ISSN:0975-9646

No Science No Humans, No New Technologies No changes "Big Data a Great Revolution"

S.Sangeetha, A.K Sreeja

Assistant Professor,

Department of Computer Science and Engineering,

B.N.M Institute of Technology, Bangalore, Karnataka, India

Abstract **- Big data has become the hot IT buzzword of today's competitive world. The term 'Big Data' is originated from the fact that we are creating enormous and mounting volumes of data every day rippling across various domains and platforms. Its existence has all been hidden from the field of computer science. Big data relates to how firm the data is rather than its complexity. In this paper, we briefly describe about Big Data's progression from RDBMS, Data Mining, Image mining, Computer vision along with its various Data Storage techniques where multiple challenging tasks had been made both for software engineers as well as Infrastructure management services. This paper also illustrates the intersection of Big Data, mobile, and cloud computing creating new opportunities to become key enabler and demand for bigger, better, and faster applications. Big data has found its applications in all sectors and is thus becoming a dominant class of applications that are deployed over virtualized environments.**

Keywords **- Big Data, RDBMS, Data Mining, Image Mining, Mobile and Cloud Computing**

I. INTRODUCTION

In 2005 Roger Mougalas from O'Reilly Media coined the term Big Data for the first time, only a year after they created the term Web 2.0. It refers to a large set of data that is almost impossible to manage and process using traditional business intelligence tools. 2005 is also the year that 2005 is also the year that Hadoop was created by Yahoo! built on top of Google's MapReduce. It's goal was to index the entire World Wide Web and nowadays the open-source Hadoop is used by a lot organizations to crunch through huge amounts of data. As more and more social networks start appearing and the Web 2.0 takes flight, more and more data is created on a daily basis. Innovative startups slowly start to dig into this massive amount of data and also governments start working on Big Data projects. In 2009 the Indian government decides to take an iris scan, fingerprint and photograph of all of tis 1.2 billion inhabitants. All this data is stored in the largest biometric database in the world. In 2010 Eric Schmidt speaks at the Techonomy conference in Lake Tahoe in California and he states that "there were 5 exabytes of information created by the entire world between the dawn of civilization and 2003. Now that same amount is created every two days." In 2011 the McKinsey report on Big Data: The next frontier for innovation, competition, and productivity, states that in 2018 the USA alone will face a shortage of 140.000 – 190.000 data scientist as well as 1.5 million data managers.In the past few years, there has been a massive increase in Big Data startups, all trying to deal with Big Data and helping organizations to understand Big Data and more and more companies are slowly adopting and moving towards Big Data. However, while it looks like Big Data is around for a long time already, in fact Big Data is as far as the internet was in 1993. The large Big Data revolution is still ahead of us so a lot will change in the coming years. Let the Big Data era begin!

II. BIG DATA CHALLENGES

Every year the data transmitted over the internet is growing exponentially. By the end of 2016, Cisco estimates that the annual global data traffic will reach 6.6 zettabytes. The challenge will be not only to "speed up" the internet connections, but also to develop software systems that will be able to handle large data requests in optimal time. Many organizations are prepared to pilot and adopt big data as a core component of the information management and analytics infrastructure. Big data presents a number of challenges relating to its complexity.

- One challenge is how we can understand and use big data when it comes in an unstructured format, such as text or video.
- Another challenge is how we can capture the most important data and deliver that to the right people in real-time.
- A third challenge is how we can store the data, and how we can analyze and understand it given its size and our computational capacity. There are numerous other challenges, from privacy and security to access and deployment.

III. CHARACTERISTICS OF BIG DATA

Big Data is characterized by the following 4 Vs:

• Volume - the vast amount of data generated every second that are larger than what the conventional relational database infrastructures can cope with.

• Velocity - the frequency at which new data is generated, captured, and shared.

• Variety - the increasingly different types of data (from financial data to social media feeds, from photos to sensor data, from video capture to voice recordings) that no longer fits into neat, easy to consume structures.

• Veracity - the disarrayed data (Facebook posts with hash tags, abbreviations, typos, and colloquial speech)

IV. CLASSIFICATION OF BIG DATA

The characteristics of the big data has been much more helpful to know about how the data is collected, analyzed, and processed. Once the data is classified, it can be matched with the appropriate big data pattern:

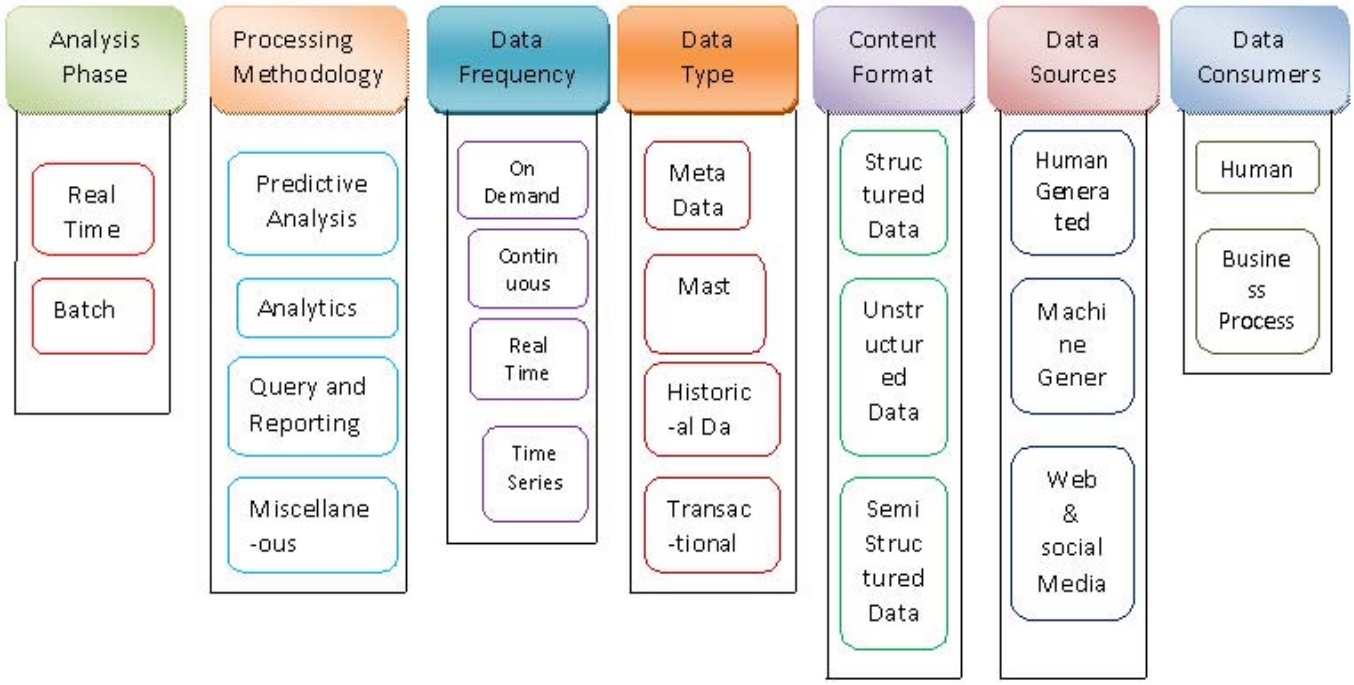

Fig 1: Classification of Big Data Characteristics

- Analysis type — Whether the data is analyzed in real time or batched for later analysis. Give careful consideration to choosing the analysis type, since it affects several other decisions about products, tools, hardware, data sources, and expected data frequency. A mix of both types may be required by the use case:
  - Fraud detection; analysis must be done in real time or near real time.
  - Trend analysis for strategic business decisions; analysis can be in batch mode.
- Processing methodology — The type of technique to be applied for processing data (e.g., predictive, analytical, ad-hoc query, and reporting). Business requirements determine the appropriate processing methodology. A combination of techniques can be used. The choice of processing methodology helps identify the appropriate tools and techniques to be used in your big data solution.
- Data frequency and size — How much data is expected and at what frequency does it arrive. Knowing frequency and size helps determine the storage mechanism, storage format, and the necessary preprocessing tools. Data frequency and size depend on data sources:
  - On demand, as with social media data
  - Continuous feed, real-time (weather data, transactional data)
  - Time series (time-based data)
- Data type — Type of data to be processed — transactional, historical, master data, and others. Knowing the data type helps segregate the data in storage.
- Content format — Format of incoming data — structured (RDMBS, for example), unstructured (audio, video, and images, for example), or semi-structured. Format determines how the incoming data needs to be processed and is key to choosing tools and techniques and defining a solution from a business perspective.
- Data source — Sources of data (where the data is generated) — web and social media, machine-generated, human-generated, etc. Identifying all the data sources helps determine the scope from a business perspective. The figure shows the most widely used data sources.
- Data consumers — A list of all of the possible consumers of the processed data:
  - Business processes
  - Business users
  - Enterprise applications
  - Individual people in various business roles
  - Part of the process flows
  - Other data repositories or enterprise applications
- Hardware — The type of hardware on which the big data solution will be implemented — commodity hardware or state of the art. Understanding the limitations of hardware helps inform the choice of big data solution.

V. ARCHITECTURE OF BIG DATA AND ITS STORAGE TECHNOLOGIES

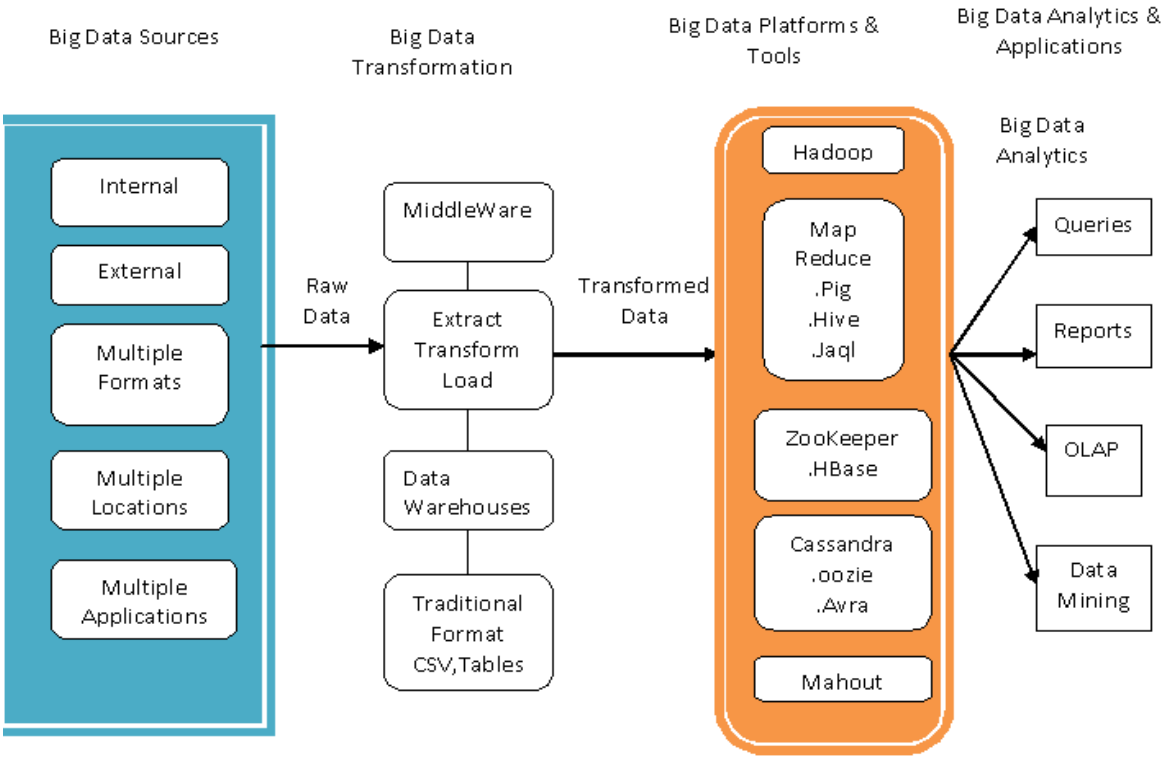

Fig 2: Architecture of Big Data and its storage technologies

Architecture of Big Data Storage techniques includes:

- Multiple clustered network attached storage(NAS) also called as scale-out NAS Clustered NAS employs storage devices attached to a network. Groups of storage devices attached to different networks are then clustered together.
- Object-based storage system distribute set of objects over a distributed storage system.
  - Hadoop is used to process unstructured and semi structured data.It uses the map reduce paradigm to locate all relevant data then select only the data directly answering the query. Hadoop works well in scale-out NAS environment. 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 MapReduce programing paradigm for managing applications on multiple distributed servers.
  - NoSql, MoongDB, and Terra Store process structured big data.
  - NoSql data is characterized by being BASE - Basically Available, Soft state(changeable), and Eventually consistent rather than the traditional db data characteristics of Atomicity, Consistency, Isolation and Durability(ACID). NoSQL focuses on a schema-less architecture (i.e., the data structure is not predefined). In contrast, traditional relation DBs require the schema to be defined before the database is built and populated.
    - Data are structured
    - Limited in scope
    - Designed around ACID principles
  - MoongDB and Terra Store are both NoSql related products used for document-oriented applications such as storage and searching of whole invoices rather than individual data fields from the invoice.
- The focus is on supporting redundancy, distributed architectures, and parallel processing
  - **Apache Avro:** designed for communication between Hadoop nodes through data serialization
  - **Cassandra and Hbase:** a non-relational database designed for use with Hadoop
  - **Hive:** a query language similar to SQL (HiveQL) but compatible with Hadoop
  - **Mahout:** an AI tool designed for machine learning; that is, to assist with filtering data for analysis and exploration
  - **Pig Latin:** A data-flow language and execution framework for parallel computation
  - **ZooKeeper:** Keeps all the parts coordinated and working together

VI. EVOLUTION OF BIG DATA IN DIFFERENT DOMAINS OF COMPUTER FIELD

A. *Transitioning from Relational databases to Big Data*

The traditional method of managing structured data includes a relational database and schema to manage the storage and retrieval of the dataset.

For managing large datasets in a structured fashion, the primary approaches are data warehouses and data marts. A data warehouse is a relational database system used for storing, analysing, and reporting functions. The data mart is the layer used to access the data warehouse. A data warehouse focuses on data storage. The main source of the data is cleaned, transformed, catalogued, and made available for data mining and online analytical functions. The data warehouse and marts are Relational databases systems[1].

B. *Evolution of Big Data in Data mining*

Big Data mining is the capability of extracting useful information from these large datasets or streams of data, that due to its volume, variability, and velocity, it was not possible before to do it. To support Big Data mining, high-performance computing platforms are required, which impose systematic designs to unleash the full power of the Big Data.

At the data level, the autonomous information sources and the variety of the data collection environments, often result in data with complicated conditions, such as missing/uncertain values. In other situations, privacy concerns, noise, and errors can be introduced into the data, to produce altered data copies. Developing a safe and sound information sharing protocol is a major challenge.

At the model level, the key challenge is to generate global models by combining locally discovered patterns to form a unifying view. This requires carefully designed algorithms to analyze model correlations between distributed sites, and fuse decisions from multiple sources to gain a best model out of the Big Data.

At the system level, the essential challenge is that a Big Data mining framework needs to consider complex relationships between samples, models, and data sources, along with their evolving changes with time and other possible factors.

C. *Evolution of Big Data in Big Data + Cloud = Analytics-as-a-Service*

The growth in Big Data, as well as the expansion in data analytics platforms in recent years such as Hadoop and NoSQL, are creating new opportunities for cloud computing to become a key enabler of Big Data analytics. Public clouds providers, such as Amazon Web Services, Google, and Microsoft, offer their own brands of big data systems in their clouds, whether NoSQL or SQL, that are cost efficient and easily scalable for businesses of all sizes. All of this points us to the reciprocal relationship between cloud and Big Data that is driven by consumer demand for bigger, better, and faster applications. In fact, the combination of Big Data and cloud computing has led to another service model known as Analytics as a Service (AaaS). This model will provide companies with faster, scalable ways to integrate, analyze, transform, and visualize various types of structured, semi-structured, and unstructured data in real time.

D. *Evolution of Big Data in Mobile + Cloud = Ubiquitous Computing*

Mobile and cloud are overlapping in so many ways that they've really blended together into what we're fast coming to know and recognize as ubiquitous computing, or the idea of computers being available anywhere. This is really all about instant, real-time services called everything as a service, or "XaaS," where X is a catch all to describe how any service you want can now be obtained in the cloud. This includes everything from cloud hosting (Dropbox) to project management (Basecamp) to email marketing (MailChimp) and much, much more. It's to the point now with Single Sign-on (SSO) through Google or Facebook that you can be up and running these services in minutes. Since mobility is the preferred channel for online engagement now, cloud and mobile are interacting and providing enormous new channels for branding, products, and services. In fact, one of the major integrations with XaaS that has everyone's attention is Internet of Things, which is spawning a host new instant service offerings that will give customers real time access on every conceivable area of their lives and surroundings. IoT is XaaS in a nutshell and the primary channel to ubiquitous computing. The continued ascendancy of mobile computing over desktops, coupled with Internet of Things and wearables, will mean that computers will soon be everywhere.

The simultaneous rise of cloud and big data technologies isn't coincidental—they're mutually reinforcing. Big data enables the cloud services we consume. For example, SaaS lets us collect data that was infeasible or impossible in a world of packaged software. An application can record every interaction from millions of users. This service in turn drives demand for big data technologies to store, process, and analyze these interactions and inject the value of the analysis back into the application through query and visualization.

E. *Evolution of Big Data in image mining*

Over more than a decade major evolution has been made in making computers learn to understand, index, and annotate pictures representing a wide range of concepts. Image mining deals with the extraction of implicit knowledge, that is, image data relationship or other patterns not explicitly stored in the images. Image mining is more than just an extension of data mining to the image domain. The fundamental role of image mining is to discover the means of an effective processing of low-level pixel representations, contained in a raw image or image sequence, to arrive at high-level spatial objects and relationships. The focus of image mining is on the extraction of patterns from a large collection of images. While there seems to be some overlap between image mining and

content-based retrieval (since both deal with large collections of images), image mining goes beyond the problem of retrieving relevant images. In image mining, the goal is to discover image patterns that are significant in a given collection of images and the related alphanumeric data. The fundamental challenge in image mining is to reveal out how low-level pixel representation enclosed in a raw image or image sequence can be processed to recognize high-level image objects and relationships.

F. Intersection of big data, mobile and cloud computing

rtunities for social The intersection of Big Data, mobile, and cloud computing has created the perfect storm for the development of innovative new applications. Mobile visualization applications are providing dynamic insight into sales, marketing, and financial data from virtually anywhere. Cloud computing services are making it possible to store virtually unlimited amounts of data and access scalable, low-cost data processing functionality at the push of a button.

Big Data, mobile, and cloud are the heavy-lifters in today's digital technologies marketplace. Each capability is providing unheard of new opportunities for innovation and creativity. One phenomenon that is happening is that we're seeing a greater level of "convergence" between Big Data, cloud, and mobile. Convergence has to do with the alignment of these capabilities in ways that are greater than the sum of their parts.

VII. APPLICATIONS OF BIG DATA

A. Big data in healthcare

Big Data plays a vital role to improve the quality and efficiency of healthcare delivery. Big Data applications are expected to have higher impact when data from various healthcare areas, such as clinical, administrative, financial, or outcome data, can be integrated. Healthcare organizations are leveraging big data technology to capture all of the information about a patient to get a more complete view for insight into care coordination and outcomes-based reimbursement models, population health management, and patient engagement. Although profit is not and should not be a primary motivator, it is vitally important for healthcare organizations to acquire the available tools, infrastructure, and techniques to leverage big data effectively or else risk losing potentially millions of dollars in revenue and profits.

B. Big Data in Finance

Big Data is seen as a years-long journey for financial organizations. Financial markets have undergone a remarkable transformation over the past two decades due to advances in technology, including faster and cheaper computers, greater connectivity among market participants, and tremendous amounts of data. Financial services, in particular, have widely adopted big data to inform better investment decisions with consistent returns. The sector is beginning to build out road maps of where Big Data could deliver the most value within this broader set of technology investments.

C. Big Data in Retail

The exponential growth of retail channels and the increasing use of social media are empowering consumers. With the information assets readily available online, consumers are now better able to compare products, services and prices. When consumers interact with companies publically through social media, they have greater power to influence other customers or damage a brand. In order for retailers to capitalize on these and other changes in the industry, they need ways to collect, manage and analyze a tremendous volume, variety, velocity and veracity of data. If retailers succeed in addressing the challenges of big data, they can use this data to generate valuable insights for personalizing marketing and improving the effectiveness of marketing campaigns, and removing inefficiencies in distribution and operations.

D. Government

Big Data can give governmental organizations access to data on a much larger extent than ever before. Governments are dealing with increasing volumes of data that have high variety of structures and helps them to increase the tax collection. Governments additionally have high potential for improving their utilization of so-called dark data; data that is available somewhere in the system but not actively used. The continuous digitalization of governmental services and communication with citizens will further accelerate the growth of data and big data technologies will play in future an important role for efficient and customer centric services.

E. Biometrics

It has become increasingly obvious that applications of Big Data are expanding immensely.

Very large-scale biometric systems are becoming mainstream in nationwide identity cards and mobile secure payment methods. As with other Big Data systems, biometric systems contend with the "four V" challenges that involve the effective managing of the complex life cycle and operations of identity information—despite the immense enrollment database size (volume) and rapid transaction response-time (velocity) requirements using potentially noisy, fraudulent (veracity), and multiple (variety) biometric identifiers. Biometric systems also provide a rich case study involving how these issues manifest and are addressed in a unique, domain-specific way. We believe that by virtue of dealing with some of the most critical entities, namely identity and entitlement, biometric systems are likely to emerge as among the most critical of the Big Data systems.

F. Agriculture

To feed the world's rapidly-expanding population in the coming decades, agriculture must produce more. Big data holds one of the keys for farmers, but it's also a weapon that could be used against them. One of the most important technologies in agriculture nowadays is to create agriculture big data. It may be created naturally, if field sensing technology is distributed widely and measured data are shared on cloud storage services. However commercial

storage services are not sustainable. Robust storage service to record permanently such important data is needed.

G. Smart cities

Cities are focussing on sustainable economic development and high quality of life, with wise management of natural resources. These applications will allow people to have better services, better customer experiences, and also be healthier. Big data is certainly enriching our experiences of how cities function, and it is offering many new interaction and more informed decision-making with respect to our knowledge of how best to interact in cities.

VIII. CONCLUSION

In this paper, we briefly discussed about the evolution of big data, its challenges, its characteristics, its classification, its architecture in various categories like hadoop, map reduce, zookeeper, Cassandra etc., which act as a good storage mechanism for big data to emerge higher in growth. This paper also explains about the relationship between big data with RDBMS, Data Mining, Image Mining, intersection of big data, mobile and cloud computing which gives brief idea of how big data play its role in different customs of various applications.

REFERENCES

[1] Sangeeta Bansal, Dr. Ajay Rana , "Transitioning from Relational Databases to Big Data", in International Journal of Advanced Research in Computer Science and Software Engineering

[2] Bernice Purcell "The emergence of “big data” technology and analytics " in Journal of Technology Research

[3] Bharti Thakur, Manish Mann, "Data Mining for Big Data: A Review" in International Journal of Advanced Research in Computer Science and Software Engineering

[4] David Feinleib, "Big Data Bootcamp" www.it-ebooks.info

[5] U. Fayyad. Big Data Analytics: Applications and Opportunities in On-line Predictive Modeling.http://big-data-mining.org/keynotes/#fayyad, 2012.

[6] Apache Hadoop, http://hadoop.apache.org

[7] Apache Mahout, http://mahout.apache.org

[8] J. Gama. Knowledge Discovery from Data Streams. Chapman & Hall/Crc Data Mining and Knowledge Discovery. Taylor & Francis Group, 2010.

[9] U. Kang, D. H. Chau, and C. Faloutsos. PEGASUS:Mining Billion-Scale Graphs in the Cloud. 2012

[10] Xindong Wu , Gong-Quing Wu and Wei Ding “ Data Mining with Big data “, IEEE Transactions on Knoweledge and Data Enginnering Vol 26 No1 Jan 2014.

[11] NoSQL Architecture a blog by Kris Zyp http://www.sitepen.com/blog/2010/05/11/nosql-architecture/

[12] Considerations for Big Data: Architecture and Approach by Kapil Bakshi – Paper published in IEEE