A Survey on Big Data and Hadoop

[1]Kawale S. M., [2]Dr. Holambe A. N.
[1]Lecturure, [2] Head of Department
[1]Department of Computer Engineering,
[1]SVERI's College of Engineering (Polytechnic), Pandharpur, Pandharpur, India.
[2]Department of Computer Science and Engineering,
[2]TPCT's College of Engineering, Osmanabad, Osmanabad, India.

***Abstract* - The term 'Big Data' describes innovative techniques and technologies to capture, store, distribute, manage and analyze larger-sized datasets with high-velocity and different structures. Big data can be structured, unstructured or semi-structured, resulting in incapability of conventional data management methods. Data is generated from various different sources and can arrive in the system at various rates. In order to process these large amounts of data in an inexpensive and efficient way, parallelism is used. Big Data is a data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it. Hadoop is the core platform for structuring Big Data, and solves the problem of making it useful for analytics purposes. Hadoop is an open source software project that enables the distributed processing of large data sets across clusters of commodity servers. Hadoop File System was developed using distributed file system design, HDFS is highly fault tolerant and designed using low-cost hardware. It is designed to scale up from a single server to thousands of machines, with a very high degree of fault tolerance.**

***Index Terms*— Big Data, Hadoop, HDFS.**

I. INTRODUCTION

Today, every field is based on digitization and it grows exponentially. Due to the high growth in digitization huge amount of structured as well as unstructured data was generated and process is going on continuously. The data is being generated and collected from different sources such as, transactions, social media, sensors, retails, audios, videos, government sectors etc. For example, in facebook every month 40 billion contents are being shared. It is necessary for organizations to mine this data continuously to survive in current market trends and become a good competitor. When data is analyzed properly it helps the organizations to define current and future strategies. The conventional data processing techniques gives degraded performance while creating, managing and analyzing big data. Hadoop provides platform for structuring and managing Big Data, and making it useful for analytics purposes. Big data analytics is important and advanced analytic techniques which operate on big data for examining large amounts of data. In analytics, data divided into different sectors to assess it according to time, and compare one sector to another. With the help of big data enterprises can develop a more systematic and perceptive understanding of their business, which helps to increase the productivity and innovation.

Defining Big Data

Big data is a term that describes the large volume of data both structured and unstructured that inundates a business on a day-to-day basis. But it's not the amount of data that's important. It's what organizations do with the data that matters. Big data can be analyzed for insights that lead to better decisions and strategic business moves.

Big data is a term that refers to data sets or combinations of data sets whose size (volume), complexity (variability), and rate of growth (velocity) make them difficult to be captured, managed, processed or analyzed by conventional technologies and tools, such as relational databases and desktop statistics or visualization packages, within the time necessary to make them useful. While the size used to determine whether a particular data set is considered big data is not firmly defined and continues to change over time, most analysts and practitioners currently refer to data sets from 30-50 terabytes(10 12 or 1000 gigabytes per terabyte) to multiple petabytes (1015 or 1000 terabytes per petabyte) as big data. Layered Architecture of Big Data System. It can be decomposed into three layers, including Infrastructure Layer, Computing Layer, and Application Layer from top to bottom.

Describing Big Data via the Three Vs

Volume of data: Volume can be referred as the size of data. Large amount of data is collected from different sources such as, transactions, social media, sensors, retails, audios, videos, government sectors etc. It ranges from terabytes to petabytes.

Variety of data: Variety of data means type of data to which Big data support. Big data supports different types of data such as structured, unstructured and semi structured.

Velocity of data: Velocity considered as the speed of data capturing, processing and visualizing it. Many time-sensitive areas big data plays very important role [1].

Figure 1: Layered Architecture of Big Data

Challenges in big data

Data Acquisition and Recording

Big Data has been generated from different data capturing sources. For example, simulation and different scientific experiments easily generates petabytes. Most of these data is not useful; it needs to be filtered. The first challenge is, data needs to be filtered in such way that important data will not be loose. The second challenge is, generate the correct metadata for stored data.

Information Extraction and Cleaning

Whatever information was collected from different sources is not in required format for analysis. This information needs to be cleaned and arrange into proper format for analysis for that require an information extraction tools that fetch the required data from the essential sources.

Data Integration, Aggregation, and Representation

Big data is heterogeneous in nature, it is not easy to store it and lob it into a repository. This data needs to structure carefully so that it will be useful for data analysis. Effective management, representation, data access policies must be considered.

Query Processing, Data Modelling, and Analysis

Different methods are available for mining knowledgeable data from big data. Querying Big Data is different from traditional techniques because it is heterogeneous, dynamic and inter-related. Querying or Mining Big data requires integrated and efficiently accessible data techniques and scalable mining algorithms.

Interpretation

Analyzing Big Data is not having value if analytical information is not presented in user friendly manner. This information must be interpreted with proper visualization and clear specification. With this interpretation additional guidelines or information must be provided for better understanding. This additional information is considered as provenance of the data [1].

Hadoop: Solution for Big Data Processing

Hadoop is a Programming framework used to support the processing of large data sets in a distributed computing environment. Hadoop was developed by Google's MapReduce that is a software framework where an application break down into various parts. The Current Appache Hadoop ecosystem consists of the Hadoop Kernel, MapReduce, HDFS and numbers of various components like Apache Hive, Base and Zookeeper. HDFS and MapReduce are explained in following points.

If there is just an area that has experienced tremendous development in recent years, that will be database management. Database has evolved tremendously with companies and an individual having the opportunity to improve the manner in which it handles or rather manages data. For instance, businesses nowadays prefer to utilize the more evolved columnar database instead of the more conventional row based database. This is primarily because of the increased flexibility that columnar database offers.

Now, talking of growth in the world of applications, Hadoop apps are one of the best that have dramatically evolved over time and consequently giving businesses and people improved solutions when it comes to file systems. Specifically, Hadoop distributed file system is the main storage system which is normally utilized by Hadoop applications. It is renowned for providing high-performance data access across Hadoop clusters. Just like the rest of Hadoop applications, HDFS has also become a fundamental tool in the management of big data as well as in offering support in massive data analytics applications.

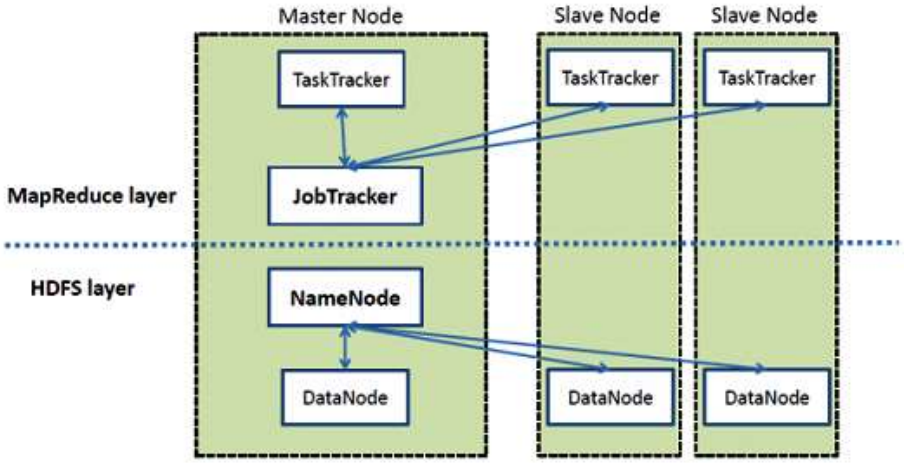
Apparently, Hadoop distributed file system is usually deployed on cheap commodity hardware. In turn, server failures are a common thing. However, the file system is originally designed to be extremely fault tolerant which is made possible by two things; Firstly, data transfer between compute nodes is made rapid and secondly, Hadoop systems are made to continue running when a compute node fails.

The aim is to enable parallel processing. Upon taking data, it breaks down the information into distinct pieces and then distributes them to the different nodes within the cluster. HDFS also copies every piece of data several times besides distributing the copies to the individual nodes. At least a copy of the data is usually placed on a separate server rack than the rest. As a result, if

data on the nodes crash, it can be accessed somewhere else within the cluster under consideration. This means that processing does not stop even when a potential failure is being resolved.

HDFS was written specifically to help provide support to applications containing large sets of data, and this includes hulking individual files which reach up to terabytes. The file system makes use of master architecture where each cluster consists of one Name Node which manages the operations of the files system besides supporting Data Nodes, which usually manage storage of data on distinct compute nodes.

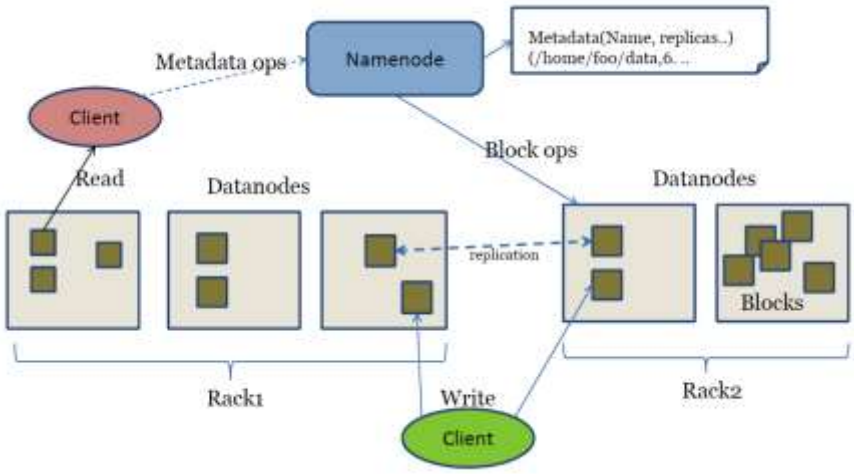
Fig. 1: Hadoop Distributed File System



1. HDFS Architecture

Hadoop includes a fault tolerant storage system called the Hadoop Distributed File System, or HDFS. HDFS is able to store huge amounts of information, scale up incrementally and survive the failure of significant parts of the storage infrastructure without losing data. Hadoop creates clusters of machines and coordinates work among them. Clusters can be built with inexpensive computers. If one fails, Hadoop continues to operate the cluster without losing data or interrupting work, by shifting work to the remaining machines in the cluster. HDFS manages storage on the cluster by breaking incoming files into pieces, called "blocks," and storing each of the blocks redundantly across the pool of servers. In the common case, HDFS stores three complete copies of each file by copying each piece to three different servers

Fig. 2: Hadoop Distributed File System



.

2. MapReduce Architecture

The processing pillar in the Hadoop ecosystem is the MapReduce framework. The framework allows the specification of an operation to be applied to a huge data set, divide the problem and data, and run it in parallel. From an analyst's point of view, this can occur on multiple dimensions. For example, a very large dataset can be reduced into a smaller subset where analytics can be applied. In a traditional data warehousing scenario, this might entail applying an ETL operation on the data to produce something usable by the analyst. In Hadoop, these kinds of operations are written as MapReduce jobs in Java. There are a number of higher

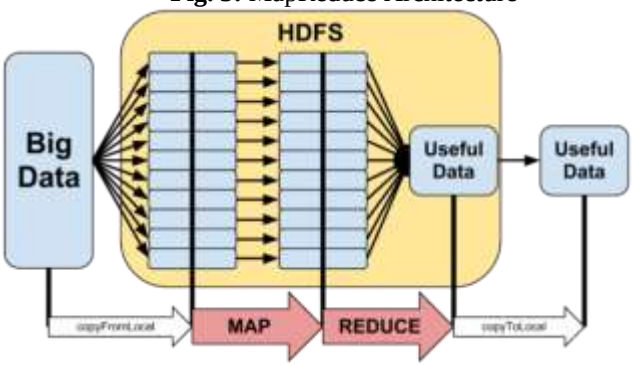
Level languages like Hive and Pig that make writing these programs easier. The outputs of these jobs can be written back to either HDFS or placed in a traditional data warehouse.

There are two functions in MapReduce as follows:

map – the function takes key/value pairs as input and generates an intermediate set of key/value pairs.

reduce – the function which merges all the intermediate values associated with the same intermediate key.

Fig. 3: MapReduce Architecture



II. LITERATURE REVIEW

S. Vikram Phaneendra & E. Madhusudhan Reddy et.al. Illustrated that in olden days the data was less and easily handled by RDBMS but recently it is difficult to handle huge data through RDBMS tools, which is preferred as "big data". In this they told that big data differs from other data in 5 dimensions such as volume, velocity, variety, value and complexity. They illustrated the hadoop architecture consisting of name node, data node, edge node, HDFS to handle big data systems. Hadoop architecture handle large data sets, scalable algorithm does log management application of big data can be found out in financial, retail industry, health-care, mobility, insurance. The authors also focused on the challenges that need to be faced by enterprises when handling big data: - data privacy, search analysis, etc [2].

Kiran kumara Reddi & Dnvsl Indira et.al. Enhanced us with the knowledge that Big Data is combination of structured , semi-structured ,unstructured homogenous and heterogeneous data .The author suggested to use nice model to handle transfer of huge amount of data over the network .Under this model, these transfers are relegated to low demand periods where there is ample ,idle bandwidth available . This bandwidth can then be repurposed for big data transmission without impacting other users in system. The Nice model uses a store –and-forward approach by utilizing staging servers. The model is able to accommodate differences in time zones and variations in bandwidth. They suggested that new algorithms are required to transfer big data and to solve issues like security, compression, routing algorithms [3].

Jimmy Lin et.al. used Hadoop which is currently the large –scale data analysis " hammer" of choice, but there exists classes of algorithms that aren't " nails" in the sense that they are not particularly amenable to the MapReduce programming model . He focuses on the simple solution to find alternative non-iterative algorithms that solves the same problem. The standard MapReduce is well known and described in many places .Each iteration of the pagerank corresponds to the MapReduce job. The author suggested iterative graph, gradient descent & EM iteration which is typically implemented as Hadoop job with driven set up iteration &Check for convergences. The author suggests that if all you have is a hammer, throw away everything that's not a nail [4].

Wei Fan & Albert Bifet et.al. Introduced Big Data Mining as 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. The author also started that there are certain controversy about Big Data. There certain tools for processes. Big Data as such hadoop, strom, apache S4. Specific tools for big graph mining were PEGASUS & Graph. There are certain Challenges that need to death with as such compression, visualization etc.[5].

Albert Bifet et.al. Stated that streaming data analysis in real time is becoming the fastest and most efficient way to obtain useful knowledge, allowing organizations to react quickly when problem appear or detect to improve performance. Huge amount of data is created everyday termed as " big data". The tools used for mining big data are apache hadoop, apache big, cascading, scribe, storm, apache hbase, apache mahout, MOA, R, etc. Thus, he instructed that our ability to handle many exabytes of data mainly dependent on existence of rich variety dataset, technique, software framework [6].

Bernice Purcell et.al. Started that Big Data is comprised of large data sets that can't be handle by traditional systems. Big data includes structured data, semi-structured and unstructured data. The data storage technique used for big data includes multiple clustered network attached storage (NAS) and object based storage. The Hadoop architecture is used to process unstructured and semi-structured using map reduce to locate all relevant data then select only the data directly answering the query. The advent of Big Data has posed opportunities as well challenges to business [7].

Sameer Agarwal et.al. Presents a BlinkDB, a approximate query engine for running interactive SQL queries on large volume of data which is massively parallel. BlinkDB uses two key ideas: (1) an adaptive optimization framework that builds and maintains a set of multi-dimensional stratified samples from original data over time, and (2) A dynamic sample selection strategy that selects an appropriately sized sample based on a query's accuracy or response time requirements [8].

Yingyi Bu et.al. Used a new technique called as HaLoop which is modified version of Hadoop MapReduce Framework, as Map Reduce lacks built-in-support for iterative programs HaLoop allows iterative applications to be assembled from existing Hadoop programs without modification, and significantly improves their efficiency by providing inter iteration caching mechanisms and a loop-aware scheduler to exploit these caches. He presents the design, implementation, and evaluation of HaLoop, a novel parallel and distributed system that supports large-scale iterative data analysis applications. HaLoop is built on top of Hadoop and extends it with a new programming model and several important optimizations that include (1) a loop-aware task scheduler, (2) loop-invariant data caching, and (3) caching for efficient fix point verification [9].

Shadi Ibrahim et.al. Project says presence of partitioning skew1 causes a huge amount of data transfer during the shuffle phase and leads to significant unfairness on the reduce input among different data nodes In this paper, author develop a novel algorithm named LEEN for locality aware and fairness-aware key partitioning in MapReduce. LEEN embraces an asynchronous map and reduce scheme. Author has integrated LEEN into Hadoop. His experiments demonstrate that LEEN can efficiently achieve higher locality and reduce the amount of shuffled data. More importantly, LEEN guarantees fair distribution of the reduce inputs. As a result, LEEN achieves a performance improvement of up to 45% on different workloads. To tackle all this he presents a present a technique for Handling Partitioning Skew in MapReduce using LEEN [10].

III. CONCLUSION

We have entered an era of Big Data. The paper describes the concept of Big Data along with 3 Vs, Volume, Velocity and variety of Big Data. The paper also focuses on Big Data processing problems. These technical challenges must be addressed for efficient and fast processing of Big Data. The challenges include not just the obvious issues of scale, but also heterogeneity, lack of structure, error-handling, privacy, timeliness, provenance, and visualization, at all stages of the analysis pipeline from data acquisition to result interpretation. These technical challenges are common across a large variety of application domains, and therefore not cost-effective to address in the context of one domain alone. The paper describes Hadoop which is an open source software used for processing of Big Data.

REFERENCES

[1] Kawale S. M., Dr. Holambe A. N., Bokefode J. D. "A Review on Big Data Concepts and various Analytic Techniques". International Journal of Computer Trends and Technology (IJCTT) V52(1):13-16, October 2017. ISSN:2231-2803.

[2] S.Vikram Phaneendra & E.Madhusudhan Reddy "Big Data- solutions for RDBMS problems- A survey" In 12th IEEE/IFIP Network Operations & Management Symposium (NOMS 2010) (Osaka, Japan, Apr 19{23 2013).

[3] Kiran kumara Reddi & Dnvsl Indira "Different Technique to Transfer Big Data : survey" IEEE Transactions on 52(8) (Aug.2013) 2348 { 2355}.

[4] Jimmy Lin "MapReduce Is Good Enough?" The control project. IEEE Computer 32 (2013).

[5] Umasri.M.L, Shyamalagowri.D, Suresh Kumar.S "Mining Big Data:- Current status and forecast to the future" Volume 4, Issue 1, January 2014 ISSN: 2277 128X.

[6] Albert Bifet "Mining Big Data In Real Time" Informatica 37 (2013) 15–20 DEC 2012.

[7] Bernice Purcell "The emergence of "big data" technology and analytics" Journal of Technology Research 2013.

[8] Sameer Agarwal, Barzan MozafariX, Aurojit Panda, Henry Milner, Samuel MaddenX, Ion Stoica "BlinkDB: Queries with Bounded Errors and Bounded Response Times on Very Large Data" Copyright © 2013ì ACM 978-1-4503-1994.

[9] Yingyi Bu _ Bill Howe _ Magdalena Balazinska _ Michael D. Ernst "The HaLoop Approach to Large-Scale Iterative Data Analysis" VLDB 2010 paper "HaLoop: Efficient Iterative Data Processing on Large Clusters.

[10] Shadi Ibrahim _ Hai Jin _ Lu Lu "Handling Partitioning Skew in MapReduce using LEEN" ACM 51 (2008) 107–113

[11] Kenn Slagter · Ching-Hsien Hsu "An improved partitioning mechanism for optimizing massive data analysis using MapReduce" Published online: 11 April 2013.