A survey on Big Data Analytics and Visualization

Ravichandra H[*1] **Dr. Amutha S**[*2]
[*1]Assistant Professor, Dept of CSE [*2]Professor, dept. of CSE
[*1,2]Dayananda Sagar College of Engineering Bangalore, India

***Abstract*— Technological impact is generating huge volumes of data from traditional and digital sources. Current technological revolutions such as social media allow us to generate facts plenty faster than ever before. Big data mainly deals with 5 V's and data flow pattern is a Process of data collection, storage, optimization, analytics and reports. Parallel computing, Hadoop distributed file system, Map reduce are few techniques which efficiently handle data processing techniques. This paper briefs about the data handling methods and architecture of Parallel computing, Hadoop distributed file system, Map reduce.**

Keywords— Parallel computing, Hadoop distributed file system, Map reduce.

I. INTRODUCTION

Advances in technologies have a greater impact in the field of data science in which, data is generated continuously from traditional and digital sources and are referred as big data. Characteristics features include volume, velocity, variety, variability and veracity of data. Processing huge quantity of data using on-hand database management tools or any traditional data processing applications is very hard and hence, an appropriate framework for big data solution is much essential.

'Big data' is data that exceeds the processing capacity of conventional database systems. Current technological revolutions such as social media allow us to generate facts plenty faster than ever before [1]. Analysis of collected data from single or multiple sources leads to the exploration of new structures, methods, and applications for effective big data management.

The perception of massive facts and its application in business intelligence have attracted widespread interest in latest years due to its great capacity in generating business influences [2]. On the other hand, integrating varied information from different sources provides a view of the domain and generates more accurate marketing intelligence.

A. BIG DATA CHARACTERISTICS AND ISSUES

The main characteristics of big data are defined with 5 V's attributes shown in Figure 1. Classifications of the data are defined with these attributes with corresponding applications.

A. Characteristics

Volume: Big data implies enormous volumes of data which is grouped under static or real-time applications and corresponding analysis can be considered. Origin of data will be from different sources and generated from machines, networks and human interaction on systems like social media constitute the volume of data. Variety: of data to this context refers to many sources and data types which can be structured and unstructured, depending on the frame work. Data originated from sources like spreadsheets and databases accumulated from emails, photos, videos, monitoring devices, PDFs, audio, etc. constitute different types of data. Varieties of unstructured and structured data pose a problem for storage, mining and analyzing, in turn needs attention.

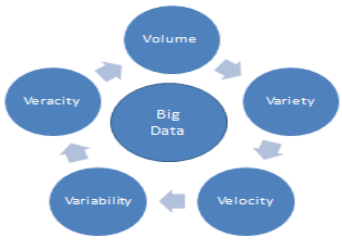

Figure 1 : Characteristics of Big data

Velocity: is the rapidity at which data flows in from different sources like a business processes, machines, networks and human interaction like social media sites, mobile devices, and etc. Flow of data is considered to be massive, continuous or static. Actual-time records can help researchers and agencies make valuable selections that provide strategic aggressive benefits.

Variability: mainly deals with applications as the data appropriate for one application may be useless in other. Hence consistency of data is difficult to obtain or maintain to meet the needs of the applicant. One of the biggest challenges of data is variability of being stored and mined for meaningful analysis and experiment. Veracity: of data is key to make right decisions that deal with defining the valid data with application. Big data veracity refers to how long is data valid and how long should it can be stored.

B. ISSUES

Processing speed: One major characteristic of Big data is volume, in which huge amount of data accumulates with respect to time. Processing speed defines the time taken for a query to complete its execution from the collected data. Issues of this kind to be noted and observation to be stated before processing.

Data Quality: Refers to the level of quality Data depending on the application. Data is commonly considered high-quality if, they're match for the intended uses in operations, choice making and decision plans.

Storage: is a method of preserving the information in need for processing in further, with appropriate queries on execution of data in which portable, semi-portable and Volatile are few popularly data storage methods known. Transportation and security: Certain domains like social media and health information accumulate huge

data about individuals. An efficient algorithm has to be developed in personalizing the data about individuals by providing security, is a key research problem.

II. THE FLOW OF MASSIVE RECORDS VENTURE

Data beyond the capability of storage, manage, and process can be termed as big data on which technological concepts can be applied for execution [3]. Process of data collection, storage, optimization, analytics and reports follow a pre defined steps on overall work flow of big data shown in Figure 2.

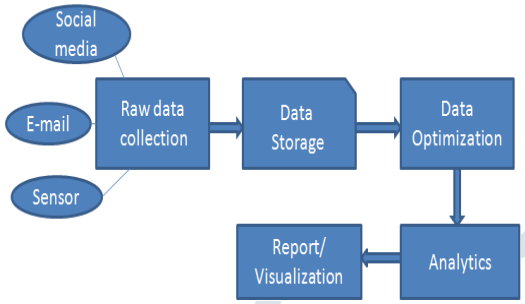

Figure 2: The flow of massive records venture

A. Data collection

Data collection is a schematic way of gathering information from a variety of sources to get a complete and accurate analysis of the result. A major source of data collection is through internet's social media like face book, twitter or any e-mail, mobile technology, online survey tools or global positioning systems of which data may be static or dynamic followed by structured or unstructured. Observing behaviors of data participating in the query, an appropriate questionnaire method or model is to be practiced in which interview modeling of scheduled data and methods for information retrieval for corresponding purpose has to be mentioned. Steps above constitute to collection of data which can be further classified into primary and secondary data.

B. Data Storage

Data storage is a technique for archiving data in a device or other forms for use in applications. It refers to storage and management of huge data collected for processing on well defined storage infrastructure for storage services and query processing. Data Storage is frequently used to mean the devices and data connected to the computer through input/output operations and to process the data by use of appropriate query. Storage are broadly classified into two types 1) Primary Storage, that store's data in the memory. 2) Secondary Storage, that store's data on hard disk. The capacity of RAM increases the speed of accessing data through devices.

C. Data Optimization

Optimization is a technique to reduce the redundancy of the data, most commonly known as a method of extracting data used by several applications in retrieving from a data sources to its destination. Duplicate and unnecessary data are to be eliminated by applying methods of optimization which helps in the fast process for fetching data. A logical schema from the data view is prepared and then mapped with an essential aspect for data integration. This system of information optimization includes data transformation or statistics mediation between a data supply and its vacation spot. Statistics mapping as a way of data optimization should translate records between diverse styles of statistics kinds and presentation formats into a unified layout used in one-of-a-kind file technology. Several techniques are available for data optimization. Mainly focused techniques are parallel computing, Hadoop distributed file system, Map Reduce.

a) Parallel computing:

Parallel data processing includes a number of interconnected processing elements which are ready to execute a single stream of instructions on multiple streams of data. Every processing detail has reminiscence capable of storing a number of bits and a mathematics unit able to processing some of bits in keeping with instruction [7]. Parallel data processing units are used independently, and operations are performed concurrently by the two processing units on different data within common memory. Conventional facts mining algorithms have been not able to fulfill huge statistics's analysis wishes. In view of this, big data mining and parallel algorithms can be used. One best parallel computing algorithm is PrePost and is based on Hadoop platform which improves the efficiency of algorithm by adding a prefix pattern, and on this basis into the parallel design ideas [6].

b) Hadoop distributed file system:

Hadoop distributed file system was invented by Google File System (GFS) [8]. HDFS is built on master slave architecture which logically separates file system and application data. This technique mainly divides the data by distributing storage and computation across many servers. As the resource growth is very high, the accumulated data will be equally divided and processed among the servers [9]. The main advantage of HDFS compared with the traditional file system is 1. High fault tolerance. 2. Large scale data processing. The Figure 3 defines the general work flow architecture of Hadoop system. Request from clients are termed as job tracker which enters the environment of Hadoop Cluster and creates a task tracker. Master and slave nodes control task tracker and requests from the clients are pleased. [13]

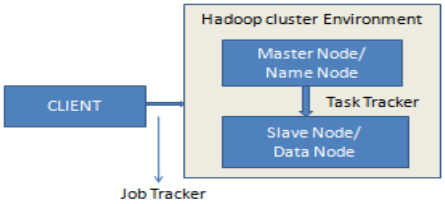

Figure 3: General Hadoop Architecture

c) Map reduce:

Map reduce is a technique of optimization of data in which functionality is divided into two phases, 1. User defined program called map and reduce 2. Framework for executing a possibly large number of instances of program. The map program reads a set of "records" from an input file, does any desired filtering and transformations, and then outputs a set of records of the form (key, data). As the map program produces output records, a "split" function partitions the records into M disjoint sets by applying a function to the key of each output record. The map program terminates with M output files [10]. Figure 4 shows the general architecture of map-reduce and data flow process as optimization technique.Face recognition is a technology the world needs in various

fields. The face recognition has been able to solve various pressing issues like - fraudulent passports, identification of criminals, prevention of fraudulent voting, banks and many more such applications.

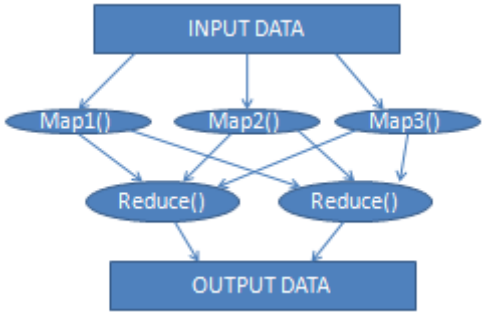

Figure 4: General work flow of Map-reduce

III. BIG DATA ANALYTICS

Analytics is the sub process defined after Data Management that consists of Modeling and analysis of data along with Interpretation on optimized data to obtain meaning full results. The technical definition of analysis includes analyze and acquire intelligence from big data through 'insight extraction'. Figure 5 shows the general application of how analytics and analysis differ and useful.

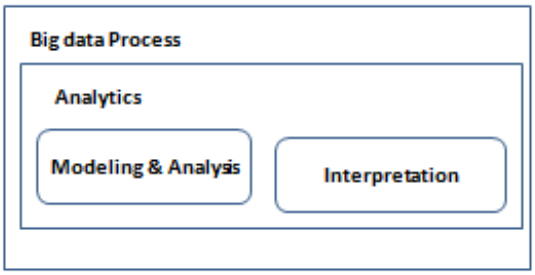

Figure 5: Analytics and Analysis frame work

Widely used analysis methods or tools includes Recommendation System, Deep learning and Network analysis which find places in applications like Deep learning, Network analysis, Topological analysis, Guilt-by-association, Shortest path analysis.

A. Report/ Visualization of results:

Data visualization is the presentation of data in a pictorial or graphical format. Huge amount of data collected and analyzed help decision makers at all levels support data visualization to predict analytical results. The advantage of report visualization prompts to identify areas that need attention or improvement, understand the factors influence customers' behavior, product placement, sales volumes and revenues. Interactive visualization is gaining importance over generating reports by providing the customers with a basic level of understanding.

REFERENCES

[1] A. McAfee, E. Brynjolfsson, T.H. Davenport, D. Patil, D Barton, Big data, the management revolution, Harv. Bus. Rev. 90(10) (2012) 61–67.

[2] H. Chen, R.H. Chiang, V.C. Storey, Business intelligence and analytics: from big data to big impact, MIS Q. 36(4) 1165–1188

[3] S. Kaisler, F. Armour, J.A. Espinosa, W. Money, Big data: issues and challenges moving forward, in: 6th Hawaii International Conference on System Sciences, (HICSS), IEEE, 2013, pp.995–1004.

[4] P. Zikopoulos, C. Eaton, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, McGraw–Hill/Osborne Media, 2011.

[5] E.W. Ngai, L. Xiu, D.C. Chau, Application of data mining techniques in customer relationship management: a literature review and classification, Expert Syst. Appl. 36(2) (2009) 2592–2602.

[6] Jinggui Liao Yuelong Zhao•, Saiqin Long, MRPrePost-A parallel algorithm adapted for mining big data, 2014 IEEE Workshop on Electronics, Computer and Applications omputer Science and Engineering.

[7] Parallel data processing system combining a SIMD unit with a MIMD unit and sharing a common bus, memory, and system controller , patent US 5355508 A

[8] S. Ghemawat, H. Gobioff, S.-T. Leung, The Googlefile system, SIGOPS Oper. Syst. Rev. 37 (2003) 29–43.

[9] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, Robert Chansler Yahoo! Sunnyvale, The Hadoop Distributed File System, California USA {Shv, Hairong, SRadia Chansler} @Yahoo-Inc.com

[10] Map Reduce: A major step backwards The Database Column

[11] TaoHuangb,1, LiangLanc,1, XuexianFanga, PengAna,d, JunxiaMind, FudiWanga , Promises and Challenges of Big Data Computing in Health Sciences. Big Data Research 2 (2015) 2–11, www.elsevier.com/locate/bdr

[12] W.G. Axinn, L.D. Pearce, Mixed Method data Collection Strategies, Cambridge University Press, 2006.

[13] A Noval Algorithm for Distributed Data Mining in HDFS, Surendar Natarajan, Sountharrajan Sehar, 2013 Fifth International Conference on Advanced Computing (ICoAC).