Research on the Training Program of Big Data Talents under the Background of Industry 4.0

1st Da Shi
College of Information Science and Engineering, *Chengdu University*
Chengdu, China
shida@cdu.edu.cn

2nd Yu He
College of Information Science and Engineering, *Chengdu University*
Chengdu, China
hhebbbassyuu@gmail.com

3rd Xi Yu
College of Information Science and Engineering, *Chengdu University*
Chengdu, China
yuxi@cdu.edu.cn

4th Lei Mou
College of Foreign Languages and Culture, *Chengdu University*
Chengdu, China
correspoding e-mail:
mulei@cdu.edu.cn

Abstract—**This paper uses text mining technology, such as data analysis and data mining, to mine and analyze job recruitment information for the talents in the field of Big Data. These source data set is got from mainstream recruitment websites in China. Through modeling and analyzing the market demand for Big Data talents, the capacity demand structure of Big Data talents was built. It is then compared with the ability structure of Big Data talents cultivated by Chinese universities. It is found that the existing Big Data talents training system can not meet the rapidly developing market demands. The curriculum applied in training Big Data talents in universities needs to be adjusted, in order to adapt to and promote the development of the Big Data industry, as well as to be prepared for the upcoming Industry 4.0 era. In the end of this paper, the opinions and suggestions on the revision of the curriculum for Big Data professionals in Chinese universities are given, which might contributes to the improvement of the quality of Big Data talents and the development of Big Data industry.**

Keywords—Big Data talents, capacity needs, training program

I. INTRODUCTION

With the gradual approach of the industrial 4.0 era, industrial Big Data environments are gradually taking shape under the impetus of new technologies such as the Internet of Things, industrial Internet, and artificial intelligence. Data has been transformed from a by-product in the manufacturing process to a strategic resource, which has received widespread attention from companies. Using Big Data technology, through the insight of data, we can predict the corresponding demand, create the value of the invisible world, solve and avoid the risks of potential problems, use the data to integrate the industrial chain and value chain, create new wealth, and promote the further development of society. Therefore, Big Data talents who master Big Data technology are increasingly becoming the focus of social talents competition. Whether there are enough numbers of Big Data talents that meet the market demand, will become an important factor affecting the country's development..

II. ANALYSIS OF MARKET DEMAND FOR BIG DATA TALENTS

Only by accurately grasping the market demand of Big Data talents in the industry 4.0 era can we cultivate high-quality talents that can better meet market demands. Since the recruitment information can reflect the talent market demand in a timely and accurate manner, this article explores the recruitment information of Big Data posts from China's carefree websites such as China Carefree, Hunting Network, and Lagou. After clearing the invalid data, a total of 5706 valid data were obtained. Then the text segmentation technology was used to extract the capability requirement feature words from the recruitment position requirements, as shown in Table 1. Referring to the relevant research literature[1,2,3], the ability of Big Data talents is divided into ten dimensions, and the response frequency RR_j of each dimension keyword is analyzed.

$$RR_j = \frac{feq_j}{\sum_{j=1}^{10} feq_j} \times 100\% \quad (1)$$

among them, feq_j is the response frequency of the j^{th} capability dimension.

Based on the response rate of formula (1), we have established a capacity requirement structure for Big Data talents that meet market needs, as shown in Figure 1.

TABLE I. CAPACITY NEEDS DIMENSIONS OF BIG DATA TALENTS

Dimension	Containing feature words
Mathematical capabilities	Mathematics, mathematical statistics, statistics, modeling, probability theory
Data mining and modeling capabilities	Association, regression, classification, clustering, svm, decision tree, Naive Bayes, NLP (Neuro-Linguistic Programming, neural network, natural language, deep learning, linear regression, logistic regression, artificial

This paper is a phased achievement of SHYFTE project, co-funded by the Erasmus+ Program of the European Union 598649-EPP-1-2018-1-FR-EPPKA2-CBHE-JP

	intelligence, data mining
Data analysis capabilities	Extraction, collection, mobile phone, data analysis, data processing, prediction, finishing, cleaning, statistics, ETL
Data analysis tool software	SPSS, SAS, Matalbe, tableabu, data analysis software, Clementine, informatica
Data processing language	Python, R, C++, C, shell, Java, Scala, MapReduce, Perl, programming, TensorFlow
Office software application	Office software, office, Word, PPT, Excel, VBA
database	Data Warehouse, Database, MongoDB, DB2, PostgreSQL, Hive, Bbase, Access, SQL, mysql, SQlserver, NoSQL, Redis, Oracle
Report writing and business analysis capabilities	Reports, pivot tables, charts, documents, copywriting, scenarios, reports, visualization, e-commerce, internet, finance, retail, business intelligence
Big Data related ability	Kafka, HDFS, Redis, Flume, YARN, Impala, Spring, Flink, SparkStreaming, Spark, SparkSQL, Hadoop, J2EE, Streaming, Kylin, ElasticSearch, Zookeeper, OLAP, kylin, Sqoop, Kettle, BI Tools, Storm, talend , Big Data processing capabilities, linux, Unix, distributed
Personal quality	Dedication, diligence, initiative, responsibility, passion, hardship, initiative, communication, writing, coordination, collaboration, expression, teamwork, thinking,

	sensitivity, innovation, insight, logical thinking, understanding, Execution, learning ability, resistance, rigor, care, pragmatism, patience, tolerance

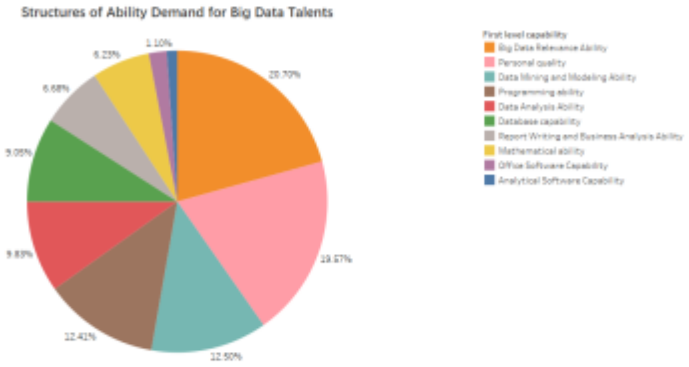

Fig. 1. Structures of Ability Demand for Big Data Talents

The analysis shows that among the capacity requirements of Big Data talents, the Big Data relevance abilities, especially Big Data parallel and distributed processing ranked the highest, accounting for 20.7% of the total capacity requirement. Secondly, the personal quality requirements of Big Data talents closely related to communication and teamwork ability account for 19.57% of total capacity demand; data mining and modeling functions and programming functions closely related to data analysis and development account for more than 10%.

If Big Data jobs are further subdivided into jobs such as data analysis, data mining, and data development, there are subtle differences in their ability requirements for Big Data talents, as shown in Figure 2. The research results show that the data development position has higher requirements for personnel. In particular, the requirements for Big Data related capabilities, database capabilities and programming capabilities are significantly higher than data analysis and data mining jobs; data mining posts require more personnel in data mining and modeling capabilities; data analysis posts require personnel Data analysis capabilities, personal communication skills, and report writing skills are more prominent.

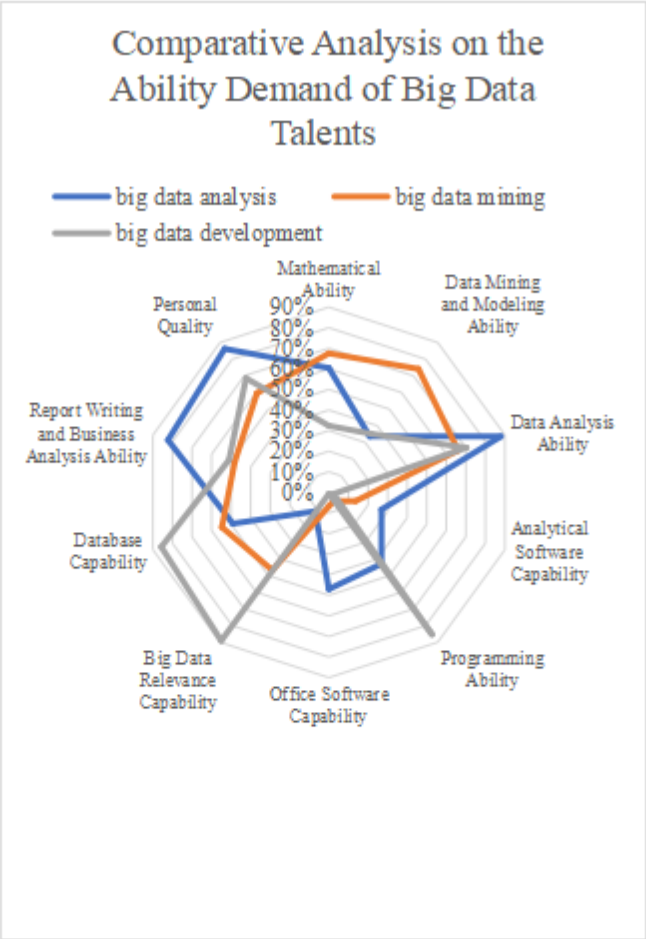

Fig. 2. Comparative Analysis on the Ability Demand of Big Data Talents

In order to help colleges and universities to grasp the ability and knowledge of Big Data talents that match the market demand more clearly and accurately, and to train the Big Data talents that meet the market demand in a timely manner, this paper further subdivides the capacity demand structure of Big Data talents. And analyze the market response rate RR_j^i for each secondary capability (knowledge) requirement.:

$$RR_j^i = \frac{freq_j^i}{\sum_{i=1}^{j} freq_j^i} \times 100\% \tag{2}$$

among them, $j = 1 \cdots 10,\ i = 1 \cdots n_j$. n_j is the j[th] number of secondary capabilities (knowledge) under the first level of capability.

According to the market response rate of formula (2), the second-level capability (knowledge) demand structure of Big Data talents in line with market demand is constructed, as shown in Table 2.

TABLE II. TWO LEVEL STRUCTURE OF ABILITY(KNOWLEDGE) DEMAND FOR BIG DATA TALENTS

First-level ability / Secondary ability or knowledge	Response Rate
Analytical Software Capability	
SPSS	36.22%
SAS	24.90%
MATLAB	21.02%
Tableau	17.86%
Big Data Relevance Ability	
Spark	32.64%
Hadoop	15.27%
Operating system	8.67%
Linux	7.64%
Distributed Architecture	6.34%
Kafka	6.05%
Storm	4.46%
HDFS	3.52%
Distributed tools	3.26%
MapReduce	3.16%
Redis	3.03%
Flume	2.65%
Flink	1.68%
ElasticSearch	1.62%
Data Analysis Ability	
Data collation	46.03%
Data analysis	37.62%
Data acquisition	11.46%
Forecast	4.88%
Data Mining and Modeling Ability	
Machine learning	32.54%
Common Mining Algorithms	31.22%
Natural language	29.49%
Artificial intelligence	6.74%
Database capability	
SQL	36.25%
Hive	30.43%
Oracle	11.05%
Redis	6.93%
Hbase	6.22%
MongoDB	3.87%
DB2	1.74%
PostgreSQL	1.33%
SparkSQL	1.26%
Cassandra	0.91%
Mathematical ability	
Mathematical foundation	38.15%
Modeling	33.39%
Probability and Statistics	28.46%
Office Software Capability	
Excel	68.55%
PPT	22.81%
Word	8.63%
Personal quality	
Communication and Writing	31.18%

Ability	
Thinking ability	27.80%
Working attitude	17.70%
Professionalism	10.94%
Collaboration	10.40%
English ability	1.97%
Programming ability	
Python	26.29%
R	24.88%
Java	24.78%
Scala	9.86%
Shell	9.23%
C++	3.37%
Perl	1.60%
Report Writing and Business Analysis Ability	
Chart	38.81%
Domain Knowledge	35.37%
Copywriting	25.82%

The analysis shows that in the Analytical Software Capability dimension, the top 3 are SPSS, SAS and MATLAB, and the response rate exceeds 20%; in the Big Data Relevance Ability dimension, the demand is prominently Spark and Hadoop; In the Data Analysis Ability Capability Demand dimension, the requirements are highlighted by Data collation and Data analysis; in the Data Mining and Modeling Ability Capability Requirements dimension, the requirements are highlighted by Machine Learning, Common Mining Algorithms, and Natural language; in the Database capability capability requirement dimension. The demand is highlighted by SQL and Hive; in the Personal quality capability requirement dimension, the requirements are prominently Communication and Writing Ability and Thinking ability; in the Programming capability capability dimension, the requirements are prominent in Python, R and Java; in Report Writing And Business Analysis dimension, it is highlighted by the order of Chart, Domain Knowledge and Copywriting.

The industry distribution of Big Data jobs were analyzed in order to understand the main domain knowledge that Big Data practitioners should have, as shown in Figure 3.

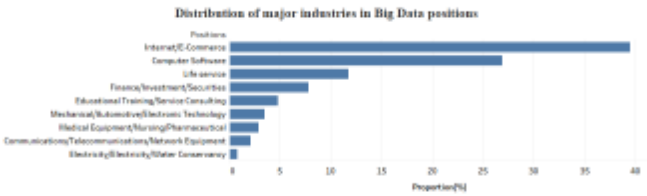

Fig. 3. Distribution of major industries in Big Data positions

As can be seen from Figure 3,the Big Data jobs are mainly distributed in the fields of Internet, e-commerce, computer software, life services and financial investment at present. In the context of Industry 4.0, Big Data jobs in industrial manufacturing are rapidly emerging.

III. Research on the Cultivation of University Big Data Talents

A. *Research on the Status of the Cultivation of Big Data Talents in Chinese Universities*

Text mining and data analysis technology are used to collect the core courses of the existing Big Data majors in Chinese universities, analyze the current situation of college Big Data talents training, and compare and analyze with market demand. The proportion of the frequency of core courses in each competency dimension in the total frequency of all core courses is taken as the training level of the ability dimension, and compared with the ability responsivity of market demand. The results are shown in Figure 4.

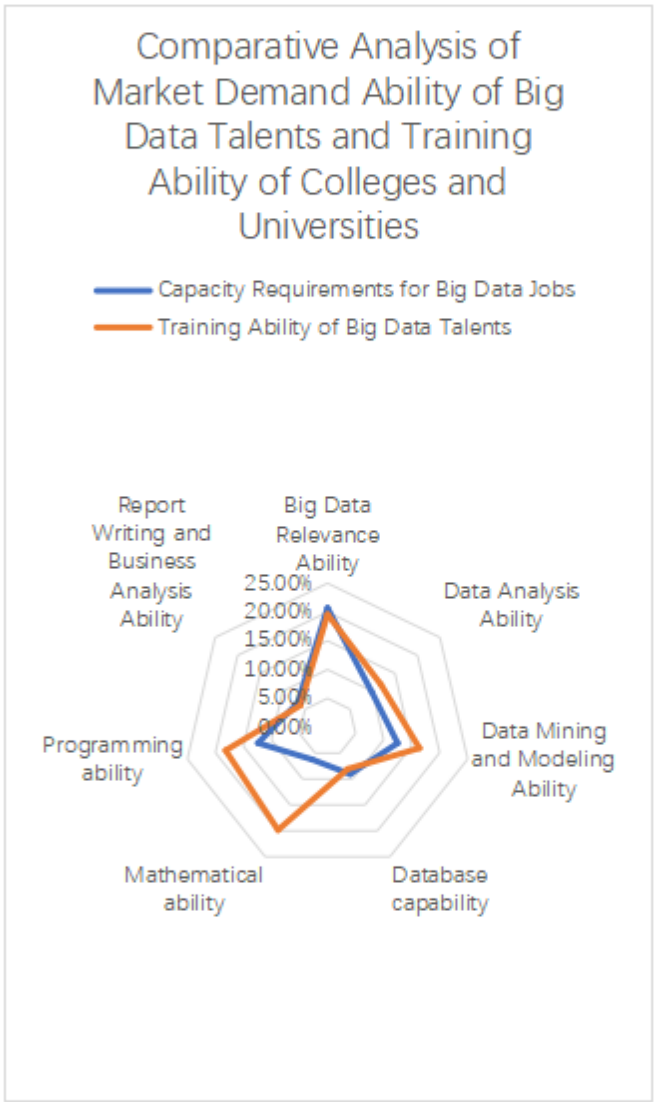

Fig. 4. Comparative Analysis of Market Demand Ability of Big Data Talents and Training Ability of Colleges and Universities

The analysis report shows that the Big Data talents cultivated by universities still have a gap with the market demand in terms of Big Data related ability, database ability and analytical report writing ability.

Hadoop accounted for only 2.82% in the Big Data capability dimension, spark accounted for 0.91%, Distributed Storage accounted for 2.17%, the Large Database and its Application accounted for 3.22% in the database capability dimension; Data Visualization accounts for 3.62%,

and domain knowledge accounts for less. Even in the overall programming capability dimension that exceeds market demand, Python only accounts for 2.82%, and R language only accounts for 0.47%.

Therefore, there are structural problems in the current curriculum of Big Data talents in Chinese universities. Computer courses are still the majority of traditional courses, while the new courses required by the Big Data industry are insufficient.

B. Suggestions for the revision of the Big Data talents training program in the context of Industry 4.0

1) Clarify the goal of talent training and adhere to the characteristics of running a school

According to the orientation of the school and the needs of the market, colleges and universities should make clear the training objectives of the major of Big Data, adhere to the accurate orientation, and run schools with characteristics.

There is a huge demand for Big Data talents in the market, but there are still differences in the ability requirements of different positions in the Big Data industry.

Colleges and Universities should ensure the category of Big Data major talents to be cultivated, regarding the objectives of the School and also the demand from industry and society. Carefully set up the corresponding curriculum system, implement the precise training of Big Data talents, and highlight the professional characteristics, to meet the requirement of the demand of market and society.

For example, undergraduate-level schools should focus on posts with relatively low data capacity requirements such as data analysis and mining, and clearly identify Big Data talents located in a certain field.

While maintaining the public foundation course for undergraduate education, universities should focus on the establishment of data analysis and mining courses, and at the same time open industry knowledge in a certain field to enhance the overall quality of students to enable students to have the computer and data capabilities required by Big Data talents, as well as domain knowledge in an industry, enhancing the competitiveness of students in Big Data processing in the industry.

At the graduate level of Big Data, universities can focus on training objectives on Big Data development positions with high ability requirements, and focus on cultivating students to become senior staff in data industries such as data scientists and Big Data project managers.

2) Keep up with market demand and optimize the curriculum system

According to the ability matching analysis of Big Data talents, in the current curriculum system of Big Data specialty, the opening rate of new curriculum related to Big Data is very low, and it is far from meeting the needs of the market. Therefore, the curriculum system of Big Data specialty in Colleges and universities needs to be adjusted vigorously to reduce the number of traditional computer courses, add new courses related to python, R, hadoop, hive, spark and other data have been added to strengthen the cultivation of the core competence of unstructured data processing, large data storage and other data.

With the advent of industrial 4.0 era, the high integration of manufacturing industry and Big Data, the continuous development of advanced productivity, the market will inevitably appear more new skills and knowledge of Big Data. The major of Big Data in Colleges and universities must keep abreast of the market demand and train more high-quality Big Data talents in line with the market demand.

3) Comprehensive talent training for cross-disciplinary integration

Big Data Posts basically require people to have the composite ability of knowledge in computer, mathematics and related fields, good thinking ability and communication and expression ability, good understanding ability of business objectives and project data of Big Data projects, skilled use of corresponding algorithms and models to complete the mining and analysis of Big Data, and discover the rules and values of Big Data hiding. The talent training program of Big Data specialty should fully reflect the characteristics of interdisciplinary integration. In addition to computer and mathematics courses, it should also combine the characteristics of the specialty, enhance the integration of knowledge in the field of industry, and enhance the comprehensive quality of students.

4) Integrating production, teaching and research, strengthening school-enterprise cooperation and realizing complementary advantages

The major of Big Data in China's colleges and universities has not been established for a long time. Teachers and teaching resources are relatively limited, which can not meet the training of high-quality Big Data talents [4,5,6]. Big Data companies focus on the Big Data industry and are more sensitive to the market, meanwhile, they master the new technologies and knowledge more quickly than universities, and by this, they accumulated huge amount of data during their work. If these professional strengths and rich resources of enterprises are fully used in the training of Big Data professionals, it will greatly promote the quality of training Big Data professionals in Colleges and universities. Therefore, it is necessary to strengthen school-enterprise cooperation, make full use of social forces, integrate production, education and research, integrate enterprise resources, scientific research resources and teaching resources of Big Data specialty, promote each other, virtuous circle, and improve the quality of Big Data specialty.

5) Strengthen practical teaching and improve students' practical ability

Big Data positions have higher requirements for practical skills. In the training of Big Data talents, colleges and universities must strengthen the cultivation of practical ability and provide students with sufficient opportunities for internship training. To this end, schools must increase the construction of laboratories, practice bases and practical teaching resources. In addition to the internship training in school, the designing of the teaching system should be done rationally, make full use of the holidays to practice in enterprises, and improve students' Big Data practical ability through a large number of practical training.

IV. ACKNOWLEDGMENT

The Shyfte project referred by "598649-EPP-1-2018-1-FR-EPPKA2-CBHE-JP" has been funded with support

from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

REFERENCES

[1] Yao Li, Zhu Longfei, Cui Chen. "Data science course construction and talent cultivation in the big data era," Computer era, vol. 317, no. 11, pp. 91-94+97, 2018.

[2] Shi da, Yang Jinhao, Zhang Zhiqiang, Chen Dan, and Wang Weijun, " Exploration and Practice of Data Engineering Undergraduate Personnel Training System," Journal of Chengdu University (social science), 2017, no. 1, pp. 112-117.

[3] Tan Linhai, "Research on Talents Training in Big Data Industry," The Chinese Journal of ICT, 2017, no. 10, pp. 91-94.

[4] Rao Linlin, Tao Juan, and Tao Guangcan. "Research on the Specialty

[5] Liu Guirong, Qin Chunrong, and Lin Yi, Research on Demand-Oriented Customized Training Model and Strategy for Big Data Talents. The Chinese Journal of ICT in Education, 2018. No.423(12): p. 82-84.

[6] Wang Yuanzhuo, "On the Construction of Big Data Teaching System under the Background of New Engineering," China University Teaching, 2018,(12):35-42.