International Journal of Engineering Education Vol. 34, No. 2(A), pp. 497–504, 2018
Printed in Great Britain

0949-149X/91 $3.00+0.00

Big Data Processing: A Graduate Course for Engineering Students*

JIANLIANG GAO, JINFANG SHENG and ZUPING ZHANG
College of Information Science and Engineering, Central South University, China. E-mail: gaojianliang@csu.edu.cn

In the era of big data, professional graduates mastering big data technology are in urgent demand for both academia and industry. Nowadays, it is timely to set up a new course about big data processing for engineering students. In this paper, we design the course of big data processing for the graduate students with engineering backgrounds. This course provides a general overview of the frontier field of big data processing. The main advantages of this new course include: (1) it is a new big data processing course and introduces the newest technology such as Hadoop and Spark, which is great helpful to engineering students; (2) the course includes theoretical knowledge, as well as practical aspects of big data processing, which makes this course more suitable for engineering graduates than the related big data training courses. The survey results of recent four years illustrate that this course is successful as an emerging course for engineering education.

Keywords: big data; emerging course; engineering education

1. Introduction

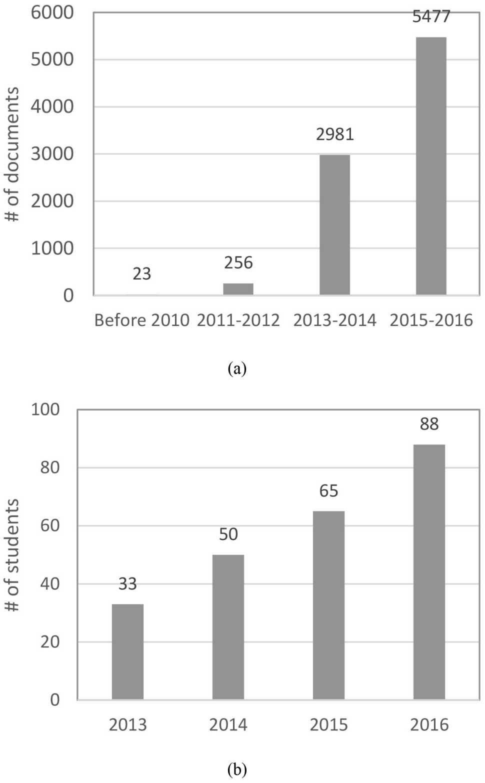
Course design is a key to train talents in the engineering field and many proposals present the approaches about how to design a new course [1, 2]. Nowadays, big data processing has attracted significant interest not only in academia, but also in a wide variety of industrial applications. Engineering students must acquire a solid foundation of frontier technology to prepare them for the wide ranging demands of their future careers [3]. In the big data era, universities and other academic transformed, need to consider the industry's increasing demand for engineers whose ability matches with the requirements of processing big data. This requirement applies to engineering disciplines such as computer sciences, electrical and electronic engineering. It is timely and necessary to set up a frontier course about big data processing for engineering students.

Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. Big data is tightly related to science, engineering, healthcare, finance, business, and ultimately our society itself. Although there are some data processing related courses such as data mining and machine learning [4], they cannot deal with the new challenges of big data processing. The features of big data can be described as five Vs: Volume, Variety, Velocity, Veracity and Value. These five Vs present challenges to existing technology which usually deals with structural, small scale data. For example, variety means various types of data, most of which are unstructured data. To obtain significant value from big data, new technologies must be learned in universities and institutes.

Besides the necessity, it is timely and feasible to set up the course of big data processing. The accumulated knowledge and technology of big data processing have paved the way toward a frontier graduate course. From the theoretical point of view, research on the method and development of big data can be considered as the first step toward this goal. Efforts in this direction have made rapid progress in both the depth and breadth of this field. As an indication of the ever-increasing interest and progress in this field, the rise in the number of scientific and technical documents related to big data is obvious. Fig. 1(a) shows the numbers of documents in IEEE Xplore Digital Library published with the word "big data" in their titles, abstracts, or index terms are 23, 256, 2981 and 5477 during four periods, i.e., before 2010, from 2011 to 2012, from 2013 to 2014, and from 2015 and 2016, respectively, which keeps a continuous increase.

Big data processing is suitable and popular as a graduate course. Different from undergraduate courses, graduate courses are encouraged to meet up with the frontier of research [6]. Graduate students might find research direction from the new course as there are still many challenges associated with big data including recognition of useful versus irrelevant data, efficient storage, privacy and security of data, intelligent analysis, etc. The authors began to set up this course for graduate students from 2013. The numbers of students who took this course have increased from 33 in 2013 to 88 in 2016, as shown in Fig. 1(b), which reflects the popular of this course. In the course survey, over 90% students thought this course was helpful to

* Accepted 12 December 2017.



(a)

(b)

Fig. 1. Continuous growth. (a) numbers of documents in IEEE Xplore digital library published with the word "big data" in their titles, abstracts, or index terms; (b) numbers of students who took this course in the authors' classes in the years from 2013 to 2016.

their research and even over 50% students would combine big data processing technology with their research. An average of 94% of the students would recommend the other students of the next year to take this course and it further reflects the degree of acceptance of this course.

A description of the course will be followed by detailed topics covered, including theoretical module and practical module. Course projects are also described and some advising projects are provided. The results of a student survey are discussed in the following. Finally, conclusions are drawn in the end of this paper.

2. Literature survey

Big data related courses are emerging in recent years. These courses are mainly divided into two categories. One category is to extend the traditional curriculums with big data background, and the other is to introduce the use of big data platforms such as Hadoop and Spark. In the following, we will survey the two categories in detail.

With the advent of big data age, many traditional curriculums such as data mining and machine learning are extended with combining big data background. For example, a data mining course which focuses on data warehousing and mining has been set up for student majoring in computer science [7]. By taking big data into account, this course includes the methods of extracting interesting knowledge from massive amount of data. In the online courses website Stanford Engineering Everywhere (SEE), traditional machine learning course is also extended with many big data applications such as autonomous navigation, bioinformatics and web data processing [8]. Combing the frontier research results, new courses are also developed in many universities. For instance, deep learning is widely used for image classification and manipulation, speech recognition and synthesis, natural language translation, self-driving cars, and many other activities. Therefore, a course of deep learning will be set up at Harvard University in spring term 2018 [9].

In addition, new courses on the use of big data platforms are also beginning to offer for engineering students. For example, a series of course in the field of data engineering and data analysis on the large scale is organized on Courser website including four concise courses: big data essential, big data analysis, big data applications on machine learning at scale, and big data applications on real-time streaming [10]. Many other online courses are also set up recently. For example, a short time Spark training course is designed on Cloudera website [11]. A course of introduction to Hadoop and MapReduce is opened to the fundamental principle behind Hadoop [12].

As a new big data processing course, both theoretical research and practical processing platforms should be included. In theoretical research, many proposals promote the development of big data from various perspectives: big data storage, big data analysis, and big data security, etc. From the practical aspect, Hadoop was first released based on Google file system paper [5]. In recent ten years, Hadoop has been deployed in numerous companies to deal with increasing big data. Furthermore, Spark is one of the newest frameworks in big data area. In order to make the contents of big data course solid, we design a big data course including both theoretical topics and practical platforms of big data processing.

3. Course description

Big data processing is a two-credit graduate-level course offered to engineering Master students over a 16-week semester, with two 50-min sessions per week. To be able to understand and analyze the engineering aspects of big data processing, students need to take three prerequisite courses prior to the big data processing course: computer algorithm, database system and computer programming [13].

Table 1. Course topics

	Topics	# of Sessions
Theoretical	Introduction of big data	2
	Big data storage: HDFS, NoSQL database	4
	Big data analysis: Uncertain, incomplete data Various type data Dynamic data	6
	Big data security, privacy	4
Practical	Hadoop programming MapReduce HBase Hive Mathout	8
	Spark programming Spark core Spark SQL Spark streaming MLib GraphX	8

In this course, students are required to understand the breadth and depth of big data. From the breadth aspect, theoretical knowledge is subsequently covered from big data storage to security. To deepen the students' understanding even more, they are required to design big data processing algorithms and implement them by programming. The topics of this course are divided into two modules: theoretical and practical. Table 1 shows the topics presented in the course along with the approximate time spent on each topic.

3.1 Theoretical module

As a new frontier course in the field of big data, it starts with introductory sessions about the background of big data. The first topic leads students to an overview of big data processing, including the necessary of new processing technology for big data, and the promotion function to industry which made by big data processing.

3.1.1 Introduction to big data

This course should first explain what big data is. In recent years, big data has become a hot term in academic and engineering area. But what is the intrinsic feature of big data? In this course, five Vs (Volume, Variety, Velocity, Veracity and Value) are adopted to explain the feature of big data. Volume refers to the vast amounts of data. It makes most data sets too large to store and analyze using traditional database technology. Variety refers to the different types of data. In big data era, over 80% of data is unstructured [19]. It requires big data technology to analyze and bring together data of different types. Velocity refers to the speed at which new data is generated and the speed at which data moves around. In this course, several streaming technologies are taught to deal with this challenge. Veracity refers to the messiness or trustworthiness of the data. It requires new technology to make up for the lack of quality or accuracy. Value refers to the ability of turning data into value and it is the ultimate goal of big data processing. Based on the features of 5Vs, the course develops the contents including big data storage, big data analysis, big data security and privacy.

3.1.2 Big data storage

Data storage is a basis of big data processing and how to store big data is the first challenge for ever-increasing big data. In the 5 Vs of big data, volume, variety and velocity are related to the challenge of storage. In this course, students are mainly taught two contents: distributed file systems and NoSQL database. Distributed file system is proposed as a solution of data storage, which requires simpler horizontal scaling to clusters of machines. In this course, Hadoop Distributed File System (HDFS) is introduced as a typical distributed file system. The difference between HDFS and traditional file systems is an emphasis. Especially, students are required to understand storage in the form of data blocks, and data blocks are distributed stored in multiple data nodes. In this part, only principle of block storage is introduced, and how to cope with HDFS is the content of practical module.

On the top of the distributed file system, NoSQL (referring to "not only SQL") database provides a mechanism for storage and retrieval of data which is modeled in means other than the tabular relations used in relational databases. The data structures used by NoSQL databases (e.g., key-value, wide column, or graph) are different from those used by default in relational databases, making some operations faster in NoSQL. In this course, three types of NoSQL databases are taught. (1) Key-value: Key-value structure is widely adopted in big data processing. It uses the associative array as fundamental data model. Data is represented as a collection of key-value pairs, such that each possible key appears at most once in the collection. It provides a convenient way to get or store data even the data set is very large. (2) Graph: As an unstructured data, graph can present many real-world networks which are consisted of elements interconnected with a finite number of relations between them [20]. The type of data could be social relations, public transport links, road maps or network topologies. (3) Wide column: Wide column storage uses tables, rows, and columns as relational database. But unlike a relational database, the names and format of the columns can vary from row to row in the same table. In the practical module, HBase is an according wide column storage database.

3.1.3 Big data analysis

After big data storage, this course moves to big data analysis. Big data analysis converges the database, artificial intelligence, machine learning, data mining, statistics and other fields of knowledge and becomes a key content in the field of data processing. Three aspects of big data analysis are included in this part teaching. (1) Analysis of uncertain, incomplete data. Uncertain and incomplete data are defining features for big data applications, which come from various sources. Each data field is no longer deterministic but is subject to some error distributions. This is mainly linked to domain specific application with inaccurate data readings and collections. Therefore, especial techniques are needed to eliminate these uncertainties or tolerate them during big data processing. (2) Analysis of various type data.

As big data applications are featured with multiple sources which results in various type data. A big data analysis system has to enable an information exchange and fusion mechanism to ensure that all distributed data sources with various types can work together to achieve a global optimization goal. (3) Analysis of dynamic data. Dynamic data is another factor for big data processing. The rise of big data is driven by the rapid increasing of complex data and their changes. Documents posted on WWW servers, social networks, communication networks, and transportation networks, and so on are all featured with dynamic data. Therefore, it is necessary to deal with dynamic data for big data processing.

3.1.4 Big data security and privacy

In this part, students first study the challenges of security and privacy of big data. Security and privacy challenges cover the entire spectrum of big data life cycle: sources of data production, the data itself, data processing, data storage, and data transport and data usage. For big data security risks, attack surface of the nodes in a cluster may not have been reviewed and servers adequately hardened. User authentication and access to data from multiple locations may not be sufficiently controlled. Significant opportunity for malicious data input and inadequate data validation. For big data privacy concerns, de-identified information can be re-identified by big data mining. Although data is anonymous before outsourcing, it is possible to re-associate anonymous data with specific individuals. In most situations, just removing the identifier is not an efficient method to protect data from attackers again. Obviously, it is a momentous issue that ensures the safety of data in big data processing. In this part, students are encouraged to propose novel methods of privacy protection and implement them as course project.

3.2 Practical module

Practical training is very important for engineering students [16]. Mastering practical big data programming is a crucial goal of setting up this course. Corresponded to the theoretical module, two big data platforms (Hadoop and Spark) are taught in practical module of this course.

3.2.1 Hadoop programming

Hadoop is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Hadoop includes mainly four parts: Hadoop Common, HDFS, Yarn, and MapReduce.

In this part, the focus of teaching is MapReduce programming. MapReduce [17] is a programming model for parallel processing of large data sets on a cluster. MapReduce calculation mode divides calculation process into two steps: Map and Reduce. In the stage of map. The original data that were input into mapper are going to be filtered and transformed. The obtained intermediate data will be regarded as input in Reduce period. After processing in the Reduce period, the terminal results will be output. Students are required to code parallel programs using MapReduce.

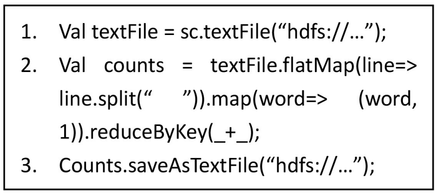
The following example illustrates the basic map and reduce functions. Each document is split into words, and each word is counted by the map function, using the word as the result key, as

```
Input: String name, String document
Output: (w, number)
foreach word w in document do
    Emit (w,1)
end
```

Algorithm 1. Map Function.

```
Input: String word, Iterator partialCounts
Output: (word, number)
Sum = 0;
foreach pc in partialCounts do
    sum += pc
end
emit (word, sum)
```

Algorithm 2. Reduce Function



```
Val textFile = sc.textFile(“hdfs://…”);
Val counts = textFile.flatMap(line=> line.split(“ ”)).map(word=> (word, 1)).reduceByKey(_+_);
Counts.saveAsTextFile(“hdfs://…”);
```

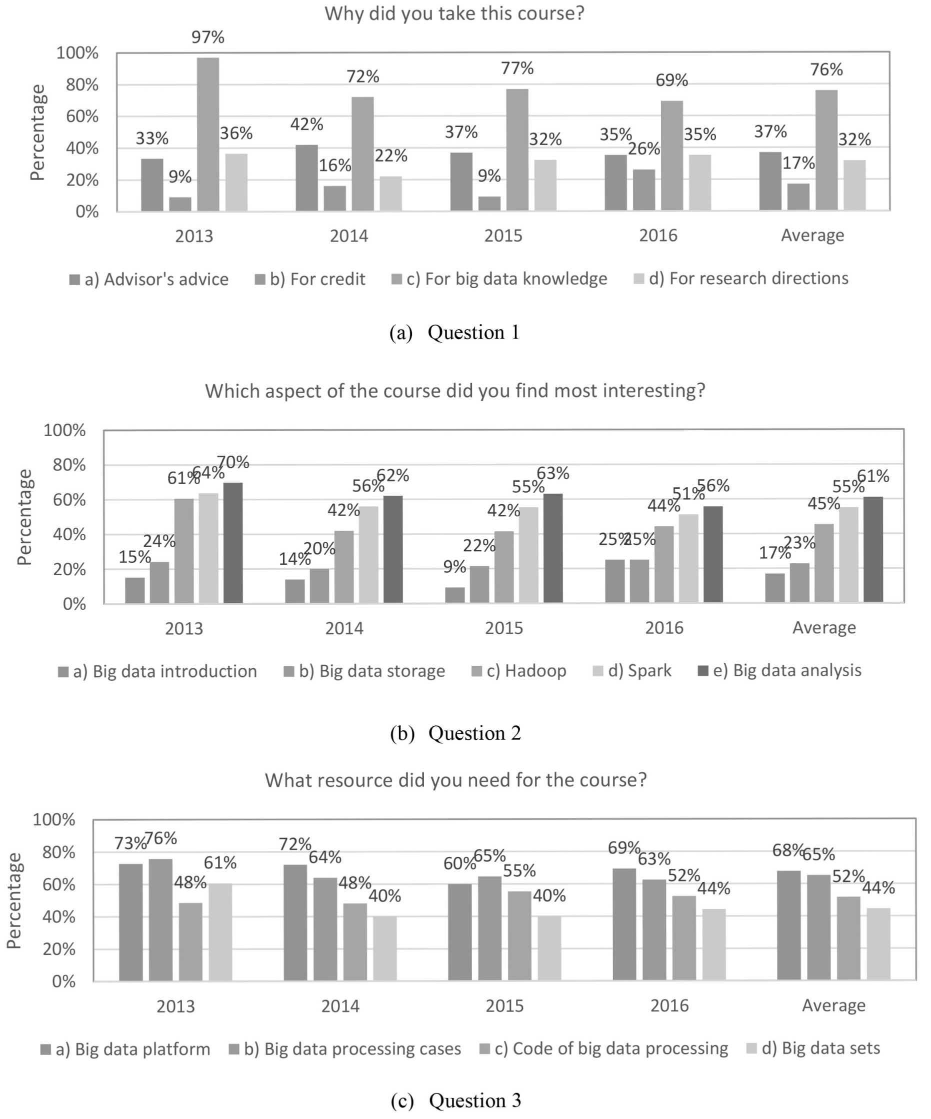
Algorithm 3. Word count using Spark.

shown in Algorithm 1. The framework puts together all the pairs with the same key and feeds them to the same call to reduce. Thus, this function just needs to sum all of its input values to find the total appearances of that word, as shown in Algorithm 2.

Base on MapReduce programming, three additional software libraries are presented as optional programming practice. The first one is HBase, which is the Hadoop database, a distributed, scalable, big data store. The second is Hive, which is a data warehouse infrastructure built on top of Hadoop for providing data summarization, query, and analysis. The last one is Mahout, which implements many distributed machine learning algorithms. Students can use partial or all three libraries to simplify basic MapReduce programming.

3.2.2 Spark programming

In order to keep up with the trend of technological development, this course introduces spark programming. Originally developed at the University of California, Berkeley’s AMPLab, the Spark was later donated to the Apache Software Foundation, which has maintained it since. Spark is developed in



(a) Question 1

(b) Question 2

(c) Question 3

Fig. 2. Student survey results of multiple-choice questions in Table 3. Questions 1–3 are provided with each individual subfigure (a)–(c), and the last group of each figure is the average results of the percentages from 2013 to 2016.

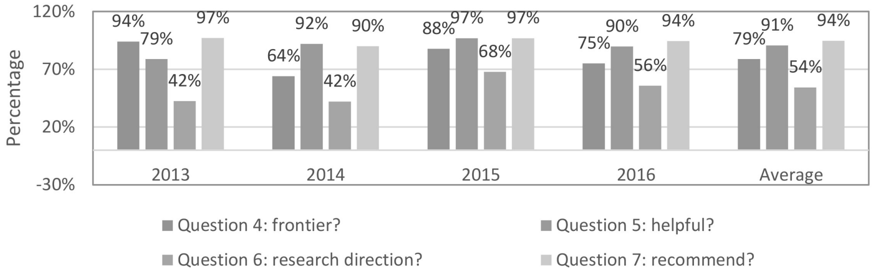


Fig. 3. Student survey results of Yes/No questions in Table 3. The percentage of answering "Yes" for Questions 4–7.

response to limitations in the MapReduce cluster computing paradigm, which forces a particular linear dataflow structure on distributed programs: MapReduce programs read input data from disk, map a function across the data, reduce the results of the map, and store reduction results on disk. With an application programming interface centered on a data structure called the resilient distributed dataset (RDD), Spark programming offers a restricted form of distributed shared memory. Algorithm 3 illustrates the basic Spark code for word count.

In addition to Spark core programming, the course includes several libraries: Spark SQL, MLlib, GraphX, and Spark Streaming. (1) Spark SQL: Spark SQL is a component on the top of Spark core that introduces a data abstraction called DataFrames, which provides support for structured and semi-structured data. It also provides SQL language support with command-line interfaces. (2) Spark Streaming: Spark Streaming leverages Spark Core's fast scheduling capability to perform streaming analysis. (3) Mlib: Spark MLlib is a distributed machine learning framework on top of Spark Core that, due in large part to the distributed memory-based Spark architecture, is as much as nine times as fast as the disk-based implementation used by Apache Mahout. (4) GraphX: it is a distributed graph processing framework on top of Spark. GraphX provides two separate APIs for implementation of massively parallel algorithms.

Table 2. Course Project

Project No.	Title	Techniques
1	Word count	HDFS, HADOOP, Spark
2	Log analysis	Streaming
3	Recommendation system	Mathout MLib
4	Big social network data analysis or Web data [19] analysis	Integrated Hadoop/ Spark

Table 3. Student Survey Questionnaire

Questions	Choices
1. Why did you take this course?	(a) Advisor's advice (b) For credit (c) For big data knowledge (d) For research directions in the field
2. Which aspect of the course did you find most interesting?	(a) Big data introduction (b) Big data storage (c) Hadoop (d) Spark (e) Big data analysis
3. What resource did you need for the course?	(a) Big data platform (b) Big data processing cases (c) Code of big data processing (d) Big data sets
4. Was this the most frontier course you have taken?	Yes / No
5. Was this course helpful to your research work?	Yes / No
6. Would you choose data processing as your research direction?	Yes / No
7. Would you recommend that next year's students take the course?	Yes / No
8. Degree of attraction	Score (1–5)
9. Degree of the uniqueness of course content	Score (1–5)
10. Degree of difficulty	Score (1–5)

4. Course projects

Students are assigned three short projects and a term project throughout the semester. Table 2 illustrates advising project topics. But students are encouraged to find projects related to their research fields such as stream data processing [14], healthcare data analysis [15], web data mining [18] and so on.

As shown in Table 2, the advising short projects include: (1) Word count. Students can learn HDFS operations and the procedure of using Hadoop and Spark by this simple project. (2) Log analysis. In this project, log records are generated in real time and students can learn streaming processing, especially Spark Streaming. (3) Recommendation system such as music or movie recommendation. This project requires machine learning and data mining techniques [4]. Students can choose Hadoop or Spark platform to implement it. The according library is Mahout and MLib.

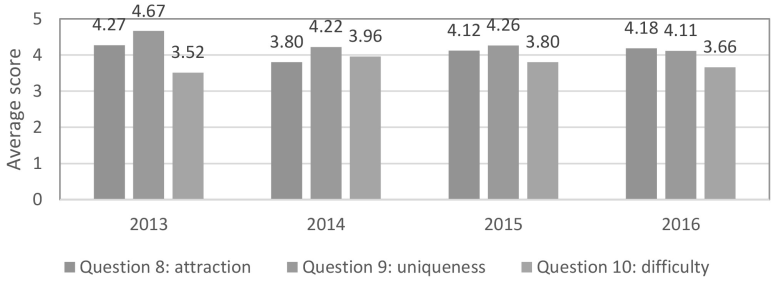


Fig. 4. Student survey results of scoring each question. The average scores for Questions 8–10 in Table 3.

For a term project, students are divided as groups of two or three persons. Each group is required to finish a real world big data processing such as social network data or web data [19]. This is an integrated project, including storage or data partition, analysis of unstructured data, and parallel computing problem [20]. Because this project deal with real world big data, students might meet various problems in the project, such as the efficiency problem. As a frontier graduate course, we encourage they solve the problems by group discussion and studying the newest documents.

5. Course survey

At the end of the course, the students were provided with a survey, shown in Table 3. There are three kinds of questions. The first three questions are multiple-choice about their reasons for taking this course, about the interesting course content, and about the resource construction of this course. For each question, students could select as many answers as they felt appropriate. For Questions 4–7, students answer just "Yes" or "No" about whether this course was frontier course or was helpful, and so on. The last three questions are to evaluate the content from the perspective of attractive, unique and difficulty. A score between 1 and 5 is given to each item. There is a steady increase of the number of students who took this course. From 2013 to 2016, the numbers of students are 33, 50, 65 and 88, respectively. It reflects that big data processing course is becoming more and more popular in recent years. In the following, we describe the survey results.

Figure 2 shows the results of Questions 1–3. These questions are all multiple-choice and the percentages of each choice from 2013 to 2016 are shown in the sub-figures. To overview the survey, we show the average results of the four years as the last group of each figure. According to Fig. 2(a), for more than 76% of the students, the course's knowledge was the key reason for taking this course. There is as high as 97% choose the item of big data knowledge as their motivation for taking this course in the first year of this course. The other main reasons for taking this course are advisor's advice and for their research directions. It reflects big data are attracting more interest not only for students, but also for advisors. Students' choice of which aspect of the course they found the most interesting is shown in Fig. 2(b). Over 60% students choose big data analysis as the aspect of the course, which reflects the correctness of taking big data mining as an important content of this course. Hadoop and Spark are the following choices as the most interesting aspects. In both industrial and academic area, Spark is becoming popular as a frontier big data processing framework. About the choice of what resource did you need for the course, the four items are all important for this course as shown in Fig. 2(c). Among them, big data platform and big data processing cases are the top two choices for Question 3. It hints the main direction for course resource.

To the second type of questions, students are required to answer Yes or No to Questions 4–7 and the results are shown in Fig. 3. For almost 80% of the students, this was their most frontier course they have taken. Most students think this course is helpful to their research work. And even over 50% students take big data related topic as their research directions. The consistent recommendation rate for this course is as high as 94%. In the four years, the recommendation rates are all over 90%. It illustrates the popularity of this course.

Figure 4 illustrates the results of the third type of questions, i.e., Questions 8–10. The scores for attractive of this course are ranged from 3.80 to 4.27. For Question 9, the scores are all over 4 in the four years. It reflects the contents of this course are unique and frontier. About the difficulty of this course, students gave their scores below 4 in all four years. It hints that the degree of difficulty is acceptable.

6. Conclusion

The volume and level of research achievements in the field of big data has reached to a point where the training of talents in engineering field must done in an organized fashion. On efficient way to do this is to develop new subject-specific academic courses at the graduate level. This paper has reported a new graduate-level course, focused on the design and development of big data processing, which provides students with an overview of the new area of frontier data science and its associated challenges. The focus of the course is then put on the engineering of big data processing.

The instructor's experience of teaching this course and the survey results over several course offerings give rise to the following concluding remarks. (1) Given the increasing interest in the field of big data processing, development of this course is not only timely, but necessary to speed its academic and industrial growth and advancement. (2) As a new graduate course for engineering students, it is correct to place an emphasis on practical programming. In the topics of the course, Hadoop and Spark programming is considerable important. Survey results also illustrate the popularity and feasibility of practical topics.

References

1. K. Nohara, M. Norton and E. Kawano, Imparting Soft Skills and Creativity in University Engineering Education through a Concept Designing Short Course, *International Journal of Engineering Education*, **32**(6), 2017, pp. 538–547.
2. G. Herman, D. Goldberg, K. Trenshaw, M. Somerville and J. Stolk, The Intrinsic-Motivation Course Design Method, *International Journal of Engineering Education*, **33**(2), 2017, pp. 558–574.
3. A. Zendler, C. Spannagel and D. Klaudt, Marrying content and process in computer science education, *IEEE Transactions on Education*, **54**(3), 2011, pp. 387–97.
4. X. Wu, X. Zhu and G. Wu, Data mining with big data*, IEEE Transactions on Knowledge and Data Engineering*, **26**(1), 2014, pp. 97–107.
5. S. Ghemawat, H. Gobioff and S. Leung, The Google file system, *ACM SIGOPS Operating Systems Review*, **37**(5), 2003, pp. 29–43.
6. G. Tartarini, M. Barbiroli and F. Fuschini, Consolidating the electromagnetic education of graduate students through an integrated course, *IEEE Transactions on Education*, **56**(4), 2013, pp. 416–423.
7. http://web.engr.illinois.edu/~hanj/cs412/bk3_slides/01Intro.pdf, accessed 1 Dec. 2017.
8. https://see.stanford.edu/Course/CS229, accessed 1 Dec. 2017.
9. www.extension.harvard.edu/academics/courses/deep-learning/25120 , accessed 1 Dec. 2017.
10. https://www.coursera.org/specializations/big-data-engineering#courses, accessed 1 Dec. 2017.
11. https://www.cloudera.com/more/training/courses/spark-training.html, accessed 1 Dec. 2017.
12. https://www.class-central.com/mooc/1470/udacity-intro-to-hadoop-and-mapreduce , accessed 1 Dec. 2017.
13. G. S. Maceda, P. D. Arjona-Villicana and F. E. Castillo-Barrera, More time or better tools? a large-scale retrospective comparison of pedagogical approaches to teach programming, *IEEE Transactions on Education*, **59**(4), 2016, pp. 274–281.
14. F. Yan and J. Gao, Reliable NoC design with low latency and power consumption, *Electronics Letters*, **53**(6), 2017, pp. 382–383.
15. J. Gao, B. Song, W. Ke and X. Hu, BalanceAli: multiple PPI network alignment with balanced high coverage and consistency, *IEEE Transactions on Nanobioscience*, **16**(5), 2017, pp. 333–340.
16. A. Gero, Y. Stav and N. Yamin, Increasing Motivation of Engineering Students: Combining "Real World" Examples in a Basic Electric Circuits Course, *International Journal of Engineering Education*, **32**(6), 2016, pp. 2460–2469.
17. J. Dean, and S. Ghemawat, MapReduce: simplified data processing on large clusters, *Communications of the ACM*, **51**(1), 2008, pp.107–113.
18. H. Chen, X. Li and M. Chau, Using open web APIs in teaching web mining, *IEEE Transactions on Education*, **52**(4), 2009, pp. 482–490.
19. Web data sets: http://webdatacommons.org, accessed 30 April 2017.
20. Y. Zhang, T. Cao and S. Li, Parallel processing systems for big data: a survey, *Proceedings of the IEEE*, **104**(11), 2016, pp. 2114–2136.

Jianliang Gao is currently with College of Information Science and Engineering, Central South University, China, as an associate professor, and he is also a visiting professor in Drexel University, USA. He received a PhD in computer science at the institute of computing, Chinese Academy of Sciences. His research interests include big data processing, parallel computing and engineering education. He is the founder of the course of big data processing at Central South University. He severed as the general chair of 2016 IEEE International Conference on Big Data.

Jinfang Sheng is currently with College of Information Science and Engineering, Central South University, China, as an associate professor. She received a PhD in computer science at Central South University. She has been a cofounder of the course of big data processing since 2013 at Central South University. Her research interests include database and education.

Zuping Zhang is currently with College of Information Science and Engineering, Central South University, China, as a professor. He has severed as the chair of the department of computer science since 2014. He received a PhD in information and science and engineering at Central South University in 2005. His research interests include big data processing and engineering education.