Big Data: Insight

Mrs. S.V. Balshetwar, Dr. R.M.Tugnayat

***Abstract*— In this world of information technology the ever growing thing is data, it can be set of facts, observations anything in the form of digitized manner. Now a day the data has been in trillion gigabytes. Every person making use of social media has abundant of data on internet, companies make use of this data to analyze many things right from sentiments of people in purchasing a product to fraud detection or technically for securing the company data. This paper gives an idea regarding big data and the technology around big data.**

***Index Terms*— Big data, Big data analytics, NoSQL, unstructured data.**

I. INTRODUCTION

The world around has large amount of data, classified data, structured data, unstructured data, homogeneous data, heterogeneous data, out of the available data which can have abundant information in it, that can be extracted on three main parameters: accuracy, timeliness and completeness of data.

In this digital world every now an than data is generated from, sensors , social media that is responsible for explosive growth of data .An survey [1] shows that the rate of data creation has increased so much that 90% of data in the world today has been created in recent two years thus data today known as *Big* data is the large and rapidly growing volume, variety and velocity, veracity of information that cannot be leveraged by existing RDBMS and data warehousing systems. Fig 1. Shows 4 Vs of Big Data

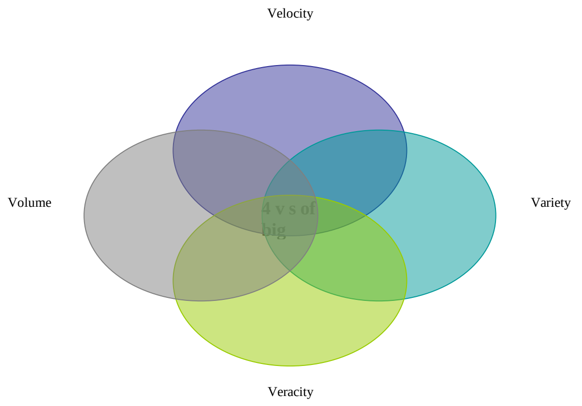

Fig. 1 4 Vs of Big Data

Manuscript received May 20, 2014.

Mrs.S.V.Balshetwar, Information Technology, Satara College of Engineering and Management, Limb, Satara, India.
Dr. R.M. Tugnayat, Principal, Shri Shankarprasad Agnihotri College of Engineering,Wardha, India

Business people are capturing and analyzing big data for adding value in the process of decision making. Nevertheless it can be used by Consumer services, government agencies, capital markets, healthcare and etc.

Web based companies are more interested in analysis of this data, it is possible because of decreasing cost of storage devices, flexibility and cost effectiveness of data centers and the recent developments of new frameworks that can get integrated with new data management systems that has improved analytical capabilities .

II. BIG DATA ANALYTICS

Traditionally the data in DW are structured, that are extracted from operational systems. Whereas big data comes from tweets, sensors, mobile which are not of the same structure rather they are multi structured or unstructured. Analysis of structured data is easier as compared to semi structured, multi structured or unstructured data. In order to extract value out of big data, analysts make use of many advanced analytical techniques.

Traditional structured data is stored in RDBMS and SQL is used as analysis tool which cannot be used in case of big data because it comes in variety and volume from various sources. Thus there are multiple trends for management and processing of big data, Hadoop and MapReduce, offer alternatives to traditional data warehousing. ADBMSs (analytic RDBMSs) and non-relational systems (sometimes called NoSQL systems) are available for processing multi-structured data [2][3].

Big data analytics is process or technique of applying advanced analytic techniques to very large unstructured data, analyzing such data can produce operational and business value data.

Fig. 2 shows the components of Big Data Analytic process.

Big data analytics investigate the lowest granular details of business operations and find their way in standard reports.

III. BIG DATA TECHNOLOGIES TO HANDLE DATA

Data at rest and data in motion both come under big data, accordingly technology for handling such data have been grouped: batch processing analytic for data at rest and stream processing analytics for data under motion.

Large volumes (gigabytes, terabytes, petabytes) of data are stored in memory or disk which we can call data at rest. Hadoop is one of the most populat technology for batch processing. The Hadoop framework make use of HDFS for storing large files and MapReduce programming model

is used to handle large scale data processing problems that are distributed and parallelized. Large Streams of responded which

data that come every second, minute or hour is called data in motion. Stream processing is a growing area of research it does not have a single dominant technology like Hadoop.

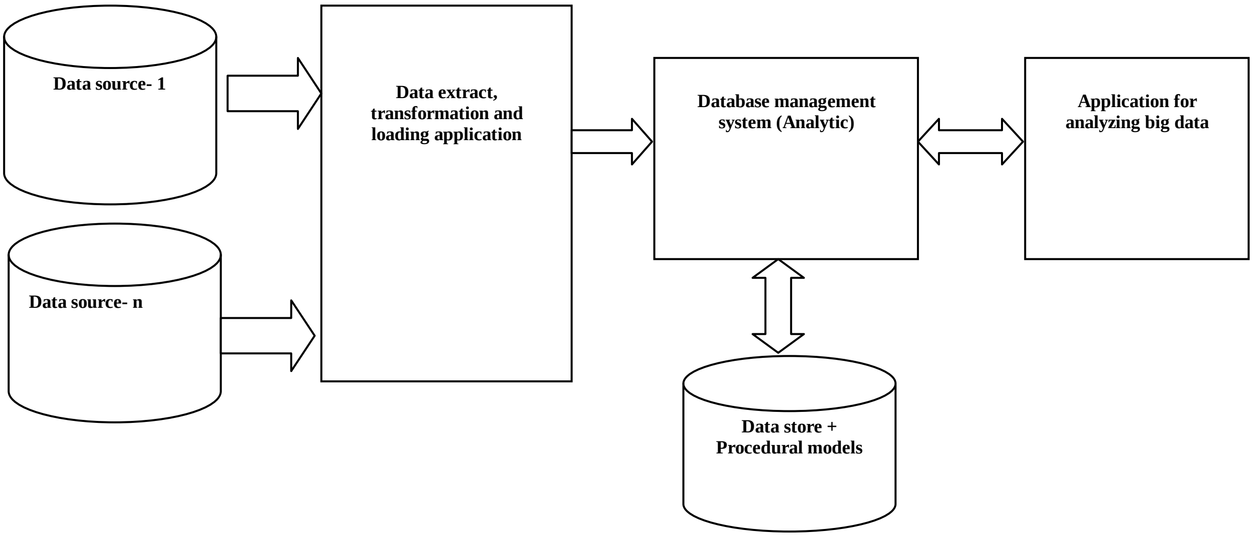
Data source- 1
Data source- n
Data extract, transformation and loading application
Database management system (Analytic)
Application for analyzing big data
Data store + Procedural models

Fig 2 Big Data Analytics: Components

IV. Databases for Big Data

Increasing volume of data is beyond the traditional RDBMS models but it is easier to map such data objects onto NOSQL databases. Databases used for big data storage are viz: Key-value databases, Document Storage database, Graph Database,Object Database, Multi-model Database, Column-oriented databases, Schema-less databases or NoSQL databases[2].

For efficient querying and storage of big data databases such as Cassandra, Couchdb, Greenplum Database, Hbase, Mongodb, Vertica, Aster Data, Hypertable, Big Table, Saphana, Infogrid, Hypergraphdb, Allegrograph, Bigdata, Versant, Db4-O, Allegrograph, Virtuoso, Terrastore, Leveldb, Couchbase, Server,Berkeley Db,Voldemort, Memcachedb, Amazon Dynamodb, Dynomite.

V. Top languages for Big Data Analytics

The most popular languages continue to be R (used by 61% of KDnuggets readers), Python (39%), and SQL (37%). SAS is stable at around 20%. The highest growth was for Pig/Hive/Hadoop-based languages, R, and SQL, while Perl, C/C++, and Unix tools and Clojure are also on high rate in popularity.

VI. Big Data Software

Platfora, Datameer, Hadoop, Spark, HP Vertica , MongoDB , Splunk , Tableau.

VII. VTechnologies associated with Big Data Analytics

NoSQL databases, Hadoop and MapReduce. These technologies form the hub of an open source software framework that supports the processing of large data sets across gathered systems [4].

VIII. Big Data Processing Requirements

For placing data, a proper structure is required and the requirements are:

1. Reduced time for loading data.
2. Speed must be increased for processing of queries.
3. Utilizing the storage space efficiently.
4. Handling dynamic, unstructured data patterns.

IX. Big Data Storage Architecture Considerations

In order to get proper solution for storage of big data following factors can be considered:

1. Big data comes at high speed so the bandwidth requirement as per application is an important consideration.
2. What is the data type: structured, unstructured or mix.
3. Is the big data considered for application is distributed or concentrated at one physical location?
4. How is the data? is it stratum or not.
5. How is the access to data? Is old data required often, does it get mix with new data access very often.

X. Hardware and Software Architecture to Handle ‘Big Data’**:**

The three primary architectures used to handle big data are [7]:

1. Symmetric Multiprocessing Solutions (SMP): This infrastructure uses multiple processors that share a common operating system and memory, it is used as the basis of most BI/DW environment.

2. Massively Parallel Processing (MPP) data warehousing appliances: In this infrastructure every processor has its own operating system and memory; it can grow horizontally and is used mainly for structured data.
3. NoSQL platforms : NoSQL database, also called Not Only SQL, is an method to data management and database design that's useful for very large sets of distributed data. NoSQL is especially useful when an enterprise needs to access and analyze massive amounts of unstructured data or data that's stored remotely on multiple virtual servers.

XI. Better use of BIG DATA

Like bacteria, big data are lurking in the stomachs of cows. Some farmers are using sensors and software to analyze it and predict when a cow is getting ill [8].

The casino company Caesars Entertainment uses data to spot when gamblers have lost so many times at the slot machines that they might not come back: "If the company can present, say, a free meal coupon to such customers while they're still at the slot machine, they are much more likely to return to the casino later."

London's Heathrow airport increased the number of on-time flights from 65% to 80% in just two months after using an algorithm to coordinate everything that goes into a flight turnaround process.

Telecom company Verizon has a unit that analyzes location data for other businesses for example, telling a basketball team where the fans at their stadium came from.

Manufacturing companies commonly embed sensors in their machinery to monitor usage patterns, predict maintenance problems, and enhance build quality. Studying these data streams allows them to improve their products and devise more accurate service cycles.

Insurance companies are now asking drivers to voluntarily contribute data that tracks their movement, locations, and where they are at various times of the day so they can develop better risk profiles for each customer. By showing that they drive the speed limit, travel in areas that incur fewer accidents, and avoid high crime areas customer can qualify for a lower cost insurance plan.

Multi-Channel Marketing and Sentiment Analysis , companies combine social media feeds, customer demographic information, psychographic data (values, attitudes, interests, or lifestyles), purchase data, and network usage data to paint a complete picture of each customer’s behavior, likes, and dislikes. Harnessing this information helps retailers to understand each potential buyer as a “market of one” and to present personalized, tailored offerings to individual customers.

XII. CONCLUSION

We have entered an era of Big Data. Analyzing new and diverse digital data streams can reveal new sources of economic value and provide fresh insights and identify market trends. Hopefully by highlighting several technologies related to Big Data and the importance of Big Data in today’s world has raised importance of Big Data in academic research also.

Acknowledgment

I would like to thank Prof. (Dr.) R.M. Tugnayat for his encouragement and comments on earlier drafts of this paper. Without the support of my loving husband V.P. Balshetwar (B.E. Electronics) & our kids Siddhesh and Katyayani it would have been impossible to complete this work.

I would also like to thanks Dhiraj Mothghare, Vijay Urade, Raman Bane who have encouraged in selecting this topic for research.

References

[1] Improving Decision Making in the World of Big Datahttp://www.forbes.com/sites/christopherfrank/2012/03/25/improvingdecision-making-in-the-world-of-big-data/
[2] 10 emerging technologies for Big Data By Thoran Rodrigues in Big Data Analytics, December 4, 2012
[3] Analytic platforms: Beyond the Traditional Data Warehouse by By Merv Adrian and Colin White Beye NETWORK Custom Research Report Prepared for Vertica.
[4] http://hadoop.apache.org/.
[5] http://www.facebook.com/press/info.php? statistics.
[6] J. Dean and S. Ghemawat, “MapReduce: Simplified data processing on large clusters,” in OSDI, 2004, pp. 137–150.
[7]http://www.gartner.com/technology/research/big-data /
[8] How companies can make better use of big data - Los Angeles Times

Mrs.BalshetwarS.V. is working as Head of Information Technology department at satara college of engineering and management, limb, satara. She has completed her M.Tech (Computer Sci. & technology.) at Shivaji University, Kolhapur, India. She received her AMIE (Computer Sci. & Engg.) from Institute of Engineers (India), Kolkata, India in 2008. Her research interest includes data security & data mining, Artificial Intelligence, Big Data.

Dr. R.M. Tugnayat, Principal, Shri Shankarprasad Agnihotri College of Engineering,Wardha, India.