RESEARCH ARTICLE OPEN ACCESS

Big Data and Big Data Management (BDM) with current Technologies –Review

Atul S, Desale Girish B, and Patil Swati P
Department of Computer Science & IT, JET's Z. B. Patil College, Dhule, Maharashtra, INDIA
Department of Computer Science, S.S.V.P.S's Science College, Dhule, Maharashtra, INDIA

ABSTRACT

The emerging phenomenon called ―Big Data‖ is pushing numerous changes in businesses and several other organizations, Domains, Fields, areas etc. Many of them are struggling just to manage the massive data sets. Big data management is about two things - ―Big data‖ and ―Data Management‖ and these terms work together to achieve business and technology goals as well. In previous few years data generation have tremendously enhanced due to digitization of data. Day by day new computer tools and technologies for transmission of data among several computers through Internet is been increasing. It‗s relevance and importance in the context of applicability, usefulness for decision making, performance improvement etc in all areas have emerged very fast to be relevant in today‗s era. Big data management also has numerous challenges and common complexities include low organizational maturity relative to big data, weak business support, and the need to learn new technology approaches. This paper will discuss the impacts of Big Data and issues related to data management using current technologies.

Keywords: Big Data, Big Data Management (BDM),e-Bussiness, Electronic Data Interchange(EDI), e-Governence, e-Commerce, Hadoop, ETL.

I. INTRODUCTION

We are aware of term e-Business, Electronic Data Interchange(EDI), e-Governence, e-Commerce (Further we will call it electronic era).This is all about to automate maximum organizational processes with the use of computer software or applications and Internet. That is why several companies and organizations had enforced automated computer applications. Today, it is assumed that an organization of any size or complexity have several computer applications for the sake of efficiency and competitiveness. A consequence after this electronic era is that many organizations of any size, domain, area and various companies now have massive volumes of data to manage and it is increasing day by day with extreme velocity and variety on account of 3Vs concept regarding Big Data i.e. Volume, Variety and Velocity. Although organizations have the skills for structured data (which is what comes out of most operational applications) today‗s unprecedented data volume and speed of generation make big data management a challenge. [1]

Big Data generating resources:

Big Data can be simply defined by explaining the 3V_s – volume, velocity and variety which are the driving dimensions of Big Data concept. Gartner analyst Doug Laney introduced the famous 3 V_s concept in his 2001 Meta group publication, ―3D data management: Controlling

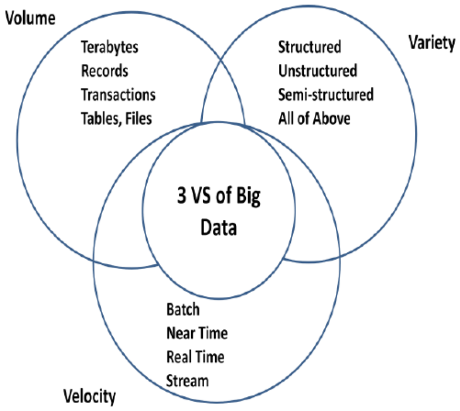

Fig.1: Schematic representation of the 3V's of Big Data

1. Volume: This essentially concerns the large quantities of data that is generated continuously. Initially storing such data was difficult because of high storage costs. This data can be easily distinguishes between structured data, unstructured data and semi-structured data. [6]

2. Velocity: In prehistoric times, data was processed in batches. However was possible only when the incoming data rate is slower than the batch processing rate. Now the speed of data generation and transmission is extremely high. For example Facebook [7] it generates 2.7 billion like actions/day and 300 million photos amongst others roughly amounting to 2.5 million pieces of content in each day while Google Now processes over 1.2 trillion searches per year worldwide. [8].

3. Variety: Today data is loosing the structure and now formats are documents, databases, excel tables, pictures, videos, and audios in hundreds of formats. Structure cannot be imposed like before for data analysis. It can be of any type- structures, semi-structured or unstructured.

Today‘s data sources may include [2]:

 Traditional enterprise data from operational systems.

- Student test data.
- Social media data.
- Institution marketing data.
- Financial business forecast data.
- Web site browsing pattern data.
- Campus sensor data.
- Data gathered from mobile devices.

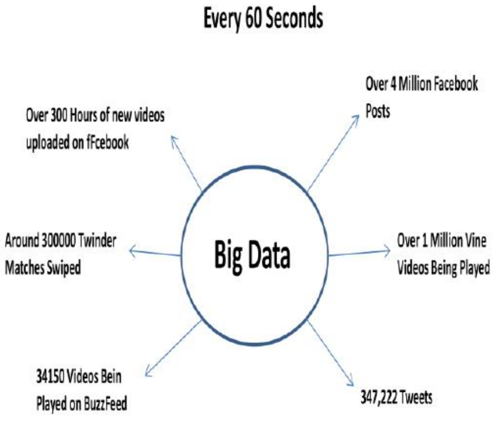
Every 60 Seconds
Over 4 Million Facebook Posts
Over 300 Hours of new videos uploaded on fFcebook
Around 300000 Twinder Matches Swiped
Big Data
Over 1 Million Vine Videos Being Played
34150 Videos Bein Played on BuzzFeed
347,222 Tweets

Fig.2: Data generation on Internet

―Big data comes from many sources, in many formats. Some industries have large, valuable stores of unstructured data, typically in the form of human language text. For example, the claims process in insurance generates many textual descriptions of accidents and other losses, plus the related people, locations, and events. Most insurance companies process this unstructured big data using technologies for natural language processing (NLP), often in the form of text analytics. The output from NLP may feed into older applications for risk and fraud analytics or actuarial calculations, which benefit from the larger data sample provided via NLP‖ [1].

―Sensors are coming online in great numbers as a significant source for big data. For example, robots have been in use for years in manufacturing, but now they have additional sensors so they can perform quality assurance as well as assembly. For decades, mechanical gauges have been common in many industries (such as chemicals and utilities), but now the gauges are replaced by digital sensors to provide real-time monitoring and analysis. GPS and RFID signals now emanate from mobile devices and assets—ranging from smart phones to trucks to shipping pallets—so all these can be tracked and controlled precisely‖[1].

Big Data:

- Includes several varieties of content formats such as text, audios, videos, photographs, images, pdfs, ppts etc.
- Can have several electronic transactions such as sharing of files through e-mails, sharing through social media, file transfers etc.
- Has various levels of engagements while using social media (share, comment, like, tweet, retweet tag etc.)
- Have scalable communication for people through one-one, one-many, many-many communications through various online forums and portals of several experts or people of similar or different community on number of issues.
- Generation is device independent it can be generated through different mobiles and smart phones, palmtops, PCs etc.
- Is also generated through real time transactions such as whether data inputs through various electronic devices or sensors from various geographic locations, satellites which is in huge amount of digital inputs generating per second for statistical analysis

BDM Technologies in use:

As data sets are in various formats and data generation through electronic means are generating in large volumes, velocity and variety per second, it has become difficult for current DBMS & RDBMS to store, manipulate and extract required information from it in real time. Big Data Management Current technologies the most commonly used framework is Hadoop. Hadoop is the combination of several other components like Hadoop Distribution File Systems (HDFS), Pig, Hive and HBase etc. "Apache Hadoop is an open-source software framework for storage and large-scale processing of data-sets on clusters of commodity hardware. There are mainly five building blocks inside this runtime environment (from bottom to top)". [9]

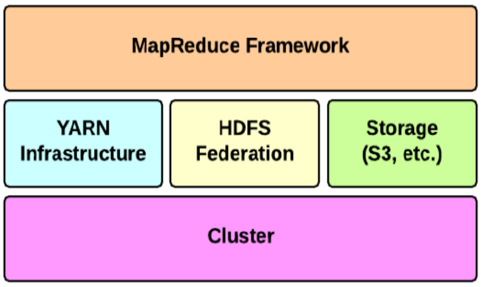

Fig.3: Hadoop Architecture [9]

Limitations for Current Technologies:

The rise of big data has come with the limitations with the management issues. Even five years ago, a company could leverage a DBMS such as Oracle for a data warehouse. However, Oracle was built in a time when databases rarely exceeded a few gigabytes in size. Along with other legacy DBMSs, it cannot perform at the scale now required[10]. While Hadoop is very new and unknown to large number of people, it is been something of a mystery to the business world. Today it is highly recognised by the technology world, but most people still are unaware or confused that what it is actually used for. Basically some technicians think that it was developed to facilitate data analytics and certain forms of batch-oriented distributed data processing and ETL (Extract, Transform, Load) operations. Where Hadoop has major limitations in its analytic functionality, and database like qualities of Hadoop are not a replacement for a true analytic platform and fundamentals of Hadoop was not designed to facilitate highly interactive analytics[10]. As HDFS was implemented to speed the processing of various web documents, and to apply the Map Reduce framework to this processing where there was no need of schema as well as it was built to operate on clusters of arbitrary size so there was no need of appropriate storage scheme. And also there was no optimizer to look into data flow and it‘s nature automatically and no check points for data recovery. This implies that the output extracted from Hadoop cluster have no guarantee to be perfect and 100% [10]. And the lacking of an expertise in handling the Hadoop Framework for BDM precisely is the biggest issue today.

II. CONCLUSIONS

In this paper, we have tried to present the significance of Big data, limitations of current technologies in analyzing Big data. Where the major issues in processing the Big data are quality of output due to open source components in analytical and management processes, lacking of expertise required in operating the current Hadoop framework, compatibility of the current technology and security.

REFERENCES

[1]. TDWI Research – TDWI Best Practices Report By Philip Russom Fourth Quarter 2013.

[2]. Big Data and Social Media to Improve the Quality of Higher Education - Dr. Savita Kumari (Sheoran)Assistant Professor, Dept. of CS&A Indira Gandhi University Meerpur, Rewari (Haryana) INDIA - 122502 IJCSMC, Vol. 5, Issue. 3, March 2016, pg.179 – 185.

[3]. http://www.3rootsstudios.com/is-big-data-under-utilized.

[4]. http://www.exist.com/wp-content/uploads/2014/10/3Vsbigdata.png.

[5]. http://www.forbes.com/sites/gartnergroup/2013/03/27/gartners-big-data-definition-consists-of-three-parts-not-to-be-confused-with-three-vs Big Data – Concepts, Applications, Challenges and Future Scope

[6]. Samiddha Mukherjee1, Ravi Shaw2 Information Technology, Institute of Engineering & Management, Kolkata, India. International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 2, February 2016 Copyright to IJARCCE DOI 10.17148/IJARCCE.2016.5215 ISSN(Online)2278-1021. ISSN(Print) 2319-5940.

[7]. http://www.internetlivestats.com/twitter-statistics/

[8]. http://www.internetlivestats.com/google-search-statistics/

[9]. http://ercoppa.github.io/HadoopInternals/HadoopArchitectureOverview.html

[10]. .https://www.em360tech.com/wp-content/files_mf/1360922634PARACCEL1.pdf