MJRET
Open Access
Multidisciplinary Journal of Research in Engineering and Technology
Journal homepage: www.mjret.in
ISSN:2348 - 6953

Prajakta Kulkarni
Dept. of Computer Department
K J College of Engineering &
Management Research
Pune,India

Pratibha Bodkhe
Dept. of Computer Department
K J College of Engineering &
Management Research
Pune,India

Kalyani Hole
Dept. of Computer Department
K J College of Engineering &
Management Research
Pune,India

Ashwini Kondalkar
Dept. of Computer Department
K J College of Engineering &
Management Research
Pune,India

Social Network Data Extraction Analysis

Abstract

Now-a-days the use of internet is increased; the means of the interaction of people has also been increased. Because of this there are large amount of data have been aggregated very fast which are related to social relationship. The data in web is not organized there are large amount of data have been aggregated very fast which are related to social relationship. Hence it has become very difficult for researches to collect data, to pre-process data, to perform different social networks analyses as well as to visualize the social networks. It requires large amount of time and resources. The proposed system will give solution to the recent problems to store and process large amount of data. The proposed system is not only data warehousing system but also data processing engine. Hadoop and Hama processes data separately with Map-Reduce and BSP algorithms respectively. Performance is checked by generating graph of processing time of Hadoop and Hama.

Keywords: *Hadoop, Hama, Map-Reduce, BSP (Bulk Synchronous Parallel)*

1. INTRODUCTION

With the accelerated growth of internet usage, intercommunication of people has also been increased. In this environment, abundant data is accumulated very fast which are related to social relationship. The data has three main properties: Velocity, Volume, and Variety.

a) Velocity: Data is growing very fast as number of internet users are increasing day by day.
b) Volume: Data is of large volume, which is very difficult to store and process. Data is measured in terabytes (10^12) and petabytes (10^15).
c) Variety: Data is in various formats like text, audio, video, images etc.

The relational database is not suitable for storing this type of data. Hence the proposed system works on HDFS (Hadoop Distributed File System) which is 'No-SQL' database.Data is processed by two different frameworks Hadoop and Hama. Hadoop uses Map-Reduce algorithm, whereas Hama uses BSP (Bulk Synchronous Parallel) algorithm.

2. BRIEF DISCRIPTION

The development and popularity of online social networks in recent years has changed the Internet environs leading to a more collaborative environment. Nowadays, hundreds of millions of Internet users participate in social networks, form communities, produce and consume media content in radical ways. Such network provides space for sharing multimedia information among neighbors in social graph. Many modern enterprises are crawling data at the most detailed level and creating data storehouses ranging from terabytes to petabytes. The proposed social network analysis system is very useful for analyzing social network data. Furthermore, the analysis results can be applied to many useful areas, such as marketing, the detection of crime and terrorists, etc.

2.1 Cloud Computing Techniques

Social network platforms have rapidly changed the way that people communicate and interact. Platforms like Facebook, Twitter, LinkedIn and Google+ enable the establishment of and participation in digital communities; and the representation, documentation and exploration of social interactions as well as relationships. Therefore, electronic relationships are quickly becoming associated with their real world counterparts. Social networks have always served as a vital means for the sharing and exchange of multimedia information, improving the understanding of relationships, improving communication between globally dispersed individuals, and more recently measuring scientific impact. Social networks can facilitate the construction of Social Clouds: the provisioning of Cloud infrastructure through social network constructs.

2.2 Hadoop

Hadoop is a set of tools which is used for storing and processing the big data on web. It is a distributed model and Linux based set of tools. Hadoop has master slave relationship.

2.2.1 Structure of Hadoop

Slaves contain two nodes: Task Tracker, Data Node.

1. Task Tracker: Job of task tracker is to process small piece of task that has been given to this particular node.

2. Data Node: Job of this data node is to manage the data that has been given to particular node.

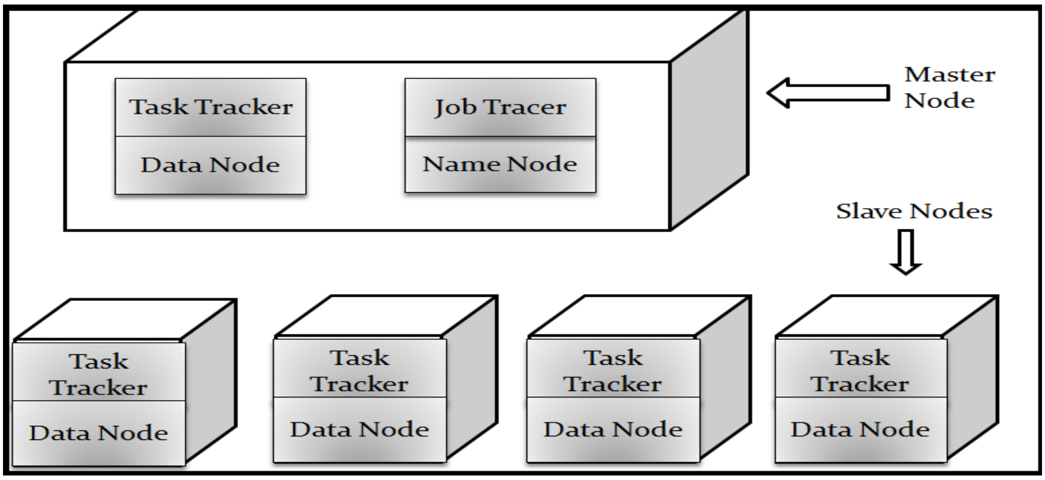

Fig. 1: Stucture of Hadoop

Master has two additional nodes: Name Node, Job tracker.

1. **Name Node:** Name node is responsible for keeping the index of which data is residing on which data node. Hence, when application contacts the name node, it tells the application to go to particular node to get required data.

2. **Job Tracker:** Role of job tracker running on master is to break the bigger task into smaller modules and send each module of computation to task tracker. Task trackers will perform small tasks and send result back to job tracker. Job tracker will combine all results and send back to the application.

The Map-Reduce algorithm which is proposed by Google is a very famous example of public cloud computing, it can be used for computer programs that need to process and generate large amount of data. The implementation of Map-Reduce is Hadoop framework. Programs written in Map-Reduce are automatically parallelized and executed on a large cluster hence strength in data locality, fault tolerant and parallel process.

2.2.2 Limitation of Map-Reduce

Map-Reduce have its weakness in mathematical graph process, as it has to deliver the state of the graph from one step to another. This causes a low-efficiency issue when dealing the processing of graphic algorithm.

2.2.3 Hama

According to graph theory, social network is a directed graph composed by objects and their relationship. Hence in social networking graph processing has become very important task. Hama is a distributed computing framework based on BSP computing techniques for massive scientific computations. Massive scientific computations are mathematical computations of matrix, graph, and networks. BSP is very good option to speed up iteration loops during the iterative process which requires several passes of messages before the final processed output is available, such as finding the shortest path, and so on. The implementations of BSP and Map-Reduce involved in the project are Apache's Hadoop and Apache’s Hama. The purpose of our project is to compare the crawler performance executed on these two major platforms. From the analysis result it shows Map-Reduce have better performance when dealing with 100 URLs. However, with the increasing of the number of URLs, Hama BSP needs less time to accomplish the designed tasks and the performance is much better than Map-Reduce.

Hama also provides an easy-to-program, as well as a flexible programming model, as compared with traditional models of Message Passing, and is also compatible with any distributed storage.

3. IMPLIMENTATION

Basically our proposed system is a comparative system which compares performance of Hadoop and Hama. The system will be implemented on java. So for the implementation purpose five modules are taken into consideration and they are as follows:

i. Module 1(Crawl the raw data from web):

The crawler is the software designed in java which crawls raw data from web.

- Input for module 1: URL (Uniform resource Locator)of any social networking website.
- Processing: It crawls all the raw data i.e. XML, HTML tags with the actual data contents. It detects links associated with the crawling page and redirect to that link. It crawls that redirected links too. Hence we get all the raw data in one folder.
- Output of module 1: Raw files stored in a folder.

ii. Module 2 (Convert raw data into fair data) :

- Input for module 2: Raw data i.e. XML, HTML tags
- Processing: Parsing of this raw data. Parser is used for converting raw data to fair data. This parsing program is designed in java. It identifies the last index of opening tag and the first index of closing tag. And hence it fetches the fair data between those last and first indices. And this fair data is given as input to the third module of this project.
- Output: Fair data i.e. data without XML, HTML tags.

iii. Module 3(a) (Processing of raw data with Hadoop):

- Data Storage: Hadoop stores all the fair data files in HDFS format by dividing the big file into small chunks. Small chunks are stored and replicated on data nodes.
- Data Processing: Hadoop processes the stored data by using MapReduce algorithm. Storing of data on data nodes is called as mapping of data. On the basis of users demand combing this divided data into one output file is called as reducing of data.
- MapReduce works on the principle 'serial in parallel'. All the data nodes work parallel but processing in each data node is done serially. Due to the 'serial in parallel' principle it works slowly.

iv. Module 3(b) (Processing of raw data with Hama)

- Data Storage: In Hama data storage is same as that of Hadoop by using HDFS.
- Data Processing: Algorithm used in Hama for data processing is 'Bulk Synchronous Parallel' (BSP).This is the inbuilt algorithm of Hama introduced by 'Apache'. It works on 'Parallel in Parallel' principle. All data nodes work parallel and each individual processing in data node is also done in parallel. Hence it works faster than MapReduce.

v. Module 4 (Generation of analytical graph):

Graph generation program is designed in java. It considers processing time of both Hadoop and Hama for the generation of graph.

- Input for module 4: Processing time of both the frameworks Hadoop and Hama.
- Output of Module 4: Graphwhich shows the comparison of processing time of these two frameworks.

4. PROPOSED SYSTEM ARCHITECTURE [10]

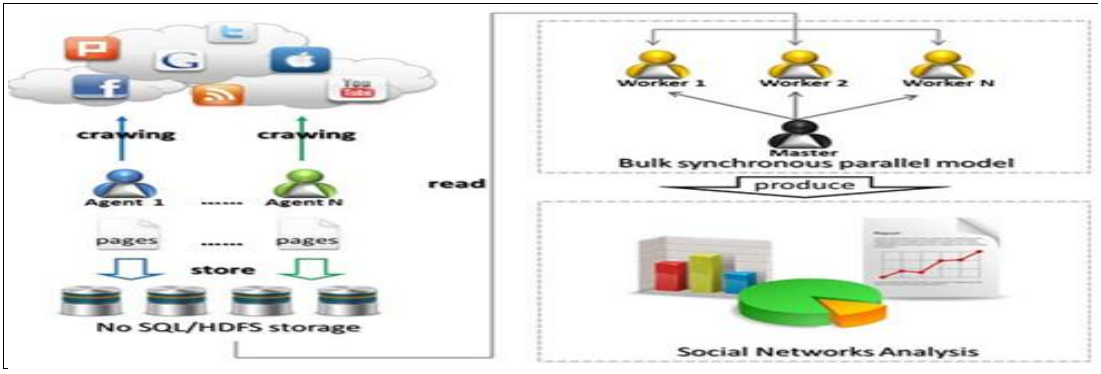

Fig. 2: Proposed system Architecture

According to different tasks in the system architecture, the components in the system can be divided into three different parts, which are front-end data collection components, intermediate system analysis components and analysis resultsproducing components.

i. **Front-End Data Collection Components:**
The collected data by crawling agents will then be stored in distributedenvironment. The advantage to use HDFS is its reliability. In distributed environment, a system frequently encounters hardware failure.

ii. **Intermediate System Analysis Components:**
After the processing of the front-end data collection components, the data that stored in the system are raw data. Then the system will provide data after different level of processing to users according to their requirements. BSP will be used as the technique to play as role to process social data according to different algorithms, which is based on the structure of Master/Worker. In the system, Master has to assign works to workers and to acquire the processing results. Workers will perform the works according to the assignment form the Master. The assignments could be data pre-processing, social networks analysis or visualization.

iii. **Analysis result producing components**
After performing appropriate algorithms by the intermediate system analysis components, the analysis result producing components will play as a role to produce results to fit the analysis requirements of users.

5. TECHNOLOGIES USED IN THE PROPOSED SYSTEM

i. Crawler
ii. Hadoop
iv. Hama
v. HDFS (Hadoop Distributed File System)

6. CONCLUSION

This application is developed using cloud techniques such as Hadoop and Hama and generation report will be shown using data warehouse. The main idea of the system is to deal with the difficulties of recent social networks data extraction. The system is designed to store social data in the data warehouse and to perform social networks analysis and other processing.

7. FUTURE SCOPE

It is not possible to develop a system that makes all the requirements of the user. User requirements keep changing as the system is being used. Some of the future enhancements that can be done to this system are: the system can be enhanced to increase the look and feel of the application. The web crawler technique for dynamic update of live score and stock update can be used.

8. ACKNOWLEGEMENT

First and foremost, we would like to thank our guide Prof. Mokashi, for his guidance and support. We will forever remain grateful for the constant support and guidance extended by guide, for completion of project report. Through our many discussions, he helped us to form and solidify ideas. The valuable discussions we had with him, the penetrating questions he has put and the constant motivation, has all led to the development of this project. We are also thankful to our family members for their encouragement and support in this project work.

REFERENCES

[1]. Ilango Sriram Ali Khajeh-Hosseini Cloud Computing Co-laboratory School of Computer Science University of St Andrews St Andrews, UK" Research Agenda in Cloud Technologies'. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[2]. Valiant, L. G. (1990) A bridging model for parallel computation. Communications of the ACM, 33(8), pp.103–111.

[3]. H. Alani, S. Kim, D. Millard, M. Weal, W. Hall, P.Lewis, and N. Shadbolt, "Automatic Ontology-Based Knowledge Extraction from Web Documents", IEEE Intelligent Systems, 2003, 18(1):14-21.

[4]. Bill Howe Magdalena Balazinska Michael D. Ernst HaLoop: Efficient Iterative Data Processing on Large Clusters.

[5]. Heer, J., and Boyd, D. (2005), "Vizster: Visualizing Online Social Network", In Proceedings of 2005 IEEE Symposium on Information Visualization, October 23-25, 2005, Minneapolis,MN USA, pp.32-39.

[6]. Hamasaki, M., Matsuo, Y., Ishida, K., Hope, T., Nishimura, T. and Takeda, H. (2006) "An Integrated Method for Social Network Extraction".

[7]. Data Warehouse, Data Mart, Data Mining, and Decision Support Resources, http://infogoal.com/dmc/dmcdwh.htm, downloaded from the web, February 8, 2007.

[8]. Grossman, R. L. (2009). The case for cloud computing. IT Professional, 11(2), pp.23–27.

[9]. Birman, K., Chockler, G., and Renesse, R. V. (2009) Toward a Cloud Computing Research Agenda. SIGACT News, 40(2), pp

[10]. Dean, J., and Ghemawat, S. (2009) Mapreduce: Simplified dataprocessing on large clusters. pp. 137–150.

[11]. I-Hsien Ting, Chia-Hung Lin, and Chen-Shu Wang "Constructing A Cloud Computing Based Social Networks Data Warehousing and Analyzing System" ,2011 International Conference on Advances in Social Networks Analysis and Mining.

[12]. BaodongJia , Tomasz WiktorWlodarczyk, ChunmingRong," Performance Considerations of Data Acquisition in Hadoop System", 2nd IEEE International Conference on Cloud Computing Technology and Science.

[13]. Steps of installation of Hadoop [Online]. Available: http://v-lad.org/Tutorials/Hadoop/03%20-%20Prerequistes.html

[14]. P. B. Gibbons, Y. Matias, and V. Ramachandran A Bridging Model for Parallel Computation.