IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-6, Issue-10, Oct 2018 E-ISSN: 2347-2693

Big Data Concepts and Techniques in Data Processing

B.Suvarnamukhi[1*], M.Seshashayee[2]

[1]Dept. of Computer Science, GITAM, Visakhapatnam, A.P, India
[2]Dept of Computer Science, GITAM, Visakhapatnam, A.P, India

**Corresponding Author: mukhi.suvarna@gmail.com,*

Available online at www.ijcseonline.org

Accepted: 15/Oct/2018, Published: 31/Oct/2018

Abstract**—** In digital world where data grows rapidly because of more use of internet and number of devices like, smart phones, laptops, personal and machines at a very increase rate. Big data is a pool of large amounts of data collected from various sources, such as social working sites like facebook and transactional data like online shopping sites. Big data is stored in distributed architecture framework. Hadoop is an open source framework for creating distributed applications that process huge amount of data. Hadoop keeps them all together under a single roof with their functionalities. Preparing of information can incorporate different activities like highlighting, indexing and searching. It is highly difficult to process in single machine, with huge amount of data. This paper focused on Concepts and Techniques in Big Data processing.

Keywords**—** Big Data , Distributed architecture,Problem, Hadoop,Distributed Processing

I. INTRODUCTION

Big Data is a field related with the analysis and processing of large information that is available from various sources. Although Big Data may appear as a new discipline, it has been developing for years. The storage and analysis of large datasets has been a long-standing problem. Big Data is very large and complex that it is really tough to process and analyze the data. With traditional approaches it is difficult. Effective data management for Big Data sets is not possible with traditional RDBMS (Relational database management systems). Due to the size of Big Data it is very difficult to extract the information in a proper and required manner. Bringing insights from the large amount of data is very much useful. In fact it is the raw data that is processed into information.

Understanding Big Data

There are Different Types of Data processed by Big Data. Solutions can be human-generated (or) machine-generated, despite the fact that it is the machines that produce the systematic outcomes.

Human-generated data is the result of human interaction with systems, such as online services and digital devices. Examples of human-generated data are structured data, video, Textual data. Machine-generated data is generated by software programs and hardware devices in response to real-world events. For instance, a log document catches an approval choice made by a security benefit, and a point-of-offer framework creates an exchange against stock to reflect things obtained by a client.

Models of machine-produced information incorporate web logs, sensor information .The primary kinds of data are:

•Structured data / Organized data
• Unstructured data
•Semi-structured data / semi-organized data

Organized Data (or) structured Data

Organized information fits in with an information model or composition, it captures relationship between different entities.

Unstructured Data

Information that does not comply with an information model or data schema is known as unstructured data. It is assessed that unstructured information makes up 80% of the information inside some random endeavour. Unstructured information has a quicker development rate than organized information**.**

Semi-organized Data

Semi-organized information has a characterized level of structure and consistency, yet isn't social in nature. Rather, semi-organized information is progressive or chart based. This sort of information is generally put away in documents that contain content. Structured data holds into a data model and it is often stored in a tabular form.

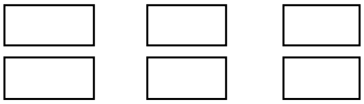

Fig 1: Structured Data

Example: Person Details like Age Gender Address, Contact Number Etc

Table 1. Organized Data

Sno	Name	Age	Gender	Contact Number
1	Mukhi	36	Female	123456789
2	Supreeth	07	Male	123321456
3	Sarayu	12	Female	900000000

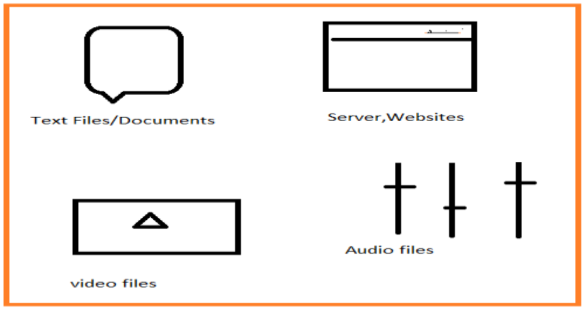

Fig 2: Unstructured Data

Types and sources of Data

Table 2. Various Types of Data

Sno	Type of Data	Source	Examples
1	Social Data	Social working sites	Facebook, Twitter, LinkedIn
2	Machine Data	Information Generated From RFID, Chips, Sensors	RFID chip,GPS
3	Transactional Data	Online shopping sites, Business to Business	Retail websites, e-bay, Amazon

Organization of the paper: In Section-I Introduction of Big Data , Section –II Big Data Processing is discussed and in Section –III Big Data Processing Techniques were discussed followed by conclusion in Section –IV.

II. Big Data Processing

In Big data processing, There are several more Big Data technologies have been developed and there are categorized into data processing concepts .In processing a lot of content must be extracted and analyzed from the collected information to serve the knowledge requirements of various business organizations, political parties and scientific research departments. The process is initiated with the retrieval of information, which can come from various sources like database, websites, documents or content management system. Hadoop, is responsible for storing massive amount of data.

Before preparing big information it must be recorded from different information creating sources. In the wake of chronicle, it must be filter and optimize. Just the pertinent information ought to be recorded by methods for channels that dispose of futile data. So as to encourage this work specific instruments are utilized, for example for example ETL. ETL method commonly combines data from multiple systems in which the data is actually loaded into the data warehouse.

Various stages in ETL

Table 3. Stages in ETL Approach

Sno	Stage	Description of stage
1	Extraction	Extraction is the initial segment of an ETL procedure which includes extracting the information from the source framework. In separating information accurately sets the phase for the achievement of ensuing procedures. The majority of the undertakings are to consolidate information with a few distinctive source frameworks.
2	Transformation	In this phase rules or principles are applied to the extracted information loads into the end target. If any information does not require any change whatsoever such information is known as "immediate move" or "go through information
3	Loading	The loading stage loads huge volume of data, loaded in a short period and should be optimized for better performance.

Challenges in Big data

Huge information because of its different properties like volume, speed, assortment, fluctuation, esteem and multifaceted nature set forward numerous difficulties.

III. Big Data Processing Techniques

Enormous information requires a few advancements to effectively process huge amounts of information within tolerable elapsed time. The Technologies being connected to enormous information incorporate hugely parallel preparing (MPP) databases, information mining matrices, conveyed record frameworks, disseminated databases, and distributed computing stages, the Internet, and versatile stockpiling frameworks. Genuine or close constant data conveyance is one of the characterizing attributes of Big Data Analytics.

Hadoop

The Apache Hadoop venture creates open-source programming for dependable, adaptable, dispersed figuring. The Apache Hadoop programming is best known for

MapReduce and its distributed file system.Hadoop framework consists of two main core components includes HDFS and MapReduce..

HDFS (Hadoop Distributed File System)
The Hadoop Distributed File System (HDFS) is a distributed file system which provides fault tolerance and designed to run on the commodity hardware. HDFS support a high throughput mechanism for accessing this large amount of information, where files are stored in the form of sequential redundant manner.

MapReduce Programming Frameworks
MapReduce is a product structure acquainted by Google in 2004 making the structured data and out of some unstructured data. MapReduce is a programming model namespace and controls access to documents by customers. Maps input key/esteem sets to an arrangement of middle of the road key/esteem sets.

The Mapper step: The master hub takes the information, allots it up into smaller sub-problems, and conveys them to worker hubs. A worker hub may do this again thus, prompting a staggered tree structure. The worker hub processes the smaller problem, and passes the appropriate response immediately to its main master hub. Map takes a list of data elements, which transforms input data to an intermediate output data element, by using Map Function. Map (A1, b1) ? List (A2, b2) Mapper reads data in the form of Key/Value pairs and its output zero or more Key/Value pairs.

Table 4. MapReduce Key, Value Pair

Phase	Input	Output
Mapper	(k,v)	(k,v)
Sort and Shuffle	(k,v)	(k,List(v))
Reducer	(k,List(v))	(k,v)

Diminish or Reduce step: The pro centre point by then assembles the reactions to all the sub-problems and goes along with them by one means or another to shape the yield. The Reduce work is then associated in parallel. It combines these values together returning a single output values in the same domain: Reduce (A2, List (b2)) ? List (b3)

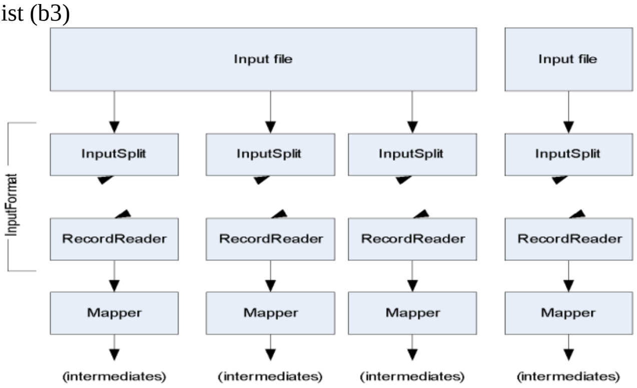

Fig 3: Getting Data to the Mapper

IV. CONCLUSION

Presently, we are travelling through Big data, As we have immense and volumes of information that are created every day, this tremendous size of data should be turned for successful handling in Expanding business and change in expectation for everyday comforts of people. This paper centres on talking about the issues, challenges, application, preparing and in addition some methods used. Future work will be on handling big data Processing methods and Techniques.

REFERENCES

[1]. Patel, Aditya B., Manashvi Birla, and Ushma Nair. "Addressing big data problem using Hadoop and Map Reduce." Engineering (NUiCONE), 2012 Nirma University International Conference on. IEEE, 2012.
[2]. A, Katal, Wazid M, and Goudar R.H. "Big data: Issues, challenges, tools and Good practices." Noida: 2013, pp. 404 – 409, 8-10 Aug,2013.
[3]. Amrit pal, Pinki Aggrawal, Kunal Jain, Sanjay Aggrawal “A Performance Analysis of MapReduce Task with Large Number of Files Dataset in Big Data using Hadoop” Forth International Conference on Communication Systems and Network Technologies, 2014.
[4] Kaur, Anureet. "Big Data: A Review of Challenges, Tools and Techniques." International journal of scientific research in science, engineering and technology 2.2 pp. 1090-1093 ,2016.
[5]. Verma, Jai Prakash, et al. "Big data analytics: Challenges and applications for text, audio, video, and social media data." International Journal on Soft Computing, Artificial Intelligence and Applications (IJSCAI) 5.1 2016.
[6]. Kaki, Gowtham, et al. "Safe Memory Regions for Big Data Processing." transfer (successorId, t, outList) pp. 17 -18 ,2016.
[7]. Big Data Black Book: Covers Hadoop 2, MapReduce, Hive, YARN, Pig, R and Data Visualization by DT Editorial Services Paperback ,2016.
[8]. Bhandari, Renu, Vaibhav Hans, and Neelu Jyothi Ahuja. "Big Data Security–Challenges and Recommendations." IJCSE pp. 93-98 ,2016.
[9]. Tanuja, A and D. Swetha Ramana. "Processing and Analyzing Big data using Hadoop." International Journal of Computer Sciences and Engineering 4.4 : pp. 91-94 ,2016.

Authors Profile

B.SuvarnaMukhi pursuing Ph.D in Dept of Computer Science at GITAM (Deemed–to-be-University) Visakhapatnam and woking as a Assistant Professor in St Marys Group of Institutions Hyderbad. Having 7 years of Teaching Experience and a Member of IAENG Interested to do research area in BIG DATA.

Dr. M.Seshashayee, Assistant Professor in Dept of Computer Science at GITAM Her Research specialization is in Image Segmentation Methods using Data Mining Techniques.Editorial Member for Honorary Editorial Board of International Journal of Computational Mathematical Ideas (IJCMI), Reviewer in International Journal of Innovations in Computer Science and Engineering.