Survey on Big Data Analytics and Hadoop Tools

Chennupati Vinay Kumar
UG Student
Department of Computer Science & Engineering
GITAM University Hyderabad India

Abstract
Big data is a collection of large data sets that include different types such as structured, unstructured and semi-structured data. This data can be generated from different sources like social media, audios, images, log files, sensor data, transactional applications, web etc. The question that arises now is, how to develop a high performance platform to efficiently analyze big data and how to design an appropriate mining algorithm to find the useful things from big data. To process or analyse this huge amount of data or extracting meaningful information is a challenging task now a days. Big data exceeds the processing capability of traditional database to capture, manage, and process the voluminous amount of data.

Keywords: Big Data, Hadoop, Map Reduce

I. INTRODUCTION

Imagine a world without data storage a place where every detail about a person or organization, every transaction performed, or every aspect which can be documented is lost directly after use. Organizations would thus lose the ability to extract valuable information and knowledge, perform detailed analyses, as well as provide new opportunities and advantages. Anything ranging from customer names and addresses, to products available, to purchases made, to employees hired, etc. has become essential for day-to-day continuity. Data is the building block upon which any organization thrives. Now think of the extent of details and the surge of data and information provided nowadays through the advancements in technologies and the internet. With the increase in storage capabilities and methods of data collection, huge amounts of data have become easily available. Every second, more and more data is being created and needs to be stored and analyzed in order to extract value. Furthermore, data has become cheaper to store, so organizations need to get as much value as possible from the huge amounts of stored data. The size, variety, and rapid change of such data require a new type of big data analytics, as well as different storage and analysis methods. Such sheer amounts of big data need to be properly analyzed, and pertaining information should be extracted.

Big data shall mean the datasets that could not be perceived, acquired, managed, and processed by traditional IT and software/hardware tools within a tolerable time. Big Data describes any massive volume of structured, semi structured and unstructured data that are difficult to process using traditional database system such as RDBMS. An example of big data may be Exabyte’s (1024 terabytes) of data consisting of trillions of records of millions of people from different sources such as websites, social media, mobile data, web servers, online transactions and so on. In the past, type of information available was limited. There was a well-defined set of technology approaches for managing information. But in today's world, the amount of data has been exploding. It has grown to terabytes and petabytes. Because in every minute, there are 280,000 tweets, more than 100 million emails are sent. 2 million searching queries in Google, and more than 350 GB of data is processed in Face book in every minute. Some of the applications of big data are in areas such as social media, healthcare, traffic management, banking, retail, education and so on.

II. CHALLENGES AND OPPORTUNITIES WITH BIG DATA

A. Volume

The volume of data, especially machine-generated data, is exploding, how fast that data is growing every year, with new sources of data that are emerging. For example, in the year 2000, 800,000 petabytes (PB) of data were stored in the world, and it is expected to reach 35 zettabytes (ZB) by 2020 (according to IBM). 108 Big Data Computing Social media plays a key role: Twitter generates 7+ terabytes (TB) of data every day. Facebook, 10 TB. Mobile devices play a key role as well, as there were estimated 6 billion mobile phones in 2011. The challenge is how to deal with the size of Big Data

B. Variety, Combining Multiple Data Sets

More than 80% of today’s information is unstructured and it is typically too big to manage effectively. It used to be the case that all the data an organization needed to run its operations effectively was structured data that was generated within the organization. Things like customer transaction data, ERP data, etc. Today, companies are looking to leverage a lot more data from a wider variety of sources both inside and outside the organization. Things like documents, contracts, machine data, sensor data, social media, health records, emails, etc. The list is endless really.
A lot of this data is unstructured, or has a complex structure that’s hard to represent in rows and columns. And organizations want to be able to combine all this data and analyze it together in new ways. For example, we have more than one customer in different industries whose applications combine geospatial vessel location data with weather and news data to make real-time mission-critical decisions. Data come from sensors, smart devices, and social collaboration technologies. Data are not only structured, but raw, semi structured, unstructured data from web pages, web log files (click stream data), search indexes, e-mails, documents, sensor data, etc. Semi structured Web data such as A/B testing, sessionization, bot detection, and pathing analysis all require powerful analytics on many peta bytes of semi structured Web data. The challenge is how to handle multiplicity of types, sources, and formats.

C. Velocity

Velocity in Big data is a concept which deals with the speed of the data coming from various sources. This characteristic is not being limited to the speed of incoming data but also speed at which the data flows. For example, the data from the sensor devices would be constantly moving to the database store and this amount won’t be small enough. Thus our traditional systems are not capable enough on performing the analytics on the data which is constantly in motion.

D. Variability, Data Quality, Data Availability

Variability considers the inconsistencies of the data flow. Data loads become challenging to be maintained especially with the increase in usage of the social media which generally causes peak in data loads with certain events occurring.

There are several challenges:
- How good is the data?
- How broad is the coverage?
- How fine is the sampling resolution?
- How timely are the readings?
- Is there data available, at all?

E. Data Discovery

This is a huge challenge: how to find high-quality data from the vast collections of data that are out there on the web.

F. Scalability

Techniques like social graph analysis, for instance leveraging the influencers in a social network to create better user experience are hard problems to solve at scale. All of these problems combined create a perfect storm of challenges and opportunities to create faster, cheaper and better solutions for Big Data analytics than traditional approaches can solve.

G. Complexity

It is quite an undertaking to link, match, cleanse and transform data across systems coming from various sources. It is also necessary to connect and correlate relationships, hierarchies and multiple data linkages or data can quickly spiral out of control.

III. MANUFACTURING – BENEFITS AND CHALLENGES

Manufacturing and production managers believe the greatest opportunities of Big Data for their function are to detect product defects and boost quality, and to improve supply planning. Better detection of defects in the production processes is next on the list. A $2 billion industrial manufacturer said that analyzing sales trends to keep its manufacturing efficient was the main focus of its Big Data investments. Understanding the behaviour of repeat customers is critical to delivering in a timely and profitable manner. Most of its profitability analysis is to make sure that the company has good contracts in place. The company says its adoption of analytics has facilitated its shift to lean manufacturing, and has helped it determine which products and processes should be scrapped. They see far less opportunity in using Big Data for mass customization, simulating new manufacturing processes, and increasing energy efficiency.

For manufacturers, opportunities enabled by big data can drive productivity gains both through improving efficiency and the quality of products. Efficiency gains arise across the value chain, from reducing unnecessary iterations in product development cycles to optimizing the assembly process. The real output value of products is increased by improving their quality and making products that better match customers’ needs. Data have become an important factor of production today on a par with physical assets and human capital and the increasing intensity with which enterprises.

A. Hadoop

Hadoop runs applications using the Map Reduce algorithm, where the data is processed in parallel with others. Hadoop is used to develop applications that could perform complete statistical analysis on huge amounts of data. Hadoop is an Apache open source framework written in java that allows distributed processing of large datasets across clusters of computers using simple
programming models. The Hadoop framework application works in an environment that provides distributed storage and computation across clusters of computers. Hadoop is designed to scale up from single server to thousands of machines, each offering local computation and storage. It is quite expensive to build bigger servers with heavy configurations that handle large scale processing, but as an alternative, you can tie together many commodity computers with single-CPU, as a single functional distributed system and practically, the clustered machines can read the dataset in parallel and provide a much higher throughput. Moreover, it is cheaper than one high-end server. So this is the first motivational factor behind using Hadoop that it runs across clustered and low-cost machines. Hadoop runs code across a cluster of computers.

This process includes the following core tasks that Hadoop perform. Data is initially divided into directories and files. Files are divided into uniform sized blocks of 128M and 64M (preferably 128M). These files are then distributed across various cluster nodes for further processing. HDFS, being on top of the local file system, supervises the processing [8]. Blocks are replicated for handling hardware failure. Checking that the code was executed successfully, performing the sort that takes place between the map and reduce stages, sending the sorted data to a certain computer, writing the debugging logs for each job.

Hadoop has two major layers namely:
- Processing/Computation layer (Map Reduce), and
- Storage layer (Hadoop Distributed File System).

B. Map Reduce

Map Reduce is a parallel programming model for writing distributed applications devised at Google for efficient processing of large amounts of data (multitier byte data-sets), on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner. The Map Reduce program runs on Hadoop which is an Apache open-source framework. It is a processing technique and a program model for distributed computing based on java. The Map Reduce algorithm contains two important tasks, namely Map and Reduce. Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples. As the sequence of the name Map Reduce implies, the reduce task is always performed after the map job. The major advantage of Map Reduce is that it is easy to scale data processing over multiple computing nodes. Under the Map Reduce model, the data processing primitives are called mappers and reducers. Decomposing a data processing application into mappers and reducers is sometimes nontrivial. But, once we write an application in the Map Reduce form, scaling the application to run over hundreds, thousands, or even tens of thousands of machines in a cluster is merely a configuration change. This simple scalability is what has attracted many programmers to use the Map Reduce model.

![Map Reduce diagram](image)

**Fig. 1: Map Reduce**

1) The stages of Map Reduce

Program generally Map Reduce paradigm is based on sending the computer to where the data resides! Map Reduce program executes in two stages, namely map stage and reduce stage.

- Map stage: The map or mappers job is to process the input data. Generally the input data is in the form of file or directory and is stored in the Hadoop file system (HDFS). The input file is passed to the mapper function line by line. The mapper processes the data and creates several small chunks of data.
Reduce stage: This stage is the combination of the Shuffle stage and the Reduce stage. The Reducers job is to process the data that comes from the mapper. After processing, it produces a new set of output, which will be stored in the HDFS.

During a Map Reduce job, Hadoop sends the Map and Reduce tasks to the appropriate servers in the cluster. The framework manages all the details of data-passing such as issuing tasks, verifying task completion, and copying data around the cluster between the nodes. Most of the computing takes place on nodes with data on local disks that reduces the network traffic. After completion of the given tasks, the cluster collects and reduces the data to form an appropriate result, and sends it back to the Hadoop server.

IV. HADOOP COMPONENT

There is an extensive list of products and projects that either extend Hadoop’s functionality or expose some existing capability in new ways, like MapReduce for distributed data processing, HDFS for distributed file system, Hive for distributed data warehouse and provides SQL based query language, HBase for distributed column based database, Pig provides an abstraction layer with the help of scripts in the language Pig Latin, which are translated to MapReduce jobs. Other examples of projects built on top of Hadoop include Apache Sqoop Apache Oozie, and Apache Flume, YARN for improvement of cluster utilization and scalability etc.

A. HDFS

The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware its architecture is shown in fig 4. It has many similarities with existing distributed file systems. However, the differences from other distributed file systems are significant. HDFS is highly fault-tolerant and is designed to be deployed on low-cost hardware. HDFS provides high throughput access to application data and is suitable for applications that have large data sets. HDFS was originally built as infrastructure for the Apache Nutch web search engine project. HDFS is part of the Apache Hadoop Core project.

B. Map Reduce

MapReduce is a Java-based system created by Google where the actual data from the HDFS store gets processed efficiently. MapReduce breaks down a big data processing job into smaller tasks. MapReduce is responsible for the analysing large datasets in parallel before reducing it to find the results. In the Hadoop ecosystem, Hadoop MapReduce is a framework based on YARN architecture. YARN based Hadoop architecture, supports parallel processing of huge data sets and MapReduce provides the framework for easily writing applications on thousands of nodes, considering fault and failure management. The basic principle of operation behind MapReduce is that the “Map” job sends a query for processing to various nodes in a Hadoop cluster and the “Reduce” job collects all the results to output into a single value. Map Task in the Hadoop ecosystem takes input data and splits into independent chunks and output of this task will be the input for Reduce Task. In The same Hadoop ecosystem Reduce task combines Mapped data tuples into smaller set of tuples. Meanwhile, both input and output of tasks are stored in a file system. MapReduce takes care of scheduling jobs, monitoring jobs and re-executes the failed task. MapReduce framework forms the compute node while the HDFS file system forms the data node. Typically in the Hadoop ecosystem architecture both data node and compute node are considered to be the same.
C. Yarn

Yarn forms an integral part of Hadoop 2.0. YARN is great enabler for dynamic resource utilization on Hadoop framework as users can run various Hadoop applications without having to bother about increasing workloads.

D. Pig

Pig is a high level scripting language that is used with Apache Hadoop. Pig enables data workers to write complex data transformations without knowing Java. Pig’s simple SQL-like scripting language is called Pig Latin, and appeals to developers already familiar with scripting languages and SQL.

Pig is complete, so you can do all required data manipulations in Apache Hadoop with Pig. Through the User Defined Functions (UDF) facility in Pig, Pig can invoke code in many languages like JRuby, Jython and Java. You can also embed Pig scripts in other languages. The result is that you can use Pig as a component to build larger and more complex applications that tackle real business problems.

E. Hive

Hive developed by Facebook is a data warehouse built on top of Hadoop and provides a simple language known as HiveQL similar to SQL for querying, data summarization and analysis. Hive makes querying faster through indexing.

F. Sqoop

Sqoop component is used for importing data from external sources into related Hadoop components like HDFS, HBase or Hive. It can also be used for exporting data from Hadoop to other external structured data stores. Sqoop parallelized data transfer, mitigates excessive loads, allows data imports, efficient data analysis and copies data quickly.

G. Flume

Flume Component is used to gather and aggregate large amounts of data. Apache Flume is used for collecting data from its origin and sending it back to the resting location (HDFS). Flume accomplishes this by outlining data flows that consist of 3 primary structures channels, sources and sinks. The processes that run the dataflow with flume are known as agents and the bits of data that flow via flume are known as events.

H. HBase

HBase is a column-oriented database that uses HDFS for underlying storage of data. HBase supports random reads and also batch computations using MapReduce. With HBase NoSQL database enterprise can create large tables with millions of rows and columns on hardware machine. The best practice to use HBase is when there is a requirement for random ‘read or write’ access to big datasets.
V. CONCLUSION AND FUTURE WORK

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.

Today, data is generated from various different sources and can arrive in the system at various rates. To process these large amounts of data is a big issue today. In this paper we discussed Hadoop tool for Big data in detail. Hadoop is the core platform for structuring Big Data, and solves the problem of making it useful for analytics purposes. We also discussed some hadoop components which are used to support the processing of large data sets in distributed computing environments. In future we can use some clustering techniques and check the performance by implementing it in hadoop.

REFERENCES