A Construction Project of Big Data Laboratory Based on Open Source Software

Pan Qinghe

School of Computer and Information Engineering, Harbin University of Commerce, Harbin 150028, China

panqinghe200@sina.com

Keywords: Open source software, Big Data, GitHub, Laboratory Construction

Abstract: With the rapid development of computer science and technology, it has a great impact on the traditional computer courses. To adapt to this trend, many colleges and universities have set up new curriculum systems around big data, artificial intelligence and blockchain technology and it brings new opportunities and challenges for the construction of relevant experimental platforms. This research discusses open source software issues in the construction of laboratory software platforms. It discusses the characteristics, role and impact of open source software and points out the great influence of open source software on students and teachers in learning and teaching process. Two examples are given. One example shows the using of open source software in student experiment. The other describes the construction of big data software platform based on open source software in our laboratory.

1. Introduction

Nowadays Big Data [1], AI [2] and Blockchain [3] are the research focus of industry and academic. These techniques and their relative theories are changing the world. Many colleges and universities reform relative curricula and design new specialties in order to conform this trend. All these techniques put emphasis on applications in various fields so new laboratory platforms must be set up to satisfy new requirements. The laboratories are open to students, teachers and researchers. Students can apply the theories and frameworks learned in classes into practical fields. By programming on these techniques, they reinforce the knowledge and obtain the basic abilities to explore further. Teachers and researchers can use laboratories as their scientific research platform to verify their new ideas or simulate theoretical results. So, the laboratory platforms must have scalability and reliability to meet the fast developments of techniques and variable demands of student and teacher. Also, the platforms need to be economical and easy-to-use. The computer techniques develop very fast so a cheap or even free building schemes are necessary to satisfy the iterative progress of techniques. For easy-to-use it means the experiments environment should be easy to use for all people by some simple trains and exercises or it is hard and time-wasting for students or teachers to learn and research. Aiming at these problems this paper puts forward a construction project of big data laboratory based on open source software. It notes that 1) the paper focuses on software construction not hardware and 2) other laboratory platforms such as AI, Blockchain, E-business can also be built on platform we put forward and they can coexist.

2. Open Source Software

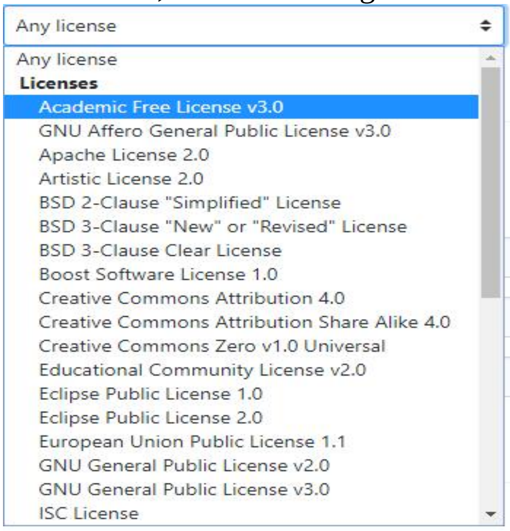
At present, many open source software can run across platforms, bringing convenience to software development. For programmers, the famous open source software includes Apache Web Server, Linux operating system, GNU C/C++ Compiler Collection, MySQL and PostgreSQL database, and Hadoop big data frame.

2.1 Open Source.

According to different writing forms, open source has two meanings [4]: the lowercase open source indicates that the source code of the program can be read and modified by other users and developers under the corresponding license restrictions. Capitalized Open Source is related to Open Source Initiative (OSI). If a software satisfies OSI approved license, then the software is Open Source. OSI gives 10 basic standards for open source software [5]:

(1) Free Redistribution
(2) Source Code
(3) Derived Works
(4) Integrity of The Author's Source Code
(5) No Discrimination Against Persons or Groups
(6) No Discrimination Against Fields of Endeavor
(7) Distribution of License
(8) License Must Not Be Specific to a Product
(9) License Must Not Restrict Other Software
(10) License Must Be Technology-Neutral

Open source software is usually released under the relevant open source licenses. You can see over 30 open source licenses on GitHub [6], as shown in Fig. 1.



Any license

Any license
Licenses
Academic Free License v3.0
GNU Affero General Public License v3.0
Apache License 2.0
Artistic License 2.0
BSD 2-Clause "Simplified" License
BSD 3-Clause "New" or "Revised" License
BSD 3-Clause Clear License
Boost Software License 1.0
Creative Commons Attribution 4.0
Creative Commons Attribution Share Alike 4.0
Creative Commons Zero v1.0 Universal
Educational Community License v2.0
Eclipse Public License 1.0
Eclipse Public License 2.0
European Union Public License 1.1
GNU General Public License v2.0
GNU General Public License v3.0
ISC License

Figure 1. The source licenses listed on GitHub

There are differences between licenses and licenses, and between different versions of the same license. Commonly used open source protocols include GPL, Apache, MIT, and BSD [7] and so on. In the laboratory construction, the chosen open source software is generally simply deployed and used, and the source code is not modified and republished, so it is not necessary to get too much entangled in the choice of the protocol. If the source code is modified and released again, it is necessary to consider the issue of open source license.

2.2 Free Software.

A concept like open source software is "free software"[8], which was advocated and put forward by Richard Stallman in 1984. The user can run free software freely, can study how the program runs and modify the program to meet its needs, and can republish the modified program. Obtaining source code is a prerequisite. Although there are some differences, open source software and free software are interchangeable concepts in many cases.

2.3 GitHub.

At present, GitHub is probably the largest open source code library in the world. In June 2018, Microsoft bought GitHub for $7.5 billion [9]. This shows that the strategic position of open source software is getting more and more attention. A lot of open source software that can be applied to construction of laboratory platforms can be found on GitHub.

Recent years in China the BAT (Baidu, Alibaba and Tencent) as the representative of innovation of new techniques are fast developing. At the same time, they contribute many high quality open source projects on GitHub. The influence is so large especially for undergrads who want to become software engineers and work in these top companies. In GitHub 2017 annual report [10] it was about seven hundred thousand Chinese users had registered in GitHub until 2017.

For undergrads in computer science and technology specialty or other relative specialties it's necessary to introduce opensource concept and helpful code repositories like GitHub to them. The benefit is obvious. They can find their interesting projects on these repositories, search projects to help themselves solve problems, and even join in some projects to contribute their code; By cloning the projects that set up by famous companies in the world, they can learn how to organize big software projects, write high quality code and grasp excellent programming practices. This is the trend in the world and GitHub also recognize this. In 2017 there were about 1.3 million students learning and 5.3 thousand teachers teaching on approximate 280 thousand GitHub repositories. The "Student Developer Pack project" [11] has helped more than 250 thousand students around the world to program with free tools and environment.

In summary, several reasons for choosing open source software can be summarized:

(1) Be economical

(2) Track the latest technology

(3) Provide more options for learning and teaching

(4) Improve students' programming skills

Beside above reasons there are two standards that should be considered when build libraries platforms.

(1) Choose more stable and community-wide software. For example, the projects with most stars or with most forks on GitHub are better choices.

(2) Pay attention to the licenses. GitHub provides advanced search capabilities at https://github.com/search/advanced. In the "With this license" drop-down list, there are about 30 optional open source licenses. Compared with other licenses, there are fewer restrictions on BSD or MIT licenses, and usually the software under these licenses are better choices.

3. Software Selection

3.1 Operating System.

The source code of most open source software is open and can be modified and released under the permission of relevant open source licenses, so it can run on Windows, *nix, Mac OS or other systems. However, since Mac OS generally run on expensive Apple computers, and the Windows systems and related software (such as office suites) need to be purchased and authorized, so the open source Linux systems are better choices for building laboratory software platform. In addition, most of the original forms of open source projects are compiled and tested under the *nix platforms so it will cause various problems when these projects are ported to Windows. For example, the various data storage and processing systems contained in the current popular Hadoop ecosystem are difficult to deploy on Windows systems. Therefore, using the open source *nix systems directly, such as Linux systems, can avoid above problems. For most Linux distributions Ubuntu [12] is easy-to-use and feature-rich. It is divided into desktop, server, and cloud service versions. The desktop and server versions are our deployment choices. The desktop version comes with many software suites and various programming IDEs. The server version is mainly used to write and debug server-side code for various types of developments.

3.2 Office and Programming Software.

In contrast to Windows systems the available open source alternatives on Ubuntu are given in Table 1. The first column represents the comparison category, the second column gives the software choices under the Windows systems, and the third column gives the representative choices of the corresponding open source software under Ubuntu. As mentioned earlier, many open source software can be used across platforms, so some of the software in the third column also has an open source version on Windows systems.

Table 1. Common software on Windows and Ubuntu

	Win7,Win10	Ubuntu
Browsers	IE	Firefox, Chrome
Office suits	Ms office	LibreOffice
Image manipulation	Photoshop	GIMP
Flow chart	Visio	Dia
IDEs		
C,C++	VC++	Codeblocks
Web development	VS, Dreamweaver	Brackets, Atom
Editor	Notepad	Gedit
Database	SQL Server	MySQL, PostgreSQL
Big data Environment	Hadoop, Spark	Hadoop, Spark
Science Computation	MATLAB	Octave

4. Examples

4.1 Example 1. Programming and Writing Experiment Report.

Students in the laboratory can use open source tools for programming and writing experiment reports. The experimental content of the computer course mainly includes the code writing, debugging, running, and writing an experiment report. There are corresponding open source tools for students to use at each stage. For Java programming students can use Eclipse to write, debug code, and use LibreOffice to write experiment reports. For C programming, students can use codeblocks to edit code and GDB to debug code. MySQL and PostgreSQL can be used to experiment with relational database systems; Linux itself can be used as an experimental reference for operating system courses. In short, the experimental content of almost any computer course can be completed with open source software.

4.2 Example 2. Big Data Processing Framework.

This example shows a big data software platform deployed in the laboratory. The operating system of all the machines in this platform is Ubuntu. Each part of the platform is composed of open source software. The platform is shown in the Fig. 2.

In general, a big data processing system consists of four parts: data acquisition, data storage, data analysis, and data visualization. The platform also consists of four parts, including:

(1) Distributed data crawling system. It consists of a crawling server and several crawling clients. The crawling clients can be deployed in the local laboratory (In Fig.2, IP address is 192.168.10.1 and 192.168.10.2, respectively) or deployed on cloud system on the Internet. The reason for deployment on the cloud is that each cloud machine can have its own IP so it can reduce the possibility of crawl blocking.

(2) Data storage system. The crawled data should be saved in an appropriate form for further analysis. The platform deploys two storage solutions: MySQL distributed cluster and Hadoop-Spark big data processing framework. They can be deployed in the same distributed cluster without affecting each other. The saved data can be either ordinary text information or multimedia information such as pictures, videos, and audios.

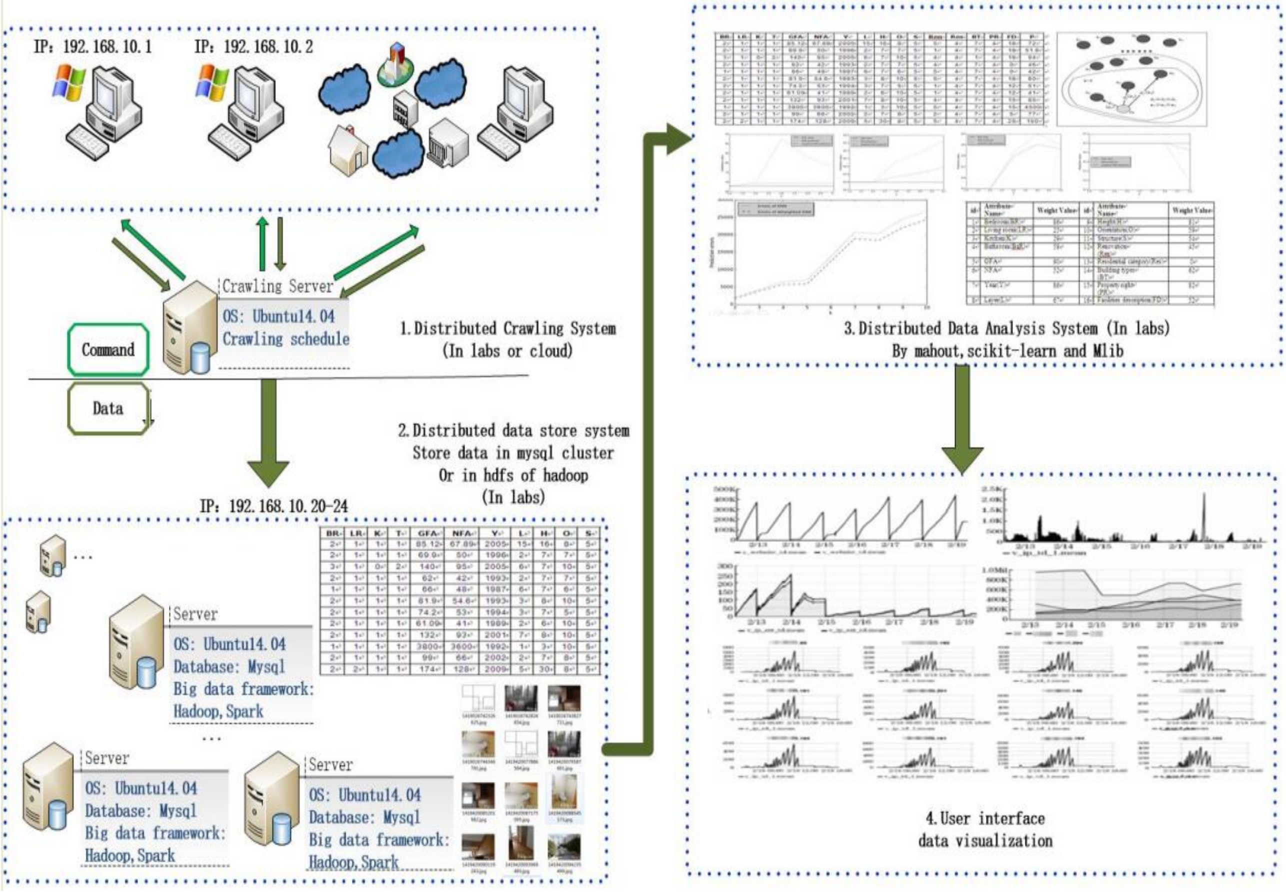
(3) Data Analysis System. It can use Mahout [13]and MLib[14] provided by Hadoop and Spark respectively as machine learning libraries, or use other machine learning libraries such as Python's scikit-learn[15]. Certainly, usually we need to write data analysis code by ourselves. The example given in top right corner of Fig.2 is an analysis that uses weighted KNN technology for housing price forecasting.

(4) Data visualization system. It is mainly used to visualize the results of data analysis and calculation. It can provide a web interface. Users can input parameters and the background services will use corresponding models to predict and classify data. The platform uses Django **[16]** as the

development framework for background and Grafana [17] as the data visualization framework.

The distributed data crawling system obtains the required experimental data, then the experimental data is transferred to a distributed storage system for storage. The distributed big data machine learning algorithm libraries or algorithms written according to requirements can be used to analyze the saved data in the future. Analysis and processing are performed, and the processed results are displayed by the data visualization system.

To sum up, the open source software used in the system mainly includes: Ubuntu, MySQL, Hadoop, Spark, Django, and Grafana.



BR	LR	K	T	GFA	NFA	Y	L	H	O	S
2	1	1	1	85.12	67.89	2005	15	16	8	5
2	1	1	1	69.9	50	1996	2	7	7	5
3	1	0	2	140	95	2005	6	7	10	5
2	1	1	1	62	42	1993	2	7	7	5
1	1	1	1	66	48	1987	6	7	6	5
2	1	1	1	81.9	54.6	1993	3	8	10	5
2	1	1	1	74.2	53	1994	3	7	5	5
2	1	1	1	61.09	41	1989	2	6	10	5
2	1	1	1	132	93	2001	7	8	10	5
1	1	1	1	3800	3600	1992	1	3	10	5
2	1	1	1	99	66	2002	2	7	8	5
2	2	1	1	174	128	2009	5	30	8	5

Figure 2. The big data software platform based open source software

5. Summary

When software platform in the big data laboratory is built, open source software can give us more choices, solve problems and save expenses. Open source is also important for learning and teaching. Students can learn source code, learn the project organization structures of large-scale software engineering, and learn to use open-source technologies to solve their own problems. Teachers can also learn from existing codes, learn the behind-the-scenes of advanced algorithms, focuses on improving code quality and operating efficiency.

References

[1] Huda M, Maseleno A, Atmotiyoso P, et al. Big data emerging technology: insights into innovative environment for online learning resources. International Journal of Emerging Technologies in Learning (iJET), 2018, 13(1): 23-36.

[2] Siau K. Impact of Artificial Intelligence, Robotics, and Automation on Higher Education, 2017.

[3] Tapscott D, Tapscott A. The BlockChain revolution and higher education. Educause Review, 2017, 52(2): 11-24.

[4] Christian Gross. Open Source for Windows Administrators. Charles River Media, Inc, 2005.

[5] https://opensource.org/osd

[6] https://github.com/search/advanced

[7] The Open Source Alternative: Understanding Risks and Leveraging opportunities. John Wiley & Sons, Inc, 2008

[8] Paul Kavanagh. Open Source Software: Implementation and Management. Elsevier Digital Press, 2008

[9] https://news.microsoft.com/2018/06/04/microsoft-to-acquire-github-for-7-5-billion/

[10] https://github.com/codefordenver/2017-Annual-Report

[11] https://education.github.com/pack

[12] Attar M R. Free Open-Source Software Has Gained the Popularity as a Software Development Method-A Review. International Journal of Advanced Research in Computer Science, 2017, 8(7).

[13] Sharma I, Tiwari R, Rana H S, et al. Analysis of Mahout Big Data Clustering Algorithms//Intelligent Communication, Control and Devices. Springer, Singapore, 2018: 999-1008.

[14] Hernández-Castaño J A, Camacho-Nieto O, Villuendas-Rey Y, et al. Experimental Platform for Intelligent Computing (EPIC). Computación y Sistemas, 2018, 22(1).

[15] Prettenhofer P, Weiss R, Dubourg V, et al. Scikit-learn: Machine Learning in Python, 2016.

[16] Dauzon S, Bendoraitis A, Ravindran A. Django: Web Development with Python. Packt Publishing Ltd, 2016.

[17] Craig Haskins. EPICS and Open Source Data Analytics Platforms. 16th Int. Conf. on Accelerator and Large Experimental Control Systems, 2018:1420-1422.