IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 18, 2019, accepted August 30, 2019, date of publication September 3, 2019, date of current version September 17, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2939196

Curriculum Reform in Big Data Education at Applied Technical Colleges and Universities in China

XIN LI[1], XIAOPING FAN[1,2,3], XILONG QU[1], GUANG SUN[1,2], CHEN YANG[1], BIAO ZUO[1], AND ZHIFANG LIAO[3]

[1]School of Information Technology and Management, Hunan University of Finance and Economics, Changsha 410205, China
[2]Big Data Institute of Finance and Economics, Hunan University of Finance and Economics, Changsha 410205, China
[3]School of Computer Science and Engineering, Central South University, Changsha 410012, China

Corresponding author: Xin Li (lixin@hufe.edu.cn)

This work was supported in part by the Undergraduate Research Learning and Innovation Project under Grant 2018YJCX036, in part by the University Teaching Quality Engineering Project under Grant 2018ZLGC03 (Hunan University of Finance and Economics), in part by the Hunan Provincial Higher Education Institutes Teaching Reform Project under Grant 2018XJG867, and in part by the 2018 Science Research Project of the Education Department of Hunan Province under Grant 18C0975.

ABSTRACT With the boom in data science, big data education has received increasing attention from all kinds of colleges and universities in China, and many of them are in a rush to offer big data education. This paper first analyzes the major areas of big data capability training and the Chinese market needs for various kinds of data science talent. Then, it discusses the curriculum design process for the ''Data Science & Big Data Technology'' bachelor's degree program, and summarizes some detailed approaches to improving teaching experiments. Finally, this paper proposes a graduating student profile for big data education at applied technical colleges and universities in China. The authors' main ideas include that, at the applied technical colleges and universities, a) a suitable graduating student orientation should be determined as the big data talent needs are hierarchical; b) the redesigned curriculum in big data education should provide students more practical capabilities and knowledge; c) the teaching of the existing mainstream big data technologies and tools should be significant components in the syllabi of big data education.

INDEX TERMS Applied Technical Colleges and Universities, big data education, curriculum reform.

I. INTRODUCTION

As described by Wikipedia, the term ''Big-data'' was firstly used in 1990s, but the big data education has always been a part of data science (DS) education since its early stages.

According to an incomplete list on Github [1], there are over 600 data science programs at over 200 universities around the world (however, the list contains only two Chinese universities in Hong Kong). On this list, there are 60 bachelor's degree programs in data science, which are mostly at universities in the United States, although the majority of those listed programs are the master's and certificate degree programs.

At the same time, many universities also introduced some special courses to launch their data science education, such as ''Data Science: Large-scale Advanced Data Analysis'' at the University of Florida in 2011, ''Data Science and Analytics Thought Leaders'' at the UC Berkeley in 2012, and ''Introduction to Data Science'' at Columbia University in 2013.

Early in 2005, the College of Charleston, South Carolina, started a four-year undergraduate program in predictive analytics, machine learning, and data mining. In 2014, Paul Anderson and some faculty members from this college reported their ten-year experience which demonstrated that ''education and training for big data concepts are possible and practical at the undergraduate level.'' [2] At the ACM's 2014 Special Interest Group on Computer Science Education (SIGCSE '14) in Atlanta, Paul Anderson joined two professors from other universities in a deep discussion on the future of data science education at the undergraduate level. Their consensus was that an undergraduate data science program should provide students with a solid foundation in mathematics, statistics and computation. These graduates must gain adequate skills in finding, conditioning, exploring,

The associate editor coordinating the review of this article and approving it for publication was Jon Atli Benediktsson.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see http://creativecommons.org/licenses/by/4.0/

warehousing, and modeling big data, as well as the visualization of data, to enter the new field of ''big data'' careers [3].

In September 2015, the National Institute of Standards and Technology (NIST) of the U.S. Department of Commerce published the Special Publication 1500-1 ''Big Data Interoperability Framework'', which stated the formal definition that ''data science…incorporates principles, techniques, and methods from many disciplines and domains including data cleansing, data management, analytics, visualization, engineering, and in the context of Big Data, now also includes Big Data Engineering'' [4]. This publication also summarized the skills needed in data science, including domain expertise, statistics & machine learning and engineering. The relation between data science and big data was adequately highlighted in this NIST publication.

To compose curriculum guidelines for an undergraduate data science program, a group consisting of 25 undergraduate faculties from a variety of institutions in the United States held a workshop in 2016 at the Park City Mathematics Institute (PCMI) [5]. One of their summary points was that ''a redesign of the curriculum …will provide a rich and effective series of courses to prepare graduates for a career in data science.'' The experts believed that many traditional courses should be redesigned in the interests of efficiency and potential synergies.

In recognition of the burgeoning development of big data, Chinese universities have also increased their pace in big data education since the 2010s. Some data science courses first appeared at a few of the top Chinese universities, such as Tsinghua University, Peking University and Renmin University of China. Their targets usually focused on introducing and tracing the latest developments in data science [6].

In 2016, China's 13th Five-Year National Plan for Economic and Social Development clearly proposed the ''National Big Data Strategy.'' Since then, big data has been regarded as a national strategic resource, and all kinds of actions to promote the development of big data have been fully implemented to accelerate the sharing, opening, developing and application of big data resources. In 2017, the Ministry of Education of the P.R.C approved as many as 250 colleges and universities to start a bachelor's degree program in ''Data Science and Big Data Technology''. In the following year, 220 colleges and universities applied to offer this program, which has become the ''hottest'' major in the recent years [7].

However, a BA (Bachelor of Arts), a BS (Bachelor of Science) and a BE (Bachelor of Engineering) will be conferred respectively by the University of International Business and Economics, the Peking University and the Central South University, the first three Chinese universities offering an undergraduate program of ''Data Science and Big Data Technology''. Obviously, there are several directions that the graduates can take even though the program has the same name, and the program curricula and syllabi should evidently be different.

In most Chinese applied technical colleges and universities, big data education always starts with some optional courses in a computer science major or is scattered among some specialized courses. Many teachers at these colleges and universities also lack systematic training and practical experience with big data education. In response to changing this situation, the Ministry of Education requested its Higher Education Committee of Computer Teaching Steering and Professional Guidance to join with many famous Chinese universities and some big data applications enterprises to offer some seminars on big data that are relevant to teaching research, teaching training and curriculum redesign. Every year, hundreds of computer science teachers can take these opportunities to enhance their understanding of and teaching skills in big data education.

Most of the teachers who have attended these seminars have come from applied technical colleges and universities, which make up the main body of Chinese higher education institutions. After these seminars, the teachers quickly realized that their tasks and targets regarding big data education are clearly different from those of the top universities, which are regarded as research-oriented universities. At most applied colleges and universities, the curriculum on big data education needs a serious redesign to ensure that their graduate students gain adequate professional capabilities and the skills necessary to satisfy the market needs. These teachers must also determine the necessary changes to the curriculum and finally implement such changes in their daily teaching.

Many Chinese colleges and universities are making an ongoing attempt to promote the emerging program, and many articles are being published to summarize the various experiences with big data education. For example, Chao Lemen has conducted an empirical study with the help of text analysis techniques to examine data science curricula in China and abroad; [6] Lu Xiaoguang has proposed improving the professional construction of ideas, perfecting the curriculum, and strengthening the practical teaching and teacher training to meet the growing demand for data science; [8] Zhang Zhiwei has conducted an exploratory research on the construction of a big data major, curriculum design, talent training and the construction of a professional teaching team in the context of new engineering [9]. Some articles also discuss the introduction of big data technology application to other non-CS (computer science) majors. Most of these papers always propose some brief and directional suggestions for big data education, and lack detailed instructions on how to redesign the new curriculum [2], [3], [5], [6], [8], [9].

This paper focuses on curriculum redesign of big date education at Chinese applied technical colleges and universities. It first analyzes the main areas and key content of big data application as well as the talent market needs for various aspects of big data talent and then discusses in detail the syllabi adjustments in the different areas of the big data education curriculum. The following section of the paper also introduces a teaching experiment that tries to promote students' big data knowledge, capabilities and consciousness. As a consequence, the paper summarizes the overall graduating student profile of big data education at Chinese applied technical colleges and universities.

This curriculum reform explicitly reflects the integration of data science talent market needs and the future strategic direction of these applied technical institutes. It also plays an active part in the national higher education reform in China.

II. MAIN AREAS OF BIG DATA APPLICATION CAPABILITIES

Big data has become an infrastructure of the information society, and many industries and social fields are increasingly applying big data. However, "big data education" is still a relatively new concept in China. By the end of August 2018, there were only three Chinese studies on the Chinese National Knowledge Infrastructure (CNKI) with "big data education" as the keyword, two of which were about education science, and only one of which was about data science. During the above big data education research seminars, experts from famous Chinese universities, big data application enterprises and training institutions also used different expressions for "big data education". In general, there is a minimum consensus that from the perspective of data science, the aims of big data education are to cultivate student's big data thinking and application capability. In terms of breadth, it could become the liberal education for various majors; and in terms of depth, it could also become the core of data science professional education.

As one of the targets of big data education, big data application capability involves at least the understanding, acquisition, storage, transmission, visualization, analysis and application of data. Among these, the core capability is data analysis ability [10].

- ***Understanding of big data:*** The main characteristics of big data are often summarized as "volume", "velocity", "variety" and "value". Some experts also define big data as "massive and complex data sets that cannot be stored centrally, and cannot be analyzed and processed within an acceptable time, or who's individual or limited parts present low value while the whole of data provides high value" [11]. An understanding of big data belongs to a basic data capability. It includes the understanding of certain concepts such as data format, data type and data lifecycle, and it also covers the capability to extract data according to business requirements, to analyze data dialectically and to take data as the basis for decision making. This is an essential part of data literacy or information literacy.
- ***Acquisition of big data***: The sources of big data are diverse, widely distributed and of uneven quality. To obtain raw data, it is necessary to determine the location, type and format of data sources in different business scenarios, which may involve many subjects such as physics, materials, electronics, etc. Among these, the most common way to obtain massive data from the Internet is to use Python to crawl, which is also the first thing to learn for beginners. Different types of web page data also correspond to crawler technology for targeted processing.
- ***Big data processing***: To better represent the value of big data, to ensure the quality of data and to improve the efficiency of data processing, raw data must be processed by means of, for example, data auditing, cleaning, conversion, integration, desensitization, reduction and labeling. Through these data processing activities, we can introduce a processor's design and thinking, concentrate the sparsely distributed value of big data and realize the data value-add process [12].
- ***Big data management***: In the lifecycle of big data, the processed data results need to be managed before they can be used for subsequent data analysis applications and long-term storage. In the era of big data, data management has also surpassed traditional relational databases. Some relatively new data management technologies, such as NoSQL and NewSQL, have provided different storage management modes for different types of data [11], [13].
- ***Big data analysis application***: Big data analysis is often based on open-source tools, often associated with particular subjects, and it is often significantly different from traditional data analysis. The most commonly used data analysis tool is the R language and the Python language.
- ***Big data visualization and product application***: Data visualization is an important representation of big data value. Its core aim is to clearly state the overall characteristics of big data and to clearly demonstrate the value of big data. Visualization can help users understand data and make decisions, and it can provide some visualized expression tools for extensive big data applications. To build a big data analysis tool to meet the needs of both the common market demand and the various user-defined requirements is obviously a complicated big data task. Students can learn to apply existing big data analytics products to solve practical problems, which is the basic goal of big data education.

Big data capabilities cover a wide range of complex content, with different depths of capability goals in different directions. When applied technical colleges and universities conceive of their big data capability training programs for computer majors, they must all do what they can, measure their talents, closely keep up with the application-oriented talent training goals, determine practical training plans and steps, and integrate big data capability training into daily courses. In daily teaching, especially in combination with some teaching practices, these students can deepen their understanding of theoretical knowledge, improve their ability to solve practical problems and cultivate macro perspectives on big data.

III. VARIOUS TALENT MARKET NEEDS

Chinese universities and colleges are normally classified into three main groups: researching, applied and professional. The top 5-10% of them belongs to the first group, which can attract the best Chinese students, and their targets aim at achieving innovative research.

At the same time, the applied and professional universities and colleges receive over 70% of the rest of the high school graduates. However, their strategic direction commonly focuses on professional engineers or skilled workers.

The authors work at the Hunan University of Finance and Economics (HUFE), one of the provincial-jurisdictional applied universities located in a south-central Chinese province. In the past two decades, Hunan has gradually changed from an agricultural province to an industrial province, and its GDP has leapt from ¥32 billion to ¥346 billion, slightly higher than the average growth rate of the whole P.R.C. During 2018, Hunan added nearly 800 thousand new enterprises, most of them small-medium sized businesses (SMBs). How to meet the large talent needs of these SMBs has become one of the foci of provincial government educational funds. As a result, most provincial-jurisdictional applied universities have also adjusted their strategic direction to take advantage of these abundant employment opportunities.

To promote their enterprise competitiveness in the Internet era, increasingly more SMB owners are very willing to take advantage of big data technologies. However, with a limited financial budget, SMBs can hardly support a technical team to conduct their own big data applications. They would rather employ one or two skilled data engineers, who are familiar with most of the existing big data tools and data processing schemes, to support their big data applications.

To keep pace with such a significant change in the talent market requirements, applied colleges and universities have made great efforts to adjust their strategic talent direction. These efforts are not only aim to obtain more governmental fund support but also to promote their graduates' professional adaptation and talent market competitiveness. As a result, alluding to big data education, this type of an adjustment effort focuses on curriculum design.

IV. CURRICULUM DESIGN

Prior to designing the curriculum for the big data program, the faculty of the university made a great effort to analyze the potential big data positions suitable for the graduates of the applied colleges and universities. This analysis revealed three main groups of big data industry talent needs: R&D systems engineer, applied engineer and data analyst. Furthermore, the most popular job offers were data scientist, data systems architect, data systems analyst, Hadoop engineer, data analyst, data mining engineer, data visualization engineer, and so on. According to the above analysis, the first three kinds of big data positions might be filled by the top research universities' graduates or even master's/Ph.D. graduates. However, the remaining four kinds of positions should leave more opportunities for graduates from applied colleges and universities. The job descriptions and capability requirements of the most popular big data positions are shown in Table 1.

TABLE 1. The most popular big data positions.

Positions	Description	Capabilities
Data Scientist	Customer-oriented, create products and processes with meaningful value-added services	Extract and synthesize data, statistical analysis, data insight and information mining, develop software.
Data Systems Architect	Big data platform construction, big data systems design, infrastructure	Computer systems architecture, network architecture, programming paradigm, file system, distributed parallel processing, etc.
Data Systems Analyst	Data security life cycle management, analysis and application	Artificial intelligence, machine learning, mathematical statistics, matrix computing, optimization methods
Hadoop Engineer	Solve big data storage problems	Hadoop, Python, Linux, etc.
Data Analyst	Extract, analyze, and present data to make business sense of the data	SPSS, STATISTIC, EViews, SAS, big data Magic Mirror and other data analysis software
Data Mining Engineer	Find patterns in big data	Linear algebra, higher algebra, convex optimization, probability theory, Hadoop & MapReduce, Python & Spark.
Data Visualization Engineer	Design the visualization scheme that meets the business requirements	Choose the right visualization technique, make and promote visual samples, sample componentization.

Through the analysis of big data industry, the university found that the new curriculum should have at least four core components or layers: a data resource layer, a basic capability layer, an analysis and display layer and an application layer, which indicate the step by step path of data value promotion. Each layer requires different data capabilities associated with certain computer courses. To start the big data education, the university chose to design a new curriculum based on these four core components of the big data industry, and the determination of all core courses also revolved around these components. The correspondence between data capabilities and related courses is shown in Table 2.

In the above four components, the data resource and basic capability layers are more dependent on traditional computer science courses, the analysis and display layer is based on an innovative combination of statistics and IT, and the application layer reflects the special characteristics and professional background of each university or college.

TABLE 2. The correspondence between data capabilities and related courses.

Components	Capabilities	Related Courses
Data resource layer	Acquisition and exchange of raw data	Information resource management, software architecture, computer interface technology.
Basic capability layer	Data storage, data processing, and databases	Programming language, program design, data structure, database principles.
Analysis & display layer	AI, statistical applications, DM, ML, visualization and data analysis	Distributed systems, artificial intelligence, applied statistics, multivariate statistical analysis, machine learning and data mining, data visualization, cloud computing, big data analysis.
Application layer	Data application combined with major characteristics and university background	Non-computer core major courses.

However, the final objective of curriculum design should guarantee the professional competitiveness of graduate students. As far as Hunan University of Finance and Economics is concerned, its traditional advanced majors include accounting, economics and finance, and most of its graduates work in banks or the finance departments of the government and enterprises. This typical finance background of the university has to be taken into account when a new curriculum is designed.

In 2016, Hunan University of Finance and Economics started its initial preparatory work for a bachelor's degree program in data science and big data technology. The Financial Big data Research Institute was swiftly set up and its faculty members focused on designing this new program. Through their efforts, the big data major curriculum was designed on a firm foundation of data science fundamentals, with the flexibility of diverse applied disciplines that can support students in addressing the talent needs of local enterprises. Two years later, this program was officially approved by the Ministry of Education of the P.R.C.

V. COURSE ANALYSIS OF A BIG DATA CURRICULUM

Data science's prime target is to obtain knowledge from raw data; thus, the big data program should be a typical engineering major. Like all engineering programs, this 4-year big data bachelor's degree program must provide not only a solid foundation of science and engineering but also an efficient communication capability and an in-depth understanding of the environmental, cultural, economic and social effects of engineering.

In the beginning of the curriculum redesign, a few guidelines for the determination of the courses were adopted as follows:

- Some foundational courses should be assigned in the first year. The curriculum must cover mathematics, statistics and data science; thus, foundational courses, such as "Higher Mathematics", "Introduction to Data Science", "Data Structure" and some basic programming courses should be assigned in the first two terms.
- The practical content should be assigned as early as possible. Based on the first-year program training, the "Python programming" should be an important computer course in the second year, along with "Data Analysis and Mining", to provide more opportunities for the students to practice their data processing skills early.
- The courses in last two years should be some combination of the previous courses. For example, "Python for Finance: Analysis of Big Financial Data" in the third year, would synthesize knowledge from several prerequisite courses, such as the Python programming course, Statistics, Data Analysis and Finance. However, it would also be an opportunity for the senior students to consolidate their previous study.

Given the financial focus of the Hunan University of Finance and Economics, this program must also cover necessary financial and economic knowledge. As a result of all these influencing factors, the curriculum of this program must have at least four course groups: a liberal education group, an engineering basic group, a major compulsory group and an optional course group. The first two groups should be mainly taken in first two years of the program, which can be thought of as the common and the foundation years, as there are many foundational courses shared with many other engineering disciplines. The curriculum of big data program is shown in Table 3 [14].

Compared with the similar curriculum of the computer science major, this curriculum makes four kinds of changes:

- ***Deleted Courses***: Some of the typical basic courses in computer science, especially those concerning computer hardware or the bottom layer of computer systems, such as "Digital Circuits and Logical Design", "Assembly Language Programming", "Single-chip Microcontroller and Interface Technology", "Embedded Systems", "Compiler Construction Principles", etc.
- ***Added Courses***: Some financial courses and big data courses have been introduced, such as "Accounting", "Finance", "Econometrics", "Python for Finance: Analysis of Big Financial Data", "R Language Programming", "Scientific Programming and Simulation Using R", "Data Visualization", "Applied Cloud Computing", etc.
- ***Moved Courses***: Some of the similar courses have been moved from the engineering basic group or the major compulsory group to the optional course group,

TABLE 3. The curriculum of the big data program.

FIRST YEAR	1	09015141-- Moral Cultivation and Fundamentals of Law
	2	09015271-- Outline of Chinese Modern History
	3	09015251-- Situation and Policy (I)
	4	09015221-- Situation and Policy (II)
	5	12020091-- Military Theory Course
	6	13043321-- Career and Development Planning
	7	6021111 -- College English (I)
	8	6021051 -- College English (II)
	9	11005461-- College Physical Education (I)
	10	11005431-- College Physical Education (II)
	11	8013081 -- College Chinese
	12	12010021-- Mental Health Education of College Students (I)
	13	12010011-- Mental Health Education of College Students (II)
	14	10023541-- Higher Mathematics (I)
	15	10023581-- Higher Mathematics (II)
	16	02033533-- Finance
	17	04036013-- Applied Computer Technology
	18	04035483-- Advanced Language Programming
	19	04030033-- Introduction to Data Science
	20	04030093-- Data Structure
	21	04035924-- Object-Oriented Programming
	22	04030064-- Computer Composition Principles
SECOND YEAR	23	09025331-- Introduction to the Basic Principles of Marxism
	24	09025381-- Introduction to Mao Zedong Thought
	25	09015231-- Situation and Policy(Ⅲ)
	26	09015241-- Situation and Policy(Ⅳ)
	27	13013321-- Entrepreneurial Foundation
	28	11005441-- College Physical Education (III)
	29	11005451-- College Physical Education (IV)
	30	01010113-- Accounting
	31	10023523-- Probability and Statistics
	32	10023783-- Linear Algebra
	33	04030043-- Discrete Mathematics
	34	04030023-- Database Principles
	35	04030514-- Python Language Programming
	36	04030524-- Data Analysis and Mining
	37	04030515-- Web Programming
	38	04030525-- Analyze Financial and Economic Data
	39	04031505-- Operating System Principles
	40	04030095-- SPSS Data Statistics and Analysis
	41	04030555-- Introduction to Software Engineering
THIRD YEAR	42	02022763-- Econometrics
	43	04030034-- Computer Networks and Applications
	44	04030534-- Data Visualization
	45	04030544-- Applied Cloud Computing
	46	04030554-- Artificial Intelligence
	47	03015185-- Management Science
	48	04030035-- Data Science Professional English
	49	04030055-- Computing Methods
	50	04030505-- R Language Programming
	51	04030145-- Algorithm Design and Analysis
	52	04030025-- Linux Operating System
	53	04030545-- Python for Finance: Analysis of Big Financial Data

TABLE 3. *(continued.)* The curriculum of the big data program.

FOURTH YEAR	54	04030565-- Scientific Programming and Simulation Using R
	55	04030115-- Introduction to Information Security
	56	04030135-- Economics Science Practice
	57	04030085-- Internet of Things
	58	00000046-- Professional Practice
	59	00000006-- Practice Outside the Campus
	60	00000036-- Graduation Thesis
	61	04030536-- Financial and Economic Data Science Capstone

which means that they have been ''abbreviated'' or ''weakened'', such as ''Operating System Principles'', ''Software Engineering'', etc. On the other hand, there have also been some opposite moves, which means that the courses have been ''emphasized'' or ''intensified'', such as ''Database Principles'', ''Artificial Intelligence'', ''Data Analysis and Mining'', etc.

- ***Replaced Courses***: For example, ''Java Programming'' has been replaced with ''Python Language Programming'', ''Introduction to Computer Science'' with ''Introduction to Data Science'', ''Mobile Internet'' with ''Web Programming'', etc.

These changes greatly highlight the big data characteristics of this new curriculum in terms of courses. However, even with the same or a similar course name, the courses in the big data curriculum should have quite different syllabi from the courses in computer science. More detailed adjustments of syllabi will be described in the next section.

VI. SYLLABI ADJUSTMENTS TO BIG DATA COURSES

As noted previously, an in-depth change must be introduced into the syllabus to achieve course redesign, even without any changes to the course name. An important purpose of the adjustments is to integrate practical knowledge and capability training into the courses as early as possible.

Referring to the big data capability training programs of other colleges and universities, the authors have categorized and summarized the basic knowledge and practice skills of the core big data courses that must be mastered. The learning content of big data courses is so rich that it has to be properly selected based on different application scenarios. Based on the graduate profile of the big data major at the applied colleges and universities, the authors have summarized the three main modules and teaching requirements of big data education, as shown in Table 4.

- ***Systems Architecture***:

This teaching module mainly includes the operating system and distributed server cluster technology, big data processing platforms and their important components, and databases and data warehouse concepts.

Although the ''Operation System Principles'' is a typical course in this module, another Linux course has been added, and their syllabi involve increasingly more practical content. It is well-known that Linux is not only the operating system platform for many big data projects but that it also

TABLE 4. Main modules and teaching requirements of big data education.

Module	Teaching Content	Requirement
Systems Architecture	Linux: system setup and configuration management Hadoop: big data processing platform HDFS: distributed file system MapReduce: computing framework Hive: core data warehouse	Ability to set up and manage a big data processing system which can run smoothly, and an understanding of data storage and data location
Data Processing & Analysis	Data warehouse and data model design Hive or Pig tools Apache Flume or Python crawler design Data cleaning, conversion, clustering	Understand the main processes of offline data mining: data acquisition and export, processing and storage, data model design, business index design
Data Application	Python tool class libraries such as Matplotlib and Pillow JavaScript HighCharts and G6 visual library Leaflet to support the class library for secondary development	Visualization of big data analysis results

supports the operation of most big data software. As a result, a solid grasp of Linux belongs to the basic competency of big data capability. Thus, the new syllabus should include systems installation, network configuration, process management, and tools installation, and it should also include comprehension of load balancing and high reliability, and the cluster concepts and skills involved in building service architecture with high concurrency and reliability. At the applied colleges and universities, the Linux courses should introduce more content about network management and shell programming [15].

Up until now, Hadoop has been the most popular big data processing platform. Its main components include the HDFS distributed file system, the MapReduce distributed offline computing framework, and the Yarn resource scheduling platform [15]. In terms of HDFS, students should understand its system architecture and write/read the data streams and some shell operations, which includes basic daily operations, the common commands and parameters of command-line client, and even Java application development in HDFS. For MapReduce, the most important content includes its working principles, distributed applications development, MapReduce programming specifications, operating mode, debugging methods, operations processes, concurrency mechanisms and other core mechanisms that can provide flexibility for various complex applications. Hive is the core data warehouse in the Hadoop ecosystem, and its systems architecture, data storage mechanisms, configuration and installation, and its computing execution mechanism and basic operations should be well-understood. In the Hadoop ecosystem, all Hive data are stored in the HDFS, and the users can use SQL-like HQL statements for data querying or to directly execute the underlying MapReduce program [15].

- ***Data Processing & Analysis***

This module covers several main processes of offline data mining: collection, export, processing, storage, data modeling, business index design, etc. In this module, students should initially master the following abilities through practice:

1) To design the basic architecture of a big data analysis system according to specific business scenarios;

2) To select the appropriate technology for each unit based on the characteristics of a specific project;

3) To design a simple data warehouse model and architecture; and

4) To realize the respective basic function of each data analysis unit.

In fact, the systems architecture of many big data analysis projects usually adopts a similar common pattern. Therefore, during the design of the systems architecture, students can refer to existing systems, and focus on modeling the design of the data warehouse and the module design of data acquisition and processing. In this case, students should have an in-depth comprehension of the core concepts and data warehouse architecture, and they should understand the steps in data warehouse establishment and data ETL (Extraction, Transformation, Loading) and master the definition, common types, design methods and architecture selection of data modeling. All these requirements must be taken into account when the syllabus of this module is being designed.

However, this module also covers the knowledge of many existing courses, such as "Database Principles". All practical activities in this module should integrate some various contents of the different courses. Practice sessions are good opportunities of using tools such as Hive or Pig to query semi-structured data on the Hadoop platform and to try to convert external data into a specified format and load it onto HDFS in the Hadoop cluster. In data processing training, students can be asked to write the MapReduce program by themselves first and then to learn to use Hive HQL statement. It is helpful to realize the convenience of efficient data processing and the leap in working mode that can be brought about by choosing the right tools [16].

In the data collection unit, it is necessary to learn to select different types of off-the-shelf tools and to compile crawler tools to capture data. For example, when collecting log data such as access logs and website events, students can use the Apache Flume to collect log data directly from various web servers and to store it in centralized storage such as HDFS or HBase. While collecting financial data, Tushare, a free and open source Python financial data interface package, can be used to complete the process of collecting, cleaning and processing financial data, such as macroeconomic data, investment reference data, stock classification data, interbank lending data, etc.

- ***Data Application***

Data presentation is the visualization of big data analysis results and the value embodiment of the big data system. Its teaching content should cover the concepts and classifications of data visualization and introduce the underlying technical specifications of visualization in the network environment and the popular visualization class library and other technical tools.

This module can also adopt the escalation learning flow. First, students can learn to use the Python class library, such as Matplotlib and the image-processing module Pillow, to practice data graphical representation programming. After that, they can extensively learn and test other kinds of important image libraries, such as Snap, SVG, Raphael that is a JavaScript based on SVG technology, Zrender that is based on Canvas, and the Three.js and SceneJS that are based on WebGL, to improve data presentation [17].

Many data visualization libraries are written in JavaScript. According to different application scenarios, students can be guided to understand the corresponding common chart library. For example, statistics class data visualization can introduce HighCharts which is currently the most widely used visualization class library with a low threshold and good compatibility; Relationship class data visualization can introduce G6, with a simple grammar, strong interactive ability and high usability and which supports multiple views. Geographical spatial data visualization can introduce Leaflet, which specifically supports mobile applications, simple functions and second development.

Data visualization involves many disciplines, including computer technology, related natural sciences, mathematical statistical analysis, computer graphics, geographic information and other disciplines. Students should be encouraged to learn from other disciplines and engage in more interdisciplinary communication, so as to make their big data presentations replete with changes and expressiveness.

VII. EXPERIMENT IN CURRICULUM REFORM

During the design of the big data curriculum, the faculty realized that more big data practical activities should be introduced even before those courses start. They made some suitable choices to give students more opportunities to exercise basic big data skills and methods as early as possible.

Since 2017, the authors and some of the students have been working together on a practical project for this university called the ''Undergraduate Research Learning and Innovative Experiment Plan'', on the topic of ''The Design and Implementation of an Online Library Based on Hadoop''. The project goal is to establish an online library management system of city-wide colleges and universities with a big data analysis function.

In the early stages of project, the teachers and students only had a preliminary understanding of big data. The design goal was mainly focused on building a cluster-based online library to solve common functions and provide services such as interschool borrowing, personalized book recommendations, borrowing rate analysis and SMS alerts for book returns, etc. The core idea still focused on the physical construction of the system, the collection of the original data and information services based on the primary data mining. The initial design goal of the project included the following aspects:

1) To put forward the overall design of interschool online libraries and to realize the distributed management of heterogeneous data sources and the diversified service management of interschool online libraries;

2) To design personalized book recommendation services based on a big data analysis;

3) To analyze the borrowing behavior, and form a recommendation list for purchasing of popular books;

4) To analyze teachers' and students' utilization of electronic materials, such as digital books, and thus provide more effective literature for scientific research;

5) To provide an SMS service such as a new book arrival and book return reminder;

6) To attempt to realize the mobile client of the online library.

In the process of this project, the authors had many opportunities in big data teaching and training to enhance their own big data capabilities. At the same time, the students who participated in the project also consciously and gradually strengthened their big data knowledge. The teachers and students jointly realized that the original project idea was biased towards digitizing the functions of a traditional library and that the project design lacked big data thinking and an in-depth mining of the massive library data. The project's major deficiencies include the following:

1) The overall design is restricted to a migration of the original system from a single server to a cluster. Although the system reliability and response speed have been improved, the data fusion of many university libraries has not been adequately considered;

2) Although the new system integrates the library data from several universities, and the data samples available for analysis have greatly increased, the data analysis function design is still based on the data from a single university;

3) The library data are important for the teaching process data as well. Their interaction with other teaching data should be considered in the teaching big data;

4) The centralized library data should be used for deeper and broader data mining, analysis and utilization.

After a discussion between the teachers and students, the project participants decided to revise the design goals to a certain extent, to review each design goal from the perspective of big data and to better integrate the big data thinking into the target system.

In this project, based on the characteristics of the existing data and the needs of the business data analysis, students learned to reconstruct a data model with a clearer theme and reasonable hierarchy with the idea of a data warehouse, and to increase their acquaintance with a big data warehouse. Generally, the library data are relatively stable: their frequency of being inserted, deleted and modified is not too high. Therefore, although the Hive query response is relatively slow, the real-time requirements of the library system are still fairly easy to satisfy due to limited online users and limited book data.

In addition to learning how to select tools, students also learned how to program their own data by fetching tools in Python or other languages. In this project, students practiced using existing tools to extract the reader access records in the access log data flow of the online library servers, they learned how to write crawler in Python to grab books information from the library web page, and they then analyzed different crawler technologies to deal with different types of web pages. This practice activity could solidly improve the ability to collect data in a complicated environment and achieve a mastery of the knowledge of the Python course through a comprehensive study.

Data processing mainly involves data cleaning, conversion, clustering and other data processing operations. In this project, the data of online libraries of different colleges and universities had to be processed to improve data quality through data cleaning and had to be transferred into a unified format more suitable for analysis. Data from different sources was aggregated into several types of data according to function. These data could be handled by the students writing their own Python programs, and the students should be encouraged to find suitable Python resource packages to build the system.

Through a joint learning and efforts of the teachers and students, an innovative project design was introduced with more big data consciousness. The students on the project not only improved their programming ability but also gained a higher and broader perspective on the big data systems design and gained hands-on practical experience with many kinds of new technologies and tools, which laid a good foundation for the formation of their own big data capability.

VIII. THE GRADUATING STUDENT PROFILE OF BIG DATA EDUCATION

To promote campus-wide curriculum reform at HUFE, this university also created a graduating student profile of big data education, which covers the following areas:

- ***Ideological and moral qualities***: A spirit of unity, love of peace, diligence, courage and self-improvement; a concept of the rule of law and the moral norms of citizens, with good moral qualities and behavioral habits.
- ***Scientific and cultural qualities***: Good humanities and artistic assessment skills, aesthetic taste, and oral and written communication abilities; good foundation in the natural sciences; a global vision regarding the developments in science and technology trends; a certain understanding of the Chinese and foreign traditional culture and thought.
- ***Professional qualities***: Ability to track cutting-edge of big data technology and industry trends; a certain innovation consciousness in terms of basic research and development of big data project design and practice; an integrated use of basic data science knowledge, theories, technologies and method in big data project development; ability to write all kinds of technical documents; ability to apply the knowledge, skills and methods for big data systems to develop various solutions with reasonable judgment and selection.
- ***Innovation and entrepreneurship qualities***: A firm understanding of the scientific concepts of innovation and entrepreneurship; innovative thinking, an entrepreneurial spirit and the practical ability to innovate and engage in entrepreneurship.
- ***Physical and mental qualities***: Sound mental and physical health.

TABLE 5. The big data curriculum for the second degree.

1	04030023-- Database Principles
2	02022763-- Econometrics
3	04030093-- Data Structure
4	04030514-- Python Language Programming
5	04030034-- Computer Networks and Applications
6	04030524-- Data Analysis and Mining
7	04030534-- Data Visualization
8	04030544-- Applied Cloud Computing
9	04030554-- Artificial Intelligence
10	04030525—Analysis of Financial and Economic Data
11	04030516--Financial and Economic Data Capstone--Hadoop
12	04030526-- Financial and Economic Data Capstone--Spark
13	04030526-- Financial and Economic Data Science Capstone
14	00000036-- Graduation Project (Thesis)

To extend the big data education to the whole campus, the university also provides a minor second bachelor's degree program for the students in other majors who are interested in big data application to their own fields. This minor curriculum is a concentrated version of the above full version on big data education and centralizes all important data science courses, as shown in Table 5.

As part of the campus-wide initiative on big data education, this minor second-degree program also greatly extends the scale of big data education. Increasingly more students in other majors have a good chance of archiving certain big data application capabilities, which always means stronger graduating student competitiveness in their own areas.

IX. CONCLUSION

This paper provided an overview of the ongoing developments in the curriculum reform in big data education at the Hunan University of Finance and Economics and attempted to provide an example of an initiative for other applied colleges and universities. After a detailed discussion of the main areas of big data application and different big data talent needs, the authors proposed that their students' core professional competencies should come from adequate capabilities to analyze and solve big data problems with existing technologies and tools. In the curriculum redesign, the course organizing was fully considered with the view of supporting the varied capability requirements of the graduating students, and the correspondence of data capabilities and related courses was clearly defined. Furthermore, after introducing the new curriculum in detail, the paper made a brief comparison between the curricula of computer science and data science (big data), and highlighted the application characteristics of big data bachelor's degree program.

During the curriculum design for the new major, the authors realized that big data practical activities in teaching projects should play a crucial role in this program. As an example, this paper also introduced the development process of certain innovative student projects, which were greatly influenced by importing of big data education, which led to substantially improved project outcomes.

To date, big data education is still a relatively new education field in China. It involves a large variety of new technologies and tools. For those data science students at applied colleges and universities, applied practice should be emphasized first. On the basis of completing an efficient program practical training in big data capability, they should focus on mastering the use of mainstream tools and software and on understanding the adaptability of other tools. In general, the training objective of big data capability for students at applied colleges and universities could be defined as "to program and solve simple problems by oneself, to deal with complex problems by applying mainstream tools and to seek appropriate technical plans for complex real problems by multidirectional efforts." Both teachers and students have to work harder to achieve this goal in a limited teaching and practice time frame.

ACKNOWLEDGMENT

This research work was implemented at the School of Information Technology and Management, Hunan University of Finance and Economics. Thanks for the ongoing support from the university and the school.

REFERENCES

[1] *College & University Data Science Degrees*. Accessed: Aug. 6, 2019. [Online]. Available: http://datascience.community/colleges

[2] P. Anderson, J. Bowring, R. McCauley, G. Pothering, and C. Starr, "An undergraduate degree in data science: Curriculum and a decade of implementation experience," in *Proc. SIGCSE*, Atlanta, GA, USA, Mar. 2014, pp. 145–150.

[3] P. Anderson, J. McGuffee, and D. Uminsky, "Data science as an undergraduate degree," in *Proc. SIGCSE*, Atlanta, GA, USA, Mar. 2014, pp. 705–706.

[4] NIST Big Data Public Working Group. (2015). *Big Data Interoperability Framework*. [Online]. Available: http://dx.doi.org/10.6028/NIST.SP.1500-1

[5] R. D. De Veaux, M. Agarwal, M. Averett, B. S. Baumer, A. Bray, T. C. Bressoud, L. Bryant, L. Z. Cheng, A. Francis, R. Gould, and A. Y. Kim, "Curriculum guidelines for undergraduate programs in data science," in *Proc. Annu. Rev. Statist. Appl.*, vol. 4, 2017, pp. 2.1–2.16. doi: 10.1146/annurev-statistics-060116-053930.

[6] C. Lemen, Y. Canjun, W. Shengjie, Z. Junpeng, and X. Mengtian, "Data science curriculums around the world: An empirical study," *Data Anal. Knowl. Discovery*, vol. 6, no. 6, pp. 12–21, 2017.

[7] Minster of Education of P.R.C. (2018). *Record and Examination and Approval Results of Undergraduate Majors of Ordinary Institutions of Higher Learning in 2017*. [Online]. Available: www.moe.gov.cn/srcsite/A08/moe_1034/s4930/201803/t20180321_330874.html

[8] L. Xiaoguang, J. Le, and C. Qingsong, "Exploration on the talent training model of data science and big data technology," *J. Huaihai Inst. Technol.*, vol. 16, no. 9, pp. 132–135, 2018. doi: 10.3969/j.issn.2095-333X.2018.09.037.

[9] Z. Zhiwei, F. Aidong, C. Lin, W. Xiaoyin, and P. Zhenggao, "Exploration on construction of big data major in the context of new engineering," *J. Panzhihua Univ.*, vol. 35, no. 5, pp. 107–111, 2018.

[10] C. Zhenchong and H. Tiantian, "Data science:the demand and development of talents," *Big Data*, vol. 2, no. 5, pp. 95–106, 2016.

[11] X. Zongben, Z. Wei, L. Lei, G. Chonghui, Y. Jian, C. Mingmin, and Z. Yangyong, "The scientific principle and development prospect of data science and big data," *Sci. Technol. Develop.*, vol. 10, no. 1, pp. 66–75, 2014.

[12] X. Hao, Q. Yue, and H. Lan, "Liberal-education-oriented DataScience curriculum construction," *Comput. Educ.*, vol. 14, no. 8, pp. 158–162, 2016.

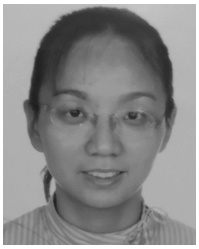
[13] L. Lirui and D. Zhonghua, "Research on data literacy education for scientific big data from the perspective of 'Internet +'," *Library*, vol. 40, no. 11, pp. 92–96, 2016.

[14] *The Curriculum of Big Data Bachelor Program*, Hunan Univ. Finance Econ., Hunan, China, 2018.

[15] Program Aha. *The Classic Learning Route of Big Data*. Accessed: Sep. 8, 2016. [Online]. Available: https://blog.csdn.net/yuexianchang/article/details/52468291

[16] Zhongqi2513. *Basic Introduction to Hive*. Accessed: Apr. 6, 2017. [Online]. Available: https://blog.csdn.net/zhongqi2513/article/details/69388239

[17] W. Chen. *Python Data Visualization Practice*. Accessed: Jul. 21, 2017. [Online]. Available: https://www.jianshu.com/p/fe4eaa20a230

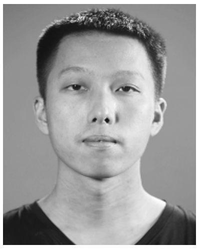


XIN LI was born in Changsha, Hunan, China, in 1969. She received the B.S. degree in computer software from the Changsha Railway Institute, in 1991, and the M.S. degree in software engineering from Central South University, in 2012. From 1991 to 1997, she was an Engineer with the Hunan Computer Factory. Since 1997, she has been a Lecturer/an Associate Professor with the School of Information Technology and Management, Hunan University of Finance and Economics. Her research interests include database systems, data mining, open-source software, big data, and so on.

XIAOPING FAN was born in Nanchang, Jiangxi, China, in 1961. He received the B.S. degree in electrical engineering from the Jiangxi University of Technology (Nanchang University), Nanchang, in 1981, the M.S. degree in traffic information engineering and control from Central South University, Changsha, in 1984, and the Ph.D. degrees in control science and engineering from the South China University of Technology, Guangzhou, and The Hong Kong Polytechnic University, Hong Kong, in 1998. He was an Assistant Professor with the School of Information Engineering, Central South University, from 1984 to 1994, and an Associate Professor, from 1994 to 1999. From 1999 to 2010, he was a Professor with the School of Information Science and Engineering, Central South University. Since 2010, he has been a Professor with the Laboratory of Networked Systems, Hunan University of Finance and Economics. He is currently the Vice President of the Hunan University of Finance and Economics, where he is also the Director of the Finance and Economics Big Data Institute. He is the author of two books and over 300 journals and conference articles. He holds 15 inventions. His research interests include wireless sensor networks, big data, data mining, and intelligent transportation systems. He is also an Associate Editor of the *Systems Engineering* journal.

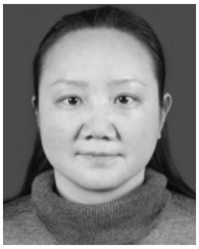
XILONG QU was born in Shaoyang, Hunan, China, in 1978. He received the Ph.D. degree from Southwest Jiaotong University, China, in 2006. He completed his Postdoctoral study at the Post-Doctoral Research Center of Computer Science and Technology, South China University of Technology. He is currently the Dean of the School of Information Technology and Management, Hunan University of Finance and Economics, China. He is also the master's degree Supervisor of Xiangtan University and the Hunan Institute of Engineering. He is the author of three books and over 30 journals and conference articles. He holds five inventions. His research interests include manufacturing informatization, distributed system integration technology, and information security.

GUANG SUN received the Ph.D. degree in computer science from Hunan University, in 2012. He is currently the Duty Director of the Finance and Economics Big Data Institute and the Duty Dean of the School of Information Technology and Management, Hunan University of Finance and Economics. He is also a Professor with the Big Data Institute. He is also a Visiting Scholar with The University of Alabama. His research interests include sensor networks security, information hiding (with a focus on software watermarking and software birth marking), big data analysis, and visualization.



CHEN YANG was born in Hengyang, Hunan, China, in 1999. Since 2017, he has been studying in computer science and technology at the Hunan University of Finance and Economics. From 2018 to 2019, he was a Student Assistant with the Innovation Training Laboratory. He has also studied with the ACM Team at the Hunan University of Finance and Economics. His research interests include C and Python programing, algorithm application, data visualization, and Web crawling in Python. He received the Third Prize in the Hunan LanQiaoBei University Student Programming Competition, in 2019, and a Third Prize in the Financial Big Data Competition at the Hunan University of Finance and Economic.

BIAO ZUO was born in Changsha, Hunan, China, in 2000. Since 2017, he has been studying in computer science and technology at the Hunan University of Finance and Economics. From 2018 to 2019, he was a Student Assistant with the Innovation Training Laboratory, Hunan University of Finance and Economics. He was in-charge of one of the projects in the University Student Innovation and Entrepreneurship Competition. He has entered the university's ACM Team. His research interests include programming, data acquisition, and data analysis and visualization. He is passionate about using Python. He received the Second Prize in the 2019' Student Programming Competition at the Hunan University of Finance and Economics.



ZHIFANG LIAO was born in Changsha, Hunan, China, in 1968. She received the B.S. degree in industry control engineering and the M.S. degree in computer science from the Changsha Railway Institute, in 1990 and 1998, respectively, and the Ph.D. degree in computer technology and application from Central South University, in 2008, where she has been a Lecturer/an Associate Professor, since 1997. From 1990 to 1997, she was an Engineer with the Hunan Computer Factory. Her research interests include open-source software, open-source software ecosystems, data mining, and so on.