IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-7, Issue-5, May 2019 E-ISSN: 2347-2693

A Preface to Big Data (Overview Of Big Data)

P. Sunanda

Department of CSE, G. Pulla Reddy Engineering College (Autonomous), Kurnool, Andhra Pradesh, India

Corresponding Author: psunandareddy@gmail.com, Tel.: +91-8686597198

DOI: https://doi.org/10.26438/ijcse/v7i5.257262 | Available online at: www.ijcseonline.org

09/May/2019, Published: 31/May/2019

***Abstract*—** A collection of facts, such as values or measurements is known to be the data. Whatever the data is present it is to be stored for future reference. To store data we need the use of databases like Oracle etc.; these come under traditional databases where the structured data i.e., data that resides in a fixed field within a record or file. The problem comes in storing other form of data i.e., unstructured data e.g., photos and graphic images, videos, streaming instrument data, web pages, pdf files, PowerPoint presentations, emails, blog entries, wikis and word processing documents. This paper presents a survey on the technology "Big Data" that is used to store unstructured data.

***Keywords*—** big data characteristics, technologies, challenges, cloud computing

I. INTRODUCTION

In the current digital era, according to massive progress and development of the internet and online world , we face a huge volume of information and data day by day from many different resources and services which were not easy to handle. Networks, Cloud Storages, Social Networks and etc., produce big volume of data and also need to manage and reuse that data or some analytical aspects of the data[1]. Although this massive volume of data can be really useful for people and corporations, it can be problematic in storing this data using traditional databases like Oracle. So, we move towards the new technology called "Big Data". The rest of this paper is organized as follows. Section 2 provides the characteristics of Big Data. Section 3 provides the emerging technologies for Big Data. Section 4 provides the Big Data Users. Section 5 deals with the Challenges of Big data. Section 6 deals with Big Data Cloud Database and Computing. And Section 7 concludes the paper.

A. Definition:

Big Data is defined to be any collection of so large and complex datasets that can't be processed easily using on-hand data management tools or traditional data processing applications [2]. Datasets is a collection of related sets of information that is composed of separate elements but can be manipulated as a unit by a computer. Or it can be described as a massive volume of both structured and unstructured data which can't be stored using traditional databases.

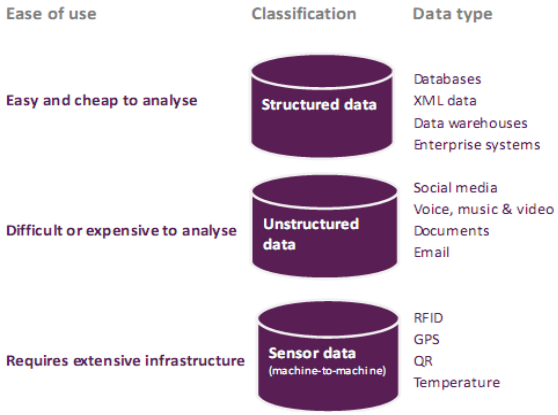

Fig: 1.1 Classification of data and Use

An example of big data might be petabytes (1,024 terabytes) or exabytes (1,024 petabytes) of data consisting of billions to trillions of records of millions of people-all from different sources (e.g. Web, sales, customer contact center, social media, mobile data and so on). The data is typically loosely structured data that is often incomplete and inaccessible [3].

B. History:

The story of how data became big starts many years before the current buzz around big data. Already seventy years ago we encounter the first attempts to quantify the growth rate in the volume of data or what has popularly been known as the "information explosion" (a term first used in 1941, according to the Oxford English Dictionary). The following are the major milestones in the evolution.

Big Data was originated in the lunch table conversations at Silicon Graphics in the mid-1990s, in which "John Mashey" figured prominently [4].

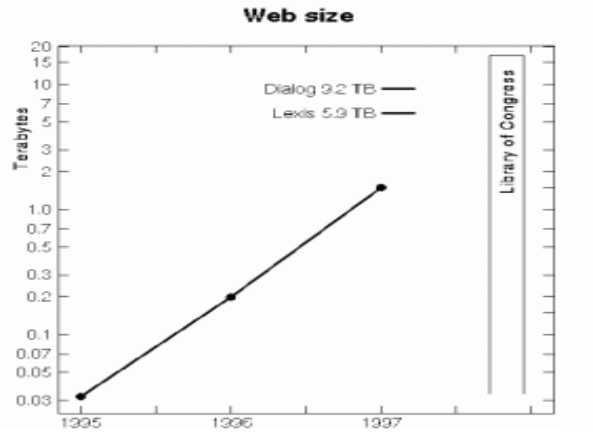

Fig 1.2 : Graph shows the growth in size of data

1997 Michael Lesk publishes "How much information is there in the world?" Lesk concludes that "There may be a few thousand petabytes of information all told; and the production of tape and disk will reach that level by the year 2000. So in only a few years, we will be able [to] save everything–no information will have to be thrown out".

November 2000 Francis X. Diebold presents to the Eighth World Congress of the Econometric Society a paper titled "'Big Data' Dynamic Factor Models for Macroeconomic Measurement and Forecasting (PDF)," in which he states "Recently, much good science, whether physical, biological, or social, has been forced to confront—and has often benefited from—the "Big Data" phenomenon.

February 2001 Doug Laney, an analyst with the Meta Group, publishes a research note titled "3D Data Management: Controlling Data Volume, Velocity, and Variety." A decade later, the "3Vs" have become the generally-accepted three defining dimensions of big data, although the term itself does not appear in Laney's note.

January 2008 Bret Swanson and George Gilder publish "Estimating the Exaflood (PDF)," in which they project that U.S. IP traffic could reach one zettabyte by 2015 and that will be at least 50 times larger than it was in 2006.

September 2008 A special issue of *Nature* on Big Data "examines what big data sets mean for contemporary science."

December 2008 Randal E. Bryant, Randy H. Katz, and Edward D. Lazowska publish "Big-Data Computing: Creating Revolutionary Breakthroughs in Commerce, Science and Society (PDF)." They write: "Just as search engines have transformed how we access information, other forms of *big-data computing* can and will transform the activities of companies, scientific researchers, medical practitioners, and our nation's defense and intelligence operations.

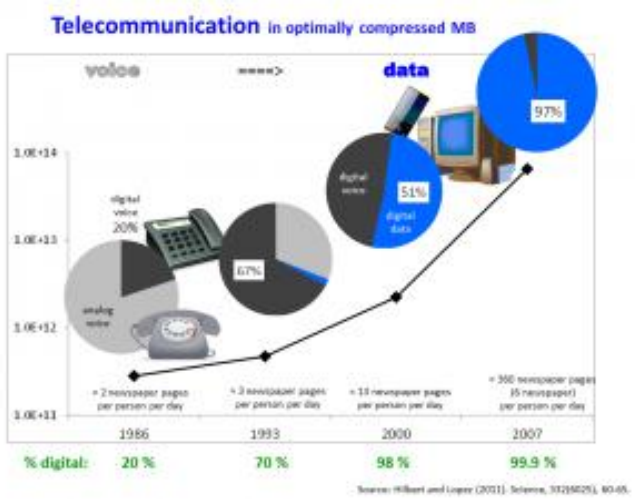

Fig 1.3: Graph show the datagrowth in telecommunication

February 2011 Martin Hilbert and Priscila Lopez publish "The World's Technological Capacity to Store, Communicate, and Compute Information" in *Science*. They estimate that the world's information storage capacity grew at a compound annual growth rate of 25% per year between 1986 and 2007.

May 2012 danah Boyd and Kate Crawford publish "Critical Questions for Big Data" in *Information, Communications, and Society*. They define big data as "a cultural, technological, and scholarly phenomenon that rests on the interplay of: technology, analysis, mythology."

II. BIG DATA CHARACTERISTICS

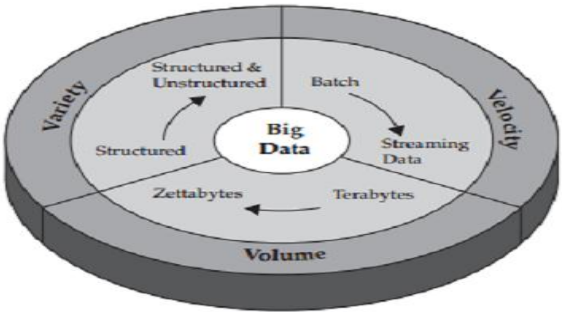

Fig: 2.1 3V Characteristics of Big Data

3V = High Volume, Velocity and Variety.

Volume: It may (but not always) involve terabytes to petabytes (and beyond) of data.

Velocity: It moves extremely fast through various sources such as online systems, sensors, social media, web click stream capture, and other channels.

- **Variety** : It's made of many types of data from many sources – structured and semi-structured, as well as unstructured (think emails, text messages, documents and the like) [5].

III. EMERGING TECHNOLOGIES FOR BIG DATA

In order to work with Big Data we need the help of technologies. Today, there are number of technologies emerging among them top 10 emerging technologies are as follows:

- ***Column-oriented databases***

Traditional, row-oriented databases are excellent for online transaction processing with high update speeds, but they fall short on query performance as the data volumes grow and as data become more unstructured. Column-oriented databases store data with a focus on columns, instead of rows, allowing for huge data compression and very fast query times. The downside to these databases is that they will generally allow batch updates, having a much slower update time than traditional models.

- ***Schema-less databases, or NoSQL databases***

There are several database types that fit into this category, such as key-value stores and document stores, which focus on the storage and retrieval of large volumes of unstructured, semi-structured, or even structured data.

- ***MapReduce***

This is a programming paradigm that allows for massive job execution scalability against thousands of servers. Any MapReduce implementation consists of two tasks:
The "Map" task, where an input dataset is converted into a different set of key/value pairs, or tuples;
The "Reduce" task, where several of the outputs of the "Map" task are combined to form a reduced set of tuples (hence the name)

- ***Hadoop***

Hadoop is an open source platform for handling Big Data. It is flexible enough to be able to work with multiple data sources, either aggregating multiple sources of data in order to do large scale processing, or even reading data from a database in order to run processor-intensive machine learning jobs.

- ***Hive***

It was developed originally by Facebook, but has been made as open source. It's a higher-level abstraction of the Hadoop framework that allows anyone to make queries against data stored in a Hadoop cluster just as if they were manipulating a conventional data store.

- ***PIG***

PIG is another bridge that tries to bring Hadoop closer to the realities of developers and business users, similar to Hive. Unlike Hive, however, PIG consists of a "Perl-like" language that allows for query execution over data stored on a Hadoop cluster, instead of a "SQL-like" language. PIG was developed by Yahoo!, and, just like Hive, has also been made fully open source.

- ***WibiData***

WibiData is a combination of web analytics with Hadoop. It allows web sites to better explore and work with their user data, enabling real-time responses to user behavior, such as serving personalized content, recommendations and decisions.

- ***PLATFORA***

PLATFORA is a platform that turns user's queries into Hadoop jobs automatically, thus creating an abstraction layer that anyone can exploit to simplify and organize datasets stored in Hadoop.

- ***Storage Technologies***

As the data volumes grow, so does the need for efficient and effective storage techniques. The main evolutions in this space are related to data compression and storage virtualization.

- ***SkyTree***

SkyTree is a high-performance machine learning and data analytics platform focused specifically on handling Big Data. Machine learning, in turn, is an essential part of Big Data, since the massive data volumes make manual exploration, or even conventional automated exploration methods unfeasible or too expensive [6].

IV. USERS OF BIG DATA

Big Data affects almost every company in the technology, media and telecoms space. Below we list a selection of TMT companies positioned to benefit from a Big Data boom.

- Accenture (USA) - IT consulting house which sees Big Data as a profitable new product line
- Adobe (USA) - Software house developing several data management products
- Amazon (USA) - Big Data pioneer. Uses in-house Big Data infrastructure for customer intelligence and AWS
- Apple (USA)-Collects Big Data via iCloud and iAd, but Big Data strategy as yet unclear
- Cisco (USA) - Redesigning its IP networking products to handle Big Data more efficiently
- Facebook (USA) -Uses in-house Big Data infrastructure for search queries and markets to advertisers.
- Fujitsu (Japan) -Makes IT, telecom and network equipment - moving into Big Data appliances

- Google (USA) -Uses in-house Big Data infrastructure for search queries.
- Infosys (India) -IT consulting house which sees Big Data as a profitable new product line
- Lenovo (China) -Hardware maker attempting to move into services. Developing Big Data appliances
- Microsoft (USA) -Uses in-house Big Data infrastructure for search queries
- Oracle (USA) -Database maker and ERP vendor that is rapidly moving into cloud and Big Data
- Red Hat (USA) -Leader in open source development - both Linux and Hadoop
- Salesforce.Com (USA) -Cloud services company moving into databases and Big Data
- TCS (India) -IT consulting house which sees Big Data as a profitable new product line
- Tencent (China) -Collects Big Data via QQ, gaming and advertising platform, but Big Data strategy unclear
- Verint Systems (Israel) - Analyses unstructured data like voice, fax, video and email.
- According to a recent IDC survey, 49% of large corporations are ready for Big Data. The three main uses for
- Big Data are:
    - *customer intelligence*
    - *risk management*
    - *fraud detection*

In the case of risk management and fraud detection, the value of Big Data is that it helps companies reach a more informed decision in real time. In the case of customer intelligence, the value of Big Data is that it allows companies to ask questions that could never be answered in the past.

V. BIG DATA CHALLENGES

If a technology is getting popular then it may have many challenges to face. In the same way Big Data is a very popular technology emerging very fastly has the following challenges:

A. *SCALE*

Whenever and wherever you want you can scale very rapidly and elastically by using Big Data.

B. *PERFORMANCE*

In an online world where nanosecond delays can cost you sales, big data must move at extremely high velocities no matter how much you scale or what workloads your database must perform. The data handling hoops of RDBMS and most NoSQL solutions put a serious drag on performance.

C. *C. CONTINUOUS AVAILABILITY*

When you rely on big data to feed your essential, revenue-generating 24/7 business applications continuous availability of data is possible.

D. *D. WORKLOAD DIVERSITY*

Big data comes in all shapes, colors and sizes. Rigid schemas have no place here; instead you need a more flexible design. You want your technology to fit your data. And you want to be able to do more with all of that data – perform transactions in real-time, run analytics just as fast and find anything you want in an instant from oceans of data, no matter what from that data may take.

E. *E. DATA SECURITY*

Big data carries some big risks when it contains credit card data, personal ID information and other sensitive assets.

F. *F. MANAGEABILITY*

Managing the Big Data is somewhat difficult and expensive.

G. *G. COST*

Working with big data the right way is an expensive task.

VI. BIG DATA CLOUD DATABASE & COMPUTING

A. *A. Background*

The rise of cloud computing and cloud data stores has been a facilitator to the emergence of big data.

Cloud computing is the co-modification of computing time and data storage by means of standardized technologies.

Cloud platforms come in several forms and sometimes have to be integrated with traditional architectures. This leads to a dilemma for decision makers in charge of big data projects. How and which cloud computing is the optimal choice for their computing needs, especially if it is a big data project? These projects regularly exhibit unpredictable, bursting, or immense computing power and storage needs.

Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction [7].

B. *B. Essential Characteristics:*

- On-demand self-service. A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed automatically without requiring human interaction with each service's provider.

- Broad network access. Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, laptops, and PDAs).
- Resource pooling. The provider's computing resources are pooled to serve multiple consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand.
- Rapid elasticity. Capabilities can be rapidly and elastically provisioned, in some cases automatically, to quickly scale out and rapidly released to quickly scale in.
- Measured Service. Cloud systems automatically control and optimize resource use by leveraging a metering capability at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth, and active user accounts) [8] [9].

C. Service Models:

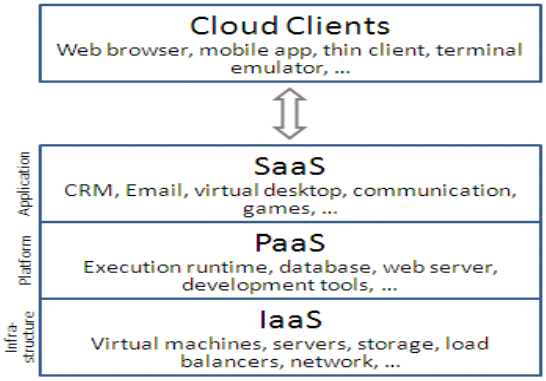

Fig 6.1: Cloud computing service model

- Infrastructure as a service (IaaS) involves offering hardware related services using the principles of cloud computing.
- Platform as a Service (PaaS) involves offering a development platform on the cloud.
- Software as a service (SaaS) includes a complete software offering on the cloud [9] [10]

D. Deployment Models:

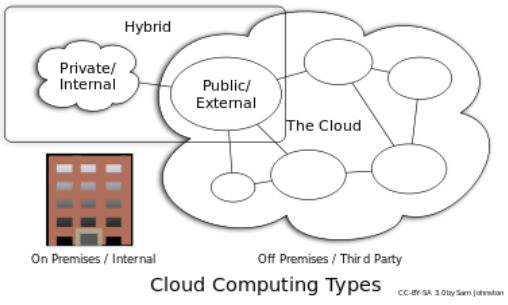

Fig 6.2: Cloud Computing Deployment Model

- Private cloud: The computing infrastructure is dedicated to a particular organization and not shared with other organizations.
- Public Cloud: Public clouds share physical resources for data transfers, storage, and processing.
- Hybrid Cloud: The hybrid cloud architecture merges private and public cloud deployments.
- Community cloud: involves sharing of computing infrastructure in between organizations of the same community [10].

C. E. Cloud Storage :

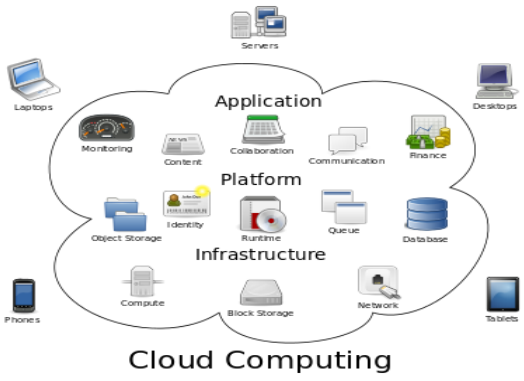

Fig 6.3: storage at various services

Professional cloud storage needs to be highly available, highly durable, and has to scale from a few bytes to petabytes. Amazon's S3 cloud storage is the most prominent solution in the space. S3 promises a 99.9% monthly availability and 99.999999999% durability per year. This is less than an hour outage per month. The durability can be illustrated with an example. If a customer stores 10,000 objects he can expect to lose one object every 10,000,000 years on average. S3 achieves this by storing data in multiple facilities with error checking and self-healing processes to detect and repair errors and device failures. This is completely transparent to the user and requires no actions or knowledge.

A company could build and achieve a similarly reliable storage solution but it would require tremendous capital expenditures and operational challenges. Global data centered companies like Google or Facebook have the expertise and scale to do this economically. Big data projects and start-ups, however, benefit from using a cloud storage service.

Cloud storage is effectively a boundless data sink. When data is copied in parallel by cluster or parallel computing processes the throughput scales linear with the number of nodes reading or writing [10].

VII. CONCLUSION

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 [11]. This survey paper provided a comprehensive study of the Big Data .In order to make Big Data more useful more technologies and tool are to be introduced for facing the security and cost challenges of Big Data.

REFERENCES

[1]. A Survey of Clustering Algorithms for Big Data: Taxonomy & Empirical Analysis A. Fahd, N. Alcatraz, Z. Tari, Member, IEEE , A. Alamri, I. Khalil A. Zomaya, Fellow, IEEE, S. Foufou, and A. Bouras
[2]. A PDF file on Big Data: Uses and Limitations by Nathaniel Schenker Associate Director for Research and Methodology National Center for Health Statistics Centers for Disease Control and Prevention Presentation for discussion at the meeting of the NCHS Board of Scientific Counselors Sept. 19, 2013
[3]. Global investment themes: telecoms, media and technology Issue No. 52 The beginner's guide to Big Data.pdf
[4]. A PDF file on a very short history on big data at www.forbes.com
[5]. http://wikibon.org/wiki/v/Big_Data:_Hadoop,_Business_Analytics_and_Beyond#A_Big_Data_Manifesto_from_the_Wikibon_Community
[6]. A PDF file on 10-emerging-technologies-for-big-data/
[7]. http://www.qubole.com/big-data-cloud-database-computing/
[8]. A PDF file on essential characteristics of cloud computing http://the cloudtutorial.com/cloudtypes.html
[9]. http://the cloudtutorial.com/cloudtypes.html
[10]. http://en.wikipedia.org/wiki/Cloud_computing
[11]. Challenges and Opportunities with Big Data -A community white paper developed by leading researchers across the United States

Authors Profile

Ms. P. Sunanda pursed Bachelor of Technology from Sri Krishnadevaraya Anatapur, in 2012 and Master of Technology from JNTUA in year 2015. She is currently working as Assistant Professor in Department of Computer Science & Engineering, G. Pulla Reddy Engineering College (Autonomous): Kurnool since 2016. Her Interested Areas: Big data, Machine Learning. She has 3 years of teaching experience.

