ISSN: 0975-766X
CODEN: IJPTFI
Available Online through
www.ijptonline.com
Research Article

SURVEY PAPER ON BIG DATA AND SCOPE OF EXISTING FRAMEWORKS

[1]Vijaya Sreenivas Kancharla , [2]Dr.T.Nalini
Research Scholar, Department of Computer Science and Engineering, Vels University, Chennai, India.
Professor, Department of Computer Science and Engineering, Bharath University, Chennai, India.

Received on 18-08-2016 *Accepted on 25-09-2016*

Abstract

Nowadays Big Data plays a vital role in managing and handling a large volume of data that both structured and unstructured. It is a collection of data sets which are not handled by the traditional database management application and tools. An objective of the Big Data is to capture, to search, to share, to transfer, to analyze, and to visualize the unstructured data.

It mainly focuses on large volume of data from various resources with different velocity. Big data is not merely a data, rather it has become a complete subject, which involves various tools, techniques and frameworks. There are plenty of programming models exits for handling the huge amount of data in an efficient manner. The technologies used for big data to handle the massive data are Hadoop, Spark, Hive, HBase and Flink.

Keywords: Hadoop, Spark, Hive, HBase and Flink,

1. Introduction

Big Data is often described as extremely largedata sets that have grown beyond the ability to manage and analyze them with traditional data processing tools. Big Data defines a situation in which data sets have grown to such enormous sizes that conventional information technologies can no longer effectively handle either the size of the data set or the scale and growth of the data set, including mobile services, retail, manufacturing, financial services, life sciences, and physical sciences.

The data in it will be of three types:

Structured data: Relational data.

Semi Structured data: XML data.

Unstructured data: PDF, Word, Text, Media Logs.

1.1 The characteristics of big data:

1. Volume:

The quantity of generated and stored data, It is the size of the data which determines the value and potential of the data under consideration and whether it can actually be considered Big Data or not. Volume represent the size of the data how the data is large. The size of the data is represented in terabytes and peta bytes.

2. Variety:

The type and nature of the data, Variety makes the data too big. The files comes in various formats and of any type, it may be structured or unstructured such as text, audio, videos, log files and more.

3. Velocity:

The term 'velocity' in the context refers to the speed at which the data is generated and processed to meet the demands and challenges, which lie ahead in the path of growth and development. Velocity refers to the speed of data processing where data comes at high speed (sometimes 1 minute is too late so big data is time sensitive).

4. Value:

This is a factor which cane a problem for those who analyse the data. Inconsistency of the data set can hamper processes to handle and manage data effectively. The potential value of Big data is huge. Value is mainsource for big data because it is important for businesses, IT infrastructure system to store large amount of values in database.

5. Veracity:

The quality of the data being captured can vary greatly and the accuracy of analysis depends on the veracity of the source data. Veracity refers to noise, biases and abnormality when we dealing with high volume, velocity and variety of data, the all of data are not going 100%correct, there will be dirty data.

2. Hadoop

Hadoop is an open-source framework that allows to store and process big data in a distributed environment across clusters of computers using simple programming models. Hadoop is an Apache open source framework written in Java that allows distributed processing of large dataset across cluster of computers using simple programming model Hadoop creates cluster of machines and coordinates work among them. It is designed to scale up from single servers to thousands of machines and consists of two components Hadoop Distributed File System(HDFS)and MapReduce Framework each offering local computation and storage Hadoop.

a) HDFS (Hadoop Distributed File System)

HDFS is a file system designed for storing very large files with streaming data access pattern, running clusters on commodity hardware. HDFS holds very large amount of data and provides easier access. To store such huge data, the files are stored across multiple machines. These files are stored in redundant fashion to rescue the system from possible data losses in case of failure. HDFS follows the master-slave architecture and it has the following elements,

The **namenode** is the commodity hardware that contains the GNU/Linux operating system and the namenode software. It is a software that can be run on commodity hardware. It is centrally placed node, which contains information about Hadoop file system. The main task of name node is that it records all the metadata & attributes and specific locations offiles & data blocks in the data nodes. Name nodeacts as the master node as it stores all the information about the system .and provides information which is newly added, modified and removed from data nodes.

The **datanode** is a commodity hardware having the GNU/Linux operating system and datanode software, It works as slave node. A datanode performs two main tasks storing a block in HDFS and acts as the platform for running jobs.

3. Spark

Spark is the heir apparent to the Big Data processing kingdom. Spark and Hadoop are often contrasted as an "either/or" choice, but that isn't really the case. They do not perform exactly the same tasks, and they are not mutually exclusive, as they are able to work together. Spark differs from Hadoop and the MapReduce paradigm in that it works in-memory, speeding up processing times and reported to work up to 100 times faster than Hadoop in certain circumstances but it does not provide its own distributed storage system. As mentioned, Spark does not include its own system for organizing files in a distributed way (the file system) so it requires one provided by a third-party. For this reason many Big Data projects involve installing Spark on top of Hadoop, where Spark’s advanced analytics applications can make use of data stored using the Hadoop Distributed File System (HDFS).

The Hadoop ecosystem can accommodate the Spark processing engine in place of MapReduce, leading to all sorts of different environment make-ups that may include a mix of tools and technologies from both ecosystems.

4. Hive

Hive is a data warehouse infrastructure tool to process structured data in Hadoop. It resides on top of Hadoop to summarize Big Data, and makes querying and analyzing easy. Hive is an open-source data warehouse system implemented by Apachefor querying and analyzing large datasets stored in Hadoop files.

Hive has three main functions as follows data summarization, query and analysis and supports queries expressed in a language called HiveQL, which automatically translates SQL-like queries into MapReduce jobs executed on Hadoop. In addition, HiveQL supports custom MapReduce scripts to be plugged into queries. Hive also enables data serialization/deserialization and increases flexibility in schema design by including a system catalog called Hive-Metastore.

Hive is dependent on Hadoop and MapReduce executions, also not designed for OLTP workloads and queries may have lag time in processing up to several minutes. This implies Hive may not be suitable for big data analytics applications that need rapid response times, typical of relational databases. It is best used for batch jobs over large sets of append-only data (like web logs).

5. HBase

HBase is a data model that is similar to Google’s big table designed to provide quick random access to huge amounts of structured data. It was created for hosting very large tables, making it a great choice to store multi-structured or sparse data. Users can query HBase for a particular point in time, making “flashback” queries possible. These characteristics make HBase a great choice for storing semi-structured data like log data and then providing that data very quickly to users or applications integrated with HBase.

6. Flink

Apache Flink is a streaming dataflow engine, aiming to provide facilities for distributed computation over streams of data. Flink is optimized for cyclic or iterative processes by using iterative transformations on collections. This is achieved by an optimization of join algorithms, operator chaining and reusing of partitioning and sorting. Flink streaming processes data streams as true streams and data elements are immediately "pipelined" though a streaming program as soon as they arrive. This allows to perform flexible window operations on streams. It is even capable of handling late data in streams by the use of watermarks.

7. Conclusions

The paper describes the concept of Big Data along with Operational vs. Analytical Systems of Big Data. The paper also focuses on Big Data processing problems. Hadoop has an ability to solve many Big Data problems but it is not a best tool for handling overall data inall perspectives. Hive not suitable for real time queries and also takes several time to provide the query response. Spark helps in speeding up processing times and reported to work multiple times faster than Hadoop but does not provide its own distributed storage system. HBase does not implement any cross data operations and joining operations without help of MapReduce again time taking process. Flink is faster then Spark, due to its underlying architecture but does not have strong community over it.

These technical challenges must beaddressed for efficient and fast processing of Big Data. Inthis paper I have tried to cover popular frameworks to process big data and limitations over them.

References

1. Apache Hadoop: http://Hadoop.apache.org
2. http://www.kdnuggets.com/2016/03/top-big-data-processing-frameworks.html
3. Dean, J. and Ghemawat, S.,“MapReduce: a flexible data processing tool”, ACM 2010.
4. http://data-flair.training/blogs/comparison-apache-flink-vs-apache-spark/
5. Hadoop Distributed FileSystem, http://hadoop.apache.org/hdfs
6. HadoopTutorial:http://developer.yahoo.com/hadoop/tutorial/module1.html
7. Jean-Pierre Dijcks, “Oracle: Big Data for the Enterprise”, 2013.
8. Greenplum Analytics Workbench ,visitus at www.greenplum.com
9. shilpa Manjit Kaur,” BIG Data and Methodology-A review” ,International Journal of Advanced Researchin Computer Science and Software Engineering,Volume 3, Issue 10, October 2013.
10. Zikopoulos, P.C., Eaton, C., deRoos, D., Deutsch, T., and G. Lapis.Understanding Big Data— Analytics for Enterprise Class Hadoop and Streaming Data. New York: McGraw-Hill, 2013.