Volume 5, Issue 7, July 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

# Insight on Big Data: Scaling Paradigm

**Deepali M Ujalambkar**
Assistant Professor Computer Engineering,
AISSMS COE, Pune, Maharashtra, India

***Abstract—over the last few years, organizations across public and private sectors have made a tactical choice to turn big data into competitive advantage. This paper discusses a nature of Big Data that may originate from different fields to describe few like scientific, industry and social activity domains and proposes improved Big Data definition that includes the following characteristics Volume, Velocity, Variety, Value and Veracity. Big Data is having challenges related to volume, velocity and variety. In order to do so insight on scaling paradigm and Apache Hadoop is focused to make good choice in the fields of Big Data analytics, data warehousing, and business intelligence. Benefit over traditional ETL tools and challenges are also discussed***

***Keywords— Big Data Ecosystem, ETL, Hadoop, variety, veracity.***

## I. INTRODUCTION

We are flooded in data today. Data is being collected at extraordinary scale in a broad range of application areas. As per worldomertrs [1] the current international population exceeds 7.3 billion [1] and over 2 billion of these people are connected to the Internet. Furthermore, 5 billion individuals are using various mobile devices, 30 billion pieces of content are shared on Facebook, 40% projected data rate according to McKinsey [2]. As a result of this technological revolution, these millions of people are generating tremendous amounts of data through the increased use of such devices. Hence big data analytics have become increasingly important in both the academic and the business communities.

There is rapid increase in information through various sources at a rate of 10x every five years. From 1986 to 2007, the international capacities for technological data storage, computation, processing, and communication were tracked through 60 analogues and digital technologies .In 2007, the capacity for storage in general-purpose computers was 2.9 × 1020 bytes (optimally compressed) and that for communication was 2.0 × 1021 bytes These computers could also accommodate 6.4 × 1018 instructions per second . However, the computing size of general-purpose computers increases annually at a rate of 58%. [3]

As a result, organizations encounter early challenges in creating, managing, and manipulating large Datasets. The amount of information individuals create themselves — writing documents, taking pictures, downloading music, etc. is far less than the amount of information being created about them in the digital universe [5].Recently, industries become interested in the high potential of big data, and many government agencies announced major plans to accelerate big data research and applications [2].Wal-Mart handles one million customers' transactions every hour feeding data more than 2.5pettabytes.Facebook is home to 40 billion photos and generates log data of over 10 PB per month. These examples indicate that this world contains unequivocally very large amount of digital data [6] .So big data business analytics is one of the major emerging technology trend.

In a survey of the state of business analytics by Bloomberg Business week (2011), 97 percent of companies with revenues exceeding $100 million were found to use some form of business analytics [7]. While the amount of large datasets is drastically rising, it also brings about many challenging problems demanding prompt solutions: – The latest advances of information technology (IT) make it more easily to generate data. For example, on average, 300 videos are uploaded to YouTube in every minute [8]. Biological data are much more heterogeneous than those in physics. With the advent of high-throughput genomics, life scientists are starting to tackle with massive data sets, encountering challenges with handling, processing and moving information that were once the domain of astronomers and high-energy physicists[9]. Such exponentially increasingly data cause a problem of how to store and manage huge heterogeneous datasets with moderate requirements on hardware and software infrastructure. In concern with heterogeneity, scalability, real time, complexity, and privacy of big data, there is a need to efficiently "mine" the datasets at different levels during the analysis, modelling, visualization, and forecasting so as to reveal its inherent property and improve the decision making. Traditional data management systems are not capable of handling huge data feeds instantaneously. This is where big data technologies come into play.

## II. DEFINING BIG DATA AND ITS CHARACTERISTICS

Big Data has come up because we are living in society that uses the exhaustive use of increasing data technology .Data is the collection of values and variables related with each other in a particular way. In recent years the sizes of databases have increased rapidly. This has lead to a growing interest in the development of tools capable in the automatic extraction of knowledge from data [1]. Data are collected and analyzed to create information suitable for making decisions. Hence

 

data provide a rich resource for knowledge discovery and decision support. A database is an organized collection of data so that it can easily be accessed, managed, and updated.

In Wikipedia, Big data is a broad term for data sets so large or complex that traditional data processing applications are inadequate to process it. The term often refers simply the use of predictive analytics or other certain advanced methods to extract value from data, and seldom to a particular size of data set. "Big data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis" [5]. "Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making."("Gartner IT Glossary") Or "Big data is a term that describes large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information."(Tech America Foundation's Federal Big Data Commission, 2012) .Hence to summarize the above definitions it states that volume, variety, velocity, veracity and value are the main characteristics of Big Data where volume, variety and velocity are 3v's which are referred as dimensions and value refers as social value. These are the main features/characteristics indentified and are crucial factors which emphasis on how to discover value from massive, variety and rapidly changing datasets.

*A.* Volume refers the size of data now is larger than terabytes and petabytes. The large scale and rise of size makes it difficult to store and analyse using traditional tools.

*B.* Variety refers to the structural heterogeneity in a dataset. Most of the firms use various types of structured, semi-structured, and unstructured data. Structured data refers to the tabular data found in spreadsheets or relational databases. For processing of such structured data typically processing RDBMS tools is suffice. Text, images, audio, and video are examples of unstructured data, which sometimes lack the structural organization required by machines for analysis. Extensible Markup Language (XML), is a typical example of semi-structured data. XML documents contain user-defined data tags which make them machine-readable.

*C.* Velocity refers to the rate at which data are generated and the speed at which it should be analyzed and acted upon. The explosion of digital devices such as smart phones and sensors has led to an extraordinary rate of data creation and is driving a growing need for real-time analytics.

*D.* Value measures the usefulness of data in making decisions.

*E.* Veracity refers having a lot of data in different volumes coming in at high speed is worthless if that data is incorrect. Incorrect data can cause a lot of problems for organisations as well as for consumers. Therefore, organisations need to ensure that the data is correct as well as the analyses performed on the data are correct. Especially in automated decision-making systems one need to be sure that both the data and the analyses are correct.

## III. BIG DATA ANALYTICS INSIGHT

There are many reasons for information outburst. One out of them is obviously the technology. As the capabilities of digital devices increased and prices dropped, lots of information is digitizing from various sources like gadgets, sensors that were not previously available. Data is not only becoming more available but also more understandable to computers. Most of the Big Data is stuff like words, images and video on the Web and those streams of sensor data. It is called unstructured data and is not typically suitable for traditional databases. Most government agencies hold or have access to an ever increasing wealth of data including spatial and location data, as well as data collected from and by citizens [19]. Such data can be utilised in ways that have the potential to transform service design and delivery so that personalised and streamlined services can accurately and specifically meet individual's needs and which can be delivered to them in a timely manner. Key areas where big data analytics may influence are data management, personalisation of services, problem solving and predictive analysis, productivity and efficiency. The following session illustrate the scaling of various big data analytics platforms. Scaling parameter for big data is in line with its inherent property of huge volume. It is the capability of the system to become habituated to increased demands in terms of data processing. To support big data processing, scaling is bifurcated in different forms like:

- Horizontal scaling and
- Vertical scaling.

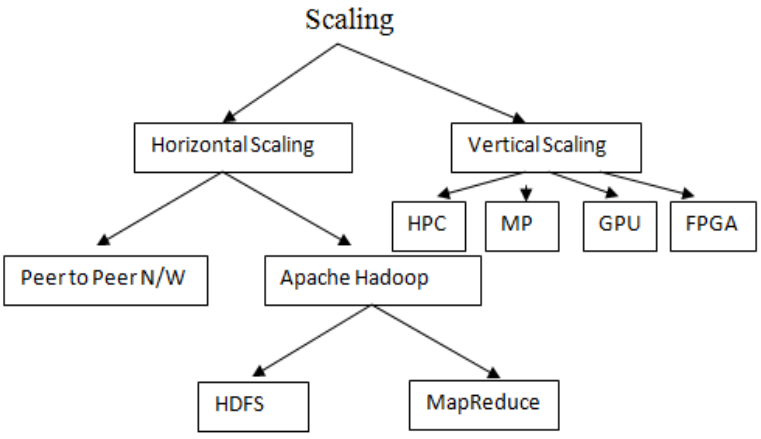


Figure 1: overview of scaling

 

Horizontal scaling involves distributing the workload across many servers and it is known as "Scale Out". Vertical Scaling involves installing more processors, more memory and faster hardware, usually, within a single server,as shown in figure 1. It is also known as "scale up". Some of the prominent horizontal scale out platforms includes peer-to-peer networks and Apache Hadoop. Mostly, Message Passing Interface (MPI) is the communication method used in such a setup to communicate and exchange the data between peers. Big data (BDMPI) is a message passing library and associated runtime system for developing out-of-core distributed computing applications for problems whose aggregate memory requirements exceed the amount of memory that is available on the underlying computing cluster. It can be used to create parallel applications.

Extract, transform, load (ETL) refers to three separate functions combined into a single programming tool. A traditional ETL process extracts data from multiple sources, then cleanses, formats, and loads it into a data warehouse for analysis. When the source data sets are large, fast, and unstructured, traditional ETL can becomes the bottleneck. Legacy systems' capacity could not store, transform, analyse and consume data from a single system. Conversely, big data solutions like Hadoop can facilitate that. Data is generated, aggregated, sorted, transformed and analysed and loaded into Hadoop as near to real time as possible. Hadoop is an ideal platform to run ETL

Apache Hadoop is an open source distributed software platform for storing and processing data. It is designed to scale up to hundreds and even thousands of nodes and is also highly fault tolerant, which is its characteristic. Using Hadoop, one can store Petabytes of data reliably on tens of thousands of servers while scaling performance cost-effectively by merely adding inexpensive nodes to the cluster [16].

Hadoop uses HDFS (Hadoop Distributed File System) and MapReduce for distributed data processing which can have structured and unstructured data. MapReduce helps programmers solve data-parallel problems in which the data set can be sub-divided into small parts and are processed independently. The system splits the input data-set into multiple chunks, each of which is assigned a map task that can process the data in parallel. Each map task reads the input as a set of (key, value) pairs and produces a transformed set of (key, value) pairs as the output. This is shown in figure 2. The framework shuffles and sorts outputs of the map tasks, sending the intermediate (key, value) pairs to the reduce tasks, which group them into final results. MapReduce uses JobTracker and TaskTracker mechanisms to schedule tasks, monitor them, and restart any that fail.

Hadoop Distributed File System (HDFS) shown in figure 3 is designed for scalability and fault tolerance. HDFS stores large files and divide them into small blocks. Then these blocks are replicated on servers. HDFS provides APIs for MapReduce applications to read and write data in parallel. Capacity and performance can be scaled by adding Data Nodes, and a single NameNode mechanism manages data placement and monitors server availability. HDFS runs across the nodes in a Hadoop cluster and together connects the file systems on many input and output data nodes to make them into one big file system. Any of these fragments or blocks can be run on any node in the cluster.

In industry, Hadoop is widely used in many big data applications, e.g., spam filtering, network searching and social recommendation engines. Hadoop is open-source software that enables reliable, scalable, distributed computing on clusters of inexpensive servers .Yahoo runs Hadoop at 42,000 servers at four data centres to support its products and services as declared in June 2012, for Ad optimization, searching and spam filtering, etc. At present, the biggest Hadoop cluster has 4,000 nodes, but the number of nodes will be increased to 10,000 with the release of Hadoop 2.0[14]. Agencies that use Hadoop are listed in [13].

For vertical scaling platforms the most popular vertical scale up paradigms are High Performance Computing Clusters (HPC), Multi core processors(MP), Graphics Processing Unit (GPU) and Field Programmable Gate Arrays (FPGA)[10].Hence, Scaling paradigm can be classified as shown in figure 1.

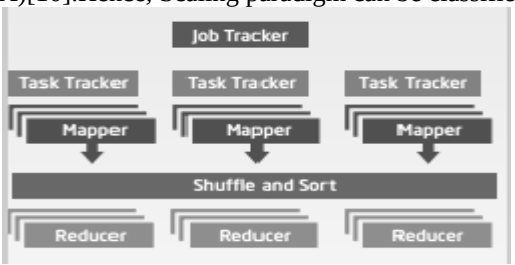


Fig 2: MapReduce

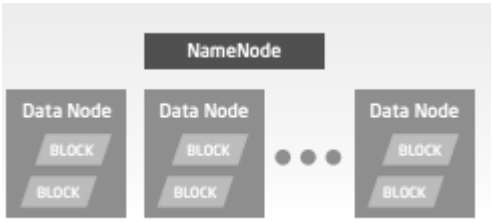


Fig 3: HDFS storage

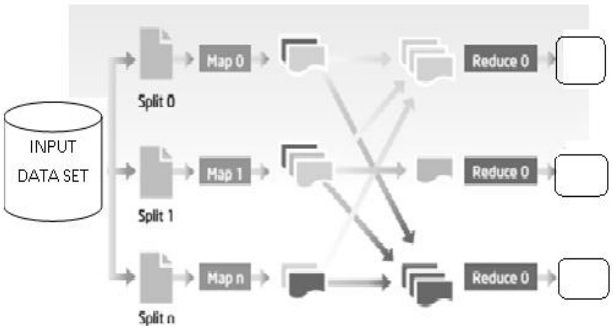


Figure 4: Hadoop Framework

 

Hadoop was inspired by Google's Map Reduce, a software framework in which an application is broken down into numerous small parts. The current Apache Hadoop ecosystem consists of the Hadoop kernel, Map Reduce, the Hadoop distributed file system (HDFS).The Hadoop framework is shown in figure 4. Hadoop ecosystem, includes. Apache Hive and Apache Pig .These are the frontrunners for extracting, loading, and transforming various forms of data. Unlike many traditional ETL tools, which are good at structured data, Hive and Pig are created to load and transform unstructured, structured, or semi-structured data into the Hadoop Distributed File System (HDFS).

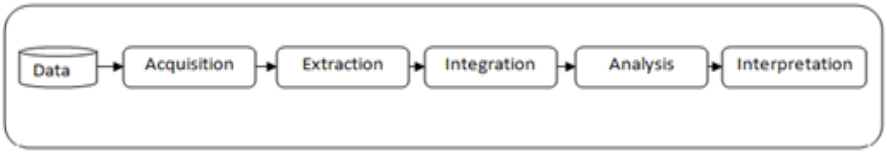


Figure 5: Big Data analysis pipeline

Figure 5 outlines the general flow about how data can be acquainted through the process of acquisition, extraction, integration and after which the big data analytic process can be followed to interpret the data which can be helpful for decision making. It runs on a cluster of industry-standard servers configured with direct-attached storage.

The challenge of Big Data is to integrate data from social media and other unstructured data into a traditional environment. The design for the systems and components that work with big data will require an understanding of both the needs of the users and the technologies that can be used to solve the problem .Not all big data and its requirements are the same [17]. It is more expensive to store and manipulate more data. As data explodes there is a need to process it and hence more computational power is required, layering on more cost. The sharply increasing data flood in the big data era brings about huge challenges on data acquisition, storage, management and analysis. Hadoop Distributed File System or HDFS lacks the ability to efficiently support the random reading of small files due to its high capacity design. As a result, it is not recommended for organizations with small quantities of data.

## IV. CONCLUSIONS

The optimistic vision of big data is that organizations will be able to crop and exploit every byte of pertinent data and use it to make the best decisions. Big data technologies not only support the ability to collect large amounts, but more importantly, the ability to understand and take advantage of its full value. Hadoop assumes that the computing capacity of each node in a cluster is the same. In such a homogeneous environment, each node is assigned to the same load, and thus it can fully use the resources in the cluster. There would not be many nodes that are idle or overloaded. However, in real-world applications, clusters are often worked in a heterogeneous environment. So there is a need to design better Data Placement Strategy due to load unbalancing. The challenges include not just the obvious issues of scale, but also heterogeneity, lack of structure, error-handling, privacy, timeliness, origin and visualization at all stages of the analysis pipeline from data acquisition to result interpretation.

## REFERENCES

[1] http://www.worldometers.info/
[2] J. Manyika, C. Michael, B. Brown et al., "Big data: The next frontier for innovation, competition, and productivity," Tech. Rep., Mc Kinsey, May 2011
[3] M. Hilbert and P. L´opez, "The world's technological capacity to store, communicate, and compute information," *Science*, vol.332, no. 6025, pp. 60–65, 2011.
[4] http://www.emc.com/collateral/analyst-reports.
[5] www.emc.com.
[6] Cukier K (2010)," Data, data everywhere", a special report on managing information, Economist Newspaper
[7] Hsinchun Chen, Roger H. L. Chiang, Veda C. Storey," *Business intelligence and analytics: from big data to big impact*", MIS Quarterly Vol. 36 No. 4, pp. 1-XX/December 2012
[8] http://www.reelseo.com/
[9] http://www.nature.com/nature/journal/v498/n7453/full/498255a.html
[10] Hadoop. http://hadoop.apache.org/
[11] Dilpreet Singh and Chandan K Reddy," *A survey on platforms for big data analytics*", Journal of Big Data 2014, Springer open access journal.
[12] Hadoop, "Hadoop," 2009, http://hadoop.apache.org/.
[13] https://wiki.apache.org/hadoop/PoweredBy
[14] http://www.informationweek.com/database/yahoo-and-hadoop-in-it-for-the-long-term/d
[15] Julie M. David, Kannan Balakrishnan, (2011), "*Prediction of Key Symptoms of Learning Disabilities in School-Age Children using Rough Sets*", Int. J. of Computer and Electrical Engineering, Hong Kong, 3(1), pp163-169
[16] White Paper: "*Extract, Transform, and Load Big Data with Apache Hadoop*"
[17] Stonebraker, M. and J. Hong. 2012. "*Researchers' Big Data Crisis; Understanding Design and Functionality*", Communications of the ACM, 55(2):10-11
[18] http://www.cra.org/ccc/files/docs/init/bigdatawhitepaper.pdf
[19] Australian Government, Department of Finance and deregulation,(March 2013) "*Big Data Strategy – Issues Paper*".

 