IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 17 December 2023, accepted 29 January 2024, date of publication 7 February 2024, date of current version 16 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3363876

RESEARCH ARTICLE

Analyzing Big Data Professionals: Cultivating Holistic Skills Through University Education and Market Demands

FANG HAN[1] **AND JIYUAN REN**[2]
[1]Education Quality Evaluation Agency, Capital University of Economics and Business, Beijing 100070, China
[2]Office of Development and Planning, Capital University of Economics and Business, Beijing 100070, China

Corresponding author: Jiyuan Ren (renjiyuan@cueb.edu.cn)

ABSTRACT This study investigates the alignment between the education of big data professionals in universities and the requirements of the labor market. Through the examination of data extracted from Chinese job advertisements, this study explores the multifaceted skill set sought by employers in the big data domain, with an emphasis on achieving a harmonious equilibrium between technical (''hard'') and soft skills. Employers expect a high level of proficiency in data processing and analysis coupled with capabilities in teamwork, communication, and leadership. The dynamic nature of big data, characterized by rapid technological advancements, underscores the importance of continuous learning. The demand for big data talent has various roles, including big data scientists, engineers, and workers, thus contributing to a diverse and competitive job market. Recognizing the potential gap between university education and the dynamic requirements of the job market, this study advocates the integration of practical skills with theoretical knowledge. It proposes that universities assert their autonomy in curriculum design while concurrently engaging with alumni and industry partners to afford students' real-world experiences. In conclusion, this study underscores the significance of cultivating a sustainable talent development system. This encourages a curriculum that adeptly balances timeliness and innovation, producing graduates armed with contemporary tools and a comprehensive understanding of data science principles. This approach fosters versatile professionals capable of effectively addressing the myriad challenges within the diverse landscapes of the industry.

INDEX TERMS Big data, skill sets, topic modeling, human resources management, job requirements, education.

I. INTRODUCTION

In recent years, the volume of data across various industries has been steadily increasing, and individuals are endeavoring to identify the optimal applications for these data. In the digital era, data resources have evolved into strategically valuable assets [1]. The research and application of big data has become a pivotal indicator of a country's strategic capabilities. In 2014, China integrated big data into the government work report and initiated an extensive talent training program for big data. The following year, the State Council issued the 'Outline of Action for Promoting the Development of Big Data,' strategically guiding the direction of big data development. In 2016, China's Ministry of Education sanctioned the establishment of Data Science and Big Data Technology specialization, marking a noteworthy milestone in the formalization of the big data talent training system. In 2022, over 40 universities in the United States offered master's programs in Big Data [2]. According to data from ShanghaiRanking, by 2023, more than 700 universities in China are projected to offer majors in data science and big data technology, conferring bachelor's degrees in science or engineering [3]. This signifies a significant increase in the number of professionals undergoing training in big data.

From a market-scale perspective, big data and its associated industries demonstrate promising prospects.

The associate editor coordinating the review of this manuscript and approving it for publication was Antonio J. R. Neves .

In November 2021, the Ministry of Industry and Information Technology released the '14th Five-Year Plan' for the development of the big data industry, forecasting that by 2025, China's big data industry will exceed a measured scale of 3 trillion yuan, with the market scale expected to continue growing steadily. According to information disclosed at the 2023 China International Big Data Industry Expo, the scale of China's big data industry reached 1.57 trillion yuan in 2022 [4]. The global big data technology market was valued at USD 309.43 billion in 2022, and the global big data analytics market size was USD 254.6 billion in the same year [5], [6]. The rapid expansion of big data is evident in the continuous generation of massive data resources, reaching approximately 328.77 million terabytes of files daily. Several studies predict that the volume of digital data will reach 150-200 zettabytes by 2025 [2], [7].

However, despite the optimistic market outlook for big data, there is a shortage of big data talent. According to a survey by Anaconda, 63% of respondents on the commercial track expressed at least moderate concern about the potential impact of talent shortage on their organizations [8]. China's statistics are comparable, indicating an overall digital talent gap of approximately 25 to 30 million, which continues to widen [9]. The demand for Data Scientists has grown exponentially in recent years, prompting the development of various educational programs and training courses to address this demand [10].

Therefore, this study focuses on the technical structure of Big Data programs and the training process of Big Data professionals. The specific research questions were as follows:

Q_1: *What are the characteristics of skill demand in the field of big data?*

Q_2: *How can higher education institutions effectively cultivate talents related to big data?*

II. LITERATURE REVIEW

In this section, we provide an overview of the literature pertinent to this study. Section II-A elucidates the multifaceted nature of Big Data skill sets, delineating discrete professional roles within the realm of big data. Section II-B delves into the paradoxical situation of the heightened demand for big data talent, which necessitates industry experience, while also addressing criticisms of the perceived misalignment between university education and the evolving needs of the business sector. Section II-C focuses on job advertisements, introducing diverse methodologies employed to comprehend the industry's demand for big data professionals, and outlining significant trends in the big data job market.

A. BIG DATA SKILL SETS

Technical "big data" skills pertain to the expertise necessary for utilizing emerging technological methods to derive valuable insights from vast quantities of data [11]. In the context of data science related careers, "Big Data" refers to addressing relevant problems in the workplace. Previous research has focused on the conceptualization of Big Data in careers and has clarified conceptual boundaries by identifying related skill sets. Categorization takes the form of several categories.

Big data skill sets encompass a range of knowledge and skills required to work with big data. By categorizing the process of big data skills, Song and Zhu revealed that their study focused on data science education in the United States [12]. The skill sets include big data infrastructure, big data analytics lifecycle, data management skills, and behavioral disciplines.

Based on professional roles, De Mauro et al. used the LDA approach to provide a profile of the job roles belonging to each Big Data job family [13]. They identified four categories within the Big Data job family: business analysts, data scientists, big data developers, and big data engineers. This clustering divides them into two separate groups: technology-enabling professionals and business-impacting professionals.

Some studies have categorized data science skills. According to LinkedIn, Zhang et al. identified the five most essential areas of technical skill: text analysis, information visualization, statistical analysis, database management, and programming [14]. Gurcan and Cagiltay outlines ten competencies for Big Data Software Engineering and analyzed the most in-demand tools, programming languages, programming tools, databases, data warehouses, and big data tools [15].

B. BIG DATA TALENT TRAINING

Despite the shortage of big data talent, many big data-related careers require candidates to possess one–three years of industry experience [16], [17]. Many companies prefer to hire graduate students with work experience because they have practical exposure to industry projects and can quickly adapt to changing industrial needs [18].

While the industry has a strong appetite for big data talent, there are also some voices criticizing the lack of relevance of universities in training talent for big data and the disconnection between education and business practices. Belloum et al. suggested that there is a need for education to align better with industry demands and incorporate practical exposure to industry settings to develop students' social and meta competencies [10].

Big data is an emerging specialty, and talent needs in big data are characterized by systematic complexity. Gardiner et al. also confirmed the complexity and multiplicity of skills required for big data jobs, including traditional development and soft skills, highlighting the value placed on a diverse skill set [19].

Talent development for big data is not entirely negatively evaluated. Yusoff et al. found through a questionnaire that students are ready to fulfill employer requirements, and they acknowledge that the university is effectively preparing them for careers in the big data profession [20]. The courses and programs offered by the university are pertinent, aligning well with the industry demands.

C. BIG DATA JOB MARKET TRENDS

Some researchers use traditional questionnaires and interviews to understand the needs of recruiters and analyze their knowledge and abilities job seekers should possess [2], [21], [22]. The training of Big Data professionals is more industry-oriented; thus, job advertisements for Big Data professionals provide an appropriate channel to analyze the industry's requirements and expectations of Big Data professionals. Various studies have analyzed job advertisements to explore the demand for careers in the industry sector [23], [24], [25].

Text clustering is a common method in text mining technology and is widely used in many fields, including topic discovery and hotspot tracking. Some studies have extracted high-quality information through data processing and further refined concepts related to the occupation in question using LDA [15], [26], Word2vec [27], [28], BERT [28], and TF-IDF [29] to understand the dynamics of the labor market.

III. METHODS

This study employs a two-stage process to collect comprehensive information related to big data jobs. Subsequently, the gathered data were pre-processing. Each phase of this methodology is described in detail in subsequent sections.

A. DATA COLLECTION

To ensure the comprehensiveness of big data-related job information, recruitment data were collected through a two-stage process. Initially, 2641 job title keywords related to "big data" are systematically crawled on the zhipin.com website using "big data" as the keyword. Based on the job title keywords, a set of 14 job titles was constructed, encompassing terms such as big data, data analysis, algorithms, data mining, deep learning, machine learning, NLP, BI, data development, data operation, user operation, product operation, cloud computing, and distributed. In the second stage, the aforementioned 14 keywords were employed as search terms on the 51job.com website to retrieve information published between September 10 and October 10, 2021. The gathered recruitment data encompasses fields such as "job title," "education," "work location," "experience requirements," "salary level," "job description," and others, resulting in a total of 102,047 collected articles.

B. DATA PROCESSION

The acquired data must undergo initial pre-processing to ensure the validity of the information. Initially, the data were cleaned to eliminate invalid information, resulting in 90,499 pieces of valid recruitment information. In the subsequent step, the job description data are tokenized by removing and standardizing the text with special characters. The jieba segmentation and stop-word filtering methods were applied. Owing to the inclusion of technical terms in the dataset, 1,379 words, encompassing proprietary terms such as SQL, NLP, TensorFlow, etc., were manually added to the self-defined dictionary. The resulting corpus comprised 64,249 words.

The Word2Vec algorithm is a prediction-based deep learning model, essentially a single hidden layer neural network that learns semantic information from text and represents similarity in vector space [30], [31]. This proves to be an efficient algorithm compared with traditional feature digitization methods. It not only extracts semantic order information from text, but also effectively addresses the challenges of dimensionality and data sparsity resulting from large datasets. Regarding the corpus data, the "job description" text undergoes feature extraction using the Word2Vec algorithm. The Continuous Bag of Words (CBOW) model architecture was employed with a neural network layer set to a dimension of 100. The minimum word frequency was set to 100, and the window size was set to 5. The words in the corpus were transformed into numerical vectors, and Table 1 displays the vectors corresponding to 3,921 high-frequency words.

TABLE 1. Dimensional classification of high-frequency words.

No.	Word	D1	D2	D3	D4	......	D100
1	Experience	1.20	-0.48	-2.32	-1.66		-1.99
2	Familiar	-1.84	2.05	-1.91	2.18		0.67
3	Development	0.32	-0.10	-0.40	-0.05		-0.38
4	Product	-0.01	-0.10	-1.43	-0.46		-0.93
5	Operations	0.00	2.59	-1.23	-1.03		-0.13
6	Ability	-0.95	-1.93	0.14	-0.43		-2.40
7	Data	2.18	-0.72	0.99	0.90		-0.95
8	Analytics	-0.19	-1.26	1.04	1.94		-3.93
9	Projects	0.38	-1.70	-0.39	-0.77		-0.95
10	Technology	-0.15	-0.73	-0.02	-0.14		-0.42
	......						
3921	resilience	0.19	-0.19	0.38	-0.09		0.05

C. DATA ANALYSIS

Descriptive statistics and cross-tabulation methods were employed to analyze fundamental data in big data talent recruitment and to investigate the correlation among education requirements, salary levels, and cities.

K-means clustering analysis is applied to cluster the vector set of "job description" text, categorizing words of distinct natures into different groups [32]. Considering the model's performance and the interpretability of clustering outcomes, the optimal number of clusters is established at six, ensuring consistency in vocabulary within each category.

The similarity index was employed to calculate the degree of association between words indicating knowledge mastery and various skill categories. The similarity score is computed using Cosine Similarity. A higher similarity score signifies greater similarity between two words in the vector space, allowing for the measurement of semantic similarity using this approach.

$$CosineSimilarity\,(v_1, v_2) = \frac{v_1 \cdot v_2}{\|v_1\| \cdot \|v_2\|}$$

IV. RESULT

This section presents the results of this study. Section IV-A analyzes big data job descriptions, highlighting a strong

TABLE 2. Crosstable for education and city in big data recruitment.

City category	High School or Below	Junior College	Bachelor	Master	Doctor	Total
First-tier cities	1961	15163	32365	4497	363	54349
New first-tier cities	1455	10508	19117	3031	218	34329
Others	854	4736	6890	801	88	13369
Total	4270	30407	58372	8329	669	102047

TABLE 3. Major frequency of big data recruitment (Top 10).

Major	Frequency
Computer Science	15288
Mathematics	7446
Automation	7056
Telecommunication Engineering	4898
Artificial Intelligence	4785
Electronics and Information Enginnering	4396
Advertisement	4310
Finance	3892
Statistics	3207
Marketing & Sales	3121

demand for multidisciplinary skills. Section IV-B outlines the demand for expertise in big data talents, categorizing keywords into segments based on the project development process. Section IV-C investigates professional skills and qualities sought in big data job seekers, employing clustering analysis based on knowledge mastery levels.

A. BIG DATA JOB DESCREPTION

Using descriptive statistics, Tables 2 and 3 present fundamental information from job descriptions in the field of big data. Recruitment for talent in big data is predominantly concentrated in first-tier and emerging first-tier cities, with educational requirements primarily emphasizing undergraduate degrees. The preferred academic disciplines include computer science, mathematics, statistics, and various fields related to engineering and business.

Using text mining enterprise recruitment information, the demand for big data talent was categorized into six groups. Table 4 presents the high-frequency keywords, demand richness (number of keywords in each category/total number of keywords), and demand intensity (total word frequency of keywords in each category/number of keywords in each category) for each category.

The highest demand (29.76%) is for candidates with diverse professional backgrounds, indicating the need for big data talent with multidisciplinary capabilities. Employers highly seek computer science, statistics, and mathematics graduates. Many current university big data programs are designed with comprehensive training rooted in these disciplines. Additionally, there is a high word frequency in finance, e-commerce, advertising, marketing, education, and other specialties. The market demands that big data talents possess knowledge across various subject areas, emphasizing enhanced industry adaptability and evident interdisciplinary training features.

The highest demand intensity is for business content (4083.17), followed by professional knowledge (1959.19), highlighting employers' preference for big data talents with proficiency in both the professional and business domains. Specifically, understanding business content is crucial for the effective application of big data expertise in enhancing business productivity, problem-solving in the industry, and ultimately improving corporate earnings. The third-highest demand intensity (1931.49) is for comprehensive quality, indicating employers' preference for big data talent with robust overall capabilities. Big data talent must engage in the entire data processing process, not only focusing on data communication but also collaborating with colleagues and leaders. The essential skills in interpersonal interactions include teamwork, communication, coordination, and execution. Moreover, big data talent should demonstrate professionalism in the field of business. Confronted with complex business issues and vast datasets, they require strong abstract thinking, logical reasoning, data sensitivity, innovation, and the ability to extract information and create value in an intricate and dynamic data landscape.

B. BIG DATA EXPERTISE NEEDS

By concentrating on expertise category keywords, a system for the demand for expertise in big data talents was established, as shown in Table 5. Keywords in the expertise demand category were categorized into five segments based on the project development process: demand analysis, system development, data storage and governance, data mining and analysis, and data report development. As the framework or components related to big data processing permeate nearly the entire project development process, possessing distinctiveness, knowledge pertaining to big data processing is listed separately as a module. Certain technologies and tools are associated with multiple segments and modules.

In the demand analysis session, big data professionals are required to conduct thorough studies on business scenarios, gain insights into market demand, and identify business issues through market research, consumer research, competitive analysis, and other methodologies. They must translate abstract problems into data-centric challenges and formulate

TABLE 4. Big data job description text clustering results.

Categories	High-frequency Keywords	Demand Richness	Demand Intensity
Job Responsibilities in Business	Management, Media, Taobao, Publicity, Supervisor, Service, Branding, Cooperation, etc.	11.73%	4083.17
Job Responsibilities in Technology	Operations, Data analysis, Product, Optimization, Project, Design, Maintenance, etc.	20.12%	1302.15
Professional Background	Computer science, Statistics, Mathmatic, E-commerce, Finance, Automation etc.	29.76%	578.12
Professional Knowledge	Big data, Machine Learning, Data Mining, SQL, Python, Linux etc.	21.42%	1959.19
Comprehensive Qualities	Team work, Communication Skills, Coordination Skills, Resilience, Innovation, etc.	8.16%	1931.49
Welfare Benefits	Two-day weekend, Insurance and Housing Fund, Promotion, Business trip, Growth, Bonus, etc.	8.80%	674.78

TABLE 5. Professional knowledge requirement system for big data talent.

Project Development Process	Concepts and Theories	Techniques and tools	Demand Richness	Demand Intensity
Requirement Analysis	Prototype, Iteration, Lifecycle, Investigation, Competitive Product, etc.	Axure, Visio, CAD, Xmind, Flowchart, etc.	5.63%	1359.69
System Development	Framework, Architecture, Frontend, Protocol, Operating System, Server, Interface, HTML, Backend, Full Stack, etc.	Java, C++, Linux, C, Spring, Shell, C#, Springboot, JavaScript, Memcached, etc.	47.19%	1844.23
Data Storage and Governance	Database, Data Structure, Data Processing, ETL, Data Warehouse, etc.	MySQL, Oracle, SQL Server, Redis, MongoDB, etc.	9.96%	3527.57
Data Mining and Analysis	Deep Learning, Machine Learning, Modeling, Data Mining, Image Processing, Control Algorithms, Natural Language Processing, 3D Reconstruction Algorithms, Algorithm Porting, Neural Networks, etc.	Python, Matlab, TensorFlow, OpenCV, PyTorch, R, Caffe, Halcon, SPSS, SAS, etc.	21.65%	2015.76
Data Report Development	BI, Visualization	Tableau, Power BI, FineReport, PowerPoint, etc.	2.16%	1657.8
Big Data Processing Modules	Big Data, Distributed, Cloud Computing, Big Data Development, Multithreading, Concurrency, Distributed Systems, Distributed Computing, PAAS, etc.	Hadoop, Spark, Hive, Kafka, HBase, Flink, Dubbo, Scala, Elasticsearch, ZooKeeper, etc.	13.42%	2469.57

solutions accordingly. The demand for knowledge at this stage lacks richness and intensity; instead, it leans towards the necessity for experience, insight, logical thinking, and other comprehensive qualitative capabilities.

During the system development stage, big data professionals must design, develop, maintain, and optimize the system or modules in alignment with the project's specific requirements. The demand for professional knowledge at this stage is diverse and rich, encompassing a wide array of theoretical knowledge and professional skills, such as front-end, back-end, client-side, and server-side development. However, the demand intensity was low, requiring more computer-related professionals. This may not be an optimal and suitable field for big data professionals.

During the data storage and governance stages, big data professionals must employ database-related software for high-performance data storage services. Additionally, they must handle data integration, governance, and quality control tools. This entails constructing data warehouses and data marts, and performing tasks such as data extraction, processing, cleaning, and verification. While knowledge richness at this stage was low, demand intensity was the highest. The storage and governance of data forms the bedrock for all data mining, BI analysis, and upper-tier applications. Therefore, proficiency in these aspects is crucial for any role in the realm of big data.

In the data mining and analysis stage, the demand for knowledge is both rich and intense, providing a space in which data science and big data technology professionals can fully leverage their advantages. Employers' expectations of big data professionals at this stage can be categorized into two primary aspects. First, employers require big data professionals to extract valuable information from massive datasets, demonstrate data in sight, and directly apply mainstream data mining or statistical modeling methods. They should be capable of integrating this knowledge with actual business scenarios, contributing to model design and algorithm selection. Leveraging massive amounts of business data, they must mine data value and support business optimization and innovation. Second, employers seek big data professionals to optimize and develop algorithmic models. These professionals need to explore the latest cutting-edge technologies, not only completing the implementation of cutting-edge algorithms and model improvement based on actual business requirements but also innovating by developing new algorithms or statistical models and applying them in real business scenarios.

During the data report development stage, big data professionals should utilize visualization tools, integrating data characteristics and modeling methods to construct suitable data presentations. Although knowledge richness at this stage is low, with less varied theories and skills and low learning

TABLE 6. Big data skills co-occurrence matrix and clustering results.

Skills	Excellence	Proficient	Acquainted	Understand	Know	Cluster
Python	0.42	0.31	0.29	0.08	0.11	1
R	0.33	0.29	0.21	-0.01	0.09	1
Matlab	0.24	0.20	0.10	-0.06	-0.02	1
SPSS	0.24	0.21	0.14	-0.19	-0.01	1
SAS	0.23	0.19	0.11	-0.14	-0.02	1
OpenCV	0.26	0.28	0.20	0.01	0.09	2
Caffe	0.26	0.15	0.19	0.01	0.04	2
PyTorch	0.24	0.16	0.19	0.05	0.05	2
TensorFlow	0.23	0.16	0.19	0.04	0.05	2
bootstrap	0.19	0.18	0.16	-0.04	0.11	2
Keras	0.20	0.15	0.17	-0.02	0.05	2
MXNet	0.18	0.11	0.14	-0.01	0.01	2
CNN	0.32	0.36	0.34	0.23	0.31	3
RNN	0.30	0.34	0.32	0.24	0.30	3
Reinforcement Learning	0.24	0.32	0.27	0.24	0.30	3
Neural Network	0.21	0.27	0.24	0.29	0.29	3
Data Mining	0.21	0.25	0.17	0.28	0.23	3
Modeling	0.23	0.23	0.20	0.24	0.14	3
Machine Learning	0.21	0.22	0.21	0.23	0.20	3
Decision Tree	0.20	0.23	0.20	0.16	0.15	3
Clustering	0.16	0.24	0.19	0.18	0.17	3
Transfer Learning	0.17	0.21	0.20	0.17	0.18	3
Deep Learning	0.16	0.25	0.17	0.18	0.19	3
NLP	0.16	0.21	0.15	0.18	0.18	3
Feature Extraction	0.14	0.23	0.13	0.15	0.12	3
Regression	0.15	0.20	0.14	0.17	0.11	3

difficulty, effectively narrating a compelling data story and fully conveying the connotation of data requires big data professionals to possess strong logical, expressive, and aesthetic abilities, along with other comprehensive qualities.

The big data processing module encapsulates rich theoretical knowledge and professional skills, with a higher intensity of demand in the job market. At this stage, big data professionals must grasp the fundamental knowledge of the five aforementioned links. They should employ big data thinking and utilize big data-related frameworks or components to execute big data processing. The concepts, modes, and methods of big data processing differ from traditional approaches in that they present a higher learning threshold and difficulty. Notably, components such as Kala, Logtash, and Flume are typical in big data transmission, while HDFS, Redis, and HBase are components related to big data storage. Additionally, Hadoop, Storm, and Spark are commonly used distributed computing frameworks.

The job market requires big data professionals with a multidisciplinary background, specialized knowledge, business acumen, and strong comprehensive qualities. The demand for big data professionals with professional knowledge extends the content of talent training programs in colleges and universities. Throughout the various stages of data project development, big data professionals must possess the theoretical knowledge and professional skills to address diverse data and business scenarios. Clearly, there remains a disparity between the training of big data professionals in colleges and universities and the requirements of the job market.

C. BIG DATA PROFESSIONAL QUALITIES

On the job recruitment website, there is a clear expression of the professional skills and qualities of jobseekers. Referring to the method developed by Li [33], refining the degree words of knowledge mastery in the descriptions on the job recruitment website, we used the following five levels of knowledge: excellence, proficient, acquainted, understand, and know. By analyzing the various skills extracted from the recruitment website, the similarity between skills and knowledge mastery vocabulary was established, and K-means clustering analysis was used for analysis. The details are listed in Table 6.

Based on the clustering results, the following three categories are obtained:

Data Analysis Tools and Programming Skills: This category reflects recruiters' demand for professional skills in the field of data analysis. This includes skills such as Python, R, MATLAB, SPSS, and SAS. These skills are mainly used for data processing, statistical analysis, modeling, and visualization. The demand for these skills may indicate that recruiters are looking for candidates with a background in data analysis and programming to support data driven decision making and problem solving.

Deep Learning and Artificial Intelligence Frameworks: The second category covers frameworks and libraries related to deep learning, artificial intelligence, and computer vision, such as OpenCV, Caffe, PyTorch, TensorFlow, Bootstrap, Keras, and MXNet. This suggests that recruiters may look for candidates with experience in artificial intelligence and deep learning, which are often applied in areas such as image recognition, natural language processing, pattern recognition, and intelligent system development.

Machine Learning and Data Science Terms: The third category includes a wide range of machine learning, data science, and related conceptual terms such as CNN, RNN, Reinforcement Learning, Deep Learning, NLP, Decision Trees, Clustering, Transfer Learning, Feature Extraction, and Regression. This indicates that recruiters may be searching for candidates with a broad background in data science and machine learning to address various complex data analysis and prediction tasks.

These three clustering results reflect the diversity of recruiter demands for professional talent in the big data field. They may need data analysis experts, deep learning engineers, and extensive data scientists to meet the requirements of different fields and projects. These results may provide guidance for recruiting strategies and training directions to ensure that recruiters acquire talent suitable for their business needs.

V. DISCUSSION

This section presents a discussion of the study. Section V-A presents an examination of job recruitment data, highlighting a robust demand for big data professionals possessing skills that span multiple domains. Section V-B delves into the knowledge structure and boundaries of big data professionals, emphasizing the imperative of systematic skill development and distinguishing between practical skills and holistic thinking paradigms. Section V-C focuses on the establishment of a sustainable talent training system, emphasizing the pivotal role of universities in balancing timely curriculum content, promoting real experiences, and fostering interdisciplinary collaboration to effectively meet the ever-evolving demands of the market.

A. DEMAND CHARACTERISTICS OF BIG DATA PROFESSIONALS

Through in-depth mining and analysis of data on job recruitment websites, the market's demand for big data talent shows the need for skills in multiple domains. Simultaneously, employers have raised higher expectations for these professionals, requiring them to possess not only rich 'hard skills' but also excel in 'soft skills' [34]. The balance between these hard and soft skills is crucial; big data professionals need to excel in key technical skills such as data processing and analysis, while actively developing soft skills such as teamwork, communication, and leadership [35]. Hard skills enable them to handle large datasets and solve complex problems, whereas soft skills assist in efficient collaboration and communication with colleagues or upper management from different departments, leading them to deal with stress in a dynamic fast-paced environment [36]. This balance reflects the comprehensive and complex nature of work in big data.

There is a significant demand from businesses and the market for big data talent, requiring various levels and types of professionals. In recent years, the demand for big data talent has continued to rise, with graduates from various educational backgrounds entering the workforce and playing a role in stratifying big data talent. The big data field encompasses several different professional roles, including big data scientists, big data engineers, and a large number of big data workers [37], each with different skill requirements and responsibilities. Diversification of roles also contributes to a more varied market demand, forming a healthy pyramid shape that facilitates the orderly development of talent.

The technical skills related to big data overlap with those of computer-related majors, and technical characteristics are characterized by rapid updates. Technologies and methods in the big data field are continuously iterated and updated, with new tools and techniques constantly emerging. This necessitates that practitioners remain vigilant, continuously learn, and adapt to keep up with industry development. For instance, some big data skills mentioned on job recruitment websites in 2018 have already deviated from current knowledge and skills [13]. This rapid iteration requires big data professionals to stay alert and continually update their skill sets to keep pace with industrial developments.

The demand for big data talent is strong, and the field possesses a certain level of specialization. However, through the presentation of job advertisements, research has identified challenges to the big data profession from other closely related fields. In addition to big data, fields such as statistics, computer science, mathematics, and finance are gradually entering the domain of big data. This increases competition in the big data talent market and creates opportunities for cross-disciplinary collisions in the professional field. This makes the big data field more diverse and stimulates practitioners to enhance their competitiveness continually.

B. KNOWLEDGE STRUCTURE AND BOUNDARIES OF BIG DATA PROFESSIONALS

This article begins by delving into job recruitment information related to the big data profession. It extracts market demands for talent in big data-related fields through key terms in job advertisements and describes the workflow structure corresponding to big data skills. Through data analysis, this study is consistent with previous research conclusions regarding the classification of concepts related to big data. This clarification highlights the kind of big data knowledge needed by businesses in the job market. This article emphasizes the need for the systematic and modular development of big data skills, providing guidance for the establishment of knowledge systems in big data majors for higher education institutions and online education organizations.

Currently, some businesses believe that the big data profession has not effectively met the market's demand for talent, and that the talent gap persists over many years. Criticisms of higher education institutions often focus on existing education models and teaching content, claiming that they fail to meet the market's talent needs. Even from the first day of enrollment, students find that the knowledge they acquire becomes outdated, making it difficult to transition into employment and address specific problems.

This prompts reflection on the disconnect between academic and industrial knowledge applications. Although this phenomenon has been discussed and criticized in many academic fields, as a relatively new discipline, big data requires careful analysis and reflection on its skills and knowledge. Recognizing the professionalism of big data should distinguish between two concepts: big data skills and knowledge. Big data skills refer more to the application of data methods and tools and the ability to address specific needs in practical environments – qualities emphasized for big data professionals mentioned on job recruitment websites. The knowledge clustering, knowledge mapping, and competency assessments discussed pertain to the business-oriented skills required in the big data field. Professional big data skills are explicit factors showcasing the specialty of the profession and are more easily understood by external parties. Big data knowledge, on the other hand, is a holistic thinking paradigm in the field of data science that includes data science thinking, data concepts, methods, and data analysis capabilities [37], [38]. These are not easily understood and ascend to the conceptualization of abstract thinking. However, these ideas and methods should be the genes of big data professionalism, constituting the core structure and logic of the related knowledge.

Reflecting on the big data profession, universities should be independent in choosing teaching content without being unduly influenced by the market. In modern universities, the level of talent cultivation is an important factor in measuring the quality of education and is one of the universities' essential missions. However, the big data profession involves not only technology and application but also its own thoughts and methods. The operation of the big data profession should adhere to the inherent laws of higher education and not enter the process of vocational technical training prematurely. Present-day higher education is already facing some delays, and the rapid iteration in the big data industry makes it difficult for education to keep up. Therefore, clarifying the knowledge boundaries of the big data profession is essential for keeping it on the right track. In addition to imparting professional knowledge and skills, it is crucial to convey long-term, continuous knowledge such as data thinking and data methods. This helps students proficient data logic and ideas, achieve sustainable knowledge inheritance, and foster a lifelong learning attitude.

C. ESTABLISHING A SUSTAINABLE TALENT TRAINING SYSTEM FOR BIG DATA PROFESSIONALS

The cultivation of big data talent is the mission and responsibility of education institutions. Universities have multiple missions including teaching, research, and social responsibility. In the current era of digital transformation, there is an urgent need for vitality and sustainable development to promote the generation and development of big data-related majors. This ensures that talent cultivated in the big data profession meets the demands of the labor market. In China, majors related to big data have emerged in less than a decade and are primarily designed to nurture talent in relevant industries. Therefore, discussing the sustainable cultivation of big data talents from the perspective of higher education involves focusing on the dual role of universities in action: disseminating new knowledge and responding to market demands.

The curriculum for Big Data majors should balance timeliness and expansiveness. From the perspective of fundamental software tools, attention should be paid to timeliness to ensure that the most commonly used software, mastered by students, remains the mainstream choice in the market upon graduation. For example, concerning the demand for quality in big data professions, the application of Python was most closely associated with the level of Excellence (0.42). In the early establishment of big data majors, some teachers were reorganized and adjusted from fields such as computer science and statistics. During teaching, they adhered to traditional software skills and missed opportunities to update and create in this new field. Therefore, apart from the solid teaching of core tools, curriculum design and teaching should closely align with industrial demands. For instance, the development of libraries such as PyTorch, TensorFlow, Keras, and MXNet mentioned earlier has not exceeded ten years. If teachers can introduce conceptual content into the classroom in advance, it will place the cultivation of big data talents at the forefront. Additionally, as a profession closely linked to digital education, the promotion and application of various MOOC platforms supplements and supports the learning of big data professionals, meeting the demand for exposure to the latest relevant skills outside formal studies.

As an emerging discipline, Big Data majors have high requirements for internships and practical training. However, the establishment of internships, practical training, and project exercises for new majors may not be mature. Classroom learning is not as clear or effective as complete participation in project training. Classroom courses often prioritize theory over practice, and assigned homework designs tend to deviate from real-world industry scenarios. Therefore, to establish a sustainable development system for cultivating Big Data talents, students should be immersed in real cases. Schools should collaborate with alumni and establish partnerships with relevant companies, allowing students to participate in project development, case studies, and other practical activities to ensure their competence in complex big data jobs after graduation.

Big data, as a concept and method, require integration with specific domain knowledge in practical work applications [39]. Therefore, big data is well suited for interdisciplinary collaboration, integrating knowledge from various fields into big data education. Combining big data with economics, education, medicine, logistics, management, and other professions can cultivate more diverse talents adaptable to different fields. This approach can effectively promote rapid development and comprehensive promotion of the big data profession.

VI. CONCLUSION

This study provides a broader understanding of the demand for talent in the Chinese big data labor market. We clarified the market's skill requirements for big data professionals during the recruitment process and analyzed the relationship between big data skills and knowledge. Additionally, the research offers implementation methods for promoting and disseminating education in the field of big data. Therefore, this study will play a significant guiding role in future applications in the education and business sectors.

Based on the analysis of skill requirements for big data professionals, research, in conjunction with the relevant concepts of the Knowledge Pyramid, has presented a novel analytical framework [40]. This framework distinguishes the extent to which different big data skills need to be mastered during the recruitment process. This analytical approach, based on skill mastery levels, can be applied in various other fields. Simultaneously, the extracted high-demand skills can effectively monitor the changing trends in the demand for skills in the big data profession. This format also provides a solid basis for curriculum development in schools, thereby ensuring the effectiveness of future talent cultivation.

REFERENCES

[1] C. Hu, Y. Li, and X. Zheng, "Data assets, information uses, and operational efficiency," *Appl. Econ.*, vol. 54, no. 60, pp. 6887–6900, Dec. 2022, doi: 10.1080/00036846.2022.2084021.

[2] A. Xu, Y. Wu, F. Meng, S. Xu, and Y. Zhu, "Knowledge and skill sets for big data professions: Analysis of recruitment information based on the latent Dirichlet allocation model," *Amfiteatru Econ.*, vol. 24, no. 60, p. 464, May 2022, doi: 10.24818/EA/2022/60/464.

[3] Shanghai Ranking. (2023). *Ranking of Data Science and Big Data Technology*. [Online]. Available: https://www.shanghairanking.cn/rankings/bcmr/2023/080910T

[4] X. Han. (Feb. 22, 2022). *China's Big Data Industry to Reach 1.57 Trillion Yuan in 2022, Up 18% Year-on-year*. Accessed: Nov. 16, 2023. [Online]. Available: https://www.gov.cn/xinwen/2023-02/22/content_5742649.htm

[5] Fortune Bus. Insights. *Big Data Technology Market Size to Surpass USD 842.17 Billion By 2030, Exhibiting a CAGR of 13.6%*. Accessed: Nov. 17, 2023. [Online]. Available: https://www.globenewswire.com/news-release/2023/09/14/2743079/0/en/Big-Data-Technology-Market-Size-to-Surpass-USD-842-17-Billion-by-2030-exhibiting-a-CAGR-of-13-6.html

[6] Straites research. (2023). *Big Data Analytics Market Size, Growth, Trends and Forecast to 2031*. SRTE114DR. [Online]. Available: https://straitsresearch.com/report/bigdata-analytics-market

[7] R. Shewale. (2023). *65 Big Data Statistics 2023 (Facts, Trends & More)*. demandsage. [Online]. Available: https://www.demandsage.com/big-data-statistics/

[8] Anaconda. (2022). *2022 State of Data Science Report*. [Online]. Available: https://know.anaconda.com/rs/387-XNW-688/images/ANA_2022SODSReport.pdf

[9] *Renrui Human Resources and Deloitte China, Report on Research and Development of Digital Talent in Industries*, Social Science Academic, Beijing, China, 2023.

[10] A. S. Z. Belloum, S. Koulouzis, T. Wiktorski, and A. Manieri, "Bridging the demand and the offer in data science," *Concurrency Comput., Pract. Exper.*, vol. 31, no. 17, Sep. 2019, Art. no. e5200, doi: 10.1002/cpe.5200.

[11] M. Gupta and J. F. George, "Toward the development of a big data analytics capability," *Inf. Manage.*, vol. 53, no. 8, pp. 1049–1064, Dec. 2016, doi: 10.1016/j.im.2016.07.004.

[12] I. Song and Y. Zhu, "Big data and data science: What should we teach?" *Expert Syst.*, vol. 33, no. 4, pp. 364–373, Aug. 2016, doi: 10.1111/exsy.12130.

[13] A. De Mauro, M. Greco, M. Grimaldi, and P. Ritala, "Human resources for big data professions: A systematic classification of job roles and required skill sets," *Inf. Process. Manage.*, vol. 54, no. 5, pp. 807–817, Sep. 2018, doi: 10.1016/j.ipm.2017.05.004.

[14] J. Zhang, T. Le, and J. Chen, "Investigation of essential skills for data analysts: An analysis based on LinkedIn," *J. Global Inf. Manage.*, vol. 31, no. 1, pp. 1–21, Jul. 2023, doi: 10.4018/jgim.326548.

[15] F. Gurcan and N. E. Cagiltay, "Big data software engineering: Analysis of knowledge domains and skill sets using LDA-based topic modeling," *IEEE Access*, vol. 7, pp. 82541–82552, 2019, doi: 10.1109/ACCESS.2019.2924075.

[16] D. Debao, M. Yinxia, and Z. Min, "Analysis of big data job requirements based on K-means text clustering in China," *PLoS ONE*, vol. 16, no. 8, Aug. 2021, Art. no. e0255419, doi: 10.1371/journal.pone.0255419.

[17] A. Persaud, "Key competencies for big data analytics professions: A multimethod study," *Inf. Technol. People*, vol. 34, no. 1, pp. 178–203, Mar. 2020, doi: 10.1108/itp-06-2019-0290.

[18] M. Almgerbi, A. De Mauro, A. Kahlawi, and V. Poggioni, "A systematic review of data analytics job requirements and online-courses," *J. Comput. Inf. Syst.*, vol. 62, no. 2, pp. 422–434, Mar. 2022, doi: 10.1080/08874417.2021.1971579.

[19] A. Gardiner, C. Aasheim, P. Rutner, and S. Williams, "Skill requirements in big data: A content analysis of job advertisements," *J. Comput. Inf. Syst.*, vol. 58, no. 4, pp. 374–384, Oct. 2018, doi: 10.1080/08874417.2017.1289354.

[20] S. Yusoff, N. H. M. Noh, N. Isa, and S. M. Nor-Al-Din, "Knowledge and skill sets for big data profession: Assessing student's quality using exploratory factor analysis," in *Proc. Int. Visualizat., Informat. Technol. Conf. (IVIT)*, Nov. 2022, pp. 272–277, doi: 10.1109/IVIT55443.2022.10033399.

[21] P. Mikalef, I. O. Pappas, J. Krogstie, and M. Giannakos, "Big data analytics capabilities: A systematic literature review and research agenda," *Inf. Syst. e-Bus. Manage.*, vol. 16, no. 3, pp. 547–578, Aug. 2018, doi: 10.1007/s10257-017-0362-y.

[22] D. H. Schmidt, D. van Dierendonck, and U. Weber, "The data-driven leader: Developing a big data analytics leadership competency framework," *J. Manage. Develop.*, vol. 42, no. 4, pp. 297–326, Jul. 2023, doi: 10.1108/jmd-12-2022-0306.

[23] A. Bäck, A. Hajikhani, and A. Suominen, "Text mining on job advertisement data: Systematic process for detecting artificial intelligence related jobs," in *Proc. 1st Workshop on AI + Informetrics (AII)*, Virtual, 2021, pp. 111–124. [Online]. Available: https://ceur-ws.org/Vol-2871/paper9.pdf

[24] I. Karakatsanis, W. AlKhader, F. MacCrory, A. Alibasic, M. A. Omar, Z. Aung, and W. L. Woon, "Data mining approach to monitoring the requirements of the job market: A case study," *Inf. Syst.*, vol. 65, pp. 1–6, Apr. 2017, doi: 10.1016/j.is.2016.10.009.

[25] I. Khaouja, I. Rahhal, M. Elouali, G. Mezzour, I. Kassou, and K. M. Carley, "Analyzing the needs of the offshore sector in Morocco by mining job ads," in *Proc. IEEE Global Eng. Educ. Conf. (EDUCON)*, Apr. 2018, pp. 1380–1388, doi: 10.1109/EDUCON.2018.8363390.

[26] J. Wei and Y. Xu, "The application of LDA model in the analysis of job talent demand under big data technology," in *Proc. Int. Conf. Artif. Intell. Everything (AIE)*, Aug. 2022, pp. 301–305, doi: 10.1109/AIE57029.2022.00065.

[27] R. Liu, W. Ye, R. Gao, M. Tang, and D. Wang, "Research on text clustering based on requirements of big data jobs," *Data Anal. Knowl. Discov.*, vol. 1, no. 12, pp. 32–40, 2017. [Online]. Available: https://kns.cnki.net/KCMS/detail/detail.aspx?dbcode=CJFD&dbname=CJFDLAST2018&filename=XDTQ201712004&v=

[28] M. Lukauskas, V. Šarkauskaitė, V. Pilinkienė, A. Stundžiene, A. Grybauskas, and J. Bruneckienė, "Enhancing skills demand understanding through job ad segmentation using NLP and clustering techniques," *Appl. Sci.*, vol. 13, no. 10, p. 6119, May 2023, doi: 10.3390/app13106119.

[29] Q. Xiao, X. Zhong, and C. Zhong, "Application research of KNN algorithm based on clustering in big data talent demand information classification," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 34, no. 6, Jun. 2020, Art. no. 2050015, doi: 10.1142/s0218001420500159.

[30] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*.

[31] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," 2013, *arXiv:1310.4546*.

[32] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, and P. S. Yu, "Top 10 algorithms in data mining," *Knowl. Inf. Syst.*, vol. 14, no. 1, pp. 1–37, 2008, doi: 10.1007/s10115-007-0114-2.

[33] X. Li, "Research and application of knowledgee question answering system in education based on knowledge graph," M.S. thesis, College Comput. Sci. Technol., Jilin Univ., Changchun, China, 2019. [Online]. Available: https://kns.cnki.net/KCMS/detail/detail.aspx?dbcode=CMFD&dbname=CMFD202001&filename=1019155246.nh&v=

[34] A. Ternikov, "Soft and hard skills identification: Insights from IT job advertisements in the CIS region," *PeerJ Comput. Sci.*, vol. 8, p. e946, Apr. 2022, doi: 10.7717/peerj-cs.946.

[35] F. Gurcan and S. Sevik, "Business intelligence and analytics: An understanding of the industry needs for domain-specific competencies," in *Proc. 1st Int. Informat. Softw. Eng. Conf. (UBMYK)*, Nov. 2019, pp. 1–5, doi: 10.1109/UBMYK48245.2019.8965457.

[36] A. Verma, K. Lamsal, and P. Verma, "An investigation of skill requirements in artificial intelligence and machine learning job advertisements," *Ind. Higher Educ.*, vol. 36, no. 1, pp. 63–73, Feb. 2022, doi: 10.1177/0950422221990990.

[37] L. Cao, "Data science: Profession and education," *IEEE Intell. Syst.*, vol. 34, no. 5, pp. 35–44, Sep. 2019, doi: 10.1109/MIS.2019.2936705.

[38] B. Baumer, "A data science course for undergraduates: Thinking with data," *Amer. Statistician*, vol. 69, no. 4, pp. 334–342, Oct. 2015, doi: 10.1080/00031305.2015.1081105.

[39] L. Cao, "Data science thinking: The next scientific, technological and economic revolution," in *Data Analytics*. Cham, Switzerland: Springer, 2018, doi: 10.1007/978-3-319-95092-1.

[40] M. E. Jennex, "Big data, the Internet of Things, and the revised knowledge pyramid," *ACM SIGMIS Database: DATABASE Adv. Inf. Syst.*, vol. 48, no. 4, pp. 69–79, Nov. 2017, doi: 10.1145/3158421.3158427.

FANG HAN was born in Mentougou, Beijing, China, in 1996. She received the B.S. degree in statistics and the M.S. degree in applied statistics from the Capital University of Economics and Business, in 2018 and 2020, respectively.

She is currently a Staff Member with the Education Quality Evaluation Agency, Capital University of Economics and Business. Her current research interests include big data analysis and higher education administration.

JIYUAN REN was born in Xicheng, Beijing, China, in 1989. He received the B.A. degree in education from Capital Normal University, in 2011, the M.Ed. degree in curriculum, teaching, and assessment from The Education University of Hong Kong, in 2013, and the Ph.D. degree in education from Capital Normal University, in 2019.

He is currently an Assistant Research Fellow with the Office of Development and Planning, Capital University of Economics and Business. His current research interests include higher education, vocational education, and family education.