



Big Data Analytics

Vikas Upadhyay, Insha Shaikh

Department of Computer Engineering
Mumbai University, Mumbai, M.S., India

Abstract- Big data encompasses large size of data which has the traits of being complex and are amalgamated from an array of autonomous sources. Data is being continuously generated from various sources which in turn can be collected to carry out useful analysis and derive a feasible output. This paper attempts to define the characteristics of big data in a concise manner. The implementation of HACE theorem and a processing model which aims to implement the architecture defined by the theorem. The paper presents the different methodologies used and various real time applications in the field of Big Data.

Keywords- Big data, HACE theorem, Processing Framework, Methodologies, Applications.

I. INTRODUCTION

What is Big Data? Many Researchers and organizations have tried to define Big Data in different ways. Gartner defines; Big Data are high-volume, high-velocity and high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization[1].

Big Data is defined as the representation of the progress of the human cognitive processes, usually includes data sets with sizes beyond the ability of current technology, method and theory to capture, manage, and process the data within a tolerable elapsed time[2].

According to Wikipedia, Big Data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools[3].

Why Big Data? Data is being generated in large amounts and this ever generating data sets are not being documented or collected to be analysed further. This data comes from sensors used to gather climate information, posts to social media sites, digital pictures and videos uploaded on internet, purchase transaction records, and cell phone conversation. All this data is Big Data. The data generated by the varied sources can be studied and analysed which in turn would help in identifying the various patterns and trends among various components of any system.

II. DIMENSIONS OF BIG DATA



Fig 2 Dimensions of Big Data

Big data is characterised by the following dimensions:

1. Volume – This applies to the huge voluminous size or volume of the data that cannot be managed via traditional designates.
2. Variety – Sizable voluminous data could be in a number of different forms such as structured or un-structured, text, images, videos etc. It can emanate from a number of different sources, contrivances, or channels such as online, offline, convivial media, mobile, sensors, cameras, TV etc.

3. Velocity – This applies to the immensely colossal volume of data that can emanate from a number of different sources at a high speed that cannot be managed via traditional denotes.
4. Variability – This applies to the state or validity of the data in cognition to the time and also refers to the data in motion such as real-time streaming of data or data at rest like those stored in the database.
5. Value - This applies to the business value which deals with the handling of Big Data and also the data we are working with is valuable for society or not.

III. EXISTING SYSTEM

Relational database management systems are not capable of handling big data, has the huge volumes of data requires massively parallel software running on thousands of servers, which is a drawback in functioning of traditional database tools. Even if servers are added to traditional systems it will increase the system expenditure and there is no gurantee of accurate processing in stipulated time period.

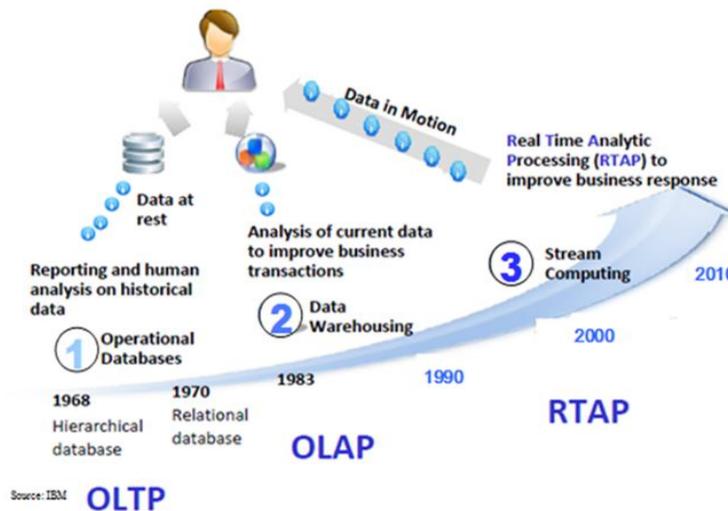


Fig 3 Traditional Database Management Tools

There are several challenges that the enterprises are faced today owing to the limitations posed by relational databases. Some of these are:

- Unstructured data that could provide a real-time business decision support remains unutilized as they cannot be stored, processed or analyzed.
- Several data islands are generated and it becomes challenging to create meaningful information from those.
- Data models are cannot be scalable and data becomes extremely difficult to manage and work with.
- The cost of data handling increases exponentially with the magnification of data

As we can see the architecture, mostly structured data is involved and is utilized for Reporting and Analytics purposes. It is observed that there are one or more unstructured sources involved, often those contribute to a very minute portion of the overall data and hence are not represented in the diagram for simplicity. However, in the case of Big Data architecture, there are various sources involved, each of which comes in at different frame of time, in different formats, and in different sizes.

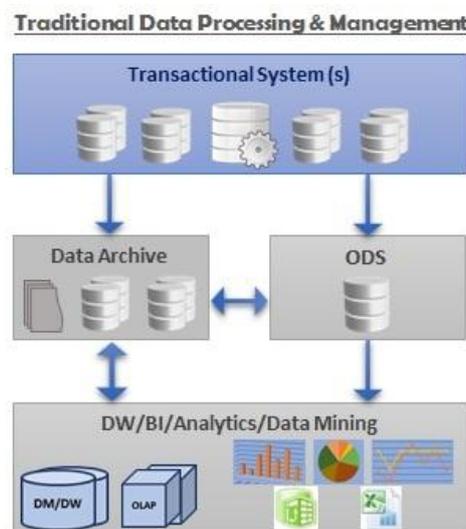


Fig 3.1 Traditional data processing management system.

The challenges include analysis, capture, search, sharing, storage, transfer, visualization, and privacy infringement. The trend to humangous data sets is due analysis of a single data set having multiple relations to other data sets rather than analysing a separate small data set having same amount of content.

IV. HACE THEOREM

The theorem was proposed in IEEE paper that we have referred and it states that, Big Data starts with large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data. These characteristics make it an extreme challenge for discovering useful knowledge from the Big Data [4][5].

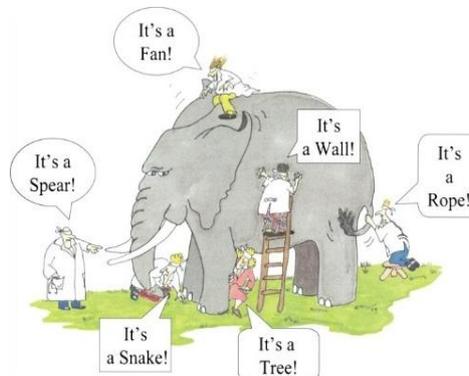


Fig 4 The blind men and the giant elephant: the localized view of each blind man leads to a biased conclusion.

The figure depicts blind men trying to conclude about the object they feel i.e. an elephant based on their localized view such as by touching its tail it may be concluded as a rope, trunk as a hose, body as a wall etc. Compare this in context of Big Data, that information is being concluded independently and also its coming from different sources in large volumes adding to uncertainties. Understanding the data flow and collection of it and using the collected data for interpretation is indeed a difficult task. Explaining the concept of HACE theorem in study of Big Data is as follows:

4.1 Huge Data with Heterogeneous and Diverse Dimensionality

Heterogeneous means from different data sources and every data collection requires a unique recording protocol. For example, each single human being in a biomedical world can be represented by using simple demographic information such as gender, age, family disease history, and so on.

4.2 Autonomous Sources with Distributed and Decentralized Control

Each data source is independent to produce and gather data without having centralized control. But, the huge volumes of the data also make an application susceptible to attacks or malfunctions, if the whole system has to rely on any centralized control unit.

4.3 Complex and Evolving Relationships

Complexity increases with varied sources of data collection and constantly increasing data size and its incoming speed evolving into relationships such as one-to-many, many-to-many. This gives useful patterns and acts as insight for analysing the data.

V. PROPOSED SYSTEM

The referred paper also give insights which shows a conceptual view of the Big Data processing framework, which includes three tiers from inside out with considerations on [5]:

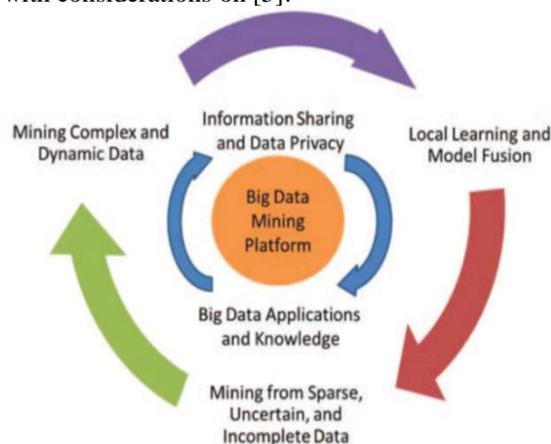


Fig 5 Big Data Processing Framework

5.1 Data accessing and computing (Tier I)

The challenges at Tier I focus on data accessing and arithmetic computing procedures. As Big Data is often stored at different locations and data volumes may continuously grow, an effective computing platform will have to take distributed large-scale data storage into consideration for computing.

5.2 Data privacy and domain knowledge (Tier II)

The challenges at Tier II center on semantics and domain knowledge for different Big Data applications. For example, depending on different domain applications, the data privacy and information sharing mechanisms between data producers and data consumers can be significantly different. Therefore, understanding semantics and application knowledge is important for both low-level data access and for high-level mining algorithm designs.

5.3 Big Data mining algorithms (Tier III)

At Tier III, the data mining challenges concentrate on algorithm designs in tackling the difficulties raised by the Big Data volumes, distributed data distributions, and by complex and dynamic data characteristics. The circle at Tier III contains three stages. First, sparse, heterogeneous, uncertain, incomplete, and multisource data are pre-processed by data fusion techniques. Second, complex and dynamic data are mined after pre-processing. Third, the global knowledge obtained by local learning and model fusion is tested and relevant information is feedback to the pre-processing stage. Then, the model and parameters are adjusted according to the feedback. In the whole process, information sharing is not only a promise of smooth development of each stage, but also a purpose of Big Data processing.

VI. SYSTEM ARCHITECTURE

Modern (Next Generation) Data Processing & Management

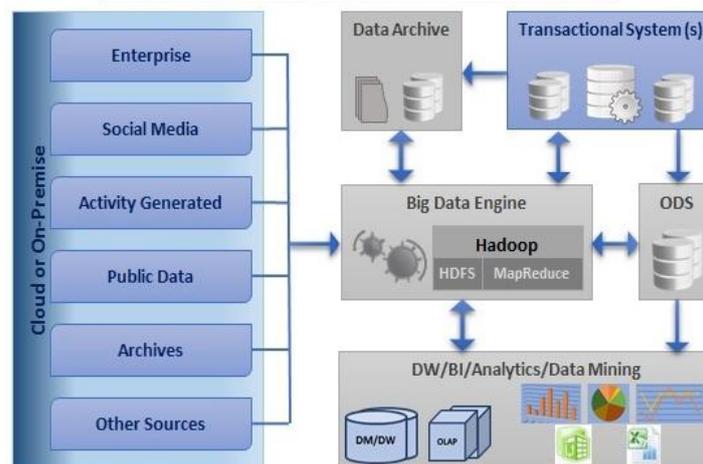


Fig 7 Big Data Processing and Management

This is the Big Data processing and management system architecture. This engine is capable of processing large volumes of data ranging from a few Megabytes to hundreds of Terabytes or even Petabytes of data of different varieties, structured or unstructured, coming in at different speeds and intervals. This engine consists primarily of a Hadoop framework, which allows distributed processing of large heterogeneous data sets across clusters of computers. This framework consists of two main components, namely HDFS and MapReduce.

VII. METHODOLOGIES

There are varieties of applications and tools developed by various organizations to process and analyze Big Data. The Big Data analysis applications support parallelism with the help of computing clusters. These computing clusters are collection of hardware connected by ethernet cables. The following are major methodologies in the area of Big Data analytics [6][7].

1. NoSql Databases: It stands for “Not Only SQL” and it differs from relational database structure by incorporating usage of wide columns, key value structure as these features are more easily managed.
2. MongoDB: It is a document orientated, based on JSON, database that can handle large number of data sets with a low maintenance and that is easy to work with.
3. Cassandra: It was originally a Facebook project, after it was release as open source. Cassandra is key and column orientated and is in many ways similar to the classic databases. It is also very close to the Google’s Big Table, offering column indexes, strong support for denormalization and materialized views.
4. BigTable: It is the solution used by Google, it is defined like a distributed store system used for managing structured data that is designed to a very large scale. As a data model, Bigtable uses a sparse, distributed and persistent multidimensional sorted map. This map is indexed by row key, column key and timestamp so that every value in the map is an uninterrupted array of bytes.

5. HBase: It is designed as an open sourced clone to the BigTable, and is very similar in most of its models and designs, supports the same data structures tables. HBase is integrated in the Hadoop project, so is easy to work using the database from a Map Reduce job.

6. MapReduce model: It is a programming model that has the purpose to process large data sets in parallel. MapReduce is a programming model for computations on massive amounts of data and an execution framework for large scale data processing on clusters of commodity servers. It was originally developed by Google and built on well-known principles in parallel and distributed processing

MapReduce program consists of two functions –Map function and Reduce function. MapReduce computation executes as follows:

1. Each Map function is converted to key-value pairs based on input data. The input to map function is tuple or document. The way key-value pairs are produced from the input data is determined by the code written by the user for the Map function

2. The key-value pairs from each Map task are collected by a master controller and sorted by key. The keys are divided among all the Reduce tasks, so all key-value pairs with the same key wind up at the same Reduce task.

3. The Reduce tasks work on one key at a time, and combine all the values associated with that key in some way. The manner of combination of values is determined by the code written by the user for the Reduce function.

7. Hadoop: It is a free, Java-based programming framework that supports the processing of large data sets in a distributed computing environment. It is part of the Apache project sponsored by the Apache Software Foundation. Hadoop was inspired by Google's MapReduce Programming paradigm

Hadoop is a highly scalable compute and storage platform. But on the other hand, Hadoop is also time consuming and storage-consuming. The storage requirement of Hadoop is extraordinarily high because it can generate a large amount of intermediate data. To reduce the requirement on the storage capacity, Hadoop often compresses data before storing it.

Hadoop takes a primary approach to a single big workload, mapping it into smaller workloads. These smaller workloads are then merged to obtain the end result. Hadoop handles this workload by assigning a large cluster of inexpensive nodes built with commodity hardware. Hadoop also has a distributed, cluster file system that scales to store massive amounts of data, which is typically required in these workloads.

VIII. APPLICATIONS OF BIG DATA

Big data applications solve and analyze real world problems using Hadoop and associated tools. Internet users and machine-to-machine connections are causing the data growth [8]. Real time areas are defined following in which big data is used:

1. Big data in healthcare:

High-performance analytics are new technologies making easier to turn massive amounts of data into relevant and critical insights used to provide better care. Analytics helps to predict disease history and its trends. Unstructured data can be captured through text mining from patient records. It means information can be collected without causing additional work for clinicians. A massive amount of data collected from different sources provides the best practices for today, and will help healthcare providers identify trends so they can achieve better results to improve medical facilities all around the world.

2. Network Security:

Big data is changing the landscape of security technologies. The tremendous role of big data can be seen in network monitoring. Big data analytics is an effective solution for processing of large scale information as security is major concern in enterprises. Fraud detection is done by using big data analytics. Phone and credit card companies have conducted large-scale fraud detection for decades. Mainly big data tools are particularly suited to become fundamental for forensics.

3. Market and business

Big Data is the biggest game-changing opportunity for sales and marketing, since 20 years ago the Internet went main stream, because of the unprecedented array of insights into customer needs and behaviours. Big data reveals customers' behaviour and proven ways to elevate customer experiences. These insights ensure your business's success.

4. Sports

Sport, in business, an increasing volume of information is being collected and captured. Technological advances will fuel exponential growth in this area for the foreseeable future, as athletes are continuously monitored. Statistics can be analyzed and collected to better understand what are the critical factors for optimum performance and success, in all facets of elite sport. Injury prevention, competition, Preparation, and rehabilitation can all benefit by applying this approach. Used consistently this is a powerful measure of progress and performance.

5. Education Systems

By using big data analytics in field of education systems, remarkable results can be seen. Data on students online behaviour can provide educators with important insights, such as if the course has to be modified or not based on students reception. This modification can be done by making students answer set of online questionnaire and track the accuracy and time taken to answer those questions.

6. Gaming industry

The amount of data that video game players are generating on a daily basis is growing rapidly. People playing video game and generated lot of data in separate areas: game data, player data and session data. In order to improve their game development, game experience, studios are turning to commercial Hadoop distributions such as MapR to analyze, collect and process data from these massive data streams. Armed with this valuable insight from big data, video game publishers are now able to enhance game player engagement and increase player retention by analyzing gamers' social behaviour, activity and tracking players' statistics, calculating rewards, quickly generating leader boards, changing game play and mechanics and delivering virtual prizes, so as to creating meaningful gaming experiences for their customers.

7. Telecommunication Industry

Today big challenges for telecommunication are volume, variety and complexity. Telcos combine ETL and traditional relational databases with big data technologies on a single platform. Telcos technology parses, transforms and integrates the vast amount of data generated by location sensors, IPv6 devices, 4G networks and machine to machine monitors' information. Telcos parse and transforms from multiple formats and sources including unstructured mobile, media, web and machine monitor provide data. Telcos masking, managing and identifying sensitive data for regulatory compliance.

IX. CONCLUSION

The big data movement has energized the data mining, knowledge discovery in data bases and associated software development communities, and it has introduced complex, interesting questions for particular data sets. As an organization collects more data at this scale, formalizing the process of big data analysis will become paramount.

Big Data is all about

- Tapping into diverse data sets
- Discovering and co-relating unknown relationships within data
- Data driven insights for faster and accurate Business decisions

Big Data is a capability

- To augment enterprises existing information fabric
- Solve today's data problems
- Transform the way business is done
- Build competitive advantage in the marketplace

In this paper, we have discussed the Hace Theorem and have analysed the main research issues of Big Data processing model and given an overview of architecture used in large data sets along with some applications. We regard Big Data as an emerging trend and the need for Big Data mining is arising in all science and engineering domains.

X. FUTURE ENHANCEMENTS

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 Software Level:

1. At the data level, the information collected from various sources posses difficult conditions such as uncertain values, introduction of noise and errors therefore altered data copies should be generated to produce data integrity is considered to be a major challenge.
2. At the model level, the key challenge is to generate global models by combining locally discovered patterns to form a unifying view. Therefore algorithms should be designed to analyse model correlations to gain the best output.
3. 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.

At Business Level:

Many major companies are investing in Big Data management which will help in analysis of data and gain insights on product usage and consumer behaviour which is indeed beneficial for any business enterprise.

The study of Big Data has the potential to create huge economic value for years to come. But the process to achieve it posses big challenges, but if efforts are directed in right manner we will see a vast changes in fields of science, medicine and business. In conclusion then, big data will change the world with the ever-growing amounts of data and large-scale analytics in relation to our ability to analyze and harness big data.

REFERENCES

- [1] "Big Data: science in the petabyte era," Nature 455 (7209):1, 2008.
- [2] Douglas and Laney, 2008, "The importance of 'Big Data': A definition".
- [3] <http://en.wikipedia.org/wiki/Big-data>
- [4] Bharti Thakur, Manish Mann, Volume 4, Issue 5, May 2014, "Data Mining For Big Data", International Journal of Advanced Research in Computer Science and Software Engineering: 469-473.

- [5] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, and Wei Ding, VOL. 26, NO. 1, JANUARY 2014, “Data Mining with Big Data”, IEEE Transactions on Knowledge And Data Engineering: 97-105.
- [6] Manish Kumar Kakhani, Sweeti Kakhani and S.R. Biradar, Volume 2 Issue 8, August 2013. “Research Issues in Big Data Analytics”, International Journal of Application or Innovation in Engineering and Management: 228-232.
- [7] Mircea Răducu TRIFU, Mihaela Laura IVAN, vol V, no 1/2014, “Big Data: present and future”, Database Systems Journal.
- [8] Sabia,SheetalKalra, 2319-2526, Volume-3, Issue-5,2014 “Applications of big Data: Current Status and FutureScope.”