IJARIIT

INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X

Impact factor: 4.295

(Volume3, Issue1)

Available online at: www.ijariit.com

Big Data: Tools and Applications

Kanchana .R
Student, Department of CSE
Cambridge Institute Of Technology.

Shashikumar D R
Department of CSE
Cambridge Institute of Technology.

Abstract: The amount of data in our industry and the world is exploding. Data is being collected and stored at unprecedented rates. The challenge is not only to store and manage the vast volume of data, but also to analyze and extract meaningful value from it. There are several approaches to collecting, storing, processing, and analyzing big data. Our analysis illustrates that the Big Data analytic is a fast-growing, influential practice and a key enabler for the social business. This paper covers the leading tools and technologies for big data storage and processing. Hadoop, Map Reduce and No SQL are the major big data technologies. These technologies are very helpful in big data management.

Keywords: Big data; Hadoop; Hadoop Distributed File System.

I. INTRODUCTION

Big data is a blanket term for the non-traditional strategies and technologies needed to gather, organize, process, and gather insights from large data sets. While the problem of working with data that exceeds the computing power or storage of a single computer is not new, the pervasiveness, scale, and value of this type of computing has greatly expanded in recent years.

In this context, "large dataset" means a dataset too large to reasonably process or store with traditional tooling or on a single computer. This means that the common scale of big datasets is constantly shifting and may vary significantly from organization to organization.

In [1] authors have categorized the big data analytics in two categories Stream processing and batch processing. In stream processing, data comes in streams and it is process as soon as possible to generate results. Storm and Apache Kafka are popular stream processing models. In batch processing first data is stored and then processed.

Volume

The sheer scale of the information processed helps define big data systems. These datasets can be orders of magnitude larger than traditional datasets, which demands more thought at each stage of the processing and storage life cycle.

Velocity

Another way in which big data differs significantly from other data systems is the speed that information moves through the system. Data is frequently flowing into the system from multiple sources and is often expected to be processed in real time to gain insights and update the current understanding of the system.

Variety

Big data problems are often unique because of the wide range of both the sources being processed and their relative quality.

Data can be ingested from internal systems like application and server logs, from social media feeds and other external APIs, from physical device sensors, and from other providers. Big data seeks to handle potentially useful data regardless of where it's coming from by consolidating all information into a single system.

Other Characteristics

Various individuals and organizations have suggested expanding the original three Vs, though these proposals have tended to describe challenges rather than qualities of big data. Some common additions are:

- Veracity: The variety of sources and the complexity of the processing can lead to challenges in evaluating the quality of the data (and consequently, the quality of the resulting analysis).

- Variability: Variation in the data leads to wide variation in quality. Additional resources may be needed to identify, process, or filter low quality data to make it more useful.

- Value: The ultimate challenge of big data is delivering value. Sometimes, the systems and processes in place are complex enough that using the data and extracting actual value can become difficult.

II. HADOOP RELATED TOOLS

Hadoop Apache's

Hadoop project has become nearly synonymous with Big Data. It has grown to become an entire ecosystem of open source tools for highly scalable distributed computing. Operating System: Windows, Linux, OS X.

Hadoop Architecture

Apache Hadoop [2 is an open source framework used to store and analyses big data which is present in Hadoop cluster. Hadoop always runs on a cluster means on homogeneous environment. Moreover homogeneous environment means all the systems which are present in cluster, their all components must be same in terms of RAM, CPU etc. Primarily Hadoop has two major components.

HDFS (Hadoop distributed File system) Map Reduce

Hadoop Distributed File System

HDFS is a file system based on the master slave architecture. When HDFS takes in data, it breaks the information down into separate pieces and distributes them to different nodes in a cluster. The file system also copies each piece of data multiple times and distributes the copies to individual nodes, placing at least one copy on a different server others. As a result, the data on nodes that crash can be found elsewhere within a cluster, which allows processing to continue while the failure is resolved.

There is also secondary name node present on Hadoop cluster .Name Node is the master node which controls all the data nodes and it contains Meta data. It manages all the file operations like read, write etc. Data nodes are the slave nodes present in Hadoop cluster. All the file operations performed on these nodes and data is actually stored on these nodes as decided by name nodes.

Map Reduce

Map Reduce 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.

The Algorithm

- Generally Map Reduce paradigm is based on sending the computer to where the data resides.
- Map Reduce program executes in three stages, namely map stage, shuffle stage, and reduce stage.
- Map stage: The map or mapper’s 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 Reducer’s 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.

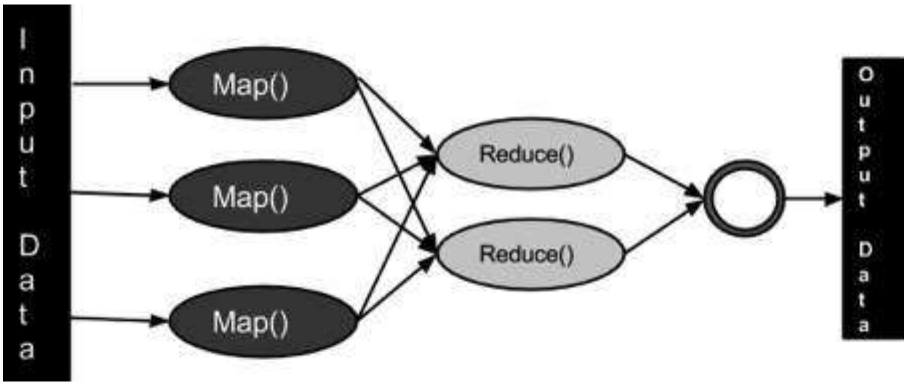

Fig 1: Map Reduce Architecture

Input: Input means that data which gets for processing and it is divided into further smaller chunks which are further allocated to the mappers.
Mapper: Mappers are the individuals that are assigned with the smallest unit of work for some processing.
Reducer: Mappers output become input for the reducers to aggregate the data in form of final output.
Output: Reducers jobs are finally collected in the form of aggregated output.

TECHNOLOGIES BASED ON HADOOP

There are many technologies which are built on the top of the Hadoop [2] by Apache which means that Hadoop is not a single project as it includes other projects also. These technologies or projects have been designed for increasing the efficiency and functionality of Hadoop. These all technologies specially designed for dealing with big data and these all are along with HDFS and Map reduce .Primarily Hadoop Eco system consists of following technologies:

Apache PIG
Apache HBase
Apache Hive
Apache Sqoop
Apache Flume
Apache Zookeeper

Table 1. Hadoop Technologies

Hadoop Technologies	Description
Apache PIG	Apache Pig is a platform for distributed big data analysis. It relies on a programming language called Pig Latin, which boasts simplified parallel programming, optimization and extensibility. Operating System: OS Independent.
Apache HBase	Designed for very large tables with billions of rows and millions of columns, HBase is a distributed database that provides random real-time read/write access to big data. It is somewhat similar to Google's Bigtable, but built on top of Hadoop and HDFS. Operating System: OS Independent.
Apache Hive	Apache Hive is the data warehouse for the Hadoop ecosystem. It allows users to query and manage big data using HiveQL, a language that is similar to SQL. Operating System: OS Independent.
Apache Sqoop	Enterprises frequently need to transfer data between their relational databases and Hadoop, and Sqoop is one tool that gets the job done. It can import data to Hive or HBase and export from Hadoop to RDBMSes. Operating System: OS Independent.
Apache Flume	Flume collects log data from other applications and delivers them into Hadoop. The website boasts, "It is robust and fault tolerant with tunable reliability mechanisms and many failover and recovery mechanisms." Operating System: Linux, OS X.
Apache Zookeeper	This administrative big data tool describes itself as "a centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services." It allows nodes within a Hadoop cluster to coordinate with each other. Operating System: Linux, Windows (development only), OS X (development only).

III. APPLICATIONS AREAS OF BIG DATA

Understanding and Targeting Customers:

This is one of the most widely used areas of Big data today. Hereig data analytics is used to understand customers and their behaviors and preferences. Companies are keen to expand their traditional datasets with social media data, browser logs as well as text analytics and sensor data to get a more complete picture of their customers. Recommendation system plays important role here. By analyzing the historical behavior of customers; it sends the personalized recommendation which is win-win condition for both customer and company.

Understanding and Optimizing Business Processes:

Big data is also widely used to optimize business processes. Retailers are optimizing their stock based on predictions generated from social media data, web search trends and weather forecasts.

Improving Healthcare and Public Health:

In health care also, we have a large amount of data coming in from various pathological reports, ultrasound and MRIs etc. Nowadays healthcare is using big data technology to predict, understand and avoid various new diseases and improving the quality of life. In the coming future all the individual data from smart watches and wearable devices can be shared with the doctors to analyze our health.

Improving and Optimizing Cities/Countries:

Many aspects of our cities/countries can be improved by Big data analytics. For example, traffic flows can be optimized based on real time traffic information as well as social media and weather data.

In agriculture:

In the coming decades agriculture will be transformed by the use of big data analytics. Sensors can be deployed on the forms and collected data is used to detect the reactions of crop on different environmental condition, soil conditions, water level etc.

Financial Trading:

Big data is used widely today in high -frequency trading. Here big data analytics is used to make trading decisions.

CONCLUSION

Handling big data efficiently is the need of the hour and one need to come up with plausible solutions to these challenges one needs to understand the concept of big data, its handling methodologies and furthermore improve the approaches in analyzing big data. With the advent of social media the need for handling big data has increased monumentally. Approximately 5 Exabytes of data has been created, from the beginning of time till 2003. The same amount is now generated every 2 days. As more and more organizations are stepping out of the traditional boundaries big data keeps growing bigger. The tools being developed are efforts for overcoming the challenges arising due to big data.

REFERENCES

1. Hu, H., Wen, Y., Chua, T. S., & Li, X. (2014). Toward scalable systems for big data analytics: a technology tutorial.Access, IEEE, 2, 652-687.
2. Apache Hadoop Project, http://hadoop.apache.org/
3. Katal, Avita, Mohammad Wazid, and R. H. Goudar. "Big data: issues, challenges, tools and good practices." Contemporary Computing (IC3), 2013 Sixth International Conference on. IEEE, 2013.
4. Aditya B Patel, Manashvi Birla, Ushma Nair, “Addressing Big Data Problem Using Hadoop and MapReduce” Nirma University International Conference on Engineering, 2012
5. Ted Garcia, Taehyung Wang “Analysis of Big Data Technologies and Methods” Seventh International Conference on Semantic Computing, IEEE 2013.
6. Jinshuang Yan, Xiaoliang Yang, Rong Gu, Chunfeng Yuan, Yihua Huang “ Performance Optimization for short MapReduce Job Execution in Hadoop” Second International Conference on Cloud and Green Computing , 2012.