# Statistical Analysis & Prediction of Web Server Logs using Big Data & Hadoop

**Shubham Yadav[1] Mr. Anwar Sarkeja[2] Versha Matre[3]**
[1,2,3]Vikrant Institute of Technology & Management, Indore, India

*Abstract*— Today, it is not uncommon to face data deluge that has brought challenges to every sector across all industries. On the other hand, social networks has brought a platform that facilitates human interaction among themselves which is creating a room to everyone to produce huge data sets using computers and smart phones as well. Moreover, data creation rate in variety of formats is yielding real challenges to traditional technologies. Big Data processing and visualization is current challenge due to data growth with high velocity in variety of data type. To tackle Big Data problems, the methodology applied is in detail investigation of current challenges, identification of technology frameworks and ecosystems, design solutions, implementation of the designed solution and test of implemented solution using Big Data set is taken place. Hadoop ecosystem which is starting point of technological shift from traditional technologies to more advanced and different has shown the change of data and technology landscape.

***Key words:*** Big Data, Hadoop, Mapreduce Hadoop Distributed File System, Visualization

## I. INTRODUCTION

In general, big data is immersed with wealth of new insights across all industries, expertise and life to provide and guide discoveries and innovations. Smart devices are the main actors in creations and utilization of big data that shifts tradition practices into modern life style and even research direction is reversed from 'theory to data' to 'data to theory' paradigm. It is estimated that total population and total mobile phones are approximately 6.8 billion and 6 billion respectively [9]. Moreover, mobile applications have changed the way people think, live and transact in smart spaces and time.

Hadoop is a successful implementation of Google's MapReduce programming model and is now an Apache Foundation open source project. It enables the processing of large volumes of structured and unstructured data using cluster of commodity hardware in a simple, scalable, economical and reliable way. Hadoop is primarily installed on Linux clusters even though it could be installed on Windows platforms using emulators like Cygwin. Hadoop provides the Hadoop distributed file system, which can store and replicate data over a cluster using the MapReduce.

## II. PROBLEM DEFINITION & PROPOSED SOLUTION

### A. *Problem Domain*

Processing speed: MapReduce require lot of time to perform these tasks thereby increasing latency. Latency:In Hadoop, MapReduce framework is comparatively slower. Caching: In Hadoop, MapReduce cannot cache the intermediate data in-memory for a further requirement which diminishes the performance of hadoop.

### B. *Proposed Solution*

#### 1) *Hadoop*

Hadoop is a framework that comprised of a number of components for its proper functioning and returning intended results. As shown in Fig. 1.1, major components are NameNode, secondary NameNode, Data Node, JobTracker and TaskTracker. Each of these components has well designed to accomplish certain task in general. NameNode is the master or brain of the whole Hadoop system and its main duties are tracking address of all stored files, listening heartbeat message of all DataNodes, manages schedules of JobTracker, holds information about inter rack status and so on. Secondary NameNode is taken as backup node which takes snapshot of NameNode in order to restore normal functioning after its failure. DataNode is a slave node where the data is deposited and data manipulation takes place before aggregation activities started. JobTracker is the one that orchestrates all tasks to be carried out throughout across task assigned nodes. TaskTracker is a slave by its very nature and its responsibility is carrying out ordered task to be performed at low level which is individual nodes or commodity machines where data is stored [25].

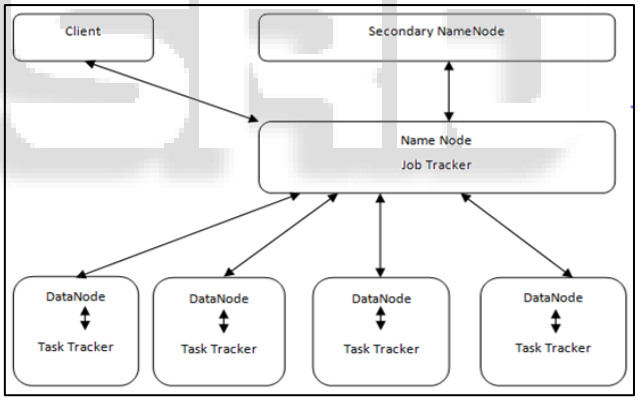


Fig. 1.1: Hadoop Components

Hadoop is also an ecosystem which consists of a set of related projects that are implemented to facilitate customization based on experience and expertise of organizations. The major projects are Hadoop Streaming which enables script writing for those who are familiar on script languages, Hadoop Hive which provides SQL writing capabilities for those who are working with SQL languages, Hadoop Pig which is purely procedural language that supports data pipeline scenarios and Hadoop HBase that stands with real time data retrieval rather than batch processing. On top of these, Hadoop Distributed File System and MapReduce are the major projects that can be taken as backbone of the ecosystem [27].

In general, Hadoop MapReduce architecture provides an environment where parallel processing is done in large set of commodity nodes. Each node is a single unit of machine which executes assigned task in full responsibilities without depending on other machines for its execution. As

 

mentioned above, Hadoop MapReduce framework is purely software solution for current limitation of space and processing capacity. Instead of putting single machine with vast space and high speed like super computer which is actually very expensive; additionally, it demands top expertise to setup and for ongoing operations as well, there comes very cheap solution that can be implemented with reasonable investment. Return on investment of new big data technologies is amazingly high in terms of insight that may be extracted from processing untapped, unstructured, data set due to traditional technological limitation. From overall dataset 90% of data is unstructured data and it is rich with insights that can shape usual practices of every industry to modern way of accomplishing activities or achieving objectives [28].

### *C. Hadoop Distributed File System (HDFS)*

File storage structure has been changed to maintain distributed file storage along with ensuring fault tolerance. In 2004 [31], Google started to change algorithm in order to boost its search capability by indexing whole files in the internet. As result, it has released white paper on Google File System which was initiated new file system, Hadoop Distributed File System, to be developed by open source community. It is a mechanism to handle large files in distributed manner over multiple of nodes in the form of chunks that each chunk will be replicated as per set replication factor at time of configuration. Whenever there is failure of one or more nodes, data will be moved from failed nodes to active nodes where accommodation space is available. In addition, it creates an environment where horizontal scaling is easily achieved to scale out to hundreds of thousands of commodity machines [25].

In addition, as shown in Fig. 1.2, HDFS is becoming the center of architectural change for current computational practice by improving performance of latency and throughput. The impact of performance improvement, at level of software (Hadoop Framework) rather than hardware, is attracting giant companies like Facebook, google, yahoo etc. so as to adopt the principle and practice as well. It enhances read/write operations of local file chunks by moving computation to where data is stored. It handles very large files which be gigabytes or more by reading or writing sequentially to/from nodes therefore there is no need to bring data to memory in order to manipulate so the role of primary memory is becoming insignificant [12].

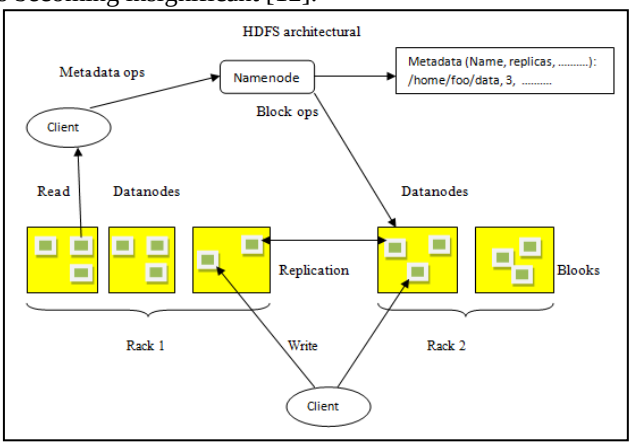


Fig. 1.2: HDFS Architectural View

#### *1) MapReduce*

MapReduce is a programing model for processing large scale datasets in a single pass in clusters of thousands of nodes by assuring fault tolerance and it supports two types of functions for different purpose of duties [33]. Map Task is a function which is used to allocate data to nodes based on replication factor set. On the other hand, Reduce Task is also a function for aggregation of data results according to request initiated by client.

Even though Map Task and Reduce Task are two functions that are clearly visible to all parties, there are other functions in between Map Task and Reduce Task to play a role for supportive activities such as splitting, sorting, shuffling etc. Map Task depends on split function before distributing chunks of a file to nodes as per replication factor. In the same fashion, Reduce Task is heavily reliance on shuffle and sort functions in order to aggregate the result. Split function accomplishes the task of chopping file into preset size of chunks so that Map Task will able to send these chunks to a designated nodes after gathering information for free space availability. Mappers create key/value pair for all coming chunks while storing them. Shuffle function, in addition, is responsible for taking input from Mappers and categorizing keys based on their groups. Sort function plays a role of sorting keys according their values before Reducers take in. Finally, Reducers combine similar keys and aggregate their values at each node which is local disk where the data resides.

#### *2) Data Processing (Technology Stack)*

In big data technology stack scenarios as depicted, style of data processing is shifted from retrieval of data from hard disk and sending it primary memory for processing to sending computation where data resides. This is great innovation for pet scale data in order to avoid disk access and network traffic bottlenecks so that results will be achieved in reasonable time.

Major data processing paradigm shifts has been brought through implementation of MapReduce framework on top of Hadoop Distributed File System. Even though this has reduced burden of data transfer and manipulation to the level of uniformity in dealing big data, it has still challenge in terms of generality to the specialists in the field by forcing them to know programming language implementation and its complexity.

Java [35] is the programming language which has been used to implement as an open source code and it is customizable by interested parties in order to adopt wherever Hadoop is used as a means to process big data. A lot of companies as well as expert communities are adopting Hadoop ecosystem and then adapting it to their own favorite environment by adding projects as depicted in Fig. 2.6. For instance, Microsoft is one of big providers of Big Data products as well as services but it has adopted Hadoop for big data storage and processing so its projects are totally dependent on Java libraries as foundation. Other programming and scripting languages are becoming part of Hadoop ecosystem as plugin onto MapReduce framework so that working on Hadoop is made as simple as performing usual projects using those languages. To mention some of these languages: Python, Ruby, SQL-like languages, Script-

 

like languages etc. all of them run on the top of MapReduce framework.

Apache Hive [33] is a project that act like data warehouse for Hive Query Language (HQL) which provides for users a capability to process data using SQL-like language. In general, it abstracts details of MapReduce implementation such that users can inject their task into MapReduce without delving how it functions. The tasks are either sending data for storage or retrieval specific result after processing data from a set of nodes, commodity hardware. Actually, Hive queries are converted into Hadoop Jobs to run whether Map Task or Reduce Task which does not mean that rational database structure is imposed on MapReduce framework rather HQL queries are interpreted as task so that users will not be forced to write Map or Reduce Task programs to achieve data analysis objectives. Even if HQL is SQL-like language, it has additional features that are completely dissimilar to SQL queries, for example structs, maps (key/value pairs) and array.

Apache Pig [33] is a scripting language that eases to write jobs and send as MapReduce jobs so as to be executed against Hadoop. It is a platform which is openly extensible for data loading, manipulating and transforming by using scripting language is called Pig Latin. It supports complex and sophisticated data manipulation though it is simple scripting language.

SQOOP [7] is one of highest projects that is used to link relational database and Hadoop projects together. So it facilitates data movement from relational databases, structured data, to Hadoop, schema less or unstructured data, and vice versa. It is plug and play extensible framework that helps developers to program through the SQOOP application programming interface (API) so as to add new connectors.

Apache HCatalog [30] has a role to abstract data view from HDFS files stored in Hadoop into tabular form. It provides integrated abstraction form for all other projects that relay on tabular structure of data view. For instance, Pig and Hive use this abstraction in order to reduce complexity of reading data from HDFS. Despite the fact that HDFS can be any data format and stored anyplace in the cluster, HCatalog gives a means for mapping to file formats and locations into tabular view of the data. In addition, it is open and extensible for proprietary file formats.

HBase [36] is a project that supports the functionality of NoSQL (Not only SQL) database on the top of HDFS. It is a storage of large column that could be limitless number of columns along with billions of rows that facilitate fast access to huge datasets or large tables which is sparsely stored. It has a functionality of Data Modification Language (DML) [37] which supports inserts, updates and deletes; however, Hadoop by its nature it is a write once and read many or infinite times. In spite of its rational database nature, it does not provide full features of relational databases such as typed columns, security, enhanced data programmability and query language capabilities.

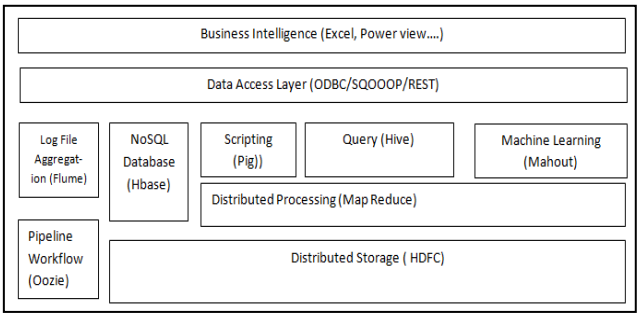


Fig. 1.3: Hadoop Ecosystem

## III. Conclusion & Future Work

In this work we applied Hadoop Map Reduce programming model for analyzing web server log files where data get stored on multiple nodes in a cluster so that access time required can be reduced and Map Reduce works for large datasets giving efficient results. In order to have summarized results for a particular web application, we need to do log analysis that will help to improve the business strategies as well as to generate statistical reports. Using Visualization tool for log analysis will provide us graphical reports showing hits for web pages, user's activity, in which part of website users are interested, traffic sources, etc. From these reports business communities can evaluate which parts of the website need to be improved, which are the potential customers, from which geographical region website is getting maximum hits, etc., which will help in designing future marketing plans. Log analysis can be done by various methods but what matters is response time. Hadoop Map Reduce framework provides parallel distributed processing and reliable data storage for large volumes of log files. Here hadoop's characteristic of moving computation to the data rather moving data to computation helps to improve response time.

## References

[1] Thanakorn Pamutha, Siriporn Chimphlee and Chom Kimpan, "Data Preprocessing on Web Server Log Files for Mining Users Access Patterns" .International Journal of Research and Reviews in Wireless Communications of Vol. 2, No. 2, ISSN: 2046-6447 ,June 2012.

[2] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, Robert Chansler, "The Hadoop Distributed File System" Yahoo, IEEE, 2010.

[3] Chris Sweeney, Liu Liu, Sean Arietta and Jason Lawrence, "HIPI: A Hadoop Image Processing Interface for Image-based MapReduce Tasks", University of Virginia, 2010.

[4] Mohamed H. Almeer, "Cloud Hadoop Map Reduce For Remote Sensing Image Analysis" Journal of Emerging Trends in Computing and Information Sciences, Vol. 3, No. 4, ISSN 2079-8407, April 2012.

[5] Muneto Yamamoto and kunihiko Kaneko, "Parallel Image Database Processing with Mapreduce and Performance Evaluation in Pseudo Distributed Mode" International Journal of Electronics Commerce Studies, Vol.3, No.2, pp.211-228, doi: 10.7903/ijecs.1092, 2012.

[6] Natheer Khasawneh and Chien-Chung Chan, "Active User-Based and Ontology-Based Web Log Data Preprocessing for Web Usage Mining" Proceedings of

 

the 2006, IEEE International Conference on Web Intelligence.
[7] Murat Ali Bayir, Ismail Hakki Toroslu, "Smart Miner: A New Framework for Mining Large Scale Web Usage Data" WWW 2009, Madrid, Spain.ACM 978-1-60558-487-4/09/04, April 20–24, 2009.
[8] P. Srinivasa Rao, K. Thammi Reddy and MHM. Krishna Prasad, "A Novel and Efficient Method for Protecting Internet Usage from Unauthorized Access Using Map Reduce". I.J. Information Technology and Computer Science, 03, 49-55, 2013.
[9] Sayalee Narkhede and Tripti Baraskar, "HMR Log Analyzer: Analyze Web Application Logs over Hadoop MapReduce", International Journal of UbiComp (IJU) vol.4, No.3, July 2013.
[10] Jian Wan, Wenming Yu and Xianghua Xu, "Design and Implement of Distributed Document Clustering Based on MapReduce", Proceedings of the second Symposium International Computer Science and Computational Technology (ISCSCT '09), pp.278-280, 26-28 Dec.2009.
[11] Http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html, NASA Logs Files.
[12] DougCutting "Hadoop Overview", http://research.yahoo.com/node/2116.
[13] Michael Cardosa, "Exploring MapReduce Efficiency with Highly-Distributed Data". MapReduce'11, ACM, USA June 2011.

 