Educational and Information Technologies (2023) 28:10209–10234
https://doi.org/10.1007/s10639-022-11558-8

Data science technology course: The design, assessment and computing environment perspectives

Azlan Ismail[1,2] · **Sofianita Mutalib**[2,3] · **Haryani Haron**[2]

Received: 17 December 2021 / Accepted: 22 December 2022 / Published online: 24 January 2023

Abstract

This article discusses the key elements of the Data Science Technology course offered to postgraduate students enrolled in the Master of Data Science program. This course complements the existing curriculum by providing the skills to handle the Big Data platform and tools, in addition to data science activities. We tackle the discussion about this course based on three main requirements, which are related to the need to exploit the key skills from two dimensions, namely, Data Science and Big Data, and the need for a cluster-based computing platform and its accessibility. We address these requirements by presenting the course design and its assessments, the configuration of the computing platform, and the strategy to enable flexible accessibility. In terms of course design, the offered course contributes to several innovative elements and has covered multiple key areas of the data science body of knowledge and multiple quadrants of the job and skills matrix. In the case of the computing platform, a stable deployment of a Hadoop cluster with flexible accessibility, triggered by the pandemic situation, has been established. Furthermore, through our experience with the implementation of the cluster, it has shown the ability of the cluster to handle computing problems with a larger dataset than the one used for the semesters within the scope of the study. We also provide some reflections and highlight future improvements.

Keywords Big Data Technology · Data Analytics · Data Science · Computing Environment · Hadoop Cluster

✉ Azlan Ismail
azlanismail@uitm.edu.my

Extended author information available on the last page of the article.

1 Introduction

Data Science and Analytics (DSA) technologies and methods are affecting the economy, as well as the job market (Miller, 2017). As the demand for DSA workers grows, it puts pressure on the supply of DSA talent to grow, in turn. In the context of Malaysia, the Malaysia Digital Economy Corporation (MDEC) stated that Big Data Analytics (BDA) is central to the Malaysian digital economy, resulting in the growth of other digital technologies (e.g., artificial intelligence (AI), the Internet of Things (IoT), and advanced automation). Furthermore, its market is expected to grow to $1.9 billion (approximately RM7.85 billion) in 2025 (Shankar, 2021). Malaysian universities have supported this agenda by offering Data Science or Data Analytics programs at both the undergraduate and postgraduate levels. It is also important to note that the Data Science and Analytics course at the postgraduate level has been offered at many reputable universities.

As part of contributing to this call, the School of Computing Sciences (formerly known as the Faculty of Computer and Mathematical Sciences), College of Computing, Informatics and Media, Universiti Teknologi MARA (UiTM) is offering a three-semester Master of Data Science program that provides in-depth knowledge of data-driven science. Its aim is to produce data professionals who are passionate about drawing meaningful insights from data using the data science approach. Within the context of the program, we focus on the course offered within the program called Advanced Data Science Technology. The course aims to provide students with the opportunity to learn and grasp the knowledge and skills of the selected big data science tools and techniques. With the knowledge and skills exercised, students can acquire the decision-making capability to choose the right tool for data science-related jobs in the future. The course is offered in the second semester and complements other courses from a technological perspective. For the job market, this course can prepare students to fill the roles of data scientists or data analysts with some data engineering skills.

From a literature perspective, several articles have reported different aspects of teaching and learning in DSA. Some of them addressed the program level of DSA. Specifically, Adams, (2020) addressed the need to balance statistical and computer science topics in a program. Oudshoorn et al., (2020) focused on developing a DSA related program based on an existing computer science program. The article by Salloum et al., (2021) promotes an interdisciplinary program related to DSA. Meanwhile, Fekete et al., (2021) introduced a data science major to a data-centric computing curriculum. Other aspects include the proposal for the development of pedagogical data sets by Bart et al., (2018), the infrastructure and tools for teaching data science throughout the statistical curriculum by Çetinkaya-Rundel & Rundel, (2018), a survey of teaching experience from the perspective of practitioners by Kross & Guo, (2019), and a recommendation of key skills for a data science curriculum, namely creating, connecting, and computing by Hicks & Irizarry, (2018).

In this paper, we focus on the design and implementation of the course. In particular, we address three main requirements or questions, given as follows:

- First, the question of "What are the innovative elements offered by the proposed course?". We need to design a course that offers necessary skills that fulfill two dimensions, Data Science and Big Data. Typically, the data science dimension emphasizes the application of the programming language and related tools to carry out data science activities (e.g. data processing and modeling). Meanwhile, the big data dimension emphasizes the use of big data frameworks or tools to execute data science activities. Our review of the published courses presented in the literature (in Section 2) has shown that there are limited courses that cover both dimensions. Therefore, a course that takes advantage of the skills of both dimensions is needed to provide more added value to students and prepare them with relevant skills (Miller, 2017). These dimensions should also be reflected in the course assessment.
- Second, the question of "How to deploy a cluster environment to support teaching and learning?". There are many tools and services that can be applied to support the teaching and learning activities. However, having these tools run on individual computers (either in the student's owned computer or the one provided in the computer lab) may limit the experience of dealing with big data ecosystems. Furthermore, setting up a big data ecosystem on the student-owned computer may be dealing with configuration issues or limited computing resources, which negatively affect the learning process. Meanwhile, setting up a big data ecosystem in the individual computer lab may require a more intensive maintenance effort. Therefore, having a computing cluster that provides the big data ecosystem is significant and can potentially improve the learning experience of students.
- Third, the question of "How can we ensure the usage of the computing platform and provide flexibility in accessing it to facilitate teaching and learning?". The need to address this question is significant, especially when dealing with a situation that limits the face-to-face learning environment (i.e., pandemic situation). Accessibility to the computing platform via the Intranet and the Internet is crucial, since learning this kind of course requires a lot of hands-on activities to grasp key concepts. Therefore, an accessibility strategy must be outlined and implemented, taking into account the current support and security aspects of the infrastructure.

The rest of the sections are organized as follows. Section 2 summarizes the related work and highlights the innovative elements of the proposed course. In Section 3, we present the design of the course and its assessment together with the computing platform and accessibility strategy. Section 4 elaborates on experience and implementation in terms of assessment deliverable, user management for cluster accessibility, and cluster utilization. We then conclude this paper with reflections as in Section 5.

2 Related work

We conducted a comparison study to determine the similarity and difference of the offered course with existing proposed courses from the literature. The aim is not to identify which is better, but rather to point out the innovative elements of the proposed course within a certain context. For this reason, we have selected a set of articles that meet the following criteria:

- The article solely focus on the course context, thus the curriculum, or review, or guideline are excluded.
- The article reported a course related to *Data Science* (DS) and/or *Big Data* (BD) terms.
- The reported course offers the practical contents (e.g. programming and utilizing tools), not only conceptual contents.

Table 1 provides a summary of these courses together with the proposed course.

As shown in Table 1, there are several courses that focus on DS (i.e., Baumer, 2015; Dichev et al., 2016; Çetinkaya-Rundel & Ellison, 2020; Donoghue et al., 2021) or only the dimension of BD (i.e., Ngo et al., 2014; DePratti et al., 2017). There are also courses that covered both dimensions (i.e., Brunner & Kim, 2016, Eckroth, 2016, 2017, 2018). On the DS side, courses commonly incorporate Python, R, and SQL programming skills. The computing environment used to support programming activities is typically the Jupyter Notebook and RStudio. For the BD side, the courses include the use of Hadoop, which runs on a cluster, cloud platform, or local machine.

The closest course to our course, which covers both dimensions, DS and BD, is the one reported by (Eckroth 2016, 2017, 2018). Similarity includes Python and SQL programming skills with the use of Hive, Spark, and MySQL for the computing environment and the application of Big Data frameworks in a cluster environment. In terms of differences, our course does not cover R programming, Storm, Big-Query, and Mahout frameworks, as well as the use of Google Cloud Platform. Having said that, our course offers the following innovative elements:

- In terms of programming, the course gives specific attention to Spark, that is, the application of Spark-Scala, PySpark, and Spark-SQL. It provides an alternative to students in choosing the appropriate Spark library when addressing the data science problem. Additionally, this course integrates the exploration of SQL in Impala as an alternative to Hive-QL. Therefore, students can appreciate the difference in terms of their ability and performance.
- In relation to the BD framework, the course also includes the use of Sqoop to instill data ingestion skills from the Relational Database Management System (RDBMS), namely MySQL.
- The computing environment for BD in this course is based on the Cloudera Hadoop distribution, which provides several distinct tools. In particular, the student can use Cloudera Manager, which acts as a cluster management tool, to explore about the hosts, and performance metrics such as application execution time and resource utilization. This exploration will help them understand the

Table 1 Comparison with Related Courses [PL: Programming language, DS: Data Science, BD: Big Data, CE: Computing Environment

Authors	PL for DS	CE for DS	BD Framework	CE for BD
(Ngo et al., 2014)	–	–	Hadoop	Hadoop-Cluster
(Baumer, 2015)	R, SQL	–	–	–
(Dichev et al., 2016)	Python	–	–	–
(Brunner & Kim, 2016)	Python, SQL	Jupyter Notebook, SQLite	Hadoop	Google Computing Platform
(Eckroth, 2016; 2017; 2018)	Python, R, SQL	OpenCV, RStudio, MySQL, Hive, Spark	Hadoop, MapReduce, Hive, Mahout, Spark, BigQuery, Storm	Hadoop-Cluster, Google Cloud Platform
(Dichev & Dicheva, 2017)	Python	Jupyter Notebook	–	–
(DePratti et al., 2017)	–	–	Hadoop, MapReduce, Spark	Local Machine
(Çetinkaya-Rundel & Ellison, 2020)	R	RStudio Cloud	–	–
(Donoghue et al., 2021)	Python	Anaconda, Jupyter Notebook	–	–
Offered Course	Python, SQL, Spark-Scala, PySpark, Spark-SQL	Hive, MySQL, Impala, Spark, Anaconda, Jupyter Notebook	Hadoop, MapReduce, Sqoop, Hive, Impala, Spark	Cloudera Hadoop-Cluster

effect of processing BD in certain contexts. Furthermore, the Cloudera distribution also provides HUE system where the students can get access to Hadoop related services (i.e. HDFS, Hive, Impala, Spark) via web interface.

In addition to the innovative elements, we also learn the opportunities to improve the course from this comparison. We mention them as part of the reflection in Section 5.

3 Methodology

This study is driven by the action research methodology (Eilks, 2018). Action research is well suited since the ultimate goal is to continuously improve the course driven by the spiral activities of action and reflection. Therefore, to address the requirements mentioned above, we present the proposed course, the benchmarking study conducted, the course assessment, the computing platform, and the strategy to enable flexible accessibility.

3.1 Course design

The aim of this course is to expose students to data science technologies within the Hadoop ecosystem. It also requires them to become familiar with the Python environment so that they have the option to choose in relation to real-world challenges. Typical learning levels attributes referred to the Bloom taxonomy include knowledge, comprehension, application, analysis, synthesis, and evaluation. Meanwhile, the psychomotor domain includes perception, set, guided response, mechanism, complex overt response, adaptation, and organization. Finally, for the affective domain, there are receiving, responding, valuing, organization, and characterization. In this paper, we share learning and teaching activities for the cognitive and psychomotor domains. In addition, there are three levels of knowledge for learning outcomes in the data science model curricula, namely the level of familiarity, the level of use, and the level of evaluation (Wiktorski et al., 2020). Through this course, students will practice the relevant skills to equip them with the selected tools for data science and analytic tasks, as well as for familiarity level and usage level learning outcomes. They will need to demonstrate the ability to develop a program to address data science problems for assessing the learning outcomes. This course applies several methods to teaching students, including lectures, lab tutorials, and problem-solving scenarios. Assessments include two assignments, one project, and one test.

The course is designed for 14 academic weeks, including assessments. It is divided into two main parts. The first part (relatively 7 weeks) focuses on the fundamental concepts, data management, and data analytics (i.e. descriptive) using the Hadoop environment. The Hadoop part covers the application and utilization of storage management (i.e. HDFS), data extraction tool (i.e., Sqoop), data warehousing tools (i.e., Hive and Impala), and data processing tool (i.e., Sparks). The main factor driving the selection of these tools was due to the training of the trainers to whom we had participated using Cloudera Quickstart VM. The training helped us realize the role

of Hadoop as the distributed data storage and computing framework, as well as its connection to other tools as an ecosystem. We believe that the framework concept should be understood by students to appreciate the platform that enables data science tasks for Big Data. The suitability of the tools is gauged through observation and evaluation.

The second part (another 7 weeks) is dedicated to allowing students to explore data science activities using Python. This part also gives the programming exposure to the students, although it is not in-depth in terms of addressing the fundamental programming skills. The students applied the programming technique to implement data engineering activities due to messy and dirty data, which includes cleaning, analyzing, and transforming data. The use of software tools may speed up the process but limit the customization of codes, especially when automation is needed. Whenever data are prepared, it can be used as input for different types of analytics, specifically descriptive and predictive analytics. Descriptive analytics are very useful for exploring trends or patterns that occur in the dataset. Its methods include exploratory data analysis and summarization. Meanwhile, predictive analytics is commonly used to predict future events or events by learning from the historical dataset or labeled data. Existing Python packages such as Scikit-Learn or Tensorflow can be used for predictive purposes. In addition to that, the second part also includes the development of the application to illustrate the use of the analytics part, typically in terms of a dashboard. At the end of the second part, students would be able to solve relevant problems by applying the knowledge gained from fundamental data management and data analytics.

3.2 Benchmarking of course contents

The benchmarking study is carried out to determine the coverage of the course content within the DSA landscape. In this section, we focus on mapping the course topics into two main references, namely, the Data Science Body of Knowledge (DS-BoK) and the Data Science Job Skills matrix.

DS-BoK (Cuadrado-Gallego & Demchenko, 2020) refers to the knowledge areas introduced to support the Data Science Competence Framework (Demchenko & Cuadrado-Gallego, 2020). DS-BoK provides the basis for defining data science-related curricula, courses, instructional methods, etc. Among the knowledge areas, we focus on the engineering knowledge areas (KAG02-DSENG) (Wiktorski et al., 2020) as it relates to the engineering principles and aspects of computer technology in this course. The knowledge areas cover software and infrastructure engineering, manipulating and analyzing complex, high volume, high-dimensionality data, structured and unstructured data, cloud-based data storage, and data management.

Meanwhile, the dimension of the job market refers to the DSA job and skill matrix discussed in (Miller, 2017) which have been classified into four quadrants, as shown in Fig. 1.

Based on these two references, we provide a mapping of the skills learned in the course, as illustrated in Table 2. As shown in the table, we can conclude that the

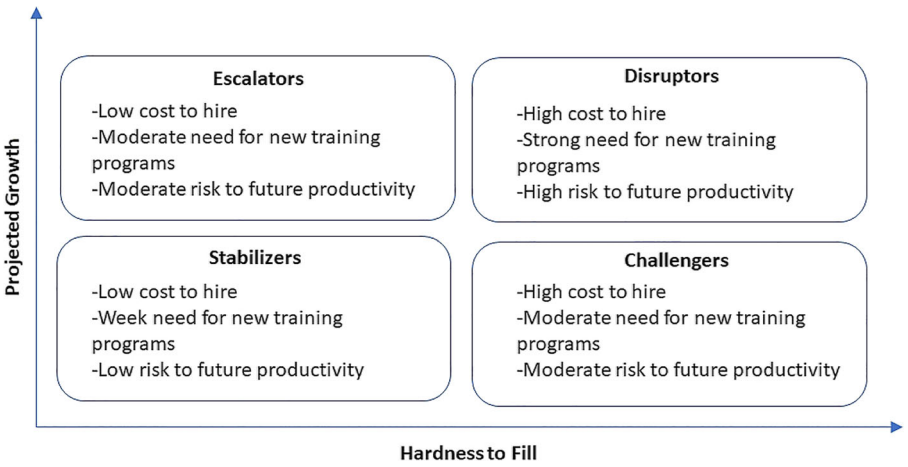

Fig. 1 DSA Job and Skill Matrix (Source taken from (Miller, 2017))

content of this course can be mapped to three aspects of the engineering knowledge areas, namely:

- KA02.02 (DSENG.02/DSIAPP) - Infrastructure and platforms for data science applications, including typical frameworks such as Spark and Hadoop, data processing models, and consideration of common data inputs at scale.

Table 2 Mapping of Course Topics to Curriculum and Job Skills Matrix Dimensions

Selected Topics	Related Tools / Libraries	Skills to learn	Skill Mapping
Hadoop ecosystem	HDFS, Cloudera Manager, HUE, YARN	Data and File Management, Cluster Monitoring	Disruptors KA02.02(KAG02-DSENG)
Database / Data Warehouse Management	Hive, Impala, Spark SQL	SQL	Stabilizers KA02.02(KAG02-DSENG)
Data Ingestion	Sqoop	Command-based Scripting	Challengers KA02.05(KAG02-DSENG)
Data Processing	Spark (Scala, PySpark, SQL) PySpark, Numpy, SciPy, Pandas	Functional Programming	Stabilizers KA02.05, KA02.06 (KAG02-DSENG)
Data Visualization	Matplotlib, Seaborn	Functional Programming	Disruptors KA02.06 (KAG02-DSENG)
Machine Learning	Scikit-learn	Functional Programming	Disruptors KA02.06 (KAG02-DSENG)

- KA02.05 (DSENG.05/BDSE) - Organization and engineering of big data systems, including approaches to big data analysis and common MapReduce algorithms
- KA02.06 (DSENG.06/DSAPPD) - Design of data science applications (big data), including languages for big data (Python, R), tools and models for data presentation and visualization

In addition, this course covers three quadrants associated with the job-skills matrix. Most of the topics have included the skills of the *disruptors* quadrant. The only quadrant that has not been covered in this course is *escalators*. However, associated skills, such as Tableau, are taught in other courses within the offered program.

3.3 Course assessments

In this course, students are evaluated through three types of assessment, namely tests, assignments, and projects. In this paper, we limit our discussion to assignments and projects. The assignments require students to apply specific technical skills to address guided problem solving. Meanwhile, for the project, students must apply a combination of technical skills to address data science problems. The assignments are done individually, whilst the project is done in a group.

The evaluation of the assessments is based on two mechanisms. First, a checklist to ensure that the submission elements are generally met in each assignment and project. Second, a rubric to deal with different quality of deliverable. The rubric follows four levels of scales, namely, Excellent (4), Good (3), Fair (2), and Poor (1). Furthermore, it comprises four elements:

- Submission requirement - The checklist created earlier is used to support the assessment for this element, which mainly focuses on what and when.
- Reporting - This element is used to assess the amount of information provided by the students to support their deliverable.
- Coding - This element is applied to assess the quality of the coding in achieving the assessment objective.
- Presentation - This element is needed to assess the articulation of the students on their deliverable. It is applicable only to the project.

We further elaborate on the evaluation details by focusing on two academic semesters, which we call semesters *Oct2020-Feb2021* and *March2021-July2021* and summarized in Table 3.

In semester *Oct2020-Feb2021*, students were assigned with two assignments and one project. The same datasets were used for the assignments and the project. This decision was made so that students could focus more on applying the tool than searching and understanding the datasets. The datasets were about a memorandum from a university (7 variables) and its relationship with the QS ranking (10 variables). Among the features of the memorandum dataset are the signing date, the collaborator,

Table 3 Summary of Assessment Outlines [α:Oct2020-Feb2021, β:March2021-July2021]

Context	Assignment 1	Assignment 2	Project
Assessment Objectives	Able to apply the techniques related to the respective services and tools for performing data analytics$^{\alpha,\beta}$	Able to apply the techniques related to the respective services and tools for performing data analytics$^{\alpha,\beta}$	Able to construct the steps programmatically with the appropriate techniques in addressing the descriptive and predictive analytics$^{\alpha,\beta}$
	Able to recognize performance differences in relation to the targeted techniques$^{\alpha,\beta}$	Able to recognize performance differences in relation to the targeted techniques$^{\alpha,\beta}$	Able to interpret the results insightfully$^{\alpha,\beta}$
Main Tasks	Data engineering, descriptive analytics$^{\alpha}$	Data engineering, descriptive analytics, performance comparison$^{\alpha,\beta}$	Data engineering, descriptive and predictive analytics$^{\alpha,\beta}$
	Data engineering, descriptive analytics, performance comparison$^{\beta}$		
Datasets (DS)	QSRanking DS, Memorandum DS$^{\alpha}$	QSRanking DS, Memorandum DS$^{\alpha}$	QSRanking DS, Memorandum DS$^{\alpha}$
	Academic Publication DS$^{\beta}$	Academic Publication DS$^{\beta}$	Staff Productivity DS$^{\beta}$
Services / Tools	HDFS, Hive DB, Hive-QL$^{\alpha,\beta}$	HDFS, Hive DB, Spark (Scala, PySpark, SQL)$^{\alpha}$	Jupyter Hub / Notebook, Python Frameworks (Numpy, Pandas, Seaborn, Matplotlib, Tensorflow, etc)$^{\alpha,\beta}$
		HDFS, Hive DB, Impala, Spark (Scala, PySpark, SQL) $^{\beta}$	

Table 3 (continued)

Context	Assignment 1	Assignment 2	Project
Techniques Involved	Hive-QL: tables creation (external, internal),counting, grouping, filtering, joining, exporting, hql execution $^{\alpha}$	Spark: reading from file and HiveDB, counting, grouping, filtering, joining, exporting$^{\alpha}$	Data extraction, handling missing values, handling duplication, correlation analysis, merging, encoding, grouping, data visualization including map visual$^{\alpha}$
	Hive-QL: tables creation (external, internal, partition, bucket), counting, grouping, filtering, joining, exporting, hql execution$^{\beta}$	Impala: counting, grouping, filtering, joining, exporting — Spark: reading from file and Hive DB, counting, grouping, filtering, joining, exporting$^{\beta}$	Data extraction, handling missing values, handling duplication, outliers treatment, encoding, Correlation analysis, data visualization, predictive modeling, model evaluation$^{\beta}$

and the initiator. The features of the QS ranking dataset that include the ranking number, the university, and the country. The details of the assessments are summarized in Table 3.

Both assignments were intended to assess their competence in relation to the Hadoop environment. The first assignment was aimed at assessing data processing and analysis skills based on Hive and using HDFS as the main storage. The datasets that were given to them require some basic cleaning tasks. They were allowed to use Excel for this reason since, in the real world, dirty datasets are inevitable, and Excel is the common tool to choose. The rest of the tasks should be done in Hive, including creating the schema and performing queries. As two datasets were given, joining is one of the important skills that students need to be familiar with. The second assignment was crafted to assess their skills in applying spark in connection with HDFS and Hive. Because they used the same datasets, they can focus more on the techniques. They were expected to perform processing and analysis tasks similar to those in the Hive into Spark environment, so they can implicitly make the technical comparison between the tools. Furthermore, they were asked to provide a comparison of execution time between the tools.

For the project, which is a group project, the assessment was designed to assess their ability to perform descriptive analytical tasks in a Python environment, as summarized in Table 3. The tasks involved exploratory data analysis tasks, data preparation, and visualization of the analysis results. As the dataset comprised university names from different countries (after merging the two datasets), they were requested to be visualized in a map visualization. Furthermore, they had to recommend potential predictive analytics that can be performed, assuming that the data can be expanded or enhanced with more features. The recommendation was to encourage them to think beyond the context of the given datasets.

In semester *March2021-July2021*, the same number of assessments was given, but with different datasets. In addition, different datasets were used for assignments and the project. The dataset used for the assignments is a publication dataset that has been gathered from Scopus over a period of time. The number of records is closed to 250,000 records. For this semester, a large dataset size was chosen, since students were expected to measure performance in executing related data analysis tasks on the server. This activity also allowed them to appreciate the different capabilities of the tools in relation to their performance in executing some commands. The summary of the assignments is presented in Table 3.

The first assignment was designed to assess the students' ability to perform data analysis in Hive and Impala, using HDFS as storage. Through this activity, we found that the bucketing concept could not be implemented in Impala as part of the cloudera version we have on the server. In addition to the dataset, the main differences compared to the previous semester were that they had to implement partitioning and bucketing concepts, and tasks had to be performed on the server. They also had to measure the performance of queries. The second assignment aimed to assess their skills related to Spark. Similar analysis tasks were required from them, but using Spark Scala, PySpark, and Spark SQL. Then, they were asked to measure and compare the performance between the tools.

The project for this semester focused on a productivity dataset that was created to represent the productivity of staff in a company. The amount of data in this dataset was less than the publication dataset for the assignments. However, the number of features was more, specifically, 36 features which can be divided into three aspects, their profiles, workloads, and tasks achieved. A column dedicated to labeling individual records had been prepared. Basically, the column determines whether a staff member is a high- or low-performance performer. This dataset was designed in such a way as to help students apply supervised learning algorithms. Furthermore, a large number of features was created to allow them to implement the correlation analysis study, so that they could choose the appropriate features in constructing the predictive models. The summary of the project is presented in Table 3.

3.4 Computing platform

In this section, we discuss the initiative taken to establish a computing platform to support the implementation of the course and instill the required skills as presented in Table 2.

Learning data science technology requires an appropriate specification of software and hardware. For this course, students can access a computer lab called the Big Data lab. The lab is provided with 30 PCs, each PC is equipped with an Intel(R) Core(TM) i7-6700 processor and 32GB RAM. On each PC, there is a Cloudera Quickstart VM that enables the students to have a computing environment with minimal Linux background and can start focusing on data analysis activities rather than spending time configuring and troubleshooting. The Cloudera Quickstart VM provides a single node environment that can run on a PC. This virtual machine is approximately 5GB in size and requires at least 4GB of memory with 1 processor, which can be configured in a virtualization tool. For this course, the students mostly use VMWare. However, when using Cloudera VM, the setting of 4GB memory with 1 processor is insufficient. Therefore, students are encouraged to set at least 8GB of memory with 2 processors. With this Cloudera VM, students can also experiment with the Linux environment whenever necessary.

In early 2021, the school received some funds to upgrade the lab with server capabilities. Therefore, a computing cluster environment was set up as an alternative platform for students, in addition to the Cloudera quickstart VM. The cluster is illustrated as in Fig. 2.

The cluster is made up of 6 nodes. Four of the nodes are virtual machines from 2 physical servers. One node is a virtual machine created from the university data center. One more node is made up of a physical server. Five nodes are configured for the Hadoop environment, and one node is dedicated to the edge node with RapidMiner Server. Of the five nodes, one is configured as the master node and the others take the role of the data node. The total resources allocated for running the cluster is 142GB memory and 27 Virtual Cores (vcores). The balance of resources is meant for the local operating system.

The master node is installed with Cloudera Manager and MariaDB. The data nodes are installed with the cloudera software agent, which will interact with the master node. The edge node is used to allow remote access to the Hadoop environment.

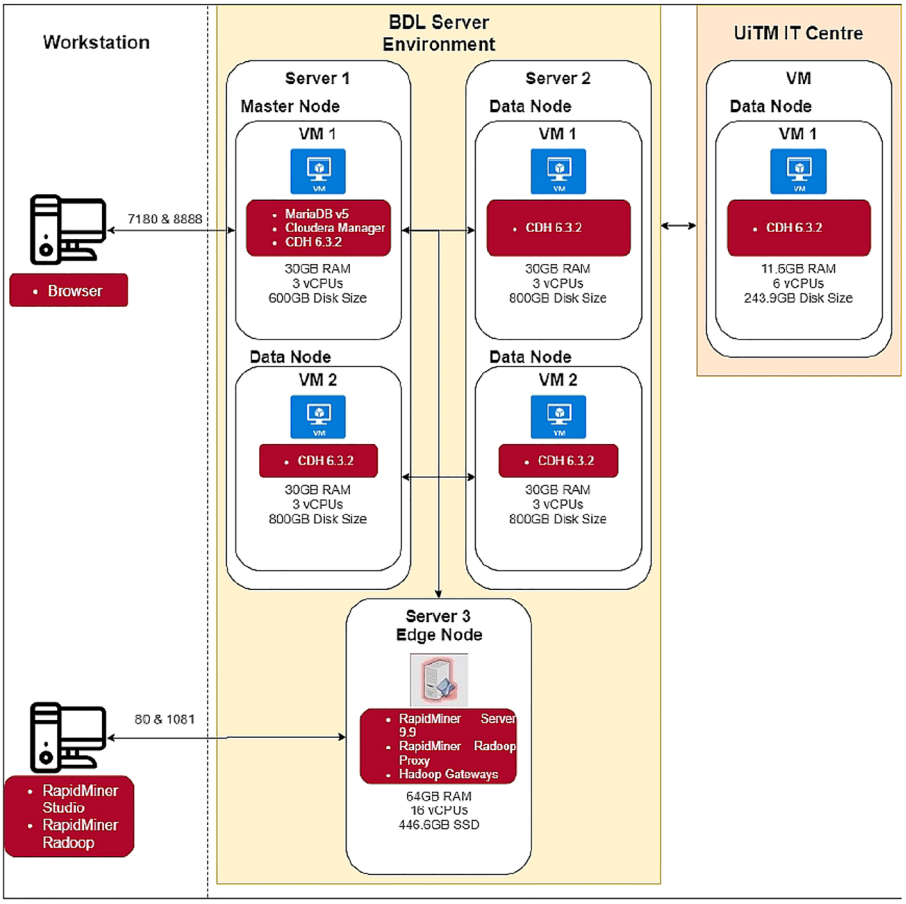

Fig. 2 Architecture of the Cluster Environment

Although the RapidMiner system is beyond the scope of this course, one of the services provided, called RapidMiner Notebooks, is used as an alternative programming editor for Python.

Students can access the Hadoop cluster in several ways. First, students can use Secure Shell (SSH) related tools, for example, Putty, to access the operating system level through the proxy machine. This access gives them the experience of dealing with Linux via the terminal environment. Once logged in, they can also access Hadoop services such as HDFS, Hive, Impala, and Spark via the terminal. Second, if they need to transfer files from the local PC to the cluster, they can use any suitable file transfer tool, e.g. WinSCP. Third, they can access the cluster through the Web application, that is, the HUE and the Cloudera Manager Web application. HUE is a web-based interactive query editor for Hive and Impala. It can also be used to manage users' accounts and files in HDFS. Cloudera Manager web application allows

users to configure and monitor the services provided as part of the Hadoop environment. However, for the cluster setting, students are only allowed to view. They can explore the configuration part with the Cloudera quickstart VM running as a single node, as introduced earlier.

3.5 Strategizing accessibility

In this section, we explain our initiative to increase accessibility to positively improve the learning experience of our students. We focus on two academic semesters, *Oct2020-Feb2021* and *March2021-July2021*, since both semesters were conducted through online learning. Figure 3 illustrates the accessibility aspects of the computing platform, specifically access to the Internet and the intranet.

3.5.1 Enabling intranet access

In semester *Oct2020-Feb2021*, when the Covid-19 pandemic occurred, students were not allowed on campus. A decision had to be made to ensure that the lab computing platform can still be used effectively for a short period of time. Therefore, we addressed the accessibility aspect by enabling remote access to these PCs (local machines). However, during this period, the cluster was not set up yet. This accessibility support is crucial, since laptops of some students are not sufficient (in terms of memory and processor) to run the Hadoop platform that we have selected, namely the Cloudera Hadoop Distribution. Details of the lab and related software were discussed in the previous section. Remote access provides them with an alternative solution and

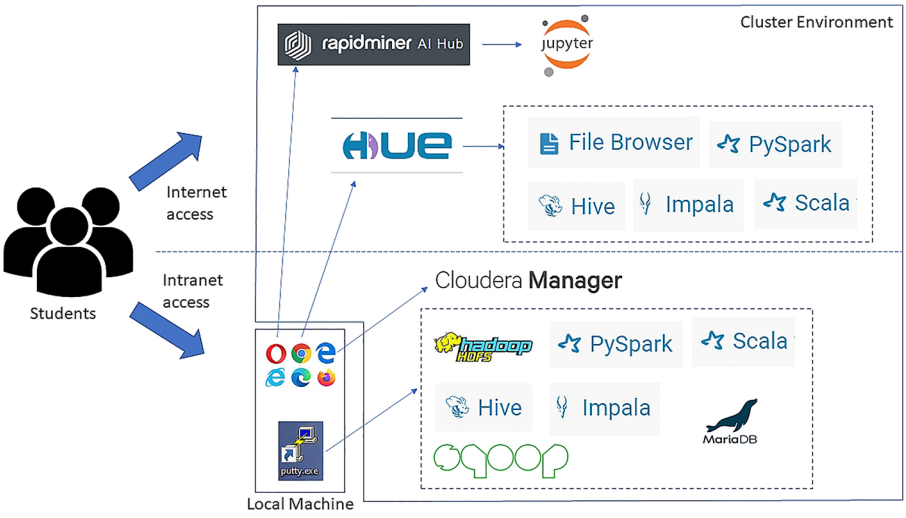

Fig. 3 Accessibility of Computing Environment for Online Learning

contributes to their positive learning experience. In addition, we also had to consider the power usage of these PCs. In general, a lab technician can help with this matter (e.g. switching on/off the PCs at certain times), but we cannot expect this person to be in the lab on a daily basis. For this reason, together with the technician in charge, we searched for a solution and found a way to perform BIOS settings on these PCs, so that each PC can automatically shutdown and boot up based on certain schedules. In the event of a power failure or unexpected shutdown of any PC, we resolve it through manual intervention.

3.5.2 Enabling internet access

In semester *March2021-July2021*, Covid-19 still affected the teaching and learning process. In terms of accessibility to the computing platform, we continue to provide remote access to students. However, we have taken another step, which is to provide a cluster environment that is based on Cloudera Cluster Management. The details of the setup have been discussed earlier. In the first part of the semester *March2021-July2021*, the accessibility of the server was set for intranet access, as shown in Figure 3 so that to turn it into Internet access, we had to go through several layers of configurations and security checks. From outside the university's network, students can access the local machines provided in the lab to use services in the cluster. Once they can enter the intranet environment, they can access all services in the cluster environment, which include the web-based systems (i.e., Cloudera Manager, RapidMiner AI Hub, HUE) and the command-line applications (i.e. HDFS, Hive, Impala, Sqoop, Spark, MariaDB). After that, we managed to open the Internet access as illustrated in Fig. 3. However, services are limited to accessing the HUE application and the RapidMiner AI Hub. With HUE, students can access the HDFS File Browser, Hive database and query editor, Impala editor, and Spark editor through the Web interface. Meanwhile, with RapidMiner AI Hub, they can get access to the RapidMiner Notebook for practicing Python programming.

4 Experience and implementation

In this section, we focus on three aspects of experience and implementation of the course. First, we present samples of the deliverables that students have produced to demonstrate the skills applied to complete the assessments. These samples serve as supporting evidence to instill the required skills within the course. Second, we highlight the user management tasks that need to be given attention to initiate accessibility. Third, we provide the utilization and health condition of the cluster, as well as the associated services. All of these aspects are part of the continuous improvement process.

4.1 Deliverable for assessments

Figures 4 and 5 represent the samples delivered. These samples illustrate the application of skills related to data engineering. In particular, Fig. 4 is a sample of Spark

```
scala> val updateSchema = StructType(
     | List(
     | StructField("dcidentifier", StringType, true),
     | StructField("publicationname", StringType, true),
     | StructField("year", StringType, true),
     | StructField("citedby_count", StringType, true),
     | StructField("articletype", StringType, true),
     | StructField("subtypedesc", StringType, true),
     | StructField("openaccessflag", StringType, true),
     | StructField("dctitle", StringType, true)
     | ));
updateSchema: org.apache.spark.sql.types.StructType = StructType(Struc
tField(dcidentifier,StringType,true), StructField(publicationname,Strik
ngType,true), StructField(year,StringType,true), StructField(citedby_c
ount,StringType,true), StructField(articletype,StringType,true), Struc
tField(subtypedesc,StringType,true), StructField(openaccessflag,String
Type,true), StructField(dctitle,StringType,true))
```

Fig. 4 Sample of Scala Spark Program for Assignment 2

Scala that contains the codes to update the table schema. This is one of the common activities for data engineering. Meanwhile, Fig. 5 is a sample of performance measurements for Assignments 1 and 2. The student had to run the respective commands and record the execution times (in seconds) of several trials to conclude the average execution time. This activity helps them recognize the performance differences as a result of architectural design (i.e., Spark with Resilient Distributed Dataset and Hive with MapReduce) and the table structure (i.e., internal and external table, nonpartitioned and partitioned table).

Meanwhile, Figs. 6, 7, and 8 are samples of project execution. These samples illustrate the application of exploratory data analysis (EDA) tasks (e.g. data distribution, outlier analysis) for the data-driven method. In particular, Fig. 6 represents the initiative of a group project to present its methodology to solve the given problem. The aim is to produce and highlight the best predictive model from the provided dataset. For this group, they had decided to tackle the imbalance aspect with the synthetic minority oversampling technique (SMOTE) and selected three machine learning algorithms, namely logistic regression, decision tree, and support vector

		Query 1						Query 2					
		1	2	3	4	5	Average	1	2	3	4	5	Average
Hive (AS1) - on server	Table Schema 1	20.856	24.579	27.558	19.536	22.756	23.057	22.619	21.198	22.035	27.234	26.505	23.918
	Table Schema 2	28.391	25.994	22.680	25.513	21.557	24.827	23.862	23.620	25.539	29.421	23.008	25.090
	Table Schema 3	22.485	22.239	20.024	20.959	24.247	21.991	22.535	22.224	22.674	21.043	26.288	22.953
	Table Schema 4	22.522	26.418	27.674	25.614	23.433	25.132	28.390	29.216	24.289	27.453	26.766	27.223
Impala(AS2) - on server	Table Schema 1	0.22	0.21	0.22	0.21	0.22	0.216	0.22	0.21	0.21	0.21	0.22	0.214
	Table Schema 2	0.32	0.23	0.22	0.26	0.23	0.252	0.31	0.31	0.32	0.31	0.31	0.312
	Table Schema 3	0.21	0.22	0.22	0.22	0.22	0.218	0.21	0.22	0.21	0.22	0.21	0.214
Spark (AS2) on server	Spark Scala	10.9	1.6	1.035	1.042	1.039	3.123	1.0	1.0	1.0	1.0	1.0	1.000
	PySpark	14.0	5.0	1.0	1.0	1.0	4.400	1.586	0.828	0.939	0.825	0.805	0.997
	Spark SQL	5	5	3	1	0.9	2.980	1.489	0.953	0.945	0.81	0.692	0.978
	Average	12.490	11.149	10.363	9.535	9.560		10.222	9.978	9.816	10.853	10.580	

Fig. 5 Sample of Performance Measurements for Assignment 1 and 2

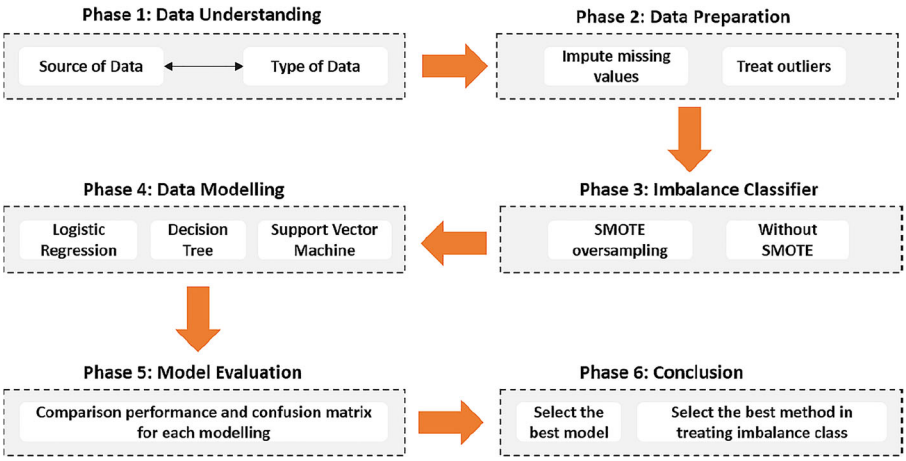

Fig. 6 Sample of Implemented Methodology for Project

machine. Then, they were able to show the performance of the predictive models with and without SMOTE implementation. Meanwhile, Figs. 7 and 8 are the visual elements produced to achieve the defined methodology. Figure 7 demonstrates their effort to understand the data distribution, while Fig. 8 illustrates the initiative to identify outliers in the dataset.

4.2 User management for cluster accessibility

There are several tasks that must be prepared to allow users to access the cluster for each academic semester. Through our experience, we identify the following administrative tasks:

- *Creating Users in Linux* - It allows students to access the cluster through the SSH mode. For this reason, we use MobaXterm to enable multiple SSH connections to virtual machines. By accessing a root account, we can create the username, password, and group for each student. Figure 9 illustrates the user creation process on multiple virtual machines using an SSH connection using MobaXterm.
- *Creating Users in HUE* - It allows students to access services such as HDFS, Hive, Impala, and Spark. To do this, we use the *Manage Users* function provided in HUE. We also initiate the HDFS home directory when creating the account. Furthermore, we set the permissions for the services that can be accessed through HUE.
- *Creating Hive Database for Each Account* - It provides a dedicated database environment for each student. Thus, they can work on exercises and assignments within their own database. It also avoids having multiple databases per account, which can result in too many databases being created and cause more maintenance activity. For this reason, we use the function *Create Database* in HUE.

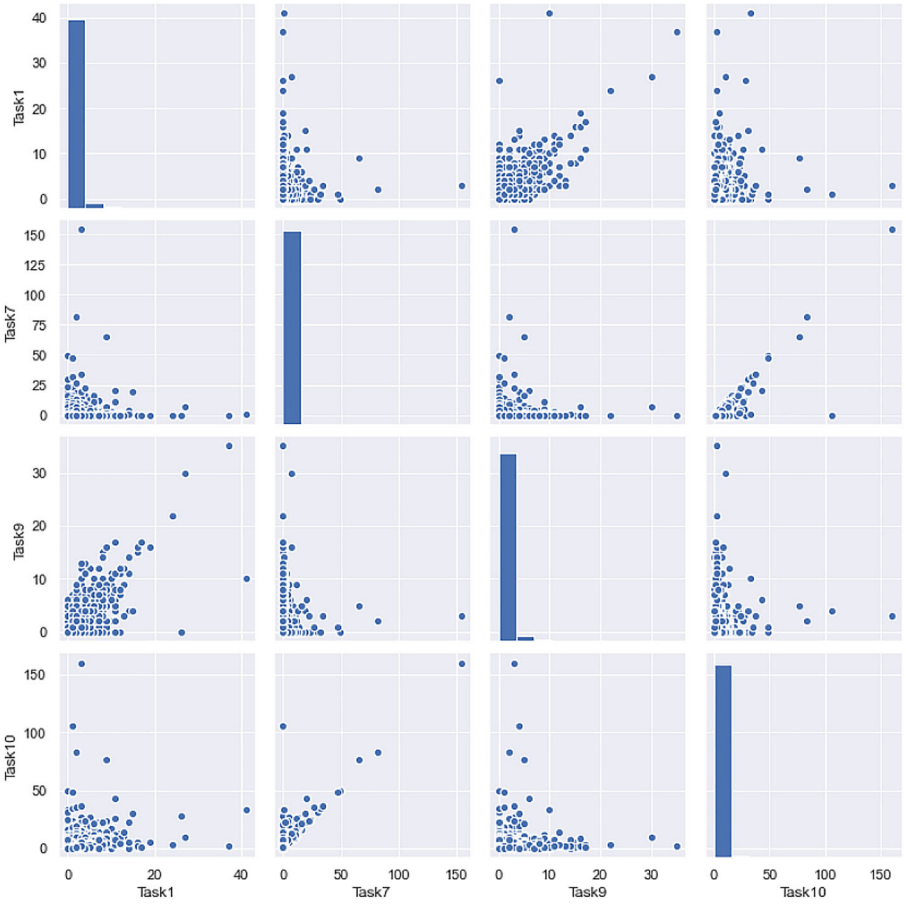

Fig. 7 Sample of Histogram as part of EDA for Project

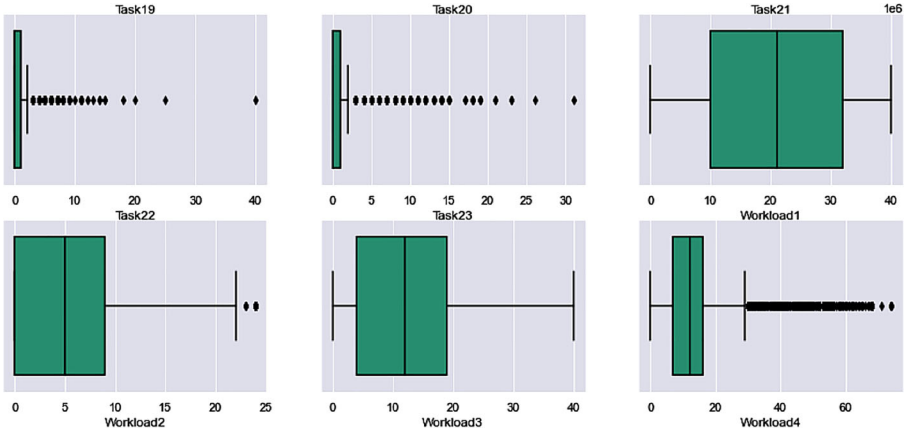

Fig. 8 Sample of Boxplot as part of EDA for Project

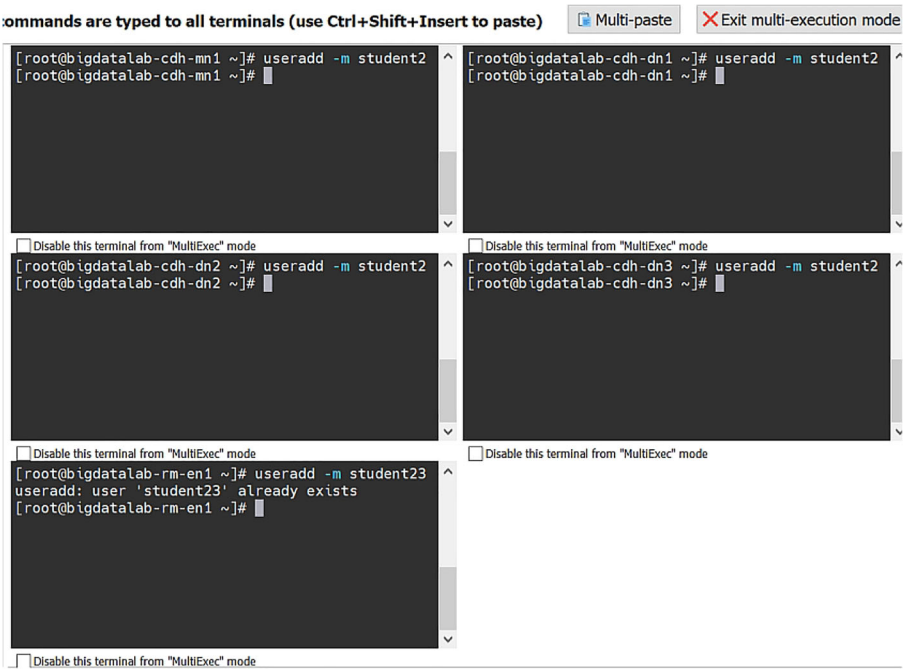

Fig. 9 Creating User Accounts using MobaXterm

- *Configuring Privileges for HDFS and Hive* - It allows some controls to access the HDFS directory and Hive database. This means that for each student, their access is restricted to their own directory and database. For this configuration, we use the *Security Browsers* provided in HUE. We set two rules as shown in Fig. 10 that are associated with each account (e.g., *student18*). The first rule is to restrict the access of each account within its own HDFS directory. The second rule is to limit the access of each account to its own database.
- *Creating Users in Rapidminer AI Hub* - It allows students to access the RapidMiner Notebook to practice Python programming. For this task, we use the *Manage Users* function provided in the RapidMiner AI Hub.

Fig. 10 Configuring Privileges in HUE

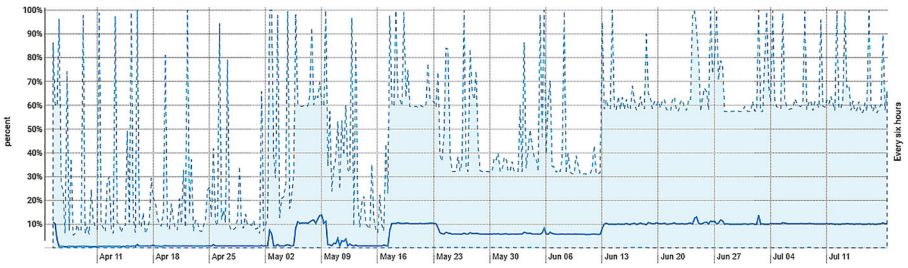

Fig. 11 CPU Utilization of the Cluster Environment

4.3 Utilization of cluster

In this section, we elaborate on the utilization of the cluster and the health conditions, as well as the utilization of the Hadoop core services, namely HDFS and YARN (including MapReduce). The data collected are based on semester *March2021-July2021* since the cluster was fully functioning during this semester. Specifically, we have taken a period of logged data from 4^{th} April 2021 to 18^{th} July 2021. The utilization of the cluster and related services is associated with 15 enrolled students for the semester *March2021-July2021*.

4.3.1 Cluster utilization

Figure 11 shows the CPU utilization of the cluster based on the stated period. The topic related to Hadoop services that require the use of cluster started on 11^{th} April 2021. As we can see, utilization is increasing toward July. Specifically, the increase in May was due to the exploration of HDFS and Hive, and students needed to work on assignment 1. Meanwhile, the increase in June was due to the exploration of Spark, and the students had to work on assignment 2.

4.3.2 HDFS health and utilization

The health condition of HDFS is also shown in Fig. 12 for the same period. HDFS is crucial for storing and managing files containing data sets. Its health can be

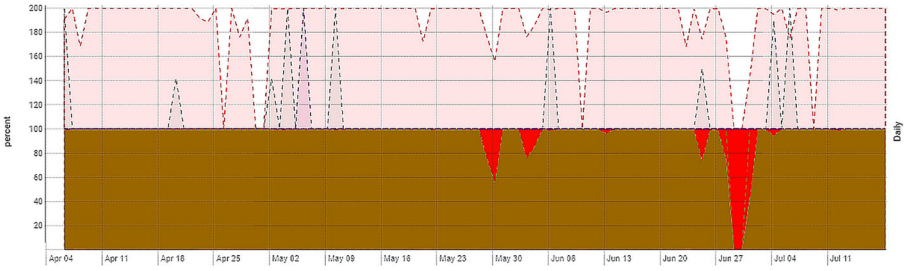

Fig. 12 Health Condition of HDFS in the Cluster Environment

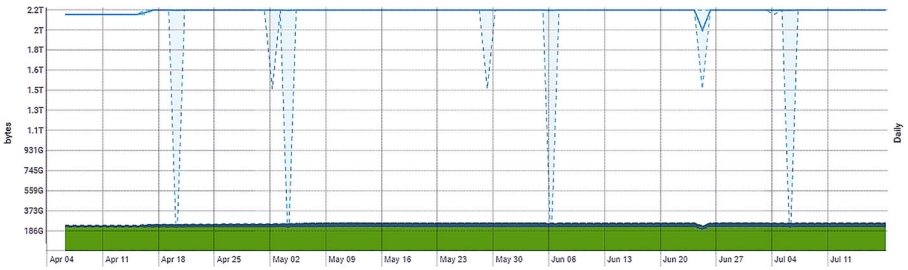

Fig. 13 Resource Utilization of HDFS in the Cluster Environment

influenced by various reasons. Most of the time, we were dealing with the health of HDFS (brown area). The concern for health is determined when the percentage of healthy nodes falls below the warning threshold. There were several periods in which HDFS was in bad health condition (red area). It is determined when the total percentage of healthy and affected nodes falls below the critical threshold. Despite this condition, the students were able to carry out their activities, especially for Assignments 1 and 2.

The utilization of HDFS resources is shown in Fig. 13 during the same period of time. The HDFS has been configured to take a maximum of 2.2 Terabytes (TiB). The mean utilization on the first day of the selected period was 215.8 Gibibyte (GiB), while the mean utilization on the last day was 217.3 GiB. HDFS has shown that it was underutilized during the period. Since we focus on the stability of the HDFS, accessibility is expected to be maintained to support a positive student learning experience. From this exercise, the incoming academic semester can then focus on increasing HDFS resource utilization by providing challenging datasets.

4.3.3 YARN health and allocation

The health of YARN is crucial to allowing applications and services to run successfully in the Hadoop cluster. As a resource manager, YARN is responsible for enabling resources to run the required applications. Its health status is also shown in Fig. 14

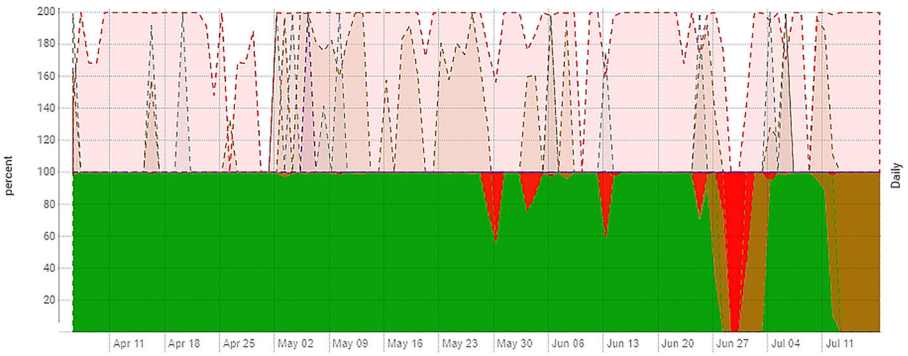

Fig. 14 Health Condition of YARN in the Cluster Environment

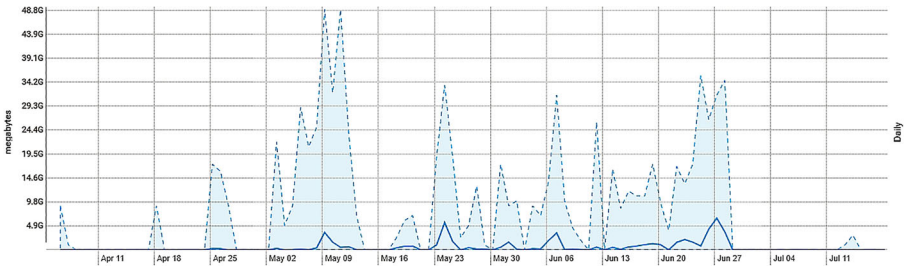

Fig. 15 Memory Allocation Across YARN Pools in the Cluster Environment

for the same period. During this period, its condition was mostly healthy (green area). It had a couple of periods of poor health (read area), especially in 27^{th} in June. This date was the submission deadline for Assignment 2. Despite this bad condition, the students managed to submit the assignment with a slight delay. Also, towards the end, we were dealing with health (brown area). However, during this period, the students no longer had access to the cluster for assessment purposes, since they focused on Python programming in the Anaconda environment.

Figure 15 illustrates the allocation of memory across the YARN pools in the cluster over the same period of time. In general, the memory allocation shows that the applications in the cluster were being utilized. From the graph, we can see that the memory allocated to YARN started to increase on 25^{th} of April since the topic of Hive was covered from this date onward. Then more memory was allocated in May because the students were working on assignment 1 due in 9^{th} in May, and then memory was allocated from time to time to the YARN pools due to the Spark topic and assignment 2 due in 27^{th} in June. After this date, the students then focused on Python programming using the Anaconda environment; thus, YARN did not need memory allocation.

5 Reflection and conclusion

In this paper, we have presented the Data Science Technology course based on a set of requirements as presented in Section 1. Compared to existing related courses, our course offers a deeper scope on Spark programming, an alternative tool for data ingestion called Sqoop, and the use of web-based tools as part of the Cloudera Hadoop distribution run on the cluster. In mapping to DS-BoK, the course contents offered have covered infrastructure and platforms(DSENG.02/DSIAPP), Big Data systems organization and engineering (DSENG.05/BDSE), and data science application design (DSENG.06/DSAPPD). Meanwhile, in relation to the job market, this course can potentially instill skills related to the stabilizer, challenger, and disruptor quadrants. These skills have been embedded in the coursework that requires students to apply them to complete assignments and project. Furthermore, a computing environment based on the Cloudera Hadoop cluster has been set up for teaching and learning. This cluster can be accessed through the Internet and the intranet. Based on

the usage of this cluster within the targeted observation period, we can conclude that it is stable in some way and has the potential to support the larger problem. However, further observation is needed to support this claim. Having said that, the use of Spark needs to be strategically handled, especially when dealing with multiple accesses, since it is memory-intensive.

Providing a cluster requires server administration skills. Furthermore, enabling access to the Internet requires network and security skills. In fact, when the cluster is up and running, it needs to be maintained, especially during the semester period. User accounts must be created, and some privilege settings must be enforced. For this reason, we are glad to have financial, technical and maintenance support to make the cluster a reality and make it usable to students. Although there were times when the cluster was not stable, in general, students can have access to it and complete their assignments and projects accordingly. Having said that, more planning is needed to ensure continuous support and to extend the capability of the cluster in terms of computing power and space.

Providing Big Data for the course is another challenge and is important for the student learning experience. The privacy of data is crucial when it comes to utilizing any existing project within the school. If there is a certain restriction to the data, that means more configuration is needed on the cluster to make it available to the students. A public Big Data set can be the alternative for this reason, but it may limit the assignments or project to the same data for all students. Therefore, a decision must be made on the suitability of the datasets for all students. This will become our concern in the coming semester. For the past two semesters, the datasets assigned to the students were related to academic ranking and publication, and the Hadoop cluster has been used to complete the tasks.

Streaming data is an interesting technological aspect that students should explore. Data may come from various sources, such as Internet of Things (IoT) devices, social media applications, and real-time applications. Dealing with this type of data can obviously broaden their skills. The preparation of the cluster to enable streaming data processing poses another technical challenge in preparing it for the learning process. In addition, course materials must be designed appropriately to provide sufficient exposure to students. Therefore, a proper plan must be made to realize this intention.

With the challenges mentioned above and the future initiatives to address them, we will continuously improve this course to allow students to acquire the knowledge and skills to enter the job markets of the DSA landscape.

Abbreviations AI, - Artificial Intelligence; BDA, - Big Data Analytics; BoK, - Body of Knowledge; DSA, - Data Science and Analytics; EDA, - Exploratory Data Analysis; HDFS, - Hadoop File System; IoT, - Internet of Things; MCO, - Movement Control Order; MDEC, - Malaysia Digital Economy Corporation; PC, - Personal Computer; UiTM, - Universiti Teknologi MARA; VM, - Virtual Machine; YARN, - Yet Another Resource Negotiator.

Acknowledgements We would like to take this opportunity to thank the School of Computing Sciences (formerly known as the Faculty of Computer and Mathematical Sciences), College of Computing, Informatics and Media, and Universiti Teknologi MARA (UiTM) for providing support to deploy the cluster for this course.

Availability of data and materials The datasets generated during and/or analyzed during the current study are not publicly available due to security reasons, but are available from the corresponding author on reasonable request.

Declarations

Competing interests The authors declare that they have no competing interests.

References

Adams, J. C. (2020). Creating a balanced data science program. In *Annual Conference on Innovation and Technology in Computer Science Education, ITiCSE, Association for Computing Machinery, pp. 185-191*.

Bart, A. C., Kafura, D., Shaffer, C. A., & Tilevich, E. (2018). Reconciling the promise and pragmatics of enhancing computing pedagogy with data science. In *SIGCSE 2018 - Proceedings of the 49th ACM Technical Symposium on Computer Science Education, Association for Computing Machinery, Inc, vol 2018-January, pp 1029–1034*.

Baumer, B. (2015). A data science course for undergraduates: Thinking with data. *American Statistician*, *69*(4), 334–342.

Brunner, R. J., & Kim, E. J. (2016). Teaching data science. In *Procedia computer science, elsevier b.v.*, (Vol. 80 pp. 1947-1956).

Çetinkaya-Rundel, M., & Ellison, V. (2020). A fresh look at introductory data science. *Journal of Statistics Education*, *2021*(S1), 16–26.

Çetinkaya-Rundel, M., & Rundel, C. (2018). Infrastructure and tools for teaching computing throughout the statistical curriculum. *American Statistician*, *72*(1), 58–65.

Cuadrado-Gallego, J. J., & Demchenko, Y. (2020). Data science body of knowledge. In J. J. Cuadrado-gallego, & Y. Demchenko (Eds.) *The Data Science Framework: A View from the EDISON Project* (pp. 43-73). Cham: Springer International Publishing.

Demchenko, Y., & Cuadrado-Gallego, J. J. (2020). Data science competences. In J. J. Cuadrado-gallego, & Y. Demchenko (Eds.) *The Data Science Framework: A View from the EDISON Project* (pp. 9-41). Cham: Springer International Publishing.

DePratti, R., Dancik, G. M., Lucci, F., & Sampson, R. D. (2017). Development of an introductory big data programming and concepts course. *Journal of Computing Sciences in Colleges*, *32*(6), 175–182.

Dichev, C., & Dicheva, D. (2017). Towards data science literacy. In *Procedia computer science, elsevier b.v.*, (Vol. 108 pp. 2151–2160).

Dichev, C., Dicheva, D., Cassel, L., Goelman, D., & Posner, M. (2016). Preparing all students for the data-driven world. In *Proceedings of the Symposium on Computing atMinority Institutions, ADMI*.

Donoghue, T., Voytek, B., & Ellis, S. E. (2021). Teaching creative and practical data science at scale. *Journal of Statistics and Data Science Education*, *29*(sup1), S27–S39.

Eckroth, J. (2016). Teaching big data with a virtual cluster. In *Proceedings of the 47th ACM Technical Symposium on Computing Science Education, Association for Computing Machinery, New York, NY, USA, SIGCSE '16, pp 175–180*.

Eckroth, J. (2017). Teaching future big data analysts: Curriculum and experience report. In *Proceedings - 2017 IEEE 31st International Parallel and Distributed Processing Symposium Workshops, IPDPSW 2017, Institute of Electrical and Electronics Engineers Inc., pp 346–351*.

Eckroth, J. (2018). A course on big data analytics. *J Parallel Distrib Comput*, *118*, 166–176.

Eilks, I. (2018). Action research in science education: a twenty-year personal perspective. *ARISE*, *1*(1), 3–14.

Fekete, A., Kay, J., & Röhm, U. (2021). A data-centric computing curriculum for a data science major. In *SIGCSE 2021 - Proceedings of the 52nd ACM Technical Symposium on Computer Science Education, Association for Computing Machinery, Inc, pp 865–871*.

Hicks, S. C., & Irizarry, R. A. (2018). A guide to teaching data science. *American Statistician*, *72*(4), 382–391.

Kross, S., & Guo, P. J. (2019). Practitioners teaching data science in industry and academia: Expectations, workflows, and challenges. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, ACM, New York, NY, USA, p 14*.

Miller, S. (2017). The quant crunch: how the demand for data science skills is disrupting the job market.

Ngo, L. B., Duffy, E. B., & Apon, A. W. (2014). Teaching HDFS/MapReduce systems concepts to undergraduates. In *Proceedings of the International Parallel and Distributed Processing Symposium, IPDPS, IEEE Computer Society, pp 1114–1121*.

Oudshoorn, M. J., Titus, K. J., & Suchan, W. K. (2020). Building a new data science program based on an existing computer science program. In *Proceedings - Frontiers in Education Conference, FIE, Institute of Electrical and Electronics Engineers Inc., vol 2020-October*.

Salloum, M., Jeske, D., Ma, W., Papalexakis, V., Shelton, C., Tsotras, V., Zhou, S., & Shelton, C. T. (2021). Developing an interdisciplinary data science program; developing an interdisciplinary data science program. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education, ACM New York, NY, USA*.

Shankar, A. C. (2021). MDEC's commissioned study shows malaysia's big data analytics market expected to grow to us$1.9b by 2025. https://www.theedgemarkets.com/article/mdecs-commissioned-study-shows-malaysias-big-data-analytics-market-expected-grow-us19b-2025.

Wiktorski, T., Demchenko, Y., & Cuadrado-Gallego, J. J. (2020). Data science curriculum. In J. J. Cuadrado-gallego, & Y. Demchenko (Eds.) *The Data Science Framework: A View from the EDISON Project* (pp. 75–108). Cham: Springer International Publishing.

Affiliations

Azlan Ismail[1,2] · Sofianita Mutalib[2,3] · Haryani Haron[2]

Sofianita Mutalib
sofianita@uitm.edu.my

Haryani Haron
harya265@uitm.edu.my

1 Institute for Big Data Analytics and Artificial Intelligence (IBDAAI), Universiti Teknologi MARA (UiTM), 40450, Shah Alam, Selangor, Malaysia

2 School of Computing Sciences, College of Computing, Informatics and Media, Universiti Teknologi MARA (UiTM), 40450, Shah Alam, Selangor, Malaysia

3 Research Initiative Group Intelligent Systems, Universiti Teknologi MARA (UiTM), 40450, Shah Alam, Selangor, Malaysia