# Construction and Experimental Application of Large Data Platform in Colleges

Hai-dong LV, Shan-kun WANG, Yong-sheng YAO, Xiao-liang ZHANG and Guang-chao LIU

City Institute of Dalian University of Technology, Dalian China

**Keywords:** Big data, Experimental teaching, Personnel training, Employment competitiveness.

**Abstract.** This paper provides fully analysis to illustrate the potential and importance of big data technologies and its application, as well as the importance and necessity of building big data laboratory at colleges and universities. The paper also discusses the key aspects on building a big data technology laboratory with the open source big data framework and technologies, including Spark, Hadoop, Flink, Mesos, Kafka etc. The construction of platform infrastructure, lab team formation and development, students leaning and development bid data application in this platform also been discussed.

## Introduction

Big data[1] is one of the most talked about tech advancements in recent years, according to occupation social networking platform LinkedIn released "2017 China the hottest internet jobs report" shows that R & D engineers, product managers, human resources, marketing, operations and data analysis are six kinds of personnel positions now in Chinese Internet industry demand the most exuberant. Among them, the demand for R & D engineers is the most, and the Big data analysis engineer is the most scarce. The report shows that the index of data analysis talents is the lowest, only 0.05, which is highly scarce. Data analysis talents also have the fastest job hopping speed, and the average speed of the job hopping is 19.8 months.

According to the statistics of China Business Federation, Specialized Committee statistics show that the gap of talent in China's basic data analysis will reach 14 million in the future, and more than 60% of BAT companies employ large data talents.

Planning for the development of large data industry in December 2016 the State Council issued "(2016-2020)" pointed out that the "big data of basic research, product development and business application of all kinds of talent shortage, it is difficult to meet the needs of the development of" specific requirement "to strengthen personnel training data, the integration of colleges and universities, enterprise and social resources, to promote the establishment of innovative talent training mode. To establish and perfect the personnel data multiple level and multiple type cultivation system".

Big data technology has rapidly penetrated into various fields of economic and social development, and with the rapid development of large data technology, the training of this kind of talent has gradually become an important content in the teaching of information technology in Colleges and universities. Most colleges and universities have established big data related disciplines, promoting the discipline construction of big data, and strengthening the training of big data talents, providing effective guarantee for the development of big data industry.

In order to further promote the application innovation of big data, deepen the integration of educational research and big data application, improve the application level of big data, promote the construction of related disciplines in big data, and innovate the training mechanism of big data talents, the college big data platform need to construct and implement, to help student learn and practice big data course and technologies.

In this paper the big data platform of our college designing and construction is discussed in detail, the application of big data that helps student to learn and practice these technologies has been developed and implemented on the platform.

The platform also provides the API to support the students submit and test their own development big data application to this platform, they can view the result of analysis with very kinds of data visualization technologies such E-Chart, D3.js etc.

**Big Data Experimental Platform Construction**

The big data platform is constructed fully with open source big data technologies, this can reduce the investment cost extremely and without have to worry about intellectual property protection. The big data technology Spark[2] as the platform core, other related technologies include Apache Mesos[3], Kafka[4], Flink[5], Cassandra[6], Zookeeper[7] are used to assist performing the big data processing job, the platform architecture[8] is showed as Figure 1.

Apache Spark is a fast and general engine for large-scale data processing, Run programs up to 100x faster than Hadoop MapReduce in memory, or 10x faster on disk, so the platform mainly use Spark as big data process engine instead of another big data engine Hadoop.

The original platform use the Spark2.21, now the latest Spark 2.3 was selected to replace the old version. Apache Spark 2.3.0 is the fourth release in the 2.x line, it adds support for Continuous Processing in Structured Streaming along with a brand new Kubernetes Scheduler backend. Other major updates include the new DataSource and Structured Streaming v2 APIs, and a number of PySpark performance enhancements. Under the latest version the big data processing task can be use the new data api such DataFrame or DateSet without using the low level api RDD, the big data application development will be easy than previous version.

For platform cluster management[9] the Apache Mesos is used instead of Apache Hadoop YARN. As the Apache Mesos abstracts CPU, memory, storage, and other compute resources away from machines (physical or virtual), enabling fault-tolerant and elastic distributed systems to easily be built and run effectively[10].

In order to save hug hardware investment, total 20 old computers out of the laboratory were used to be the platform cluster node which runs as the big data process executor, each with 8GB and 1 CPU, under the Mesos management and controlling the platform total has 160GB and 20 CPU executor nodes. The cluster manager use a powerful HUAWEI server with 2 CPU and 96GB ram for improving the big data. The Ubuntu server 16.04 was installed in all the computers of platform.

For real-time streaming paradigm big data processing the Apache Kafka was used as real time data exchange center[11].

Kafka is a distributed streaming platform that is used publish and subscribe to streams of records. Kafka is used for fault tolerant storage. Kafka replicates topic log partitions to multiple servers. Kafka is designed to allow your apps to process records as they occur. Kafka is fast and uses IO efficiently by batching and compressing records. Kafka is used for decoupling data streams. Kafka is used to stream data into data lakes, applications, and real-time stream analysis systems.

Kafka is often used in real-time streaming data architectures to provide real-time analytics. Since Kafka is a fast, scalable, durable, and fault-tolerant publish-subscribe messaging system, Kafka is used in use cases where JMS, RabbitMQ, and AMQP may not even be considered due to volume and responsiveness. Kafka has higher throughput, reliability, and replication characteristics, which makes it applicable for things like tracking service calls (tracks every call) or tracking IoT sensor data where a traditional MOM might not be considered.

Kafka can work with Flume/Flafka, Spark Streaming, Storm, HBase, Flink, and Spark for real-time ingesting, analysis and processing of streaming data. Kafka is a data stream used to feed Hadoop BigData lakes. Kafka brokers support massive message streams for low-latency follow-up analysis in Hadoop or Spark. Also, Kafka Streaming (a subproject) can be used for real-time analytics.

Kafka is used for stream processing, website activity tracking, metrics collection and monitoring, log aggregation, real-time analytics, CEP, ingesting data into Spark, ingesting data into Hadoop, CQRS, replay messages, error recovery, and guaranteed distributed commit log for microservices.

In order to let the students to have a wider range of big data knowledge, the platform also use the Apache Flink framework as the another real-time stream data processing. Apache Flink is an open source platform from Apache Software Foundation for large-scale distributed stream and batch data

processing that provides data distribution, communication, and fault tolerance for distributed computations over data streams. Flink can be integrated with other open-source and big data processing tools such as Spark and Hadoop both for data input and output as well as deployment.

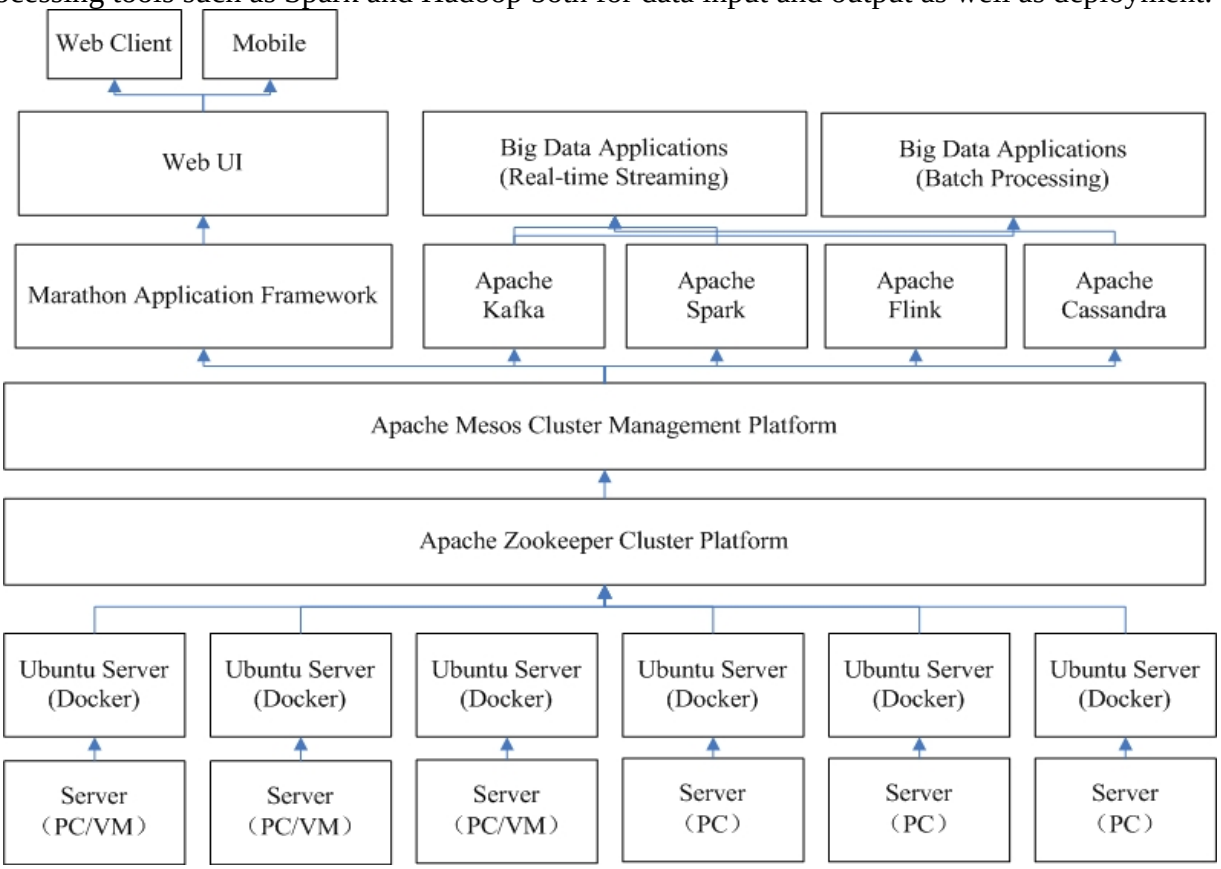


Figure 1. The Architecture of a Big Data Analytics Experimental Platform

**Design and Development of Experimental Cases**

Countless practices have proved that learning with examples is the best way to learn new technologies, the learning big data technologies are no exception. In order to give the student the big data learning examples, the teachers in our college have developed many big data processing application project and deployed in the platform.

All the big data examples can be viewed through the web and mobile client which is developed with web framework Angular 4 and server side web platform Node.js and Express.

When the student submit one kind of these big data examples to the big data platform, the example code will send to the platform with EclairJS framework.

EclairJS provides JavaScript and Node.js developers with an API for Apache Spark, and enables them to take advantage of Spark's unique data-processing environment that includes streaming, SQL, Machine Learning, and a graph database. Using EclairJS, developers can write applications entirely in JavaScript, some of which will be executed in the local JavaScript environment such as Node.js, and some of which will be executed on Spark.

EclairJS is composed of a client component that runs in the local JavaScript environment, and can be simply installed from npm, and server components that can be remote from the client and handle JavaScript in Spark.

All the big data processing application example code can be viewed on line in the platfrom, so the students can understand how the example is developed, and they can download the example project file from the platform to their local computer, then they can develop their own big data application by learning the downloaded example code.

With using this big data platform the practice has showed that the student's interest is enhanced and the teaching quality is also improved.

**Improving Students Ability of Big Data Analysis**

The platform provides the web UI interface for upload big data application package file and submit the application to big data engine, the analytical result can be showed with visualization tools such as EChart and D3.js which have been built in the platform.

The students can develop big data application with all the languages that the Spark can support, such as Scala, Python, Java and R. After student login the platform, they can upload application packaging files to the platform with application code window, meanwhile the data analysis example file can also upload with the data window.

The platform installation and configuration guiding file has been published in the document center in the platform, all the students can build their own big data processing platform with their laptop computer. By using Oracle VM VirtualBox student can setup at least three Linux virtual server, then they can install and configure Mesos, Spark, Kafka, ZooKeeper, Flink etc within these virtual server, finally has their own big data platform.

For student the most difficult part is understand the clustering cooperation framework ZooKeeper. It is a centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services. All of these kinds of services are used in some form or another by distributed applications. Each time they are implemented there is a lot of work that goes into fixing the bugs and race conditions that are inevitable. Because of the difficulty of implementing these kinds of services, applications initially usually skimp on them, which make them brittle in the presence of change and difficult to manage. Even when done correctly, different implementations of these services lead to management complexity when the applications are deployed

Up to the present a lot of students has used the platform and developed many big data application, through this platform, they learned the variety skills of big data development. In the future, there will be a very strong employment competitiveness in the job market.

**Summary**

Through the construction and application of the platform, on the one hand, it improves the scientific research ability of the large data of our teachers, and on the other hand, it provides the students with the processing platform for large data learning. In the procedure of designing and implementing of the big data platform, the teachers can master the integration of various big data techniques and frameworks and can improve their knowledge level in theory teaching, so they can solve the most question in teaching big data technologies. In this big data platform not only teachers but also the students can develop various kind of big data projects and applications.

The final result is to improve the large data teaching and scientific research ability of the whole college and cultivate a large number of big data talents.

**Acknowledgement**

This research was financially supported by the 2017 Fund subject of study of education and teaching foundation of City Institute of Dalian University of Technology (ID:JXYJ2017010).

**References**

[1] Rayan Dasoriya. A review of big data analytics over cloud, 2017 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia), Year: 2017, Pages: 1-6.

[2] Information on http://spark.apache.org/.

[3] Information on http://mesos.apache.org/.

[4] Information on http://flink.apache.org/.

[5] Information on http://kafka.apache.org/

[6] Information on http://cassandra.apache.org/.

[7] Information on http://zookeeper.apache.org/.

[8] Jong-Hoon Lee, Young Soo Kim, Jong Hyun Kim, Ik Kyun Kim, Ki-Jun Han. Building a big data platform for large-scale security data analysis, 2017 International Conference on Information and Communication Technology Convergence (ICTC), Year: 2017, Pages: 976 - 980.

[9] Bao Rong Chang, Yo-Ai Wang, Yun-Da Lee, Chien-Feng Huang, Development of multiple big data analysis platforms for business intelligence, 2017 International Conference on Applied System Innovation (ICASI),Year:2017,Pages:1930-1933.

[10] Chien-Heng Wu; Franco Lin; Wen-Yi Chang; Whey-Fone Tsai; Hsi-Ching Lin; Chao-Tung Yang, Big data development platform for engineering applications, 2016 IEEE International Conference on Big Data (Big Data),Year:2016,Pages:2699-2702.

[11] Yushui Geng; Xianzhao Yan. Spark standalone mode process analysis and data skew solutions, 2017 IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC),Year:2017,Pages:647-653.