Volume 5, Issue 7, July 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Next Generation Eco-System of Big Data

Ch. Ratna Babu[1], Dr. M. Sreelatha[2], K. Praveen Kumar[3]
[1]Associate Professor, Department of CSE, RVR & JC College of Engineering, Guntur, Andhra Pradesh, India
[2]Professor, Department of CSE, RVR & JC College of Engineering, Guntur, Andhra Pradesh, India
[3]Assistant Professor, Department of CSE, Mallineni Lakshmaiah College of Engineering, Guntur, Andhra Pradesh, India

Abstract— Today we are living in computing world. With increased digitalization the amount of structured and unstructured data being created and stored is exploding. The data is being generated from various transactions, digital images, videos, audios and click streams for domains including healthcare, retail, energy and utilities. In addition to business and organizations, individuals contribute to the data volume. For instance, 30 billion content are being shared on face book every month; the photos viewed every 16 seconds in Picasa could cover a football field. So Organizations need an efficient way to deal with this data explosion for this purpose big data came in to picture. In this paper various concepts related to big data are discussed briefly.

Keywords— Big Data, Eco-System, Reference architecture , Hadoop.

I. INTRODUCTION

The term Big Data refers to large scale information management and analysis technologies that exceed the capability of traditional data processing technologies. Big Data is differentiated from traditional technologies in three ways: the amount of data (volume), the rate of data generation and transmission (velocity), and the types of structured and unstructured data (variety). Human beings now create 2.5 quintillion bytes of data per day. The rate of data creation has increased so much that 90% of the data in the world today has been created in the last two years alone. This acceleration in the production of information has created a need for new technologies to analyse massive data sets.[1]

II. PROPERTIES OF BIG DATA

There are lot of properties of Big Data which are described in short notation as V's. These V's can be 3,4 or even 5 as they are described by different data organisations. In this document we are describing only 3 important V's.

A. Volume

Big data uses massive data, including for example meta-data from internet searches, credit and debit card purchases, social media postings, mobile phone location data, or data from sensors in cars and other devices. The volume of data being produced in the world continues to increase rapidly. The Boston Consulting Group estimates total growth of 2.5 Exabyte's, which equals 2.5 billion gigabytes, per day . It is increasingly possible to hold very large datasets, due to the decreasing cost of storage and the availability of cloud-based Big data and data protection services. These datasets may be so large that they cannot be analysed using 'traditional' methods, such as MS Excel spread sheets, relational databases and SQL queries, but new tools have been developed to analyse them such as No SQL and the open source software Hadoop.

B. Variety

Big data often involves bringing together data from different sources. Currently it appears that big data analytics mainly uses structured data eg: in tables with defined fields but it can also include unstructured data. For example, it is possible to obtain a feed of all the data coming from a social media source such as Twitter. This is often used for 'sentiment analysis', to analyse what people are saying about products or organizations. A retailer might combine this data with their own in-house data collected from point-of-sale terminals and loyalty cards, to produce rich and detailed information for \marketing. From an IT perspective, combining data from different sources in this way presents particular challenges. Technologies have been developed for big data that do not require all of the data to be put into a single database structure before it can be analysed.

C. Velocity

In some contexts, it is important to analyse data as quickly as possible, even in real time. Big data analytics can be used to analyse data 'in motion', as it is produced or recorded, as well as data 'at rest' in data stores. A potential application of 'in motion' analysis is in credit card payments. For example, Visa is looking at using big data analytics to develop a new ways of authorizing credit card payments. The volume, variety and velocity of Big Data causes performance problems

when being created, managed and analysed using the conventional data processing techniques. Using conventional techniques for Big Data storage and analysis is less efficient as memory access is slower. The data collection is also challenging as the volume and variety of data has to be derived from sources of different types[2]. The other major challenge in using the existing techniques is they require high end hardware to handle the data with a huge volume, velocity and variety.

III. BIG DATA ANALYTICS

The process of analysing and mining Big Data can produce operational and business knowledge at an unprecedented scale and specificity. They need to analyze and leverage trend data collected by businesses is one of the main drivers for Big Data analysis tools The technological advances in storage, processing, and analysis of Big Data include (a) the rapidly decreasing cost of storage and CPU power in recent years; (b) the flexibility and cost effectiveness of data centres and cloud computing for elastic computation and storage; and (c) the development of new frameworks such as Hadoop, which allow users to take advantage of these distributed computing systems storing large quantities of data through flexible parallel processing. These advances have created several differences between traditional analytics and Big Data analytics[3].

IV. HADOOP USAGE IN BIG DATA

Hadoop is a distributed file system and data processing engine that is designed to handle extremely high volumes of data in any structure. Hadoop has two components. The Hadoop distributed file system (HDFS), which supports data in structured relational form, in unstructured form, and in any form in between the Map Reduce programming paradigm for managing applications on multiple distributed servers. The focus is on supporting redundancy, distributed architectures, and parallel processing. Hadoop works on the principle of Map-Reduce algorithm, the following is a map reduce algorithm for searching a word in document.

$$\mathrm{tf_i} = \frac{n_i}{\sum_k n_k}$$

$$\mathrm{idf_i} = \log \frac{|D|}{|\{d : t_i \in d\}|}$$

$$\mathrm{tfidf} = \mathrm{tf} \cdot \mathrm{idf}$$

D: Total no of documents

{d:t_i€d}: No of documents where t_i appears

V. BIG DATA & ANALYTICS NEXT GENERATION ARCHITECTURE

A. Unified Information Management

1) *High Volume Data Acquisition:* The system must be able to acquire data despite high volumes, velocity, and variety It may not be necessary to persist and maintain all data that is received. Some may be ignored or discarded while others are kept for various amounts of time.
2) *Multi Structured Data Organization and Discovery:* The ability to navigate and search across different forms of data can be enhanced by the capability to organize data of different structures in to a common schema Using this form of organization, the system can relate structured data such as model numbers and specifications, semi structured data such as product documents, and unstructured data such as installation videos. In addition, new business opportunities can be discovered by looking at different forms of data in new ways.
3) *Low Latency Data Processing:* Data processing can occur at many stages of the architecture. In order to support the processing requirements of Big Data, the system must be fast and efficient[4].
4) *Single Version of the Truth:* When two people perform the same form of analysis they should get the same result As obvious as this seem , it isn't necessarily a small feat, especially if the two people belong to different departments or divisions of a company. Single version of truth requires architecture consistency and governance.

B. Real-Time Analytics

Real Time Analytics the business to leverage information and analysis as events are unfolding At a high level this includes:

1) *Speed of Thought Analysis:* Analysis is often a journey of discovery, where the results of one query determine the content of the next. The system must support this journey in an expeditious manner. System performance must keep pace with the users thought process[5].
2) *Advanced Analytics:* Advanced forms of analytics, including data mining, machine learning, and statistical analysis enable businesses to better understand past activities and spot trends that can carry forward into the future. applied in real time, advanced analytics can enhance customer interactions and buying decisions, detect fraud and waste, and enable the business to make adjustments according to current conditions.
3) *Event Processing:* Real time processing of events enables immediate responses to existing problems and opportunities. It filters through large quantities of streaming data, triggering predefined responses to known data patterns.

4) *Intelligent Processes:* Intelligent Processes a key objective for any Big Data and Analytics program is to execute business processes more effectively and efficiently This means channeling the intelligence one gains from analysis directly into the processes that the business is performing. At a high level this includes:
5) *Application Embedded Analysis:* Many workers today can be classified as knowledge workers; they routinely make decisions that affect business performance Embedding analysis into the applications they use helps them to make more informed decisions.
6) *Guided User Navigation:* Some processes require users to take self directed action in order to investigate an issue and determine a course of action whenever possible the system should leverage the information available in order to guide the user along the most appropriate path of investigation.
7) *Performance and Strategy Management:* Analytics can also provide insight to guide and support the performance and strategy management processes of a business. It can help to ensure that strategy is based on sound analysis. Likewise, it can track business performance versus objectives in order to provide insight on strategy achievement.

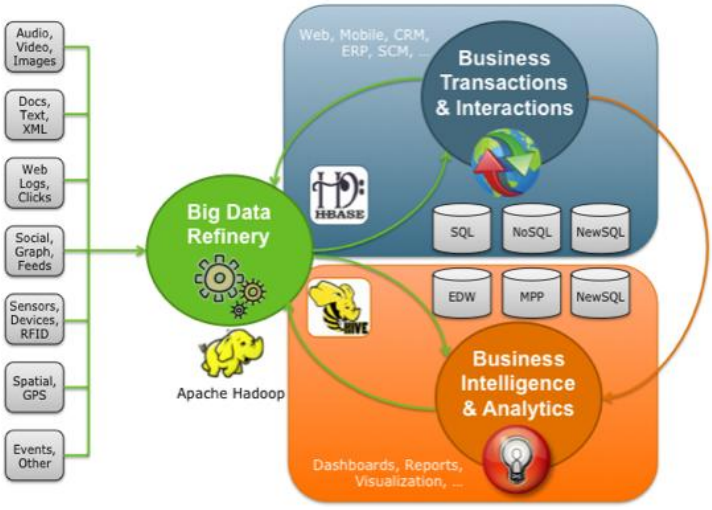

Fig1. Next Generation Reference Architecture of BIG DATA

VI. APPLICATIONS AND COMPLETE WORKING ECO-SYSTEM OF BIG DATA

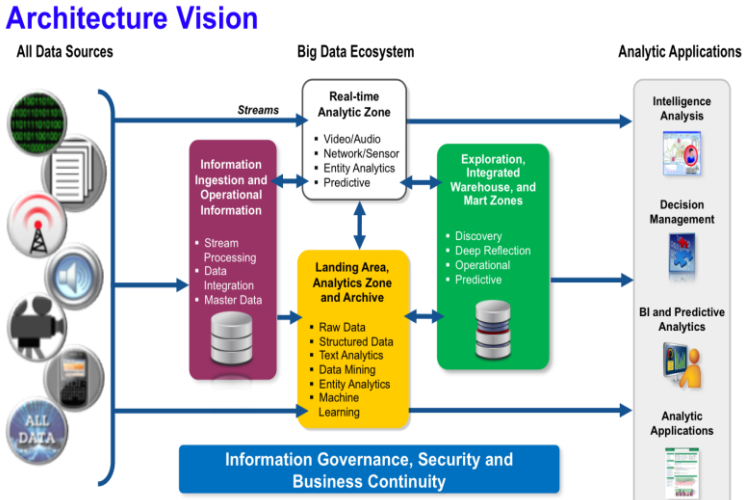

Fig 2: BIG DATA Eco-System Architecture

The above architecture of Eco-System contains following
Components:

1. Data Sources: These are the sources from which all data is gathered and stored in ware house.
2. Big Data Eco-System: Explains about various areas big data is used.
3. Analytical Applications: it explains about the different type of applications and where really Big Data is used.

VII. CONCLUSION

The above concepts which we have discussed are in still in theoretical phase. In future Big Data along with other concepts like Cloud Computing and Hadoop going to solve present Data Storing problems which are increasing minute to minute. But the main problem of using Concepts of Big data and Cloud Computing are how the data can be protected is a big question which is regulating the growth of these concepts to some extent

REFRENCES

[1] Big data and data protection" 20140728 Version:1.0" https://ico.org.uk/media/for-organisations/documents/1541/big-data-and-data-protection.pdf

[2] "Big data: The next frontier for innovation, competition and productivity" by McKinsey Global Institute june 2011.

[3] "Big Data Analytics for Security Intelligence" by Big Data Working Group September 2013.

[4] "Big Data Spectrum " by Infosys.

[5] "Big Data & Analytics Reference Architecture" by Oracle Enterprise September 2013. http://www.oracle.com/technetwork/topics/entarch/oracle-wp-big-data-refarch-2019930.pdf

[6] "Architecture Framework and Components for the Big Data Ecosystem Draft Version 2.0:" by Yuri Demchenko, Canh Ngo, Peter Membrey September.2013.http://www.researchgate.net/profile/Yuri_Demchenko/publication/221276559_Security_Services_Lifecycle_Management_in_OnDemand_Infrastructure_Services_Provisioning/links/0deec5217e22860c3e000000.pdf

