Technical University of Sofia
Faculty of Computer Systems and Technologies

Eastern Macedonia and Thrace Institute of Technology

PROCEEDINGS

EIGHTH INTERNATIONAL SCIENTIFIC CONFERENCE

COMPUTER SCIENCE' 2018

UNITe 2018 WORKSHOP

Big Data and Smart Industry

CONFERENCE CHAIRS
Prof. Milena Lazarova,
Technical University of Sofia, Bulgaria
Prof. Dimitrios V. Bandekas,
Eastern Macedonia and Thrace Institute of Technology, Greece

HONORARY CO-CHAIR:
Ognyan Nakov,
Technical University of Sofia, Bulgaria

ISBN: 978-619-167-177-9

Contents

Framework for Big Data Analytics of Moodle Data Using Hadoop in the Cloud

Aleksandra Nikolovska, Aleksandar Velinov, Stojance Spasov, Zoran Zdravev
Goce Delcev University, Stip, Republic of Macedonia, anikolovska11@gmail.com, aleksandar.velinov@ugd.edu.mk, stojance.spasov@ugd.edu.mk, zoran.zdravev@ugd.edu.mk

***Abstract**: Big Data includes huge volume, high velocity and extensible variety of data, both structured and unstructured. Analysis and examination of this data plays a big role in making a better decisions and lead to strategic business moves. In this research, we are focusing on analysis of Moodle e-learning platform database with aim of improving the educational process. Using the concept of learning analytics we are trying to better understand teaching and learning processes in order to get as much as possible from educators and learners which are the main actors in the mentioned activities. The standard procedures that are specific to the databases will take a lot more time to do this. To facilitate this process, big data processing tools are used. Here, we also present the possibility of analyzing big data in the cloud. For the purpose of this research, we developed a framework for big data analysis.*

***Keywords:** Moodle, Hadoop, Big Data, Analysis, Learning Analytics.*

1. INTRODUCTION

Due to the advent of new technologies, devices and communication means, the amount of data produced is growing rapidly every year. Big Data involves the data produced by different devices and applications like social media data for example, or search engine data. Thus we can say that Big Data includes huge volume, high velocity and extensible variety of data, both structured and unstructured. Analysis and examination of this data plays a big role in making a better decisions and lead to strategic business moves. Big data affects organizations across practically every industry: banking, government, health care, education, manufacturing and many others. The database of the Moodle E-Learning platform at the Goce Delcev University, will be used for the purposes of this research. The database contains data for a period of six years (from 2012 to 2018). The size of the database that contains the data from this period is 6.9 GB. Besides the database, we will also analyze logs from the server that have a size of about 10GB. In this research, we are focusing on usage of Big Data analysis with aim of improving the educational process. While speaking of Big Data in education, we are coming to the idea of learning analytics. Learning analytics is defined as measurement, collection, analysis and reporting of data about learners and their contexts, in order to understand and optimize learning and the environments in which it occurs [1]. Actually, by utilizing learning analytics tools, our intention is to enhance the courses results and establish more effective teaching methods and learning procedures. As we are explaining in Section 2 and Section 4 of this research, our starting point is to prepare a good dataset for further examination. In order to do that, or in other words, to build a strong basis for reaching our goal, we have used tools for storing and analysis of Big Data, which are mentioned in Section 3 of this paper. In this section, we are presenting the features of the tools we have used, and our experiences while using these tools. In Section 5

of this research, we are focusing on explaining our defined framework for Big Data analytics in order to find meaningful relations in a huge amount of data and make conclusions based on extracted knowledge.

2. MOODLE DATABASE AND LOG FILES

Moodle [2] is a platform, which offers the possibility of accessing learning content to all students, creating courses, reporting forms, assignment submissions to university courses. Moodle is one of the most widely used e-learning tools. It is open source and anyone can use it and change its code. Moodle is already widespread in all segments of education, from primary schools to universities [3]. Moodle consists of set of functionalities grouped into two groups: modules and resources [4]. Resources are educational materials that are in digital form and attached to the e-learning platform. Examples of resources are PDF documents, images, Word documents, PowerPoint presentations and so on. The modules are integrated parts of the Moodle platform that enable the interaction of teachers and students. Examples of module are Forums, Assignments, Glossary, Lessons, Survey, Quizzes, Choices, Chats, and so on. All students and teachers data are stored in the Moodle database.

For our research, we plan to analyze selected tables from the Moodle database. The database contains a lot of data. Moodle has a MySql database and the size of the database is expressed in gigabytes (GB). Data mining tools are used to discover patterns in data sets. The problem occurs when the number of data in the tables in the databases becomes extremely large. It is common for the systems that have a large number of users, such as the Moodle e-learning platform. Then data management becomes more complicated. This also complicates the process for extracting knowledge from databases. The standard procedures that are specific to the databases will take a lot more time to do this. To facilitate this process, big data processing tools are used. There are a number of tools that are specific to databases and which can also extract knowledge. We will look at some of these tools in Section 3 of this research. First, we must investigate the Moodle database.

We plan to analyze the Moodle logs that also contain a lot of data. The logs contain data for all user actions on Moodle platform. Like the Moodle database, the size of the logs is also expressed in gigabytes (GB), although this is all depending on the time period in which the Moodle platform is used. We expect to get knowledge from the analysis of Moodle logs.

3. TOOLS FOR BIG DATA ANALYTICS

Data analytics is the process of examining, cleaning, transforming and modeling large data sets with the goal of discovering hidden patterns, unknown correlations and collect useful information. For the purpose of this research, we investigated the Cloudera Hadoop [5] – an open source framework for distributed storage and distributed processing of large data sets. The key component of Hadoop framework is HDFS (Hadoop Distributed File System) which can store a huge amount of structured, semi-structured and unstructured data. MapReduce is the second core component of this framework and it represents the processing layer, which programming model is designed for handling large volumes of data in parallel by dividing the work into a set of independent tasks. This actually represents the heart of

Hadoop and its power and efficiency is due to parallel processing done in this layer of the framework. Querying and analysis of data stored in HDFS is provided by Hive, powerful data warehouse infrastructure software built on top of Hadoop. Hive chooses respective database servers to store the metadata of tables, columns in a table, their datatypes and HDFS mapping. Because Map-Reduce framework is very low level and requires writing custom programs, which are hard to maintain and reuse, Hive provide SQL-like language called HiveQL for querying the structured data. The conjunction part of HiveQL and MapReduce is Hive execution engine, which compiles the queries into Map-Reduce jobs that are executed in Hadoop. Besides HiveQL, for our research, we also investigated Impala, which is massive parallel processing SQL query engine for handling huge volumes of data that is stored on Hadoop cluster. It provides high performance and low latency compared to other SQL engines for Hadoop. We have examined the time needed for query to be executed while running it on HiveQL and Impala. When we compared the query execution time, we concluded that Impala is faster than Hive. The reason for this is the fact that Impala is not based on MapReduce algorithm. Instead, it implements a distributed architecture based on daemon processes that are responsible for all the aspects of query execution that run on the same machines.

With the increasing popularity of cloud computing, we can think of Hadoop in the cloud. This "Hadoop in the cloud" paradigm denotes the execution of Hadoop clusters on resources offered by a cloud provider [6]. The cloud instance is identical to a server to which we access remotely, with root access, some number of CPU cores, and some amount of disk space [6]. There are several reasons why to run Hadoop in the cloud such as flexibility, speed of change, lower risk, worldwide availability, cloud provider features, capacity, efficiency and reduced cost. We can use some cloud services to use Hadoop in the cloud and analyze big data.

Cloudera can deploy Cloudera Distribution including Apache Hadoop (CDH) in the public cloud environments: Amazon Web Services (AWS), Google Cloud Platform (GCP), Microsoft Azure [6].

Cloudera has the following products that can work with big data in the cloud:

- Data engineering
- Analytic Database
- Operational Database

Data engineering provides CDH components for fast and cost-effective storing and processing of data, including Spark and Navigator. Analytic Database provides CDH components for discovering, analyzing, and understanding data, including Impala, Hive-on-Spark, and Navigator. Operational database provides CDH components for building data-driven applications, including HBase, Spark Streaming, and Navigator [6].

We can use the Amazon Elastic Map Reduce (EMR) to create a fully configured elastic clusters on Amazon EC2 running Hadoop and the other tools that are related with Hadoop [6]. Hadoop MapReduce is used to break down the job into smaller pieces that can be distributed across nodes in the Amazon EMR cluster [7]. The Amazon Simple Storage Service (S3) can be used as a data layer for Hadoop. Amazon S3 has a characteristics such as low cost, scalability and durability. It is perfect for storing big data. In this way we will separate the storage layer (Amazon S3) from the processing layer (Amazon EMR). In the future, to save costs we can terminate Amazon EMR when it is idle. The data will remain in the Amazon S3.

Amazon EMR File System (EMRFS) can be used for reading and writing files from Amazon EMR to Amazon S3 [8]. EMRFS is implementation of Hadoop Distributed File System (HDFS). All EMR clusters use it as a file storage system. We can store our data in Amazon S3, and we can process and analyze them using Hadoop in Amazon EMR. HDFS, which stores the data across local disks of the cluster, can also be used as data storage layer. We can store our data in the cloud and later using Hadoop and other tools we can create the dataset.

4. CREATING DATASET (QUERY, TABLES)

Moodle database has 392 tables. Some of them are empty while some contain a lot of data. In the first phase of our research, we looked at all the tables in the databases.

We have identified potential research tables [10]. Some of them are given in Table 1. Because the tables contain a large number of data, it is a good basis for gaining knowledge from them. The tables also contain personal data. We need to omit these data in order to protect the privacy of the users. This process is called depersonalization.

Using the potential research tables, our goal is to create a dataset that will be the target of the analysis. Using queries and linking tables, we can first create sub-tables.

Tab. 1: Some of the potential research tables

Table name	Description
mdl_assign	Assignment that a course creator has included in a course
mdl_assign_submission	Here are records of submissions for each assignment
mdl_assign_grades	Grades for the submitted assignments
mdl_choice, mdl_answers and mdl_option	Each instance of the Choice activity is recorded here. Here are also the configuration settings for each Choice activity.
mdl_course	There is an entry in this table for every course that has been created in your Moodle site. There are 33 fields for each entry, including creation date and language.
mdl_forum	Every time a forum is created (note that there is a default one for every course), a record is added in this table. There are several other tables that are used in conjunction with forums.
mdl_forum_discussions	Discussions of the users on forums
mdl_posts	Blogging posts are stored here
mdl_lesson	For each Lesson activity that you create for any course, a record is entered into this table
mdl_quiz	Quiz activity in courses
mdl_user	Each Moodle user has an entry in this table
mdl_logstore_standard_log	Almost all visitor interactions are logged in this table
mdl_role_assignment	The user role in Moodle

The idea behind the sub-tables is to facilitate the process of creating complex queries that require a lot of time to execute. These tables, will be later used for other research that has a different context. After completing this dataset, we will announce it to be publicly available, and anyone who wants, can use it.

5. FRAMEWORK FOR BIG DATA ANALYTICS

As we said earlier in this paper, our main goal is to develop a good framework for Big Data analysis. According to some research [11] the process of data pre-processing is considered the most crucial phase in the whole data analytics process

and it can take more than a half of a total time spent in solving the data mining problems. That is the reason why we devoted more attention and effort to build a rich dataset that will be used afterwards in order to gather useful information and gain knowledge from our Big Data. We have done a lot of investigation before we decide to use the tools mentioned in Section 3 of this paper. Before that, we also did some exploring of the tables from Moodle database, so we can make a decision which one to be used as a basis for constructing our dataset. Here, because our starting point is Big Data collected from LMS (learning management system) the usage of learning analytics tools is inevitable in order to reach our final target - locate the drawbacks in the teaching processes and try to fix and adjust them. There are number of reports, blocks and other plugins for Moodle that provide learning analytics. Some of them are logs, activity, activity completion, course participation, course overview, course completion status, events list, live logs and so on [12]. In order to find appropriate correlations and collect valuable information from the finalized dataset, we are planning to use some of the tools for data science and machine learning. At the time of writing, our intention is to use Rapid Miner or WEKA as tools for clustering our dataset. After we get the expected results from the clustering, our idea is to build reports based on obtained information and knowledge.

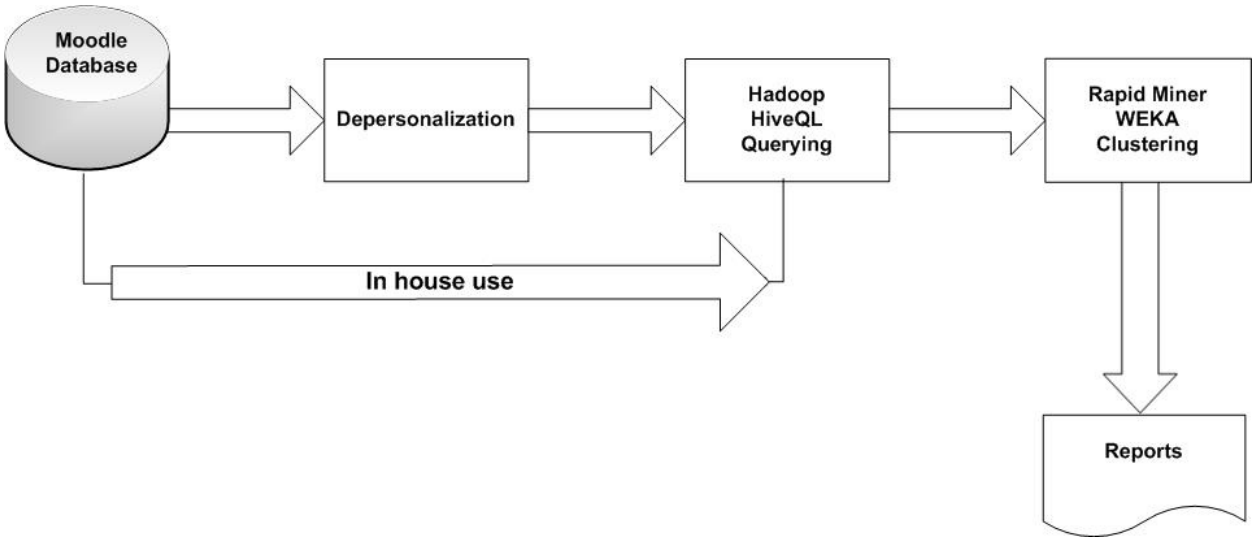

Fig.1 Framework for Big Data analytics

6. CONCLUSION

Nowadays there is abundant of tools for Big Data analysis, modeling or transforming data with the goal of discovering worthy information and supporting decision making process. Starting from the fact that the amount of data is growing rapidly, it is not so difficult to predict that these tools will become a compulsory part of almost every organization and institution. In that sense, we would like to emphasize the essential role of choosing the right tools and frameworks in the field of data science, so to extract as much as possible knowledge from the large volumes of data. In our case we are dealing with Big Data originated from the educational processes, we can speak of learning analytics which in simple words is collecting the traces that learners leave behind and using those traces to improve learning [13]. Having this in mind, by applying the learning analytics tools on Big Data gathered from learning management systems, we can predict future student achievements, we

can define group of students which are not likely to succeed, or we can modify teaching styles and methods. Apart from defining the at-risk students, learning analytics also can help teachers to find out whether their students are understanding the material in the courses. In that way, with the usage of learning analytics software, we get the directions on how instructors and students should proceed with their activities, in order to have better results in the teaching processes and educational systems. It is important to remember that regardless the area in which we are researching, and no matter how we collect the Big Data, it's primary value comes not from the data in its raw form, but from the processing and analysis of it and the insights, products and services that emerge from analysis. The importance of the Big Data does not revolve around how much data we have, but what we do with it.

7. REFERENCES

[1]. Scapin, R. (2015). Learning Analytics in Education: Using Student's Big Data to improve Teaching. Dawson College.
[2]. (2018) Moodle. [Online]. Available: https://moodle.org/
[3]. Umek, L., Keržic, D., Tomaževic, N., & Aristovnik, A. (2015). Moodle E-Learning System and Students' Performance in Higher Education: The Case of Public Administration Programmes. International Association for Development of the Information Society.
[4]. Costa, C., Alvelos, H., & Teixeira, L. (2012). The use of Moodle e-learning platform: a study in a Portuguese University. Procedia Technology, 5, 334-343.
[5]. (2018) Cloudera. [Online]. Available: https://www.cloudera.com/
[6]. Havanki, B. (2017) Moving Hadoop to the cloud. Sebastopol; USA: O'Reilly Media, Inc.
[7]. (2018) Apache Hadoop on Amazon EMR. [Online]. Available: https://aws.amazon.com/emr/details/hadoop/
[8]. (2018) Using EMR File System (EMRFS). [Online]. Available: https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-fs.html
[9]. (2018) Cloudera Enterprise in the Cloud. [Online]. Available: https://www.cloudera.com/documentation/director/cloud/topics/cloud_overview.html
[10]. (2018) Moodle database tables. [Online]. Available: https://schoolitexpert.com/web/moodle/584-moodle-database-tables
[11]. Romero Morales C., Applying data mining in Moodle, Department of Computer Sciences and Numerical Analysis, University of Cordoba
[12]. (2015) Moodle News. [Online]. Available: https://www.moodlenews.com/2015/learning-analytics-tools-available-in-moodle-moodleresearch-moodleworld/
[13]. Duval, E. (2012). Learning Analytics and Educational Data Mining [Online]. Available: https://erikduval.wordpress.com/2012/01/30/learning-analytics-and-educational-data-mining/