IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 22, 2019, accepted June 12, 2019, date of publication June 20, 2019, date of current version July 9, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2924075

Big Data Software Engineering: Analysis of Knowledge Domains and Skill Sets Using LDA-Based Topic Modeling

FATIH GURCAN[1] **AND NERGIZ ERCIL CAGILTAY**[2]
[1]Department of Computer Engineering, Faculty of Engineering, Karadeniz Technical University, 61080 Trabzon, Turkey
[2]Department of Software Engineering, Faculty of Engineering, Atilim University, 06830 Ankara, Turkey

Corresponding author: Fatih Gurcan (fgurcan@ktu.edu.tr)

ABSTRACT Software engineering is a data-driven discipline and an integral part of data science. The introduction of big data systems has led to a great transformation in the architecture, methodologies, knowledge domains, and skills related to software engineering. Accordingly, education programs are now required to adapt themselves to up-to-date developments by first identifying the competencies concerning big data software engineering to meet the industrial needs and follow the latest trends. This paper aims to reveal the knowledge domains and skill sets required for big data software engineering and develop a taxonomy by mapping these competencies. A semi-automatic methodology is proposed for the semantic analysis of the textual contents of online job advertisements related to big data software engineering. This methodology uses the latent Dirichlet allocation (LDA), a probabilistic topic-modeling technique to discover the hidden semantic structures from a given textual corpus. The output of this paper is a systematic competency map comprising the essential knowledge domains, skills, and tools for big data software engineering. The findings of this paper are expected to help evaluate and improve IT professionals' vocational knowledge and skills, identify professional roles and competencies in personnel recruitment processes of companies, and meet the skill requirements of the industry through software engineering education programs. Additionally, the proposed model can be extended to blogs, social networks, forums, and other online communities to allow automatic identification of emerging trends and generate contextual tags.

INDEX TERMS Big data software engineering, competency map, knowledge domains and skill sets, topic modeling, latent Dirichlet allocation.

I. INTRODUCTION

Today's digital world is also called the era of big data. Accordingly, big data systems are causing a transformation in the architecture and methodologies of software engineering [1]. The volume and variety of data generated and shared is increasing exponentially [2], [3]. In general terms, big data refers to operations based on processing huge amounts of data to reveal hidden patterns and correlations, and offer other insights [2]. The valuable insights and implications derived from big data-oriented services applications are used in intelligent processes, such as guiding decision-making strategies in business, science, society, and government [3]–[5].

Increasing demands for data-oriented services and applications in all industrial and social areas has led to intensification of software development activities related to big data [3]. Big data-driven software applications are increasingly dominant in today's technological life cycle. Recently, big data-oriented software systems have been embedded in many modern products and services; thus, they have become more and more significant around the world [1], [3], [4], [6], [7]. The economy and industry are experiencing a computerized transformation toward software- and service-based businesses for which modern software systems can provide valuable inferences from big datasets [3], [4], [8]. Throughout this transformation process, software engineering has undertaken an important mission in the modernization of many industries [3], [9]–[11]. However, the advent of big data systems has led to the emergence of new issues and challenges that need to be resolved using technology-based disciplines, especially software engineering [3], [12]–[14]. Therefore, 'big data software engineering' (BDSE) has taken its place in the literature as a

The associate editor coordinating the review of this manuscript and approving it for publication was Yongwang Zhao.

new discipline and has become a widespread research and application field in recent years [12], [13].

As a concept, big data is defined around 5Vs of volume, variety, velocity, veracity, and variability [6], [15]. BDSE involves advanced processes, methodologies, and approaches that adapt the traditional software development life-cycle according to these main characteristics (5V) of big data [15], [16]. The big data-driven software development process is conducted by combining big data operations into the software development life-cycle phases [3], [12], [17]. During the development of 'big data-oriented software applications' (BDSAs), the most challenging period in terms of BDSE is concerned with the scalable design of software systems [10]. In the development life-cycle of BDSAs, the system must first be designed with scalable architecture [5], [13], [18]. In keeping with scalable architecture, a BDSA is initially designed on a small data scale, and then expanded to big data systems. This scalable architecture allows faster and more efficient implementation of the development steps of BDSAs, such as design, prototyping, testing, diagnosis, and quality-assurance [3], [18]. Especially in BDSAs with dynamic data streams, complications, such as data replication, scaling, inconsistency and latency, as well as real-time processing can be resolved using more sophisticated software architecture [5], [6], [13], [19].

Given this background, BDSAs require a wide range of vocational knowledge, skills, and competencies, unlike traditional software applications [10], [11], [18], [20]. The software industry is a dynamic working environment based entirely on qualified human resources [3], [9], [11], [20], [21]. The quality of BDSE-based products and services is closely related to the competencies of BDSE specialists [3], [8], [22]. As BDSAs become more widespread, the demand for qualified BDSE specialists is increasing day by day. For this reason, BDSE is a growing field of employment in today's labor market. In this respect, a prediction report, published by the McKinsey Global Institute, underlines that ''the United States alone faces a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts to analyze big data and make decisions based on their findings'' [23]. Parallel to this report, research indicates that ''inadequate staffing and skills are the leading barriers to Big Data Analytics'' [24]. This data-oriented technological transformation has led to the emergence of an innovative field of expertise in terms of software engineering [14]. Therefore, the number of BDSE-oriented job postings in online employment platforms are increasing significantly [3], [25]. When viewed from this perspective, it is envisaged that the identification of up-to-date knowledge and skills sought by the BDSE industry will provide significant contributions to meet the requirements for qualified labor in this area [21].

This study aimed to identify the knowledge domains and skill sets required for BDSE. A semi-automatic methodology was proposed to analyze the collections of online job advertisements (ads). Our methodology is based on the semantic analysis of BDSE job ads using Latent Dirichlet Allocation (LDA), a probabilistic topic model used to discover latent semantic patterns (topics) in a textual corpus. Based on the topics discovered by LDA, we revealed the essential knowledge and skills required for BDSE. A competency taxonomy for BDSE was then developed by mapping the topics according to competency domains. Furthermore, the technologies required for BDSE, such as programming languages and tools, databases, data warehouses, and big data tools were extracted.

II. BACKGROUND

The methodology of this study was based on a content analysis of the textual content of BDSE job ads using probabilistic topic models in order to reveal the knowledge domains and skill sets required for BDSE. Therefore, the background of the study is addressed under two subheadings: ''big data software engineering'' and ''topic models''.

A. BIG DATA SOFTWARE ENGINEERING

Ever since the term big data emerged in the scientific literature, it has been a phenomenon closely associated with software engineering. In particular, the increasing demands and challenges related to BDSAs, which have emerged as a natural consequence of the widespread use of big data, have been a fundamental issue frequently highlighted in the literature, especially in the last few decades. Recent studies indicate that products and services using BDSAs require more advanced and specific professional knowledge and skills in terms of software engineering principles, procedures and processes [8], [13], [18], [20]. Although software engineering is a data-driven discipline, software development processes concerning big data systems require the use of more progressive knowledge and skills, such as scalable architecture, real-time data processing, real-time coding, integration, and testing. [3], [7], [8], [10], [11], [22], [25], [26].

BDSE is principally based on the simultaneous use of software engineering processes with big data processing and analytics [3], [4], [8], [17]. In particular, online analytical processing and business intelligence are commonly employed in most BDSAs [4]. Development of software applications based on real-time data processing and analysis processes using dynamic data streams has created a new area of expertise in software engineering [3]. This is why meeting the growing need of qualified specialists in BDSE has emerged as a challenge that software engineering must resolve in the near future [3]. This challenge, commonly defined as a big data skills gap, has been discussed by a number of scientific studies and industrial reports [3], [4], [8]–[11], [17], [21], [23]. Considering research based on the analysis of job ads, it is observed that a limited number of specific studies have been carried out to reveal the knowledge and skill requirements for the software engineering industry. A number of these studies are only related to big data competencies [8], [17], [25], while the rest solely focus on software engineering [9], [27]–[29]. To the best of our knowledge, this is the first experimental study that specifically sets out the knowledge domains and

skill sets required for BDSE in line with emerging market demands. Our further analysis of online job ads is expected to provide significant contributions in terms of identifying and understanding the knowledge and skills needed in the industry and meeting these requirements [8], [9].

B. TOPIC MODELING

Topic modeling is a probabilistic approach used for the discovery of latent semantic patterns, called topics, in the collection of unstructured documents. Topic modeling refers to describing the semantic structure of documents in keeping with the discovered topics. This approach is based on the principle that topics have a percentage of random probability distribution in words within a document [30], [31]. Specific words are expected to be seen more frequently in a document because a document is intuitively related to a specific topic [31], [32]. The topics discovered by topic modeling are actually semantic clusters created by words that are often used together in a document [31]. In the identification of latent semantic structures, the probability distribution of each topic is calculated along with the distributions of topics per document and topic assignments per word in each document [30], [31].

LDA, a probabilistic and generative model, is one of the topic modeling algorithms commonly used in text mining [30]. The term ''latent'' in LDA is related to the discovery of the semantic content of documents by analyzing the latent semantic structures within the documents [33]. The generative approach in LDA is defined as the process of assigning the words in a document to random variables, and semantically clustering them using a repetitive probabilistic process based on a Dirichlet distribution [31]. LDA does not require any labeling or training set, as it uses unsupervised learning approaches [30]. Therefore, an LDA model can be effectively applied to huge collections of documents in a given text corpus to discover semantic patterns [30], [31], [34].

Recently, LDA has been widely used in text mining studies conducted in different contexts, such as natural language processing, information extraction, sentiment analysis, literature research, and social trend analytics [31], [35]. Likewise, this model has been employed as an effective method in some studies that performed the analysis of online job advertisements in various industries [9], [25]. The topic models that were initially developed for the analysis of textual data are now applied to different types of data, including genetic data, images, videos, and social networks [31], [33]. For these reasons, in this study, LDA was used as the method for topic modeling.

III. METHODOLOGY

In this study, a semi-automatic methodology was proposed to analyze the content of online BDSE job ads. In this context, the methodology consisted of four main phases: data collection, data preprocessing, data analysis, and interpretation (see Figure 1). Each phase of this methodology is described in more detail in the following sections.

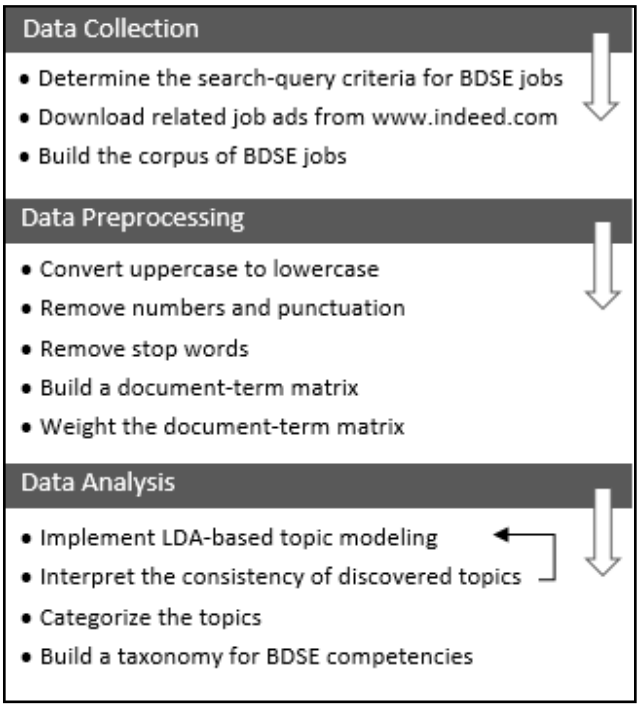

FIGURE 1. An overview of the research methodology.

A. DATA COLLECTION

For this study, indeed.com [36], an online employment site offering comprehensive search and filter options, was selected as the data source. The expertise areas of BDSE include a wide range of job titles and descriptions. The query and filter options were utilized to acquire the job ads that only covered software and application development for big data. Job ads were queried and filtered to find the ads containing ''big data'' and ''developer'', ''big data'' and ''software'', and ''big data'' and ''application'' statements in job titles; e.g., those with job titles, such as ''big data developer'', ''big data software engineer'', ''big data application developer'', and ''big data software developer''. In this way, a more specific data set was created within the scope of the study. Finally, using an API developed by indeed.com, a data set consisting of 2,638 online job ads published in a three-month period from May 2018 to July 2018 was created.

B. DATA PREPROCESSING

The preprocessing stage is very important to improve the quality analyses of unstructured text data [2], [37], [38]. Data preprocessing is an operation that must be undertaken especially in the analysis of web-based unstructured textual contents [2], [37]. In this study, the preprocessing applied to the experimental data set consisted of several sequential steps. Initially, the textual content was divided into words (tokens), known as tokenization, to obtain meaningful attributes [39]. Thus, each textual content in the data set was represented by a word vector. Then, punctuation, web links, private tags, and meaningless characters were deleted. Stop words were then removed from the texts to reduce the word space. In the

preprocessing stage, ''tm'' package, a framework for text mining applications within R, was used to remove punctuations, numbers, stop words, html tags, and white space, and convert all text to lower case. In order to eliminate the loss of meaning in the text content, the stemming process was not applied since the data set contained technical jargon, and the word space representing the data set was composed of many technical words.

With the completion of the preprocessing steps, each text (job ad) in the data set was defined as a word vector. As a result of preprocessing, the size of the word space for the entire data set was reduced from 13,877 to 10,432. The data set consisting of BDSE job ads was characterized by 10,432 unique words, which also referred to the size of the word vector for each ad. The number of vectors was 2,638, which was also the number of job ads. In this way, all text in the data set was characterized as a word vector in the vector-space model. The (word) vectors belonging to each ad were combined to create a document-term matrix (DTM) on which to perform a quantitative analysis [39]. The creation of DTM is based on the ''bag of words'' approach, and it provides information about word frequencies without considering their order [31], [32]. The DTM created for this analysis consisted of 2,638 rows and 10,432 columns. In other words, the DTM indicated that 2,638 job ads were represented by a word space comprising 10,432 terms. The weighting process of the DTM was performed by taking into account the frequency of the words.

C. IMPLEMENTATION OF LDA-BASED TOPIC MODELING

This phase of the experimental analysis covered the implementation of topic models of the data set to reveal the knowledge domains and skill sets required for BDSE in a comprehensible manner. LDA is an effective topic model suitable for the characteristic of this study based on the semantic analysis of job ads. In LDA-based topic modeling, a topic is defined by a group of related words with different probabilities [30]. For example, if a topic contains words, such as human, genome, dna, rna, genetics, and gene, this word cluster describes a topic related to genetics [31]. In LDA-based topic modeling, the distribution of topics in documents and the distribution of words in topics are independent of each other. More specifically, the same words may appear at different rates in different topics. Similarly, the same topics may appear at different rates in different documents. This assumption in the LDA model is based on a Bayesian joint probabilistic model. The aim of undertaking LDA-based topic modeling in this study was to reveal the latent semantic structures (word clusters) in the textual corpus consisting of job ads.

In the implementation process of the LDA model, the probability distribution for each topic was calculated using the Bayesian estimation technique along with a Dirichlet distribution. For the implementation of the LDA model in this experimental analysis, we used MALLET [40] tool, which is an implementation of the LDA model that employs the Gibbs sampling algorithm [41]. MALLET was applied with different iteration numbers and stabilized for 2,200 Gibbs sampling iterations. Another important parameter for the application of LDA is the number of topics indicated by T [31], [32], which is a user-specified parameter that adjusts the granularity level of the discovered topics. Different values ranging from 30 to 60 were tried, and the desired modeling level was obtained from $T = 48$. In the selection of the ideal number of topics, the semantic consistency of the discovered topics, and the distribution of the descriptive keywords in these topics were taken into consideration. In the LDA model, the researchers manually assigned topic names to the discovered topics, consistent with the descriptive keywords. However, one of the two main approaches was followed during the naming process: The topic names were mainly given through a meaningful combination of the first four keywords. In some cases, the keywords all together defined a specific domain. For such cases, the topic names were assigned by considering the general meaning of all keywords. Therefore, the assigned topic names might slightly vary according to the perspective of the researchers.

IV. RESULTS

In order to better understand the competencies for BDSE, first, the skill sets were identified from the data sets by topics. Then, these skill sets were mapped into the competency domains. Additionally, most in-demand tools for BDSE, programming languages, programming tools, databases, data warehouses, big data tools, and their combinations were all analyzed to better identify the competencies. In the following section, the results of the analysis are presented and discussed.

A. IDENTIFICATION OF KNOWLEDGE AND SKILLS BY LDA

The job ads in the data set comprised a wide spectrum of knowledge, skills and capabilities in different expertise areas. This wide range of BDSE jobs extended the coverage of the discovered topics. Therefore, the LDA-based topic modeling analysis was performed on the job ads using different parameters, and as a result, 48 trending topics were determined to reveal the knowledge domains and skill sets of BDSE with optimal granularity. These topics are presented in Table 1, together with the descriptive LDA keywords and topic rates. The topics are listed in the table in descending order of their rate. The names of the discovered topics were assigned manually considering the descriptive keywords and their frequencies. In Table 1, the first word was the most seen and the last word was the least seen word in a topic. In this context, the topic names were commonly assigned taking into consideration the first four keywords. On the other hand, for some topics, the keywords all together defined a specific domain. The following six topic names were assigned considering all keywords: data-driven languages, big data tools, license fields, scripting programming, databases, and web services.

Table 1 indicates that experience, communication skills, and hadoop ecosystem were among the competencies with

TABLE 1. Discovered topics.

Topic name	LDA keywords	Rate %
Experience	experience year skill work vocational relevant minimum preferred related practice	4.35
Communication skills	skills communication ability written strong effectively verbal excellent presentation	4.24
Hadoop ecosystem	hive hadoop ecosystem spark hbase hdfs pig mapreduce sqoop kafka impala flume	3.98
Software development	development software code writing coding implementation program execute	3.96
Testing	test testing software code integration control cases components unit qa check	3.42
Cloud	cloud azure aws services platforms paas iaas computing google settings saas	3.26
Deep-domain knowledge	knowledge deep domain big data understanding strong successful good background	3.19
Analytical skills	analytical background skills knowledge analysis statistics models data predictive	3.15
Data-driven languages	java python r scala languages big data driven spark programming focused	3.11
Problem solving	problem technical solving solve ability skills strong level analytical concepts proven	2.99
Big data tools	bigdata hadoop tools platforms aws etl cloudera preferred hortonworks mapr	2.93
Project management	project management delivery planning manage lead level managing activities organize	2.83
Design and scaling	design scale flexible architecture application scaling modeling patterns prototyping	2.74
License fields	license science computer related ope bachelor field engineering equivalent discipline	2.51
Team-working	team work working group growing strong part communication join motivated	2.50
Scripting programming	java python scala programming languages scripting shell language scripts perl ruby linux	2.47
Streaming-data integration	integration data stream connect flow set query internal resource host warehouse	2.25
Data structures	complex data structures large scale set designing unstructured multi project developing	2.19
Databases	sql nosql database relational server oracle cassandra mongodb stores db rdbms	2.12
Platform	technologies platform source open software build generation latest stack including	2.04
Real-time programming	real time concurrency programming response latency system process synchronize event	2.02
Distributed systems	distributed systems networking peer spread networks wan remote access location	1.86
Real-time data	time full real data flow online processing pipelines ingestion stack event streaming	1.81
Scalable programming	scalable develop programming scala prototype spark extend scalability enlarge modular	1.74
Web services	restful apps framework web user api java rest developing applications spring	1.74
Agile development	software agile development scrum practices cycle methodology life sdlc devops	1.72
Streaming processing	streaming frameworks processing kafka data storm cassandra messaging stream spark	1.64
Leadership	technical leadership provide business role senior leader relationships guidance direction	1.59
Data reporting-visualization	data reporting etl visualization modeling tools display warehousing intelligence tableau	1.54
Machine learning	learning machine optimization models analytics advanced predictive statistical algorithms	1.52
Collaborative environment	environment collaborative work fast paced joint multiple dynamic manage tasks	1.52
Big data warehousing	data big warehouse lake enterprise integration lineage movement etl client	1.51
Team-member	team members role build join related engineers work developers scientists closely	1.48
Business requirements	business requirements understand problem meet define objectives solution ensure users	1.45
Mapreduce programming	mapreduce hadoop programming parallel distributed java spark data map tasks	1.41
Continuous integration	continuous integration deployment devops jenkins automation control git docker maven	1.33
Scaling	systems scalable principles data scaling technologies distributed storage file extend	1.32
Professional training	training professional skill career technical growth self-development organization build	1.32
Decision-support	information decision support intelligence mission systems government services enable	1.29
Storage	systems storage file networking big data distributed computing scalable stream load	1.20
Initiative	initiatives creativity functional vision lead cross technical common develop goals	1.17
Master degree	preferred master degree related graduate programs field engineering computer science	1.16
Computational thinking	computational logical thinking skill cognitive algorithms decomposition solution pattern	1.13
Business intelligence	business analytics intelligence solution team drive insights improve strategies leading	1.11
Big data processing	big data processing batch hadoop hdfs bounded spark mapreduce yarn pig	1.08
Object-oriented	object oriented design architecture prototyping integration application java oop model	1.05
Verification	verification software process control evaluation version validation check testing unit	1.04
Quality-assurance	quality testing assurance practices maintain process ensure policies implement confirm	1.02

the highest demand in the BDSE industry. Other knowledge and skills in the top ten were software development, testing, cloud, deep-domain knowledge, analytical skills, data-driven languages, and problem solving. The discovered topics also covered various emerging trends, such as big data tools, software architecture, programming languages, frameworks, platforms, technologies, and competencies that shed light on the priorities and demands in the ever-growing BDSE industry.

B. MAPPING OF KNOWLEDGE AND SKILLS BY COMPETENCY DOMAINS

This phase of the analysis aimed to categorize and present the knowledge and skills in a more comprehensible manner. To this end, a mapping process was performed by associating the knowledge and skills with the competency domains and workflows. The BDSE knowledge and skills revealed by 48 topics were mapped into 10 core competency areas, and a competency map was developed for BDSE. The distribution

TABLE 2. Competency map.

ID	Competency areas	Knowledge and skills	Rate %	Total %
1	Big data frameworks	Hadoop ecosystem	3.98	11.22
		Big data tools	2.93	
		Data structures	2.19	
		Databases	2.12	
2	Big data processes	Streaming-data integration	2.25	7.82
		Data reporting-visualization	1.54	
		Big data warehousing	1.51	
		Scaling	1.32	
		Storage	1.20	
3	Big data analytics	Analytical skills	3.15	7.07
		Machine learning	1.52	
		Decision-support	1.29	
		Business intelligence	1.11	
4	Data processing types	Real-time data	1.81	4.53
		Streaming processing	1.64	
		Big data processing	1.08	
5	Software development lifecycle	Software development	3.96	14.96
		Testing	3.42	
		Design and scaling	2.74	
		Business requirements	1.45	
		Continuous integration	1.33	
		Verification	1.04	
		Quality-assurance	1.02	
6	Programming	Data-driven languages	3.11	11.80
		Scripting programming	2.47	
		Real-time programming	2.02	
		Scalable programming	1.74	
		Mapreduce programming	1.41	
		Object-oriented	1.05	
7	Software development frameworks	Cloud	3.26	10.62
		Platform	2.04	
		Distributed systems	1.86	
		Web services	1.74	
		Agile development	1.72	
8	Vocational background	Experience	4.35	12.53
		Deep-domain knowledge	3.19	
		License fields	2.51	
		Professional training	1.32	
		Master degree	1.16	
9	Soft skills	Communication skills	4.24	11.19
		Problem solving	2.99	
		Project management	2.83	
		Computational thinking	1.13	
10	Work style	Team-working	2.50	8.26
		Leadership	1.59	
		Collaborative environment	1.52	
		Team-member	1.48	
		Initiative	1.17	

of the knowledge and skills according to the competency areas are given in Table 2 with their percentages.

As shown in Table 2, the first four of the competency areas are related to big data discipline, which consist of big data frameworks, big data processes, big data analytics, and data processing types. The total rate of these competency areas related to big data is 31%. The next three competency areas are related to the software engineering discipline, comprising the software development lifecycle, programming, and software development frameworks. The total rate of these competency areas is 37%. The last three concern the interdisciplinary areas, consisting of vocational background, soft skills, and work style at a rate of 32%. These 10 competency areas are discussed in detail below.

The first competency area, big data frameworks (11.22%), contains four knowledge and skill items consisting of the Hadoop ecosystem, big data tools, data structures, and databases. The second, big data processes (7.82%), has five items, namely streaming-data integration, data reporting-visualization, big data warehousing, scaling, and storage. The third, big data analytics (7.07%), contains the four items of analytical skills, machine learning, decision-support, and business intelligence. The fourth, data processing types (4.53%), has three items: real-time data, streaming

processing, and big data processing. The fifth, software development lifecycle (14.96%), consists of seven items: software development, testing, design and scaling, business requirements, continuous integration, verification, and quality-assurance. The sixth, programming (11.80%), has six items comprising data-driven languages, scripting programming, real-time programming, scalable programming, mapreduce programming, and object-oriented. The seventh, software development frameworks (10.62%), has five items, namely cloud, platform, distributed systems, web services, and agile development. The eighth, vocational background (12.53%), contains five consisting of work experience, deep-domain knowledge, license fields, professional training, and master degree. The ninth, soft skills (11.19%), comprises four items: communication skills, problem solving, project management, and computational thinking. Finally, work style (8.26%) contains five items, namely team-working, leadership, collaborative environment, team-member, and initiative.

C. IDENTIFICATION OF THE MOST IN-DEMAND TOOLS FOR BDSE

In today's collective environments of software development, a wide range of tools and technologies, such as programming languages, frameworks, databases, and data tools are used together. In order to reveal the tools and technologies required for BDSE, the data set was analyzed using the keyword indexing technique [29]. The findings of this analysis were divided into the four main categories of programming languages, programming tools, database technologies, and big data tools, which are discussed in detail in the following sections.

TABLE 3. Programming languages.

Programming languages	Rate %
Java	27.8
Python	21.3
Scala	16.3
R	6.8
C++	4.9
JavaScript	4.1
.Net	3.6
Ruby	2.7
C	2.5
Perl	2.3
C#	2.3
Typescript	2.2
Go	1.5
Julia	0.9
Php	0.8

1) PROGRAMMING LANGUAGES

Being elementary tools of application development, many programming languages have been developed for different purposes from past to present. In order to identify the programming languages used in BDSE, the data set consisting of job ads was analyzed by keyword indexing. As a result, the top 15 programming languages required for BDSE were identified and are presented in Table 3 with their rates. According to the results, Java is the most leading programming language in this field, followed by Python and Scala. The total rate of these three programming languages is 65%, a high percentage indicating their leadership. The results show the significant competition between Java and Python in this field. Besides, the R programming language seems to have a growing trend in data science in recent years.

TABLE 4. Programming tools.

Programming tools	Rate %
Jenkins	18.9
Maven	16.0
Spring MVC	13.7
SVN	9.2
Github	8.2
Hibernate	5.8
Node.js	5.4
Angular.js	5.4
JQuery	3.9
Backbone.js	2.7
Sprint	2.6
Lucene	2.6
NumPy	2.2
Ant	2.0
Flask	1.5

2) PROGRAMMING TOOLS

Programming tools are often used in conjunction with programming languages to develop software applications more easily. These tools contain various types of utilities, such as frames, libraries, and applications. As seen in Table 4, Jenkins, a continuous integration tool developed in Java, is the most in-demand programming tool in the BDSE. Maven, a tool built for java projects, ranks second, and it is followed by Spring MVC, an application and control tool for the Java platform. The fourth is Apache Subversion, abbreviated as SVN, is a software versioning and revision control system distributed under an open source license. GitHub, a development platform and version control system built around the Git tool, ranks fifth among the tools. The sixth is Hibernate, an open source object relational mapping (ORM) tool that provides solutions for Java environments to map object-oriented domain models. Moreover, JavaScript libraries containing node.js, angular.js, jquery, and backbone.js seem to be significantly utilized in the field of BDSE.

3) DATABASES AND DATA WAREHOUSES

Databases and data warehouses are an integral part of BDSE with their active role in processes, such as data acquisition, processing and storing. There are many databases and data warehouses designed for specific tasks. The 15 most in-demand databases and data warehouses in BDSE identified in this study are given in Table 5 with their rates. The results reveal that Hive, a data warehousing tool, ranks first. This is followed by SQL, a database query tool; Hbase, a distributed column-oriented database; NoSQL, a new generation database; and Cassandra, a distributed database system.

TABLE 5. Databases and data warehouses.

Databases and data warehouses	Rate %
Hive	22.0
SQL	17.3
Hbase	13.9
NoSQL	13.2
Cassandra	8.8
Oracle	5.9
MongoDB	5.3
MySQL	2.6
SQL Server	2.5
Teradata	2.3
Sap Hana	1.6
Netezza	1.5
PostgreSQL	1.3
Redis	1.0
IBM Db2	0.7

TABLE 6. Big data tools.

Big data tools	Rate %
Hadoop	20.0
Spark	17.5
Kafka	9.8
Aws	9.1
Mapreduce	5.5
Pig	5.4
Hdfs	5.2
Azure	4.5
Storm	4.2
Impala	4.0
Cloudera	3.5
Sqoop	3.5
Redshift	2.8
Flume	2.5
Yarn	2.4

The total rate of these five items is 66%, showing their dominance in this field.

4) BIG DATA TOOLS

Unlike traditional data processing, there are a large number of tools used in big data processing for various purposes. The existing big data tools focus on three processing paradigms: big data processing, real-time data, and hybrid processing. Table 6 presents the 15 most in-demand big data tools in BDSE identified in this study. As presented in Table 6, Hadoop, a widespread tool for big data processing, ranks first. This is followed by Spark used both as a batch and as a real-time processing tool. Then comes Kafka, which is a distributed-streaming tool. This is followed by Aws, a cloud service platform. The fifth tool is mapreduce which is a programming model for big data processing. The total rate of these five tools is 62%, which clearly demonstrates the leading role of these big data tools in the BDSE field.

D. IDENTIFICATION OF THE MOST IN-DEMAND COMBINATIONS

Today's dynamic and collaborative software development environments require combining many tools and technologies, such as programming languages, programming tools, databases, and data warehouses. In this context, to identify the combinations of tools and technologies with a high demand in the BDSE industry, a further analysis was carried out using keyword indexing. Considering the in-demand tools and technologies given in the previous section, the 15 most in-demand triple combinations consisting of programming languages and database or data warehouses were identified. In Table 7, the triple combinations are sorted in descending order of their frequency of occurrence.

TABLE 7. In-demand combinations of programming languages and databases or data warehouses.

Combinations	Rate %
Java, Python, Hive	11.7
Java, Python, Scala	10.7
Java, Python, Sql	9.6
Java, Scala, Hive	9.4
Java, Python, Hbase	7.8
Java, Scala, Sql	7.3
Python, Scala, Sql	6.6
Java, Scala, Hbase	6.5
Java, Python, Nosql	5.8
Java, Python, Cassandra	4.9
Java, Scala, Nosql	4.6
Python, R, Sql	4.0
Java, Python, R	4.0
Python, Scala, Nosql	3.9
Python, Scala, R	3.2

TABLE 8. In-demand combinations of tools and databases or data warehouses.

Combinations	Rate %
Hadoop, Spark, Hive	10.8
Hadoop, Spark, Kafka	8.6
Hadoop, Spark, Hbase	7.4
Hadoop, Spark, Aws	7.2
Hadoop, Hive, Hbase	6.7
Hadoop, Hive, Pig	6.6
Spark, Hive, Kafka	6.6
Hadoop, Hive, Kafka	6.4
Spark, Hive, Hbase	6.3
Hadoop, Hive, Aws	5.8
Hadoop, Spark, Pig	5.8
Spark, Hive, Pig	5.7
Spark, Hive, Aws	5.5
Hadoop, Spark, Mapreduce	5.3
Hadoop, Spark, Storm	5.3

According to the results, the highest in-demand triple combination is 'Java + Python + Hive, followed by Java + Python + Scala, and Java + Python + Sql. As clearly seen in Table 7, the dominance of Java and Python in the BDSE industry is noteworthy. The table also reveals combinations of these two programming languages with different databases or data warehouses. R and Scala are the two other programming languages commonly seen in these combinations, and the most common databases or data warehouses are Sql, Hive, Hbase, Cassandra, and Nosql.

The 15 most in-demand triple combinations including big data tools and database or data warehouses were also identified and are presented in Table 8. The combination with the

highest demand was Hadoop + Spark + Hive, followed by Hadoop + Spark + Kafka, and Hadoop + Spark + Hbase. As big data processing tools, the leadership of Hadoop and Spark in the BDSE industry is clearly seen in Table 8.

V. DISCUSSION

In this study, in order to better understand the competencies for BDSE, first, the skill sets arranged by topic were identified from the data sets created using online job ads in this field. Then, these skill sets were mapped into competency domains. Based on these results, the following ten competencies were identified:

1. Big data frameworks
2. Big data processes
3. Big data analytics
4. Data processing types
5. Software development lifecycle
6. Programming
7. Software development frameworks
8. Vocational background
9. Soft skills
10. Work style

According to the results, in the BDSE field, Java, Python and Scala are the most demanded programming languages; Jenkins, Maven and Spring MVC presented as the most demanded programming tools; Hive, SQL, Hbase and NoSQL are listed as the most demanded databases; Hadoop and Spark are the most demanding big data tools; and finally, Java + Python + Hive, Java + Python + Scala, and Hadoop + Spark + Hive are tool combinations with the highest demand. The results of this study have important implications for software engineering programs, which are summarized below.

A. TOWARD REAL-TIME AND SCALABLE ARCHITECTURE

The discovered topics indicate a transformation from the traditional software architecture to real-time and scalable architecture. This revolution in software architectures aims to eliminate the challenges encountered in the development process of BDSAs considering that velocity and volume are the main characteristics of big data. In this process, real time and scalable architecture is considered as a solution to overcome the difficulties experienced due to high velocity and high volume of big data [26], [42]. Real-time applications on continuous systems using big data streams require effectively implementing processes, such as coding, testing, integration, and automation in a real-time and scalable manner. In particular, the topics of design and scaling, real-time programming, scalable programming, real-time data, and streaming processing clearly demonstrate the tendency toward such architecture (see Tables 1 and 2).

B. DOMINANCE OF CLOUD-BASED SERVICES AND APPLICATIONS

Among the major themes of the discovered topics demonstrating the BDSE knowledge domains and skill sets, many were related to cloud-based services and applications, indicated by the high rates of the topics of cloud and distributed systems. Likewise, the fact that numerous tools and technologies used to develop BDSA are cloud-based supports this idea. The tools revealed by our analysis, such as Python, Go, Perl, Php, Kafka, Storm, Flume, Aws, Hbase, Cassandra, and Pig are commonly used for the development of cloud-based services and applications. These findings indicate that developers prefer cloud-based platforms over traditional platforms to develop BDSAs, as also stated in other studies [1], [42]. This is because cloud computing resources provide numerous solutions for BDSE specialists to store, manage and process big data. Consequently, the findings obtained from the analysis present cloud-based services and applications as an innovative area offering numerous opportunities of research, practice, and employment for specialists. For these reasons, the impact of cloud-based services and applications on BDSE cannot be disregarded.

C. FROM SOFTWARE ENGINEER TO DATA SCIENTIST

Software engineering has been a pioneering discipline, which is closely related with data science. With the recent developments in data-driven technologies, the functional role of software engineering in data science has become even more evident. In particular, the advent of big data phenomenon has led to noteworthy changes in data-driven knowledge and skills needed in software engineering. As seen in the findings presented in this paper (Table 2), of all knowledge and skills, those related to big data have a total rate of 31%. More specifically, according to the findings, there are four main competencies in BDSE related to big data, consisting of big data frameworks, big data processes, big data analytics, and data processing types, in addition to software engineering knowledge and skills. In light of these findings, it can be concluded that expertise in BDSE requires a strong data-driven background, as well as traditional software engineering skills [5], [18]. Besides, the extensive use of data-oriented programming languages, such as Python, R, Scala, and Julia, and big data tools is another indicator of the big data-driven evolution in software engineering. From this perspective, today's software engineers can also be seen as data scientists.

D. TRANSITION FROM DATABASES TO DATA WAREHOUSES

A relational database is defined as a system of manageable data stores that are queried and updated using data management language expressions. A database is the basic building block for data solutions. Databases are designed to easily access, read, write, search and delete the data they hold. They are mostly used for online transaction processing (OLTP). On the other hand, a data warehouse is a database specifically designed for storing, filtering, retrieving, and analyzing huge data collections or a group of databases. A data warehouse is a system that combines data from many different sources within an organization for the purposes of analysis, visualization and reporting. The analysis reports obtained from

complex queries contained in a data warehouse are also utilized in decision support systems. Data warehouses are used for online analytical processing (OLAP) that involves complex queries for analysis and reporting, rather than OLTP.

As a result, data warehouses and databases are both relational data systems designed for different purposes. Data warehouses are no better than databases, or vice versa. They perform various different functions, each designed for a specific task. Ultimately, decision-support and strategy-setting mechanisms in today's data-driven business environments necessitate fast and comprehensive data analysis. Especially with the advent of big data, the transition from databases to data warehouses has accelerated since the products and services based on BDSAs mostly consist of processing, analytics, and reporting operations of data [35]. As also indicated by our findings, the topics of streaming-data integration, data reporting-visualization, big data warehousing, and scaling (see Table 2) in the competency domain of big data processes are closely related to data warehousing processes. This transition is also revealed by the four topics in the competency domain of big data analytics; i.e., analytical skills, machine learning, decision-support and business intelligence'' (see Table 2). In addition, the fact that Hive, a data warehousing tool, ranks first among databases and data warehouses is another important finding that supports this idea (see Table 5).

E. POWER OF SOFT SKILLS

In the most general sense, soft skills are defined as skills, abilities and attitudes relevant to cognitive and personality traits required for the implementation of technical knowledge and skills in a workplace [8], [20]. The need for soft skills and technical skills required for information technology professionals is a key topic that has been frequently discussed in recent scientific research and industrial reports [8], [11], [20]. In a similar study conducted on big data jobs [8], soft skills concerning working in a team, leadership, and communication skills were identified. In an another study [9], the researchers emphasized the necessity of commutation skills for software developers. Taking a different perspective, another study investigated the soft skills required for processes in the software development life cycle to reveal the effects of these skills on the software development process [20]. The findings reported confirmed that soft skills were among the most in-demand skills by the industry.

According to our findings, a wide range of soft skills required for BDSE specialists are highly demanded. The findings related to soft skills include the topics of communication skills, problem solving, project management, and computational thinking (see Table 2). In particular, communication skills has the second highest rate among all topics. Furthermore, the topics included in the competency domain of work style can also be considered as soft skills. When viewed from this perspective, the total rate of soft skills is approximately 20% in all topics (see Table 2), which clearly reveals their power among all knowledge domains and skills.

F. NECESSITY OF COLLABORATION AND TEAM-WORKING

Collaboration and team-working refer to the capability to work together in synchronization with other team-members to achieve comprehensive projects requiring the collective use of individual skills through an effective sharing of tasks, knowledge, and experiences. Collaboration and team-working constitute a functional methodology effectively implemented in recent software development projects [20]. Effective collaboration and well-designed team-working are the key building blocks of an operational software project. Many researchers have discussed the essential concepts and approaches related to team roles and collaboration dynamics, which are necessary for today's dynamic software development environments [8], [20]. Likewise, our findings related to the competency areas of work style (see Table 2), namely team-working, leadership, collaborative environment, team-member, and initiative reveal the necessity of collaboration and team-working skills for BDSE.

G. THE WIDE SPECTRUM OF KNOWLEDGE DOMAINS AND SKILL SETS

Software engineering industry is one of the most demanded and fastest growing professional fields all over the world. This industry has extremely dynamic and competitive working environments with an ever-changing and progressing demand for knowledge, skills, and abilities t. Our analysis revealed the highly demanded knowledge domains and skill sets for BDSE. The findings of the analysis demonstrate that expertise in BDSE requires a wide spectrum of knowledge domains, skill sets, and abilities. Considering these findings, a conceptual competency map is proposed to organize these knowledge and skills. This map consists of the following ten competency domains: big data frameworks, big data processes, big data analytics, data processing types, software development lifecycle, programming, software development frameworks, vocational background, soft skills, and work style (see Table 2).

The discovered competencies demonstrate that expertise in BDSE has an interdisciplinary proficient background that requires the combined use of an extensive collection of both technical and soft skills [9]. Although the priorities of competencies differ from one position to another, employers in the BDSE industry generally demand a collection of both technical and soft skills, defined as the employability skill set. In this regard, our findings offer a more comprehensive viewpoint for BDSE employers to identify the employability skill set necessary for the effective assessment of employee candidates. The knowledge domains and skill sets also indicate the necessity of a demand-driven educational approach based on interdisciplinary collaboration in order to achieve competency-based software engineering education [14]. Our findings are also consistent with the industry reports and academic research that emphasize the use of technical and soft skills together based on an interdisciplinary background containing business science, data science, software engineering,

computer science, mathematics, statistics, and communication science [3], [8]–[11].

H. INSIGHTS INTO THE USE OF TOOLS AND TECHNOLOGIES

BDSE specialists use an extensive collection of tools and technologies containing programming languages, programming tools, databases, data warehouses, and big data tools and platforms to develop BDSAs in a more efficient way. In the choice of these tools, the market needs and trends are generally considered as a guide. The findings of this study provide wide-ranging insights into the use of these tools, technologies, and platforms targeting BDSE. The findings reveal that the most in-demand programming languages for BDSE are Java, Python and Scala. The importance of these three programming languages for big data development environments is also underlined by other studies [17], [43]. Likewise, the findings also show that Jenkins, Maven and Spring MVC are the most in-demand programming tools for BDSE. Besides, Hive is the most in-demand data warehouse tool, and SQL is the most demanded database tool for BDSE. Finally, Hadoop and Spark present as the most in-demand big data tools for BDSE, consistent with other studies [6], [43], [44]. Regarding the combined use of programming languages and database or data warehouses, Java + Python + Hive' is the most in-demand triple combination. For the combined use of big data tools, the Hadoop + Spark + Hive triple combination has the highest demand in the BDSE field.

VI. CONCLUSION

Based on the results, it can be concluded that to satisfactorily meet the industrial requirements, the BDSE tools and skills identified in this research can be considered in software engineering education and related life-long learning programs. Additionally, the method proposed in this study can be improved to automatically analyze the contents of different job ads and regularly update the list of competencies based on the new requirements of the industry. In this way, the communication level of the industry and software engineering education programs can be improved significantly, and graduates of software engineering programs can gain appropriate skills to fulfill the requirements of the industry.

REFERENCES

[1] M. D. Assunção, R. N. Calheiros, S. Bianchi, M. A. S. Netto, and R. Buyya, "Big Data computing and clouds: Trends and future directions," *Inf. Sci.*, vol. 275, pp. 314–347, Dec. 2013.

[2] M. Kantardzic, *Data Mining: Concepts, Models, Methods, and Algorithms*, 2nd ed. Hoboken, NJ, USA: Wiley, 2011.

[3] A. Metzger, "Software engineering: Key enabler for innovation," NESSI, Berlin, Germany, White Paper, 2014.

[4] H. Chen, R. H. L. Chiang, and V. C. Storey, "Business intelligence and analytics: From big data to big impact," *MIS Quart.*, vol. 36, no. 4, pp. 1165–1188, Dec. 2012.

[5] X.-W. Chen and X. Lin "Big data deep learning: Challenges and perspectives," *IEEE Access*, vol. 2, pp. 514–525, 2014.

[6] C. L. P. Chen and C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on big data," *Inf. Sci.*, vol. 275, pp. 314–347, Aug. 2014.

[7] X. Wu, H. Chen, G. Wu, J. Liu, Q. Zheng, X. He, A. Zhou, Z.-Q. Zhao, B. Wei, M. Gao, Y. Li, Q. Zhang, S. Zhang, R. Lu, and N. Zheng, "Knowledge engineering with big data," *IEEE Intell. Syst.*, vol. 30, no. 5, pp. 46–55, Sep./Oct. 2015.

[8] 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, Mar. 2018.

[9] F. Gurcan and C. Kose, "Analysis of software engineering industry needs and trends: Implications for education," *Int. J. Eng. Educ.*, vol. 33, no. 4, pp. 1361–1368, 2017.

[10] J. Bosch, "Speed, data, and ecosystems: The future of software engineering," *IEEE Softw.*, vol. 33, no. 1, pp. 82–88, 2015.

[11] A. M. Moreno, M. I. Sanchez-Segura, F. Medina-Dominguez, and L. Carvajal, "Balancing software engineering education and industrial needs," *J. Syst. Softw.*, vol. 85, no. 7, pp. 1607–1620, 2012.

[12] N. H. Madhavji, A. Miranskyy, and K. Kontogiannis, "Big picture of big data software engineering: With example research challenges," in *Proc. 1st Int. Workshop Big Data Softw. Eng. (BIGDSE)*, 2015, pp. 11–14.

[13] K. M. Anderson, "Embrace the challenges: Software engineering in a big data world," in *Proc. 1st Int. Workshop Big Data Softw. Eng. (BIGDSE)*, 2015, pp. 19–25.

[14] V. Garousi, K. Petersen, and B. Ozkan, "Challenges and best practices in industry-academia collaborations in software engineering: A systematic literature review," *Inf. Softw. Technol.*, vol. 79, pp. 106–127, Nov. 2016.

[15] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *Int. J. Inf. Manage.*, vol. 35, no. 2, pp. 137–144, 2015.

[16] N. Khan, I. Yaqoob, I. Abaker T. Hashem, Z. Inayat, W. K. M. Ali, M. Alam, M. Shiraz, and A. Gani, "Big data: Survey, technologies, opportunities, and challenges," *Sci. World J.*, vol. 2014, Jul. 2014, Art. no. 712826.

[17] S. Debortoli, O. Müller, and J. vom Brocke, "Comparing business intelligence and big data skills," *Bus. Inf. Syst. Eng.*, vol. 6, no. 5, pp. 289–300, Oct. 2014.

[18] I. Gorton, A. B. Bener, and A. Mockus, "Software engineering for big data systems," *IEEE Softw.*, vol. 33, no. 2, pp. 32–35, 2016.

[19] B. Shahzad, A. M. Abdullatif, N. Ikram, and A. Mashkoor, "Build software or buy: A study on developing large scale software," *IEEE Access*, vol. 5, pp. 24262–24274, 2017.

[20] P. Holtkamp, J. P. P. Jokinen, and J. M. Pawlowski, "Soft competency requirements in requirements engineering, software design, implementation, and testing," *J. Syst. Softw.*, vol. 101, pp. 136–146, Mar. 2015.

[21] H. Gasmi and A. Bouras, "Ontology-based education/industry collaboration system," *IEEE Access*, vol. 6, pp. 1362–1371, 2017.

[22] S. Miller, "Collaborative approaches needed to close the big data skills gap," *J. Org. Des.*, vol. 3, no. 1, pp. 26–30, 2014.

[23] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers, *Big Data: The Next Frontier for Innovation, Competition, and Productivity*. New York, NY, USA: McKinsey, 2011, p. 156.

[24] P. Russom, "Big data analytics," *TDWI Best Practices Rep.*, vol. 19, no. 4, pp. 1–34, 2011.

[25] 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, 2018.

[26] H. Hu, Y. Wen, T.-S. Chua, and X. Li, "Toward scalable systems for big data analytics: A technology tutorial," *IEEE Access*, vol. 2, pp. 652–687, Jul. 2014.

[27] A. Aken, C. Litecky, A. Ahmad, and J. Nelson, "Mining for computing jobs," *IEEE Softw.*, vol. 27, no. 1, pp. 78–85, Jan./Feb. 2010.

[28] B. Prabhakar, C. R. Litecky, and K. Arnett, "IT skills in a tough job market," *Commun. ACM*, vol. 48, no. 10, pp. 91–94, Oct. 2005.

[29] D. Smith and A. Ali, "Analyzing computer programming job trend using Web data mining," *Issues Informing Sci. Inf. Technol.*, vol. 11, no. 1, pp. 1–12, 2014.

[30] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[31] D. M. Blei, "Probabilistic topic models," *Commun. ACM*, vol. 55, no. 4, pp. 77–84, Apr. 2012.

[32] H. M. Wallach, "Topic modeling: Beyond bag-of-words," in *Proc. ICML*, 2006, pp. 977–984.

[33] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proc. Nat. Acad. Sci. USA*, vol. 101, no. 1, pp. 5228–5235, 2004.

[34] T. L. Griffiths, M. Steyvers, and J. B. Tenenbaum, "Topics in semantic representation," *Psychol. Rev.*, vol. 114, no. 2, pp. 211–244, 2007.

[35] F. Gürcan, "Major research topics in big data: A literature analysis from 2013 to 2017 using probabilistic topic models," in *Proc. Int. Conf. Artif. Intell. Data Process. (IDAP)*, 2018, pp. 1–4.
[36] *Job Search | Indeed*. Accessed: Mar. 26, 2018. [Online]. Available: https://www.indeed.com/
[37] A. Kyriakopoulou and T. Kalamboukis, "The impact of semi-supervised clustering on text classification," *Inf. Process. Manag.*, vol. 50, no. 1, p. 180, Jan. 2013.
[38] A. N. Srivastava and M. Sahami, *Text Mining: Classification, Clustering, and Applications*. Boca Raton, FL, USA: CRC Press, 2009.
[39] A. Karl, J. Wisnowski, and W. H. Rushing, "A practical guide to text mining with topic extraction," *Wiley Interdiscipl. Rev. Comput. Stat.*, vol. 7, no. 5, pp. 326–340, 2015.
[40] A. K. McCallum. (2002). *MALLET: A Machine Learning for Language Toolkit*. [Online]. Available: http://Mallet.Cs.Umass.Edu
[41] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-6, no. 6, pp. 721–741, Nov. 1984.
[42] A. Fernández, S. del Río Victoria, L. Abdullah, B. María, J. del Jesus José, M. Benítez, and F. Herrera, "Big data with cloud computing: An insight on the computing environment, MapReduce, and programming frameworks," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 4, no. 5, pp. 380–409, 2014.
[43] L. Rodríguez-Mazahua, C.-A. Rodríguez-Enríquez, J. L. Sánchez-Cervantes, J. Cervantes, J. L. García-Alcaraz, and G. Alor-Hernández, "A general perspective of big data: Applications, tools, challenges and trends," *J. Supercomput.*, vol. 72, no. 8, pp. 3073–3113, 2016.
[44] F. Gürcan and M. Berigel, "Real-time processing of big data streams: Lifecycle, tools, tasks, and challenges," in *Proc. 2nd Int. Symp. Multidisciplinary Stud. Innov. Technol. (ISMSIT)*, 2018, pp. 1–6.

FATIH GURCAN received the B.S. degree in statistics and computer science, the M.S. degree in computer engineering, and the Ph.D. degree in computer engineering from Karadeniz Technical University, Trabzon, Turkey, in 2001, 2009, and 2017, respectively. He was an Instructor with the Department of Informatics, Karadeniz Technical University, from 2001 to 2014, where he has been an Instructor with the Center for Research and Application in Distance Education, since 2015. His research interests include trend analysis, sentiment analysis, statistical topic modeling, engineering education, data mining, machine learning, big data analytics, and text mining.

NERGIZ ERCIL CAGILTAY received the degree in computer engineering and the Ph.D. degree in instructional technologies from Middle East Technical University, Turkey. She worked for commercial and government organizations as a Project Manager for more than eight years in Turkey. She was also with the Indiana University Digital Library Program as a System Analysis and a Programmer for four years. She has been with the Software Engineering Department, Atilim University, Turkey, since 2003, as an Associate Professor. Her main research interests include information systems, medical information systems, engineering education, instructional systems technologies, distance education, e-learning, and medical education.