A REVIEW OF CURRENT TECHNIQUES OF BIG DATA ANALYSIS

Dr.Anamika Bhargava[1], Pooja Goyal[2]

[1]*Associate Professor, DAVIM, Faridabad, Haryana(India)*

ABSTRACT

Big data is a term that is used to manage large volume of data – both structured and unstructured – that shows different pattern and structure on the basis of daily transaction and operational data .Amount of data that's not important ,important is what the organizations do with that data. Big data can be analyzed for insights that lead to better decisions and strategic business moves. Data is same weather it is 'Big-data' or 'Small-data', but every day date is increasing in exponentially in bigger consequently requires different approaches: techniques, tools & architectures to solve a New problem and old problems in a better way.' According to the Forum for Innovation, 90% of the world's data was created in the last two years. To find the value in all this data, businesses and IT have been eagerly experimenting with a host of new analytical techniques in addition to storage, processing, and integration technologies. Everyone is keenly aware of Big Data as it is at the heart of nearly every digital transformation. Big Data is a commercial success; in thousands of cases, it has increased brand loyalty, uncovered truths, predicted the future, revealed product reliability, and discovered real accountability. This paper argue about why, what, when and where use the big data technology and tell future scope of Big Data .

Keywords: Big data ,Hadoop , Volume, Velocity, Variety.

I. INTRODUCTION

Data is easier to capture, manage and access through third parties such as Face book, D&B, and others preprocessing techniques. The use of the data is rapidly changing the nature of communication, shopping, advertising, entertainment, and relationship management. It is difficult to work with huge data using relational databases and desktop statistics, requiring instead "massively parallel software running on tens, hundreds, or even thousands of servers". For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management

The process of managing big data: **Big data =Transaction+ Integration +Observation.**

Big data usually includes data sets with sizes beyond the ability of commonly-used software tools to capture, curate, manage, and process the data within a tolerable elapsed time. It is a set of techniques and technologies that require new forms of integration to uncover large hidden values from large datasets that are diverse, complex, and of a massive scale. Big data sizes are a constantly moving target, from a few dozen terabytes to many petabytes of data in a single data set. With this difficulty, a new platform of "big data" tools has arisen to handle sense making over large quantities of data, as in the Apache Hadoop Big Data Platform.

II. CHARACTERSTICS OF BIG DATA

Big Data is not just about the size of data but it includes different volume, variety and velocity of data. As per [1] these three attributes work together to form the three Vs of Big Data.

- Volume tells about the quantity doesn't sample. It just observes and tracks the result in term what happens.
- Velocity tells that how big data is often available in real-time.
- Variety tells the quality of captured data can vary greatly. Accurate analysis depends on the veracity of source data. It draws from text, images, audio, video; plus it completes missing pieces through data fusion.

Variability refers to inconsistency the data can show at times—which hampers the process of handling and managing the data effectively.

Handling the three Vs helps organizations extract the value of Big Data. The value comes in turning the three Vs into the three is:

1. Informed intuition: predicting likely future occurrences and what course of actions is more likely to be successful.

2. Intelligence: looking at what is happening now in real time (or close to real time) and determining the action to take

3. Insight: reviewing what has happened and determining the action to take.

III. HISTORY OF BIG DATA

It all started with the World Wide Web. As the web grew in the late 1900s and early 2000s, search engines and indexes were created to help locate relevant information amid the text-based content.

- In the early years, search results really were returned by humans. But as the web grew from dozens to millions of pages, automation was needed.
- Web crawlers were created, many as university-led research projects, and search engine start-ups took off (Yahoo, AltaVista, etc.).
- One such project was an open-source web search engine called Nutch – the brainchild of Doug Cutting and Mike Cafarella.
- They wanted to invent a way to return web search results faster by distributing data and calculations across different computers so multiple tasks could be accomplished simultaneously.
- During this time, another search engine project called **Google** was in progress. It was based on the same concept – storing and processing data in a distributed, automated way so that relevant web search results could be returned faster.
- In 2006, Cutting joined Yahoo and took with him the Nutch project as well as ideas based on Google's early work with automating distributed data storage and processing. The Nutch project was divided. The web crawler portion remained as Nutch.
- The distributed computing and processing portion became Hadoop (named after Cutting's son's toy elephant). In 2008, Yahoo released Hadoop as an open-source project.

- Today, As per [7] Hadoop's framework and ecosystem of technologies are managed and maintained by the non-profit Apache Software Foundation (ASF), a global community of software developers and contributors.

IV. IMPORTANCE OF BIG DATA

As per [3] ,One of the top reasons that organizations shift to Hadoop that's its ability to store and process any type of huge amount of data in any format quickly. As per [9] Volume and variety of data is constantly increasing like Megabyte =>gigabyte ->terabyte ->petabyte
Different Internet product and social media are consideration key factors. It gives other benefits also as

- **Computing Power** distributed computing model quickly processes big data. The more computing nodes you use the more processing power you have.
- **Flexibility** Unlike traditional relational databases, you don't have to preprocess data before storing it. You can store as much data as you want and decide how to use it later. That includes unstructured data like text, images and videos.
- **Fault Tolerance.** Data and application processing are protected against hardware failure. If a node goes down, jobs are automatically redirected to other nodes to make sure the distributed computing does not fail. And it automatically stores multiple copies of all data.
- **Low Cost.** The open-source framework is free and uses commodity hardware to store large quantities of data.
- **Scalability**. You can easily grow your system simply by adding more nodes. Little administration is required.

V. BIG DATA CHALLENGES

One of the very basic challenges is to understand and prioritize the data from the garbage that is coming into the enterprise. Ninety percent of all the data is noise, and it is a daunting task to classify and filter the knowledge from the noise. As per [4] the search for inexpensive methods of analysis, organizations have to compromise and balance against the confidentiality requirements of the data. The use of cloud computing and virtualization further complicates the decision to host big data solutions outside the enterprise. But using those technologies is a trade-off against the cost of ownership that every organization has to deal with. Data is piling up so rapidly that it is becoming costlier to archive it. Organizations struggle to determine how long this data has to be retained, as some data is useful for making long-term decisions, while other data is not relevant even a few hours after it has been generated. With the advent of new technologies and tools required to build big data solutions, availability of skills is a big challenge. A higher level of proficiency in the data sciences required to implement big data solutions today because the tools are not user-friendly yet. They still require computer science graduates to configure and operationalize a big data system.

VI. HADOOP, THE OPEN SOURCE HEART OF BIG DATA

Hadoop is almost synonymous with the term “Big Data” in the industry and is popular for handling huge volumes of unstructured data. With Hadoop, it is possible to store and analyze unstructured data in a much smaller time frame using the power of distributed and parallel computing on commodity hardware. More important, the line indicating the boundary of data that can be utilized and data that cannot, is dropping, leading to a much greater peak and hence, in more possible value. Together with its free license, huge community and open source techniques, many initiatives using Hadoop have emerged, also indicating its success.. Many big IT organizations started to distribute their own commercial version of Hadoop by adding enterprise support, additional functionalities and tools and even bundled with specific hardware.

As per [2] the Hadoop Distributed File System enables a highly scalable, redundant data storage and processing environment that can be used to execute different types of large-scale computing projects. For large volume structured data processing, enterprises use analytical databases such as EMC’s Greenplum and Teradata’s Aster Data Systems. Many of these appliances offer connectors or plug-ins for integration with Hadoop systems. Big Data technology can be broken down into two major components are – Hardware component, Software component. The hardware component refers to the component and infrastructure layer. The software component can be further divided into data organization and management software, analytics and discovery software, and decision support and automation software.

VII. BIG DATA ARCHITECTURE

Analogous to the cloud architectures, the big data landscape divided into four layers shown vertically in Fig-1

- **Infrastructure as a Service (IaaS):** This includes the storage, servers, and network as the base, inexpensive commodities of the big data stack. This stack can be bare metal or virtual (cloud). The distributed file systems are part of this layer.
- **Platform as a Service (PaaS):** The NoSQL data stores and distributed caches that logically queried using query languages form the platform layer of big data. This layer provides the logical model for the raw, unstructured data stored in the files.
- **Data as a Service (DaaS):** The entire array of tools available for integrating with the PaaS layer using search engines, integration adapters, batch programs, and so on in this layer.
- **Big Data Business Functions as a Service (BFaaS):** Specific industries—like health, retail, ecommerce, energy, and banking—can build packaged applications that serve a specific business need and leverage the DaaS layer for cross-cutting data functions.

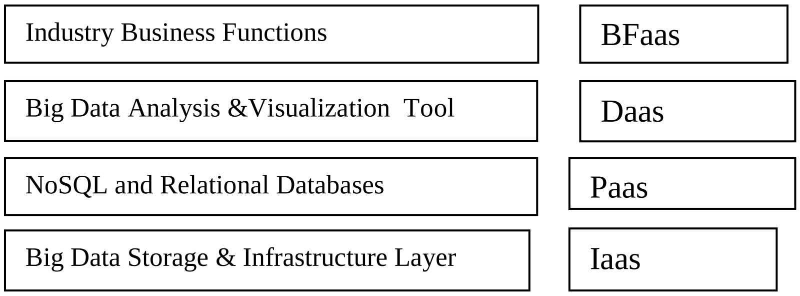
Industry Business Functions
BFaas
Big Data Analysis &Visualization Tool
Daas
NoSQL and Relational Databases
Paas
Big Data Storage & Infrastructure Layer
Iaas

Fig 1 : Big Data Architecture Layers

VIII. CONCLUSION

Today many technologies are emerging in the field of Big Data. Hadoop file system is one of them. Apache Hadoopis an open Source software framework that supports data -Intensive distributed applications, licensed under the Apache v2 license. It supports the running of applications on large clusters of commodity hardware.Big data is directed to continue rising during the next year and every data scientist will have to handle a large amount of data every year.This data will be more miscellaneous, bigger and faster. We discussed in this paper several insights about the subjects and what we think arethe major concern and the core challenges for the future. Big Data is becoming the latest final border for precise data research and for business applications.. The vital challenge is that a Big Data mining structure needs to consider complicated interaction between data sources, samples and models along with their developing changes with time and additional probable factors. A system wants to be cautiously designed so that unstructured data can be connected through their composite relationships to form valuable patterns, and the development of data volumes and relationships should help patterns to guess the tendency and future.

REFERENCES

[1] Knulst, “ De stand van Hadoop”, Incentro, 2012.

[2] Russom, “ Big Data Analytics”, TDWI Research, 2011.

[3] Bloem, J. Doorn, M. V. Duivestein, S. Manen & Ommeren, “Creating clarity with Big Data”, Sogeti, 2012.

[4] Vinayak Borkar, Michael J. Carey, Chen Li, “Inside “Big Data Management”: Ogres,Onions, or Parfaits?”, EDBT/ICDT 2012 Joint Conference Berlin, Germany, 2012.

[5] Domingos, Pedro, and Geoff Hulten. "Mining high-speed data streams."Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2000.

[7] Sawant, Nitin, and Himanshu Shah. "Big Data Application architecture “big data application architecture Q and A après, 2013.

[8] Marko Grobelnik marko.grobelnik@ijs.si

[9] Ovum. What is Big Data: The End Game. [Online] Available from: http://ovum.com/research/what-is-big-data-the end-game/ [Accessed 9th July 2012].