Big Data Analysis: Organization and Mapping of Data

Srikanth Bethu[1,] B. Sanakara Babu [2]
[1]Assistant Professor, Department of Computer Science and Engineering, Gokaraju Rangaraju Institute of Engineering & Technology, Hyderabad,India.
[2]Professor, Department of Computer Science and Engineering,
Gokaraju Rangaraju Institute of Engineering & Technology, Hyderabad,India.

Abstract

Social networking sites (Facebook) generate 500TB of Data per day. This data has to be segregated, distributed and to be made easy to access in future. Making all these data segregated and provide data as per their necessity is termed as Big data analysis. The problem is inappropriate allocation and organization of Data which results in either excess in one Data format or deficient in another Data format. Some data might be running with less contingency when mapped into it. This can be resolved by the clear analysis of the data organization and utilization.

Keywords: *BigData, Data Analysis, Organization and Mapping of Data.*

1. Introduction

Organization of data on a large scale where the end user is the common man. Big Data is a new emerging domain in which the various types of datasets are grouped into single unit which can be used to perform several tasks and operation in an organized manner. It is also a storage mechanism in which complex and huge size data can be stored in a larger manner at the same time it can be accessed and processed in an efficient way.

The primary aim of big data analytics tools is to assist the various companies to get more aware of the decisions related to the industries or business by allowing data scientists, predictive modelers and related analytics professionals to examine a huge amount of trans action data with the help of available business intelligence programs. These tools can be used in social media, network activities, customer reports, sensors networks and emails. The importance of these tools also service serves various domains such as cloud computing, sensor based networks, internet of things and others Information or data related domains.

Big Data

Big data is a term for both structured and unstructured data. The data is of large volume with high complexity, which cannot be processed by conventionally available applications (Excel, RDBMS, etc.). The big data challenges include two things: 1) Data collection, storing data, and analysis of data. 2) Security and privacy of the information stored. The term big data extends its use to advanced methods of machine learning and deep learning to extract valuable information from data. With the use of advanced algorithms, accuracy in Big data analytics can be improved. This leads to confident decision making, and better decisions which result in the reduction of risks which in turn reduces costs and improved operational efficiency.

Big data is not purely a data, slightly it has become a entire subject, which is concerned with various devices, approaches and frameworks. It is a expansive phrase for arranging a data so large or complex which is difficult to deals with universal data applications. The information posted by the millions of people can view the info through various social media across the world.

Example: Face book and Twitter. Objection includes capture, duration, analysis, research, allocation, repository, transmission, decision, and data isolation. The marketing department is collecting the information about the feedback of their advanced mediums, such as social media like Face book. Analysis of arranging a data can find new interactions, to "spot business trends, prevent diseases, and combat crime and so on". The framework is required to manage and process large amount of structured and unstructured data in actual time and can preserve data privacy and security in order to harness the potential of big data. The drawback affects web research, investment and business information services. Professional of Media and Broadcasting and Governments, Scientists etc are frequently encountering the drawback due to arranging a large amount of data in many fields. The most significant attribute of the Big Data is the large amount of data is represented by diverse dimensionalities and heterogeneous. Twitter, MySpace, Orkut and LinkedIn etc. were the various sites from which large amount of data is generated.

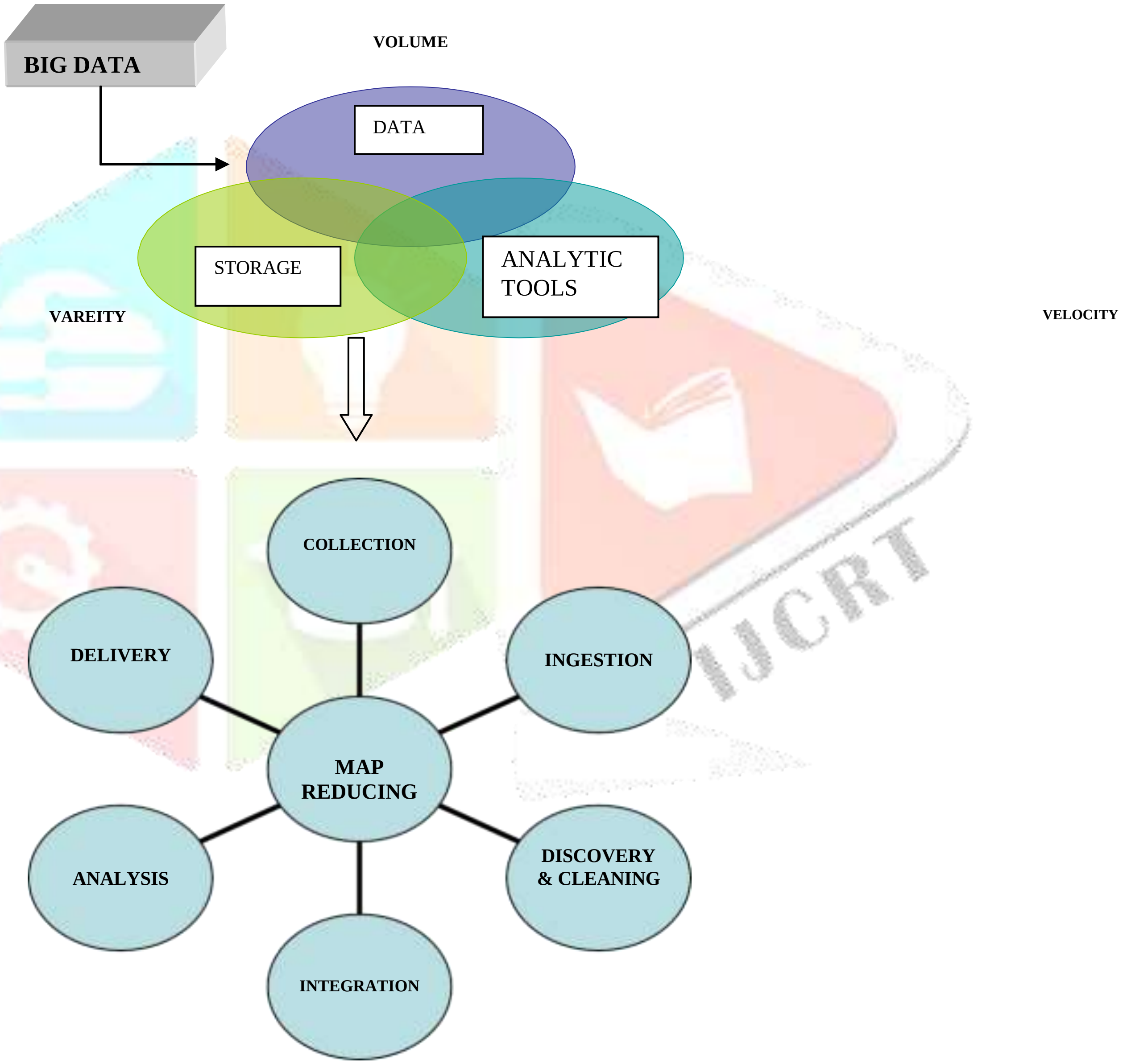


Fig.1. Big Data Analysis

The above Fig.1 refers the following analysis of Big data

- Collection refers to structured, Unstructured and Semi-Structured data from multiple sources
- Ingestion refers to loading vast amounts of data onto a single data store
- Discovering & Cleaning refers to understanding and content, cleanup and formations.
- Integration refers to linking entity extraction, entity resolution, and indexing and data fusion.
- Analysis refers to intelligence statistics, predictive and text analytics, machine learning.
- Delivery refers to querying, visualization real time delivery on enterprise class availability.

Map Reducing

Map Reduce is a software framework for processing big data sets in a distributed environment consisting of several machines on a cluster. The core idea behind Map Reduce is mapping your data set into a collection of [key, value] pairs, and then reducing overall pairs with the same key. The overall concept is simple, but is actually quite expressive when you consider that

a) Almost all data can be mapped into [key, value] pairs somehow, and

b) Your keys and values may be of any type: strings, integers, dummy type and, of course, [key, value] pairs themselves.

c) The canonical Map Reduce use case is counting word frequencies in a large text, but some other examples of what you can do in the Map Reduce framework include: Distributed sort, Distributed search, Graph traversal and Machine learning.

d) To implement Map Reduce, two scripts are written: The mapper script and the reducer script. The rest will be handled by the Amazon Elastic Map Reduce (EMR) framework when deployed on Amazon cloud. When we start a map/reduce workflow, the framework will split the input into segments, passing each segment to a different machine. Each machine on the cluster then runs the mapper script on the portion of data attributed to it. The map script (which you write) takes some input data, and maps it to [key, value] pairs according to your specifications.

Ex: For example (Fig. 2), if we wanted to count word frequencies in a text, we would have [word, count] be our [key, value] pairs. Emitted [key, value] pairs are then shuffled (to use the terminology in the diagram below), which basically means that pairs with the same key are grouped and passed to a single machine, which will then run the reduce script over them. The reduce script (which you also write) takes a collection of [key, value] pairs and reduces them according to the user specified reduce script. In our word count example, to get the frequency, we count the number of occurrences of a word. Thus, we would want our reduce script to simply sum the values of the collection of [key, value] pairs which have the same key.

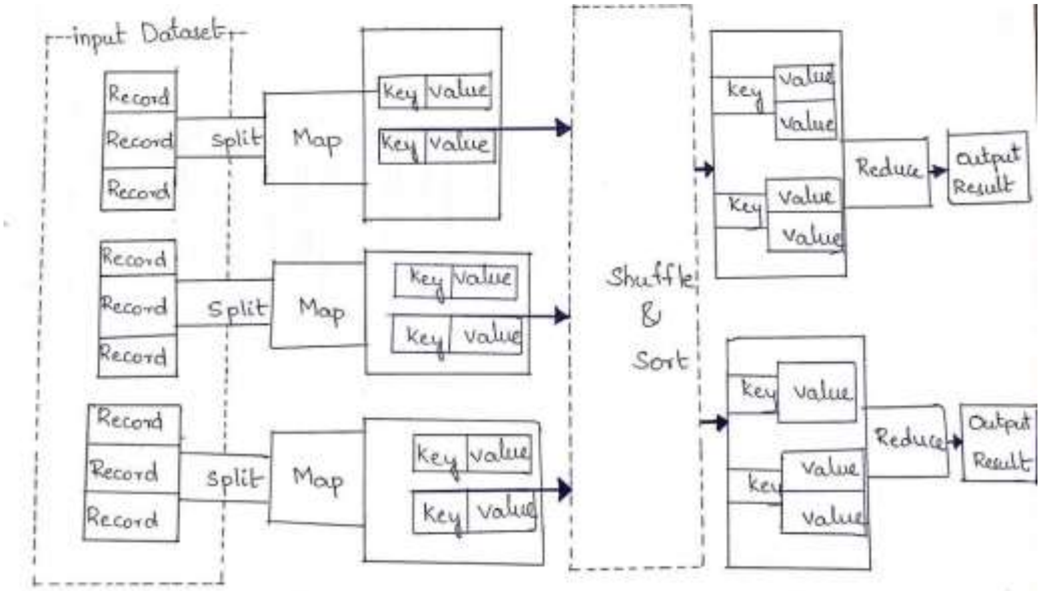


Fig.2. Map-reduce method

Map-Reduce is a parallel data processing model introduced by Google dividing computation in two functions, Map and Reduce. In Map-Reduce programming model, a Map- Reduce job consists of a map function, a reduce function. When a function is called the below steps of actions take place. Map-Reduce will first divide the data into N partitions with size varies from 16MB to 64MB. Then it will start many programs on a cluster of different machines. One of program is the master program; the others are Workers, which can execute their work assigned by master. The main steps Map-Reduce would take to process the data would be:

- Get the input data.
- Split the data into separate blocks.
- Assign the blocks to Map tasks.
- Sort the output of the Map tasks.
- Reduce the sorted data using the Reduce tasks.
- Store the output in the file system.

2. Literature Survey

- B. Hindman, quoted that on shared environments it plans to improve cluster utilization but is focused on HPC. Mesos allows scheme to achieve local information. It shares resources in a [me grained fondant.
- D. Gmach , the predicting permanent usage patterns of such assignments has been studied. This deals with the issues of performance modeling, capacity planning.
- Dean et aI., proposed the large amount of data processing assignments such as Map Reduce. In order to perform simple computations the author designed a new absorption that allows expressing the simple computations, but hides some messy details. Describes implementation of Map Reduce.
- Yahoo!, developed the capacity scheduler, a pluggable scheduler. The multi- holder cluster is designed in user friendly format while maximizing the throughput, which has been designed to run Hadoop applications. Supports Hierarchical queues.
- M. Zaharia et ai, proposed that to each holder at the cluster has equal number of shares which is allocated by Hadoop's fair scheduler. Addresses the problem of how to robustly perform speculative execution to maximize performance.

- Polo et ai, proposed that the implementation of soft deadline which supports for Map Reduce jobs / tasks can be done by Adaptive Scheduler.
- l. Wolf et aI., quoted in order to provide Service-Level Agreement (SLA) guarantees the scheduler suggested extra data to the fair scheduler. The main aim of the author is to optimize each variety of standard scheduling theory metrics.
- R.ahmed and G.karypis propose that it stores the information which increase rapidly above the usual level and as a capacity to exchange the data.

3. Problem Statement and Implementation

Let us have a discussion on Bus Data Analysis of any State Government. The problem is inappropriate allocation of buses which results in either excess in one route or deficient in another route. Some buses might be running with less passengers boarding into it. This can be resolved by the clear analysis of the tickets issued. Using the data obtained from the depot, we analyze on ticketing information and try to figure out the difference between actual requirement and the currently running buses.

Existing system

- There is no particular software implementation for analyzing the routes and timings based on the ticketing data.
- Due to lack of predictive analysis, Road Transport Corporation (RTC) is presently facing many problems.
- Thorough analysis can help RTC get out of jam.
- City buses are spread throughout the city.
- Monitoring the crowd rush has never been considered
- Analyzing and predicting the peak hour traffic load of the passengers in the buses is never done
- The frequencies at which the buses run do not match the need of the buses

Proposed system

- The stated obstacle can be eliminated by the clear analysis of the tickets issued.
- Using the data obtained from the depot, we analyze on ticketing information.
- Try to figure out the difference between actual requirement and the currently running buses.
- Firstly, we are using Hadoop platform for the existing system through map reduce framework which handles data accurately.
- Tools used: Hive for retrieving data.
- Finally the output is displayed in the form of graphs.

Implementation

Algorithm of Map Reduce for Mapping

- Read small file token.
- The key denotes the name of the file and the value denotes the content of the file.
- If the size of the file is greater than frequency, then avoid to add in the list.
- The value of default threshold is equal to 80 percent of hadoop block size and the default block size is equal to 64 mega bytes.
- It manages the list of name nodes for combining, which is done by reducer. (Output limit)

Algorithm of in Map Reduce for Reducer

- Inputs obtained from the Mapper.

- Threshold considering files are merged.

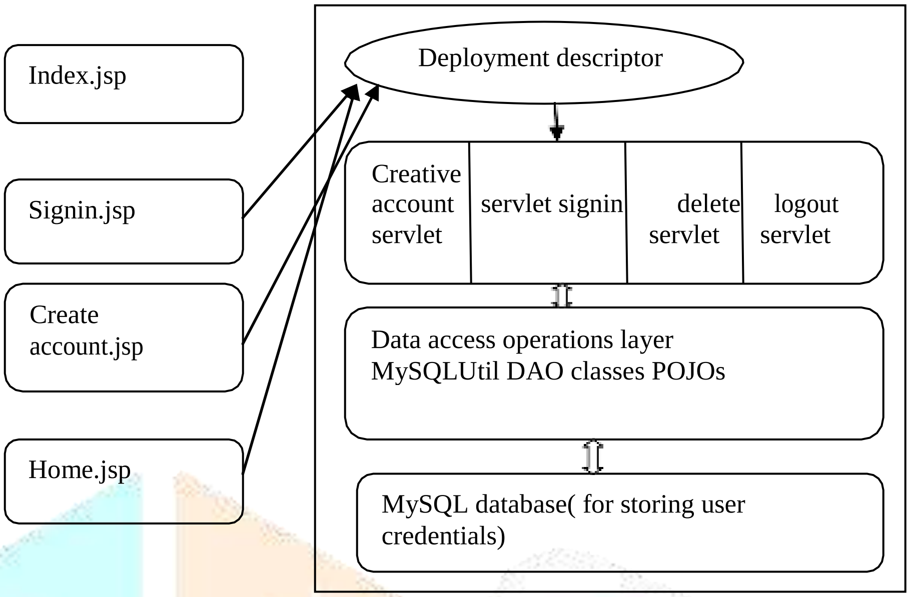


Fig.3. Flow diagram for the data access between the OS and merger map reduce components

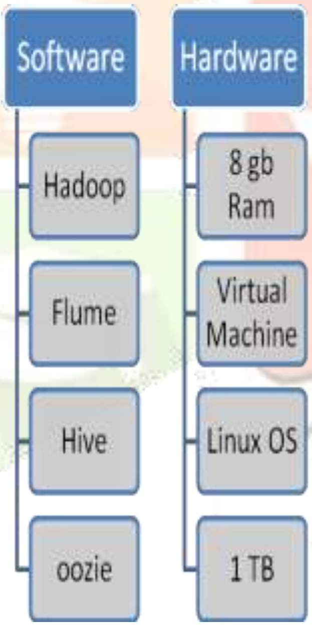
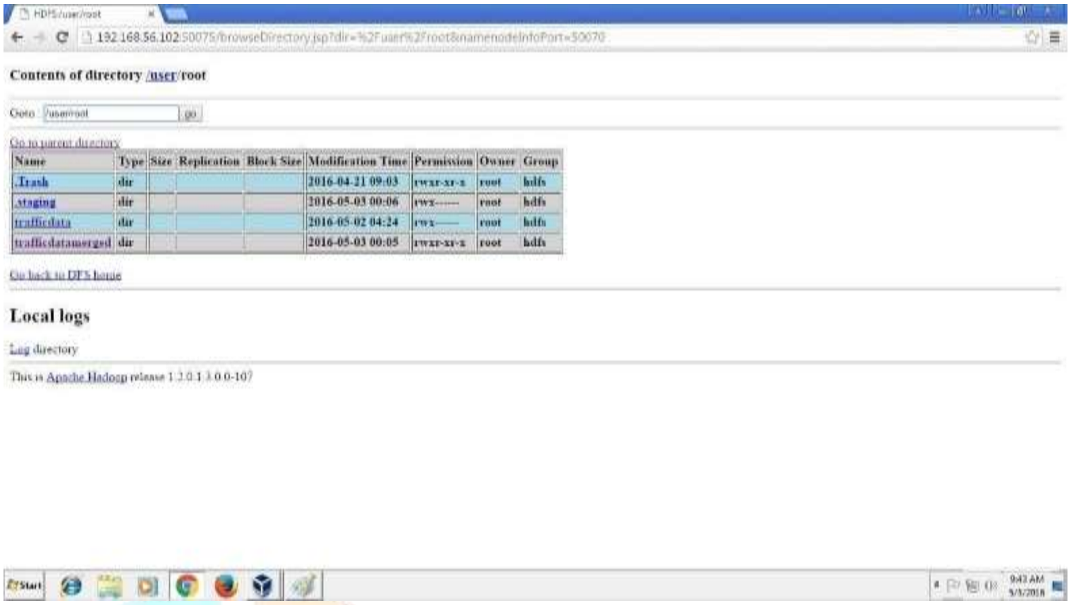


Fig.4. Software and Hardware requirements

4. Experimental Results

The traffic data merger contains the files that implement the small file problems for the documents; this is the resultant structure of the architecture. The traffic data is generated through an in built or available API known as ALEXA, this generates the website information. The information generated is in form of its rank, Google index, the revenue, initial investment and total outcome with the number of users visiting the site.


HDFS:/user/root
192.168.56.102:50075/browseDirectory.jsp?dir=%2Fuser%2Froot&namenodeInfoPort=50070

Contents of directory /user/root

Goto : /user/root go

Go to parent directory

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
.Trash	dir				2016-04-21 09:03	rwxr-xr-x	root	hdfs
.staging	dir				2016-05-03 00:06	rwx------	root	hdfs
trafficdata	dir				2016-05-02 04:24	rwx------	root	hdfs
trafficdatamerged	dir				2016-05-03 00:05	rwxr-xr-x	root	hdfs

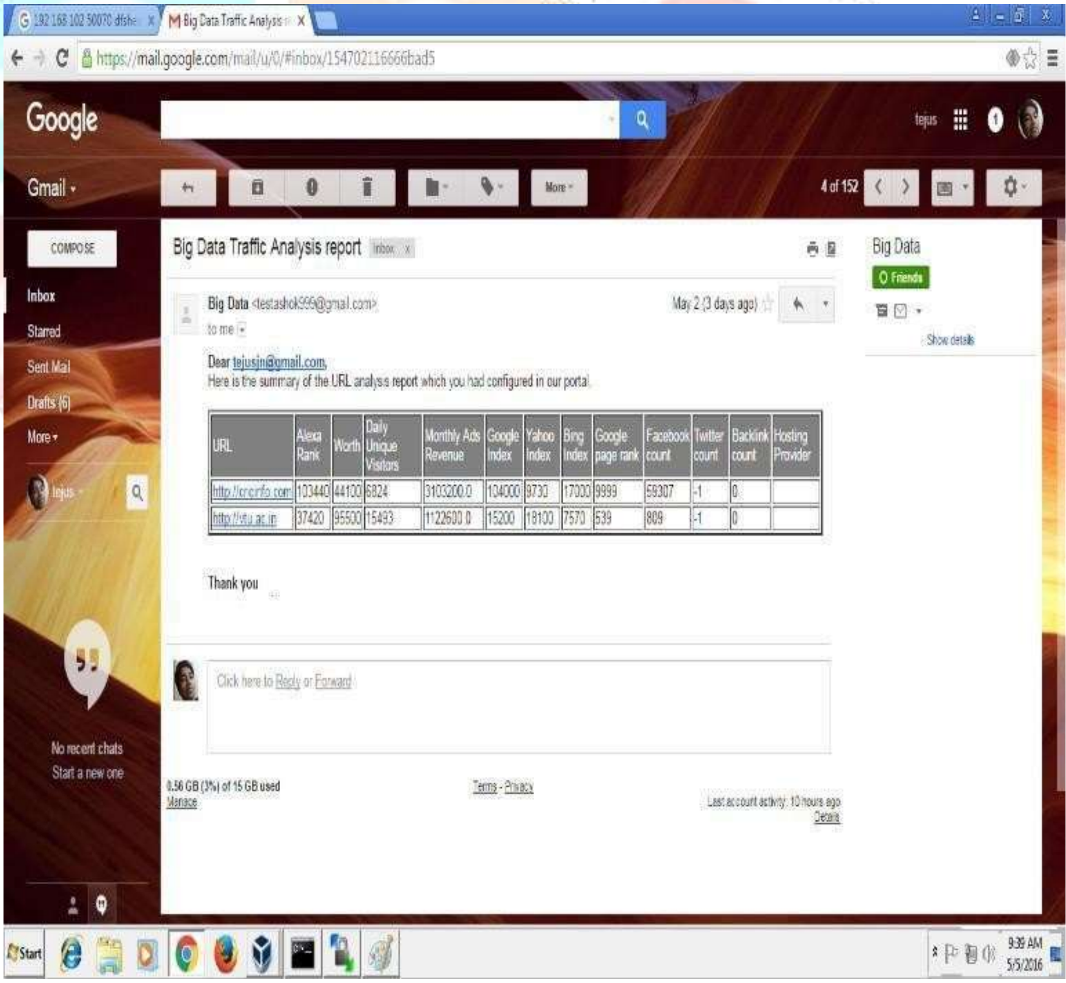
Go back to DFS home

Local logs

Log directory

This is Apache Hadoop release 1.2.0.1.3.0.0-107

Fig.4. Hadoop joblist



Big Data Traffic Analysis report Inbox x

Big Data <testashok999@gmail.com>
to me

May 2 (3 days ago)

Dear tejusjn@gmail.com,
Here is the summary of the URL analysis report which you had configured in our portal

URL	Alexa Rank	Worth	Daily Unique Visitors	Monthly Ads Revenue	Google Index	Yahoo Index	Bing Index	Google page rank	Facebook count	Twitter count	Backlink count	Hosting Provider
http://cricinfo.com	103440	44100	6824	3103200.0	104000	9730	17000	9999	59307	-1	0	
http://vtu.ac.in	37420	95500	15493	1122600.0	15200	18100	7570	539	809	-1	0	

Thank you

Click here to Reply or Forward

0.56 GB (3%) of 15 GB used
Manage

Terms - Privacy

Last account activity: 10 hours ago
Details

Fig. 5. Bigdata Analysis report

The above fig.4. and fig.5. explains the working of Hadoop and Data analysis report of Bus based on the ticket issuing at every stop.

5. Conclusion

Performance testing is a key element for production. Good performance makes it possible to ensure that the deployed solution answers the expectations in terms of response time. By comparing the experiment results on different benchmarks and by analyzing them, we found that Hive is a very strong contender and faster by processing data in-memory while Hadoop Map Reduce persists back to the disk after a map or reduce action, so Spark should outperform Hadoop Map Reduce. We plan in our future work to set up Hadoop and Hive on a bigger cluster and deploy these frameworks on the Cloud Computing to test the scalability of each platform as a Cloud service.

References

[1] Bincy P Andrews,Binu A,A Perusal On Hadoop Small File Problem, International Journal of Computer Science Engineering and Information Technology Research Vol 3, Issue 4, Oct 2013, pp.221-226.

[2] Garry Turkington Book, Hadoop Bigner's Guide , first edition, February 2013.

[3] Robert D. Schneider, Hadoop Buyer's Guide by ubuntu ,October 2013. An approach for MapReduce based Log analysis using Hadoop , Hemant Hingave Prof. Rasika Ingle,IEEE xplore paper.

[4] Tom White, Hadoop the definative guide Yahoo Press" third edition, 2012.

[5] Min Chen. Shiwen Mao, Yunhao Liu, "Big Data: a survey", Mobile Networks Appl. 19 (2) (2014) 171-209.

[6] J. Yan, X. Yang, R. Gu, C. Yuan, and Y. Huang, "Performance optimization for short MapReduce job execution in Hadoop," in 2012 Second International Conference on Cloud and Green Computing (CGC), November 1-3, pp. 688-694.

[7] K. Wang; X. Lin; W. Tang, "Predator - An experience guided configuration optimizer for Hadoop MapReduce," Cloud Computing Technology and Science (CloudCom), 2012 IEEE 4th International Conference on, vol., no. , pp.419,426, 3-6 Dec. 2012.

[8] S. Huang, 1. Huang, 1. Dai, T. Xie and B. Huang. 'The HiBench Benchmark Suite: Characterization of the MapReduce- Based Data Analysis. " Lecture Notes in Business Information Processing, 2011, Volume 74, Part 3, 209-228. 2011.

[9] M. Zaharia, M. Chowdhury, Michael 1. FrankIin, S. Shenker, 1. Stoica "Spark: Cluster Computing with Working Sets" HotCloud'IO Proceedings of the 2nd USENIX conference on Hot topics in cloud computing Pages 10-10, 20 I O.

[10] M. Zaharia, M. Chowdhury, T. Das, A. Dave, 1. Ma, M. McCauley, Michael 1. Franklin, S. Shenker, T. Stoica "Resilient Distributed Datasets: A Fault-Tolerant Abstraction for ln-Memory Cluster Computing" Technical Report UCB/EECS-2011-82, EECS Department, UC Berkeley, 2011.

ACKNOWLEDGMENT

SRIKANTH BETHU has completed M.Tech in Computer Science and Engineering in 2011 at Osmania Univeristy, Hyderabad, India. He was currently working as Assistant Professor in Department of Computer Science and Engineering at Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad. He was very fond of Data Mining concepts. He has published several research papers in the areas of Data Mining and Data Engineering. Attended several National and International conferences.

Dr. B. Sankarababu was a doctorate from Aacharya Nagarjuna University, Guntur, Andhrapradesh, India, and completed Ph.D in 2016. Currently working as a Professor in Department of Computer Science and Engineering at Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad .He has done his research on Data science and Data algorithms. He has published several research papers on Data mining concepts.