cassandra, Redis, Riak, CouchDB Neo4j, Hbase	OS Independent
PENTAHO	Commercial and Open Source	Supports structured data.	Mongo DB, Cassandra, Redis, Riak, CouchDB, MapR.	OS Independent
SPLUNK	Commercial	Unstructured data Time-series textual	Relational IBM Database 2, SAP, Sybase	Windows XP, Vista, 7 and 8
TABLEAU	Commercial	Structured and Unstructured data	MySQL, Microsoft SQL Server, Oracle, EMC, GreenPlum	MS Windows 8.1, Vista or Server 2012 R2, 2012, 2008 or 2003
KARMASPHERE	Commercial and Open Source	Structured, Semi-Structured and Unstructured data	Base HDFS file data	Red Hat/Cent OS/Ubuntu Linux
TALEND	Open Source	Structured, Semi-Structured and Unstructured data	Mongo DB, Cassandra, Redis, Riak, CouchDB, MapR.	OS Independent
SKYTREE	Open Source	Structured, Semi-Structured and Unstructured data	RDBMS, HDFS, CSV	Linux

5. CONCLUSIONS

Big data technologies have emerged as biggest asset to handle un-structured and semi structured data, though it is necessary to analyze the proper selection of tool and technology. Buyers in market are keen to know about tools which are beneficial for their organization and profit generated assets. So far we have analyze the some famous tool like jasper tools, Pentaho and Splunk. Tableau etc. Dependency factors are very important while selecting a tool to adopt for current business transactions. By analyzing our comparison of various technologies one should be able to identify its need from personal and commercial point of view. Cost factor and operating system dependency is major issue while selecting any tool for corporate sectors. It has been seen that majority of mid size scale industry prefer to choose open source technologies due to cost factors. Essentially Apache Hadoop is quite famous in this field.

Data handling technologies play vital role on mobile platform. Enormous amount of data is generated data by day from various rich sources like twitter, facebook and instagram. Big data tool are proved to be a boon for such companied data handling issue are big hurdle in this path.

REFERENCES

[1] M. A. Beyer and D. Laney, "The importance of "big data": A definition," Gartner, Tech. Rep., 2012.
[2] X. Wu, X. Zhu, G. Q. Wu, et al., "Data mining with big data," IEEE Trans. on Knowledge and Data Engineering, vol. 26, no. 1, pp. 97-107, January 2014.Rajaraman and J. D. Ullman, "Mining of massive datasets," Cambridge University Press, 2012.
[3] Z. Zheng, J. Zhu, M. R. Lyu. "Service-generated Big Data and Big Data-as-a-Service: An Overview," in Proc. IEEE BigData, pp. 403-410, October 2013. A . Bellogín,I. Cantador, F. Díez, et al., "An empirical comparison of social, collaborative filtering, and hybrid recommenders," ACM Trans. on Intelligent Systems and Technology, vol. 4, no. 1, pp. 1-37, January 2013.
[4] W. Zeng, M. S. Shang, Q. M. Zhang, et al., "Can Dissimilar Users Contribute to Accuracy and Diversity of Personalized Recommendation?," International Journal of Modern Physics C, vol. 21, no. 10, pp. 1217- 1227, June 2010.
[5] X. Liu, G. Huang, and H. Mei, "Discovering homogeneous web service community in the user-centric web environment," IEEE Trans. on Services Computing, vol. 2, no. 2, pp. 167-181, April-June 2009.
[6] Zielinnski, T. Szydlo, R. Szymacha, et al., "Adaptive soa solution stack," IEEE Trans. on Services Computing, vol. 5, no. 2, pp. 149-163, April-June 2012.
[7] F. Chang, J. Dean, S. mawat, et al., "Bigtable: A distributed storage system for structured data," ACM Trans. on Computer Systems, vol. 26, no. 2, pp. 1-39, June 2008.
[8] V. Gupta, G. S. Lehal, "A Survey of Common Stemming Techniques and Existing Stemmers for Indian Languages," Journal of Emerging Technologies in

Web Intelligence, vol. 5, no. 2, pp. 157-161, May 2013.
[9] T. Niknam, E. Taherian Fard, N. Pourjafarian, et al., "An efficient algorithm based on modified imperialist competitive algorithm and K-means for data clustering," Engineering Applications of Artificial Intelligence, vol. 24, no. 2, pp. 306-317, March 2011.
[10] C. Platzer, F. Rosenberg, and S. Dustdar, "Web service clustering using multidimensional angles as proximity measures," ACM Trans. on Internet Technology, vol. 9, no. 3, pp. 11:1-11:26, July, 2009.
[11] G. Adomavicius, and J. Zhang, "Stability of Recommendation Algorithms," ACM Trans. on Information Systems, vol. 30, no. 4, pp. 23:1-23:31, August 2012.
[12] Yamashita, H. Kawamura, and K. Suzuki, "Adaptive Fusion Method for User-based and Item-based Collaborative Filtering," Advances in Complex Systems, vol. 14, no. 2, pp. 133-149, May 2011.
[13] Y. Zhao, G. Karypis, and U. Fayyad, "Hierarchical clustering algorithms for document datasets," Data Mining and Knowledge Discovery, vol. 10, no. 2, pp. 141-168, November 2005.
[14] http://www.maxmunus.com/page/Hadoop-Training
[15] https://databrio.com/big-data
[16] https://www.eno.com/training-tutorials-courses/cloudera-training-courses/
[17] https://www.eno.com/training-tutorials-courses/cloudera-training-courses/
[18] https://data-flair.training/blogs/setup-hadoop-cdh3-on-ubuntu-single-node-cluster/
[19] https://hadooponlinetraining2k3.blogspot.com/
[20] http://conscientia.co.in/workshops.html
[21] https://gauravag1112.wordpress.com/
[22] http://spectraminds.wikidot.com/horton
[23] http://mrclogic.com/Big%20Data%20%20Analytics%20and%20Hadoop.php