Building an Experimental Platform for Cloud and Big Data Education

Genlang Chen
Ningbo Institute of Technology, Zhejiang University, Ningbo, China
cgl@zju.edu.cn

Shiting Wen*
Ningbo Institute of Technology, Zhejiang University, Ningbo, China
wensht@foxmail.com
*Corresponding author

Jinqiu Yang
Ningbo Institute of Technology, Zhejiang University, Ningbo, China

GuanHui Song
Ningbo Institute of Technology, Zhejiang University, Ningbo, China

***Abstract*—The mission of Higher Education is to foster specialized talent with innovative spirit and practical ability. As regards the discipline of computer science, hands-on practice plays a very important role in the process of innovative and quality education. However, with the rapid development of computer technology, traditional computing education is unable to adapt to the current cloud-based environments and large data distributed computing technologies. Currently, distributed computing education only emphasizes theoretical knowledge, and never focuses on practical exercise. In this paper, we present an experimental platform to address this urgent problem. Our platform enables users to perform cloud computing and big data processing technology, whether they are campus students or distance learners. Moreover, our experimental platform can improve students' interest in leaning and build a bridge between campus education and industrial application requirements in distributed computing technologies. The experimental results show that students may have high self satisfaction and satisfaction with their company while training on this platform.**

Keywords-Cloud Education; Lab Platform; Big data Education

I. INTRODUCTION

We are surrounded by data, tens of thousands of videos and pictures which every moment are uploaded to the Internet. In 2003–2006, Google published three academic articles – GFS[1], MapReduce[2] and Bigtable[3] – to address the challenge of big data. Once published in various studies, this immediately attracted wide attention, because many other companies were also encountering the challenge of data expansion. Subsequently, Doug Cutting had the opportunity to lead a project to develop an open source version of MapReduce, called Hadoop[4]. In time, other Internet companies such as Yahoo and Facebook responded. Further, some Internet companies adapted Hadoop as a core part of their technologies stack. Today, many traditional industries, such as telecommunications, mobile and traditional media industries, also use Hadoop system as a basic data processing platform. Hadoop has had a revolutionary impact on Internet technology. It has also promoted the development of cloud computing technology. A few decades ago, relational databases and network security were considered optional programmer skills and knowledge, but they have become compulsory requirement for an efficient programmer. Similarly, understanding distributed data processing will soon become an indispensable skill for every programmer. Cloud computing has been regarded as a promising technology since its unique parallel programming models can help make more efficient processing big data, and easily build a data centre[6]. Cloud computing provides on-demand computing power, storage space and information service. All services are safe and reliable, and can be shared, scalable, large-scale and low price. Due to these advantages, cloud computing has become a main big data solution.

Now, Hadoop is the most successful open source cloud computing platform. Therefore, many large Internet companies such as Taobao, Baidu and Jingdong deploy Hadoop to address their concurrent data access bottlenecks.

More and more enterprises expect to enroll new employees who have mastered Hadoop to operate and manage their Hadoop system. However, the skilled developers of Hadoop are currently hard to find. As a direct result, the salary of Hadoop developers is much higher than other ordinary programmers. We present this experimental teaching platform for distributed technical education [7]. The main contributions are the following:

- We design our experimental teaching system combined with background and industry trend, as well as the characteristics of the school and majors. We also investigate student traits and update the experimental content so that our teaching model can cultivate innovative talents.
- Within the limited teaching hours, we try our best to help students master distributed programming and problem solving skills. We balance the industry requirements and time limitation to improve students'

comprehensive quality and innovation ability. This is of great scientific significance for enhancing the development of teaching and the curriculum.

The rest of this paper is organized as follows: Section 2 introduces the experimental platform; Section 3 illustrates the deployed environment. Section 4 examines the satisfaction of students and employers. Section 5 concludes the paper.

This template provides authors with most of the formatting specifications needed for preparing electronic versions of their papers. All standard paper components have been specified for three reasons: (1) ease of use when formatting individual papers, (2) automatic compliance to electronic requirements that facilitate the concurrent or later production of electronic products, and (3) conformity of style throughout a conference proceedings. Margins, column widths, line spacing, and type styles are built-in; examples of the type styles are provided throughout this document and are identified in italic type, within parentheses, following the example. PLEASE DO NOT RE-ADJUST THESE MARGINS. Some components, such as multi-leveled equations, graphics, and tables are not prescribed, although the various table text styles are provided. The formatter will need to create these components, incorporating the applicable criteria that follow.

II. SYSTEM FRAMEWORK OF EDUCATION PLATFORM

A. System Overview

The experimental platform uses Hadoop as a base platform[5]. The physic node uses a general-purpose computer. This experimental platform has many advantages, such as scalability, low cost and so on. Thus, it is very suitable for teaching. The specific topology is shown in Figure 1.

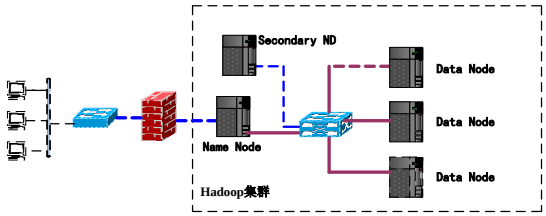

FIGURE I. TOPOLOGY OF HADOOP CLUSTER

It is of great advantage that Hadoop is very convenient and simple to operate. Thus, the programmers can effectively build and run distributed procedures to process big data sets. Further, building a Hadoop cluster is very cheap since we can use a general computer as physic nodes and almost all software are open source. For these reasons, it is very suitable for teaching. On the other hand, due to its robustness and scalability, it can be competently used in a Yahoo and Facebook stringent production environment. These features make Hadoop very popular both in academia and industry. At the same time, as open source software, it is our first choice for teaching.

There are four basic roles of Hadoop nodes: NameNode, Secondary NameNode, DataNode and Tasktracker Node. The Tasktracker Node and DataNode can be merged in teaching experimental platform.

The experimental platform deploys 10–20 general computers as physic nodes (we can use existing computer experiments for cost saving). We only deploy one better PC as NameNode. A wireless router and other infrastructure hardware are also deployed accordingly. The detailed configuration is shown in Table 1. The whole system has clear structure, is cheap and has high throughput meeting our teaching needs especially for undergraduates' distributed programming practice.

TABLE I. DETAIL CONFIGURATIONS

HARDWARE(TYPE)	***Function***
Dell PowerEdge R610 1U Rack Octal Core-XEON L5520*2 Memory Dell DRR3-4G*4 HardDisk SATA 2T*2	DataNode
Dell PowerEdge R720 2U Rack Intel XEON E5-2600*2 Memory Dell DRR3-32G*2 HardDisk SATA 4T*2	NameNode
H3C S5120-52P-SI 48	Communication

Here, we explain the function of NameNode and DataNode. Name Node is deployed for storing the metadata information of the system (metadata) and scheduling jobs. It is the centre of the whole system. DataNode is deployed for storing user data and performing user tasks. The secondary NameNode is deployed as clod backup and for taking log compression combination.

B. Deploy Open Source Software

The family of the Hadoop project has a number of sub open source systems. This is very convenient to our teaching. We can freely access, use, modify and distribute these open source software as long as there is compliance with Apache protocol without charging any fee. Figure 2 shows deployed software systems in our experimental platform.

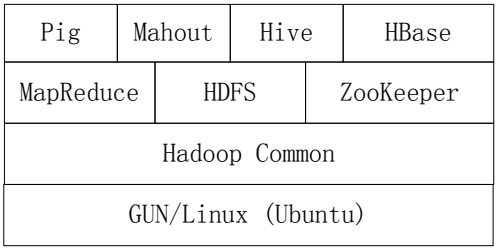

FIGURE II. TECHNOLOGY STACK OF EXPERIMENTAL PLATFORM

- Hadoop Common: this public part of Hadoop provides a variety of tools for other sub-projects including system configuration, remote procedure call, serialization mechanism, abstract file system and so on.
- HDFS: a distributed file system for data storage in Hadoop. This is a fault-tolerant file system that can detect and respond to hardware failure, for low-cost generic hardware. HDFS simplifies the file access model using streaming data access API. A master-slave structure, it

consists of a NameNode and a set of DateNodes. The NameNode manages the file system metadata, while the DataNode stores the user data. HDFS is an open source implementation of GFS.

- MapReduce: deals with massive data using a parallel framework. The "divide and conquer" programming includes two steps, i.e., Map function and Reduce function. The task is performed by a JobTracter and several TaskTracters; JobTracter is responsible for scheduling and managing TaskTracter, and TaskTracter is responsible for executing concrete tasks. Hadoop Common, HDFS and MapReduce constitute the three major components of the early Hadoop project. MapReduce is suitable for processing big data in a distributed parallel environment for a large number of computers.

- HBase: an open source implementation of Bigtable. HBase can provide scalable, highly reliable, high-performance, distributed data storage and management. It is a column-oriented dynamic database.

- Hive: a data warehousing tool, first designed by Facebook that provides data storage manipulation and query like traditional SQL language, called HiveQL.

- Pig: runs on Hadoop for large datasets for analysis and evaluation. It provides a high-level, domain-oriented abstraction language – Pig Latin – which enables users to generate automatically MapReduce programming to deal with big data.

- Mahout: the objective of Mahout is to create scalable machine learning algorithm. It is also an apache top-level project based on MapReduce for implementing data mining algorithm programming. The Mahout now includes clustering, classification, recommendation engine and frequent item sets mining and other widely used data mining algorithms.

Currently, the experimental platforms that focus on cloud technologies and big data processing in distributed learning for undergraduate education are relatively-speaking lacking in our country. So building an experimental platform is urgent.

III. ENVIRONMENTAL DEPLOYMENT

After building a physic experimental platform, we need prepare the datasets for training students.

A. Data Sets

Only the data allow HADOOP to do meaningful work. From the Internet we can acquire many free public datasets. On the website of the National Bureau of Economics Research (NBER),1 there are many small datasets for testing. Each of the datasets is about 250M, which is very suitable for test teaching in class by a standalone or pseudo-distribution mode.

There are also available some large free-open datasets, which can be downloaded from the following websites:

Wsdream: a Web service invocation dataset, which collects from the Internet. This dataset can be used to predict the QoS of Web service to recommend service to users. The dataset includes 339 users invocating 2,139 services to produce more than 1.5 million service call records.

Netfix: an online movie rental site. Its core business is to recommend movies to users based on similar users' evaluations. It publishes datasets to encourage users to develop better recommendation algorithm. The datasets are about 2GB, including 480,000 users, 17,000 movies and over 100 million movie ratings.

Amazon Public Datasets: Amazon EC2 [8] provides several large datasets for free development, mainly including biological, chemical and economic categories.

In addition, we can use the open source web crawler software Craw4j, crawling appropriate data, according to our requirements, from the Internet to build our own datasets.

B. Practice Process

First, we need to deploy system hardware and software. Then, the system administrator installs a no password login system, i.e., SSH login system. The administrator Login the NameNode of the cluster and starts the cluster system. For each student, we need to create a personal folder on the system (HDFS). It is recommended using students' ID numbers to name personal folders. We also need to set permissions for each student and allow them only to modify and delete the folder of data. This can effectively control the security of the cluster.

IV. DISCUSSION

According to the above illustration, our experiment platform can meet the needs of about 100 students' simultaneous operations on the terminal. Our cluster processing a 100MB–100G data sorting operation takes about 10–50s of the time. Currently, there are many excellent commercial cloud computing systems. However, there exist a large number of limitations, so it is not suitable for our teaching.

TABLE II. COMPARING WITH EC2

Type	Self Hadoop Cluster	EC2
Cost	Once built, it can be used multiple times	Charge by numbers, and has to use int. credit card
Open	All Software are Open Source	No. hidden
Operation	Easy	Difficult
Management	Fair	Fair

We built and put into use this experimental platform at the Graduate Summer School in 2012. In 2011 and 2012, there were more than 20 graduates a year engaged in cloud computing related work. Through these two units graduate employers pay a return visit to the students at the entry-month satisfaction as shown in figure 3 (left: 2011, right figure 2012).

As shown in figure 3, students trained in our system obtained high satisfaction from their employers.

At the same time, we surveyed the salary of students six months after their recruitment, as shown in figure 4 (left: 2011, right: 2012).

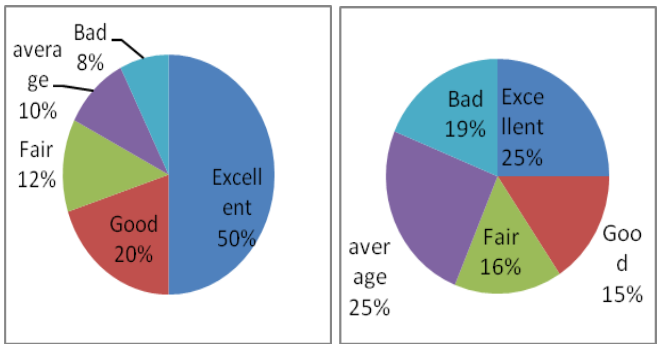

FIGURE III. EMPLOYER'S SATISFICATION

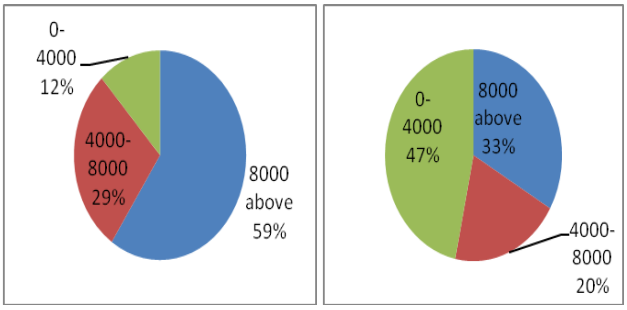

FIGURE IV. STUDENTS' SALARY

From figure 4, we can see the students who, trained on our platform, obtained a high salary. In summary, our experimental platform can improve students' programming capability, and build a bridge between companies' requirements and school teaching. Our platform is not only welcomed by students, but also improves their interest and efficiency. Moreover, as with students, our platform is welcomed by employers, and saves time and expense training new employees. Therefore, this platform development is very important for teaching.

V. CONCLUSION

We propose a Hadoop-based distributed undergraduate teaching platform for cloud computing and big data education. Our platform has scalability, reliability and low cost characteristics. In this teaching platform, all software are open source, so it greatly reduces our investment costs. Open source code is also very suitable for teaching, research and analysis. The students who are trained on the platform will obtain higher salary than others. Moreover, these students reduce the training time and costs of the production companies.

VI. ACKNOWLEDGEMENT

The work described in this paper was fully supported by the Higher Education Reform Project in Zhejiang Province kg2013520.

REFERENCES

[1] S. Ghemawat, H. Gobioff, S. Leung, The Google File System, ACM SIGOPS operating System Review, 37(5):29-43, 2003.

[2] J. Dean, S. Ghemawat, MapReduce: Simplified data processing on large clusters, Communications of the ACM, 51(1):107-113, 2008

[3] F. Chang, J. Dean etc., Bigtable: A distributed storage system for structured data, 26(2): 2008.

[4] Apache Hadoop, http://hadoop.apache.org, accessed 2013,11,01.

[5] J. Leverich, C. Kozyrakis, On the enery in efficiency of hadoop clusters, ACM SIGOPS Operating System Review, 44(1):61-65, 2010

[6] S. Nabil, Cloud computing for education: a new dawn?, Internal Journal of Information Management, 30(2):109-116, 2010.

[7] T. Carter, P. Hauselt, M. Martin, and M. Thomas, Building a big data research program at a small university, Journal of Computing Science in Colleges, 28(2): 95-102, 20102.

[8] O. Agmon, B. Muli, S. Assaf, Deconstructing amazon ec2 spot instance pricing, pp.304-311,2011.