A Hadoop based Weblog Analysis System

Chen-Hau Wang
Department of Computer Science
National Chiao-Tung University
Hsinchu, Taiwan
chwang612@gmail.com

Ching-Tsorng Tsai
Information Engineering
Tunghai university
Taichung, Taiwan
cttsai@thu.edu.tw

Chia-Chen Fan
Department of Computer Science
National Chiao-Tung University
Hsinchu, Taiwan
wandy260178@yahoo.com.tw

Shyan-Ming Yuan
Department of Computer Science
National Chiao-Tung University
Hsinchu, Taiwan
smyuan@cs.nctu.edu.tw

***Abstract*-In recent years, cloud computing has been an important issue in the field of research. Cloud computing employs distributed storage and distributed computing technology to achieve a large number of stored data, as well as fast data analysis and processing. As the rapid development of Internet technology, digital data showing explosive growth, the face of massive data processing, the traditional text software and relational database technology has been facing a bottleneck, presented the results are not very satisfactory. For this problem, the concept of cloud computing is a more appropriate choice. In this paper, based on the architecture of Hadoop with HDFS (Hadoop Distributed File System) and Hadoop MapReduce software framework and Pig Latin language, we design and implement an enterprise Weblog analysis system. Experimental results, by analyzing daily Weblog records, we get Application Server traffic trends, performance of program statistical reports, and performance reports of different intervals and different actions of program by user request. The main purpose of this system is to assist system administrators to quickly capture and analyze data hidden in the massive potential value, thus providing an important basis for business decisions.**

Keywords—Cloud Computing, Distributed File System, Pig programming language

I. INTRODUCTION

In recent years, internet and mobile have become very public. These devices generated a lot of data. Cloud Computing is a hot topic to deal with this information. Cloud computing technology utilizes the integration of resources to reduce costs and simplify computing platform. It affects our life and work. Google has developed a number of cloud computing technology and architecture, such as MapReduce, Google File System etc. These technologies and architecture has a feature that developers don't know how to place the data and cut the computing on distributed systems. Developers focus his mind on service development and underlying architecture handle the data and dispersing and cutting of computing. These technologies and architectures can increase the speed of development services. In this era, database application is very important such as train schedules inquiries and reservations, social network (Facebook) and Gmail etc. These applications are all needed database technologies. Both academia and industry are great importance the research of distributed cloud computing technology and investment substantial resources in developing cloud-related products. Various sectors of the IT staff have also started to explore how to use cloud technology to get better at working up efficiency. Relational Database Technology Architecture of cloud architecture is very different from traditional architecture. It is difficult in the technically when large enterprises migrate traditional database systems to cloud-based or creating a cloud computing database. We can overall designed and consider the advantages of cloud computing technology to integration of the internal resources if we development a new systems. The system uses the most efficient and the easiest manage to provide a table and reliable computing and storing energy services.

II. RELATED RESEARCH AND BACKGROUND

A. Hadoop

Hadoop [1] is the open-source program under Apache Software Foundation. It comes from a part of internet search engine in Apache Nutch [2]. Hadoop was developed from java. It is a distributed computing software platform that user easier writing and handling a lot of data.

Hadoop software architecture features:

1. It can handle and storage amount of data

2. Using distributed File System obtain quick response.

3. Automatically obtained backup data and deployment of computing resources when the error occurred in a node.

The Hadoop architecture of core is Hadoop Distribution File system (HDFS) [4] and Google MapReduce of open Source [3], [8]. Google proposes the Cloud key technology to implement Hadoop. Google compared with Hadoop is in the TAVLE I. HDFS can automatic acquisition the backup data and deployment of computing resources when the node error. Hadoop can deploy in cheap hardware to form a distributed system. MapReduce programming model allows users to application development when don't understand the underlying details of the distributed system.

TABLE I. GOOGLE COMPARED WITH HADOOP

	Google	**Hadoop**
Develop Group	Google	Apache
Algorithm Method	MapReduce	MapReduce
File System	GFS	HDFS
Storage System	BigTable	HBase
Resource	open document	open source

978-1-4799-4266-4/14 $31.00 © 2014 IEEE
DOI 10.1109/U-MEDIA.2014.9

B. HDFS

HDFS is distributed File System in Hadoop. It is easier to expand. This purpose is analysis a lot of data operates on inexpensive hardware and provide fault tolerance. HDFS is master-slave model architecture (Master/ Slave). A set of Hadoop cluster is composed of a Namenode and many Datanode. Namenode is management of the file system namespace, tree structure of maintenance the file system and all contents information. File system is stored in the Datanode. Datanode usually served by a number of different nodes and return to Namenode currently storage block list information. HDFS in order to ensure data consistency so provide a single directory system and file namespace. The access mode is once writing and many times access. A data file be cut into several smaller blocks [13] and stored in a different Datanode when the user writing the HDFS requirement. Each block has several replications storing in different nodes to achieve high fault-tolerance features. The HDFS architecture is in the Fig. 1.

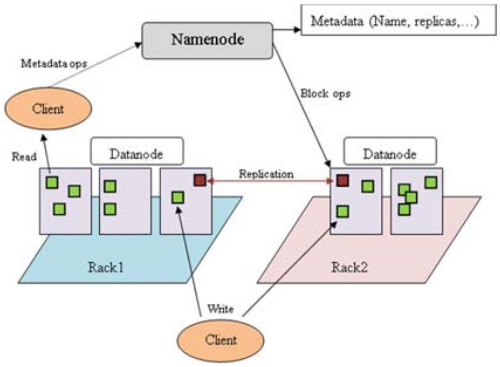

Fig. 1. HDFS architecture

C. MapReduce

Google proposed an important technique to compute a lot of data is MapReduce. This technique uses distributed computing. The work requires a lot of computing resources, cut into many sub work across multiple computing unit, and finally merge the results of multiple computing unit for the final result. The MapReduce data flow and control flow is in the Fig. 2.

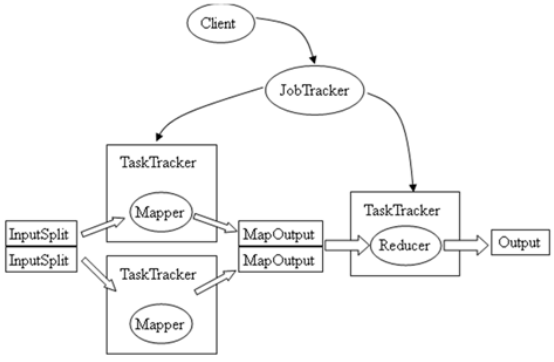

Fig. 2. MapReduce data flow and control flow

D. Pig

Pig [5] [6]is SQL-like language. It is a high-level data processing language and suitable for processing large data sets. Pig language is more focused on data query and analysis, rather than the data to modify and delete operations.

Pig is mainly performed in a distributed cloud platform. Pig has the advantage of very high speed, can handle a lot of information in a short time. Such as system log, specific web data, a large database files, etc.

III. METHODOLOGY AND IMPLEMENTATION

A. Weblog System architecture format

This paper used the finance the company's core in domestic. The core system architecture is in the Fig. 3. The business models of the financial sector is not only working in the office but also visiting outside customers and supply immediate services. Both requirements of staff and handling customer business are using this core system. Application server contains two sections that internal network and external network. An intranet network consists of three Application Servers of cluster architecture in Fig. 3. The three Application Servers are 3AP, 4AP, and 5AP respectively. They have a load-balancing mechanism. External network consists of two Application Servers of cluster architecture in Fig. 3. The two Application Servers are 2AP and 6AP. They have a load-balancing mechanism too. All Application Servers access the same one DB Server. Considering information security, they aren't directly connected between Server and Server. They are needed communication through controlling a firewall.

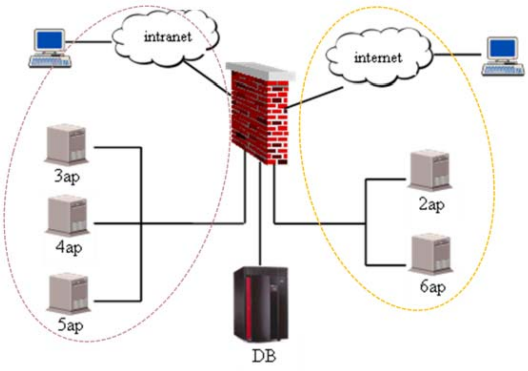

Fig. 3. Existing system architecture

Five Application Servers are documented all program execution trace in the Fig. 3. These documents are called a Weblog. A Weblog is text format saving in on each machine for convenient analysis. The business volume increase, so weblog very large daily. It is very time consuming and impractical to analysis the data. Therefore, it is very important that design a system to handle and analyzed the huge Weblog by the characteristics of cloud computing technology. The system is important when performance tuning, system expansion, equipment upgrades or future procurement planning.

B. System environment and parameter settings

This paper uses Hadoop building a set of application system to seeking the solution of weblog data analysis. The system compiles internal norms and doesn't affect company's operations. It hopes that have reference value in enterprise IT architecture. The experimental architecture simulates Hadoop in the VMware environments. First, the main is function test and program flow. Second, the information moves to real cluster architecture after confirming the data. Because Hadoop is an ecosystem and it contains many collaborative projects. It has compatibility issues between different versions of different project. Therefore experimental environment used in stable version. Experimental environment hardware and software specifications are as follows:

VMware version: VMware-server-2.0.2-203138

Namenode and Datanode: 1GB of RAM, system Ubuntu 10.04 and kernel version 2.6.32-38

Hadoop platform used hadoop-0.20.2 and pig-0.10.0, respectively. Hadoop package provides a profile and described in the following TABLE II.

TABLE II. HADOOP CONFIGURATION FILES OF ENVIRONMENTAL PARAMETERS

Files	Explanation
hadoop-env.sh	When the operation of Hadoop environment
core-site.xml	Hadoop core configuration, such as I/O settings of HDFS and MapReduce; IP location of master
hdfs-site.xml	Background service settings of HDFS, the number of namenode, datanode, replication
maprd-site.xml	Background service settings of MapReduce, jobtrackers and tasktrackers
master	Implement secondary namenode machine list
slaves	Implement datanode and tasktracker machine list

Hadoop can be started after completion of the relevant settings and observe functioning properly:

It starts Hadoop and execution of bin/start-all.sh in the installation directories. It will open Namenode, Datanode, JobTracker, TaskTracker related information in the process. We use jps instruction inquiry into process id of each component. The Hadoop start sample is in the following Fig. 4.

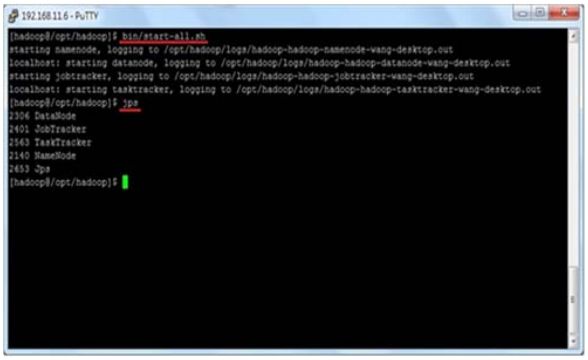

Fig. 4. Hadoop start sample

The installation of the Pig is very easy. The prepare Java environment then to unzip the Pig into planned path and to set related PATH in environment variable. The Pig environment variable setting is in the following Fig. 5. Pig can direct execution when the batch scheduling or immediate need.

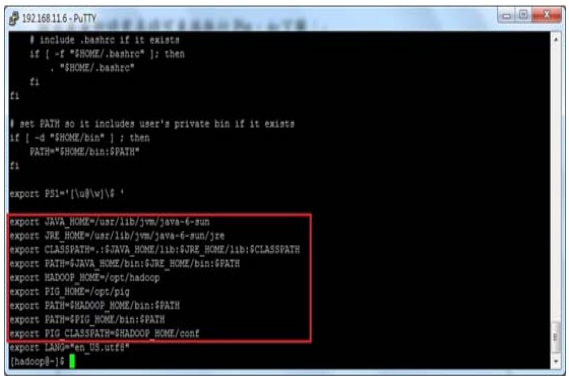

Fig. 5. Pig environment variable setting

C. System architecture and processes

In system implementation, all of the professions records are dispersed in the different AP server. First, AP server's Weblog spread Linux through the FTP protocol doing preliminary collation. In the same profession units integrated into a log file such as Fig. 6 step (a). After the files have finished cutting, designation block size upload into HDFS specified directories through command such as following Fig. 6 step (b). It doesn't change the data after Weblog file stores in HDFS. System administrator develops Pig program to load the log and copy analysis result to restore for Linux directories such as following Fig. 6 step (c). Finally, System administrator create different dimensions analysis reports according results such as sales side, Application Server flow and data etc. such as following Fig. 6 step (d).

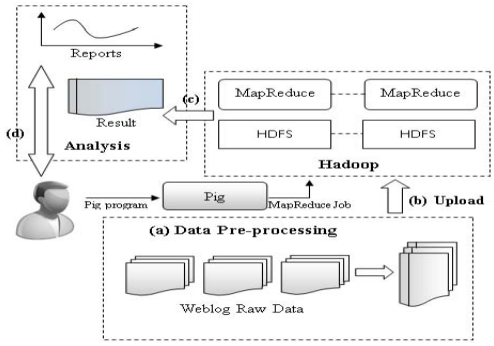

Fig. 6. System flow chart

IV. SYSTEM IMPLEMENTATION

A. Program Development

The system functions are divided into two categories. It is batch analysis and interactive input conditions. The main purpose of batch analysis is traffic statistics and conduct multidimensional analysis. Interactive function is analysis specific business and specific range. The basic program

structure is a shell script frame to receive different needs parameters and to call the Pig script. The program flow is in the following Fig. 7.

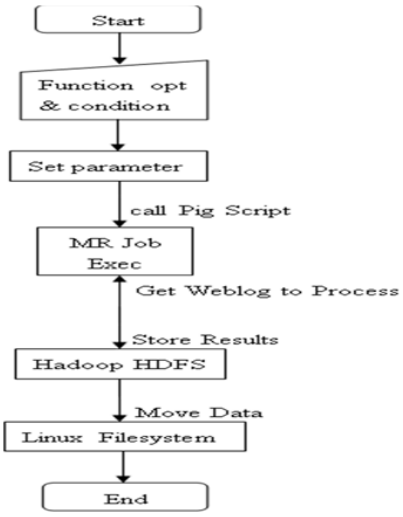

Fig. 7. Program flow

B. System traffic trends Batch Analysis

The example is a Weblog file at March 1 to March 5, 2013. It is selected important business for analysis objects. The result is in the following:

The main purpose of daily system traffic statistics can observe the growth trend of Server loading. The data is the important information of system upgrade or expand. The following Fig. 8 is the analysis of early March. For this analysis result, we found a potential problem for system administration to reference or review that stability of load balancing mechanism.

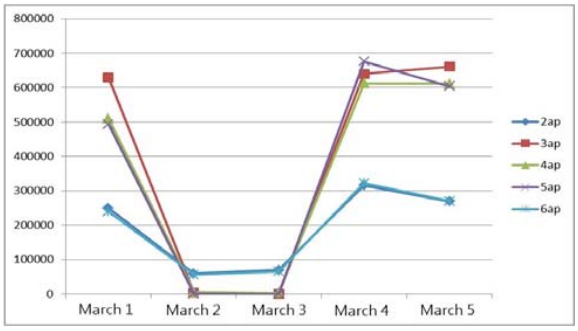

Fig. 8. Daily system flow char

Another analysis is observed daily traffic trends basis on each business such as following Fig.9. Because of the characteristics of financial operations and internal policy factors can be summed up some important dates. Such as the data are pay deadline, performance balance sheet date, and product promotion. The system load will be significantly different when important dates. The purpose of Fig.9 provides considerations for systems management or future systems assessment planning before the special data.

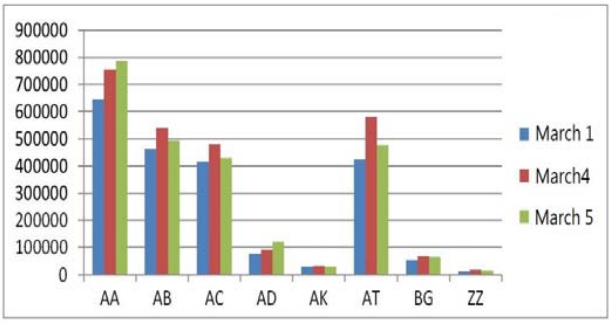

Fig. 9. The business flow charts

C. System flow specified interval analysis

In addition to the statistics of daily batch generating, there are temporary needed for query specific interval of system flow. Therefore, the paper developed an interactive function of input conditions. This function is mainly provided for system management such as following Fig. 10.

```
[hadoop@~/pig_scripts]$ ./cust_analysis_main.sh
Please choose：(1)Program computing count statics;
(2) AP server flow statics; ==>  2
Please enter dates (Format YYMMDD):
130304
For more information, please enter the starting time
(Format HHMM):
1300
For more information, please enter the end time
(Format HHMM):
1700
Please enter a Business Code:
AT
```

Fig. 10. The input screen of system flow conditions specified

D. Program performance statistics

In addition to focus on system-flow data, the paper is analyzed the program features side. The daily batch statistical reports show the daily total number of executions for each business program such as the TSBLE III. and the average execution time of sales program such as the TABLE IV..

TABLE III. THE EXECUTION TIME AND STATISTICS OF SINGLE SALES PROGRAM

Ordered list	Program Name	Computation time (ms)
1	com.caxxxy.aa.a0.trx.AAA0_0100	54950
2	com.caxxxy.aa.a0.trx.AAA0_0200	52054
3	com.caxxxy.aa.z0.trx.AAZ0_0200	46371

TABLE IV. THE AVERAGE EXECUTION OF SINGLE SALES PROGRAM

Ordered list	Program Name	Computation time (ms)
1	com.caxxxy.aa.b0.trx.AAB0_0202	132708
2	com.caxxxy.aa.z7.trx.AAZ7_0100	12747
3	com.caxxxy.aa.b9.trx.AAB9_0200	5432

Program performance analysis provided interactive input conditions features. This system uses system retrieves records to analysis the function response speed. The result shows action total computing times and statistical data on the AP Server. The data are computing times of action, average computing time, the most computing time and the shortest computing time respectively. The specified conditions analysis results of program performance are in the follow TSBLE V. :

TABLE V. THE SPECIFIED CONDITIONS ANALYSIS RESULTS OF PROGRAM PERFORMANCE

Query date: 2013/3/4			**Query time: 08:00~12:00**			
Program Name: com.caxxxy.aa.e0.trx.AAE0_0700						
The total times statistics						
AP Server	2ap	3ap	4ap	5ap	6ap	Subtotal
Computing times	5	491	439	665	12	1612
query action statistics						
AP Server	2ap	3ap	4ap	5ap	6ap	Subtotal
Action computing times	1	483	430	650	6	1570
Average computing time	118	166	234	157	353	181
Most computing time	118	3079	4048	1703	1086	4048
Shortest computing time	118	28	43	86	123	28

E. Performance comparison

In this section, we focus on the HDFS performance and test two programs performance of Pig and Shell scripts.

Test 1: HDFS is a distributed file system. The data is centralized management (Namenode) and distributed Storage (Datanode). HDFS files system partition size and copy block size is 64MB and stored in Datanode. It affects the number of Map in Mapreduce and efficiency what file size and block size. The number of map relationship between the files and block size is following:

Total data size / block size = map number

We are according to block size cut the files when the files over block size. After cutting we distribute to different node for processing. Large block size let hard disk data transfer time longer than search for a file starting position seek time. This makes the data transfer speed of HDFS and hard disk transfer speed closer. Smaller block size let big files cutting to more small files. This cause a lot of copy, read, write, and more disk seek time then drag on Map performance.

We test the effect on performance of different block size and try to search the most suitable for block size on this system. Test Conditions: A Weblog files about 300MB is cut 64MB, 128 MB, 256 MB and 512MB block size then uploaded to HDFS and implement System traffic trends batch program. The test result is in the TABLE VI..

TABLE VI. BLOCK SIZE PERFORMANCE COMPARISON

Block size	**Cutting the number of files**	**Computing time**
64MB	5	2 minutes 19 seconds
128MB	3	2 minutes 4 seconds
256MB	2	1 minute 18 seconds
512MB	1	1 minute 12 seconds

If the big files block size is so small that cutting too many file and metadata stored in memory of Namenode can't be released so take up too much resources. This system's weblog size is average larger than 256MB for a single sales file so this system block size defaults to 256MB.

Test 2: Linux environment can use very commands or tool for processing and analysis the word files. In recent years, files are almost more several hundred MB or GB. Content analysis requirements are variable, too. It is very impractical and difficult to reach users that only use sample Linux instruction and Shell scripts. We wrote a Shell scripts to simulate program computing count statistics function. Using this function tested different file size and compare with Hadoop and general word processing tools performance differences. The result is in the following Fig .11.

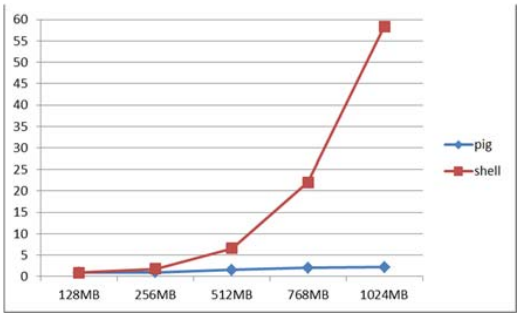

Fig. 11. Compare Pig and Shell processing time (minute)

Hadoop is a distributed architecture and suitable for handling large files. Hadoop performance is better than shell command when the files over 512MB. Pig is a high-level programming language. It provides so lots of data analysis functions that to meet the different needs. Thus, Hadoop architecture is suitable to solution this system's analysis requirements.

V. CONCLUSION

This paper proposes a Weblog analysis system based on Hadoop. The implementation shows that MapReduce program Structure can effective solution very large Weblog files in the Hadoop environment. Pig programming language can be easily analyzed the requirement of log and better performance. This system provided two functions for Application Server traffic analysis and program performance statistics. AP Server traffic Statistics provided system administrators monitoring system, found the potential problems, and predict the future trend of the system. Program performance statistics provides a reference for performance tuning.

REFERENCES

[1] Apache Hadoop. Available: http://hadoop.apache.org/
[2] Apache Nutch. Available: http://nutch.apache.org/
[3] Hadoop MapReduce. Available: http://hadoop.apache.org/mapreduce/
[4] Hadoop Distributed File System. Available: http://hadoop.apache.org/hdfs/
[5] Apache Pig. Available: http://pig.apache.pig/
[6] Pig Latin. Available: http://wiki.apache.org/pig/PigLatin

[7] NCHC Cloud Computing Research Group. Available: http://trac.nchc.org.tw/cloud

[8] Hadoop0.20 Documentation-Map/Reduce Tutorial, The Apache Software Foundation, 2008

[9] D. Borthakur, "The Hadoop Distributed File System: Architecture and Design," The Apache Software Foundation, 2007.

[10] J. Dean and S. Ghemawat, "MapReduce: simplified data processing on large clusters," in Proc. of the 6th USENIX Symposium on Operating Systems Design & Implementation (OSDI), 2004.

[11] Y. Yu, M. Isard, D. Fetterly, M. Budiu, U. Erlingsson, P. K. Gunda,and J. Currey, "DryadLINQ: A system for general-purpose distributed data-parallel computing using a high-level language," 2008.

[12] C. Olston, B. Reed, U. Srivastava, R. Kumar, and A. Tomkins,"Pig latin: a not-so-foreign language for data processing," in Proc. of the SIGMOD Conf, 2008, pp. 1099–1110.

[13] Mackey, G. Sehrish, S. Jun Wang, "Improving metadata management for small files in HDFS" IEEE International Conference on luster Computing and Workshops, pp.1–4, Aug. 2009.

[14] A. Gates, O. Natkovich, S. Chopra, P. Kamath, S. Narayanam, C. Olston, B. Reed, S. Srinivasan, and U. Srivastava., "Building a High-Level Dataflow System on Top of Map-Reduce:The Pig Experience." Proc. of the VLDB Endowment, vol. 2,no. 2, 2009.

[15] R. Chaiken, B. Jenkins, P.-A. Larson, B. Ramsey, D. Shakib, S. Weaver, and J. Zhou., "Easy and Efficient Parallel Processing of Massive Data Sets," Proc. of the VLDB Endowment,vol. 1, no. 2, 2008.