https://doi.org/10.17163/abyaups.204.2

Towards Data-Driven Higher Education: Designing a Learning Analytics Model with Big Data Architecture

Joe Carrión Jumbo
Universidad Internacional SEK, (Ecuador)
Orcid: https://orcid.org/0000-0003-3632-5352

Rosángeles Carrión Flores
Universidad Internacional SEK, (Ecuador)
Orcid: https://orcid.org/0009-0000-3446-4207

Verónica Beltrán García
Universidad Internacional SEK, (Ecuador)
Orcid : https://orcid.org/0009-0009-8794-4816

Ruth Patricia Sánchez Montero
Universidad Internacional SEK, (Ecuador)
Orcid: https://orcid.org/0009-0002-3596-0899

Introduction

Educational institutions make significant efforts in curricular planning and pedagogical adaptation to ensure students achieve the highest level of competence in key areas such as Language and Literature and Mathematics. However, state indicators (Ineval 2022) show that a high percentage of high school students do not reach a satisfactory level. This background may impact students' performance in university education.

For universities, early identification of these factors is crucial to address them in their micro-curricular planning. A resource that can help achieve this goal is the thorough analysis of student performance through learning analytics (LA: Learning Analytics). Data-driven analysis powered by learning analytics and artificial intelligence enables educational systems to become more agile and responsive to students' immediate needs (Elhussein et al. 2024). As Buckingham (2023) notes"Learning analytics has emerged as one of the most common terms for the community seeking to understand the implications of these developments in how we analyze learning data". Mostly educa-

tional institutions have Learning Management Systems (LMS) as they are key tools for teaching and learning, and their use has increased due to digital innovation stemming from the pandemic (Wirani & Manurung, 2020).

Learning Analytics (LA) has become a viable solution to analyze performance results recorded in learning management platforms, facilitate curricular adaptation, and improve academic performance (Siemens, 2013). This article presents a work in progress on a Learning Analytics project in a higher education institution, processing historical data from 2020 to 2024 in applied sciences, which requires handling approximately 2GB of data per year. For this purpose, a Big Data architecture has been designed to extract Key Performance Indicators (KPIs) for Learning Analytics. In the first stage, the KPIs are obtained through a systematic literature review and will subsequently be selected based on a specific study of the higher education institution. KPIs can be extracted through the analysis and massive processing of LMS data. Educational institutions use LMS as the main component for recording and tracking teaching activities (Lajoie, 2013). As Ruiz (2019) notes, "Learning analytics can be applied to various aspects of human existence; in education, it cannot be conceived outside the framework of educational research".

The purpose of this work is to develop and validate a comprehensive Learning Analytics (LA) model for higher education institutions, focused on applied sciences, through a literature review of relevant indicators (KPIs), their prioritization-based selection, and the design of a managerial visualization tool. Additionally, a robust Big Data hardware and software infrastructure is required to support both the volume of data generated by the educational process and the processing demands of LA algorithms.

Materials And Methods

Literature Review: To compile and synthesize the most relevant articles related to the objectives of this work, an exhaustive systematic review was conducted using a protocol based on PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses). The synthesis considered recommendations from Moher et al. (2009) and Moraga C. & Cartes-Velásquez (2015), using the Scopus database, a 5-year date range definition, inclusion criteria, and quality assessment by a team of 5 researchers. An initial group of 166 articles was selected, with an exhaustive review of 70 articles. The search protocol defined queries for the following terms: "index, indicator, kpi" and "analytics, learning, higher, education", limited to journal articles, conference papers, conference reviews, or book chapters published in English and Spanish. The search protocol was adapted from Kitchenham (2004), resulting in the following defined search string:

Figure 1
Literature search string applied to Scopus.

```
TITLE-ABS-KEY ( ( index OR indicator OR kpi ) AND analytics AND learning
AND higher AND education ) AND ( LIMIT-TO ( DOCTYPE , "ar" ) OR LIMIT-TO (
DOCTYPE , "cp" ) OR LIMIT-TO ( DOCTYPE , "cr" ) OR LIMIT-TO ( DOCTYPE ,
"ch" ) OR LIMIT-TO ( DOCTYPE , "re" ) OR LIMIT-TO ( DOCTYPE , "er" ) OR
LIMIT-TO ( DOCTYPE , "ed" ) OR LIMIT-TO ( DOCTYPE , "dp" ) ) AND (
LIMIT-TO ( PUBSTAGE , "final" ) ) AND ( LIMIT-TO ( LANGUAGE , "English" )
OR LIMIT-TO ( LANGUAGE , "Spanish" ) )
```

The results applied in February 2025 yielded 166 documents. Subsequently, articles from the last 5 years were filtered requiring a minimum length of 5 pages and inclusion of at least 20 references. These exclusion criteria resulted in a refined list of 110 documents. After reviewing titles and abstracts to verify relevance to the current study's scope, 40 articles were discarded, leaving a final list of 70 articles for exhaustive evaluation.

Following the PRISMA framework to address the first research question, we implemented a verifiable quality validation process (Lame, 2019) through individual article assessment. This evaluation examines five key criteria: (1) Explicit statement of Objectives, (2) Clear formulation of Hypotheses, (3) Defined Research Questions, (4) Detailed Methodology description, and (5) Comprehensive Results Discussion. The scoring system applies the following rigorous scale: (0 points) Criterion not present, (1 point) Criterion minimally addressed in text, (2 points) Criterion present but requires improvement, and (5 points) Criterion fully and properly developed. This multi-tiered approach ensures systematic quality control while maintaining methodological transparency throughout our analysis.

We trained five reviewers to thoroughly read all documents and select those exceeding at least 50% of the total quality score. At the time of this publication, we completed 50% of the article evaluations (35 out of 70). Among these 35 assessed articles, 26 (74.3%) met the minimum 50% quality threshold

Systematic Review Synthesis: The presented synthesis, following the typology defined by Okoli (2015) constitutes an independent literature review. The selected articles were analyzed to identify Learning Analytics (LA) indicators applied in higher education.

One key finding is LA's ability to personalize learning by adapting to students' individual needs to enhance academic performance. Through the analysis of data from digital platforms - particularly Learning Management Systems (LMS) - LA can identify behavioral patterns, engagement levels, and interaction metrics that enable both performance prediction and early detection of at-risk students (those likely to underperform or drop out). This capability supports timely, targeted interven-

tions, proving particularly valuable in distance or hybrid education contexts.

The reviewed literature emphasizes the use of advanced techniques, such as data mining, machine learning (ML), and predictive analytics, as inputs for curricular planning, with student performance serving as another key data input. Learning Analytics (LA) analyzes not only academic performance but also aspects such as online collaboration, engagement in digital activities, and self-regulated learning management. For example, alert dashboards and predictive models based on logistic regression or neural networks enable educators to make more informed decisions and dynamically adapt pedagogical strategies. The ongoing systematic review aims to build a theory of relevant KPIs to address the first research question, incorporating recommendations from Jaccard & Jacoby (2019).

Articles that specifically define KIP are shown in Table 1.

Table 1
Articles that define KPIs

ID	Year	Title
1	2020	Implementing a Machine Learning Approach to Predicting Students Academic Outcomes
2	2021	Are We Living in LA (P)LA Land? Reporting on the Practice of 30 STEM Tutors in Their Use of a Learning Analytics Implementation at The Open University
3	2021	Multiclass Prediction Model for Student Grade Prediction Using Machine Learning
4	2022	A Novel Multi-Dimensional Analysis Approach to Teaching and Learning Analytics in Higher Education
5	2022	Learning Analytics in Open and Distance Higher Education: The Case of the Open University UK
6	2022	Towards a Learning Analytics Model in Higher Education Applied to a LMS in Virtual Education Based on Motivation Indicators for Permanence
7	2023	Data-Driven Analysis of Student Engagement in Time-Limited Computer Laboratories
8	2023	Learning Analytics for Data-Driven Decision Making: Enhancing Instructional Personalization and Student Engagement in Online Higher Education
9	2023	Predicting At-Risk Students' Performance Based on LMS Activity using Deep Learning
10	2023	Prediction of Dilatory Behavior in eLearning: A Comparison of Multiple Machine Learning Models
11	2023	Process Indicators for Grading Group Essays: Learning Analytics of Assessment Data and Online Behaviour
12	2023	Proposal of Indicators for Measuring Collaborative Writing in a Digital Learning Environment
13	2023	Surviving and thriving: How changes in teaching modalities influenced student satisfaction before, during and after COVID-19
14	2024	Academic engagement and management of personalised active learning in higher education digital ecosystems
15	2024	The Mediating Role of Learning Analytics: Insights into Student Approaches to Learning and Academic Achievement in Latin America

Indicators for Learning Analytics in Higher Education

The LA indicators described in the systematic review encompass a wide range of metrics to evaluate student performance, engagement, and satisfaction in educational environments, with key indicators including:

Satisfaction and perception

SBJ (course satisfaction), TEA (teaching satisfaction), PLN (content and competencies), MTD (teaching methodologies), WLD (workload), and OBL (teaching compliance), which measure students' subjective experience (Ortiz-Beltrán et al. 2023).

Participation and collaboration

Balance of Contribution (contribution balance), Co-writing (cowriting), and collaborative writing strategies, along with connectivity, acquisition, productivity, and interactivity metrics. Activity and engagement: (sessions, duration, active days), Course Accesses (course accesses), Interactions (interactions), and Average Minutes (average time), measuring participation frequency and quality (Haddouche et al. 2023: Salas-Martinez et al. 2024; and Villalobos et al. 2024).

Academic performance

Final Exam Grade (final grade), Quizzes (quizzes), Short Tasks (short assignments), Extensive Assignment (extensive assignments), and Attendance (attendance), along with predictive metrics such as precision, recall, and F-measure (Alé-Ruiz et al. 2024).

Demographic data and progress

Applicant profile, admission, enrollment, demographics, academic progress, and graduate employment, which help contextualize performance (Li et al. 2022).

These indicators enable comprehensive learning assessment, facilitating identification of at-risk students, teaching personalization, and curricular planning improvement.

KPI Analytics process

The KPI Analytics process follows the CRISP-DM (Cross Industry Standard Process for Data Mining) framework (Wirth & Hipp, 2000), a widely adopted methodological standard in data mining and advanced analytics, which enables systematic LA implementation through interconnected phases; this study adapts CRISP-DM for educational data analysis from an LMS, focusing on three core phases: (1) Business and Domain Understanding, (2) Data Preparation, and (3) Modeling and Evaluation.

Results

Relevant LA Indicators

The KPIs identified in the literature review have been grouped into 3 criteria

- Criterion 1: Educational Quality
- Criterion 2: Engagement
- Criterion 3: Academic Risk

The following details selected KPIs from the literature review relevant to higher education, presenting 24 indicators divided into 6 groups:

Table 2
List of KPIs by domain

ID	Domain	Nro.	KPI	Article
1	Student Satisfaction and Perception	1	Overall course satisfaction	13
		2	Overall teaching satisfaction	
		3	Teaching plan contents and competencies	
		4	Teaching and learning methodologies	
		5	Perceived workload	
		6	Teacher obligation compliance	
2	Collaboration and Learning Strategies	7	Contribution balance	12
		8	Co-writing	
		9	Collaborative writing strategies (summative and integrative)	c
3	Study Time and Behavior	10	Time dedicated to schedule planning	6
		11	Time dedicated to reviewing additional information	
		12	Delegation activity intensity	
		13	Behavior strategies	
		14	Outlier detection (unusual sessions)	
4	Platform Activity and Participation	15	Sessions: duration, actions, active days/weeks	15
		16	Activity proportion by activity type	
		17	Entropy and action distribution	
		18	Daily/weekly action entropy	
5	Content Access and Interaction	19	Unique users/courses	8
		20	Course/item/tool/content access	
		21	Average time and resource usage	
6	Academic Performance	22	Final grade	14
		23	Overall course satisfaction	
		24	Overall teaching satisfaction	

The 10 domains are related to the 3 criteria, described in Table 3, where an X marks the relationship between each domain and criterion

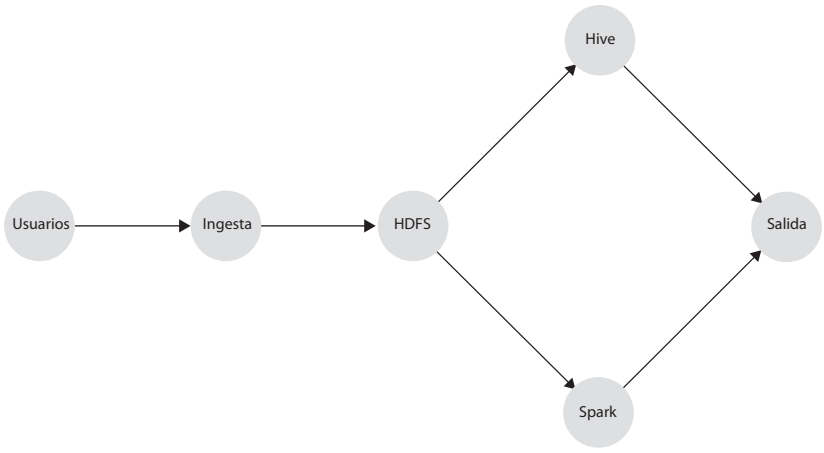
Table 3
List of KPIs according to the basic criteria

ID	Domain	Education Quality	Engagement	Academic Risk
1	Student Satisfaction and Perception	X		
2	Collaboration and Learning Strategies	X		
3	Study Time and Behavior		X	X
4	Platform Activity and Participation		X	X
5	Content Access and Interaction		X	
6	Academic Performance	X		X

Big data Architecture

The implemented Data Lake provides a scalable and flexible infrastructure for storing and processing large volumes of data for Learning Analytics in Higher Education, an approach to using a data lake for higher education analysis that has also been examined by Rahim et al. (2024) and Adegbenjo et al. (2023). Its design is based on using Hadoop as a distributed file system (HDFS) and Hive for data querying, enabling integration with analytics tools like Spark. The following details the implemented system architecture, including hardware, software, and data flow.

Fig. 1
Big data Architecture



The Data Lake environment was implemented on a virtual machine configured to ensure adequate performance for data extraction, storage, and query processes, with the following specifications:

The virtualization platform used was Proxmox Virtual Environment with 24GB RAM, 4 processors (2 sockets, 2 cores), and 300 GB allocated storage and Ubuntu operative system.

Data Lake Software and Technologies

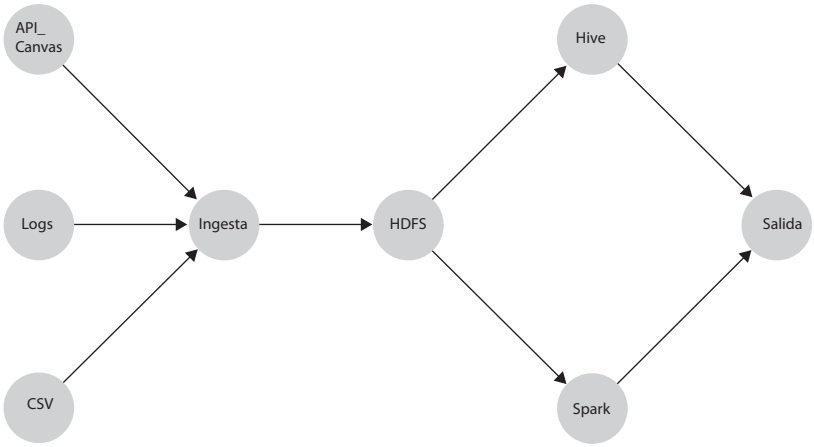
The Data Lake was built using an advanced Big Data technology ecosystem including:

- Apache Hadoop as the framework for large-scale data processing managing distributed storage through HDFS.
- HDFS as Hadoop's distributed file system.
- Apache Hive as the database engine enabling data reading/writing/management via HQL on HDFS.
- Python for object-oriented programming and automation script development.
- Spark for in-memory distributed processing.
- External APIs for Canvas LMS data extraction.

Data Lake Workflow

Data Lake implements an agile, optimal, and precise data flow that achieved the project's objectives through these stages: data ingestion extracting information from the Canvas LMS API, storing it in CSV format while generating system logs; data storage preserving raw information in the implemented HDFS directory; processing utilizing tools like Hive and Spark to generate key insights for Higher Education Learning Analytics; and querying/analysis employing Python, Pandas, and advanced data visualization tools for pattern discovery and decision-making in Higher Education.

Fig. 2
Data Lake Workflow.



Discussions

Indicator Relevance

Student Satisfaction and Perception: Evaluates aspects such as humanism and education quality, measuring student satisfaction and applied methodologies while contributing to comprehensive education through content assessment and teaching ethics.

Collaboration and Learning Strategies KPIs

Assess collaborative learning, promoting educational innovation through balanced individual work and co-writing with summative and integrative evaluation criteria.

Study Time and Behavior

Evaluates responsibility in assignment submission, fostering professional development through autonomous work and student self-management.

Platform Activity and Participation along with Content Access and Interaction

Measures student engagement and autonomy development, assessing pedagogical effectiveness and scientific production, where participation pattern analysis strengthens educational quality when examining changes across academic periods.

Academic Performance

Evaluates educational quality, verifies competency achievement, identifies improvement needs, enhances teaching-learning processes by knowledge gap detection, and facilitates timely remediation through support programs, student retention, and graduation rates.

BigData Architecture

The implementation of the Hadoop-based (HDFS) Data Lake with Hive and Spark provides a robust technical solution for processing educational KPIs in LA, with the proposed architecture supporting the data volume (2GB/year) and variety (LMS, historical, logs) while enabling efficient extraction, storage, and querying of key indicators, where scalability is expected to handle complex Canvas LMS datasets through distributed processing with Spark optimizing metric calculations like ‘con-tent access frequency’ or ‘average interaction time’ required for student engagement KPIs.

A critical contribution is the HDFS storage (with folders like /LMS and /DataHist) enabling historical data traceability for comparative analysis (e.g. cohort performance), though initial CSV data ingestion required manual preprocessing, suggesting future integration of automated ETL pipelines, while the Hive query layer allows synthesizing complex KPIs (e.g. ‘dropout risk’ based on activity + performance).

Conclusions

The literature analysis identified 24 key KPIs for LA in higher education, grouped into educational quality, engagement, and academic risk. These indicators are validated by recent scientific evidence, enabling student performance monitoring and data-driven pedagogical interventions for institutional decision-making.

The implementation of Data Lake with Hadoop, Hive, and Spark is a scalable solution for processing massive educational data (>2GB/year), with this fault-tolerant architecture supporting LA algorithms and managerial visualization to ensure availability and efficiency in resource-constrained university environments.

Future work aims to develop an interactive KPI visualization service (for educational quality, engagement, and academic risk) through real-time dashboards, allowing educational managers to monitor trends and execute timely data-based interventions.

Acknowledgments

The research received internal funding from Universidad Internacional SEK through the P062425 (Análisis del Aprendizaje de Ciencias Aplicadas durante el periodo 2020 al 2024) research project.

References

Adegbenjo, A. A., Onuiri, E. E., Kalesanwo, O. B., Agbaje, M. O., Abel, S. B., Fatade, O. B., Amusa, A. I., Umeaka, K. C., Ehioghae, E., & Onamade, K. O. (2023). Design and analysis of an automated IoT system for data flow optimization in higher education institutions. Journal Européen des Systèmes Automatisés, 56(5), 889–897. https://doi.org/10.18280/jesa.560520

Alé-Ruiz, R., Martínez-Abad, F., & del Moral-Marcos, M. T. (2024). Academic engagement and management of personalised active learning in higher education digital ecosystems. *Education and Information Technologies*, *29*(10), 12289–12304. https://doi.org/10.1007/s10639-023-12358-4

Buckingham, S. S. (2012). *Learning analytics* (UNESCO Policy Brief). UNESCO.

Elhussein, G., Hasselaar, E., & Lutsyshyn, O. (2024, abril). Shaping the future of learning: The role of AI in Education 4.0. World Economic Forum. https://www.weforum.org/publications/shaping-the-future-of-learning-the-role-of-ai-in-education-4-0/

Haddouche, A. M., Djelil, F., Mandran, N., Hoffmann, C., & d'Ham, C. (2023). Proposal of indicators for measuring collaborative writing in a digital learning environment. En *Proceedings of the 15th International Conference on Computer Supported Education (CSEDU 2023)* (pp. 495–502). Science and Technology Publications. https://doi.org/10.5220/0011757700003470

Jaccard, J., & Jacoby, J. (2019). *Theory construction and model-building skills: A practical guide for social scientists* (2nd ed.). Guilford Publications.

Kitchenham, B. (2004). *Procedures for performing systematic reviews* (Keele University Technical Report TR/SE-0401). Keele University.

Lajoie, S. P. (2013). Computer environments as cognitive tools for enhancing learning. En S. P. Lajoie & S. J. Derry (Eds.), *Computers as cognitive tools* (pp. 261–288). Routledge.

Lame, G. (2019). Systematic literature reviews: An introduction. En *Proceedings of the 22nd International Conference on Engineering Design (ICED19)* (pp. 1633–1642). Cambridge University Press. https://doi.org/10.1017/dsi.2019.169

Li, Q., Duffy, P., & Zhang, Z. (2022). A novel multi-dimensional analysis approach to teaching and learning analytics in higher education. *Systems*, *10*(4), Artículo 96. https://doi.org/10.3390/systems10040096

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Annals of Internal Medicine*, *151*(4), 264–269. https://doi.org/10.7326/0003-4819-151-4-200908180-00135

Moraga C., J., & Cartes Velásquez, R. (2015). Pautas de chequeo, parte II: QUOROM y PRISMA. *Revista Chilena de Cirugía*, *67*(3), 325–330. https://doi.org/10.4067/s0718-40262015000300015

Okoli, C. (2015). A guide to conducting a standalone systematic literature review. *Communications of the Association for Information Systems*, *37*, 879–910. https://doi.org/10.17705/1cais.03743

Ortiz Beltrán, A., Hernández-Leo, D., & Amarasinghe, I. (2023). Surviving and thriving: How changes in teaching modalities influenced student satisfaction before, during and after COVID-19. *Australasian Journal of Educational Technology*, *39*(6), 72–88. https://doi.org/10.14742/ajet.8958

Rahim, S. A. A., Sidi, F., Affendey, L. S., Ishak, I., & Nurlankyzy, A. Y. (2024). Leveraging data lake architecture for predicting academic student performance. *International Journal on Advanced Science, Engineering and Information Technology*, *14*(6), 2121–2129. https://doi.org/10.18517/ijaseit.14.6.12408

Ruiz, I. I. B. (2019). Las analíticas de aprendizaje como cultura digital en las universidades: diagnóstico de su aplicación en el sistema de educación a distancia de la UNAM a partir de una escala compleja. *Revista Iberoamericana de Educación*, *80*(1), 89–116.

Salas Martinez, A., Ramirez Martinell, A., & Martinez Ramos, S. (2024). Towards a learning analytics model in higher education applied to a LMS in virtual education based on motivation indicators for permanence. En *2024 12th International Conference in Software Engineering Research and Innovation (CONISOFT)* (pp. 349–355). IEEE. https://doi.org/10.1109/CONISOFT63288.2024.00052

Siemens, G. (2013). Learning analytics: The emergence of a discipline. *American Behavioral Scientist*, *57*(10), 1380–1400. https://doi.org/10.1177/0002764213498851

Villalobos, E., Hilliger, I., Gonzalez, C., Celis, S., Pérez Sanagustín, M., & Broisin, J. (2024). The mediating role of learning analytics: Insights into student approaches to learning and academic achievement in Latin America. *Journal of Learning Analytics*, *11*(1), 6–20. https://doi.org/10.18608/jla.2024.8149

Wirani, N., & Manurung, A. A. (2020). The importance of using a web-based learning model to prevent the spread of COVID-19. *AlAdzkiya International of Education and Social (AIoES) Journal*, *1*(1), 16–24.

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. En *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining* (pp. 29–39). Springer.