Research on Architecture of Education Big Data Analysis System

Jinhua Chen
School of Education
Shaanxi Normal University
Shaanxi, China
School of Computer Science
Sichuan Normal University
Chengdu, China
e-mail: csdcjh@126.com

Jing Tang, Qin Jiang, Yuxin Wang, Chunmei Tao, Xu Zhang, Jingya Liao
School of Computer Science
Sichuan Normal University
Chengdu, China
e-mail: 1612706385@qq.com; 1505787857@qq.com; tangjing80@sina.cn

***Abstract*—As the big data Era, the amount of data in education shows explosive growth. Big data analysis system architecture intends to provide a reference for large educational data analysis. Through literature analysis and network investigation, This paper reviewed the existing mature system of large educational data analysis system .From two aspects of generality and difference to compare, this paper proposed research framework from several aspects, including analysis of the common thinking, the generality of open source thought, analysis of the common areas, technology framework, the core technology and special functions. This paper compares 3 mainstream data analysis system, summarizes the enlightenment for the analysis of large data, puts forward a kind of intelligent analysis system of education big data, and sums up the development prospect of big data analysis system.**

Keywords-education big data; data analysis; system architecture

I. INTRODUCTION

With the development of Internet, Internet of things, cloud computing and intelligent reading terminal, the education sector has entered the area of big data. Data cannot be captured, managed, and processed by using conventional software tools within an acceptable time horizon. [1] The generation of big educational data including educational equipment purchase records, educational web log, large educational activities and so on. These data are characterized by massive and unstructured data. The emergence of education large data, the amount of data to reach PB (1024 TB, 1TB= 1024 GB), EB (1024PB) or ZB (1024EB) level, [2] With Volume (mass), Velocity (high speed), Variety (diversity), Value (high value) 4V features, which makes the traditional database management tools for data Retrieval, collection, storage, filtering, transmission, computing, sharing, analysis and visualization to face a lot of problems. So the rapid development of education data has aroused widespread concern and attention at home and abroad, scientific and effective data analysis of education has become one of the most core problems in the field of Education. The difficulty of big educational data analysis is not its large amount of data, but in the massive, complex, unstructured data analysis. It is difficult to complete the analysis task within the specified time without the help of professional analysis tools, or it is difficult to find the hidden educational value in the big data in a short time. Education data from all types and levels of schools and educational administrative departments, It is necessary to select a suitable analysis tool, when face the large and complex educational data. To be good at something, it must be beneficial to its devices. A good tool can not only make our work more effective, but also let us in the increasingly fierce competition in the area of big data and cloud computing, which can tap the value of education big data, timely adjustment of strategic direction, and improve the ability to analyze large data.

Big data analysis technology is a new generation of technology and architecture; it is low cost, fast acquisition, processing and analysis technology, from a variety of large scale data extraction value .Education big data analysis has been widely studied. Generally speaking, we can sum up 3 development trends: First, real-time data calculation .In the context of educational background, as a supplement to the batch calculation, which was shorten the processing time of the PB data to the second level, the real-time computation is paid more and more attention. Second, Diversity of data analysis system. Since 2008, The Google and MapReduce clones of GFS and Apache Hadoop, in that have been widely accepted by Internet companies. Which had become the factual standard in the field of big data processing in Education. The appearance of Scribe, Flume, Kafka, Storm, Drill, Impala, TEZ/Stinger Presto, Spark/Shark, [3] which expanded the ecological environment of big data technology education and promoted the development of ecological environment to benign and integrity. Third, data processing engine, big data analysis of education need to get rid of the traditional general system to reduce costs and improve energy efficiency. The use of specialized system architecture has become a trend. At present, the analysis of big data is divided into five aspects: Date Mining Algorithms, Semantic Engines, Predictive Analytic, Capabilities Analytic Visualization, Data Quality Management. The core of the application of educational data is adaptive teaching, discovery of educational law and support of precision management, which can be summarized into five levels, namely, learning, teaching, research, management and policy [4].

978-1-5090-3619-6/17/$31.00 ©2017 IEEE

II. EDUCATIONAL BIG DATA ANALYSIS SYSTEM ARCHITECTURE

Educational big data analysis of the prototype of the system architecture. The current mainstream education big data calculation model and analysis of system architecture, such as Table 1.

TABLE I. MAINSTREAM EDUCATION BIG DATA CALCULATION MODEL AND ANALYSIS OF THE SYSTEM ARCHITECTURE [5]

System	Calculation mode	Describe	Developers
Hadoop	Batch computing	The first open source implementation of Map Reduce paradigm	Apache
Samza	Flow calculation	Linked In open source apache distributed flow computing system	TApache
Spark	Batch computing	Support data memory and the latest analysis system of resilience	UC Berkeley AMP Lab

- Hadoop: Yahoo, Facebook, Amazon and domestic baidu, Alibaba and many other Internet companies are based on the Hadoop to build their own distribution[6]. Hadoop uses Map Reduce distributed computing framework, and according to GFS developed HDFS distributed file system, according to Big Table developed HBase data storage system. Despite the same principles as distributed computing systems used internally by Google, Hadoop is still out of reach of Google's standards in terms of speed. However, the open source nature of Hadoop makes it the international standard for distributed computing systems.
- Samza: Samza is a distributed stream processing framework, dedicated to real-time data processing. The difference is that Samza is based on Hadoop and uses Linked In's own Kafka distributed message system. In the Samza stream data processing, each Kafka cluster is connected to a cluster that can run Yarn and processes Samza jobs. Samza is ideal for real-time streaming data processing applications, such as log services, real-time services, data tracking applications, it can help developers to high-speed message processing, but also has good fault tolerance [7].
- Spark: Spark supports the latest analysis system for memory data and recovery capabilities. It is based on Hadoop on a number of architectural improvements. Spark and Hadoop biggest difference is that Hadoop use hard disk to store data, and Spark use memory to store data, so Spark can provide more than Hadoop100 times the speed of operation[8]. However, due to the Memory outage will lose data, Spark cannot be used to process data that requires long-term storage.

III. THE COMMONALITY ANALYSIS OF EDUCATIONAL BIG DATA ANALYSIS SYSTEM

Hadoop, Samza and Spark are currently the most popular systems for big data analysis. Hadoop is used for off-line and complex educational big data processing. Samza is mainly to solve the high data rate and education of large amounts of streaming data processing. Spark is often used for off-line rapid education of big data processing. Through the education of big data analysis system architecture in-depth analysis, and summarizes its beneficial reference and enlightenment to the educational data analysis. The study found that the structure of educational data analysis system is more common in the areas of analysis thought, open source thinking and analysis.

A. Analysis of the Commonality of Thinking

Hadoop, Samza and Spark are all based on Google's three papers on big data generated, which later became the important foundation for the development of education big data, and open the door to algorithms for educational data analysts. According to the time of the first paper should be published in 2003 Google File System, which is a distributed file system. Basically: the file is divided into many blocks, use redundant means stored in the business machine cluster; here have to say that basically every article on Google is about "commercial models". Followed by the 2004 Map Reduce was announced, and now Map Reduce basically has represented the education of large data. Legend, the Google use it to calculate their search index. However, Mikio L. Braun believes that the way it works is that Google places all the crawled pages on their clusters and uses Map Reduce to recalculate each day. Bigtable released in 2006, inspired numerous No SQL database, such as: Cassandra, HBase and so on. Half of the Cassandra architecture is modeled on Bigtable, which includes the data model, SSTables, and write-ahead logging(The other half is imitating Amazon's Dynamo database, using peer-to-peer clustering). Analysis of the commonness of thought as shown in Figure 1.

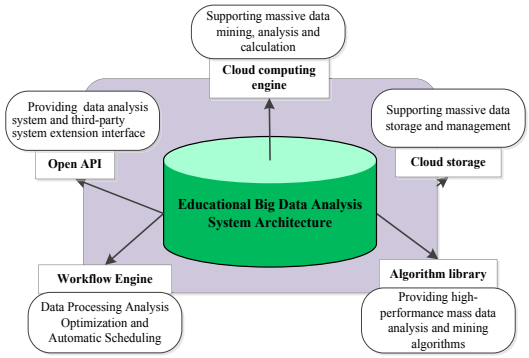

Figure 1. The generality of the theory of educational big data analysis system.

B. The Commonness of Open Source

Education big data analysis system development language is based on the JVM framework of open source language, which determines the education of big data analysis system architecture are common open source. Because open source thought is the result of computer

software, development up to now, the type and quantity of Computer class of open source products are so many and widely used. In the field of operating systems, Linux occupy a large share in the server market, and continued expansion. About 75 percent of IBM's thin-film servers are running the Linux operating system. On the Web, more than 50 percent of Web servers in the world use open source Apache systems. In this paper, the education of big data analysis system are under the Apache Foundation project. In the database area, My SQL is a lightweight database for Internet applications, PostgreSQL is for large applications, and Education Big Data Analysis System is based on a typical open source architecture.Speaking of Big data Processing System of Open Source Education, we have to think Hadoop, which is the originator of this field and it's an open source implementation of GFS and Map Reduce. Thanks to the ease of use and fault tolerance of the Map Reduce framework and the inclusion of both storage and computing systems, Hadoop has become one of the cornerstones of educational data-processing systems. Hadoop can meet most needs of the off-line storage and off-line computing, and its performance is impressive. As a result, Hadoop was able to meet more than 90 percent computing requirements of the off-line storage and off-line in the early days of building a large data-processing system, becoming the first choice for major systems.

C. *The Commonalities of Analysis Field*

While education field tends to apply big data, we can find many wonderful uses of education big data every day from which education can be really benefit. Most teachers, students and schools will be influenced by the big data analysis. And then, there are 3 high-value uses of educational big data which are critical common applications of educational big data analysis system architecture.

1) Providing services for teachers and students: It is one of the largest, most well-known big data applications in education. The focus here is to use educational data to better understand teachers and students as well as their teaching and learning behavior preferences. Schools are keen to collect educational media data, browser logs, text analysis and sensor data to comprehensively understand the teachers and students. In most cases, the general objective here is to create a prediction model.

2) Improving education quality: The computing capability of educational big data analysis makes it possible to decode the entire DNA in a few minutes so that we can find new education methods, and to better understand and predict the education mode. As everyone can benefit from the data poduced by smart watches and wearable device, education big data can also help students learn better. The future education will not be limited to the small sample, but service for each of the teachers and students. Educational big data technology has been used to monitor pre-school, primary and secondary school students. By recording and analyzing students' behavior, teachers are now able to predict any learning situation. In this way, teachers will be able to timely correct their bad learning habits.

3) Improving safety: Educational big data is widely used to improve school safety. The U.S. National Security Agency use big data analysis to fight terrorism activities, and even to monitor our life. While the America education administration and school use big data technology to detect and prevent education network attacks, and use educational big data tools to prevent students or juvenile delinquency, detect student IC card fraud and sex trading.

The 3 uses above are the most common application fields of educational big data. Of course, with the growing popularity of educational big data analysis tools, these fields will achieve amazing results.

IV. COMPARISON OF MAINSTREAM ARCHITECTURE OF EDUCATION BIG DATA ANALYSIS SYSTEM

Architecture of big data analysis system focused on different industries and different technical fields in market, using their own technical strength and strategic analysis capabilities to open up their respective areas of information analysis. The paper mainly analyzes the technical structure, core technology and characteristic function of the education big data analysis system, and summarizes the education big data analysis and application which can learn from.

A. *Comparison of Technical Structure*

1) Hadoop. Hadoop is a software framework which can distributed processing of large amounts of data. Hadoop project consists of three parts, namely Hadoop Distributed File System (HDFS), Hadoop Map Reduce programming model and Hadoop Common. Figure 2 below is the latest technology architecture of Hadoop system. The use of Hadoop is about IBM, Alibaba, BAT, Facebook and so on.

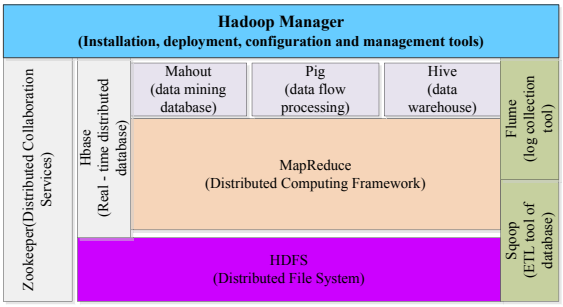

Figure 2. Technology architecture of Hadoop system [9].

2) Samza. In Samza, the data stream is cut apart, and each part consists of a sequence of read-only messages, each of which has a specific ID (offset). This system also supports batch processing, that is, successive processing of the same data flow partition multiple messages. When Samza processes a data stream, it processes each received message one at a time. Samza's stream unit is neither a tuple nor a Dstream, but rather each message. Samza's implementation and data flow modules are pluggable, although Samza's

features is Yarn (another resource scheduler) which rely on Hadoop and Apache Kafka. Because Samza places the storage and processing on the same machine, it does not load additional memory while maintaining processing efficiency. This framework provides a flexible pluggable API: you can accord to your choice on its default execution, message delivery and storage engine operations which can be replaced at any time. Figure 3 below is Samza technology core architecture. The use of Samza is about Metamarkets, Intuit, Linked In and so on.

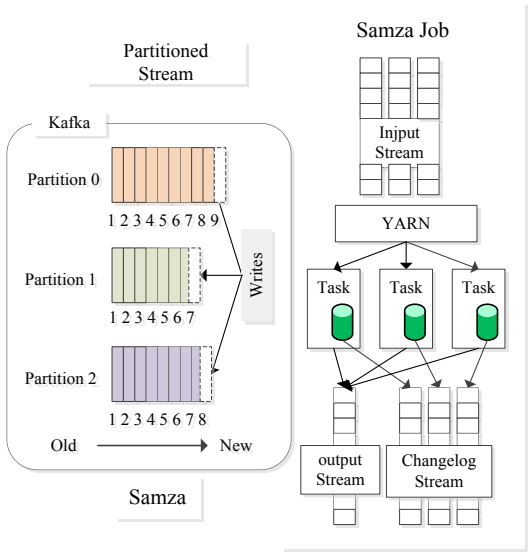

Figure 3. Core architecture of Samza technology.

3) Spark. Spark is an open source clustering computing system based on memory computing, the purpose is to analyze the data more quickly. Spark has formed a mature ecosystem, as shown in Figure 4.

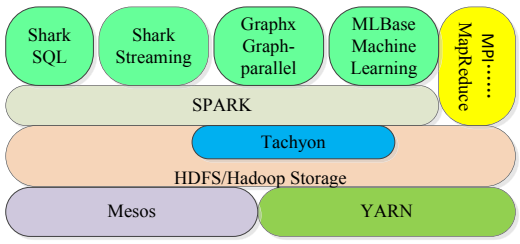

Figure 4. technology architecture of Spark system.

Spark ecosystem uses Spark as the core, based on RDD, which can build a large memory system based on educational computing, providing all-in-one data processing solutions. The use of Spark is about NASA JPL，eBay，Yahoo and so on.

B. *Comparison of Core Technology*

Architecture of big data analysis system only develop the core technology about collection, analysis, service and security control of education big data, so that we can obtain useful educational information from the massive educational data sources. The core technology of three major mainstream education data analysis systems is shown in Table 2.

TABLE II. COMPARISON OF THE CORE TECHNOLOGY OF THREE MAJOR MAINSTREAM EDUCATION DATA ANALYSIS SYSTEMS

Mainstream systems	Core technology
Hadoop	Map Reduce, HDFS and HBase correspond the most core technology of Google, which are Map Reduce, GFS and Bigtable. Big data analysis of the originator of actual combat.
Samza	Kafka, Resource management, Performance monitoring, High fault tolerance.
Spark	Resilient Distributed Dataset (RDD), RDD is the most basic abstraction of Spark, is the abstract use of distributed memory, Implementation of an abstract implementation of a distributed data set in a way that operates local collections.

C. *Comparison of Characteristic Function*

The pertinence of education big data analysis systems are relatively strong, Hadoop and Spark focus on batch processing, Samza focus on stream processing. Special features refers to the places that data analysis has some innovative and prominent features to improve, in addition to the core technology of education big data analysis system architecture. These features are the further development and improvement of the educational data analysis system, which can enhance the ease of use of data analysis, including the integrity of data, the ease of operation of big data analysis, timeliness of information services and so on. The characteristic function of three major mainstream education data analysis systems is shown in Table 3.

TABLE III. COMPARISON OF THE CHARACTERISTIC FUNCTION OF THREE MAJOR MAINSTREAM EDUCATION DATA ANALYSIS SYSTEMS

Mainstream systems	characteristic function
Hadoop	①high efficiency：By distributing the data, the Hadoop can process them on the whole nodes parallel, which makes processing very fast.② low cost： can distribute and process data through a server farm of common machine. Those server clusters total up to thousands of nodes. ③ Capacity expansion： Can reliably store and process gigabytes of data.
Samza	Samza's stream unit is neither a tuple nor a Streams, but rather each message. Samza's features are Yarn (another resource scheduler) which rely on Hadoop and Apache Kafka.
Spark	①For performance requirements of the table, to provide distributed Cache system will table data in advance Cache to memory, the subsequent query will directly access the memory data, no longer need disk overhead.②As online services to perform tasks, to avoiding task process startup and destruction overhead.

V. ENLIGHTENMENT AND PROSPECT OF BIG DATA ANALYSIS IN EDUCATION

A. *Enlightenment of Big Data Analysis in Education*

The exploration of information analysis method has a new breakthrough because of the development of big data and the development of the big data analysis system.

Scholars have made innovations in educational data analysis techniques and methods, including data modeling for education big data prediction, data visualization based on educational data, and efficient and secure cloud storage system for processing big data of education.

This paper puts forward an intelligent analysis system on the basis of studying the structure of the three kinds of mainstream education data analysis system. The system architecture is shown in Figure 5.

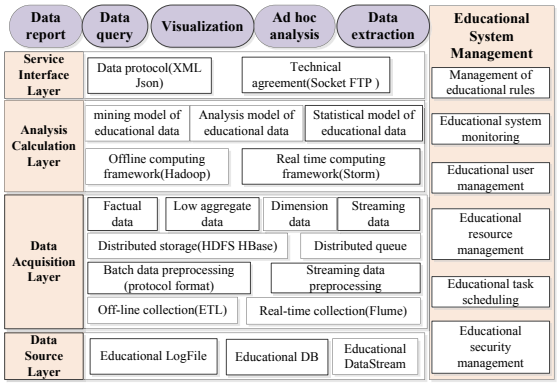

Figure 5. Technology architecture of spark system.

Educational big data tells us that education data analysis is the key to the analysis of data lies in the massive, complex, unstructured data. It is difficult to complete the analysis task within the specified time without the help of professional analysis tools, or it is difficult to find the hidden educational value in the big data in a short time.

B. Prospect of Big Dataanalysis in Education

Education big data analysis system architecture is an important component of the era of big data analysis of educational data tools, and it solves the core problem of data analysis in the analysis of educational data. With the emergence and development of big data analysis system architecture technology, we deal with massive educational data more easily, more cheaply and rapidly. The development of educational data analysis system architecture can be divided into five directions.

1) big data acquisition in Education: The diversity of educational data lead to the differences in the quality of educational data, which seriously affects the availability of educational data.

2) big data storage: Large scale data storage, complex management, need to take into account the structured, unstructured and semi-structured data is the main problem. The distributed database in the education of big data technology is effective to solve these problems.

3) Education big data calculation: The need for diversity in education has led to the emergence of a variety of typical computing models. Batch processing of big data computing (such as Hadoop Map Reduce), query analysis and calculation (such as Hive), streaming computing (such as Storm), graph computing (such as Pregel), memory computing (such as Hana) and iterative computing (such as Ha Loop), the combination of these computational models will be an effective means.

4) Educational big data mining: With the rapid expansion of the amount of educational data, not only the depth of data analysis and mining, but also the need for automated analysis is getting higher and higher. More and more big data analysis system architecture came into being, such as the R version of the Hadoop for the education of big data mining, data mining algorithms based on the development of Map Reduce.

5) Visualization of education big data: It is helpful to help people explore and interpret complex data for decision makers to excavate the educational value of data and to promote the development of big data.

With the rapid development of education big data, the education big data analysis system will serve as the important content of the service to the various educational institutions to provide creative and valuable analysis results, and is conducive to the education management decision-making.

ACKNOWLEDGMENT

The work is supported by the outstanding doctoral dissertation foundation of Shaanxi Normal University. No: X2014YB10 and Sichuan Provincial Social Science Planning foundation "based on MOOCS mode to the minority areas in Sichuan transport high-quality educational resources mechanism research". No: SC15B067.

REFERENCES

[1] Chen Chen. Construction of big data analysis and decision support platform based on cloud computing. Library theory and practice, vol 05, pp.101-104, 2016.

[2] Li Xuelong. Overview of big data systems. Chinese Science: Information Science, vol 01 , pp.1-44, 2015.

[3] Meng Xiaofeng, Ci Xiang. Big data management: concepts, technologies and challenges. Computer research and development, vol 01, pp.146-169, 2013.

[4] Sun Hongtao, Zheng Qinhua. The core technology, application status and development trend of educational data. Journal of distance education, vol 05, pp. 41-49, 2016.

[5] Cheng Xueqi, Jin Xiaolong, Yang Jing and Xu Jun . The progress and development trend of big data technology. Science and technology review, vol 14, pp.49-59, 2016.

[6] Cheng Xueqi, Jin Xiaolong, Wang Yuanzhuo, Guo Jiafeng, Zhang tie win and Li Guojie. Big data system and analysis technology. Journal of software, vol 09, pp.1889-1908, 2014.

[7] Xia Jingbo, Wei Zekun, Fu Kai and Chen Zhen. Overview of research and application of Hadoop technology in cloud computing. Computer science, vol 09, pp.6-11+48, 2016.

[8] Fang Lulu. Recommendation system based on big dataanalysis, Beijing : Beijing University of Posts and Telecommunications, 2015, pp.67.

[9] Dong Xinhua, Li Ruixuan, Zhou Wan Wan, Wang Cong, Xue Zhengyuan and Liao Dongjie. Overview of Hadoop system performance optimization and function enhancement. Research and development of computer science, vol S2, pp. 1-15, 2013.