International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 2, Issue 11, November 2015

Business Driven On Big Data Technologies

[1] Sartaj Fatima, [2] Mohammed Sajjad Ali
Lecturer, IT/ Project Manager
[1] sartaj1122@gmail.com, [2] mdsajjadali@gmail.com

Abstract: **"Big data" causes "big buzz," with widely varying opinions about what the phrase means. Some database professionals are even skeptical about what value, if any, their organizations can obtain from big data. This paper examines the growing prevalence of big data across nearly every industry and explains why being good at using big data is critical for today's businesses. Some technology vendors in the legacy database or data warehouse spaces say "big data" simply refers to a traditional data warehousing scenario involving data volumes in either the single or multi-terabyte range. Such is not the case. Today, big data isn't limited to traditional data warehouse situations, but includes real-time or line-of-business data stores used as the primary data foundation for online applications that power key external or internal business systems.**

Index Terms **— Big data, Velocity, Variety, Volume, Business, Bytes**

I. INTRODUCTION

The term 'Big Data' describes innovative techniques and technologies to capture, store, distribute, manage and analyze petabyte- or larger-sized datasets with high-velocity and different structures. Big data can be structured, unstructured or semi-structured, resulting in incapability of conventional data management methods. Data is generated from various different sources and can arrive in the system at various rates. In order to process these large amounts of data in an inexpensive and efficient way, parallelism is used. Big Data is a data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it. (Bifet, 2013)

Big data is a term that refers to data sets or combinations of data sets whose size (volume), complexity (variability), and rate of growth (velocity) make them difficult to be captured, managed, processed or analyzed by conventional technologies and tools, such as relational databases and desktop statistics or visualization packages, within the time necessary to make them useful. While the size used to determine whether a particular data set is considered big data is not firmly defined and continues to change over time, most analysts and practitioners currently refer to data sets from 30-50 terabytes(10 12 or 1000 gigabytes per terabyte) to multiple petabytes (1015 or 1000 terabytes per pet byte\ [-p\') as big data. Figure it can be decomposed into three layers, including Infrastructure Layer, Computing Layer, and Application Layer from top to bottom. (Reddy, 2014)

II. 3 VS OF BIG DATA

Volume of data*:* Volume refers to amount of data. Volume of data stored in enterprise repositories have grown from megabytes and gigabytes to petabytes.

Variety of data: Different types of data and sources of data. Data variety exploded from structured and legacy data stored in enterprise repositories to unstructured, semi structured, audio, video, XML etc.

Velocity of data: Velocity refers to the speed of data processing. For time-sensitive processes such as catching fraud, big data must be used as it streams into your enterprise in order to maximize its value.

III. TYPES AND SOURCES OF BIG DATA

Executives need to be cognizant of the types of data they need to deal with. There are three main types of data, regardless of whether or not a company is using big data – unstructured data, structured data, and semi structured data. Unstructured data are data in the format in which they were collected; no formatting is used (Coronel, Morris, & Rob, 2013). Some examples of unstructured data are PDF's, e-mails, and documents (Baltzan, 2012). Structured data are formatted to allow storage, use, and generation of information (Coronel, Morris, & Rob, 2013). Traditional transactional databases store structured data (Manyika et al., 2011). Semistructured data have been processed to some extent (Coronel, Morris, & Rob, 2013). XML or HTML-tagged text is examples of semistructured data (Manyika et al., 2011). Business executives with traditional database management systems need to broaden their data horizons to

include collection, storage, and processing of unstructured and semi structured data. (Howe, 2014)

Data collection of unstructured and semi structured data is done through several internet-based technologies. Chui, Löffler, and Roberts (2010) describe sensors providing big data as being part of the Internet of Things. The Internet of Things is described as sensors and actuators that are embedded in physical objects that provide data through wired and wireless networks (Chui, Löffler, & Roberts, 2010). Some industries that are creating and using big data are those that have recently begun digitization of their data content; these industries include entertainment, healthcare, life sciences, video surveillance, transportation, logistics, retail, utilities, and telecommunications (Chui, Löffler, & Roberts, 2010). Devices generating data in these industries include IPTV cameras, GPS transceiver, RFID tag readers, smart meters, and cell phones (Chui, Löffler, & Roberts, 2010).

IV. LITERATURE REVIEW

The process of the research into complex data basically concerned with the revealing of hidden patterns. Sagiroglu, S.; Sinanc, D. (20-24 May 2013),"Big Data: A Review" describe the big data content, its scope, methods, samples, advantages and challenges of Data. The critical issue about the big data is the privacy and security. Big data samples describe the review about the atmosphere, biological science and research. Life sciences etc. By this paper, we can conclude that any organization in any industry having big data can take the benefit from its careful analysis for the problem solving purpose. Using Knowledge Discovery from the big data easy to get the information from the complicated data sets [6].

The overall Evaluation describe that the data is increasing and becoming complex. The challenge is not only to collect and manage the data also how to extract the useful information from that collected data. According to the Intel IT Center, there are many challenges related to Big Data which are data growth, infrastructure, data variety, data visualization, data velocity. Garlasu, D.; Sandulescu, V. ; Halcu, I. ; Neculoiu, G. ;(1719 Jan. 2013),"A Big Data implementation based on Grid Computing", Grid Computing offered the advantage about the storage capabilities and the processing power and the Hadoop technology is used for the implementation purpose. Grid Computing provides the concept of distributed computing. The benefit of Grid computing center is the high storage capability and the high processing power. Grid Computing makes the big contributions among the scientific research, help the scientists to analyze and store the large and complex data.

In this the performance of SF-CFS is compared with the HDFS using the SWIM by the face book job traces .SWIM contains the workloads of thousands of jobs with complex data arrival and computation patterns. Aditya B. Patel, Manashvi Birla, Ushma Nair (6-8 Dec. 2012) "Addressing Big Data Problem Using Hadoop and Map Reduce" reports the experimental work on the Big data problems. It describe the optimal solutions using Hadoop cluster, Hadoop Distributed File System (HDFS) for storage and Map Reduce programming framework for parallel processing to process large data sets.

Real Time Literature Review about the Big data According to 2013, facebook has 1.11 billion people active accounts from which 751 million using facebook from a mobile. Another example is flicker having feature of Unlimited photo uploads (50MB per photo), Unlimited video uploads (90 seconds max, 500MB per video), the ability to show HD Video, Unlimited storage, unlimited bandwidth. Flickr had a total of 87 million registered members and more than 3.5 million new images uploaded daily. (Swanson, 2012)

V. OBJECTIVES

Like many new information technologies, big data can bring about dramatic cost reductions, substantial improvements in the time required to perform a computing task, or new product and service offerings. Like traditional analytics, it can also support internal business decisions. The technologies and concepts behind big data allow organizations to achieve a variety of objectives, but most of the organizations we interviewed were focused on one or two. The chosen objectives have implications for not only the outcome and financial benefits from big data, but also the process—who leads the initiative, where it fits within the organization, and how to manage the project.

- Cost Reduction from Big Data Technologies
- Time Reduction from Big Data
- Developing New Big Data-Based Offerings
- Supporting Internal Business

VI. METHODOLOGY

A. Data collection methods:The data collection methods that were found in our review of past empirical studies included the following five methods: test or quiz, questionnaire, interview or focus group discussion, observation, and content analysis.

***B. Test or quiz*:** This data collection method may include pretest and posttest. Pretest is an instrument used to gather participants' baseline performance data prior to an intervention or treatment of some sort. An example of pretest data is participants' English vocabulary test or quiz scores before studying materials via mobile

phone email. Posttest, on the other hand, is an instrument used to gather participants' performance data after the conclusion of an intervention or treatment of some sort. An example of posttest data is participants' English vocabulary test or quiz scores after studying materials via mobile phone email.

***C. Questionnaire*:** A Likert-type scale of items used to collect data on participants' satisfaction, or attitudes about a specific issue; for example students' satisfaction with using mobile phones in learning English vocabulary.

D. Interview or focus group: We refer to interview as a verbal exchange or conversation between the researcher and an individual participant done either face to face or through the telephone. On the other hand, in a focus group interview, the researcher questions several individuals in small groups simultaneously (Fontana & Frey, 2000).

E. Observation: A data collection method in which the researcher directly watched participants in natural contexts or in contexts that are contrived to be realistic in order to get an indication of their behaviours or activities (Knupfer & McLellan, 1996)

F. Content analysis: A data collection method used by researchers to study participant behaviour or activity indirectly by gathering and examining the written contents of a communication (e.g. project plans, reflection logs, journals, emails, student worksheets, time logs, or text messages), usually through a process of comparison, and categorisation (Fraenken & Wallen, 2006; Schwandt, 1997). Results showed that 31.4% of all data collection methods used in previous studies was questionnaire, 22.5% were test or quiz items, 20.6% were content analysis, 18.6% were interview or focus group, and 6.9% were observation. (Niketan Pansare1, 2011)

VI. CONCLUSION

We are in the development area of big data. There are various challenges and issues regarding Big data. There must support and encourage fundamental research towards these technical issues if we want to achieve the benefits of big data. Big-data analysis fundamentally transforms operational, financial and commercial problems in aviation that were previously unsolvable within economic and human capital constraints using discrete data sets and on-premises hardware. By centralizing data acquisition and consolidation in the cloud, and by using cloud based virtualization infrastructure to mine data sets efficiently, big-data methods offer new insight into existing data sets.

RECOMMENDATION

The new applications are generating vast amount of data in structured and unstructured form. Big data is able to process and store that data and probably in more amounts in near future. Hopefully, Hadoop will get better. New technologies and tools that have ability to record, monitor measure and combine all kinds of data around us, are going to be introduced soon. We will need new technologies and tools for anonym zing data, analysis, tracking and auditing information, sharing and managing, our own personal data in future. So many aspects of life health, education, telecommunication, marketing, sports and business etc that manages big data world need to be polished in future.

BIBLIOGRAPHY

Anon. (2015). PARALLEL PROCESSING OF HEALTHCARE DATASETS USING MAP REDUCE. International Journal of Advance Research In Science And Engineering .

Bifet, A. (2013). Mining Big Data In Real Time. Journal of Management .

Grzeda, P. C. (2014). Parallelization of electrophysiological phenomena in myocardium. In Springer-Verlag, editor, Proceedings of Euro PVM/MPI 2004, .

Howe, Y. B. (2014). The HaLoop Approach to Large-Scale Iterative Data Analysis" . HaLoop: Efficient Iterative Data Processing .

Niketan Pansare1, V. B. (2011). Online Aggregation for. Journal of Management .

Olmsted, J. P. (2014). Scaling at Scale: Ideal Princeton. Institute for Computational Science and Engineering .

Reddy, S. P. (2014). "Big Data- solutions for RDBMS problems- A survey". Journal of Management .

Swanson, C. H. (2012). Matchmaking: A New MapReduce Scheduling. 10th IEEE International Conference on Computer and Information Technology .