Review and Challenges of Big Data Analytics with Hadoop Distributed File System

Jebeula. T [#1],Jebamalar Tamilselvi. J[#2]

[#1]*Department of Computer Applications, #2Department of Computer Applications*

Bharathiar University, Jaya Engineering College

Coimbatore, Chennai.

[1]jebeula@gmail.com

[2]jjebamalar@gmail.com

***Abstract* -Big data is a unique term that describes a collection of data sets which are large and complex, growing data sets with various, independent sources; it contains both unstructured and structured data. Big data is large and complex data sets for traditional data analysis. The data is increased exponentially because of social media, email and document and sensor data. The growth of this data affects the business and other science of world. Big data analytics is the process of extracting large amounts of structure, semi structure and unstructured data for decision making and strategic business planning. The Hadoop Distributed File System (HDFS) has parallel processing for handling big data. This paper presents a study of various challenges, algorithms and tools of big data analytics with HDFS.**

***Keywords*--Big Data, Hadoop, Hadoop Distributed File System, Big Data Analytics tools.**

I. INTRODUCTION

Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision-making. Big data is often boiled down to a few varieties including social data, machine data, and transactional data. Social media data is providing remarkable insights to companies on consumer behavior and sentiment that can be integrated with CRM data for analysis, with 230 million tweets posted on Twitter per day, 2.7 billion Likes and comments added to Facebook every day, and 60 hours of video uploaded to YouTube every minute (this is what we mean by velocity of data).

To gain value from this data, choose an alternative way to process it. Big Data is the next generation of data warehousing and business analytics and is poised to deliver top line revenues cost efficiently for enterprises. Big data is a popular term used to describe the exponential growth and availability of data, both structured and unstructured. Big data analytics can reveal insights hidden previously by data too costly to process, such as peer influence among customers, revealed by analyzing shoppers' transactions, social and geographical data.

Google has introduced Map Reduce framework for processing large amounts of data on commodity hardware. Apache's Hadoop distributed file system (HDFS) is evolving as a superior software component for cloud computing combined along with integrated parts such as Map Reduce. Hadoop, which is an open-source implementation of Google Map Reduce, including a distributed file system, provides to the application programmer the abstraction of the map and the reduce. With Hadoop it is easier for organizations to get a grip on the large volumes of data being generated each day, but at the same time can also create problems related to security, data access, monitoring, high availability and business continuity.

Apache Hadoop is a software framework which is open-source and used to store, manage and process datasets using MapReduce programming model. Apache Hadoop has two parts; HDFS and MapReduce. HDFS is used for storing data in distributed environment. MapReduce is popular for its simplicity, scalability, and fault-tolerance. MapReduce is one of the key approaches to meet the demands of computing massive datasets.

II. RELATED WORKS

Over the last five years, many research works was carried out and published paper in Big Data. In enterprise big data analytical tools, author reviewed the different big analytics tools and compared with different functionalities. In that, proprietary data analytical tools such as Citus DB, Google BigQuery, Greenplum HD and Hadapt and open source data analytical tools such as Hadoop, Apache Drill and Dremel, Apache Hive, Cloudera Impala and Graph are reviewed .

D.P.Acharjya et al. has discussed some current techniques for analyzing big data. This paper reviewed three important emerging tools namely MapReduce, Apache Spark, and Storm with three factors like batch processing, stream processing, and interactive analysis. Sabitha et al. has discussed about Big Data and some of the emerging Big Data tools and supported databases.

III.CHALLENGES OF BIG DATA ANALYTICS

Various challenges of big data analytics are available to handle big data with different computational techniques in different platforms. The challenges are classified into four categories such as data storage, computational environment, algorithms and tools in big data analytics [6].

Categories

- **Data storage** means that Database technology (including parallel databases) to be neither well-suited nor cost-effective for big data. To handle the challenge of Web-scale storage, the Google File System (GFS) was created for extremely large files whose content can span hundreds of machines in shared-nothing clusters created using inexpensive commodity hardware [5]. It should be able to handle structured, un-structured and semi-structured high volume data. As well as the newly updating information have to be supported for analysis process.

- **Computational environment** deals with processing modes for decision-making power with the large volume and high growth of information assets. The processing mode defines the working platform of big data such as single system or multiple systems with distributed and parallel processing.

- The various **algorithms and tools** are available for handling big data. These algorithms and tools of big data to exploring the hidden value in the big data with high accurate rate.

This paper analyses how these challenges can be achieved for high-quality, up-to-date technology related to big data analytics.

IV. ALGORITHMS OF BIG DATA ANALYTICS

A variety of algorithms are available for analyzing valuable information in big data depending upon the user need. The algorithms are classified into three categories based on their functionalities. They are Clustering, Classification and Collaborative filtering.

Clustering is a technique for grouping of similar or dissimilar objects together. Classification techniques classify the category of a given 'object' belongs to with respect to several input attributes.

Some of the algorithms are used to reduce the iteration in the process. These algorithms will run as one-pass clustering algorithm in a single machine. Other algorithms are used to process the data in parallel and distributed environment and run as multi-pass clustering using MapReduce concept. These parallel and distributed algorithms use multiple machines to speed up the computation and increase the scalability to handle big data.

The algorithms and platforms which are supported are summarized in Table I.

TABLE I
ASSESSMENT OF ALGORITHMS AND PLATFORM

Type	**Algorithms**	**Platform**
Collaborative Filtering	User-Based Collaborative Filtering	Single Machine
	Item-Based Collaborative Filtering	Single Machine/MapReduce
	Matrix Factorization with Alternating Least Squares	Single Machine/MapReduce
	Matrix Factorization with Alternating Least Squares on implicit Feedback	Single Machine/MapReduce
	Weighted Matrix Factorization, SVD++, Parallel SGD	Single Machine
Classification	Logistic Regression-Trained Via SGD	Single Machine
	Navie Bayes/Complementary Navie Bayes	MapReduce
	Random Forest	MapReduce
	Hidden Markov Models	Single Machine
	Multilayer Perceptron	Single Machine
Clustering	Canopy Clustering	Single Machine/MapReduce
	K-Means Clustering	Single Machine/MapReduce
	Fuzzy K-Means	Single Machine/MapReduce
	Streaming K-Means	Single Machine/MapReduce
	Spectral Clustering	MapReduce
	CLARANS: Fighting with Exponential Search Space	Single Machine

	IRCH: Fighting with Limited Memory	Single Machine
	URE: Fighting with the Irregular Cluster	Single Machine
	DisCo: Co-Clustering with MapReduce	Single Machine/MapReduce
	BoW: Subspace Clustering with MapReduce	MapReduce
	DBDC: Density-Based Clustering	MapReduce

V. BIGDATA ANALYTICS TOOLS

This section presents the Big Data Analytics tools that were found in the survey. It describes about the functionalities, benefits and drawbacks for commodity and open source Data Analytics tools. Refer Table II to understand the benefits and drawbacks for commodity Data Analytics tools. Refer Table III to understand the benefits and drawbacks for open source Data Analytics tools.

TABLE II
ASSESSMENT OF FUNCTIONALITIES, BENEFITS AND DRAWBACKS ABOUT COMMODITY TOOLS

Tools	Functionalities	Advantages	Disadvantages
Citus DB	Supports parallel and distributed data.	Scalable and robust analytical tool. Fast and flexible access to massive volumes of data. Partitions massive data and executes queries efficiently.	Real time insertion and deletion is not available in Citus DB. Does not support real time analytics.
Google BigQuery	Data are extracted on a periodic basis from source systems. Fully-managed and cloud based interactive query service for massive datasets.	Reducing time and expense to query data. Run ad hoc queries on multi-terabyte data sets in seconds.	Handles only structured data using SQL It will not update existing data.
Greenplum HD	Revamps Hadoop to operate more like a relational database. Modular and flexible solution for analyzing data.	Easiest, most dependable, and fastest Hadoop distribution on the planet. Improves Reduce phase performance drastically.	Can't use dump utilities to dump external table contents. Not able to select data from single node Hadoop on the same cluster.
Hadapt	Tackles "big data" analysis in the cloud Analyzes both structured and unstructured data in one platform Runs SQL queries 50x faster than Apache Hadoop+Hive	Best way to perform complex analytics on structured and unstructured data in Hadoop. Adapt's Adaptive Analytical Platform requires no connectors.	Complex queries can perform poorly because of unpredictable node performance and availability.

TABLE III
ASSESSMENT OF FUNCTIONALITIES, BENEFITS AND DRAWBACKS ABOUT OPEN SOURCE TOOLS

Tools	Functionalities	Advantages	Disadvantages
Hadoop	Allows the distributed processing of massive data sets across different sets of servers. Can detect and handle faults at the application layer without depending on high end hardware.	Scalable. Flexible. Fault tolerant. Cost effective**.**	Not suited for small data. MapReduce requires a lot of time to perform these tasks thereby increasing latency.

Apache Drill and Dremel	Executes large-scale, ad-hoc queries, with lower latencies. An interactive data analysis tool for large datasets.	Drill does not require schema or type specification for the data in order to start the query execution process. Do not need to create and manage tables/views in a metadata repository, or rely on a database administrator group for such a function.	Cannot read xml data (only json, csv, parquet..) No single-row sub query support. Limitations on Join No schema, so it might create some confusion.
Apache Hive	Designed to enable easy data summarization, ad-hoc querying and analysis of large volumes of data. HiveQL statements are automatically translated into MapReduce jobs.	Supports external tables which make it possible to process data without actually storing in HDFS. Fits the low-level. Interface requirement of Hadoop perfectly.	Does not offer real-time queries. Does not offer row-level update. Provides acceptable latency for interactive data browsing. Sub-queries are not supported.
Cloudera Impala	Impala is the highest performing SQL engine. The fastest way to access data that is stored in Hadoop Distributed File System.	Don't need data transformation and data movement for data stored on Hadoop. Processes data that is stored in HDFS at lightning-fast speed.	No support for Serialization and Deserialization. Need to refresh the tables always. Does not provide any support for triggers.
Giraph	A Framework for performing offline batch processing of semi-structured graph data on a massive scale.	Recursive problems are nicely solved iteratively.	Each job is executed N times. Mappers send PR values and structure. Extensive IO at input, shuffle & sort, output.
Apache Mahout	Build and support a community of users and contributors such that the code outlives any particular contributor's involvement or any particular company or university's funding. Focus on real-world, practical use cases as opposed to bleeding-edge research or unproven techniques.	Provides scalable and commercial machine learning techniques for large scale and intelligent data analysis applications.	The collection of algorithms implemented for use continues to expand with time.
Apache Spark	Allocates the resources and the worker nodes to do the data processing in the form of tasks. Helps to run an application in Hadoop cluster, up to 100 times faster in memory, and 10 times faster when running on disk.	Supports SQL queries, streaming data, machine learning, and graph data processing. Provides support for deploying spark applications in an existing Hadoop clusters Supports iterative computation, improves speed and resource utilization.	Consumes a lot of Memory and issues around memory consumption are not handled in a user friendly Manner. Apache Spark would take large resources.
Dryad	Implementing parallel and distributed programs for handling large context bases on dataflow graph. Provides a large number of functionality including generating of job graph, scheduling of the machines for the available processes, transition failure handling in the	Users do not need to know anything about concurrent programming.	Needs to create a graph for virtual nodes and the communication is done via local write/distant read.

	cluster, collection of performance metrics, visualizing the job.		
Storm	A distributed and fault tolerant real time computation system for processing large streaming data.	Able to process over a million jobs in fraction of a second on a node Integrated with Hadoop to harness higher throughputs. Easy to implement and can be integrated with any programming language.	Does not run on Hadoop clusters but uses Zookeeper and its own minion worker to manage its processes.
Jaspersoft	It can quickly explore big data without extraction, transformation, and loading (ETL).	Supports structured and un-structured types of data. Supports database, like mango DB, Cassandra, Redis, Riak, Hbase, etc. It is independent of the operating system used.	Report developers need to manually manage the dependency between the master report and sub-report files. Too many sub-reports can result in very poor performance because each sub-report opens its own database connection, thread, and queries.
Splunk	A real-time and intelligent platform developed for exploiting machine generated big data. It combines the up-to-the-moment cloud technologies and big data. It helps user to search, monitor, and analyze their machine generated data through web interface.	Can be used for anyone within an organization. Lots of plugins customizations. Impressive dashboard with search and charting tools.	Expensive for large data volumes. Optimizing searches for speed is more.
Gridgain	This is an alternative of MapReduce and this also supports HDFS. This is used for fast analysis of real time data using in memory processing.	It is compatible with Hadoop DFS and it offers a substitute to Hadoop's MapReduce.	It is not designed for fast parallel processing and endless expansion
HPCC	Its expansion is High performance computing cluster. Both paid version and open source is available.	It can provide a data intensive supercomputing environment which can helps to work with large amount of data sets. It provides an efficient processing to solve the problems of big data.	Rate and licensing
Apache Cassandra	Wide-column store based on ideas of Big Table and DynamoDB.	This is a high performance, scalability and high availability software. It has a good built in Cache.	Cassandra supports row level atomic operation and Atomic and Isolation per partition using Light Weight Transactions.
KNIME	It can discover the hidden potential of your data, mine for fresh insights, and can predict new futures by analysing the data.	Intermediate results can be inspected at any time and new nodes can be inserted. The data tables are stored together with the work flow structure and the nodes settings.	Preliminary results are not available as soon as possible as if real pipeling were used.
KarmaSphere	It provides a good environment for handling the big data sets of structured and unstructured data.	Provides data analysts immediate entry to structured and unstructured data on Hadoop, through SQL and other familiar languages. Can make ad-hoc queries, interact with the results, and iterate – without the aid of IT.	High Complexity

Tableau	It is very suitable for generating the information about cash in the memory.	Supports structured and un-structured types of data. Excellent mobile support. Low-cost solution to upgrde.	Lack of predictive capabilities. Risky security Change Management Issues
Hbase	It provides real time access to Hadoop and it provides distributed and scalable data set. It is modeled after Google's BigTable and it used Java for programming.	Perfect for real-time querying of Big Data. HBase enjoys Hadoop's infrastructure and scales horizontally using off the shelf servers.	Does not natively support secondary indexes. Scale limitlessly as load and performance demands increase simply by adding server nodes.

VI. TOOLS USED IN INDUSTRIES

By now, many companies have decided that big data is not just a buzzword, but a new fact of business life -one that requires having strategies in place for managing large volumes of both structured and unstructured data. And with the reality of big data comes the challenge of analyzing it in a way that brings real business value. Business and IT leaders who started by addressing big data management issues are now looking to use big data analytics to identify trends, detect patterns and glean other valuable findings from the sea of information available to them. The following table lists some tools used in various industries to manage big data.

TABLE IV
TOOLS USED IN VARIOUS INDUSTRIES

Company	Big Data Analytics Tools
IBM	Apache Hadoop, InfoSphere, JAQL
Cloudera	CDH, Cloudera Standard, Cloudera Enterprise
Oracle	Oracle Big Data Appliance
Google	BigTable, DRIME
Yahoo!	Sherpa, PIG
Amazon	SimpleDB, Dynamo
Microsoft	Dryad
Facebook	Apache Cassandra, HIVE
Hypertable	HyperTable
ASF	Apache CouchDB
Apache	HBASE

VII. CONCLUSIONS

Due to Increase in the amount of data in the various applications, it becomes difficult to handle the data, to find Associations, patterns and to analyze the large data sets. In this paper, we explore the challenges of Big Data Analytics. Compared to the findings in the last survey of Big Data Analytics, there has been a strong development. Many more tools are available and their maturity has improved. In the current survey, four commodities and eighteen Open Source Big Data Analytics tools were considered. The Big Data Analytical tools category has improved a lot compared to the previous survey. Thus, in this paper various tools with functionalities, benefits and limitations are reviewed.

REFERENCES

[1] Sabitha M.S, Dr.S.Vijayalakshmi, "Big Data–Literature Survey", International Journal for Research in Applied Science & Engineering Technology (IJRASET), Volume 3, Issue XII, December 2015 IC Value: 13.98 ISSN: 2321-9653.

[2] Satanand Mishra, Vijay Dhote, "Challenges in Big Data Application: A Review", International Journal of Computer Applications (0975–8887) Volume 121–No.19, July 2015.

[3] K.Sharmila, Dr.S.A.Vethamanickam, "Survey on Data Mining Algorithm and Its Application in Healthcare Sector Using Hadoop Platform", International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, ISO 9001:2008, Volume 5, Issue 1, January 2015.

[4] C. Lakshmi, V. V. Nagendra Kumar, "Survey Paper on Big Data", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 6, Issue 8, August 2016. ISSN: 2277 128X.

[5] M.Dhavapriya, N.Yasodha, "Big Data Analytics: Challenges and Solutions Using Hadoop, Map Reduce and Big Table", International Journal of Computer Science Trends and Technology (IJC S T), Volume 4 Issue 1 , Jan - Feb 2016, ISSN: 2347-8578.

[6] D.P.Acharjya, Kauser Ahmed P, "A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 2, 2016.

[7] Dr.R.Kousalya, T.Sindhupriya, "Review On Big Data Analytics And Hadoop Framework", International Journal of Innovations in Scientific and Engineering Research(IJISER), ISSN: 2347-9728(print), Vol 4 Issue 3MAR 2017/101.

[8] Shlomi Dolev, Patricia Florissi, "A Survey on Geographically Distributed Big-Data Processing using MapReduce", IEEE TRANSACTIONS ON BIG DATA, June 2017.

[9] Priya Dahiya, Chaitra.B, "Survey on Big Data using Apache Hadoop and Spark", International Journal of Computer Engineering In Research Trends" Volume 4, Issue 6, June-2017, pp. 195-201 ISSN (O): 2349-7084.

[10] Dr.R.Saravanakumar and Dr.C.Nandini, "A Survey on the Concepts and Challenges of Big Data: Beyond the Hype", Advances in Computational Sciences and Technology ISSN 0973-6107 Volume 10, Number 5(2017) pp. 875-884.