Optimal organisation of a big data training course: big data processing with BigQuery and setting up a Dataproc Hadoop framework

Meruert Serik†, Gulmira Nurbekova† & Meiramgul Mukhambetova‡

L.N. Gumilyev Eurasian National University, Nur-Sultan, Kazakhstan†
Khalel Dosmukhamedov Atyrau University, Atyrau, Kazakhstan‡

ABSTRACT: Nowadays, a high growth rate of data flows stored in various computing devices creates the need for big data analysis and increases the demand for specialists with competencies in this area in the socio-economic sphere. The training of big data specialists requires a comprehensive approach and the introduction of innovative technologies, hardware and software, programming environments into the content of education. In this article, the authors present a training course on big data processing and analysis prepared by them. During the preparation, they examined and selected optimal solutions for its implementation using modern hardware and software complexes, including cloud technologies. The authors describe data processing from various sources, data collection using ERA 500 controllers and sensors, processing and analysing big data through the BigQuery service from Google Cloud Platform (GCP), the Google Data Studio visualisation tool, configuring and using the Hadoop framework for high-performance computing distributed through the Dataproc service. The training course was conducted for students in Kazakhstan universities.

INTRODUCTION

The processing of large amounts of data for the analysis of specific activities has become increasingly widespread during the 2010-2020 decade, and the analysis of large amounts of data has started to be introduced in education to improve the learning process and its results. The emergence of big data has caused a revolution in all spheres of society [1]. There is a demand to work with exabyte-sized data now and in the future. The exabyte itself is one billion gigabytes or 10^{18} (quintillion) bytes, and the volume of the daily data stream distributed over the Internet is measured in exabytes, the volume of which is growing hourly. Zettabytes and yottabytes are measured in bytes 10^{21} and 10^{24}, even higher than exabytes. The comparative indicators of information measurement show how much the volume of data has increased as a result of global technological progress.

Analysing the world experience, Kazakhstan gives priority to the development of a system of intelligent analysis and forecasting based on big data to improve decision-making processes at the state level. The creation of a technological centre for big data analysis - a single *place of data collection* and assistance would ensure the reliable functioning, safety, integrity of state information resources, including those on the basis of existing initiatives [2]. However, there are contradictions between the requests of employers and educational programmes that are lagging behind. This problem causes difficulties in finding employment for young specialists not fully prepared for the big data challenge.

The purpose of the research: to prepare students for future professional activity by introducing a training course on big data processing and analysis.

During the analysis of relevant training courses in Kazakhstan and abroad, as well as various teaching aids, a small number of practice-oriented activities were observed.

The content of the training course developed by the authors of this article has been implemented in several stages through the following technological processes:

1. Data collection using the ERA 500 controller and sensors.
2. Processing and analysis of big data via the Google Cloud Platform (GCP) BigQuery service.
3. Implementation of data visualisation work through the Data Studio tool.
4. Configuring the Hadoop framework for high-performance computing distributed through the Dataproc service.

The training course with the new content was conducted for 4th-year students of the specialty 6B01511-Informatics at *L.N.Gumilyov* Eurasian National University, Nur-Sultan, Kazakhstan, and for undergraduates in the educational programmes 7M01514-SMART-city Technologies and 7M01525-STEM-Education in the same institution.

LITERATURE REVIEW

The high level of development of the big data industry imposes new requirements on educational institutions in terms of its integration into the curricula, and thus creating new opportunities for graduates. Big data can be viewed in the context of educational policy and learning analytics. Hence, for the systematic and effective conduct of activities in this direction, it is beneficial to involve practitioners, scientists and politicians as stakeholders and advisors in education [3].

In addition to the use of big data, digital transformation is leading to profound changes in modern higher education. It is also necessary to consider changing the basic institutional value of education to better meet the needs of students using big data and digital technologies [4].

In the process of learning big data, students perceive this concept differently, so it is necessary to analyse user profiles and their tendency to use big data, as well as examine what factors influence the use of big data [5].

The essence of professional education is the analysis of the existing model of basic education, the combination of certain forms of education, the analysis of the interpenetration of big data and education, the promotion of integration and development of big data [6].

Higher educational institutions offering computer science and engineering courses have started to introduce big data into the programmes, and conditions are being created for the training of highly competent specialists in the future [7]. It follows from this that the introduction of big data into the educational process is a modern and necessary step. In this work it is essential to analyse various methods and technologies and take into account the profile of students.

COURSE DESIGN

While working with big data, along with its volume, an important parameter is the variety and speed of processing. Large data structures can be message forms, images in a social network and their updates, data from sensors, GPS signals from mobile phones, and much more.

During the organisation of the training course, the authors of this article considered the ERA 500 network controller as an example for generalising data received from sensors as initial data. In the following steps, the GCM functions were used as a software solution for processing and analysing big data. The transition of modern production organisations to the use of cloud technologies, involves getting rid of technical work, purchasing licensed programs, using only the relevant resources and providing the necessary services [8][9].

Another feature of cloud technologies is the ability to work with the environment necessary for complex calculations, both extracurricular and in distance learning conditions. It is worth noting that this has especially helped students during on-line training in the context of the current pandemic. The model of practical implementation of the big data training course is presented in Figure 1.

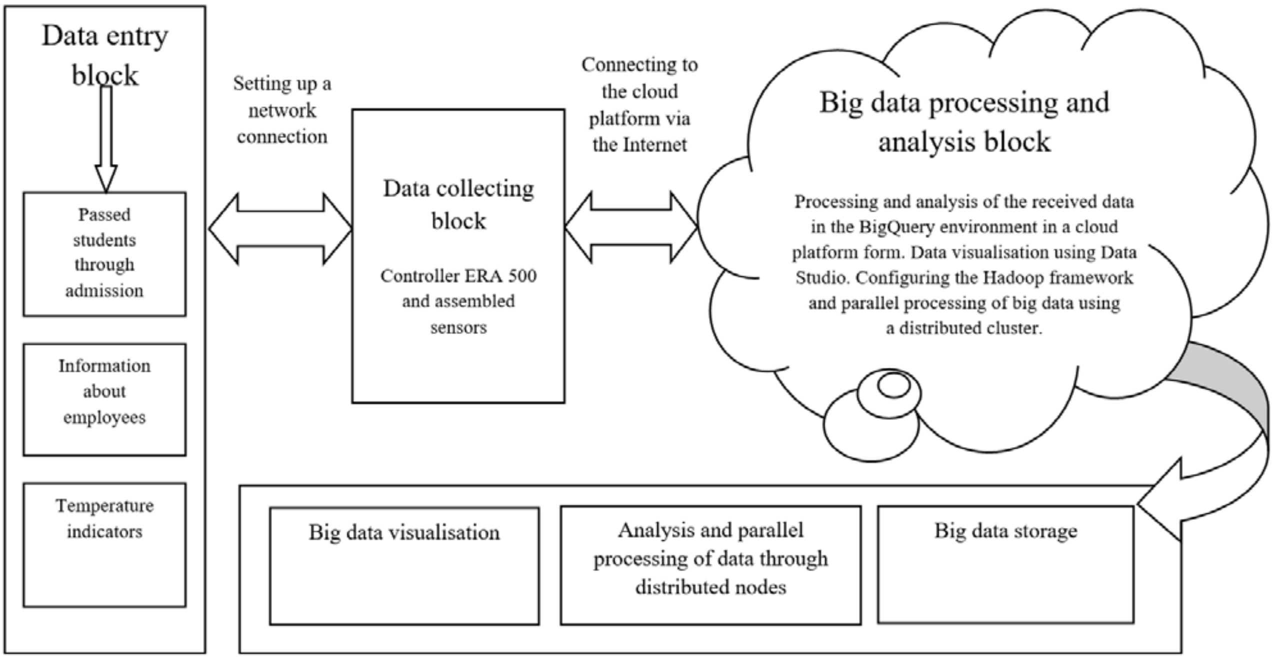

Data entry block
Passed students through admission
Information about employees
Temperature indicators
Setting up a network connection
Data collecting block
Controller ERA 500 and assembled sensors
Connecting to the cloud platform via the Internet
Big data processing and analysis block
Processing and analysis of the received data in the BigQuery environment in a cloud platform form. Data visualisation using Data Studio. Configuring the Hadoop framework and parallel processing of big data using a distributed cluster.
Big data visualisation
Analysis and parallel processing of data through distributed nodes
Big data storage

Figure 1: Structural model of the big data training course.

In the next section, the authors focus on the results of the training course in accordance with the developed model.

RESEARCH RESULTS

The training course *Big Data and Cloud Computing*, implemented in the educational process for students and undergraduates of *L.N. Gumilyov* Eurasian National University in order to master such methods as storage, processing and transmission of large amounts of data, had five credits (one lecture, two practical classes, two independent works) in the educational programmes of the Faculty of Information Technologies: 6B01511-Informatics, 7M01514-SMART-city Technologies and 7M01525-STEM-Education.

At the first stage of the special course on working with big data, the ERA 500 network controller collecting data from sensors was taken as an example. The ERA 500 network controller allows one to work with such activities as data collection, storage and processing in the educational process. The program that supports this device allows one to calculate the number of incoming and outgoing customers in small firms and large enterprises, track the number of customers inside the building and calculate working hours. The simple configuration of the device sorts data for 500 to 10,000 people (Figure 2).

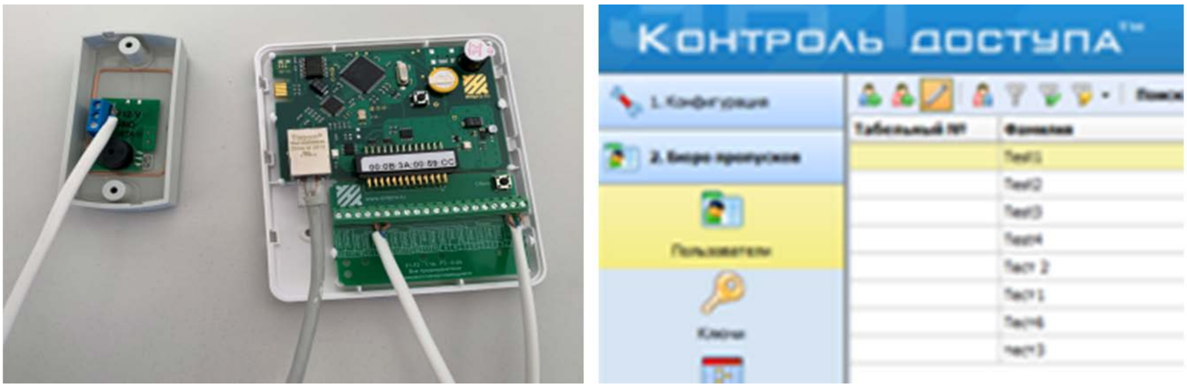


Figure 2: Hardware and software part of the ERA 500 network controller.

The program maintains a list of employees and generates different passage schedules for different groups. The working time is also taken into account, it is possible to control access and save videos. The system also contains further additions [10]. For example, combining several entry and exit points into one passage zone, the ability to prohibit repeated passage, automatically sending reports on the time spent at work to employees, etc. For further processing of the received data, the possibilities of importing data into MS Word, MS Excel, convenient storage and printing of any data are provided.

For processing and analysing the received data at the second and third stages, the BigQuery environment provided by GCP was used, respectively, and the Data Studio tool for developing data visualisation. BigQuery is a cloud database that can process large amounts of data from various sources generated in order to solve the problems of database management systems and to support working with large amounts of data. This solution, which performs SQL-like queries, allows one to analyse large amounts of data in real time [11].

It is known that cloud platforms are implemented through special subscriptions and a fee is charged for the resource used. However, cloud platform providers also offer their own special packages for academic purposes and for novice users in this field. When performing practical work in the educational process, the authors used a free GCP subscription.

To start processing data through BigQuery, a new project is first created, according to which a dataset is formed. After creating a new dataset (in the case presented in this article, the name Big data_univer_edu is assigned), the location area of the data array (the selected territory) is selected. To upload data from different sources to the BigQuery environment, the data must be saved in CSV, JSON, Avro, Parquet and ORC types. In practical work with students from *L.N. Gumilyev* Eurasian National University, the data received from the ERA 500 controller was saved in CSV format, loaded into BigQuery, and at this stage was created a table corresponding to the project.

As a result of the data collected using the turnstile network controller, requests were made in the form of employee working hours, authentication of visiting times and images of students and employees, temporary exclusion of persons with temperatures above 37 degrees Celsius from the building using a specially connected temperature sensor for a pandemic situation and identification of workers with high body temperature in different time periods.

BigQuery has the following ways to store data from a single table processed:

- up to 1 GB on Google Drive in CSV or JSON format;
- up to 16,000 records in CSV or JSON format in the local network;
- in the form of a BigQuery table;
- copy to Google Sheets or Clipboard save data up to 16,000 records.

These criteria were proposed for a free subscription, which was used in this study for educational purposes. However, the amount of data storage in paid subscriptions will be higher.

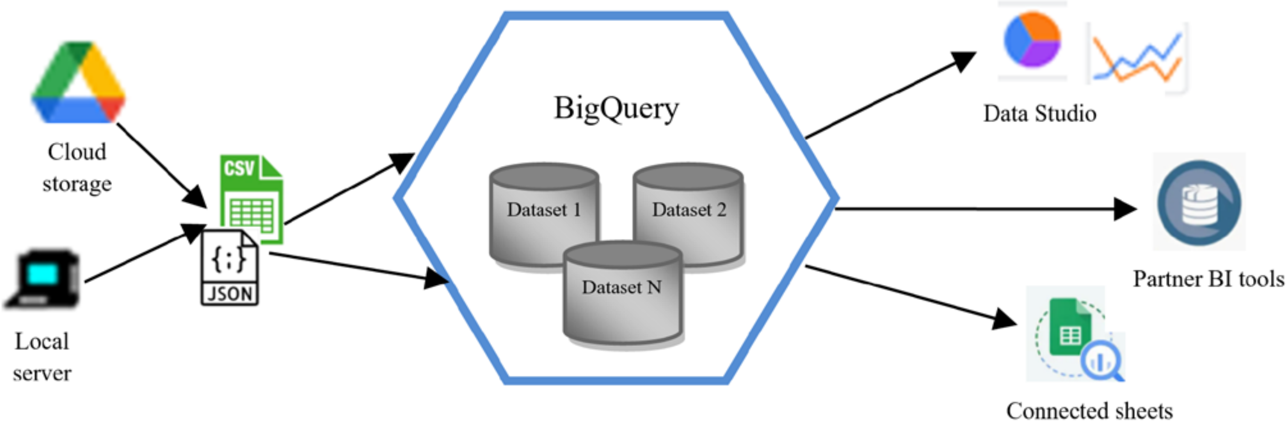
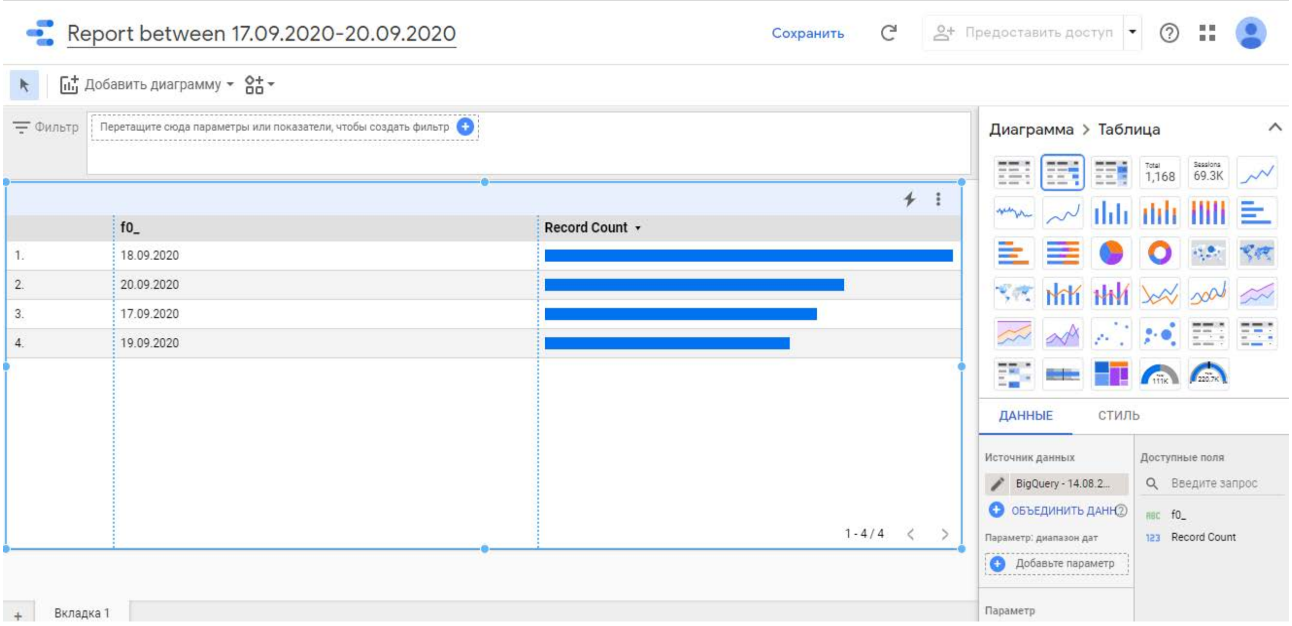


Figure 3: The process of processing big data in BigQuery.

The third step is to create a data visualisation to produce results in a way that is understandable and visible to users after the requests are executed and the data is analysed. BigQuery provides Data Studio, Looker, connected sheets and Partner BI tools for data visualisation (Figure 3). The Data Studio platform, which is considered in practical works within the scope of the training course, is a highly scalable information panel that allows to create data visualisation and business analytics. To simplify data visualisation work in BigQuery, the data export command is provided in the work area to the Data Studio environment. Figure 4 below illustrates a request to analyse the days during the week when employees showed high temperature levels.


Report between 17.09.2020-20.09.2020

	f0_	Record Count
1.	18.09.2020	
2.	20.09.2020	
3.	17.09.2020	
4.	19.09.2020	

1 - 4 / 4

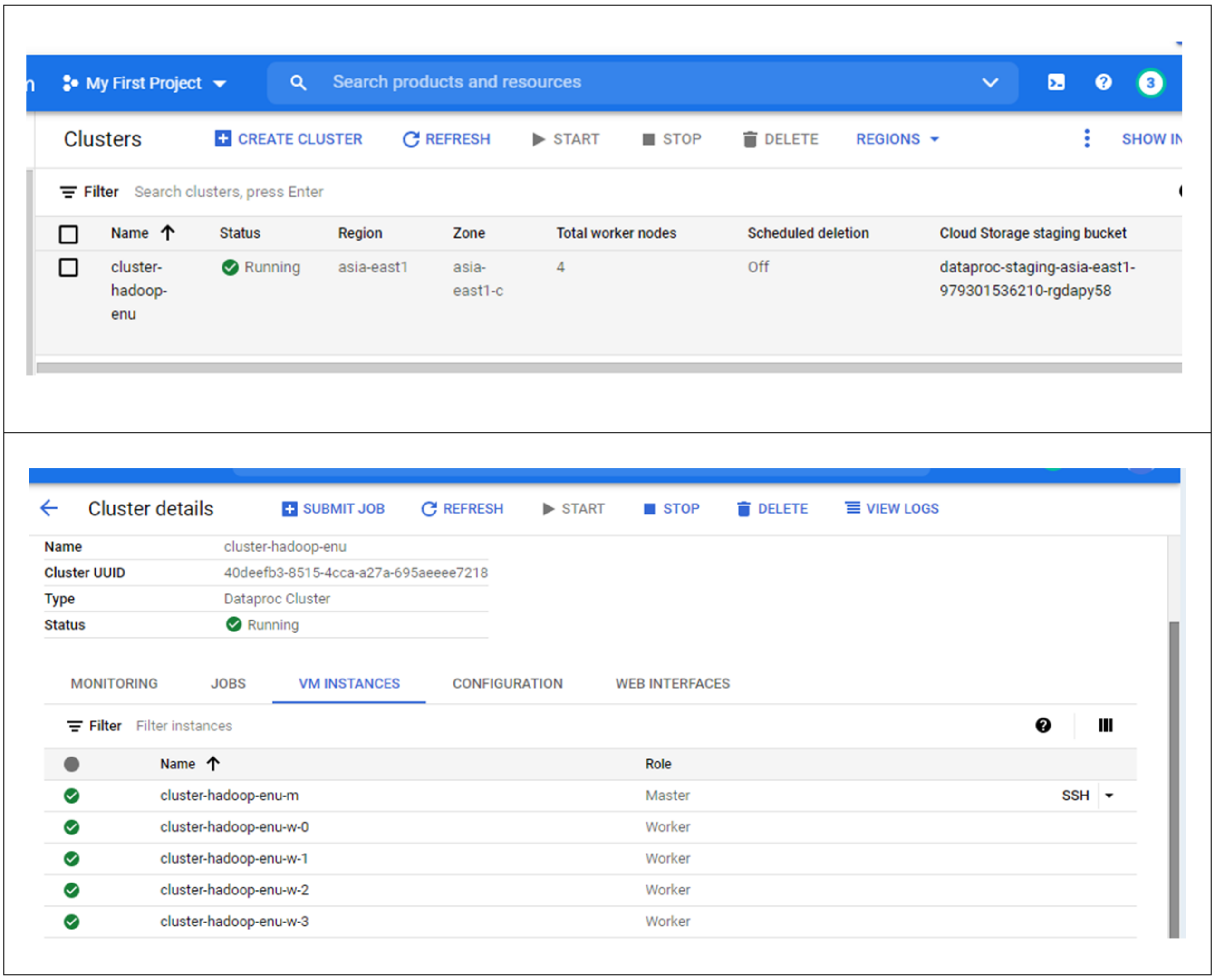
Figure 4: Visualisation of the data collected from the temperature sensor in the network controller.

The exchange of visualisations created in Data Studio with other users is also provided in the system.

The fourth stage, including powerful and basic technologies for storing and processing big data, was devoted to working with the Hadoop framework. The basis of Hadoop technology is access to processing large amounts of data, creating a cluster as a result of combining distributed nodes. To use a set of Hadoop utilities, one can implement it on a local network or on a cloud platform. In this study, the configuration of a cluster consisting of distributed nodes via Hadoop was carried out on GCP. In the case referred to in this article, the cluster is a computing environment necessary for large-scale calculations. So, with the help of the capabilities of the cloud platform, the authors were able to quickly and conveniently perform work that requires many complex technical settings. Previously, the issues of setting up a parallel computing cluster in a local network were considered by the same authors [12][13].

Thus, working with the Hadoop framework on GCP is included in the group of commands of the State Data service. Using this service, one can configure a new cluster in less than two minutes. The cloud platform offers three types of cluster creation: 1) standard (one master, *n* workers), 2) single node (one master, *0* workers) and 3) high availability (three masters, *n* workers). In the case presented in this article, the standard type was chosen to explain the essence of big data processing using parallel computing based on the cluster principle. Using this type, the authors first set up a cluster that would be one master and four workers (Figure 5).

The volume and capability of the working environment for processing big data, created under the name Cluster-hadoop-enu, can be transformed according to one's calculations. Subscribing to a cloud platform also imposes some restrictions for increasing the cluster nodes. However, the proposed limits for students to master the principles of working environments for large-scale computing were sufficient.



□	Name ↑	Status	Region	Zone	Total worker nodes	Scheduled deletion	Cloud Storage staging bucket
□	cluster-hadoop-enu	✓ Running	asia-east1	asia-east1-c	4	Off	dataproc-staging-asia-east1-979301536210-rgdapy58

Name	cluster-hadoop-enu
Cluster UUID	40deefb3-8515-4cca-a27a-695aeeee7218
Type	Dataproc Cluster
Status	✓ Running

●	Name ↑	Role	
✓	cluster-hadoop-enu-m	Master	SSH ▾
✓	cluster-hadoop-enu-w-0	Worker	
✓	cluster-hadoop-enu-w-1	Worker	
✓	cluster-hadoop-enu-w-2	Worker	
✓	cluster-hadoop-enu-w-3	Worker	

Figure 5: The process of configuring the Hadoop cluster via GCP.

The instant execution of complex technical steps through cloud services can lead to students' misunderstanding of the training content without taking into account the architectural features of the topics. In this regard, the authors also carefully considered the appropriate combination of theoretical material with practical tasks. The training course covered basic concepts on databases and their logical concepts, processing structured, unstructured and weakly structured data in BigQuery, differences between SQL and NoSQL, organisation of distributed data and network technologies based on the Hadoop framework, setting up a cluster of high-performance parallel computing and cloud technology architecture. The course was focused on developing a good understanding of this area by students and the formation of the necessary skills.

CONCLUSIONS

Processing and analysing large amounts of data is a complex process that requires huge technical resources. The introduction of bid data into educational programmes of information and communication technology courses can lead to spending a lot of time on the organisation of hardware and software, technically complex processes, configuration settings and writing large program codes. The improved services offered by cloud technologies bring a new impetus to the big data processing situation. It consists in organising a complex process with several technical stages, by the cloud system itself, with the priority of the main working environment.

The bid data training course outlined in this article, along with the necessary theoretical knowledge, contributes to the improvement of students' practical skills using modern software solutions. Thus, it supports the formation of competent graduates who can successfully compete in the labour market of the future. The authors hope that their study will be useful for teachers and students, scientists and practitioners-researchers in this field.

REFERENCES

1. Yu, W. and Su, C., The path of higher education management in the era of big data. *Proc. Inter. Conf. on Applications and Techniques in Cyber Security and Intellig.*, Springer, Cham, 448-453 (2020).
2. The Programme *Digital Kazakhstan*, Approved by the Government of the Republic of Kazakhstan No. 827 (2017), 4 August 2021, https://primeminister.kz/kz/documents/gosprograms
3. Wang, Y., Big opportunities and big concerns of big data in education. *Tech. Trends*, **60**, 381-384 (2016).
4. Park, Y.E., Uncovering trend-based research insights on teaching and learning in big data. *J. of Big Data*, 7, **1**, 1-17 (2020).
5. Matas-Terron, A., Leiva-Olivencia, J.J. and Negro-Martínez, C., Tendency to use Big Data in education based on its opportunities according to Andalusian education students. *Social Sciences*. 9, **9**, 164 (2020).
6. Jiang, J. and Chen. T., Research on the value of smarter education in the era of Big Data. *Proc. 5th IEEE Inter. Conf. on Big Data Analytics*, Xiamen, China, 42-45 (2020).
7. Vidal-Silva, C.L., Madariaga, E.A., Rubio, J.M. and Urzúa, L.A. Study of the reality and viability of the education in big data in the Chilean academy. *Infor. Technol.*, 30, **5**, 239-248 (2019).
8. Serik, M., Mukhambetova, M. and Yeskermessuly, A., Improving the content of a client-server technology training course: set up and collaborative implementation of local and cloud-based remote servers. *Inter. J. of Emerg. Technol. in Learning*, 14, **21**, 191-204 (2019).
9. Alim, N., Md Saad, M.S., Mahmud, H. and Gunawan, F., Usability and satisfaction of Google Classroom as an instructional teaching and learning medium: the students' perspective. *World Trans. on Engng. and Technol. Educ.*, 19, **1**, 16-20 (2021).
10. Application Guide the Era of New Techn. Access Control (2020), https://allsee.kz/p42376586-era-500-setevoj.html
11. Fernandes, S. and Bernardino, J., What is bigquery? *Proc. 19th Inter. Database Engng. & Applic. Symp.*, Yokohama, Japan, 202-203 (2015).
12. Serik, M., Karelkhan, N., Kultan, J. and Zulpykhar, Z., Setting up and implementation of the parallel computing cluster in education. *Inter. J. of Emerg. Technol. in Learning*, 14, **6**, 4-17 (2019).
13. Yerlanova, G., Serik, M. and Kopyltsov, A., High performance computers: from parallel computing to quantum computers and biocomputers, *J. of Phys.: Conf. Ser.*, 1889, **3**, 032032 (2021).