ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

An Initial Imperative Study on Big Data

Mayushi Chouhan[1], Makhan Kumbhkar[2]

Student, Department of Computer Science, Acropolis Institute of Technology & Research, Indore, MP, India[1]

Asst. Professor, Department of Computer Science, Christian Eminent College, Indore, MP, India [2]

ABSTRACT: Big data has become possible due to low cost storage, high performance servers, high-speed networking. Big data is a word for data sets that are so huge or intricate that conventional data processing application software is not enough to deal with them. This paper presents Big data Techniques, different big data Sources and Hadoop Distributed file System.

KEYWORDS: BIG data analysis techniques, big data sources, types of data, HFDS,EJB

I. INTRODUCTION

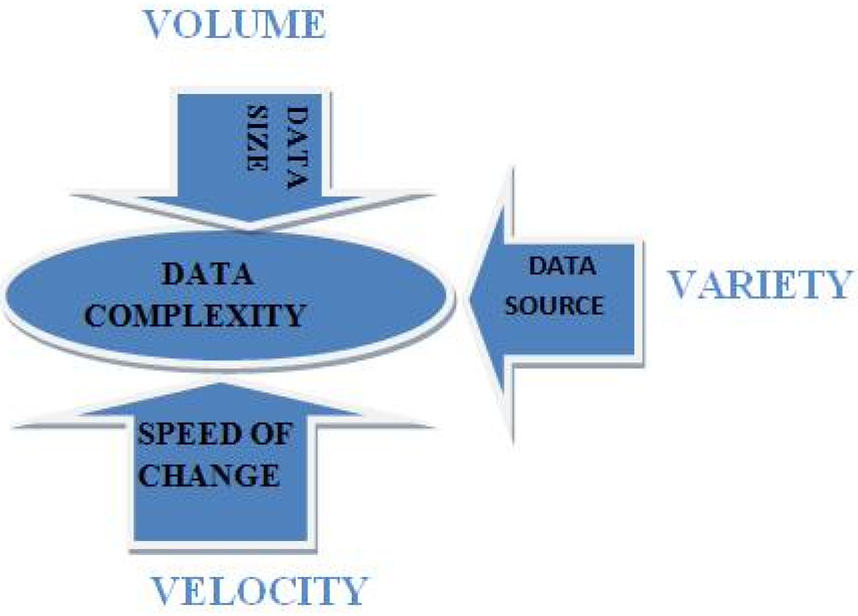
Big data is really big, reaping value from lot of data is real challenges for enterprise that practice Big Data.
Enterprises are collecting huge amount of data from various sources, but falling to evaluate it properly and use it for data Analytics and business decisions.
It is necessary for enterprises to analysis the data efficiently and derives workable business strategies that derive business performance.
It can be best leveraged with efficiently predictive analytics tools and gain deeper insights into customer preferences and market dynamics.
Big data may well be the Next Big Thing in the IT world. The first organization to embrace it was online start-up firms. Firms likes Google, eBay, LinkedIn, and Face book build around big data from the beginning .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. It is similar to small data but bigger in size. but having data bigger it requires different approaches .it generate values from the storages and processing of very large quantities of digital information that cannot be analyzed with traditional computing technique.

II. ATTRIBUTES OF BIG DATA

A) **Big Data Volume:** a typical pc might have had 10 gigabytes of storage in 2000. Today face book ingests 500 terabytes on new data every day .The smart phones, the data they create and consume; sensors embedded into everyday objects will soon result in billions of new, constantly-updated data feed contain environmental, location, and other information including video.

B) **Big Data Veloci ty:** Click streams and ad impressions capture user behaviour at millions of events per second high frequency stock trading algorithms reflect market changes within milliseconds .machine to machine process exchange data between billions of devices.

C) **Big Data Variety**: Big data is not just a number, date and strings. Big Data is also geospatial data, 3D data, audio and video**,** and unstructured text including log file and social media.



Fig(1) : attributes of big data

III. STORING, SELECTING AND PROCESSING OF BIG DATA

Big data store: The structured and unstructured types of data is also reason enough why we need to look for tools that can help in storing the data and mine it properly for best results. Hadoop and Cloudera are equipment that enterprises should utilize in order to store the structured and unstructured data. They generate an enterprise hub, where the data is stored, and wherein security is given highest priority.

There are many Store of Big data like Key Value, graph, document, Hadoop Distributed file System.

Selecting Big Data Store: When selecting big data technologies, there are a lot of elements in the solution to consider. Selecting a storage platform that can meet the needs of big data can be a challenge. Choosing the big data Stores based on our data characterises .moving code to data .aligning business goals to the appropriate data store.

Processing Big Data: Apache Hadoop is a distributed computing framework modelled behind Google MapReduce to process large amounts of data in parallel. Hadoop Distributed File System (HFDS) modelled on Google GFS is the essential file system of a Hadoop cluster. HDFS works more powerfully with a few large data files than numerous small files. In general, data flows from components to components in an enterprise application. This is the case for application frameworks (EJB and spring framework), integration engines (Camel and Spring Integration), as well as ESB (Enterprise Service Bus) products. Nevertheless, for the data-intensive processes Hadoop deals with, it makes better sense to load a big data set once and perform various analysis jobs locally to minimize IO and network cost, the so-called "Move-Code-To-Data" philosophy. When we load a big data file to HDFS, the file is split into small blocks (or chunk or file blocks) through a master node (centralized Name Node) and resides on individual slave nodes (Data Nodes) in the Hadoop cluster for parallel processing.

IV. BIG DATA ANALYSIS TECHNIQUES

Big data analysis techniques have been reaching lots of awareness for what they can reveal about clients, market trends, marketing programs, equipment presentation and other business fundamentals. For many IT decision makers, big data analytics tools and technologies are now a top priority. These stories highlight trends and perspectives to help you manage your big data implementation. However, techniques listed below may be applied to big data and, in commonly,

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

larger and more spread datasets can be used to generate more numerous and insightful results than smaller, less diverse ones.

A) **Association rule learning**: Association rule learning is a method for discovering interesting correlations between variables in large databases. It was firstly used in large chains of supermarket to determine interesting relations between products, using data from supermarket POS (point-of-sale) systems.

B) **Classification tree analysis:** Statistical classification is a method of identifying categories that a new observation belongs to. In other words, It requires a training set of correctly identified observations like historical data.

C) **Genetic algorithms:** Genetic algorithms are inspired by the way evolution works – that is, through mechanisms such as inheritance, mutation and natural selection. These mechanisms are used to “evolve” useful solutions to problems that require optimization.

D) **Machine learning:** Machine learning includes software that can learn from data. It gives computers the ability to learn without being explicitly programmed, and is focused on making predictions based on known properties learned from sets of “training data.”

E) **Regression analysis:** regression analysis involves manipulating some independent variable (i.e. background music) to see how it influences a dependent variable (i.e. time spent in store). It describes how the value of a dependent variable changes when the independent variable is varied. It works best with continuous quantitative data like weight, speed or age.

F) **Sentiment analysis:** it helps innovators to determine the sentiments of speakers or writers with respect to a theme, subject or matter.

G) **Social network analysis:** this is a technique which was firstly used in the telecommunications sectors, and then quickly adopted by sociologists to study interpersonal relationships. It is now being applied to analyze the relationships between people in many fields and commercial activities. Nodes represent individuals within a network, while ties represent the relationships between the individuals.

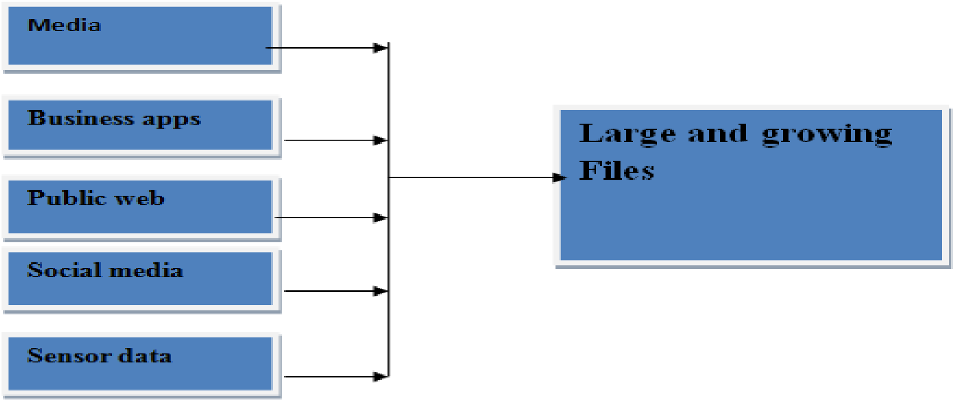
V. BIG DATA SOURCES

Media: Media exists in-out of your organization, can connect with APIs (think an API to collects images from Pinterest and insert them into a product page) and is fairly structurally ordered.

Business apps: Business apps are structurally designed, and using APIs user may pull data from both inside and outside. Internally, think integrating CRM or Web Content Management with ecommerce system. Externally, using Weather Co. or Weather Underground data for local personalization.

Public web: it is cool and useful external applications may be mashed up with it.

Social media: it is data with high velocity, high volume that may use to identify trends, analyze sentiment about brand, customer service and competitor to social accounts that match with the customer email addresses consists in file.



Fig(2): big data sources

DOI: 10.15680/IJIRCCE.2017. 0505147

VI. BIG DATA FEATURE AND SCOPE IN FUTURE

Following can be achieved using Big data technologies, because easy to store and analyze it, improving things using technology.

Health Care: If person want to take care of their future health and diseases it can be cured before he get diseased by analyzing his history data. He/she has to share information of his health changes and weekly activity using the Big Data Techniques.

Management: By Gathering information from past changes and comparing the current changes. What thing has changed company like number of employs, technologies and domain expertise, projects? How many required for future projects accordance with data and we can train and ready for it. After analyzing, what thing are effected we will come to know by past data, by comparing current data.

Product Improvement: Once product released in market we can improve it by taking a comments and feedback from the customer. By using the current, history data and suggesting the new products which are related to the product which people buys. By analyzing product will be improved and relevant product sale by giving adds suggestion.

Environment Study: We can predict the kind to changes in future environment. By using the same history current data by this we can take the preposition before it occur At least we can guise what to do, before it may happen.

Education: Taking current and past data of staff from each State and analyzing on it.

From this data get there education. Taking every parent reviews, suggestion and in which subjects there kid is weak. After analyzing the data, we can improve both staff and kids knowledge.

Future of Big Data: The Future data sets will continue to grow storage unit costs will continue to decrease Processing Costs will decrease Networks Capacity will continue to grow Data growth may exceed processing capacity .this industry on its own is worth more than $100 billion and growing at almost 10% a year which is roughly twice as fast as software business as a whole.

VII. CONCLUSION

Today Economics Environment demands that business be driven by default, accurate and timely information. The world of big data is solution to the problem. there are always business and it tradeoff to get to data and information in a most cost –effective way. In this paper I simply introduce basic concepts of Big data as well as Big data Technique. We have entered an era of Big Data. Through better analysis of the large volumes of data that are becoming available, there is the potential for making faster advances in many scientific disciplines and improving the profitability and success of many enterprises.

REFERENCES

[1] www.webopedia.com

[2] https://en.wikipedia.org/wiki/Big_data

[3] G. Noseworthy, Infographic: Managing the Big Flood of Big Data in Digital Marketing, 2012 http://analyzingmedia.com/2012/infographic-big-flood-of-big-data-in-digitalmarketing.

[4] H. Moed, The Evolution of Big Data as a Research and Scientific Topic: Overview of the Literature, 2012, ResearchTrends, http://www.researchtrends.com.

[5] Dr. Krishnakumar et al. Perspectives on big data analytics and techniques, International Journal Of Pharmacy & Technology.