Analysis of student activity on the e-learning course based on «OpenEdu» platform logs

N.D. Barsukov, ORCID: 0000-0003-3962-9087 <nik0xff@gmail.com>
I.M. Sysoev, ORCID: 0000-0001-5748-5529 <ivanabc97@gmail.com>
A.A. Pereskokova, ORCID: 0000-0002-6937-150X <alina.alexandrovna.sh@gmail.com>
I.V. Nikiforov, ORCID: 0000-0003-0198-1886 <igor.nikiforovv@gmail.com>
D. Posmetnijs, ORCID: 0000-0001-9573-9286 <posmetnijs@gmail.com>

Peter the Great St.Petersburg Polytechnic University,
29, Polytechnicheskaya, St.Petersburg, 195251, Russia

Abstract. A lot of people nowadays use online education platforms. Most of them run on the free «Open edX» open-source software platform. Using the logs that the platform provides us, we can get psychometrics of students, which can be used to improve the presentation of material or other things, which can increase the quality of online courses. We provide a ready-to-use tool that will help figure out how and for what purpose you can analyze the log files of platforms based on «Open edX».

Keywords: Big Data; e-learning; analytics; logs; online educational platforms; microservices

Анализ активности студентов на курсах онлайн-обучения на основе логов платформы «OpenEdu»

Н.Д. Барсуков, ORCID: 0000-0003-3962-9087 <nik0xff@gmail.com>
И.М. Сысоев, ORCID: 0000-0001-5748-5529 <ivanabc97@gmail.com>
А.А. Перескокова, ORCID: 0000-0002-6937-150X <alina.alexandrovna.sh@gmail.com>
И.В. Никифоров, ORCID: 0000-0003-0198-1886 <igor.nikiforovv@gmail.com>
Д. Посметныйс, ORCID: 0000-0001-9573-9286 <posmetnijs@gmail.com>

Санкт-Петербургский политехнический университет Петра Великого,
195251, Россия, Санкт-Петербург, ул. Политехническая, д. 29

Abstract. В настоящее время многие люди используют образовательные онлайн-платформы. Большинство из них работают на бесплатной программной платформе с открытым исходным кодом «Open edX». Используя логи, которые предоставляет нам платформа, мы можем получить психометрические данные студентов, которые можно использовать для улучшения представления материала или других вещей, которые могут повысить качество онлайн-курсов. Мы предоставляем готовый инструмент, который поможет выяснить, как и с какой целью вы можете анализировать лог-файлы на основе «Open edX».

Ключевые слова: большие данные; онлайн-обучение; аналитика; логи; платформы для онлайн-обучения; микросервисная архитектура

1. Introduction

Online electronic educational learning platforms are very popular nowadays. One of the biggest and widely used platforms is «Open edX» [1-3]. This is an open-source software platform that provides off-the-shelf tools for educational services. One of the important features of the platform is that it generates student and teacher activity log files. But the disadvantage of the platform is that it does not provide any data analysis tools for monitoring educational progress and success.

One of the most popular educational platform based on «Open edX» in our region is «Open Education» [4]. The problem of the Open edX platform and in particular of one of its implementations «Open Education» is that the teachers, conducting courses on this platform, are missing the tools for analyzing the educational process, which leads to missing control on the educational process and decreasing its efficiency. On the other hand, the students, who use the platform, are also not able to monitor their academic performance.

It's important to provide teachers, course administrators, and students with educational analytics tools that help them to make the educational process more efficient [5] Improving online educational platforms can make online learning at universities more friendly, easy, and happy for all the involved actors.

Our work and project aim consist of several important parts:

- make research on the structure and format of the Open edX platform logs for applicability for automatic analysis of the students' performance. The result of the research showed that all required and important user activity actions (audits) are presented in the logs, so the analytics is possible. Also, we realized that the size of the logs is huge, and they contain millions of actions logged, which makes us think about the usage of Big Data technologies for analytical purposes;
- create and formulate analytic tasks, that can be solved on the logs. The result of that activity is that there is a list of 18 analytics tasks that help students and teacher to monitor the progress;
- implement the software solution, that demonstrates the idea and all the possibilities that Open edX logs provide. As an outcome, there is a tool for extracting, transforming, and preserving logs from a specific course / courses from this platform and the number of analytics tasks implemented in that tool. The result of the analysis is presented in the form of files with metrics, and graphs based on the data obtained.

The paper has the following structure. Section 2 shows the related work. Section 3 describes the system design. Section 4 discusses the implementation process and result of the pilot project. Section 5 presents the conclusion.

2. Related work

There are some articles facing the similar problem of online courses activity analysis.

In [6] (below, Article 1), authors take Moodle as a target platform of further analysis. Authors also use log files as a data source for further analysis. The files are stored in a database, processed and visualized to provide data implementation for teachers who are the final users of that tool. The metrics which authors take for analysis are "the grades of online assignments, reading time, the total number of login times, the total number of online discussions" and others. The choice of these metrics is based on log files content – such values may be easily extracted from raw data.

Speaking about analytics, authors of [7] (Article 2) use advanced machine learning methods such as Random Forest to provide analytics of the online learning process. Particularly, the work describes the prediction of a student's dropout from a course.

Authors of [8] (Article 3) also suggest a method for detection of students who seem to be expelled at the end of the course. Authors use machine learning methods to make the prediction of the further academic performance of students. The prediction is based on logged data of the educational platform.

In [9] (Article 4), authors make statistical analysis of the online course data. The course is running on Moodle platform. Authors also use a self-made logging system to extend log data provided by the platform with new types of recorded events.

Authors of [10] (Article 5) make the visualization of LMS log data. Firstly, they make preprocessing of logs and then draw the scatter plot of student activity within a specific class and the plot of whole faculty activity during one online course.

We've made an analysis of these papers using 5 characteristics. These characteristics describe each of the solutions proposed by papers' authors. The characteristics are:

- *name of educational platform used for analysis* – the name of online educational platform used in article;
- *log files were used for analysis* – did authors use log files to make analysis or not;
- *prediction methods were used in analysis* – do authors use prediction methods in their analysis or not;
- *data visualization was made* – did authors make a visual interpretation of their results or not;
- *eeady-to-use tool was developed* – have authors developed a ready-to-use tool for third-party usage or not.

The result of related works analysis is presented in Table 1.

Table I. Articles' comparison analysis

	Paper 1	*Paper 2*	*Paper 3*	*Paper 4*	*Paper 5*
Name of educational platform used for analysis	Moodle	Not mentioned	Not mentioned	Moodle	Moodle
Log files were used for analysis	+	—	+	+	+
Prediction methods were used in analysis	—	+	+	+	—
Data visualisation was made	+	+	+	+	+
Ready-to-use tool was developed	+	—	—	—	—
Legend: "+" - supported; "—" - unsupported					

According to the result of related works analysis, all of the authors make analysis of student activity during one or more online courses and all of them make the visual interpretation of the results. Most authors base their results on log files data from educational platforms. Talking about platforms, Moodle is the most popular one within paper authors. Some papers also contain descriptions of prediction methods to make forecasts of student academic performance.

However, only one paper describes a ready tool which contains all analytics methods described there and which can be utilized by other people. Also, any of these articles doesn't describe work with the Open edX platform.

Within our work we are going to implement a tool for analysis of online courses data based on log files of the Open edX platform.

Ultimately, our tool differs from others in that we directly process the platform logs, which allows flexibility in the approach to the analysis of what happened on the course. Thus, the teacher can get an answer to the question on a very specific task, in contrast to other tools.

3. System design

In our solution, we are using microservice architecture presented in fig. 1. This allows us to conveniently implement and modify the logic of the service rebuild and redeploy a small part of the tool instead of full application rebuilding. We are using Docker and other DevOps practices. Docker effectively helps us in leveraging microservices architecture [11]. We see three microservices here.

- UI Service – allows the end-user to interact with the application.
- ETL Service – is responsible for receiving, transforming unstructured logs, and loading them into the Database; This service can receive logs from the local machine or directly from the platform based on «opened». In our case, this is the platform – «Open education».
- The analytical service contains a set of analytical scripts that work with the database and a module for building the output file - result or report, which will be sent to the appropriate user interface.
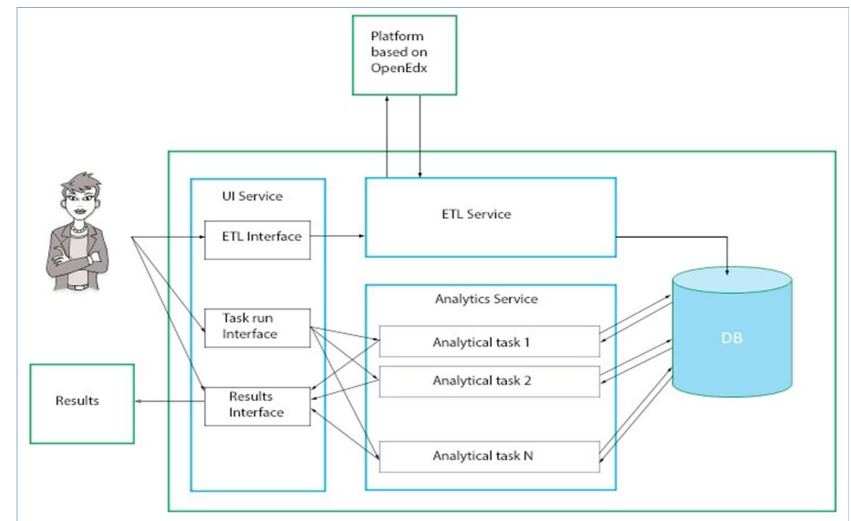

Fig. 1. Architecture of the application

3.1 Log-file structure

All educational platforms can maintain the activity (actions) of users on the platform while undergoing learning on the course or performing test and examination tasks By this way, it becomes possible to analyze user behavior and, based on the obtained analytics, improve educational courses and receive psychometric [12-13] data of students. Due to the improvement of the courses, it will be more convenient and easier for students to learn the material obtained, it will be easier for teachers to distinguish distinguished students and more accurately set final grades.

Барсуков Н.Д., Сысоев И.М., Перескокова А.А., Никифоров И.В., Посметныйс Д. Анализ активности студентов на курсах онлайн-обучения на основе логов платформы «OpenEdu». *Труды ИСП РАН*, том 32, вып. 3, 2020 г., стр. 91-100

Barsukov N.D., Sysoev I.M., Pereskokova A.A., Nikiforov I.V., Posmetnijs D. Analysis of student activity on the e-learning course based on «OpenEdu» platform logs. *Trudy ISP RAN/Proc. ISP RAS*, vol. 32, issue 3, 2020, pp. 91-100

The typical log file is presented in fig. 2. It is a JSON file describing the events occurring in the LMS system [14]. An example of a log-file is shown in figure 1. An event is an entity that describes individual user activity in a course (for example, enrolling in a course, watching a video lecture, sending a response during testing, etc.). In the log file, the event is represented by a JSON object and contains a set of fields.

```json
{
    "ip":"anon_22e8ad25256ffdc60a8cb8957e4ebdb7dc8f3f62a2eac77a285b1bdb1c0d82ea",
    "host":"courses.openedu.ru",
    "page":null,
    "time":"2018-09-09T06:35:17.172512+00:00",
    "agent":"Mozilla/5.0 (Windows NT 6.3; WOW64) AppleWebKit/537.36 (KHTML, like Gecko)
    "event":"{\"POST\": {}, \"GET\": {}}",
    "context":{
        "path":"/api/extended/calendar/course-v1:spbstu+PHYLOS+fall_2018",
        "org_id":"",
        "user_id":955006,
        "course_id":""
    },
    "referer":"https://openedu.ru/my/",
    "username":"alive-proud-snarespenguin",
    "event_type":"/api/extended/calendar/course-v1:spbstu+PHYLOS+fall_2018",
    "event_source":"server",
    "accept_language":"ru,en;q=0.9"
}
```

Fig. 2. Log example

Among the fields that are most interesting as part of the analysis tasks, one can single out the time field (the time the event was recorded in the log file), user_id (user identifier - the initiator of the event), course_id (course identifier) and event_type (the type of event listed in the documentation). A complete list of events used in Open edX is given in the documentation.

3.2 User interfaces

The latest stable version of our tool provides the CLI (command-line interface). After starting the tool, the user is asked about the logs that he would like to analyze, and all available log names are shown to the user in a list. The user must select the name of the log and enter it. The next step is to select an analysis task. The user will see all available tasks and will have to enter the number of the selected task for analysis.

Some tasks require additional input parameters to run. If the user selects such a task, he will be given the corresponding messages in the console, and the user will have to enter the necessary parameters, such as launching the task for all users or only one selected user. After starting this task, the user must wait for its completion. The user will receive a message with information about the placement of the results, and if any graphs are created after the task is completed, they will be automatically opened in the browser.

In addition, we are developing a new graphical user interface that is not yet included in the stable version. It consists of several pages on which the user can see the instructions, select the logs, start the analysis, and see the results. The graphical user interface is more user-friendly.

We have our log analysis algorithm. Using a query in the database, we get the information necessary for analysis from the log, then we analyze it using various mathematical techniques. In one of the tasks, we use some innovation.

3.3 Innovative aspects of the design

Since the tasks of analytics depend on the requirements of the customer, it is necessary to create such a backend architecture in which we can easily integrate new tasks for analytics. OOP architecture and Reflections API is well suited for solving this problem, so we can add new tasks inherited from an abstract task by adding only the database queries without changing the architecture of the project.

At each startup, the system overloads the log, which makes it resistant to user crashes and software error implementation.

4. System implementation

4.1 Description of the implementation

We use laptops for development and production deployment. Our tool is designed for teachers who create their courses and for administrators of educational platforms based on the Open Source "Open edX" software platform. We designed our tool for any hardware platform which meets minimum requirements: 4GB RAM, 2.5Ghz processor.

For user interface we use React, Redux libraries, webpack, and Babel for modules bundling and converting JavaScript code to backward-compatible representation. Npm is a package manager we use for handling packages in development. Lodash is a modern JavaScript utility library delivering modularity, performance, and extras. (See Table 2 for licensing information.)

For analytical and ETL services we are using Java 8 with Spring 5 [15]. For better readability of Java code, we use the Lombok library [16], which allows you to reduce the boilerplate code (constructors, «Object»-methods and etc.) by using annotations. Reflection allows us to look at existing tasks in the project, create a list of tasks and give it to the UI. For documentation we use swagger, it allows us to create documentation in a semi-automatic mode for our services.

We use PostgreSQL [17] as a database. Our build system – Gradle [18-19]. Interaction between database and service is provided by JDBC [20] driver. Our system is RESTful. All microservices use the REST paradigm to interact with each other.

Table 2. Components acquired from external sources

Library	License
Spring	Apache License 2.0
Springfox-swagger	Apache License 2.0
Reflections	BSD 2-clause
Lombok	MIT License
React	MIT License
React-bootstrap	MIT License
Redux	MIT License
Webpack	MIT License
Lodash	MIT License

4.2 Innovative aspects of the implementation

In the implementation of our system, we decided to record each log file (which is a JSON-file) as every row in our PostgreSQL database. So, we use it like a No-SQL database. SQL query for logs selection is presented in fig. 3.

PostgreSQL has many standard functions such as working with strings or JSON-files, which allow working with logs more effectively. We are thinking about transferring our database to No-SQL [21] in the future if it is proved that it will be more productive.

```
SELECT log_line -> 'username' as user_name,
       log_line #>> '{context, user_id}' AS user_id
FROM logs
WHERE log_line -> 'username' != 'null'
```

Fig. 3. Select query example

5. Results

5.1 System output

As a result, we received a solution that provides course administrators the following functionalities:

- download log files from educational platforms;
- choose log files between different courses on a platform;
- store log files in a database;
- run analytic tasks based on downloaded logs, a few of them:
  1) Calculate total user time on the course and user time distributed per day;
  2) Show activity type for all users (or for a particular user) on course depending on the date;
  3) Show the user way over the pages;
  4) Show amount of video play events per day;
  5) Show words from pdf search field;
  6) Get video watching durations by elements of course.

The system gives the calculations in the form of tables saved in XLSX or CSV format and generated graph from these tables.

On the graph in figure 4 we can see the rule that the user visited on the course. We can conclude that this user watches the lectures for two months. After that he decided not to continue studying but we can write him a letter and find his opinion perhaps he didn't like this course or problem was in the poor presentation of material.
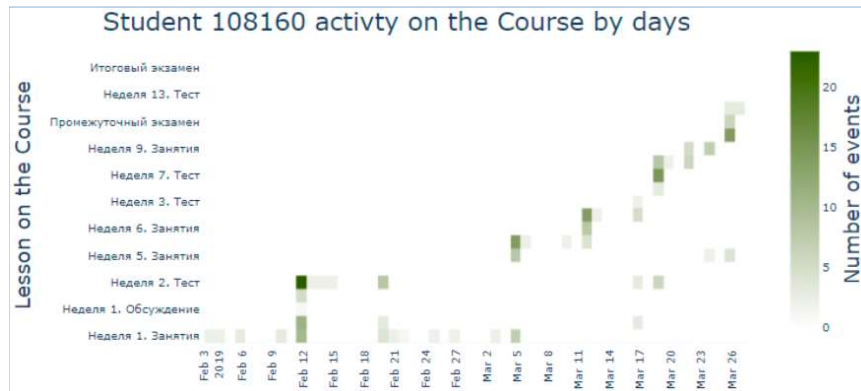

Fig. 4. User activity on course graph

5.2. Challenge and issues

The main challenge was to study the logs and their loading, conversion, and analysis. It was required to read and learn a lot of Open edX documentation. On the other hand, the initial log file that we took was 18Gb to analyze, which made it almost impossible to leant it manually, so we had to create

simple pasting scripts to learn it fast, and only then we were able to retrieve a small log for testing purposes.

Our current database design has limitations that do not allow us to process the logs as fast as we wanted. Therefore, we have found a solution that is now improving. Logs can be larger than 18 GB and contain more than 10 000 000 events generated with semi-structured information inside, so the problem is to process the logs quickly enough.

Open edX platform doesn't provide a good and quick API for logs downloading for offline analysis, so that feature request and feedback has been provided to the platform administrators, but unfortunately, until today they have not implemented the required functionality for us.

References / Список литературы

[1]. Open edX official website, Edx Documentation Resources. Available at: https://docs.edx.org/

[2]. Blagojević, M., Milošević D.: Massive open online courses: EdX vs Moodle MOOC. In Proc. of the 5th International Conference on Information Society and Technology, 2016, pp. 346-351.

[3]. Sriram M. Comparative Analysis of Massive Open Online Course (MOOC) Platforms. In Proc. of the 4th International Conference on Global Business, Economics, Finance and Social Sciences, 2015, pp. 1-7.

[4]. Open Education official website, About the Project. Available at: http://npoed.ru/about (in Russian).

[5]. Krasnov S., Kalmykova S., Abushova E., Krasnov A. Problems of Quality of Education in the Implementation of Online Courses in the Educational Process. In Proc. of the International Conference on High Technology for Sustainable Development (HiTech), 2018, pp. 1-4.

[6]. M. Furukawa, K. Yamaji, Y. Yaginuma and T. Yamada. Development of learning analytics platform for OUJ online courses. In Proc. of the IEEE 6th Global Conference on Consumer Electronics (GCCE), 2017, pp. 1-2.

[7]. B.B. Mishra and S. Mishra. Quality Improvements in Online Education System by Using Data Mining Techniques. In Proc. of the 2nd International Conference on Data Science and Business Analytics (ICDSBA), 2018, pp. 532-536.

[8]. Nobuhiko Kondo, Midori Okubo, Toshiharu Hatanaka. Early Detection of At-Risk Students Using Machine Learning Based on LMS Log Data. In Proc. of the 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI), 2018, pp. 198-201.

[9]. P. Esztelecki . G. Korosi. Analysis of a short on-line course through logged data recording by a self-developed logging module. In Proc. of the International Conference on Computer, Information and Telecommunication Systems (CITS), 2018, pp. 1-5.

[10]. R. Raga, Jennifer Raga. A comparison of college faculty and student class activity in an online learning environment using course log data. In Proc. of the IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation, 2017, pp. 1-5.

[11]. Jaramillo D., Nguyen D., Smart R. Leveraging microservices architecture by using Docker technology 2016. In Proc. of the IEEE Region 3 South East Conference (SoutheastCon), 2016, pp. 1-5.

[12]. Kalmykova S.V., Chapaykina M.D., Shirokova S.V. Application of a systematic approach to the project management implementation of e-learning in a classical university (on the example of the Peter the Great University). Obrazovatel'nye tehnologii, no. 2, 2018, pp. 67-74 (in Russian) / Калмыкова С.В., Чапайкина М.Д., Широкова С.В. Применение системного подхода к управлению проектом внедрения электронного обучения в классическом университете (на примере ФГАО ВО СПбПУ Петра Великого). Образовательные технологии, no. 2, 2018 г., стр. 67-74,

[13]. Kravchenko D. Psychometrics in online education. University Book: Journal of Information and Analysis, no. 3, 2019, pp. 52-55 (in Russian) / Кравченко Д. Психометрика в онлайн-образовании. Университетская книга: информационно-аналитический журнал, no. 3, 2019 г., стр. 52–55.

[14]. Edx, Student Events. Available at: https://edx.readthedocs.io/projects/devdata/en/latest/internal_data_formats/tracking_logs/student_event_types.html

[15]. Spring official website, Microservices. Available at: https://spring.io/microservices.

[16]. Project Lombok Official website, Lombok Features. Available at: https://projectlombok.org/features/all.

[17]. PostgreSQL official website, Documentation. Available at: https://www.postgresql.org/docs/10/index.html

[18]. Gradle official website. Gradle Features. Available at: https://gradle.org/features/.

[19]. Voinov N., Rodriguez Garzon K., Nikiforov I., Drobintsev P. Big Data Processing System for Analysis of GitHub Events. In Proc. of the 2019 XXII International Conference on Soft Computing and Measurements (SCM), 2019, pp. 187-190

[20]. Oracle Help Center, Lesson: JDBC Introduction. Available at: https://docs.oracle.com/javase/tutorial/jdbc/overview/index.html.

[21]. Strauch Christof. NoSQL Databases. 2012. Available at: https://www.christof-strauch.de/nosqldbs.pdf.

Information about authors / Информация об авторах

Nikita Dmitrievich BARSUKOV is a graduate student at the Institute of Computer Science and Technology. Research interests: development of highly loaded software, development of corporate software systems, big data analytics, financial technologies, software verification.

Никита Дмитриевич БАРСУКОВ – студент магистратуры Института компьютерных наук и технологий. Область интересов: разработка высоконагруженного программного обеспечения, разработка корпоративных программных систем, аналитика больших данных, финансовые технологии, верификация программного обеспечения.

Ivan Mikhailovich SYSOYEV – graduate student at the Institute of Computer Science and Technology. Research interests: Big Data, Big Data Analytics, Data Model Building, Data Visualization.

Иван Михайлович СЫСОЕВ – студент магистратуры Института компьютерных наук и технологий. Область интересов: большие данные, аналитика больших данных, построение моделей данных, визуализация данных.

Alina Aleksandrovna PERESKOKOVA – graduate student at the Institute of Computer Science and Technology. Area of interest: big data, big data analytics, distributed data storage systems, interface development, big data display.

Алина Александровна ПЕРЕСКОКОВА – студентка магистратуры Института компьютерных наук и технологий. Область интересов: большие данные, аналитика больших данных, системы распределенного хранения больших данных, разработка интерфейсов, отображение больших данных.

Igor Valerievich NIKIFOROV – candidate of technical sciences, associate professor of the Higher School of Software Engineering at the Institute of Computer Science and Technology. Research interests: parallel data processing, distributed data storage systems, big data, software verification, test automation.

Игорь Валерьевич НИКИФОРОВ – кандидат технических наук, доцент Высшей школы программной инженерии Института компьютерных наук и технологий. Сферы научных интересов: параллельная обработка данных, системы распределенного хранения данных, большие данные, верификация программного обеспечения, автоматизация тестирования.

Deniss POSMETNIJS – graduate student at the Institute of Computer Science and Technology. Research interests: Big Data, Big Data Analytics, Development of End User Interfaces (Front End), Data Visualization.

Денисс ПОСМЕТНЫЙС – студент магистратуры Института компьютерных наук и технологий. Область интересов: большие данные, аналитика больших данных, разработка интерфейсов конечного пользователя (фронтенд), визуализация данных.