LASyM: A Learning Analytics System for MOOCs

Yassine Tabaa
Information and Communication Systems Laboratory
College of Sciences, Abdelmalek Essaadi University
Tetouan, Morocco

Abdellatif Medouri
Information and Communication Systems Laboratory
College of Sciences, Abdelmalek Essaadi University
Tetouan, Morocco

***Abstract*—Nowadays, the Web has revolutionized our vision as to how deliver courses in a radically transformed and enhanced way. Boosted by Cloud computing, the use of the Web in education has revealed new challenges and looks forward to new aspirations such as MOOCs (Massive Open Online Courses) as a technology-led revolution ushering in a new generation of learning environments. Expected to deliver effective education strategies, pedagogies and practices, which lead to student success, the massive open online courses, considered as the "linux of education", are increasingly developed by elite US institutions such MIT, Harvard and Stanford by supplying open/distance learning for large online community without paying any fees, MOOCs have the potential to enable free university-level education on an enormous scale. Nevertheless, a concern often is raised about MOOCs is that a very small proportion of learners complete the course while thousands enrol for courses. In this paper, we present LASyM, a learning analytics system for massive open online courses. The system is a Hadoop based one whose main objective is to assure Learning Analytics for MOOCs' communities as a mean to help them investigate massive raw data, generated by MOOC platforms around learning outcomes and assessments, and reveal any useful information to be used in designing learning-optimized MOOCs. To evaluate the effectiveness of the proposed system we developed a method to identify, with low latency, online learners more likely to drop out.**

Keywords—Cloud Computing; MOOCs; Hadoop; Learning Analytics.

I. INTRODUCTION

Nowadays, Cloud Computing [1] has laid the ground for a new generation of educational systems, by providing scalable anytime/anywhere services simply accessed through the Web from multiple devices without worrying how/where those services are installed, maintained or located. The Web [2] ushered in a new era of possibilities and expectations for transforming education as it was stated by many studies and reports. With its promise of virtually "infinite resources", Cloud Computing has consolidated the ubiquity of the Web in several learning aspects [3] and made feasible widened access to quality educational materials and courses. Thus, any educational institution can exploit and share its teaching expertise and learning resources through a global online presence. In 2012, new endeavors such as edX [4], Coursera[5] and Udacity [6] introduced more than 200 online costless college courses made accessible to any person connected to the Internet. These courses are called MOOCs, Massive Open Online Courses [7], and they exploit web technologies [2] to offer free online education to as many persons as possible. In May 2012, Harvard and MIT inaugurated the non-profit edX and, since then, the University of Texas and the University of California Berkeley have joined them. The for-profit MOOC platform, Courseara, was initiated after the joint of 33 colleges and it exposes contributions from Princeton, Stanford, Penn, Duke, Ohio State, the University of Virginia and other colleges. Another for-profit MOOC platform, Udacity, was co-created by Stanford professor Sebastian Thrun, David Stavens, and Mike Sokolsky. Although actually most MOOCs do not offer credit, students can learn at their own pace and receive electronic certificate of accomplishment.

Leading US massive open online course providers have each almost increased the number of universities offering courses; for instance, Coursera delivers some 332 courses to its 3.1 million students since its launch in 2012, thereby generating big data as Web logs of activities and learning operations. However, course completion rates have gotten a lot of attention: as reported in Katy synthesis [8] which compares the ratio of students completing a course to total number of students registered on a variety of courses provided by several MOOC providers, for some courses the rate does not reach 2%. Accordingly, two interesting aspects may well be enhanced in future designed MOOCs; namely, decreasing the high MOOCs' dropout rates, and optimizing learning operations through MOOC platforms.

In this paper, we propose LASyM, a Learning Analytics System for MOOCs, whose core aim is to mine MOOCs' big data, essentially generated by user through learning operations on MOOC platforms, using a Cloud based Hadoop [9] to ultimately analyse students' behaviour with the intent of increasing the impact of analytics on teaching and learning in such environments.

The rest of the paper is structured as follows. Section 2 presents briefly MOOCs and their types. Section 3 discusses benefits of learning analytics coupled with big data. Section 4 presents a background overview and discusses related work in the context of MOOCs, Learning Analytics and Hadoop based platforms. Section 5 introduces the proposed system and describes a small-scale environment. Finally, section 6 concludes the paper and describes the future research directions.

II. MOOCS : AN OVERVIEW

The progress of both information technologies and the education context run in a parallel course. In particular, educational exchange means knew an exponential growth around the end of the 20th century. By the 21st century, these means became more sophisticated and innovative [10]. Basically, Internet-based learning stood in lieu of any

educational transfer means used antecedently. Lately, mobile technologies joined the learning environment virtualizing classrooms and education sources. In this virtual numeric learning environment, the responsibility of the instructor has become an administrative one, and the educative material is simply advocated based on a general interesting context. Moreover, the number of students that an instructor can successfully manage, the main instructor's capacity indicator, is no more an issue of significance. The use of sophisticated information technologies for information transfer and student activities evaluation destroys the obstacles of human competences' limitedness, and makes the concept of unlimited class sizes achievable [10].

There is no commonly accepted definition of MOOCs, even during the period that this paper was being written the Wikipedia definition of MOOCs evolved. On September 2012, Wikipedia defined a MOOC as "a course where the participants are distributed and course materials are also dispersed across the web", adding that "this is possible only if the course is open, and works significantly better if the course is large. The course is not a gathering, but rather a way of connecting distributed instructors and learners across a common topic or field of discourse" [7]. By January 2013, the definition had become: "a type of online course aimed at large-scale participation and open access via the web. MOOCs are a recent development in the area of distance education, and a progression of the kind of open education ideals suggested by open educational resources" [7].

Massive Open Online Courses were perceived by Stephen Downes, and George Siemens, as an approach to address information excess, react to students' inquiries for pertinent knowledge, integrate IT progress, and decrease education's fee [11]. The intended objectives of this suggested online educational model was to gather unlimited number of learners, course materials, and information transfer means. The proposed model would not be subject to any limitations except for technological capabilities and their related costs. While MOOCs are considered a relatively new initiative, the concept was first discussed in 2008, but wasn't really taken up to any great extent until the last couple of years. The term MOOC (Massive Open Online Course) was coined by Dave Cormier [12] and created a buzz in 2012 which has already been described as the "Year of the MOOC".

There seems to be many definitions of MOOCs; however, two key features characterize this new educational technology: open accessibility and scalability. Thus, MOOC participants do not need to be registered in a school or a university nor paying fees in order to take part of a MOOC course. Indeed, there are two types of courses offered through the MOOCs platforms: cMOOCs and xMOOCs [13]. The first type [12], described as the good MOOCs by George Siemens, who, with Stephen Downes , early put forward these courses in Canada, is essentially based on a philosophy of connectedness and sustains the social dimension of learning and active practices; thereby, this type of course encourages knowledge production rather than knowledge consumption. While xMOOCs, the most adopted by higher education worldwide [4-6], consider the instructor-guided lesson as the centre of the course and offer to large numbers of students the opportunity to study high quality courses with prestigious universities.

A MOOC system is consisted of five main elements [14]: Instructors, learners, topic, material, and context.

Instructors – Simplify the learning process via making available appropriate material, initiate communication between learners, and manage evaluations with regards to intended learning outcomes.

Learners – Anyone who wants to learn about the topic. Learners could be pursuing a formal degree or not. Learners who are simply interested with no precise objective are as well authorized to enrol.

Topic – The topic is discovered through the learner, instructor, material, and context. It is introduced all over the learning system and not just residing in a warehouse. It is adequately limited to allow emphasis but adequately wide to provide extensive coverage.

Material – Resides in diverse sites and is of multiple types and is accessed via various technological solutions.

Context – Represents the different actors forming a learning environment. This can incorporate online social networks, IT solutions, conventional information origins, diverse kinds of information transfer schemes, communication systems, intended learning outcomes, and the group constituting every course offering.

In MOOC platforms, information provided to learners is considered starting points from which they can jump off and pursue an information trajectory in accordance with their concerns. Accordingly, learners are able to communicate with one another through forums set up to help them discover common fields, find help and extra materials, and constitute particular groups so as to investigate shared topics more thoroughly. Indeed, the objective is to conceive a community of learners whereby everyone contributes by information and perspectives besides those provided by the instructor, and to get in an exploration ride. A course offered through a MOOC platform can be subject to a predefined time schedule or not, and can incorporate videos of different sources, links to websites and other online resources, some extra study materials, support forums, and all this can be accessed through multiple devices connected to the internet over wired, wireless, or cellular connections [11]. The learner chooses through which mean information is transferred may it be class forums, online social networks, or any other virtual domain. The strongest feature of a MOOC platform is elasticity [14].

III. BACKGROUNG AND RELATED WORK

A. Learning Analytics

Learning Analytics was defined in the 1st international Conference on Learning Analytics and Knowledge (LAK 2011) as "The measurement, collection, analysis and reporting of data about learners… for purposes of understanding and optimising learning and the environments in which it occurs". The main goal of LA (Learning analytics) in distance learning is primarily improving learning efficiency and learning operations effectiveness, as well as providing educators,

learners, and decision makers with actionable insight to online course level activities. Specifically, learning analytics centres on the learning process through online platforms, including the analysis of the relationship between learners, contents, and eventually instructors.

B. Big data

Big Data is defined as huge amount of unstructured information and content that can be gleaned from "infinite" activity on the internet, generally non-traditional sources, such as web logs, click streams, social media, emails, sensors, images, and videos. The ability to analyze and exploit big data offers massive opportunities for real-time intelligence about responses to products, services and even political decisions.

Thus, several business activities can benefit from opportunities that big data can engender. Common use cases include, but are not limited to: sentiment analysis, marketing campaign analysis, fraud detection, and research and development. Nowadays, big data analytics is considered as a top IT priority for most organizations. Certainly, learning analytics and big data will have a significant role to play in the future of higher education.

C. Apache Hadoop

Hadoop [9] is an open source project sponsored by the Apache Software Foundation. Inspired by Google's MapReduce [15] paradigm, it's a Java-based programming framework that supports the processing of large data sets in a distributed computing environment. Cloud based Hadoop offers the possibility of running scalable applications on systems with thousands of nodes dealing with thousands of terabytes.

Hadoop framework is actually used by major players including Facebook, Twitter, Yahoo and IBM, largely for applications involving analytics, search engines and advertising. Hadoop MapReduce is a system for parallel processing of large data sets, and a number of related projects such as Apache HBase, Hive and Zookeeper.

D. related work

Andrew NG and Daphne Koller, two Stanford University computer professors, founded Coursera [5] MOOC in 2012. Since then, it has attracted an international student community of some 3.1 million students to its 332 courses. With the array of international partners, Coursera has already begun to offer courses in Spanish, Chinese, French and Italian. MITx [16] is the first prototype designed by MIT to support MOOC courses. "Circuits and Electronics", also known as 6.002x, is the first course available on MITx, debuted in May 2012. EdX [4], successor of MITx, is a nonprofit organization put forward by Harvard and M.I.T., that allows a large community of academic institutions to take advantage of the MITx infrastructure and offers MOOC courses.

Most of the work in the Learning Analytics has focused on the LMS (Learning Management Systems), such as Morris et Al. [17] who concluded that frequency of participation and time spent on tasks are important for successful online learning through LMS. Macfadyen and Dawson [18] advocate for early-warning reporting tools that can identify and flag "at-risk" students on LMS based platforms and allow instructors to develop early intervention strategies. Yanyan et Al. [19] proposed an improved mix framework for opinion leader identification in online learning communities, the study rank opinion leaders based on four distinguishing features: expertise, novelty, influence, and activity.

In the Cloud, Amazon ElasticMapReduce [20] offers Apache Hadoop as a hosted service on the Amazon AWS (Amazon Web Services) cloud environment that provides resizable compute capacity. Tabaa and Medouri [21] proposed a novel implementation of cloud computing platform SPC (Scientific Private Cloud) which offers a highly scalable data intensive distributed computing to perform complex tasks on massive amounts of data such as clustering, matrix computation, data mining, information extraction, etc.

IV. Learning Analytics for MOOCs

Before introducing the proposed system, we would like, first, to analyze the lifecycle of the Big Data generated by MOOCs. Then, we classify student types or MOOC profiles that we shall identify while analyzing data using a simple method explained below based on "MOOC"ers behaviour.

A. Lifecycle of MOOCs' big data

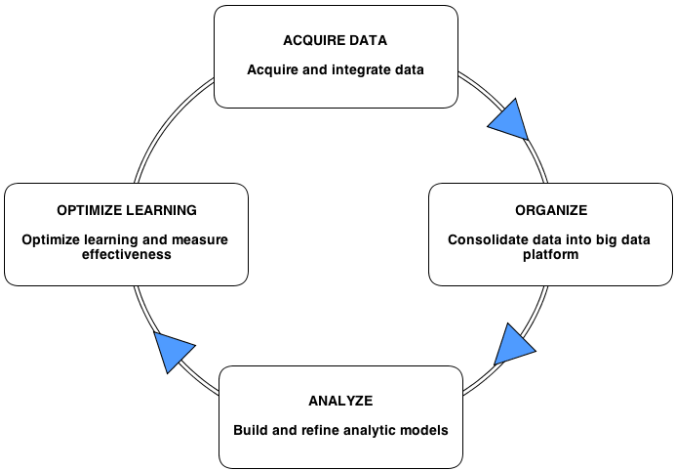

Fig. 1. Lifecycle of MOOCs' big data

As depicted in figure 1, the life cycle of big data generated by MOOCs can be described as follows:

a) Acquire: Generated data are captured periodically at source, typically as part of learning operations such as viewing materials, posting, surveys, user profiles, social media…etc.

b) Organize: Data is transferred from various sources and consolidated into a big data platform in order to prepare it for processing.

c) Analyze: Data stored in the big data platform is processed using various analysis modules, either in batches or a real-time processing.

d) Optimize Learning: The results of the "Analyze" phase are presented to MOOCs' stakeholders, enabling actions and automated interventions to be taken to provide early assistance to "at-risk" learners.

B. MOOC student patterns

Based on the Phil classification [22] of student types in a coursera-MOOC style, we redefine selected groups in the following modified classification list, that presents learners' profiles usually found in MOOC environments:

"Ghosts" – As long as a MOOC course is activated, this category of students registers to the course but at no time signs in. This category is usually the largest in terms of the number of enrolled students.

"Observers" – This category of students actually registers for the course, signs in, and might as well explore course materials. However, they do not carry out any kind of evaluations apart from basic quizzes found on lecture videos.

"Non-completers" – The majority of students fall into this category; they have recourse to MOOCs course materials to assist them study for and succeed in other courses. Essentially, those students attempt to use different course resources but do not accomplish the whole course.

"Passive Participants" – These students might consume each course material: watch lectures, complete quizzes, and interact with other learners and lecturers. Nevertheless, they do not participate in the course homework and projects.

"Active Participants" – Active participants are students who actually planned to take part of a MOOC course; they attend the lectures, accomplish the homework, interact with other participants, and complete all evaluations forms.

Following, we consider a learner as a more likely profile to dropout or "at-risk" student if he belongs to one of the following categories: "Ghosts", "Observers", "Non-completers" or "Passive Participants".

C. A method to identify "at-risk" students in MOOC environments

As depicted in Fig. 2, in order to identify "at-risk" students, we suggest a simple method based on two principal characteristics: interaction and persistence. These indicators can be measured by essentially analyzing learners' behaviour and activities, such as the number of viewed videos, downloaded lectures, and replayed quizzes and surveys.

Persistence indicates user's concentration stability on a course in temporal terms. Although, it is a complex phenomenon that results in student completion of an online course, we can measure students' persistence through an important indicator, namely the number of viewed course materials. Thus we define the persistence of a student as:

$$P_c(s) = {|VM_c(t)|}\Big/{(1-\alpha)|NM_c(t)|}$$

$|VM_c(t)|$ denotes the number of viewed materials (namely slides, lecture sequences, tutorials…etc), of a course (c) in an instant $(t)$and $|NM_c(t)|$ is the total number of materials of a course (c) released in an instant (t) . α is an adjustment factor with a value range of [0, 1]. On the other hand, interaction indicator is used to measure students' interaction on a specific course. Scoring of course-specific student is performed based on two assumptions: (1) the more assessments and surveys a user has participated in, the more interest he /she has on this course; and (2) the more correct submitted assessments, the more comforting to continue learning operations.

$$I_c(s) = {(1-\beta).(AC_c(s)+SR_c(s))}\Big/{(AP_c(s)+SP_c(s))}$$

Where β is an adjustment factor with a value range of [0, 1].

Student's interaction in a course (c) is calculated based on the number of students' participation in assessments and surveys, which is designated respectively as $AP_c(s)$ and $SP_c(s)$. $AC_c(s)$ and $SR_c(s)$ denotes respectively the number of correct submitted assessments and replayed surveys.

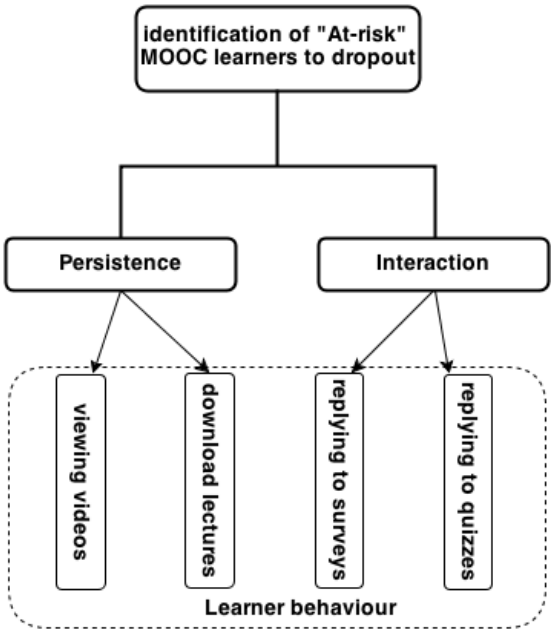

Fig. 2. Method to identify "at-risk" learners

We define $ED_c(s)$, the Engagement Degree of a learner in a MOOC course (c) as:

$$ED_c(s) = P_c(s) \times I_c(s)$$

This will result in a $ED_c(s)$ value range of [0, 1] that can be evaluated according to an adjustable threshold to identify if a learner is a potential "at-risk" profile.

V. LASyM ARCHITECTURE AND DESIGN

Fig. 3 depicts the general organization of LASyM, which envisions a learning analytics system, that enables MOOC stakeholders and providers to adjust content and provide support to their users by leveraging a private cloud based Hadoop [16], capable of processing the huge amount of captured learner-produced and learner-related data from MOOC platforms in order to minimize the time delay between the capture and use of data.

The core component of the proposed system is the analytics engine. This module of LASyM acts as a processing engine by

supplying building and deploying distributed learning analytics applications based on multiple frameworks such as MapReduce. Indeed, the main role of this component is first classifying and then processing the huge amount of hidden information in MOOC users' data.

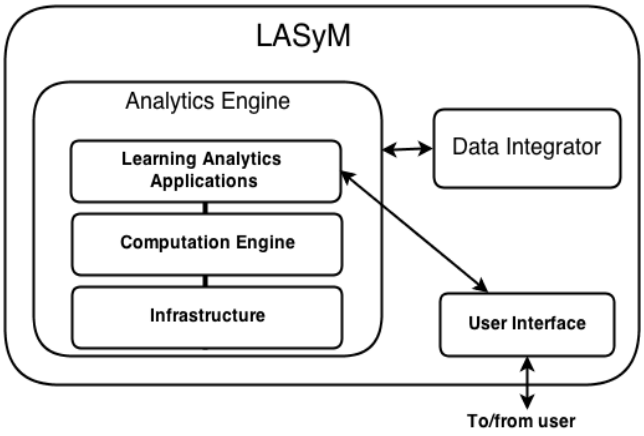

Fig. 3. LASyM architecture

The Data Integrator component is responsible of capturing data at sources before being transferred to the analytics platform. Data sources vary from students' engagement and behaviour to students' interests and preferences. Using various analysis modules, massive amounts of data will be provided and that can be mined to better optimize online learning experiences.

Finally, LASyM implements a user interface for accessing the corresponding learning analytics applications through a Web interface which also enables users to submit learning analytics jobs and explore results.

VI. LASyM : Implementation and Evaluation

In this section, we present the evaluation of our LASyM prototype. The section starts with the description of the experimental environment and infrastructure where the LASyMs' components were deployed. Then, in order to show the effectiveness of the proposed system, a small-scale scenario which implements a MapReduce-based application to identify "at-risk" learners is set up. Subsequently, evaluation of results will be presented.

A. Experimental setup

The experiments were conducted on a small scale implementation through operating a private cloud based Hadoop already deployed [21], composed of one master acting as the Resource Manager and 12 nodes, each node is a virtual machine with 2.4 GHz, 2 GB of RAM memory and 20 GB disk space allocated for HDFS (Hadoop Distributed File System), and embedding additional component, namely the data integrator and a MapReduce-based application to identify "at-risk" learners. Thus, we established a prototype of the LASyM system as shown in the Fig. 4. To execute the experiment, we used a sample of amplified dataset gathered from a typical MOOC deployed on jamiaati.org platform based on Stanford Class2go open source, itself deployed in the same private cloud. In our experiment, the parameters α and β were set at 0.2 and 0.1 respectively.

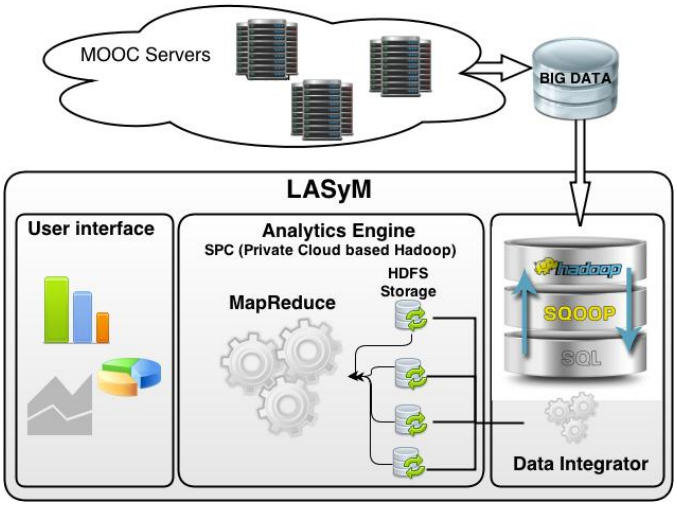

Fig. 4. LASyM architecture

The MapReduce-based application in LASyM which implements the previously developed method to identify "at-risk" learners is reduced to a single MapReduce job, meaning there is no overhead from executing multiple jobs in sequence.

B. Analytics engine

As can be seen above, the implemented LASyM analytics engine is a Hadoop-based component deployed on a private Cloud. It's a batch-oriented delivery system of analytics based on the SPC (Scientific Private Cloud) [16] previously deployed to provide researchers with a next generation of scientific platform based on the new generation of Hadoop, commonly named Hadoop 2.0. It's the main component of the LASyM, which will process Map/Reduce jobs to analyze all data acquired and organized by the data integrator component.

Indeed, there are several benefits of using such infrastructure which include the following:

- Rapid provisioning: Deploying analytics engine in the Private Cloud based Hadoop in few minutes.
- Multi-tenant frameworks: Using Cloud based Hadoop for other ends than MapReduce as a processing framework.
- High Availability: High availability with no single point of failure.
- Multi-purpose cloud infrastructure: Sharing infrastructure between Hadoop and non-Hadoop learning analytics applications.

C. Data Integrator

The Data Integrator captures data at sources, generally from relational databases, before being transferred to the analytics platform. In this experiment, the Data Integrator is responsible of extracting information from user profiles and orders stored within a relational database. Then, data is consolidated and

transferred into the analytics engine respecting the HDFS (Hadoop Distributed File System) file system. The Data Integrator implements the Apache sqoop [23] open source originally developed by cloudera [24] and currently an Apache top line project. It's a tool designed for efficiently transferring massive data between Hadoop and structured data stores such as relational databases.

Therefore, the data integrator component will collect and then import data from the MySQL database into the Hadoop Distributed File System (HDFS) [9], extracts results from processing Hadoop jobs and then exports data to MySQL tables to be able to easily explore analytics results. This component also provides the ability to schedule and automate import/export tasks.

D. Evaluation

After organizing and gathering data from the MySQL database using the Data Integrator component, an experiment was conducted in order to evaluate the performance of our LASyM implementation.

TABLE I. RUN TIMES FOR "AT-RISK" IDENTIFIER APPLICATION USING LASYM UNDER VARYING NODES NUMBER

LASyM nodes	Learners enrolled in course (c) [a]				
	10 K	*100 K*	*1M*	*2M*	*4M*
1 node	132	232	495	1578	4649
2 nodes	101	155	266	957	2760
4 nodes	88	113	234	668	1017
8 nodes	79	96	198	361	489
12 nodes	64	79	163	278	365

[a] The experiment uses a dataset collected in the 5th week of a MOOC course

Knowing that the average duration of a MOOC course is 5 weeks, we executed the developed MapReduce-based application into LASyM in different number of parallel nodes, as shown on Table 1, to be able to calculate learning analytics speedup. Learning analytics speedup measures how many times processing learning analytics through LASyM is faster than running the same MapReduce jobs on a single node. If its value is greater than 1, it means there is at least some gain from doing the work using LASyM system. A speedup equal to the number of nodes is considered ideal and means that the system has a perfect scalability.

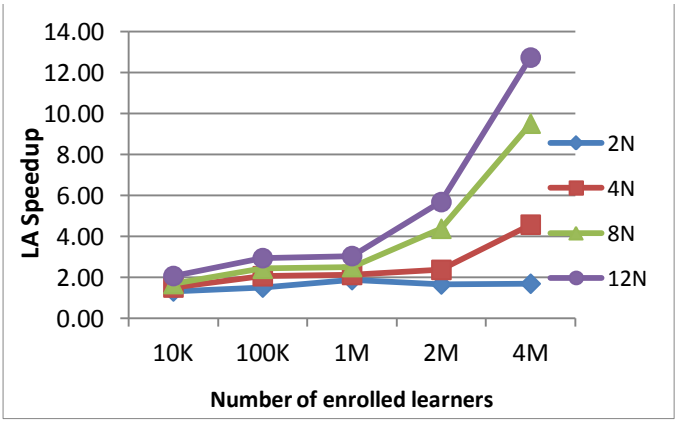

Fig. 5. Learning analytics speedup using LASyM

Fig. 5 depicts the calculated speedup. Thus, we can see that when the number of enrolled learners is small, there is only a small advantage in using LASyM: the speedup is slightly above 1 for 2 LASyM nodes and only reaches 2.06 in 12 nodes. However, when dealing with an important number of enrolled MOOC learners, the speedup grows significantly. With a number of 4 millions enrolled learners, the Learning analytics speedup using eight nodes into LASyM reaches 9.51, that can be interpreted as an ideal gain from using the proposed system to identify "at-risk" learners even if the number of enrolled learners is very high. With such system, composed of a larger number of nodes, the speedup reach the number of nodes, indicating that learning analytics process got the full benefit from using 12 nodes. The calculated results of learning analytics speedup for all cases suggest that this system has a good scalability to deal with such operations.

VII. CONCLUSION AND PERSPECTIVES

Most of the previous work related to learning analytics has focused on the learning analytics for LMS (learning Management Systems) or simply draw attention to the high dropout rate on MOOCs. However, few studies have addressed the issue of "at-risk" students' identification. As LA programs grow to include analysis of unstructured data, universities will need to develop skill and capacity to offer Hadoop based platforms and retrieval services. Indeed, the purpose of this work was, first, to design a learning analytics system capable to deal with the huge amount of unstructured data generated by a MOOC platform, as well as to develop a Hadoop MapReduce based application for automatic identification of "at-risk" students in MOOC environments.

To evaluate the performance of the proposed system, we conducted an experiment on an amplified typical MOOC dataset, where each learner is observed in terms of two features, namely interaction and persistence. We conclude that by using LASyM and our "at-risk" identification method implemented into this proposed system, we greatly reduced the latency time to analyze the huge amount of MOOCs' generated data, allowing us to identify "at-risk" learners at different stages of learning operations through the MOOC platform in a reasonable time.

In our case, we experimented LASyM for MOOC environments, which use MySQL as their DBMS, nevertheless, the system can be implemented for alternative SQL and/or NoSQL based MOOCs.

In future work we will take in consideration more indicators capable of identifying in a more precise manner "at- risk" profiles. Also, we will develop a component that enables MOOC providers to implement early intervention strategies in order to minimize the high rates of dropout in MOOC platforms.

REFERENCES

[1] R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, "Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility," Future Generation Computer Systems, vol. 25, no. 6, pp. 599-616, 6//, 2009.

[2] C. Allison, A. Miller, I. Oliver, R. Michaelson, and T. Tiropanis, "The Web in education," Computer Networks, vol. 56, no. 18, pp. 3811-3824, 12/17/, 2012.

[3] N. Sultan, “Cloud computing for education: A new dawn?,” International Journal of Information Management, vol. 30, no. 2, pp. 109-116, Apr, 2010.

[4] M. Harvard. "edX," https://www.edx.org/.

[5] "Coursera," https://www.coursera.org/.

[6] "Udacity," https://www.udacity.com/.

[7] Wikipedia. "Massive open online course," 2012-10 and 2013-01; http://en.wikipedia.org/wiki/Massive_open_online_course.

[8] K. Jordan. "MOOC completion rates," http://www.katyjordan.com/MOOCproject.html.

[9] "Apache Hadoop web page," http://hadoop.apache.org/.

[10] C. Long. "A new higher education online business model: Open and non-profit," http://blogs.reuters.com/muniland/2012/09/15/a-new-higher-education-online-business-model-open-and-non-profit/.

[11] M. Jenny, J. M. Sui Fai, and W. Roy, “Networked Learning Conference 2010.”

[12] S. D. George Siemens, and Dave Cormier. "CMOOCs initiators."

[13] J. Daniel, Making Sense of MOOCs: Musings in a Maze of Myth, Paradox and Possibility.

[14] R. Kop, “The challenges to connectivist learning on open online networks: learning experiences during a massive open online course,” The International Review of Research in Open and Distance Learning, vol. 12, no. 3, 2011.

[15] J. Dean, and S. Ghemawat, “Mapreduce: Simplified data processing on large clusters,” Communications of the Acm, vol. 51, no. 1, pp. 107-113, Jan, 2008.

[16] "MITX Home," http://www.mitx.org/.

[17] L. V. Morris, C. Finnegan, and S.-S. Wu, “Tracking student behavior, persistence, and achievement in online courses,” The Internet and Higher Education, vol. 8, no. 3, pp. 221-231, 0/3rd/, 2005.

[18] L. P. Macfadyen, and S. Dawson, “Mining LMS data to develop an “early warning system” for educators: A proof of concept,” Computers & Education, vol. 54, no. 2, pp. 588-599, 2//, 2010.

[19] Y. Li, S. Ma, Y. Zhang, R. Huang, and Kinshuk, “An improved mix framework for opinion leader identification in online learning communities,” Knowledge-Based Systems, vol. 43, no. 0, pp. 43-51, 5//, 2013.

[20] "Amazon Elastic MapReduce," http://aws.amazon.com/fr/elasticmapreduce/.

[21] Y. Tabaa, and A. Medouri, “Towards a next generation of scientific computing in the Cloud,” IJCSI International Journal of Computer Science Issues, vol. 9, no. 3, 2012.

[22] P. Hill. "Emerging Student Patterns in MOOCs," http://mfeldstein.com/emerging-student-patterns-in-moocs-a-revised-graphical-view/.

[23] "Apache Sqoop," http://sqoop.apache.org/.

[24] "Cloudera plateform," http://www.cloudera.com/content/cloudera/en/home.html.