Volume 3, Issue 3, March 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Performance of Data Warehousing Based Social Network Analysis in Cloud Computing Using Crawler Algorithm

J.Gobinath[1]
Asst Professor, IT Dept.
Rajiv Gandhi college of Engineering
Sriperumbudur, Chennai, India

D.Revathi[2]
Asst Professor, MCA Dept.
Vel Tech High Tech DR Rangarajan DR Sakunthala
Engineering College, Avadi, Chennai, India

Abstract: Cloud computing has been recognized as extremely time-consuming tasks as well as large storage space and it is necessary for order to store the social data, especially to deal with the data in the World Wide Web. We proposed a data warehousing and analyzing system which is based on the concept of cloud computing. The system has also been implemented and evaluated under the proposed environment with different cloud computing approaches.

Keywords: Social Networks Analysis, Cloud Computing, Map Reduce, BSP

I. Introduction

Social Networks Analysis (SNA) [3] is a very suitable and important research method for analyzing the social data structure to understand the characteristics of a network. Furthermore, the analysis results can be applied to many useful areas, such as marketing, the detection of crime and terrorists, recently, the web data are the main target in many researches about using information technology in the area of social networks analysis. However, the data in the web are usually very messy and noisy, as well as the size of the data is always extremely large.

Thus, researchers usually need to spend large amount of resources and time to collect data, to pre-process data, to perform different means of social networks analyses (very high dimensional matrix computation and the processing of networks graph) as well as to visualize the social networks. In the other hand, large storage space is also necessary to store the large amount of social data. Cloud computing is the current buzzword in the air [2]. The name "Cloud Computing "was inspired by the cloud symbol that's often used to represent the Internet in flow charts and diagrams. In short cloud computing means using the Internet for all computer needs. Cloud computing is a natural evolution of the widespread adoption of virtualization, service-oriented architecture, autonomic and utility computing [1]. Cloud computing is therefore a very suitable solution for SNA and which can also be used to enhance the efficiency and performance of related researches.

To overcome the recent problems and difficulties in SNA by applying the techniques. Therefore, a system has been proposed which is based on the concept of cloud computing. The system not only a data warehousing system but also a SNA engine can be used to perform different SNA analyses with high performance, the system will be implemented as well as the performance will be tested by using different techniques of cloud computing.

II. Literature Review

A. Social Network Analysis

Social network analysis is developed to understand the relationship between "actors", and the term actor can be a person, an organization, an event or an object. In a social network, each actor is presented as a node and each pair of nodes can be connected by lines to show the relationships. The social network structure graph is a graph that formed by those lines and nodes, and social network analysis is therefore a methodology the relationships. The social network[3][4] structure graph is a graph that formed by those lines and nodes, and social network analysis is therefore a methodology that used to understand the graph and the relationships and actors in the social network .

There are three important elements that included in a social network: actors, ties, and relationships. Actors are the essential elements in the social network to define the people, events or objects. Ties are used to construct the relationship between actors by using a mean of path to establish the relationship directly or indirectly. Ties can also be divided into strong tie and weak tie according to the strength of the relationships; they are also useful for discovering the subgroups of the social network. The most important measurements of SNA include network size, diameter, density, centrality and structure holes. Size is a measurement to measure the amount of nodes or links in a network, and the measurement of diameter is to measure the amount of nodes between two nodes in a network. Density is used to calculate the closeness of a network. These measurements are common used in many social network related researches and will be used in this paper as well.

B. Cloud Computing

Cloud computing is now a very hot topic in the field of Internet applications and researches [1][2]. It has been defined as an Internet service which provides extensible services dynamically over the Internet. According to the provided services, cloud computing can be categorized as **SaaS (Software as a Service), PaaS (Platform as a service) and IaaS (Infrastructure as a Service)**[7]. Proposed another cloud computing categorization which categorizes cloud computing as "pubic cloud" and "private cloud" with a consideration of the integration of hardware Equipment and software services .

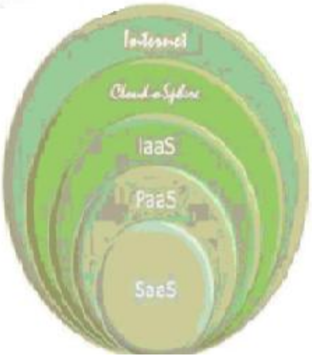

Fig: 1: The architecture of cloud computing

The Map Reduce technique which is proposed by Google is a very famous instance of public cloud computing, it can be used for computer programs that need to process and generate large amount of data. For example, Map Reduce has been used to generate the index of Google and it also has the strength in data locality, fault tolerant and parallel process. This strength is helpful for Map Reduce to enhance the performance. However, Map Reduce has its weakness in mathematical graph process. In social networks analysis researches, the computation and processing of graph is essential. Therefore, how to deal with this problem is a very important issue to implement cloud computing for social networks analysis.

The problem that discussed above about Map Reduce can be solved by applying a BSP (Bulk Synchronous Parallel) model to perform "super step" repeatedly and using the concept of parallel processing . BSP to solve the problem of Map Reduce. The proposed system architecture try to apply and implement a BSP based cloud computing technique for the data processing requirement of social networks analysis.

III. Proposed System Architecture And Analysis

According to the research background and literature review, we proposed system architecture for social data warehousing and analyzing. 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 results producing components.*

1) Front-end data collection components

In the system, well-designed crawling programs are included as one of the front-end data collection components. The data front-end data collection system collects social data from different social networking websites in different locations. First will collect data from some famous social Networking.

Websites,suchasFacebook[5] (www.facebook.com), plurk (www.plurk.com) and wretch (www.wretch.cc) etc. The collected data by crawling agents will then be stored in distributed environment.

2) Intermediate system analysis components

After collecting front and data collection and then data are 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 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.

3) Analysis result producing components

After performing appropriate algorithms bythe Intermediate system analysis components, the analysis result producing components will play as a role to produce results to fit the analysis requirements of users. The users can either send requests to the provided web based interface or API (Application Program Interface). The main concern of the analysis result producing components is the ability of cross- platform and the web based interface and API will be able to communicate between different platforms.

4) System Design

In the analysis result producing components, therefore designed a web based interface to provide the interface for user to interact, send queries, and acquire the analysis results. The system is designed based on the idea of "Interoperability" which allows the users to manipulate data and perform analysis by using different operating systems or platforms.

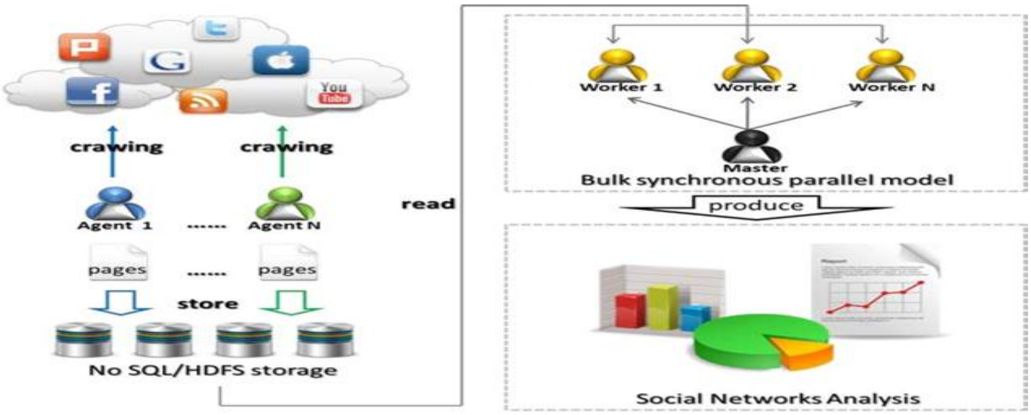

Fig: 2: The architecture of the proposed system

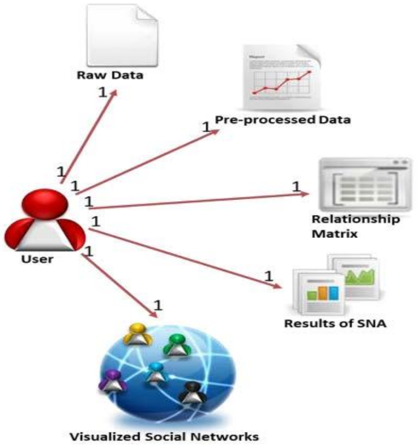

Fig: 3: The use case diagram of the proposed system

It shows the use case diagram of the proposed system. By using the system, the users can extract raw data in the social data warehouse or pre-processed data after data cleaning, or relationship matrix which transformed from the raw data. Users can also send requests to ask the cloud computing platform to perform social networks analysis and to show the analysis results. The social network analysis measurements [5][6] that included in this system include: network diameter, degree centrality, closeness centrality, between centrality. In additional, the users can also acquire the visualized network graph of appropriate social networks.

IV. Experiments Design And Performance Evaluation

To test the proposed system as well as to compare the performance when adopting different cloud computing in the system, a serial of experiments has been designed. In this section, we will first introduce the design of the experiments and then present the results of the performance evaluation for discussion.

1. The Design of the Experiments

In order to test the system performance,[8] we have designed a serial of experiments. The front-end data collection components have been chosen as the test target, as it could be the most time-consuming tasks in the entire system. Therefore, [8][9]the experiment designed is by running crawling process on two different platforms. One bases on Map Reduce, the other Bulk Synchronous Parallel (BSP). The former creates a simple programming model for developers easily manipulating large data set, the latter consists of simple steps addressing the distributed processing real life graphs. The implementations of BSP and Map Reduce involved in the experiment are Apache's Hadoop Map Reduce and Apache Hama.[9] The purpose of the experiment is to compare the crawler performance executed on these two major platforms. The procedure for such testing starts with:

1st, the client submits a job which encapsulates crawling algorithm for Map Reduce and BSP platform respectively.
2nd, the crawler fetches the designated pages according to the given URL list.

3rd, after finishing all pages required, the job begins to save content into Hadoop Distributed File System, inspired by the Google file system.

The pseudo code of the algorithm with Map Reduce

Input : A set of framework names F ={Map Reduce, Hama}. An input url list path points to a file, which contains a list of websites L = <l_1, l_2... l_n> such that I=string. An output path p points to a file, which will store web pages to be crawled, p=string.

Output : A file which contains crawled web pages P = <p_1, p_2, ..., p_n>

```
if f = Map Reduce
   then map phase
        for every URL <- $l_1$ to $l_n$
        do $p_n$ <- crawl (URL)
        collect (URL, $p_n$)
        reduce phase
        for every URL <- $l_1$ to $l_n$
        do save (URL, $p_n$) to p
else if f = Hama
    then read url-list
        for every URL <- $l_1$ to $l_n$
        do $p_n$ <- crawl (URL)
        collect (URL, $p_n$)
        bsp sync ()
     while URL <- $l_1$ to $l_n$
        do save (URL, $p_n$) to p
output p
```

It is the performance comparison of implementing BSP and Map Reduce in the system

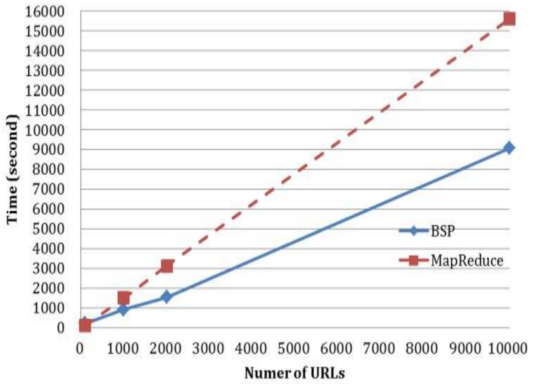

Fig: 4: Performance Evaluation

X-axis shows the Number of URLs and the y-axis shows the execution time by second. From the analysis result it shows Map Reduce has 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.

V. Conclusion and Future Research

A system architecture has been proposed which is based on the concept of cloud computing. The main idea of the system is to deal with the difficulties of recent social networks analysis researches. The system is designed to store social data in the data warehouse and to perform social networks analysis and other processing. To evaluate the performance of the system and by this to decide the main cloud computing technique that used in the system, it has designed experiments to perform a crawling algorithm under specific environment. The evaluation results show that BSP have better performance than Man Reduce.

REFERENCES

[1] Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Andy Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., and Zaharia, M. A view of cloud computing. Communications of ACM, 53(4):50–58, 2010.

[2] Birman, K., Chockler, G., and Renesse, R. V. (2009) Toward a Cloud Computing Research Agenda. SIGACT News, 40(2), pp.68–80.

[3] Boyd, D., and Ellison N. B. (2007) Social network sites: Definition, history, and scholarship. Journal

of Computer- Mediated Communication, 13(1), 2007.

[4] Fu F., Chen X., Liu L., and Wang, L. (2007), " Social Dilemmas in an Online Social Network: The Structure and Evolution of Cooperation", Physics Letters A, Vol 371, 2007, pp.58

[5] God bole, N., Srinivasaiah, M., Skiena, S. (2007), "Large-Scale Sentiment Analysis for News and Blogs", in Proceedings of ICWSM 2007, Boulder, Colorado, USA.

[6] Goodreau, S. M. (2007), "Advances in Exponential Random Graph (p*) Models Applied to A Large Social Network", Social Network, Vol. 29, 2007, pp.231-248.

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

[8] Jin, Y. Z., Matsuo, Y., and Ishizuka, M. (2007), "Extracting Social Networks among Various Entities on the Web", In Proceedings of the Fourth European Semantic Web Conference, 2007.

[9] Ting, I. H. (2008) "Web Mining Techniques for On-line Social Networks Analysis" In Proceedings of the 5th International Conference on Service Systems and Service Management, Melbourne, Australia, 30 June-2 July 2008, pp. 696-700.

[10] Xue, W., Shi, J. W., and Yang, B... (2010) X-RIME: Cloud- Based Large Scale Social Network Analysis. In Proceedings of the 7th International Conference on Service Computing (IEEE SCC), July 5-10, 2010, Miami, Florida, USA