Hit Count Analysis Using BigData Hadoop

Navonil Sarkar
Department of Computer Engineering
DPCOE
Pune, India
navonilsrkr34@gmail.com

Jitin Michael
Department of Computer Engineering
DPCOE
Pune, India
g3jitin87@gmail.com

Vivek Kumar
Department of Computer Engineering
DPCOE
Pune, India
vivekumar1995@gmail.com

Rahul Chaurasia
Department of Computer Engineering
DPCOE
Pune, India
rahulchaurasia_3895@yahoo.com

Abstract - *Log files are generated every second and are increasing at an alarming rate. Every time we access anything from internet we generate a log file at the corresponding server. These logs contain various information about the user and the website. These logs are stored and analyzed to extract various useful information about user and are used for various reasons such business decisions, security, performance analysis etc. In the past few years the number of internet users has increased drastically. As a result the log files are also generated in huge amount regularly. Terabytes and Petabytes of log files are generated every day and storing such huge amount of data is a challenge. Traditional tools are unable to store and process such huge amount of data. To overcome this problem Big Data is the solution. The main purpose of this paper is to overview the storing and processing of log files in Big data environment. To store the log files, HDFS is used. Hadoop clusters are used for efficient storage of log files. Hadoop framework is used to analyze the log files and summarize the same using some front end tools. Different Hadoop based algorithms are used to extract the necessary information from the log files and present them for various clients*

Key Words: **Bigdata , Hadoop, Map, Reduce, Pig, Hive, HDFS**

1. INTRODUCTION

In today's time, the use of internet is increasing day by day. Almost everyone today is using internet for a variety of purposes like social networking, gaming, gaining knowledge, spreading knowledge, entertainment etc. etc. With this spread of usage of internet millions of Terabytes and Petabytes of data is being generated almost every minute around the world. When someone likes a photo on Facebook data gets generated, someone uploads something on Facebook data gets generated, someone makes a Tweet data gets generated, Someone sends an email data gets generated. Now this data needs to be stored somewhere and as well as managed. Hadoop Distributed File System (HDFS) can be used to store and manage this data efficiently since Hadoop is an open-source software framework used for distributed storage and processing of very large data sets. It consists of computer clusters built from commodity hardware. It can incorporate large amount of clusters. A number of machines can be organized in the form of clusters and Millions of Terabytes of data can be stored in these clusters and managed. Hadoop is the core platform for structuring Big Data, and solves the problem of formatting it for subsequent

analytics purposes. Hadoop uses a distributed computing architecture consisting of multiple servers using commodity hardware, making it relatively inexpensive to scale and support extremely large data stores. These severs are highly powerful and distributed across many platforms. Thus we can implement these Hadoop clusters in a distributed manner to increase scalability.

Map Reduce: Now how we get data is through Map Reduce as name implies it is a two step process. There is a Mapper and Reducer programmers will write the Mapper function which will go out and tell the cluster what data point we want to retrieve. The Reducer will then take all of the data and aggregate. Hadoop is a batch processing here we are working on all the data on cluster, so we can say that Map Reduce is working on all of data inside our clusters. There is a myth that one need to be understand java to get completely out of clusters, in fact the engineers of Facebook built a subproject called HIVE which is SQL interpreter. Facebook wants a lot of people to write adhoc jobs against their cluster and they are not forcing people to learn java that is why team of Facebook has built HIVE, now anybody who is familiar with SQL can pull out data from cluster.

Now it is clear that Map reduce will be used for analyzing the data and more importantly reducing the data by eliminating the data which is not that much of importance to the data which is of utmost importance. Once we have organized and analyzed this data we will be looking to present this data to the users in an organized and structured way so that users can make the best use of this data according to their business or personal needs.Before talking about how we are going to present this data, Let's talk about some of the uses of this data. Consider someone running a small scale or a large scale business can use these data sets. Someone can use these datasets to decide a particular website on which he should advertise to excel his business. He would be able to gather information about what kind of users are accessing a web site from which locations and how many number of them are accessing. And based on this knowledge he would have a fine idea of on which websites he should be advertising to make sure that his advertisements would bring maximum business to him. For Example - A man running a pest control farm in Pune would like to put his advertisement on a website which gets maximum hits from Pune.

2. Literature Survey

We have stepped into data age due to exponential growth of data from various sources. Various companies like Google, Yahoo, Microsoft etc. those have a huge number of users are facing much difficulties with their data. Conventional storing and processing systems are unable to handle such huge amount of data. In 2004, Google introduced Map Reduce and Google File System to scale up the data processing needs. Google's MapReduce is then implemented by many search engines and ultimately adopted by Hadoop. Now, Hadoop has become the core part of computing in many web companies.The data generated by users are now in terms of Petabtes and Zetabytes and storing that amount of data calls for Hadoop. Hadoop can handle terabytes or petabytes of data thus it is called as a Big Data technology.

Today everything is going online. From ticket reservation to exam form submission every single field is growing and are shifted to the digital world i.e. Web. In order to get more and more customers, they need to analyze which prior customers are interested, where it is more popular, which kind of service people are interested in, etc. Such kind of analysis helps in improving advertize of less popular services, promote popular services in order to make business scale. Thus the data is flowing at a very higher rate. To control such huge amount of data , Hadoop is the solution. Hadoop have the capability to handle such big datasets. HDFS comes into picture when we are dealing with bigdata.

2.1 Difficulties with the existing system

Previously when data was generated in a steady rate, the main goal was to increase the processing power of the system. Different scheduling algorithms were used to increase the computational power of the systems. Then evolved distributed system which allowed a single job to run on multiple machines. From many years, High Performance Computing and Grid Computing have been doing data processing using SAN. The main disadvantage of using SAN is the single point of failure. Hadoop is then introduced, which have the capability to store and process huge amount of data. HDFS is used for efficient storage of data. Various data nodes are used to store data in HDFS. Multiple clusters can be used to store the data. Analyzing those data is also a challenge. Mapreduce algorithms are used to analyze those big datasets. Hadoop MapReduce also differs from traditional relational databases the many ways. Relational databases can handle only some gigabytes of data where Hadoop can handle terabytes or petabytes of data. Secondly, relational databases works over structured data only, there is a static schema. Hadoop MapReduce is suitable for unstructured data such as text file as well as for semi structured data .

2.2 Proposed Work

Our proposed work begins with the generation of log files from different websites upto the presentation of the log files in a desired format in a web page. Input to the system is web application log file . Log file is a simple text file consisting of URL, date, hit, age, country, state, country etc. Before storing the log files in HDFS, we have to clean the raw data. This process is known as Preprocessing. Preprocessing phase involves separation of fields using separator (like #) and removing unwanted noisy data which could be multimedia files, style sheets etc. The log files are then stored in HDFS and used for further processing. All the data is presented on a website. All the processed data after being analyzed is presented in tabular forms on a website. The details of the data is presented in form or graphs or may be in form of tables. It contains information about how many Hits are being recorded on a particular domain, Locations, Time Stamps etc. Now the user can view this data and understand it very easily and use this data according to their needs.

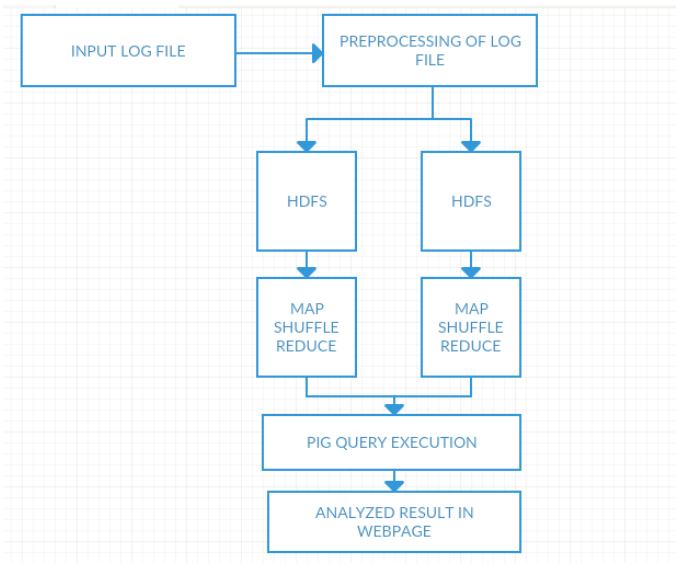

Figure 1: Workflow

The main phases are explained in detail below:

2.2.1 Log file Processing*:*

Log files generated from server contains many kind of noises and unwanted fields. Such unwanted things are removed from the log files and are arranged in a specific sequence. The log files are then stored in HDFS. Hadoop breaks down input file into smaller blocks of equal size and distributes these blocks over multiple nodes in the Hadoop cluster. Namenode and Datanode concepts are used in HDFS. Datanodes are those which contains the small chunks of data. Namenodes are those which contains the information about the data in datanodes. Now these data are given to the Map-Reduce algorithm.

2.2.2 MapReduce Algorithm:

MapReduce is a programming model which comprises of two main parts i.e. Mapping and Reducing. MapReduce takes log file as an input and feeds each record in the log file to the Mapper. Mapper processes all the records in the log file and Reducer processes all the outputs from the Mapper and gives final reduced results**.**

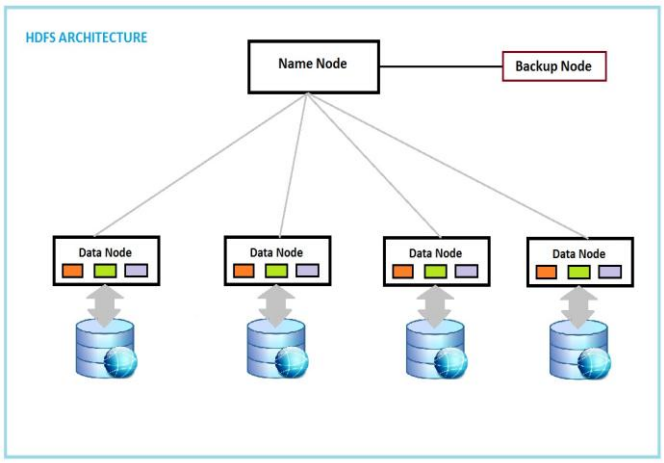

Figure 2: HDFS Architecture

Map Function:

Map function takes the log files and produces the result as (key,value) pair. Here Key is a non repetive data and it can have different values.

Reduce Function:

Input to reduce method is (key, value) pair. It sums together all counts emitted by map method.

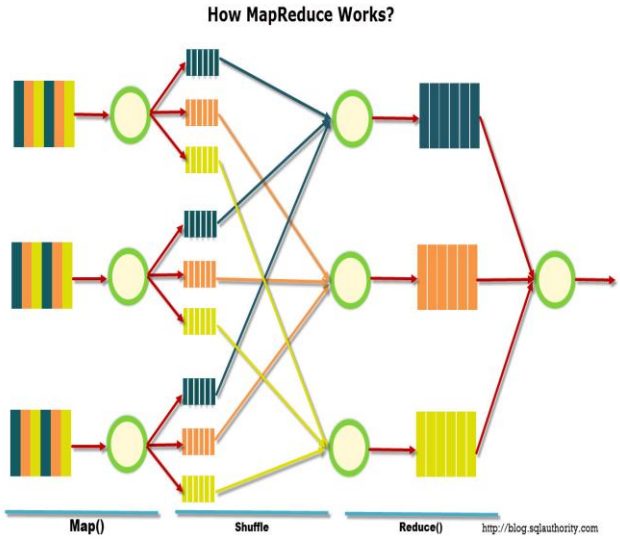

Figure 3: MapReduce

2.2.3 Presenting the processed data:

Map Reduce algorithm produces the data in form of (Key,Value) pair.Pig queries are used to store those data in our conventional way in our local system. The processed data can be then represented in form of graphs, charts or any other form. We will represent those data in any form according to the user. The data will be represented in a web page .

3. CONCLUSIONS

We have presented the log files using the Hadoop implementation. Map Reduce algorithm is used to process the log files. We have presented the processed data in different formats like graphs, charts etc. Statistical record of analysis is shown in various charts as bar chart and pie chart which gives hit count for various parameters in log file. Hit counts on various web sites that they receive is analyzed and organized with the help of Map Reduce. Information about when and form where a particular web site was accessed and by how many users will be analyzed and organized for the users. We have tested the performance on different cluster by varying the number of nodes or by changing the size of the nodes. Thus we have presented the logs with their corresponding hits. The web page shows hits with respect to time, location etc.

REFERENCES

[1] Shvachko K, Kuang H, Radia S and Chansler R. The Hadoop Distributed. File System in Proceedings of the 26th IEEE Symposium on Massive Storage Systems and Technologies, 2010.

[2] Feng Wang et al. Hadoop High Availability through Metadata Replication, IBM China Research Laboratory, ACM, 2009.

[3] S.Vikram Phaneendra & E.Madhusudhan Reddy "Big Data- solutions for RDBMS problems- A survey" In 12th IEEE/IFIP Network Operations & Management Symposium (NOMS 2010) (Osaka, Japan, Apr 19{23 2013).

[4] Kenn Slagter · Ching-Hsien Hsu "An improved partitioning mechanism for optimizing massive data analysis using MapReduce" Published online: 11 April 2013

[5] Sameer Agarwal†, Barzan MozafariX, Aurojit Panda†, Henry Milner†, Samuel MaddenX, Ion Stoica "BlinkDB: Queries with Bounded Errors and Bounded Response Times on Very Large Data"

[6] GrzegorzMalewicz, Matthew H. Austern, Aart J. C. Bik, James C.Dehnert, Ilan Horn, NatyLeiser, and GrzegorzCzajkowski,Pregel: A System for Large-Scale Graph Processing, SIGMOD"10, June 6–11, 2010, pp 135-145.

[7] Changqing Ji, Yu Li, Wenming Qiu, Uchechukwu Awada, Keqiu Li, Big Data Processing in Cloud Computing Environments, 2012 International Symposium on Pervasive Systems, Algorithms and Networks.

[8] Big Data' has Big Potential to Improve Americans' Lives, Increase Economic.

[9] Opportunities, Committee on Science, Space and Technology (April 2013).