Constructing A Cloud Computing Based Social Networks Data Warehousing and Analyzing System

I-Hsien Ting
Department of Information Management
National University of Kaohsiung
Kaohsiung City, Taiwan
iting@nuk.edu.tw

Chia-Hung Lin
Department of Information Management
National University of Kaohsiung
Kaohsiung City, Taiwan
ch1501@nuk.edu.tw

Chen-Shu Wang
Graduate Institute of Information and Logistics Management
National Taipei University of Technology
Taipei City, Taiwan
wangcs@ntut.edu.tw

Abstract—**The research area of Social networks analysis has been recognized as extremely time-consuming tasks as well as large storage space is always necessary in order to store the social data, especially to deal with the data in the World Wide Web. Therefore, how to design an architecture and environment for performing social networks analysis is very essential. In this paper, 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, MapReduce, BSP, Hadoop, Hama

I. INTRODUCTION

With the rapid growth of Internet and the concept of Web 2.0, the means of the interaction of people has also been increased. Under this background, there are large amount of data have been aggregated very fast which are related to social relationship. However, these valuable data have not been used and analyzed very well in traditional researches. Social Networks Analysis (SNA) is a very suitable and important research method for analyzing the social data structure to understand the characteristics of a network [30]. Furthermore, the analysis results can be applied to many useful areas, such as marketing, the detection of crime and terrorists, etc [9].

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 has now becoming a very popular term in the area of Internet research in recent years. With the rise of the concept of cloud computing, more and more researchers are focusing on how to apply suitable technology to deal with the time-consuming computing problem [3]. 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.

In this paper, we would like 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. In this paper, the system will be implemented as well as the performance will be tested by using different techniques of cloud computing. The analysis results might be helpful for researches who what to design similar system as a reference

The structure of this paper is organized as below: In section 1, the background and introduction will be introduced. Some related literatures of social network analysis and cloud computing will be reviewed in section 2. A system architecture of the proposed social data warehousing and analyzing system will be proposed in section 3 as well as the introduction of the components in the system. In section 4, we will design an experiment and focus on applying different techniques of cloud computing. In section 6, this paper will be concluded with the suggestions for future research.

II. LITERATURE REVIEW

A. Social Network Analysis

The research methodology of 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 [35]. In a social network, each actor is presented as a node and each pair of nodes can be connected by lines to show

978-0-7695-4375-8/11 $26.00 © 2011 IEEE
DOI 10.1109/ASONAM.2011.25

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 that used to understand the graph and the relationships and actors in the social network [26][31][34].

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. Relationships are used to illustrate the interactions and relationship between two actors. Furthermore, different relationships may cause the network to reflect different characteristics [12][13][17][18].

The most important measurements of SNA include network size, diameter, density, centrality and structure holes [4]. 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 [5][10][14][24]. These measurements are common used in many social network related researches and will be used in this paper as well.

Traditionally, researches about SNA are mainly focus on small group of actors and are process manually in most cases. [16] However, with the rapid growth of Internet and web techniques, more and more data have been collected and it has become a hard task to process these data by only the mean of manually [25]. Therefore, the scholars of information technology and computer science are starting to devote related researches to deal with these research issues [28][29]. Currently, the researches of computer science in SNA can be divided into four main topics, including *social networks construction*, *social networks extraction*, *social networks analysis* and *visualization* [8][20][21][22][23].

B. Cloud Computing

Cloud computing is now a very hot topic in the field of Internet applications and researches. It has been defined as an Internet service which provides extensible services dynamically over the Internet [1][3]. 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). Armbrust et al. 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 [2].

The MapReduce 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 [19][33]. For example, MapReduce has been used to generate the index of Google and it also has the strength in data locality, fault tolerant and parallel process. These strength is helpful for MapReduce to enhance the performance. However, MapReduce has its weakness in mathematical graph process [37]. Essentially, MapReduce is a function and therefore it has to deliver the state of the graph from on step to another. This causes a low-efficiency issue when dealing the processing of graphic algorithm. 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 [27] [36].

The problem that discussed above about MapReduce can be solved by applying a so-called BSP (Bulk Synchronous Parallel) model to perform "superstep" repeatedly and using the concept of parallel processing [32]. Pregel is developed by Google shows the problem of the processing of graphic algorithm can be improved. Pregel is a technique based on the concept of BSP to solve the problem of MapReduce. However, it isn't an open source algorithm and can't be adopted and modified for public use. Thus, the community of Apache developer is now running a BSP based project which is called "Hama". Hama and Hadoop MapReduce both use the architecture of Mast/Worker, however, it faces a problem about availability.

The proposed system architecture in this paper will therefore try to apply and implement a BSP based cloud computing technique for the data processing requirement of social networks analysis.

III. Proposed System Architecture

In section, the overview of the system architecture will firstly be introduced as well as the system components. Then, we will use a use case diagram to show the functions of the system.

B. The system architecutre and system components

According to the research background and literature review, we therefore proposed a system architecture for social data warehousing and analyzing. The proposed system architecture is shown in figure and the detail of each component will be discussed in detail below.

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. In this paper, we firstly will collect data from some famous social networking websites, such as Facebook (www.facebook.om), Plurk (www.plurk.com) and Wretch (www.wretch.cc), etc. The collected data by crawling agents will then be stored in distributed environment [7].

The advantage to use HDFS is its reliability [6]. In distributed environment, a system frequently encounters hardware failure. In order to cope with this situation, the strategy that HDFS employees is by breaking incoming files into `blocks,' storing them across machines in the cluster environment. This prevents failure in acquiring blocks from one data node to affect the integrity of the whole file; thus the robustness can be achieved.

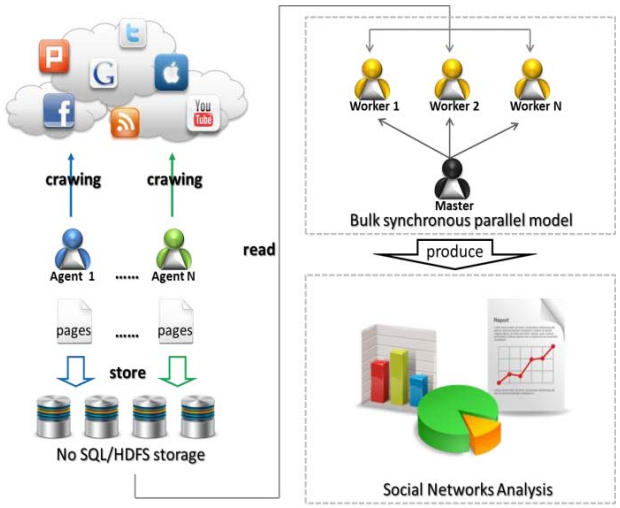

Figure 1. The architecture of the proposed system

2) 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.

3) 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. 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.

C. System design

As described in the analysis result producing components, we 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.

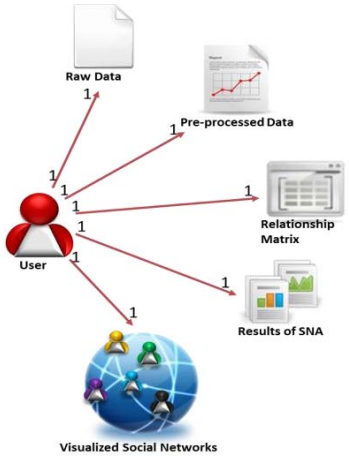

Figure 2. The use case diagram of the proposed system

Figure 2 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 that included in this system include: network diameter, degree centrality, closeness centrality, betweeness centrality. In additional, the users can also acquire the visualized network graph of appropriate social networks.

IV. Experiments Design and Performance Evaluation

In order 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.

A. The design of the experiements

In order to test the system performance, 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, the experiment designed is by running crawling process on two different platforms. One bases on MapReduce [11], the other Bulk Synchronous Parallel (BSP)[32].

The former creates a simple programming model for developers easily manipulating large dataset, the latter consists of simple steps addressing the distributed processing real-life graphs [26]. The implementations of BSP and MapReduce involved in the experiment are Apache's Hadoop MapReduce and Apache Hama. 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 MapReduce 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[15].

The pseudo code of the algorithm with MapReduce and Hama is shown in figure 3

```
Input  : A set of framework names F = {MapReduce, Hama}. An input url
list path points to a file, which contains a list of websites L = <l1, l2, ..., ln>
such that li ∈ string. An output path p points to a file, which will store
web pages to be crawled, with p ∈ string.

Output  : A file which contains crawled web pages P = <p1, p2, ..., pn>
with pn ∈ string.

Crawler(f, url list, p)
1 if f = MapReduce
2   then map phase
3       for every url <- l1 to ln
4        do pn <- crawl(url)
5          collect(url, pn)
6       reduce phase
7       for every url <- l1 to ln
8        do save(url, pn) to p
9  else if f = Hama
10    then read url-list
11     for every url <- l1 to ln
12      do pn <- crawl(url)
13        collect(url, pn)
14     bsp sync()
15    while url <- l1 to ln
16      do save(url, pn) to p
17 output p
```

Figure 3. The pseudo code of the crawling program with MapReduce and Hama implementation

The cluster comprises 4 desktop machines. 3 machines equips with Intel Core 2 Duo CPU, one with Intel Core Quad CPU. Networks cards includes 2 Intel 82566DM-2 Gigabit Network Connection, Realtek RTL8111/8168B Gigabit Ethernet controller, and Marvell 88E8071 Gigabit Ethernet Controller.

The CPU lists of the four machines:

- Intel(R) Core(TM)2 Quad CPU Q8400 @ 2.66GHz
- Intel(R) Core(TM)2 Duo CPU E4500 @ 2.20GHz
- Intel(R) Core(TM)2 Duo CPU E4500 @ 2.20GHz
- Intel(R) Core(TM)2 Duo CPU E8400 @ 3.00GHz

After performing the algorithm under the designed environment, the analysis results are presented in table 1 and figure 4. Table 1 is the performance comparison of implementing Hama BSP and MapReduce in the system and figure is the x-y plot to visualize the same data in table 1.

In table 1 and figure 4, we perform four different experiments by using the crawling program in figure 3 to retrieve and store data from100, 1000, 2000 and 10000 different URLs. The experiments are evaluated in millisecond to test the performance

Table 1. The performance comparison of implementing Hama BSP and MapReduce in the system

	Hama BSP	MapReduce
100 (URLs)	217511 (Millis)	120458
1000	920010	1486860
2000	1542519	3119418
10000	9077645	15606370

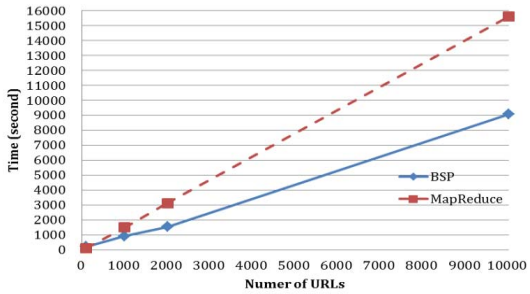

Figure 4. The performance comparison of implementing Hama BSP and MapReduce in the system

In figure 4, the x-axis shows the Number of URLs and the y-axis shows the execution time by second. From the analysis result it shows MapReduce 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 MapReduce.

V. Conclusion and Future Research

In this paper, a system architeture 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.

In order to evaluate the performance of the system and by this to decide the main cloud computing technique that used in the system, we have designed experiments to perform a crawling algorithm under specific environment. The evaluation results show that Hama BSP have better performance than ManReduce. This would be of help and interest for the researchers who want to devote in this research area.

Future researches are suggested to evaluate the performance when performing other algorithms of social networks analysis, such as matrix transformation, visualization and social networks analysis measurements, etc. The users' experience when using the system is also a good topic and research direction to evaluate the system.

REFERENCES

[1] Agrawal, R., Ailamaki, A., Bernstein, P. A., Brewer, E. A. Carey, M. J. Chaudhuri, S., Doan, A., Florescu, D., Franklin, M. J., Molina, H. G., Gehrke, J. Gruenwald, L., Haas, L. M., Halevy, A. Y., Hellerstein, J. M., Ioannidis, Y. E., Korth, H. F., Kossmann, D., Madden, S., Magoulas, R., Ooi, B. C., O'Reilly, T., Ramakrishnan, R., Sarawagi, S., Stonebraker, M., Szalay, A. S., and Gerhard Weikum, G. (2008) The Claremont Report on Database Research, SIGMOD Rec., 37(3), pp. 9–19.

[2] 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.

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

[4] Burt, R.S., (1992). "Structural Holes", Harvard University Press, Cambridge, MA.

[5] Borgatti, S. P., and Everett, M.G. (2002), "Ucinet for Windows: Software for Social Network Analysis", Harvard: Analytic Technologies.

[6] Borthakur, D. HDFS Architecture. (2011), Avaliable at:http://hadoop.apache.org/common/docs/r0.20.2/hdfs_design.pdf (Access date: March 6, 2011)

[7] Boyd, D., and Ellison, N. B. (2007) Social network sites: Definition, history, and scholarship. Journal of Computer-Mediated Communication, 13(1), 2007.

[8] Bird, C., Gourley, A., Devanbu, P., Gertz, M. and Swaminathan, A. (2006), "Mining Email Social Networks", In Proceedings of MSR 2006, May 22-23, 2006, Shanghai, China.

[9] Chen, H., Schroeder, J,, Hauck, R. V., Ridgewat, L., Alabakhsh, ., Gupta, H., Boarman, C., Rasmussen, K. and Clements, A. W. (2003), "COPLINK Connect: information and knowledge management for law enforcement", Decision Support Systems, Volume 34, Issue 3, February 2003, pp. 271-285

[10] Chin, A. and Chignell, M. (2006), "Finding Evidence of Community from Blogging Co-Citations: A Social Network Analytic Approach", In Proceedings of the IADIS International Conference on Web Based Communities 2006, San Sebastian, Spain, February 26-28, 2006.

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

[12] Dingt C. H. Q., Zha, H., Husbands, P., and Simont, H. D. (2004), "Link Analysis: Hubs and Authorities on the World Wide Web ", SIAM Review, Vol. 46, No. 2, pp. 256-268.

[13] 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. 58-64.

[14] Furukawa, T., Matsuo, Y., Ohmukai, I., Uchiyama, K., Ishizuka, M. (2007) "Social Networks and Reading Behavior in the Blogosphere" In Proceedings of ICWSM 2007, Boulder, Colorado, USA, pp. 51-58

[15] Ghemawat, S., Gobioff, H., and Leung, S., "The Google file system," In Proc. of ACM Symposium on Operating Systems Principles, Lake George, NY, Oct 2003, pp 29–4

[16] Glaser, J., Dixit, J., Green, D.P.,(2002). "Studying hate crime with the internet: what makes racists advocate racial violence?", Journal of Social Issues, 58 (1), 177–193.

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

[18] 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.

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

[20] 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.

[21] Hamasaki, M., Matsuo, Y., Ishida, K., Hope, T., Nishimura, T. and Takeda, H. (2006) "An Integrated Method for Social Network Extraction" In Proceedings of 2006 Internet World Wide Web Conference, May 23-26, Edinburgh, Scotland.

[22] Jun, T., Kim, J. Y., Kim, B. J. and Choi, M. Y. (2006). "Consumer Referral in A Small World Network", Social Network, Vol. 28, Issue 3, July 2006, pp. 232-246.

[23] 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.

[24] Jiyang Chen, Osmar Zaïane, Randy Goebel, "Local Community Identification in Social Networks", (ASONAM), In Proceedings of 2009 International Conference on Advances in Social Network Analysis and Mining, July 20-22, 2009, pp. 237-242.

[25] Laumann, E., Marsden, P., and Prensky, D. (1983). "The boundary specification problem in network analysis", In R. Burt and M. Minor (Eds.), Applied network analysis, Beverly Hills, CA: Sage, 18-34.

[26] Lento, T., Welser, H. T., Gu, L., and Smith M. (2006) "The Ties that Blog: Examining the Relationship Between Social Ties and Continued Participation in the Wallop Weblogging System" In Proceedings of the 15th International World Wide Web Conference, May 23-26 2006, Edinburgh, Scotland

[27] Malewicz, G., Austern, M. H., Bik, A. J. C., Dehnert, J. C., Horn, I., Leiser, N., and Czajkowski, G. (2009) Pregel: a system for large-scale graph processing. In Proceedings of the 28th ACM symposium on Principles of distributed computing, page 6, New York, NY, USA.

[28] Matsuo, Y., Tomobe, H., and Nishimura, T. (2007), "Robust Estimation of Google Counts for Social Network Extraction", In Proceedings of Twenty Second Conference on Artificial Intelligence (AAAI-07), July 22-26 2007, Vancouver BC Canada.

[29] Mutton, P. (2004), "Inferring and Visualizing Social Netwokrs on Internet Relay Chat", In Proceedings of the Eighth International Conference on Information Visualization, pp. 25-43.

[30] Scott, J. (2000), "Social Network Analysis: A Hand Book (2nd ed.)", SAGE publication, 2000.

[31] 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

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

[33] Vise, D. and Malseed, M. (2005), "The Google Story", Delacorte Press, USA.

[34] Wellman, B. (1982), "Studying personal communities", In P. Marsden and N. Lin (Eds.), Social structure and network analysis, Beverly Hills, CA: Sage, 61-80.

[35] Wellman, B. and Berkowitz, S. D. (ed.), (1988), "Social structures:A network approach", Cambridge University Press, pp. 19-61.

[36] 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