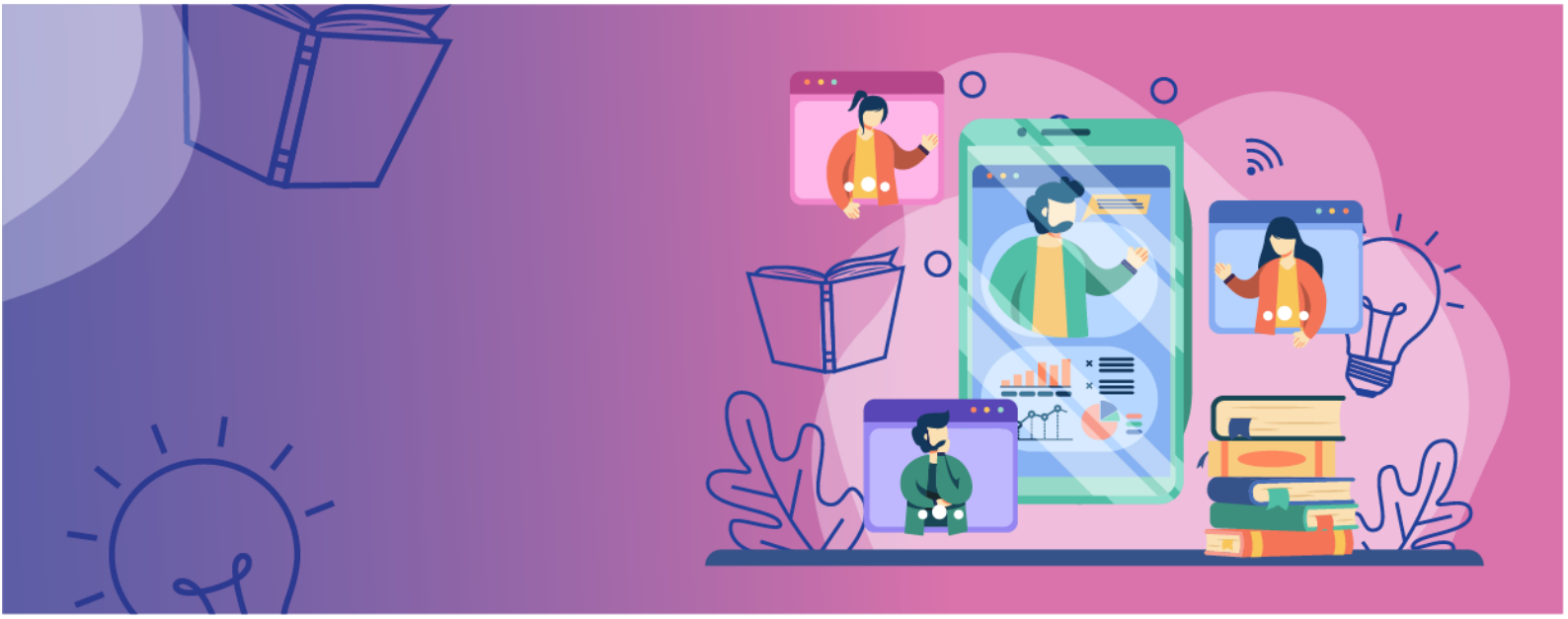
DIGITAL FIRST

DIGITAL FIRST

Digital Tech as the First Language:

Informatics for Digital Natives

D7.2 Preconditions for data modelling in schools for learning analytics for informatics (data storage, data structure, data harmonisation)



Co-funded by
the European Union

Project 101132761

D7.2 Preconditions for data modelling in schools for learning analytics for informatics (data storage, data structure, data harmonization)

Partner:	UTU
Work Package:	WP7: Data for success: Modelling learning analytics for informatics
Task:	T7.2 Defining the necessary classroom architecture for meaningful educational data mining T7.3 Defining the data structure and the necessary data points for educational data mining for informatics T7.4 Identifying data storage and data harmonization for educational data mining for informatics
Due date:	27/02/2026
Work package Leader:	UTU

Start date of project: 1 December 2023 **Duration:** 36 months

DOCUMENT HISTORY		
Version	Date	Changes
1.1	21/02/2026	

DISSEMINATION LEVEL		
PU	Public, fully open	X

Co-funded by the European Union

SEN	Sensitive (limited under the GA conditions)	
CLASS	EU classified, confidential	

MAIN AUTHORS	
Name	**Organization**
Daranee Lehtonen	UTU
Valentina Dagienė	VU

QUALITY REVIEWERS	
Name (Quality Reviewer 1, Quality Reviewer 2)	**Organization**
Maja Šarić, Maja Brkljačić	Algebra University

Co-funded by the European Union

PREFACE

This guidebook is Part 1 of a three-part series on introducing Learning Analytics (LA) in informatics education. It aims to assist educators, administrators, and decision-makers in leveraging LA to enhance teaching, learning, and educational planning. Whether you are considering LA deployment or already underway, the guide provides essential insights and practical tools. While focused on informatics, much of the content is applicable to other educational contexts.

Developed under the Digital First Project, this guidebook results from collaboration between the Turku Research Institute for Learning Analytics (TRILA), University of Turku, Finland; Vilnius University; and other project partners. Content draws on a literature review, TRILA expertise, and project activities across 10 EU countries (Bulgaria, Croatia, Cyprus, Finland, Greece, Italy, Lithuania, Portugal, Slovenia, Spain), including a grey literature review of 35 policy papers, strategy documents, and project reports; 31 interviews with teachers, researchers, policymakers, and education specialists involved in LA and informatics education; and expert group discussions. For more information on the grey literature review and interviews, see D7.1 Theory and Practice of Learning Analytics for Informatics: Research and Results (https://digitalfirstnetwork.eu/deliverables/).

Guidebook Part 1 comprises four chapters:

1. Introduction to Learning Analytics and Data Mining
2. Classroom Architecture for Meaningful Educational Data Mining
3. Data Structures and Necessary Data Points
4. Data Storage and Data Harmonization

Our deepest gratitude goes to all contributors, particularly Prof. Jari Metsämuuronen, for guidance, and to the Algebra University team for reviewing and producing the manuscript.

Co-funded by the European Union

TABLE OF CONTENTS

Co-funded by the European Union

LIST OF TABLES

LIST OF FIGURES

Co-funded by the European Union

1 Introduction to Learning Analytics and Data Mining

As educational systems increasingly integrate digital technologies, such as digital devices and digital learning platforms, they generate vast volumes of data about learners, learning processes, and learning environments. This transition coincides with the evolution of the educational landscape from Web 1.0 to Web 2.0, where interaction, collaboration, and data generation are more prominent. Learning Analytics (LA) and Educational Data Mining (EDM) have become essential for making sense of this data, playing a pivotal role in this evolution. The shift toward a data-driven education supports the broader application of analytics in educational systems, enabling the tracking of student interactions, engagement patterns, and success rates (Zhang et al., 2021).

With the rise of big data, the opportunities for educational analytics to monitor and enhance teaching and learning are immense. By leveraging vast streams of information, LA and EDM facilitate personalized learning journeys, equipping educators to provide tailored support in real-time based on data-driven insights as well as informing instructional design. However, the increasing complexity of educational data also presents challenges, requiring robust tools and strategies to interpret and apply findings effectively.

This chapter introduces the definitions and core concepts of LA and EDM, clarifies their relationship with traditional assessment practices, and situates them within the K–12 school context. It aims to provide general audience, teachers, schools, policymakers, and researchers with a conceptual foundation for understanding how data can be meaningfully collected, analyzed, and applied to enhance learning outcomes and instructional decision-making.

1.1 What is Learning Analytics?

Learning Analytics (LA) is formally defined as the measurement, collection, analysis, and reporting of data about learners and their contexts, with the goal of understanding and optimizing both learning and environments in which it occurs (e.g., Ferguson & Buckingham Shum, 2011; Shum & Ferguson, 2012; Nistor & Hernández-García, 2018). LA relies on digital traces; therefore, learning analytics can describe only those aspects of learning processes that leave data within various digital applications or environments. At its core, LA transforms raw educational data into actionable knowledge, allowing educators to make evidence-based decisions that enhance teaching effectiveness and support student learning.

LA aims to support and enhance teaching and learning by monitoring learning and providing actionable insights to inform decision-making. For learners, LA offers opportunities for formative, immediate feedback and tasks customized to their skill level and needs. For teachers, LA empowers teachers to monitor student participation, pinpoint challenges and misconceptions, identify signs of early disengagement, and adjust instruction to better meet individual learner needs. By leveraging insights from LA, teachers can enhance their responsiveness to students' unique challenges and adapt teaching strategies in real time. At the institutional level, LA serves broader objectives, such as supporting curriculum design, driving professional development initiatives, and maintaining quality assurance processes. Its application helps educational institutions align their strategies with evolving learner needs while fostering continuous improvement. Thanks to its versatile data collection and analysis methods, LA extends beyond practical application in classrooms and can also serve as a powerful tool in scientific research. By analyzing large datasets, researchers can gain deeper insights into learning behaviors, educational outcomes, and the effectiveness of instructional strategies, further advancing the field of education.

LA should be viewed as a socio-technical practice, integrating educational research, data science, and pedagogy. While relying on sophisticated systems and tools and utilizing methodologies such as statistics, machine learning, and data visualization, LA remains firmly grounded in learning theories and instructional design principles. While data

collection and analysis are essential components, the true value of LA lies in transforming data into actionable insights—predicting student behaviors, identifying learning outcomes, informing interventions, and enhancing overall educational quality (Yu, 2025; Awad et al., 2024).

Its effectiveness depends not only on the technical processes but also on the interpretation of data, the communication of insights to educators, and their thoughtful integration into everyday classroom practices and decision-making.

LA can be broadly categorized into four types, each serving distinct purposes in understanding and improving educational experience (e.g., OKM, 2024):

- **Descriptive LA**: Focuses on summarizing what is happening within the learning process and outcomes. By analyzing data from past and present activities, it describes learner behavior and performance, for example, as a histogram, a pile chart, or as statistical indicators. Descriptive LA helps observe, for example, the number of completed tasks, the time spent on tasks, the perceived difficulty of tasks, changes in grades, as well as the completion rates of courses or degree programs.
- **Diagnostic LA**: Explores the underlying reasons behind certain learning behaviors or outcomes derived from descriptive LA, such as why some students struggle with particular concepts or why engagement levels drop. Diagnostic LA, for example, seeks to identify trends in the data, generate profiles from data streams, and highlight anomalies.
- **Predictive LA**: Utilizes predictive modelling and machine learning techniques to anticipate the future development of learning and education. Predictive LA provides a basis for anticipating probable future events, such as forecasting classroom occupancy rates, identifying which students may be at risk of academic failure or dropping out, estimating the number of completed theses, predicting academic performance and future achievement, and anticipating the level of task difficulty.
- **Prescriptive LA**: Building on predictive LA, Prescriptive LA offers actionable recommendations aligned with established goals to enhance learning experiences and outcomes. It can suggest optimal classrooms based on forecasted occupancy rates; guide resource allocation based on predicted performance or anticipated task difficulty, and recommend learning materials and tasks based on projected future learning outcomes. Prescriptive LA is often integrated into automated systems, enabling it to remind students, notify teachers, and tailor learning activities based on individual needs.

Advancements in artificial intelligence (AI) have significantly enhanced the capabilities of LA. AI-driven approaches can analyze massive datasets, allowing teachers and institutions to gain deeper insights into student performance and engagement. Studies have shown that LA, when effectively applied, correlates with improved learner achievements (Awad et al., 2024). Ethical considerations are equally critical. Addressing issues such as data privacy, transparency, and fairness is vital to ensure the responsible use of analytics while fostering stakeholder trust and accountability.

1.2 What is Educational Data Mining?

Educational Data Mining (EDM) is an emerging interdisciplinary field closely tied to LA and often serves as its methodological foundation. While LA emphasizes interpreting data to inform pedagogical decisions, EDM focuses on developing and applying computational techniques to identify patterns within educational data. These fields overlap significantly and are commonly used in complementary ways to enhance educational practices.

EDM analyses the vast and complex datasets generated by educational institutions to uncover meaningful patterns and trends. By doing so, it supports informed decision-making for educators and policymakers, ultimately enhancing student outcomes and teaching strategies (Özdağoğlu et al., 2018; Triayudi et al., 2024). Examples of such data include student demographics, academic performance, behavioral activities, and interactions logged within, for example, Learning Management Systems (LMS), digital textbooks, classroom technologies, and assessment tools (Ampadu, 2023; Penteado et al., 2018; Wanjau et al., 2016).

The significance of EDM extends beyond mere performance analytics; it plays a crucial role in fostering data-driven decision-making in educational institutions. At its core, EDM aims to uncover hidden relationships and trends within datasets, facilitating a deeper understanding of the factors that influence learning. This insight enables data-driven approaches that positively impact teaching methods, learning, as well as institutional practices and policies.

EDM involves a systematic process aimed at uncovering actionable insights from educational datasets. By following key steps, institutions can leverage EDM to enhance decision-making and improve learning experiences:

1. Problem Definition: The first step in EDM is defining specific educational questions or problems to be addressed. These might include predicting student dropouts, identifying performance trends, or understanding the factors affecting learning outcomes (Arifin et al., 2022; Pal, 2012; Ruby & David, 2017). A clear problem definition ensures that the analysis remains focused and aligned with institutional goals.

2. Data Collection: After defining the problem, the next step is gathering data from various sources within educational institutions. Common data sources include log files and interaction traces from Learning Management Systems (LMS), such as student demographics, academic records, attendance, and engagement metrics (KumarYadav & Pal, 2012). Assessment results, forum discussions, and multimodal data, including sensor inputs, video recordings, and audio collected in classrooms, provide additional insights into student performance and behavior. These datasets form a solid foundation for meaningful analysis.

3. Data Preparation: Before analysis, the collected data must undergo cleaning and preprocessing to ensure its validity and reliability, particularly in heterogeneous educational environments. This step includes handling missing values, removing duplicates, and transforming variables for consistency (Mukherjee, 2022; Sharma, 2020).

4. Data Analysis: This step involves applying data mining techniques to uncover meaningful patterns and trends within the data. Common techniques include clustering (grouping data points that share similar characteristics, such as to identify groups of learners with similar learning behaviors), classification (using algorithms like Decision Trees and Naive Bayes to predict outcomes, such as student performance or dropout risks based on historical data), and association rule mining (identifying relationships among variables) (Ashaduzzaman et al., 2018; Meghji et al., 2019; Wanjau et al., 2016).

5. Interpretation of Results: Insights derived from pattern discovery are interpreted to produce actionable educational recommendations, for example, developing intervention strategies for struggling students or identifying areas where the curriculum or teaching methods can be improved (Qu et al., 2019; Saneifar & Abadeh, 2015).

6. Deployment: The final step is integrating the insights into real-world educational settings. This involves deploying strategies such as targeted interventions, adaptive learning systems, or changes in teaching methods. Continuous monitoring and evaluation ensure the effectiveness of the implementations and guide refinements over time (Arifin et al., 2022; González-Brambila & González, 2016).

While EDM offers significant benefits, its implementation comes with challenges. Ensuring data quality is a critical concern, as unreliable or incomplete data can compromise the accuracy of insights. Privacy and security issues

surrounding student data must also be addressed to maintain compliance with regulations and protect sensitive information. Another challenge is integrating EDM findings into educational policies and practices effectively to drive meaningful change (Arifin et al., 2022; Mukherjee, 2022; Pal, 2012).

Moreover, interpreting and utilizing the insights generated from data analysis requires collaboration among diverse stakeholders, including educators, administrators, and data scientists. This collaboration is essential to ensure the alignment of EDM applications with educational goals, resources, and institutional needs (Ashaduzzaman et al., 2018; KumarYadav & Pal, 2012).

1.3 Distinction Between Learning Analytics and Traditional Educational Assessments

Advancements in technology and pedagogical methodologies have transformed educational systems, resulting in clear distinctions between traditional educational assessments and LA. Understanding these differences is essential for improving educational outcomes and addressing diverse student needs.

Traditional educational assessments typically rely on standardized tests and teachers' observational evaluations to measure student achievement. While these approaches have played a central role in education systems, they are often limited in their ability to capture the complexity of learning processes. Traditional assessments typically focus on outcomes rather than processes, providing snapshots of performance at specific points in time. Common formats include multiple-choice tests, essays, and practical examinations.

LA differs fundamentally from traditional assessment in several ways. First, it emphasizes process-oriented data. Instead of only measuring what students know at the end of a unit, learning analytics captures how they arrive at that knowledge, including patterns of engagement, sequences of actions, time on task, and iterative problem-solving strategies. This allows teachers to develop a more nuanced understanding of learning as it unfolds (Raković et al., 2023).

Second, LA enables timelier feedback. Traditional assessments often provide delayed feedback, limiting opportunities for immediate instructional adjustment. In contrast, analytics-driven systems can offer real-time or near-real-time insights to support formative assessment practices and to enable teachers to intervene when difficulties arise (Khajuria et al., 2025; Janardhana et al., 2025).

Third, LA supports a more holistic view of learning. Beyond academic performance, it can incorporate indicators of engagement, motivation, collaboration, persistence, and self-regulation. This aligns with contemporary educational goals that emphasize learner agency and continuous development rather than solely summative evaluation (Knight, 2020; Russell & Wallis, 2019).

Fourth, by incorporating predictive analytics, LA can anticipate potential academic difficulties. This proactive approach allows educators to implement interventions before students fall behind, which is a significant advancement over traditional methods (Kumar et al., 2025).

Finally, LA increasingly integrates AI and big data techniques, allowing for predictive modeling and personalized learning pathways. These capabilities mark a significant shift from static, one-size-fits-all assessment approaches toward adaptive and personalized educational support tailored to individual student needs (Silva et al., 2024).

The key distinction between LA and traditional educational assessments lies in their approach to data and feedback (Kanth et al., 2018). While traditional assessments offer standardized evaluation methods with limited real-time adaptation, LA emphasize a dynamic, data-driven approach that provides timely, personalized feedback and

predictive insights. This evolution not only enhances educational quality but also aligns with modern educational goals of personalizing learning and fostering continuous development.

1.4 Learning Analytics in K–12 Education

Global Context

Interest in LA has grown significantly in the past decade, with over 1,000 articles published between 2014 and 2024 (Kılıç & İzmirli, 2024). While initially focused on higher education and corporate training, LA research has expanded into primary and secondary schools (K–12), reflecting a growing emphasis on data-informed education. Despite this progress, the implementation of LA in K–12 settings remains limited and inconsistent compared to higher education contexts. Greater teacher engagement, alignment with curriculum standards and pedagogical goals, and adaptation of LA to school-specific contexts are essential to maximize its potential (Kılıç & İzmirli, 2024; Sousa et al., 2021; Zamecnik et al., 2022).

Research has shown that LA has significant potential to enhance engagement, personalize learning, and support decision-making in schools. Dashboards, for example, can track student progress, provide real-time feedback, and increase awareness of learning processes, enabling targeted actions (Valtonen et al., 2025). However, its adoption in school settings remains underexplored (Sousa et al., 2021). Cross-national studies (e.g., Finland, China, South Africa, Uruguay, USA) reveal that interpreting and applying analytics insights is more complex than the process of data collection, hindered by cultural, technological, and pedagogical diversity in school systems, which complicates large-scale implementation (Aguerrebere et al., 2022). Furthermore, practical challenges such as time constraints limit teachers' ability to use formative assessment tools effectively (Karademir et al., 2024).

European Context

Our review of documents published across 10 EU countries offers an overview of how LA is understood and applied within the EU educational context. The review reveals that the overarching goal of LA implementation is to leverage data-driven approaches to optimize teaching, learning, and educational management tailored to the diverse needs of K–12 schools across Europe. In the partner countries, the primary purpose of using LA is to support teachers in their pedagogical practice. Key objectives include facilitating formative assessment, identifying at-risk students, enabling personalized and timely instruction, enhancing course design, and understanding student performance. These goals aim to empower teachers to implement data-driven strategies that effectively address the needs of individual learners.

LA also serves to benefit students directly by improving learning experiences, fostering personalized learning, enabling timely interventions, and enhancing learning outcomes. Furthermore, it encourages self-regulated learning and student reflection through personalized feedback and recommendations, which help strengthen motivation and sustain academic progress.

At the institutional level, LA provides actionable data to support decision-making, such as preventing school dropouts, aligning resources with strategic plans, informing curriculum development, and enhancing institutional quality processes. By delivering insights to guide educational strategies and policies, LA contributes to broader system-wide improvements.

Additionally, another emerging goal of LA is to foster transparency and home–school collaboration. By granting parents real-time access to their children's progress, LA enables families to play a more active role in supporting their children's academic development.

The document review highlights broad stakeholder involvement in LA implementation across partner countries, including policymakers, education authorities, institutional leaders, teachers, and students as primary users. Researchers contribute to analysis, evaluation, and the development of frameworks, while school leaders and administrators focus on coordination within institutions. Technical teams, developers, and IT specialists play a key role in ensuring system functionality and data integrity. Efforts to engage parents also foster transparency and home–school collaboration. Additionally, international organizations, such as EU institutions and the World Bank, are referenced as key stakeholders in advancing LA initiatives.

In addition to the document review, we interviewed teachers, researchers, policymakers, and education specialists involved in LA and informatics education. Their insights highlight the evolution of LA practices in Europe, focusing on improving teaching, enabling data-informed decision-making, and connecting research, policy, and practice to support the development of digital education.

Across the ten partner countries, LA implementation varies significantly, reflecting both advanced integration and emerging practices. Finland and Lithuania exemplify well-developed, systematic LA ecosystems that leverage robust digital infrastructures and ethical frameworks. Their integration spans institutional, research, and classroom levels, enabling detailed monitoring, diagnostic assessments, and tailored instructional design. Teachers and policymakers use insights to identify at-risk learners, personalize support, and inform system-level decisions, with GDPR compliance ensuring transparency.

Conversely, Bulgaria, Croatia, Cyprus, Greece, and Portugal are transitioning toward broader adoption. Bulgaria and Greece are gradually integrating LA across classrooms, institutions, and policies, with growing recognition of LA pedagogical potential. Portugal relies largely on administrative data for early identification of at-risk students. Croatia focuses on dashboards for tracking and adaptive assessment, while Cyprus uses platform-based monitoring tools embedded in Learning Management Systems (LMS). These countries face challenges in awareness and infrastructure but share a trajectory toward deeper integration.

In Italy, Slovenia, and Spain, LA practices remain informal or fragmented. In Italy, implementation is teacher-driven, relying on classroom observation and limited digital traces. Slovenia uses Learning Management System (LMS) features for formative assessment, focusing on monitoring, but lacks widespread integration. Spain is expanding its use of LMS dashboards and diagnostic tools, driven by EU initiatives. Together, the partner countries highlight varying readiness levels and the gradual shift from anecdotal or pilot use to structured, policy-backed LA strategies.

The interviews reveal insights into LA implementation at the institutional and national levels. Across the ten partner countries, support for LA varies significantly due to differences in systemic commitment, policy clarity, and resource allocation. Nordic and Baltic countries, such as Finland and Lithuania, demonstrate higher adoption rates supported by university-led initiatives, strategic programs, and EU funding. Finland leverages its robust research and digital infrastructure, including the ViLLE digital learning platform (Laakso et al., 2018), though it lacks formal national policies dedicated to LA. Lithuania benefits from initiatives such as EdTech LT and TŪM, but still lacks targeted funding mechanisms and institutional structures for LA.

Southern and Mediterranean countries, including Bulgaria, Greece, Italy, Spain, and Portugal, show fragmented LA implementation. Bulgaria benefits from ministry initiatives, but tangible support for LA remains limited. Greece has extensive digital infrastructure, but its LA adoption is mainly experimental. Italy relies on regional initiatives, like the National Digital School Plan, but lacks dedicated policies, funding, or widespread training. Spain participates in national and EU projects promoting digital education despite an absence of specific LA policies. Portugal's efforts

focus on broader digital transformation programs and small-scale pilots, with universities driving progress in LA adoption.

Eastern European countries, Croatia and Slovenia, demonstrate minimal systemic support for LA, underscoring the need for clear policy frameworks, strategic funding, and institutional backing to transition LA to mainstream practice. Croatia's progress remains hindered by political and institutional challenges, while Slovenia lacks institutional and national policies and funding, depending instead on informal, educator-led initiatives. Overall, despite EU-funded projects, regional initiatives, and university efforts, the lack of dedicated policies and sustainable mechanisms continues to hinder the deep integration of LA into national education systems across the partner countries.

1.5 Learning Analytics in Informatics Education

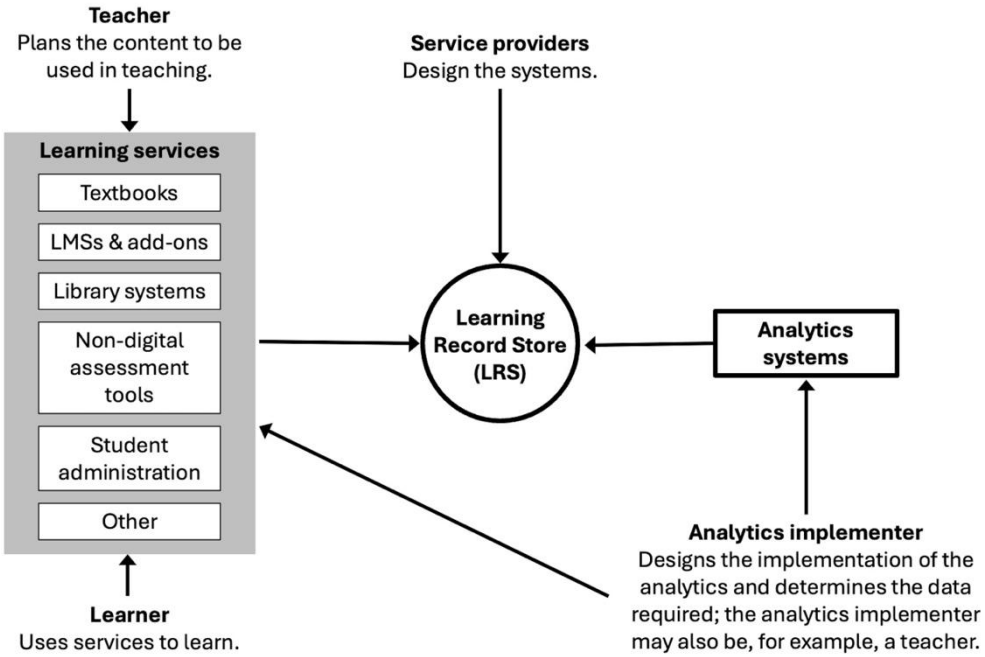
LA has transformative potential across diverse educational fields, including informatics education, where digital environments naturally produce rich data through programming platforms, simulations, and interactive problem-solving tools. These environments generate detailed traces of student actions, enabling a deeper understanding of learning processes. LA can help teachers analyze how students develop problem-solving strategies, debug code errors, and apply computational concepts. Additionally, analytics reveals common error patterns in programming tasks, allowing educators to address misconceptions proactively. Beyond understanding coding performance, LA supports the assessment of higher-order cognitive skills, such as abstraction, decomposition, and algorithmic reasoning, by analyzing students' interaction behaviors rather than relying solely on final code submissions.

Empirical studies demonstrate that LA's integration into technology-rich learning environments enhances student performance and engagement, especially when paired with effective pedagogical approaches (Awad et al., 2024). Projects like CT&MathABLE (https://www.fsf.vu.lt/ct-math-able) showcase how LA can create meaningful connections between analytics, assessment, and classroom practice, promoting impactful teaching and learning experiences.

1.6 Key Stakeholders in Learning Analytics

The LA ecosystem involves different stakeholders, such as learners, teachers, parents, education providers, policymakers, service providers, and analytics implementers. Successful implementation of LA requires collaboration among these stakeholders, as demonstrated in national initiatives in countries like Finland and Sweden, where coordinated efforts and clear governance structures are essential (Apiola et al., 2019), as shown in Figure 1.

Figure 1: Key Stakeholders in Learning Analytics (OKM, 2024, p. 73).



Learners are the primary beneficiaries of LA while providing data and granting consent for its use (or parents for underage learners). LA offers personalized learning experiences, timely feedback, and targeted support, improving learning outcomes. Summaries and visualizations of individual and group activities, combined with analytics-informed feedback, help learners understand their learning processes in relation to the group, fostering self-regulation and key metacognitive skills essential for academic growth (Brown, Rhomsen, & Al-Farouqi, 2025). Automated assessment further enables rapid feedback, guiding learners in adapting their approach. By personalizing instruction through recommendations for materials, tasks, or courses aligned with skill levels, LA helps improve learning outcomes, reduce dropout risks, and enhance learner engagement.

Teachers play a central role in using and interpreting data from LA systems to inform instructional decisions and support learning (Selwyn et al., 2023). Digital dashboards provide timely feedback by visualizing student learning processes—often difficult to observe directly in digital or blended contexts—along with performance and engagement. These insights help teachers monitor learners' performance and progress, identify their difficulties and misconceptions, assess the effectiveness of their own teaching and refine it to better align with student needs and objectives (van Leeuwen et al., 2021). LA data supports planning and improving instruction, targeted resource allocation (e.g., additional support or greater challenges for specific learners), and predictive modelling for tracking learning progress. Automated or semi-automated assessment tools further optimize teaching time by reducing manual evaluation, giving teachers more time for teaching and guidance. To fully harness LA benefits, teachers need strong data literacy—the ability to collect, analyze, and interpret data for actionable instructional decisions (Celik et al., 2022; Howard et al., 2022; Mandinach & Abrams, 2022; Mandinach & Gummer, 2016). This skill not only helps address individual learner needs effectively but also drives the digital transformation of education.

Parents also benefit from LA through real-time access to their children's progress, allowing them to actively engage in their education. This transparency strengthens home–school collaboration and enables parents to provide tailored support to enhance academic achievement.

Education authorities and providers play a crucial role in supporting LA by offering infrastructure and ensuring alignment with institutional and national objectives. They facilitate teacher development, establish ethical practices,

and ensure policy compliance. LA helps analyze the efficiency of courses, study pathways, and degree programs using data such as grades, dropout rates, completion rates, and learner demographics (e.g., major, age, entry pathway). It can uncover connections between courses to optimize study programs and better serve students. Additionally, LA supports quality assurance, curriculum evaluation, and the review of teaching practices, highlighting areas for teacher training and support. It also guides evidence-based decision-making in resource allocation, strategic planning, and operational improvement, enhancing overall educational outcomes.

Policymakers at regional and national levels develop strategic directives for incorporating LA into education systems. They play a critical role in shaping regional and national goals, allocating funding, and fostering coherent data governance frameworks. Ministries of education can promote large-scale initiatives and partnerships that integrate LA into curricula and informatics education.

Service providers are essential in creating effective, accessible, interoperable LA tools that integrate seamlessly across diverse educational settings. In addition to functionality, technology providers prioritize data privacy and security, ensuring compliance with regulations like GDPR. By delivering innovative solutions, providers enhance LA's technical capacity while empowering educators and administrators to use analytics effectively for improving teaching, learning outcomes, and decision-making at all levels.

Developers of learning materials and learning environments can use LA to gain insights into the effectiveness of their materials and environments. The collected data can reveal problematic areas and guide improvements. For example, tracking how learners interact with materials can identify sections that receive either excessive or insufficient attention. When combined with task completion data, this information can help identify the knowledge required for tasks and examine learners' information-seeking strategies. Similarly, data from task completion can be used to assess task suitability and support the development of better tasks. Data protection legislation does not restrict the collection of anonymous data by service providers, who may be authorized to gather usage information as long as it remains anonymous.

Analytics implementers, including data scientists and researchers, bring critical expertise to LA ecosystems, ensuring that data is accurately collected, analyzed, and interpreted. Data scientists focus on developing scalable and precise analytical models that handle large datasets efficiently, improving the reliability and usability of insights. Researchers enhance these efforts by exploring patterns, interpreting findings, and identifying key trends. Together, they enable stakeholders to make well-informed decisions to improve teaching, learning, and policies (Nouri et al., 2019). Researchers are also responsible for publishing their findings, allowing others to learn and benefit from the results.

Table 1 presents examples of areas where LA can have an impact on different stakeholders, along with examples of actions that can be taken and the desired outcomes. In practice, there are often multiple actions that can be used to address a single area, and several possible improvement goals. For example, the first row shows that students' cognitive performance can be supported through personalized feedback, with the goal of improving their cognitive skills.

Table 1: Examples of how LA can support different stakeholders, including key areas of impact, related actions, and intended outcomes (TRILA, 2021, pp. 10–11).

Stakeholder	Impact Areas	Actions	Desired Outcomes
Students	Cognitive performance	Personalized feedback / Recommendations	Enhancing cognitive skills
	Academic progress	Adaptive assignments	Supporting academic progress
	Motivation	Reminders / Encouragement messages	Increase student engagement
Teachers	Course planning	Analysis of learning materials	Improve subject understanding
	Interventions	Identifying students at risk of dropping out	Increase course completion rate
Curriculum Planners	Pedagogical models	Comparing study plans	Improve subject understanding
	Study program	Analyzing courses and course materials	Support student progress
Administration	Degree program	Analyzing student engagement	Increase the number of graduates
	Resource allocation	Identifying resource gaps	Influence decision-making

Our interviews highlight varying stakeholder involvement, levels of maturity, and systemic support across the ten partner countries. Cyprus, Finland, and Lithuania stand out for their integrated, multi-level collaboration, where policy, pedagogy, and technology intersect. Active participation of teachers, students, researchers, and policymakers has been observed. Formal support systems, such as professional training, co-design activities, and participatory approaches, are more established. This collaborative ecosystem aligns national strategies, educational practices, and technological infrastructures, ensuring the effective and sustainable deployment of LA systems.

In countries like Portugal, Slovenia, Greece, and Spain, stakeholder engagement in LA remains informal and fragmented. Teachers are primary users, often relying on voluntary participation, ad hoc training, pilot projects, or digital resources. Student involvement is limited, and policymakers and researchers typically engage in project-based activities rather than systemic processes. Professional development and stakeholder feedback are scarce, highlighting the need for stronger, inclusive collaboration and clear engagement pathways.

The LA landscape in the partner countries ranges from collaborative ecosystems to isolated, pilot-driven efforts. Deeply embedded LA practices emerge when teachers, students, and other stakeholders actively engage in co-design, feedback, and interpretation. Conversely, superficial, sporadic, or top-down approaches limit adoption and impact. Systematic, participatory, and sustained stakeholder engagement is essential for realizing LA's full potential in enhancing teaching and learning.

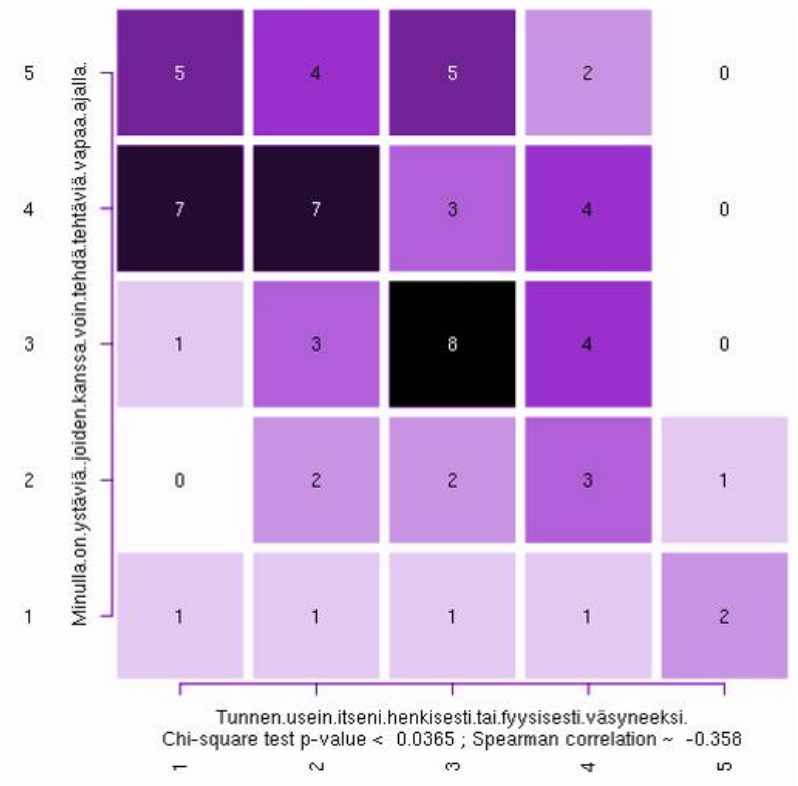
1.7 Examples of Learning Analytics Applications

LA offers versatile applications across various educational contexts, enabling educators, institutions, and policymakers to make data-informed decisions. From improving teaching practices to enhancing institutional operations, LA can serve different purposes depending on the goals and data available. Below are four examples illustrating how LA can be applied to support diverse educational needs and objectives (https://en.learninganalytics.fi/analytics#case__1).

Analysis of Study Habits

Questionnaires are valuable tools for collecting data in LA and can be even more insightful when combined with other data, such as from learning platforms or sensors. However, questionnaires alone can offer important insights into learner behaviors and experiences. Figure 2 showcases data from two questions from a questionnaire measuring students' study habits. The results reveal that students who report having friends to collaborate with completed assessments and experienced significantly fewer instances of feeling mentally or physically tired.

Figure 2: Students' responses to two questionnaire items on their study habits, emphasizing the importance of collaboration with friends for physical and mental well-being.

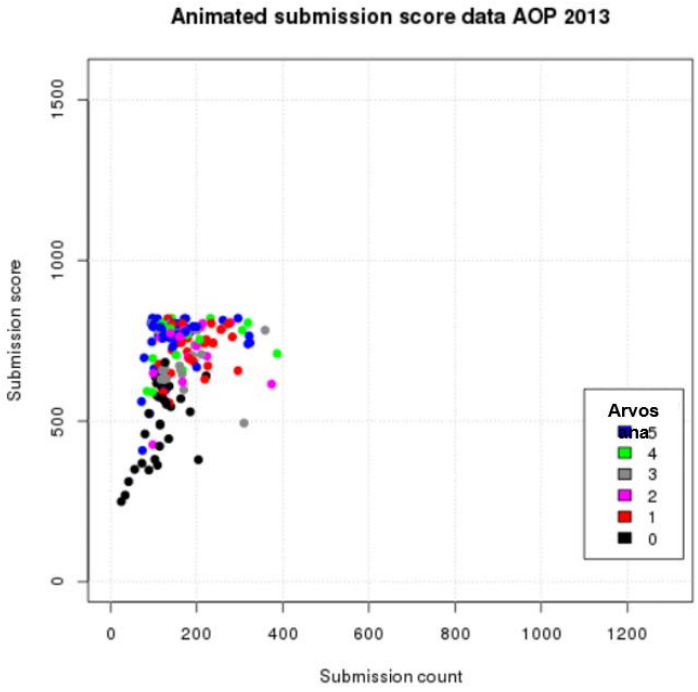


Facilitating the formation of supportive peer groups within educational settings could improve students' academic experiences by reducing stress and promoting engagement. Incorporating this insight into classroom management or collaborative learning designs could strengthen interpersonal relationships among students and optimize their learning outcomes.

Predicting Course Grades

Continuous assessment and comprehensive tracking of course achievements enable a predictive model for forecasting course outcomes. Figure 3 illustrates data collected during the first two weeks of an eight-week course. Each dot represents a student, and the colour-coding indicates their final grade. Using this predictive model, 80% of students likely to fail the course could be identified within the first two weeks, offering educators an invaluable opportunity to intervene early.

Figure 3: Predictive modelling of student learning outcomes based on course achievements during the first two weeks of an eight-week course.

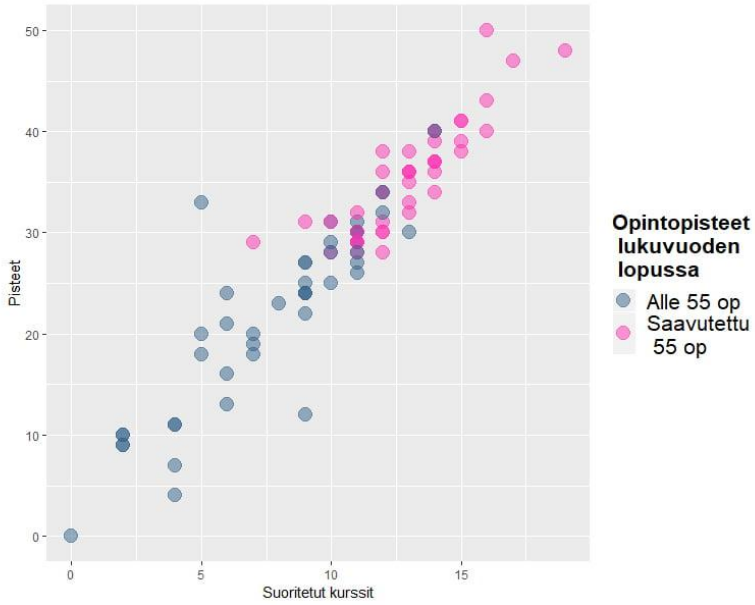


This type of analytics can inform teaching adjustment, additional support, curriculum adjustments, and resource allocation. Such interventions, supported by LA, transform data insights into proactive measures that optimize both student success and overall course effectiveness.

Achieving Curriculum Objectives

A common application of LA is developing predictive models from existing data to ensure students meet curriculum objectives. Figure 4 shows an analysis of yearly study point accumulation, which allows educators and institutions to forecast future progress and identify students at risk of falling behind.

Figure 4: Analysis of yearly study point accumulation used to project student progress and identify those at risk of lagging behind.

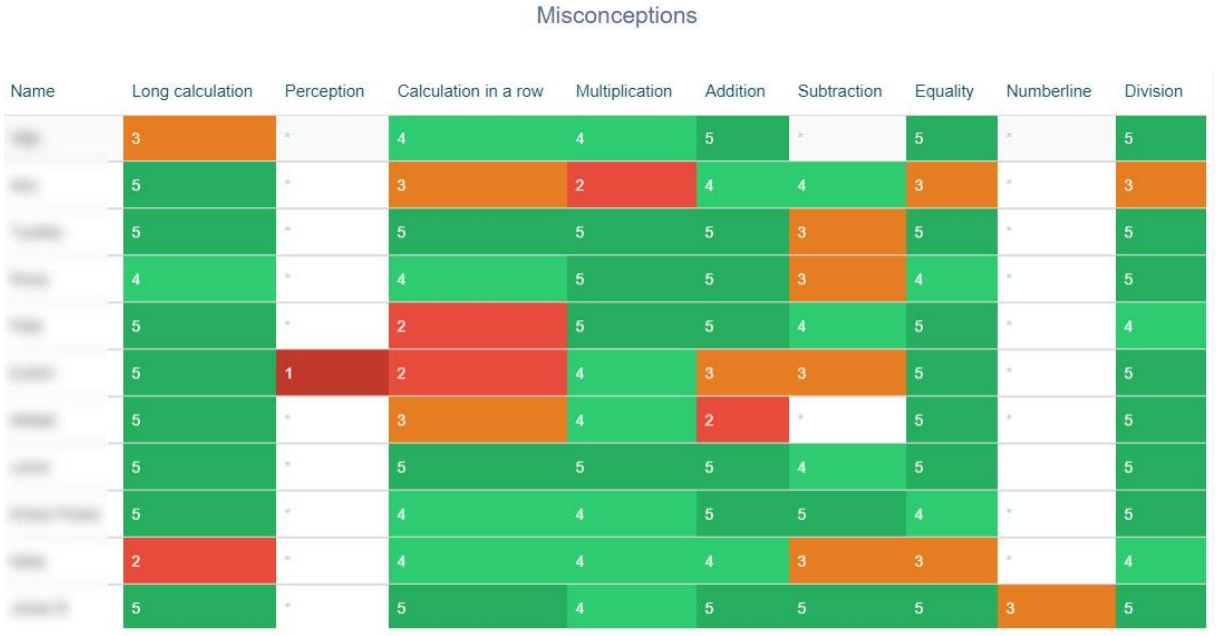


Early identification of these challenges allows for timely intervention, such as tailored support, workload adjustments, or additional guidance. This proactive approach helps prevent issues like delayed graduations or dropouts and keeps students on track to achieve their educational goals.

Automating Learning Analytics

The ViLLE learning platform employs automatic analytics to identify students' learning misconceptions in mathematics using data from their submissions (Figure 5). A study demonstrates that ViLLE's algorithms predict learning misconceptions as effectively as traditional pen-and-paper tests.

Figure 5: Automatic analytics in ViLLE platform detecting students' learning misconceptions in mathematics based on submission data.



Compared to pen-and-paper tests, automatic analytics provides the additional advantage of instant, actionable feedback, allowing teachers to address challenges promptly without the need for separate assessments.

Key Definitions & Terminology

Educational Data Mining (EDM): The development and use of methods to explore the unique and increasingly large-scale data that come from educational settings to better understand students and the educational context.

Learning Analytics (LA): The measurement, collection, analysis and reporting of data about learners and their contexts for purposes of understanding and optimising learning and the environments in which it occurs.

Learning Management System (LMS): A software application or web-based tool (e.g., Moodle, Blackboard, Canvas, Google Classroom) used to plan, implement, and assess a specific learning process.

Key Questions about LA

(OKM, 2024, pp. 24–29)

- How can LA be classified?

LA can be classified into three categories based on the advancement of its processes: visual analytics, profiling analytics, and automated decision-making. Additionally, LA focused-on data processing can be divided into four phases that describe its progression: descriptive analytics, diagnostic analytics, predictive analytics, and prescriptive analytics.

- Does the data protection regulation prohibit the use of LA?

The data protection regulation does not prohibit the use of LA but establishes boundaries for processing personal data. Personal data, defined as information directly or indirectly linked to identifiable individuals, requires a legal basis for processing. For education providers, this basis is typically the statutory task of providing education, meaning consent cannot serve as a basis for processing data required for LA. Lawfulness also relies on the criterion of necessity, ensuring data use complies with legal and ethical standards.

- How can security and privacy considerations associated with LA be addressed?

The data protection regulation mandates the implementation of technical and organizational measures to ensure context-appropriate security before processing personal data. Fundamental safeguards include access controls and regular security updates. In specific cases, such as processing health-related data, a data protection impact assessment is required in line with the regulation's risk-based approach.

- Is it worthwhile to start using LA?

LA can deliver valuable results when goals are clearly defined, and problems are well-understood beforehand. However, quick wins are rare, as success depends on establishing a culture of active LA use to inform and improve practices. Implementing LA should be seen as a long-term, continuous process rather than a one-off project, since changes influenced by analytics also shape future results, which can further enhance operations.

References

Aguerrebere, C., He, H., Kwet, M., Laakso, M.-J., Lang, C., Marconi, C., Price-Dennis, D., & Zhang, H. (2022). Global perspectives on learning analytics in K–12 education. In C. Lang, A. F. Wise, A. Merceron, D. Gašević, & G. Siemens (Eds.), *The handbook of learning analytics* (2nd ed., pp. 223–231). Society for Learning Analytics Research. https://doi.org/10.18608/hla22.022

Ampadu, Y. (2023). Handling Big Data in Education: A Review of Educational Data Mining Techniques for Specific Educational Problems. *AI Computer Science and Robotics Technology, 2*. https://doi.org/10.5772/acrt.17

Apiola, M., Karunaratne, T., Kaila, E., & Laakso, M. J. (2019). Experiences from digital learning analytics in finland and sweden: a collaborative approach. In *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (pp. 627–632). IEEE.

Arifin, M. and Widowati, W. (2022). Using Education Data Mining (EDM) and Tracer Study (TS) Data as Materials for Evaluating *Higher Education Curriculum and Policies. Kne Social Sciences.* https://doi.org/10.18502/kss.v7i14.11948

Ashaduzzaman, M., Sagor, M., Rahman, M., & Pritom, A. (2018). An Analysis of Students' Academic Record Using Data Mining Techniques and Identification of Key Factors to Aid Students' Performance. *Gub Journal of Science and Engineering, 5*(1), 45–50. https://doi.org/10.3329/gubjse.v5i1.47900

Awad, I., ALGHAMDI, A., & Al-Farani, L. (2024). The Impact of Artificial Intelligence Technologies in Educational Informatics on Improving Learners Performance: A Meta-Analysis. *Journal of Umm Al-Qura University for Educational and Psychological Sciences, 16*(3), 349–364. https://doi.org/10.54940/ep41266805

Brown, S., Rhomsen, R., & Al-Farouqi, H. (2025). Evaluating the Use of Learning Analytics in Formative Assessment. *International Journal of Post Axial: Futuristic Teaching and Learning*, 204–213.

Celik, I., Dindar, M., Muukkonen, H., & Järvelä, S. (2022). The promises and challenges of artificial intelligence for teachers: A systematic review of research. *TechTrends, 66*(4), 616–630. https://doi.org/10.1007/s11528-022-00715-y

Ferguson, R., & Buckingham Shum, S. (2011, February). Learning analytics to identify exploratory dialogue within synchronous text chat. In *Proceedings of the 1st international conference on learning analytics and knowledge* (pp. 99–103). https://doi.org/10.1145/2090116.2090130

González-Brambila, S., & González, J. (2016). Discovering Relationships among Personal and Academic Factors with Academic Performance using Association Rules. *Research in Computing Science, 118*(1), 9–17. https://doi.org/10.13053/rcs-118-1-1

Howard, S. K., Swist, T., Gašević, D., Bartimote, K., Knight, S., Gulson, K., Apps, T., Peloche, J., Hutchinson, N., & Selwyn, N. (2022). Educational data journeys: Where are we going, what are we taking and making for AI? *Computers and Education: Artificial Intelligence, 3*, Article 100073. https://doi.org/10.1016/j.caeai.2022.100073

Janardhana, D., Arif, H., U, H., Hanumanthappa, H., & Gaikwad, V. (2025). Personalized Learning Systems Using AI Enhancing Adaptive Education. *International Journal Environmental Science*, 668–676. https://doi.org/10.64252/ps80c806

Kanth, R., Laakso, M. J., Nevalainen, P., & Heikkonen, J. (2018). Future educational technology with big data and learning analytics. *In the Proceeding so the IEEE 27th International Symposium on Industrial Electronics (ISIE)* (pp. 906–910). IEEE.

Karademir, O., Di Mitri, D., Schneider, J., Jivet, I., Allmang, J., Gombert, S., Kubsch, M., Neumann, K., & Drachsler, H. (2024). I don't have time! But keep me in the loop: Co-designing requirements for a learning analytics cockpit with teachers. *Journal of Computer Assisted Learning, 40*(6), 2681–2699. https://doi.org/10.1111/jcal.12997

Khajuria, S., Devi, S. I., & Galgotra, M. (2025). Enhancing Student Achievement through Modern Assessment Methods and Feedback Mechanisms in Higher Education: Aligning with NEP 2020. *International Journal for Multidisciplinary Research, 7*(2). https://doi.org/10.36948/ijfmr.2025.v07i02.40419

Kılıç, A. H., & İzmirli, S. (2024). A Systematic Literature Review of Articles on Learning Analytics. *Asian Journal of Distance Education, 19*(2), 187–202. https://doi.org/10.5281/zenodo.13985017

Knight, S. (2020). Augmenting Assessment with Learning Analytics. *The Enabling Power of Assessment*, 129–145. https://doi.org/10.1007/978-3-030-41956-1_10

Kumar, R., Balla, R., Chahal, D., Yadav, R., Manzer, S. M., Kadaiyan, R., ... & Singh, G. (2025). Creating Digital Environment Using Data Analytics and AI for Evaluation: A Conceptual Study. *International Journal of Environmental Sciences*, 3435–3444. https://doi.org/10.64252/9dfkj869

KumarYadav, S., & Pal, S. (2012). Data Mining Application in Enrollment Management: A Case Study. International *Journal of Computer Applications, 41*(5), 1–6. https://doi.org/10.5120/5534-7581

Laakso, M.-J., Kaila, E., & Rajala, T. (2018). ViLLE – collaborative education tool: Designing and utilizing an exercise-based learning environment. *Education and Information Technologies, 23*, 1655–1676.

Mandinach, E. B., & Abrams, L. M. (2022). Data literacy and learning analytics. In C. Lang, A. F. Wise, A. Merceron, D. Gašević, & G. Siemens (Eds.), *Handbook of learning analytics* (pp. 196–203). Society for Learning Analytics Research. https://doi.org/10.18608/hla22.019

Mandinach, E. B., & Gummer, E. S. (2016). What does it mean for teachers to be data literate: Laying out the skills, knowledge, and dispositions. *Teaching and Teacher Education, 60*, 366–376.

Meghji, A., Mahoto, N., Unar, M., & Shaikh, M. (2019). Predicting Student Academic Performance using Data Generated in Higher Educational Institutes. *3c Tecnología_glosas De Innovación Aplicadas a La Pyme*, 366–383. https://doi.org/10.17993/3ctecno.2019.specialissue2.366-383

Ministry of Education and Culture, Finland (OKM), Learning Analytics Division. (2024). *Framework for Learning Analytics: Best practices in the implementation and utilisation of learning analytics*. https://urn.fi/URN:ISBN:978-952-263-731-4

Mukherjee, T. (2022). Improving Data Quality for Educational Data Mining (EDM) for Indian Ed-Tech Start-Ups. *International Journal of Science Engineering and Management, 9*(8), 32-33. https://doi.org/10.36647/ijsem/09.08.a005

Nouri, J., Ebner, M., Ifenthaler, D., Sqr, M., Malmberg, J., Khalil, M., ... & Berthelsen, U. D. (2019). *Efforts in Europe for Data-Driven Improvement of Education–A review of learning analytics research in six countries*.

Nistor, N., & Hernández-García, Á. (2018). What types of data are used in learning analytics? An overview of six cases. *Computers in Human Behavior, 89*, 335–338. https://doi.org/10.1016/j.chb.2018.07.038

Özdağoğlu, G., Öztaş, G., & Çağlıyangil, M. (2018). An application framework for mining online learning processes through event-logs. *Business Process Management Journal, 25*(5), 860-886. https://doi.org/10.1108/bpmj-10-2017-0279

Pal, S. (2012). Mining Educational Data to Reduce Dropout Rates of Engineering Students. *International Journal of Information Engineering and Electronic Business, 4*(2), 1–7. https://doi.org/10.5815/ijieeb.2012.02.01

Penteado, B., Paiva, P., Morettin-Zupelari, M., Isotani, S., & Ferrari, D. (2018). Toward Better Outcomes in Audiology Distance Education: An Educational Data Mining Approach. *American Journal of Audiology, 27*(3S), 513–525. https://doi.org/10.1044/2018_aja-imia3-18-0020

Qu, S., Li, K., Wu, B., Zhang, S., & Wang, Y. (2019). Predicting Student Achievement Based on Temporal Learning Behavior in MOOCs. *Applied Sciences, 9*(24), 5539. https://doi.org/10.3390/app9245539

Raković, M., Gašević, D., Hassan, S., Ruipérez-Valiente, J., Aljohani, N., & Milligan, S. (2023). Learning analytics and assessment: Emerging research trends, promises and future opportunities. *British Journal of Educational Technology, 54*(1), 10–18. https://doi.org/10.1111/bjet.13301

Ruby, J., & David, K. (2017). An Analysis on Academic Performance of Students using a Hybrid Model for Higher Education. *International Journal of Engineering and Technology, 9*(3), 2175–2182. https://doi.org/10.21817/ijet/2017/v9i3/1709030146

Russell, D. L. and Wallis, S. E. (2019). Designing a Learning Analytic System for Assessing Immersive Virtual Learning Environments. *Virtual Reality in Education*, 26–51. https://doi.org/10.4018/978-1-5225-8179-6.ch002

Saneifar, R. and Abadeh, M. (2015). Association Rule Discovery for Student Performance Prediction Using Metaheuristic Algorithms. In Jan Zizka et al. (Eds.): *ICAITA, SAI, CDKP, Signal, NCO*, pp. 115–123. https://doi.org/10.5121/csit.2015.51510

Selwyn, N., Hillman, T., Bergviken-Rensfeldt, A., & Perrotta, C. (2023). Making sense of the digital automation of education. *Postdigital Science and Education*, *5*(1), 1–14.

Sharma, G. (2020). Tendency of Educational Data Mining in Digital Learning Platform. *Current Trends in Computer Sciences & Applications, 2*(1). https://doi.org/10.32474/ctcsa.2020.02.000127

Shum, S. B., & Ferguson, R. (2012). Social learning analytics. *Journal of Educational Technology & Society, 15*(3), 3–26.

Silva, R., Godwin, G., & Jayanagara, O. (2024). The Impact of AI on Personalized Learning and Educational Analytics. *International Transactions on Education Technology (ITEE), 3*(1), 36-46. https://doi.org/10.33050/itee.v3i1.669

Sousa, E. B. D., Alexandre, B., Ferreira Mello, R., Pontual Falcão, T., Vesin, B., & Gašević, D. (2021). Applications of learning analytics in high schools: A systematic literature review. *Frontiers in Artificial Intelligence*, *4*, 737891.

Triayudi, A., Aldisa, R., & Sumiati, S. (2024). New Framework of Educational Data Mining to Predict Student Learning Performance. *Journal of Wireless Mobile Networks Ubiquitous Computing and Dependable Applications, 15*(1), 115–132. https://doi.org/10.58346/jowua.2024.i1.009

Valtonen, T., Paavilainen, T., López-Pernas, S., Saqr, M., & Hirsto, L. (2025). Elementary and Secondary School Teachers' Perceptions of Learning Analytics: A Qualitative Approach. *Technology, Knowledge and Learning*, 1–19. https://doi.org/10.1007/s10758-025-09847-5

van Leeuwen, A., Knoop-van Campen, C. A. N., Molenaar, I., & Rummel, N. (2021). How teacher characteristics relate to how teachers use dashboards: Results from two case studies in K–12. *Journal of Learning Analytics, 8*(2), 6–21. https://doi.org/10.18608/JLA.2021.7325

Wanjau, S., Okeyo, G., & Rimiru, R. (2016). Data Mining Model for Predicting Student Enrolment in STEM Courses in Higher Education Institutions. *International Journal of Computer Applications Technology and Research, 5*(11), 698–704. https://doi.org/10.7753/ijcatr0511.1004

Yu, H. (2025). Application of Learning Analytics Technology in Architectural Education Informatics. *Journal of Progress in Engineering and Physical Science, 4*(4), 31–37. https://doi.org/10.56397/jpeps.2025.08.04

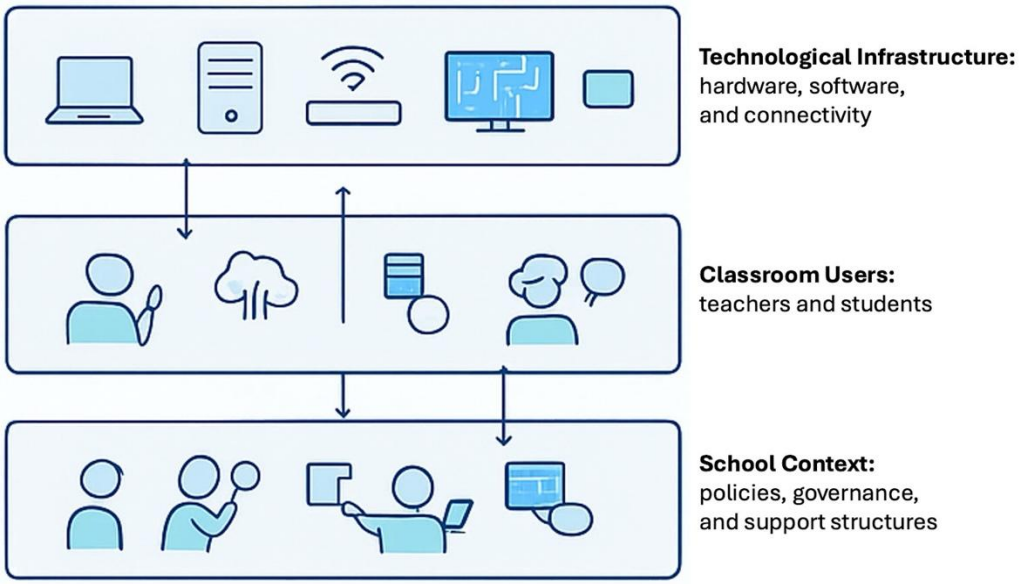
Zamecnik, A., Kovanović, V., Grossmann, G., Joksimović, S., Jolliffe, G., Gibson, D., & Pardo, A. (2022). Team interactions with learning analytics dashboards. *Computers & Education*, *185*, 104514.

Zhang, J., Shi, J., Liu, X., & Zhou, Y. (2021). An Intelligent Assessment System of Teaching Competency for Pre-service Teachers Based on AHP-BP Method. *International Journal of Emerging Technologies in Learning, 16*(16), 52. https://doi.org/10.3991/ijet.v16i16.17891

2 Necessary Classroom Architecture for Meaningful Educational Data Mining

The success of Learning Analytics (LA) and Educational Data Mining (EDM) in K–12 education extends beyond analytical methods or algorithms. It requires well-designed classroom architecture that aligns pedagogy, technology, and ethics. Classroom architecture, in this context, is broadly defined to include physical spaces, technological infrastructure, digital environments, instructional design and practices, and supportive organizational conditions. Figure 6 visually represents the EDM/LA ecosystem in K–12 classrooms. It highlights the interaction among three core dimensions: (1) technological infrastructure (hardware, software, and connectivity), (2) classroom users (teachers and students), and (3) the school context (policies, governance, and support structures). Together, these interconnected elements form an ecosystem that enables effective data collection, interpretation, and application to improve teaching, learning, and decision-making. Successful LA/EDM requires alignment across these dimensions—technology that supports pedagogy, engaged people, and robust, ethical policies.

Figure 6: The EDM/LA ecosystem in K–12 classrooms: technology, users, and governance (AI generated image).



This chapter identifies the key architectural elements needed to implement meaningful LA and EDM in school education. It focuses on the following aspects:

1. **Technological Infrastructure**: The hardware, software, and connectivity needed to support real-time data collection, analysis, and visualization.
2. **Modern Learning Environments**: The integration of physical and digital spaces, alongside flexible instructional designs that facilitate learning and LA.
3. **Enabling and Constraining Factors**: Institutional and systemic conditions that either support or limit the use of LA, such as policy alignment, teacher training, and data privacy concerns.
4. **Case Examples from Educational Contexts**: Examples illustrating effective classroom architecture setups across diverse educational contexts.

By examining these interconnected components, this chapter provides guidance on creating optimal environments for leveraging LA and EDM to enhance teaching, learning, and decision-making in K–12 education.

2.1 Technological Infrastructure

A robust and reliable technology infrastructure, including hardware, software, and connectivity, is essential for meaningful EDM. In K–12 contexts, this infrastructure must be age-appropriate, secure, scalable, and aligned with pedagogical goals rather than being driven solely by technological innovation (Paolucci et al., 2024).

Hardware

Classrooms require networked devices such as laptops, tablets, or desktop computers that students and teachers can seamlessly integrate into learning activities. Mobile devices, interactive whiteboards, projectors, and shared displays enhance collaboration and whole-class engagement, while generating valuable learning data. Emerging sensor-based technologies, such as classroom response systems, wearable devices like eye trackers, and environmental sensors, are increasingly used to collect data on student engagement, movement, and interaction patterns. In informatics education, specialized tools like educational robots and programming kits can be connected to analytics systems to monitor and support learning in programming and computational thinking. These tools provide rich data on student performance and offer unique opportunities for targeted interventions while fostering hands-on, problem-solving skills. Additional hardware can support specific purposes, enhancing accessibility and security in learning and assessment activities. Headphones, for example, allow students to listen to task instructions, accommodating diverse preferences and supporting those with reading challenges or visual impairments. For strong authorization, such as in high-stakes assessments, biometric technologies such as fingerprint or facial recognition ensure secure, efficient student identification. These systems offer greater security than traditional methods, reduce academic dishonesty, and streamline administrative processes.

Software

Software infrastructure is equally critical for enabling effective implementation of EDM/LA in classrooms. Learning Management Systems (LMS) serve as centralized platforms for organizing learning materials, assignments, assessments, and communication. These systems generate detailed log data that can be analyzed to understand learner engagement, progress, and challenges. Additionally, digital learning platforms, adaptive learning systems, and intelligent tutoring systems enhance data collection by capturing fine-grained interaction traces, such as problem-solving steps, time spent on tasks, or response patterns. These tools not only facilitate personalized learning but also provide teachers with actionable insights to refine instruction and address student needs more effectively. Together, these systems form the software backbone supporting meaningful data-driven practices in education. To support EDM, these systems should be interoperable and capable of securely and standardly sharing data across platforms. Data standards and APIs play an important role in enabling integration while avoiding fragmented data silos.

Implementing technological infrastructure in classrooms requires reliable hardware, software, and connectivity, supported by thoughtful management to ensure scalability, regular updates, and maintenance. For effective digital transformation in education, understanding the readiness of teachers and infrastructure is paramount. Husna's research (2025) reveals that a high percentage of teachers see the need for innovative tools and recognize gaps in current infrastructure. With adequate training to use LA technologies such as dashboards and data visualizations effectively, teachers can effectively leverage educational technology, enriching the data collected through various learning platforms. By focusing on enhancing teacher competencies alongside robust infrastructure while addressing privacy and ethical concerns, educational institutions can create more valuable and insightful datasets for analysis. Components like secure servers, encrypted connectivity, and appropriate data storage and management

are crucial for protecting sensitive information. By balancing innovation with security and educational relevance, technology infrastructure can unlock transformative applications of LA and EDM in K–12 education.

2.2 Modern Learning Environments

Effective classroom architecture directly influences the type of data that can be mined to improve educational practices. Modern learning environments go beyond traditional classrooms, offering spaces designed to support active, collaborative, and technology-enhanced learning. These environments are ideal for LA, as they naturally generate rich data from learner interactions with the learning environments (Beerwinkle, 2021) and enable seamless data collection. Smart classrooms are a prime example, integrating digital tools, networked devices, and platforms into authentic learning activities, such as collaborative problem-solving, inquiry-based projects, and formative assessments, embedding data collection into the learning process.

Digital Resources

Digital resources, such as simulations, programming environments, textbooks, and multimedia tools, play a central role in modern learning environments. These tools generate detailed data traces about student actions and outcomes, enabling real-time monitoring of learning processes. Coupled with LA, the collected data empowers teachers to adapt their instruction to address challenges and meet learners' needs more effectively.

Physical Facilities

The adequacy of physical classroom facilities (e.g., layout, accessibility, and design) and management of classroom environments is crucial for effective learning and teaching (Munawwaroh, 2026). Designing learning spaces that enhance learning, collaboration, flexibility, and student well-being fosters an environment conducive to rich educational outcomes. For instance, classrooms built with adaptive and flexible designs allow for diverse teaching strategies, which can lead to more nuanced data reflective of different learning experiences (Javaherikhah et al., 2021). Approaches like project-based, inquiry-based, and collaborative learning foster richer interactions, leading to more robust data for LA (Aguerrebere et al., 2022). Additionally, when classroom designs align with students' psychological needs and preferences, data mining efforts become more robust, revealing deeper insights into student engagement and learning behaviors. Classroom designs that support active learning can lead to higher student efficacy and engagement, vital metrics for data mining analyses (Chan et al., 2023). Thus, schools must prioritize physical space investments to promote a richer data-driven approach to education.

Ambient Intelligence

Ambient intelligence refers to a seamless integration of technology into educational environments, allowing for a context-aware setting that enhances the learning experience. Margetis et al. (2011) propose an architecture called ClassMATE, which provides the necessary mechanisms for pervasive computing in classrooms, facilitating context-aware learning. Such environments not only support traditional educational methods but also allow for sophisticated data collection that informs educators about student behaviors and interactions within the classroom.

Acoustics

The auditory environment of classrooms plays a vital role in educational settings. Park and Haan (2021) stress the importance of creating optimal acoustical conditions to improve speech intelligibility, thereby enhancing the learning experience. Without addressing noise levels and reverberation times, data collected through audio learning activities might be flawed or insufficient for meaningful analysis. An integrated approach that considers acoustic standards would thus improve both learning quality and the quality of data insights derived from classroom interactions.

The architecture of classrooms significantly impacts the effectiveness of data collection and analysis in educational settings. The integration of adequate physical facilities, ambient intelligence, optimal acoustics, and advanced technological readiness collectively contribute to enhanced learning environments. Adaptability supports scalability and future needs, crucial in the face of rapid technological evolution. Modular furniture, mobile devices, and reconfigurable digital infrastructure enable schools to adjust spaces and practices as education systems increasingly incorporate data-informed education. These elements not only improve educational outcomes but also enrich the datasets utilized for educational data mining, yielding more precise insights for educational improvement.

2.3 Enabling and Constraining Factors

The adoption of EDM/LA hinges on factors that must be reflected in classroom architecture. A well-designed environment aligns technology, pedagogy, and governance, creating an ecosystem where data can be collected, interpreted, and used effectively to improve teaching and learning.

Enabling Factors

Technology and methods: Advances in AI, machine learning, and data visualization provide actionable insights with minimal technical expertise. Architectural implications include integrated tools within LMSs, dashboards accessible to teachers, and reliable hardware and networks.

Ready-to-use examples and equipment: Case-based tutorials, hands-on lesson plans, and runnable dashboards that illustrate how data informs instruction, support plug-and-play implementation. Equipment that works out of the box with clear setup instructions and ongoing technical assistance reduces start-up time and ensures reliability. When these resources align with the curriculum, teachers can replicate and adapt them across classrooms, enabling scalable, consistent practice.

Professional development: Ongoing data-literacy training and coaching on the pedagogical use of analytics sustain adoption and empower teachers to integrate LA/EDM in the classroom. Schools should allocate time and resources for professional development and foster communities of practice to support collaboration, ongoing learning, and shared implementation.

Policy frameworks: Clear data governance, ethical guidance, and interoperability standards guide infrastructure design, ensure consistent practices, and enable scalable and reliable integration across platforms, providing teachers and schools with a stable foundation for incorporating LA into their educational ecosystems.

Constraining Factors

Resource limitations: Limited access to devices, unreliable connectivity, and insufficient support impede data collection and maintenance. Due to high costs and tight budgets, many informatics education classrooms may be unable to acquire specialized tools such as educational robots or programming kits. Additionally, even when technologies exist, they are often neglected or underused due to inadequate maintenance, training, or incentives.

Data fragmentation: Datasets from LMS, assessments, and external tools often use incompatible schemas and identifiers, making integration costly and error-prone. Incompatible data sources and systems hinder integration; harmonization and standardized metadata are needed to unify analytics across classrooms.

Teacher readiness: Teachers' limited familiarity with LA/EDM and unclear benefits hinder adoption. Addressing this requires targeted, practical training, ready-to-use examples, and ongoing support (e.g., coaching) to build confidence, demonstrate value, and sustain data-driven teaching.

Workload and curriculum misalignment: If analytics tools demand extra time or fail to align with standards or increase teacher workload, teacher adoption declines. Interfaces should be intuitive and context-specific, fitting existing workflows and classroom practices, and reflecting curricula to encourage sustained use.

Policy clarity and practicality: Without explicit guidance on data scope, governance roles, access, retention, and compliance, schools' risk inconsistent practices and diminished trust, hindering adoption. Provide ready-to-use templates (data inventories, DPIA checklists, data-sharing agreements) and simple decision aids to standardize procedures. Clear, usable policies enable scalable, sustainable data practices and ensure alignment of technology, pedagogy, and governance across classrooms, schools, and districts.

Ethical concerns: Diverging expectations about data ownership, use, and transparency can provoke resistance from teachers and parents, slowing adoption. Classroom architecture should embed ethics by incorporating clear consent, data minimization, governance, and accountability features, and by promoting transparent communication about data practices with all stakeholders.

Others: School culture strongly influences classroom architecture for EDM/LA. A culture that values collaboration, reflection, and evidence-based practice encourages active analytics use for pedagogy. When data are framed as accountability or surveillance, teacher buy-in declines, undermining data-driven classrooms. Teachers may concern that government authorities might use the data to evaluate their performance, creating resistance to its adoption despite its potential to improve student learning. Parental concerns about LA extend beyond privacy, which include fear of direct comparisons, stigma, and anxiety. Address this by framing LA as tracking individual progress with contextualized, non-comparative reporting and offering opt-out options. Present LA as a tool to support each learner's growth without ranking to boost parental engagement and facilitate smoother adoption.

2.4 Case Examples from Educational Contexts

Real-world cases across diverse educational contexts demonstrate how classroom architecture can support EDM/LA in practice. In the project partner countries, the technological infrastructure for EDM/LA prioritizes reliable connectivity, interoperable systems, and secure data environments to enable effective data collection and analysis. Many partner countries have developed institutional infrastructures using cloud-based or national data systems and learning management systems (LMS) like Moodle and Blackboard as primary platforms for data collection, learning delivery, and analytics integration. Dashboards are widely utilized to visualize learning progression, engagement, and institutional performance, often complemented by predictive analytics, diagnostic tools like SELFIE, and custom monitoring systems. Croatia, Lithuania, and Italy highlight the need for high-speed internet access, well-equipped classrooms, and robust public infrastructure for nationwide LA implementation. Greece, Slovenia, and Finland focus on ensuring reliable ICT environments for staff and students, supported by local and national digital networks. Countries like Bulgaria, Lithuania, and Spain emphasize LA-specific technologies, such as learning record stores (LRS), xAPI-SG standards, AI tools, and recommender systems for data interoperability and personalized feedback. Across the partner countries, LMS-based ecosystems enhanced by dashboards and predictive tools are central to leveraging learning data for actionable insights. Maturity levels vary, ranging from advanced, interoperable infrastructures to foundational initiatives focused on connectivity and classroom readiness.

Globally, large-scale digital learning platforms have been implemented to enhance learning, formative assessment, and continuous feedback in primary and secondary education (Arantes, 2021; Ogata et al., 2024; Reich, 2022). These platforms combine automated assessment, immediate feedback, and analytics dashboards, allowing teachers to detect misconceptions and learning gaps early. Regional or national systems exemplify how classroom-level data can be aggregated to inform both individualized teaching decisions and broader system-level insights (Aguerrebere et al., 2022; Macarini et al., 2020). These platforms integrate classroom architectures with wider educational ecosystems, ensuring secure data flow from individual learning activities to school and policy levels while preserving teacher autonomy.

ViLLE Digital Learning Platform

ViLLE (https://en.learninganalytics.fi/ville), developed by the Turku Research Institute for Learning Analytics (TRILA), Finland, in collaboration with teachers (Laakso et al., 2018), is a leading learning platform in the Finnish education system. It provides both exercises (Figure 7) and LA for students (Figure 8) and teachers (Figure 9) via mobile phones,

tablets, and computers, offering instant feedback for students and insightful learning data for teachers. ViLLE includes exercises in programming, computational thinking, mathematics, languages, and other subjects. Teachers can freely select materials or use built-in editors to create and share personalized exercises. Most exercises are automatically assessed, reducing workload and allowing teachers to focus on student support. ViLLE enables students to learn at their own pace while giving teachers more time for teaching. ViLLE high security standards ensure data privacy: student data is stored securely on university servers, accessible only by schools and teachers, with pseudonymized data available to TRILA statisticians and researchers. ViLLE's success in Finnish schools is remarkable: by 2025, it offered more than 250,000 exercises, was used by 70% of schools and over 35,000 teachers, and completed 300 million tasks during the school year. ViLLE's quality and usability have made it a foundational tool for LA-driven education transformation in Finland.

Figure 7: Examples of math exercises in ViLLE

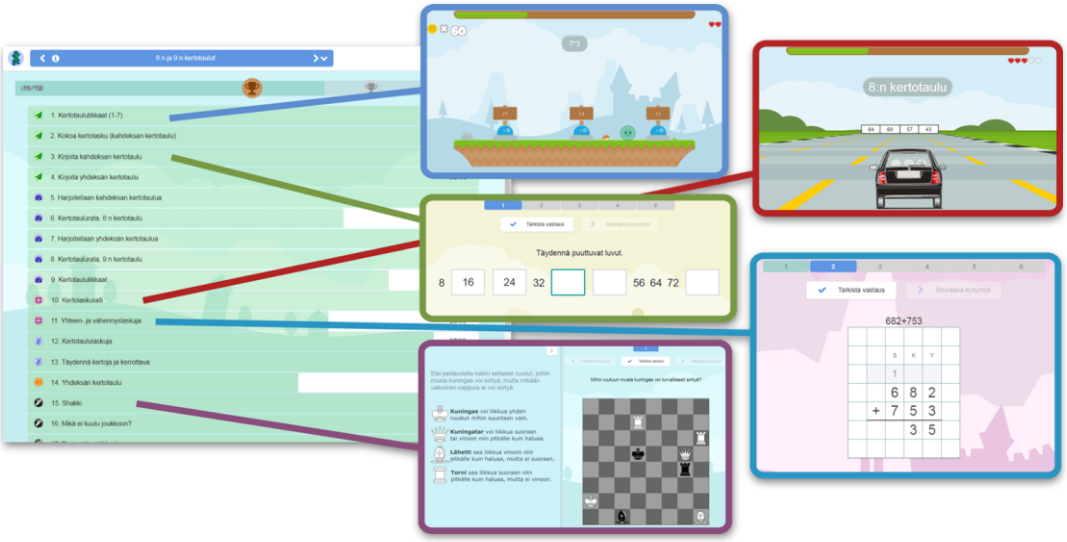


Figure 8: ViLLE analytics dashboard for students

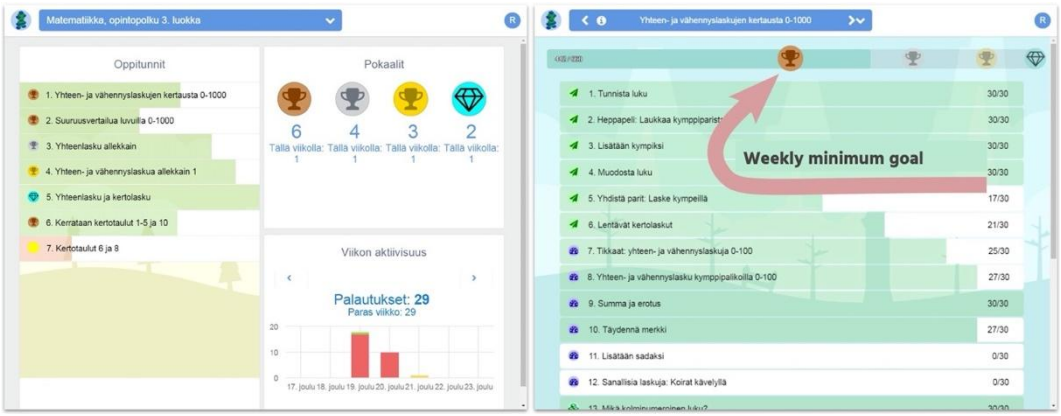
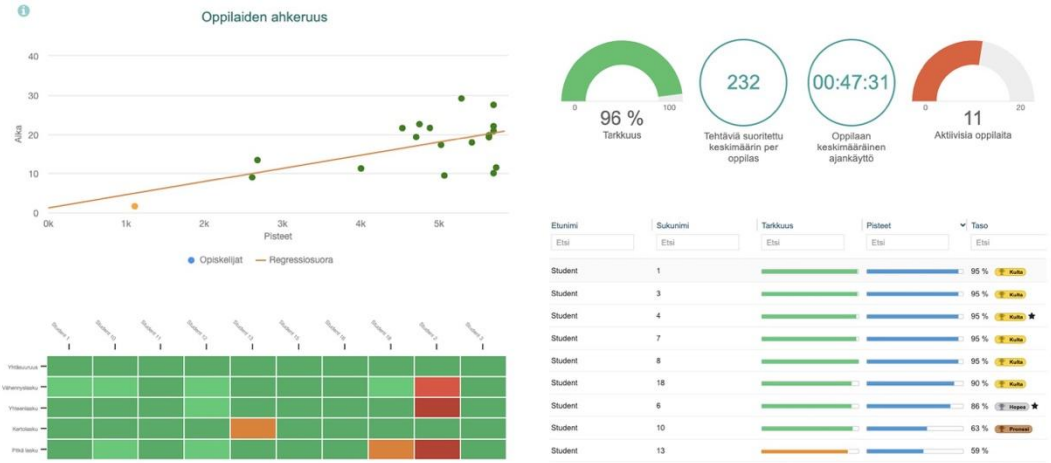


Figure 9: ViLLE analytics dashboard for teachers



Informatics classrooms equipped with educational technologies can generate rich data for EDM and LA. These technologies enable real-time data collection and analysis, fostering a deeper understanding of student learning processes and challenges while enhancing assessment practices and instructional strategies. The following

examples illustrate how LA and EDM can be integrated into informatics classrooms, demonstrating their potential to support student learning and empower teachers to create targeted, data-driven interventions.

Programming Class

Context: Students engage in programming challenges using an online coding platform such as Scratch. Each student works independently on small coding tasks that build essential programming skills over time.

Data Process: The platform automatically logs students' activities, including time spent on tasks, the number of code runs, frequency of debugging attempts, and error types encountered. Teachers can access a dashboard displaying real-time progress, student engagement, and prevalent error patterns. After class, teachers can review the analytics to identify challenging concepts (e.g., loops or variables) and design the next lesson to address these difficulties.

Process Flow: Student coding actions → platform logs interactions → server processes data → teacher dashboard visualizes key insights → teacher refines lesson plan.

Key Insight: This example highlights the importance of integrated systems, such as coding platforms with analytics capabilities, teacher dashboards, and reliable connectivity. Clear data-sharing protocols and robust systems ensure effective data tracking and seamless classroom implementation.

Robotics Activity with Logging Data

Context: Students collaborate in teams to program small robots equipped with sensors to navigate a maze. Each robot tracks movement paths and collision events.

Data Process: Robot sensors record data on movements and collisions, which is synced to tablets via Bluetooth. A classroom app logs team activity, including coding attempts, adjustments, and success rates. The app generates visual reports showing how students improved their algorithms through iterations. Teachers access analytics to evaluate team problem-solving approaches, collaboration, and persistence, offering targeted feedback during the debrief session.

Process Flow: Robot sensors → synced data to tablets → app aggregates team logs → visual analysis of problem-solving patterns → teacher provides feedback.

Key Insight: This example merges physical and digital architectures, linking sensor data, tablet interfaces, and teacher dashboards. It captures both cognitive and behavioral learning, emphasizing collaboration and real-time analytical support in hands-on informatics activities.

References

Aguerrebere, C., He, H., Kwet, M., Laakso, M. J., Lang, C., Marconi, C., ... & Zhang, H. (2022). Global perspectives on learning analytics in K–12 education. *by Charles Lang, Alyssa Friend Wise, Agathe Merceron, Dragan Gašević, and George Siemens. 2nd ed. Vancouver, Canada: SOLAR*.

Arantes, J. A. (2021). Learning Analytics at scale: collaborative or commercialized? *The International Journal of Technologies in Learning*, *29*(1), 17.

Beerwinkle, A. L. (2021). The use of learning analytics and the potential risk of harm for K-12 students participating in digital learning environments. *Educational Technology Research and Development*, *69*(1), 327–330.

Chan, D., Lam, E., & Adabre, M. (2023). Assessing the Effect of Pedagogical Transition on Classroom Design for Tertiary Education: *Perspectives of Teachers and Students. Sustainability, 15*(12), 9177. https://doi.org/10.3390/su15129177

Husna, H. (2025). Assessing Teachers' Readiness and Infrastructure Needs for Digital Classroom Development through a Service-Learning Approach. *Help Journal of Community Service, 2*(3), 140–150. https://doi.org/10.62569/hjcs.v2i3.228

Javaherikhah, A., López, M., & Álvarez, M. (2021). THE ARCHITECTURE OF IDEAL LEARNING ENVIRONMENTS., 1, 4687–4698. https://doi.org/10.21125/iceri.2021.1079

Laakso M., Kaila, E., & Rajala, T. (2018). ViLLE–collaborative education tool: Designing and utilizing an exercise-based learning environment. *Education and Information Technologies, 23*, 1655–167

Macarini, L. A., Lemos dos Santos, H., Cechinel, C., Ochoa, X., Rodés, V., Pérez Casas, A., ... & Díaz, P. (2020). Towards the implementation of a countrywide K–12 learning analytics initiative in Uruguay. *Interactive Learning Environments*, 28(2), 166–190.

Margetis, G., Leonidis, A., Antona, M., & Stephanidis, C. (2011). Towards Ambient Intelligence in the Classroom. *Lecture Notes in Computer Science*, (LNISA, volume 6768), 577–586. https://doi.org/10.1007/978-3-642-21657-2_62

Munawwaroh, Z. (2026). Evaluation of the Adequacy of Classroom Facilities and Learning Environment at MTsS Al-Ihsan Pamulang. *Pendas Jurnal Ilmiah Pendidikan Dasar, 11*(01), 113–130. https://doi.org/10.23969/jp.v11i01.40499

Ogata, H., Majumdar, R., Flanagan, B., & Kuromiya, H. (2024). Learning analytics and evidence-based K12 education in Japan: usage of data-driven services for mobile learning across two years. *International Journal of Mobile Learning and Organisation*, *18*(1), 15–48.

Paolucci, C., Vancini, S., Bex II, R. T., Cavanaugh, C., Salama, C., & de Araujo, Z. (2024). A review of learning analytics opportunities and challenges for K–12 education. *Heliyon*, *10*(4).

Park, C. and Haan, C. (2021). Initial Study on the Reverberation Time Standard for the Korean Middle and High School Classrooms Using Speech Intelligibility Tests. *Buildings, 11*(8), 354. https://doi.org/10.3390/buildings11080354

Reich, J. (2022). Learning analytics and learning at scale. *Handbook of learning analytics*, 188–195.

3 Data Structures and Necessary Data Points for Educational Data Mining

Educational Data Mining (EDM) applies data mining techniques to educational settings to understand learning and improve outcomes. EDM combines data-driven insights with theory to illuminate learners' needs, behaviors, and patterns associated with success or risk (Pei, 2017).

The growing use of digital learning platforms has resulted in large volumes of data being generated as students interact with online environments. Learning management systems (LMS), massive open online courses (MOOCs), and intelligent tutoring systems (ITS) routinely record students' activities, such as logins, resource access, task completion, assessment results, and patterns of engagement. These digital traces form an important foundation for EDM and learning Analytics (LA) as they provide detailed information about learning processes and outcomes (Papadogiannis et al., 2024).

Data points and data structure are two core concepts underpinning EDM. The primary distinction between them lies in their roles: **data points**, the individual data elements collected about learners, represent the actual data elements, while **data structures** define how that data is organized. The interplay between data points and data structures shapes data collection, processing, analysis, and interpretation. Aggregating data points enables broader insights through clustering and predictive modelling, helping educators identify at-risk students and tailor interventions (Feng et al., 2022). Effective data structures enable the organization of data points in a way that makes them easy to access, analyze, and interpret, facilitating more efficient data mining (Penteado et al., 2018; Ang et al., 2020). Through structured data, analysts can extract meaningful patterns and trends from clusters of data points, ultimately leading to predictive models that inform educational strategies and interventions (Simionescu et al., 2024; Tavares et al., 2017). A clear understanding of how data points relate within structures is essential to effectively extract actionable insights from large educational datasets.

3.1 Necessary Data Points for Educational Data Mining

3.1.1 Understanding Data Points

A data point is the smallest meaningful unit of information used in EDM. It represents a single observation about a student, a learning activity, or a learning context at a particular moment in time. For example, a student's attendance, a quiz score, the time spent on a task, the number of times a student revisits a learning resource, number of submissions, survey responses, or demographics (age, gender, program), can each function as a data point. On their own, individual data points may appear simple or even trivial. However, when combined and interpreted thoughtfully, they help reveal patterns about learning processes, engagement, and progress.

Data points are generated from multiple sources, including logs from LMS and digital learning platforms, assessment tools, classroom observations, surveys, and sensor data, and administrative records. Each data point typically contains specific attributes or features, such as a student identifier (like a student ID), demographic information, academic indicators (such as courses taken, test, scores, grades, and credits earned), behavioral metrics (including logins, a timestamp, attendance, participation rates), and interaction (such as platform interactions, discussion forum activity, and collaboration patterns). When combined, these attributes create rich, multidimensional datasets (Feng et al., 2022; Penteado et al., 2018). By analyzing multiple data points together, we can uncover important relationships, such as the link between student attendance and academic performance, or how time spent on tasks can predict skill mastery. Looking at aggregated data (like averages or distributions) and observing patterns over time (using time series data) can help identify trends and student progress.

Data points do not carry meaning in isolation. A score, click, or timestamp only becomes informative when interpreted in relation to learning objectives, pedagogical design, assessment criteria and the broader classroom context. The same value may signal progress in one setting but difficulty in another. Understanding what constitutes a data point, how it is generated, and how context shapes its meaning is therefore essential for learning analytics responsibly and effectively in schools. When interpreted carefully, such insights can meaningfully support teaching and learning (Feng et al., 2022; Simanjuntak, 2023).

3.1.2 Types of Data

In LA, data can be categorized in several ways. One approach is to consider their **functional role in the learning process**, such as demographic, interaction, performance, or psychometric data. Each category provides a different lens for understanding learners and their educational experiences. Data can also be grouped according to their **source within the educational ecosystem**. Common data sources include digital learning environments, register and student information systems, background surveys, and case-specific collections. A third approach is to classify data by their **measurement format**, most commonly as quantitative or qualitative. Understanding these different ways of categorizing educational data helps teachers and school leaders make informed decisions about what to collect and how to interpret results. In practice, effective LA considers all three perspectives—what the data represent, where they come from, and how they are expressed—to develop a balanced and meaningful understanding of teaching and learning.

Based on Functional Role

Demographic data describes students' background characteristics, such as age, gender, socio-economic status, and educational history, providing contextual information. These data are often used to explore patterns and relationships between students' socio-demographic factors and learning outcomes (Papadogiannis et al., 2024). For instance, researchers can examine how poverty rates influence access to educational resources or how students' background characteristics relate to levels of their performance (Chen & Chang, 2018). Such insights can inform the design of personalized learning approaches that strengthen connections between students and teachers.

Interaction data capture students' engagement with learning environments and educational technologies, including learning management systems (LMS), online courses, and educational software (Zhang et al., 2021). These data provide information about how long, how often, and in what ways students interact with digital learning environments (Papadogiannis et al., 2024). Analyzing interaction data helps educators understand how students actually engage with e-learning materials and activities, thereby informing the design of more effective instruction (Romero & Ventura, 2013). For instance, clickstream data (e.g., what did students click on, in what order?) or attendance logs (presence, absences, tardiness) from an LMS can reveal which resources students use most frequently, supporting more responsive and targeted content delivery.

Performance data include grades, test scores, other assessment results, and course completions that indicate levels of achievement and progress (Papadogiannis et al., 2024). These data help evaluate the effectiveness of instruction and identify areas where students may need additional support (Romero & Ventura, 2020). Performance data may be collected through traditional assessments, such as paper-and-pencil examinations and quizzes, as well as through real-time analytics embedded in online platforms (Romero & Ventura, 2013). Careful interpretation is essential, as high-quality and valid measures are needed to inform decision-making and to predict future achievement reliably.

Psychometric data focus on underlying psychological constructs, such as cognitive abilities, personality traits, motivation, self-efficacy, or emotional states (Papadogiannis et al., 2024). These data are often collected through

questionnaires, standardized instruments, or structured observations (Zhang et al., 2021). Psychometric data provides insight into the less visible factors that influence learning processes, such as students' confidence, motivation, or stress levels (Chen & Chang, 2018). When combined with other data types, they support the development of more comprehensive models of learner behaviour and educational processes.

A recent systematic review indicates that academic performance data are the most frequently used for EDM, followed by behavioral interaction and demographic data, among others (Choi et al., 2023). This suggests that while multiple data types are available, their integration remains an important area for further development.

Based on the source within the Educational Ecosystem

In LA, data can also be grouped according to their source within the educational ecosystem. From this perspective, four main categories are commonly identified: (1) Data from digital learning environments, (2) Register and student information system data, (3) Background survey data, and (4) Case-specific data (OKM, 2024, pp. 34–35):

Data from digital learning environments form a digital footprint of students' interactions with platforms and other users. These may include information about assignment submissions and feedback (automatic, teacher, peer, or self-assessment), time spent, materials and multimedia accessed, participation in discussion forums, and use of dashboards or activity visualizations. Digital learning environments typically generate large amounts of data about students' activity. For example, the ViLLE learning platform can collect millions of data points per student each year. These data include students' answers, correctness, time on task, number of attempts, and progress over time. The information is compiled and visualized in student dashboards that show individual progress and performance levels to support self-monitoring of learning. Teachers can view similar data through teacher dashboards, often in aggregated forms such as class averages or task comparisons. When used thoughtfully, these data help teachers identify learning patterns, monitor engagement, and adjust instruction to better support students.

Register and student information system (SIS) data typically include course enrollments, completed and failed courses, assessment dates, and accumulated credits. Some systems also allow students to create personal study plans, showing their intended study pace and how closely these plans are followed. Information about students' prior learning—such as previous credits, grades, or the type of school attended—helps educators understand how past experiences influence current progress and supports more targeted guidance and personalized learning. Access to such administrative data is often regulated, so permitted uses and protections must be clearly defined.

Background survey data may include demographic information, such as age, place of residence, mother tongue, prior education, or family background, as well as insights into students' motivation, attitudes toward learning, study strategies, and other factors related to academic success. For example, knowing students' learning difficulties or parents' support can help teachers understand why they may struggle with learning, while information about motivation and attitudes can guide personalized support, mentoring, or the design of targeted interventions to improve learning experience and outcomes.

Case-specific data are collected for particular teaching or research contexts and provide detailed insights into learning and instruction. Examples include classroom observations, student feedback on lessons or learning materials, and research measures such as eye-tracking or physiological responses. In addition, these data can capture students' interactions in other digital environments, such as social media, helping educators understand engagement, behavior patterns, and the effectiveness of specific learning activities.

LA often involves combining information from different sources and purposes. This fragmentation presents challenges for data integration, as systems may use different formats and operate under different access rules. The

source of the data also determines how it can be legally and ethically used. In addition, age and educational level must be considered, as not all data types are appropriate or permissible in early childhood education or compulsory schooling.

Table 2 provides examples of educational data commonly used in LA practices and research. The examples are grouped according to the aspect of the learner regarding social, educational, learning and behavioral dimensions. The table also indicates how each type of data can be collected and which stakeholder is responsible for providing or managing it. This helps schools see what data may be available and how it can be accessed in practice.

Co-funded by
the European Union

Table 2: Examples of educational data for LA, their nature, collection method, and stakeholders involved (TRILA, 2021, pp. 10–11).

Division	Parameters	Data	Method	Stakeholder
Social	Demographics	Race/Ethnicity, Residence, Gender, Age, Languages, Disability	Survey, Registry	Admins
	Social	Hobbies, Athletic status	Survey	Students
	Economic	Employment, Expectations, Debts, Tuition fees, Outlays	Survey	Students
	Personal	Interests, Expectations, Goals, Perceptions, Parental education, Relationship status	Survey	Students
	Interaction with Others	Meetings, E-mails, Text messages	LMS, Digital Platform	Students, Teachers, Admins
Educational	Academic Progress	Grades, Credits, Achievements, Feedback, Enrolments, Dropouts, Attendance, Academic history, Admission pathway, Course selection	LMS, Registry	Admins, Teachers
	Skill Acquisition	Knowledge gains, knowledge retention, knowledge transfer	LMS	Students
	Self-Regulation of Learning	Learning strategy, Time-to-task, Time-to-learn	LMS, Survey	Students
	Aptitudes / Abilities	Self-assessment, Self-efficacy, (Duckworth) Grit	LMS, Survey	Students
	Consultation of Resources	Views of videos, Webpage visits, Number of file downloads	LMS, Social Media	Students
Learning	Instructional Design	Pedagogical practices, Learning platform / tools	Survey	Teachers, Designers
	Artifacts	Essays, Notes, Sketches, Code, Slides	LMS	Students
	Assignments	Essays, Reports, Shared documents, Manuals	LMS	Students
	Digital platform	Log-in time, Log-in frequency, Frequency of use, Time spent on tasks, Frequency of resource use, Posts, Comments	LMS, Forum, Blog, Social Media	Students
	Group Work	Face-to-face interaction, Messages, E-mails	LMS, Motion Tracking Devices, Social Media	Students
	Course	Load, Performance, Resources, Peer-Evaluation	LMS, Registry	Teachers, Students
Behavioral	Natural Human Signals	Gaze, Gesture, Speech	Motion Tracking Devices	Students

	Psychometrics	Attitude, Personality traits, Academic motivation, Behavior, Affect, Confidence	Survey	Students
	Off-Task Actions	Texting, Talking, Drawing, Inattention	Motion Tracking Devices	Students

Based on Measurement Format

Quantitative data are numerical and can be measured or counted. Examples include test scores, number of logins, time spent on a task, completion rates, or frequency of participation in online discussions. Because these data are structured and comparable, they are particularly suited to EDM techniques such as pattern detection, prediction modelling, and trend analysis. Quantitative data allow teachers and schools to identify changes over time, compare groups, and monitor progress at scale.

Qualitative data, in contrast, are descriptive and provide insight into meaning, experience, and process that cannot easily be reduced to numbers. These may include written feedback, student reflections, open-ended survey responses, classroom observations, or recorded discussions. In LA, qualitative data help explain the "why" behind numerical patterns. For example, a decline in engagement metrics may only be fully understood by examining student comments or contextual classroom factors.

Both types of data play complementary roles. Quantitative data is powerful for identifying patterns and generating alerts, while qualitative data provide depth, interpretation, and pedagogical insight. Effective LA integrates both, ensuring that decisions are not driven solely by numbers but are grounded in a rich understanding of teaching and learning processes.

3.1.3 Levels of Data

Data can be collected and analyzed at different levels. Recognizing these levels helps clarify what kinds of questions can be asked, what methods are appropriate, and what types of decisions can be informed by the findings. From an **educational practice perspective**, data may be organized according to where they are generated and used: at the student level, teacher level, the course or module level, the institutional level, and the national or international level (OKM, 2024). This perspective reflects how teachers, school leaders, and policymakers typically engage with data in their daily work. From an **EDM perspective**, levels are often described as operating at micro-, meso-, and macro-level data (Papadogiannis et al., 2024). Considering both perspectives helps ensure that LA is aligned with practical needs while remaining methodologically sound.

Educational Practice Perspective

At the **student level**, data points describe individual learners' background, behaviors, misconceptions, performance, engagement, and progress. These data are often used to provide personalized feedback, identify support needs, tailor instruction, and report achievement.

At the **teacher level**, data describe both individual students and the class. These data, collected from learning environments and digital tools used, are analyzed to support teaching and learning. Analysis can reveal how well students meet learning objectives, participate in activities, use learning materials, and engage with feedback. Teacher-level data also help monitor individual students' progress over time, enabling personalized support, while providing insights into overall class performance to guide instructional planning and improve learning outcomes.

At the **course, module, or study program level**, data combines information from all students within a unit or across similar units taught by different instructors. This aggregation allows comparisons within and between courses, helping teachers or program heads evaluate instructional design, check alignment between learning objectives and assessments, and identify patterns such as low participation or common difficulties. These insights support targeted improvements in teaching and learning across the whole course or program.

At the **institutional level**, data relates to learners, educational programs, and the education provider as a whole. These data support evidence-informed quality assurance, planning, and management. They are typically drawn from

learning management systems, feedback, student information systems, and administrative records. Institutional-level data may include retention rates, achievement trends, progression patterns, and resource use. Such analyses help streamline educational pathways, monitor and predict degree completion, and identify students at risk of dropout. The insights gained can inform school improvement strategies, professional development planning, policy decisions, and the development of targeted support services, such as academic advising and study guidance.

At the **national and international levels**, data extend beyond individual institutions to provide system-wide perspectives on education. These data are often derived from large-scale assessments, national registers, and cross-system comparisons. They enable benchmarking between education providers, comparisons among institutions with similar admission criteria, and analyses of regional differences or performance against national standards. Such large-scale datasets support curriculum reform, strategic planning, and evidence-informed policymaking. Register-based data has a long tradition in many education systems and offers valuable comparative insights for national decision-making. However, ensuring high data quality and improving consistency across sources remain ongoing challenges at this level.

In EDM and LA, the level of analysis shapes both the methods used and the interpretation of results. Student-level data often enables fine-grained modelling and personalized insights, whereas higher-level data relies on aggregation and comparative analysis. Importantly, patterns observed at one level do not automatically explain phenomena at another. Responsible use of LA, therefore requires careful consideration of the appropriate level of data for the intended purpose.

EDM Perspective

Micro-level data are generated through direct interactions between students and digital learning environments like Moodle and educational apps. These data capture fine-grained learner actions (e.g., clicks, attempts, response times, and sequences of activity) within specific contexts. Such detail enables timely feedback and adaptive support, for example, by adjusting task difficulty or providing targeted hints. Micro-level data are typically validated through real-time system logs or coding procedures (Papadogiannis et al., 2024). Analytical approaches such as Bayesian Knowledge Tracing and Performance Factor Analysis are commonly used to estimate students' knowledge and predict learning outcomes (Pham Kim, 2017).

Meso-level data are primarily derived from student-generated texts in learning management systems (LMS) or social media platforms. Using Natural Language Processing (NLP), researchers and practitioners can examine cognitive, behavioral, emotional, and social dimensions of learning (Papadogiannis et al., 2024). Applications include, for example, automated feedback and grading, enhancement of the learning experience, and improvements in course design. However, the reliability of analytical tools and the influence of contextual factors remain important considerations (Baker & Yacef, 2009).

Macro-level data are typically collected over extended periods and include demographic information, enrolment records, and academic histories (Papadogiannis et al., 2024). These data are mainly used to inform institutional planning and policy decisions. For example, they support early warning systems that identify students at risk of dropout, as well as guidance systems that recommend suitable courses or learning pathways (Ferreira-Mello et al., 2019; Chaturapruek et al. 2021). Macro-level analyses also help administrative personnel evaluate curriculum effectiveness and identify broader patterns of student progression and success (Papadogiannis et al., 2021).

Taken together, these levels form a hierarchical structure that enables analysis ranging from detailed learner interactions to system-wide trends, offering a more comprehensive understanding of learning processes (Bousbia & Belamri, 2014). However, the volume and complexity of data require careful preparation and management (Baker,

2015). While different data sources provide valuable opportunities for meaningful insight, they also necessitate robust pre-processing and ethical data governance (Papadogiannis et al., 2024). Data must be systematically collected, cleaned, and anonymized to ensure accuracy, reliability, and the protection of privacy.

3.2 Data Structures for Educational Data Mining

Data structures refer to how data are organized and stored within a computer system so that they can be accessed, managed, and analyzed efficiently. A data structure can be understood as a collection of data values, the relationships among those values, and the operations that can be performed on them (Wegner & Reilly, 2003). In other words, it defines not only what data are stored, but also how they are connected and how they can be used.

In the context of EDM and LA, data structures play a crucial role in enabling effective analysis. They determine how data points relate to one another, for example, how individual student records are linked to courses, assessments, or time stamps. Well-designed data structures support the efficient implementation of data mining algorithms and help ensure that analyses are accurate and meaningful.

Different data structures are designed to organize information in ways that support particular types of tasks and analyses. While teachers and school leaders are not expected to implement these structures themselves, understanding their basic logic helps clarify how educational data are stored and analyzed. Common examples include (Seymour, 2014):

Arrays and lists are among the simplest data structures. An array stores elements in a specific order, typically of the same type (for example, a list of test scores), whereas a list can contain elements of different types in a linear sequence. These structures are well-suited to managing collections of information such as student records, grades, attendance data, or other performance indicators (Baek & Doleck, 2022; Pei, 2017).

Trees represent hierarchical relationships. They are particularly useful in education for modelling structured dependencies, such as course prerequisites, curriculum frameworks, or dependency charts for skills acquisition (Amriza et al., 2025; Simionescu et al., 2024). Tree structures enable efficient searching, sorting, and hierarchical representation of data, supporting analyses that explore relationships between learning inputs and outcomes (Cai & Li, 2024; Chen, 2022).

Graphs are used to model the relationships between complex data points. In educational contexts, graphs can represent patterns of student collaboration, social interactions, or conceptual connections within knowledge domains (Bravo et al., 2018; Tavares et al., 2017). By mapping how elements are interconnected, graph-based analyses help visually reveal interdependencies in learning behaviors and knowledge development.

Hash tables are designed for fast retrieval of information using keys (such as a student identification number). They allow rapid access to stored data and are widely used in database systems and data preparation processes. In adaptive learning environments, hash tables can support real-time access and modification of student data (Özdağoğlu et al., 2018; Wang, 2021).

Overall, data structures provide the foundation for efficient data handling and analysis. As educational data continues to grow in volume and complexity, robust data structures are essential for supporting advanced data mining techniques (Baek & Doleck, 2022). By enabling effective storage, organization, and retrieval of information, these structures strengthen the analytical capacity of educational data systems and, ultimately, the insights available to educators.

Key Questions about LA

(OKM, 2024, pp. 26–27, 93–94)

- What data are used in LA?

LA uses data from several sources. These include information generated in digital learning environments (such as learning platforms), administrative or student record systems, background surveys, and data collected for specific teaching or research purposes. Because these data come from different systems, combining them can be challenging. Each system has its own technical limits and rules about how data are stored and accessed. For this reason, decisions about what personal data are collected, how long they are kept and accessed, and how they are used should always be carefully planned in relation to educational goals. The source of the data also affects what it can be used for, so ethical and legal considerations must guide all data use in schools.

- How do pseudonymisation and anonymisation affect the use of LA?

Pseudonymised data (for example, replacing names with ID numbers) is still personal data, since individuals can be identified using additional information. Simply removing names does not remove data protection responsibilities. **Anonymised data**, however, cannot be linked back to individuals and is no longer covered by data protection laws. When planning LA, schools should consider whether anonymised data would be sufficient for their goals. This is often the safest option. It is important to remember that anonymisation must be irreversible, and that the act of anonymising data is itself a form of data processing. Any data use should always align with the original purpose for which the data were collected.

- Who owns the data used in LA?

In LA, the question of data ownership sometimes arises. Legally, information itself cannot usually be "owned" in the same way as physical property. The same data can be held by multiple parties at the same time, and it can be copied and shared easily. Unlike a physical object, such as a computer, which can only be in one person's possession at a time, data can be duplicated and distributed without limit.

Although data cannot typically be owned in a traditional sense, its use may still be restricted. For example, some data may be protected as trade secrets or subject to copyright. In LA, such restrictions are relatively rare and more often apply to the tools or software used, rather than to the learning data itself.

References

Amriza, R., Chou, T., & Ratnasari, W. (2025). Beyond the Classroom: Understanding the Evolution of Educational Data Mining with Key Route Main Path Analysis. Computer Applications in Engineering Education, 33(2). https://doi.org/10.1002/cae.70010

Ang, L., Ge, F., & Seng, K. (2020). Big Educational Data & Analytics: Survey, Architecture and Challenges. *IEEE Access, 8*, 116392-116414. https://doi.org/10.1109/access.2020.2994561

Baek, C. and Doleck, T. (2022). Educational Data Mining: A Bibliometric Analysis of an Emerging Field. *IEEE Access, 10*, 31289–31296. https://doi.org/10.1109/access.2022.3160457

Baker, R.S. (2015). *Big Data and Education*, 2nd ed.; Teachers College, Columbia University: New York, NY, USA.

Baker, R. S. J. D., & Yacef, K. (2009). The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining, 1*(1), 3.

Bousbia, N.; Belamri, I. (2014). Which Contribution Does EDM Provide to Computer-Based Learning Environments? In *Studies in Computational Intelligence. Educational Data Mining*; Springer: Cham, Switzerland, pp. 3–28.

Bravo, J., Bonilla, C., & Seoane, I. (2018). Data mining in foreign language learning. *Wiley Interdisciplinary Reviews Data Mining and Knowledge Discovery, 10*(1). https://doi.org/10.1002/widm.1287

Cai, J. and Li, Y. (2024). Fuzzy association rule mining for Personalized English Language Teaching from higher education. *Journal of Computational Methods in Sciences and Engineering, 24*(6), 3617–3631. https://doi.org/10.1177/14727978241296748

Chaturapruek, S., Dalberg, T., Thompson, M. E., Giebel, S., Harrison, M. H., Johari, R., Stevens, M. L., & Kizilcec, R. F. (2021). Studying Undergraduate Course Consideration at Scale. *AERA Open, 7*(1). https://doi.org/10.1177/2332858421991148

Chen, Y. (2022). Quality Evaluation of Student Education Management Work Based on Wireless Network Data Mining. *Mathematical Problems in Engineering, 2022*, 1–12. https://doi.org/10.1155/2022/9182420

Chen, Y., & Chang, H.-H. (2018). Psychometrics Help Learning: From Assessment to Learning. *Applied Psychological Measurement, 42*(1), 3–4. https://doi.org/10.1177/0146621617730393

Choi, W.-C., Lam, C.-T., & Mendes, A.J. (2023). A systematic literature review on performance prediction in learning programming using educational data mining. *In Proceedings of the 2023 IEEE Frontiers in Education Conference (FIE)*, USA, pp. 1–9.

Feng, G., Fan, M., & Chen, Y. (2022). Analysis and Prediction of Students' Academic Performance Based on Educational Data Mining. *IEEE Access, 10*, 19558–19571. https://doi.org/10.1109/access.2022.3151652

Ferreira-Mello, R., André, M., Pinheiro, A., Costa, E., & Romero, C. (2019). Text mining in education. *Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery, 9*(6), e1332-n/a. https://doi.org/10.1002/widm.1332

Ministry of Education and Culture, Finland (OKM), Learning Analytics Division. (2024). *Framework for Learning Analytics: Best practices in the implementation and utilisation of learning analytics*. https://urn.fi/URN:ISBN:978-952-263-731-4

Özdağoğlu, G., Öztaş, G., & Çağlıyangil, M. (2018). An application framework for mining online learning processes through event-logs. *Business Process Management Journal, 25*(5), 860-886. https://doi.org/10.1108/bpmj-10-2017-0279

Papadogiannis, I., Wallace, M., Poulopoulos, V., Karountzou, G., & Ekonomopoulos, D. (2021). A First Ever Look into Greece's Vast Educational Data: Interesting Findings and Policy Implications. *Education Sciences, 11*(9), 489. https://doi.org/10.3390/educsci11090489

Papadogiannis, I., Wallace, M., & Karountzou, G. (2024). Educational Data Mining: A Foundational Overview. *Encyclopedia, 4*(4), 1644–1664. https://doi.org/10.3390/encyclopedia4040108

Pei, Z. (2017). Educational Data Mining for Teaching and Learning. *DEStech Transactions on Social Science Education and Human Science*, (iced). https://doi.org/10.12783/dtssehs/iced2017/15101

Penteado, B., Paiva, P., Morettin-Zupelari, M., Isotani, S., & Ferrari, D. (2018). Toward Better Outcomes in Audiology Distance Education: An Educational Data Mining Approach. *American Journal of Audiology, 27*(3S), 513–525. https://doi.org/10.1044/2018_aja-imia3-18-0020

Pham Kim, C. (2017). Evaluating Student Teachers in Micro-Teaching with Analysis of Video Recording Lesson by Boris Software at Vietnam National University. *Scientific Publishing Center: Sociosphere, 8*, 67–74.

Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery*, *3*(1), 12–27. https://doi.org/10.1002/widm.1075

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery*, *10*(3), e1355-n/a. https://doi.org/10.1002/widm.1355

Seymour, L. (2014). *Data structures* (Revised first ed.). New Delhi, India: McGraw Hill Education.

Simanjuntak, R. (2023). Model of Education Technology for Language Pedagogy in Higher Education. *E3S Web of Conferences, 426*, 02044. https://doi.org/10.1051/e3sconf/202342602044

Simionescu, C., Danubianu, M., Grădinaru, B., & Măciucă, M. (2024). Educational Data Mining in European Union – Achievements and Challenges: A Systematic Literature Review. *International Journal of Advanced Computer Science and Applications, 15*(3). https://doi.org/10.14569/ijacsa.2024.0150386

Tavares, R., Vieira, R., & Pedro, L. (2017). A preliminary proposal of a conceptual educational data mining framework for science education: Scientific competences development and self-regulated learning. *SIIE: 2017 International Symposium on Computers in Education: Lisbon (Polytechnic Institute of Lisbon, School of Education), Portugal*, pp. 1–6. https://doi.org/10.1109/SIIE.2017.8259644

Turku Research Institute for Learning Analytics (TRILA). (2021). *A Pocket-Sized Handbook for Learning Analytics*. https://projects.tuni.fi/app/uploads/2021/10/a607d223-oppimisanalytiikan-kasikirja-apoa.pdf

Wang, W. (2021). Model Construction and Research on Decision Support System for Education Management Based on Data Mining. *Computational Intelligence and Neuroscience, 2021*(1). https://doi.org/10.1155/2021/9056947

Wegner, P., & Reilly, E.D. (2003). *Encyclopedia of Computer Science*. Chichester, UK: John Wiley and Sons. pp. 507–512.

Zhang, Y., Yun, Y., An, R., Cui, J., Dai, H., & Shang, X. (2021). Educational Data Mining Techniques for Student Performance Prediction: Method Review and Comparison Analysis. *Frontiers in Psychology*, *12*, 698490. https://doi.org/10.3389/fpsyg.2021.698490

4 Data Storage and Data Harmonization for Educational Data Mining

The rapid growth of digital learning environments, mobile technologies, and information and communication technologies (ICT) has significantly increased both the volume and diversity of educational data generated by students, teachers, and administrative systems (Ang et al., 2020; Gul et al., 2021; Moscoso-Zea et al., 2019). To make meaningful use of these data in Educational Data Mining (EDM) and Learning Analytics (LA), robust data storage and effective data harmonization are essential (Hamoud et al., 2021; Koliçi et al., 2014; Moscoso-Zea et al., 2019). After collecting, data must be securely stored to ensure data storage's quality, organization, and controlled access. Before analysis, data also needs to be organized, structured, and processed appropriately. Because educational data often comes from multiple and heterogeneous sources, they must be harmonized into consistent formats, definitions, and structures to enable reliable analysis.

This chapter outlines the key principles of data storage and harmonization in EDM and LA. It provides an overview of core concepts, their role in EDM applications, and discusses current challenges and future directions.

4.1 The Nature and Sources of Educational Data

Diversity of Data Sources

Educational data are generated across a wide range of digital systems used in teaching and learning. These include learning management systems (LMS), Massive Open Online Courses (MOOCs), learning object repositories, open educational resources (OER), social media platforms, linked data repositories, and mobile learning environments (Ang et al., 2020). Each system produces data in different formats and structures, with varying storage and access requirements (Buenaño-Fernández et al., 2019). In practice, a single learning environment often cannot accommodate all learning activities. For example, students may complete exercises in an LMS, provide feedback through a separate survey platform, and take assessments in another system. As a result, data are collected from multiple sources, requiring harmonization to create a complete picture of learning.

For schools and teachers, this means that educational data do not come in one consistent form. For example, LMS platforms typically generate structured data such as login records, activity logs, assessment results, and course completion information, while social media or discussion platforms may produce unstructured data, such as text posts, images, or videos (Ang et al., 2020; Jittawiriyanukoon, 2019; Liu & Yu, 2022). Large-scale virtual campuses and MOOCs can generate substantial volumes of log data on a daily basis.

This diversity creates practical challenges for data storage and use. Any system designed to support EDM or LA must be able to store, organize, and integrate data from multiple sources (Ang et al., 2020; Moscoso-Zea et al., 2019). In addition to large volumes, differences in file formats, data structures, naming conventions, and levels of detail make data harmonization complex (Ang et al., 2020; Daniel, 2014; Koliçi et al., 2014). Before analysis can take place, data often needs to be cleaned, standardized, and aligned so that information from different systems can be meaningfully combined and compared.

Volume, Velocity, and Variety of Data

In the context of big data, educational data are often described through three key characteristics: volume, velocity, and variety. Together, these features create practical challenges for data storage and harmonization.

Volume refers to the large amount of data generated in educational settings. Every login, quiz attempt, assignment submission, click, and feedback comment adds to the total. With the increasing use of digital platforms, mobile devices, and connected technologies, schools and institutions now collect far more data than traditional storage

systems were originally designed to handle (Ang et al., 2020; Moscoso-Zea et al., 2019). This requires scalable storage solutions that can manage growing datasets reliably.

Velocity refers to the speed at which data are generated. In online and blended learning environments, information is produced continuously, for example, when students participate in live quizzes, submit assignments, or interact in discussion forums. To be useful for timely feedback or early intervention, data systems must be able to process and update information quickly (Ang et al., 2020; Koliçi et al., 2014; Liu & Yu, 2022).

Variety refers to differences in data formats and meanings. Educational data may include numerical scores, written responses, attendance records, system logs, images, or videos. In addition to format differences (structural heterogeneity), the same concept may be defined differently across systems (semantic heterogeneity). For example, "student performance" might be recorded as percentages in one system, grades in another, and competency levels elsewhere (Buenaño-Fernández et al., 2019; Obilikwu & Ogbuju, 2020).

When schools use multiple platforms that store data separately, comparing or combining information becomes more complex (Buenaño-Fernández et al., 2019; Obilikwu & Ogbuju, 2020). For LA and EDM to work effectively, these differences must be addressed through harmonization (aligning formats, definitions, and structures) so that multiple data from different systems can be meaningfully analyzed together (Ang et al., 2020; Obilikwu & Ogbuju, 2020).

4.2 Data Storage for Educational Data Mining

Educational Data Warehouses

An **Educational Data Warehouse (EDW)** is a central repository where data from different sources within a school or institution are collected, stored, and organized consistently (Hamoud et al., 2021; Moscoso-Zea et al., 2019). Unlike traditional transactional databases, which simply record day-to-day operations, EDW is designed to support analysis and decision-making. By consolidating data from multiple systems—such as LMS platforms, student information systems, assessment tools, and surveys—it creates a unified view of educational data, enabling data mining and analysis (Buenaño-Fernández et al., 2019; Moscoso-Zea et al., 2019).

Schools and institutions often maintain distributed information systems, which can make LA and EDM difficult. The EDW addresses this by providing a consistent, centralized data environment, essential for effective data-driven decision-making (Buenaño-Fernández et al., 2019). For example, grades stored in an LMS, attendance records from a student information system, and participation in discussion forums can all be integrated into the EDW.

A key feature of EDWs is **Online Analytical Processing (OLAP)**. OLAP allows users to analyze data across multiple dimensions—such as time, course, student demographics, and performance—helping educators gain nuanced insights and make informed, holistic decisions (Hamoud et al., 2021; Moscoso-Zea et al., 2019).

In short, EDWs provide robust, organized, accessible, and reliable storage for educational data, supporting meaningful insights and informed decision-making at both classroom and institutional levels.

Data Marts

A **data mart** is a smaller, focused subset of a larger Educational Data Warehouse (EDW) that is designed to meet the specific analytical needs of a department, team, or functional area (Hamoud et al., 2021). While an EDW consolidates all data across an institution, a data mart provides a simpler, more manageable and cost-effective alternative, making it easier and faster to analyze specific types of information.

In schools and institutions, data marts are often used to track and analyze student academic performance. For example, a data mart might focus solely on assessment scores, attendance records, and course completion for a

particular grade or subject. By narrowing the scope, teachers and school leaders can quickly access relevant data without navigating the entire EDW.

Data marts often use a **star schema**, where multiple dimension tables—such as students, courses, and time periods—connect to a central fact table containing the main measures, like grades or participation counts (Hamoud et al., 2021). The **staging table** serves as temporary storage during the ETL (Extract, Transform, Load) process, holding data before it is loaded into the data mart (Hamoud et al., 2021).

While data marts usually store summary information, they can also be linked to operational systems to access detailed records when needed. This balance of focus and flexibility makes data marts a practical tool for schools and institutions to perform targeted analyses and support informed decision-making.

Hybrid Infrastructure

No single data storage system can meet all the diverse needs of educational data analysis. To address this, **hybrid infrastructures** have been developed to combine multiple storage approaches, such as Educational Data Warehouses (EDWs), data marts, and other repositories, into a single integrated system (Fernández et al., 2014; Moscoso-Zea et al., 2019).

A hybrid infrastructure allows schools and institutions to store and manage both detailed transactional data and aggregated, analytical data in one environment. For example, an EDW can consolidate student performance, attendance, and LMS interaction data, while a linked data mart provides teachers with a simplified view of grades and engagement for a particular class. This combination supports both broad institutional analysis and targeted, department-level insights.

Hybrid systems also enable advanced analytics and experimentation, such as using predictive models or multidimensional analysis of EDM and LA, to explore student progress, course effectiveness, or engagement patterns. By integrating multiple storage technologies, hybrid infrastructures make it easier to visualize and analyze information across different parts of the school or institution, from classroom activities to administrative processes (Moscoso-Zea et al., 2019).

In practice, hybrid infrastructures provide flexibility and scalability, allowing educational organizations to combine analytical power with practical, accessible insights for teachers, school leaders, and administrators.

Large-Scale Data Storage

The rise of big educational data has made traditional storage systems insufficient. Modern educational platforms generate massive amounts of data—from assessment scores and LMS interactions to discussion forums and social media activity—at high speed and in varied formats (Ang et al., 2020; Liu & Yu, 2022). To manage this, schools and institutions increasingly rely on distributed storage and processing frameworks that can handle large, fast, and diverse datasets.

Distributed architecture typically includes multiple layers: data sources, processing frameworks, data warehouses, analytics tools, and reporting systems (Ang et al., 2020). Open-source solutions like **Apache Hadoop** are widely used for big educational data. Its distributed file system (HDFS) allows massive datasets to be stored safely and processed efficiently across multiple servers (Liu & Yu, 2022). Tools like **Apache Sqoop** can transfer structured data from traditional databases into Hadoop for further analysis (Liu & Yu, 2022).

Non-relational (NoSQL) databases complement these frameworks by handling multiple data structures, making them ideal for the heterogeneous nature of educational data, such as combining assessment records, video

interactions, and social media content (Ang et al., 2020; Liu & Yu, 2022). **Relational SQL databases** can still be used for classical, structured data, such as student enrollment or grades.

Cloud-Based Data Storage

Cloud computing has become a natural platform for storing and analyzing large-scale educational data. **Cloud storage** is flexible, scalable, and cost-effective, making it easier for educational institutions to manage growing datasets and computational needs (Fernández et al., 2014; Liu & Yu, 2022). As a large-scale distributed storage system, cloud storage is increasingly used in educational settings, with learning resources and student data stored on remote servers. Services such as **Amazon S3, Google Cloud Storage, and Microsoft Azure** provide robust storage for MOOCs, LMS platforms, and other educational systems, with reliable backup and disaster recovery options. By combining cloud storage with distributed frameworks and harmonized data pipelines, schools can efficiently manage, analyze, and extract insights from vast educational datasets.

In short, cloud-based and distributed storage approaches provide the scalability, flexibility, and reliability needed for modern EDM and LA, ensuring that large-scale educational data can be securely stored, harmonized, and used to inform teaching and institutional decision-making.

4.3 Data Harmonization for Educational Data Mining

4.3.1 What is Data Harmonisation?

In EDM, **data harmonization** refers to the process of making heterogeneous data from different systems consistent and comparable so they can be analyzed together (Buenaño-Fernández et al., 2019; Obilikwu & Ogbuju, 2020). In practice, this means aligning differences in formats, structures, definitions, and data quality across multiple educational data sources.

For example, one school system may record grades as percentages, another as letter grades, and a third as competency levels. Before these data can be compared, such as when examining achievement across several schools, they must be converted into a shared format and interpreted using common definitions. Similarly, student information stored in separate platforms (e.g., LMS, student information systems, assessment tools) must be organized so that records referring to the same learner or course can be linked accurately.

Data harmonization typically includes practical steps such as combining data from different sources into a single repository (integration), converting them into common formats and schemas (standardization and transformation), correcting errors or inconsistencies (cleaning), and agreeing on shared definitions or data models (Al-Yadumi et al., 2021; Obilikwu & Ogbuju, 2020).

Without harmonization, data remain fragmented and difficult to interpret. This limits the reliability of analyses and weakens the insights that EDM and LA can provide (Moscoso-Zea et al., 2019; Buenaño-Fernández et al., 2019). Harmonization therefore forms a critical foundation for meaningful use of educational data.

4.3.2 Core Processes Behind Data Harmonization

Extract, Transform, Load (ETL) Processes

The quality of the ETL process directly determines the quality of the data available for EDM, making it a foundational concern for any educational data harmonization (Buenaño-Fernández et al., 2019; Moscoso-Zea et al., 2019).

A key process in data harmonization is known as **Extract, Transform, Load (ETL)**. In simple terms, ETL describes how data are collected from different systems, cleaned and organized, and then stored in a shared location for analysis (Al-Yadumi et al., 2021; Hamoud et al., 2021; Moscoso-Zea et al., 2019). The process has three main steps:

- **Extract** – Data are gathered from various sources, such as learning management systems (LMS), student information systems, or assessment tools.
- **Transform** – The data are cleaned and converted into a consistent format. This may involve correcting errors, aligning grading scales, standardizing dates, or ensuring that terms, such as "course completion" are defined in the same way across systems. The transformation stage is especially important because this is where most harmonization takes place. If data are not cleaned and aligned properly, analyses may be inaccurate or misleading.
- **Load** – The prepared data are stored in a central database or data warehouse, where they are ready for analysis.

In practice, schools and institutions sometimes focus on collecting data but give less attention to organizing and preparing them for analysis (Moscoso-Zea et al., 2019). However, the quality of the ETL process directly affects the quality of the data used in EDM. If data are not carefully extracted, cleaned, and standardized, the results of the analysis may be incomplete or misleading. Investing in these processes ensures that the insights generated are trustworthy and useful for decision-making.

Data Preprocessing and Cleaning

Before educational data can be analyzed, they must be carefully prepared. **Data preprocessing and cleaning** involve checking and improving raw data so they are accurate, consistent, and ready for use (Moscoso-Zea et al., 2019; Ramaswami et al., 2019). In practice, raw data from different systems are often incomplete or inconsistent. For example, some student records may be missing assessment results, dates may appear in different formats, or the same student may be listed twice under slightly different names. Effective preprocessing includes:

- filling in or addressing missing values,
- removing duplicate records,
- standardizing formats, such as dates, student ID, or grading scales ("A/B/C" vs. "85/75/65"), and
- resolving inconsistencies between systems (Moscoso-Zea et al., 2019; Ramaswami et al., 2019).

Another common task is aligning data labels and categories. For instance, older datasets may use different field names or split information across several columns that now need to be combined. To compare data across years or platforms, these elements must be mapped into a shared structure.

High-quality preprocessing is essential because analytical results depend directly on the quality of the underlying data. Predictive models used to identify students at risk of underachievement, for example, are only as reliable as the data on which they are built (Ramaswami et al., 2019). Combining both historical records and real-time engagement data can further improve insights—but only if the data have been properly cleaned and harmonized. In short, careful data preprocessing ensures that analyses are trustworthy and that decisions based on educational data are well-founded.

Schema Mapping and Data Reduction

When data comes from different systems, they are often organized in different ways. **Schema mapping** is the process of aligning these different structures so that information can be combined meaningfully (Al-Yadumi et al., 2021; Obilikwu & Ogbuju, 2020). This process requires both technical expertise and domain knowledge to ensure that the semantic meaning of data elements is preserved during transformation (Obilikwu & Ogbuju, 2020).

For example, one system may store "Student ID" in a column called *SID*, another may use *Learner_Number*, and a third may separate first and last names instead of using a single full name field; "last_login" in one system and "recent_access" in another should be treated as the same concept. Schema mapping ensures that these fields are correctly matched so that records referring to the same student can be linked accurately. This requires not only technical alignment but also an understanding of what each data element actually means. If definitions differ, for instance, if "course completion" includes different criteria in different systems, these differences must be clarified before merging the data.

Data reduction involves simplifying datasets while preserving the information that is most important for analysis. By reducing unnecessary detail, schools can lower storage demands and make data easier to manage and interpret (Moscoso-Zea et al., 2019).

In practice, the same information may be stored in more than one system. For example, grades might appear in both a learning management system (LMS) and a student information system (SIS). Data reduction in this case means merging these records into a single, consistent version and removing duplicates. It may also involve summarizing detailed logs, such as hundreds of individual clicks, into clearer indicators, such as weekly participation levels. This

makes data easier to analyze while preserving key patterns. When carried out carefully, schema mapping and data reduction together help ensure that datasets remain accurate, manageable, and suitable for meaningful analysis.

Batch Loading and Automated Harmonization

When large amounts of educational data need to be transferred from one system to another, this is often done through **batch loading**. Instead of moving records one by one, data is transferred in bulk at scheduled times, for example, nightly or weekly, from source systems into a central database (Sun, 2011). This approach is especially useful for schools or institutions that manage large datasets across multiple platforms.

Batch loading supports automation. Once set up, the process can regularly collect and harmonize data without requiring constant manual work. This improves efficiency and makes it possible to handle growing data volumes more reliably (Sun, 2011). However, automation does not remove the need for careful quality checks. Errors such as incorrect labels, inconsistent terminology, or spelling mistakes can be transferred along with the data. If not automatically detected, they may lead to ineffective searches and unreliable analytical results (Sun, 2011). For example, if the same course is labelled differently in two systems, an automated process may treat them as separate courses unless rules are in place to detect and correct such inconsistencies.

For this reason, robust automated harmonization including built-in validation steps that check for errors and inconsistencies before data are finalized is needed (Moscoso-Zea et al., 2019; Sun, 2011). When designed carefully, batch loading helps ensure that large datasets remain accurate, consistent, and ready for meaningful analysis.

4.3.3 Using Shared Data Standards

LA often relies on data collected from many different digital sources, such as learning management systems (LMS), student information systems (SIS), assessment tools, and AI-based platforms. Because these systems can collect, process, and store data in different ways, combining and analyzing information can be challenging. Data may not have been designed for analytics, and variations between IT systems add further complexity. To overcome these challenges, common standards and frameworks that define how data should be organized and described are essential for making systems interoperable (Obilikwu & Ogbuju, 2020; Thajchayapong et al., 2025).

What are shared standards and why do they matter?

Shared standards are agreed upon rules and practices for how data is structured, described, and shared across systems. They act like a common template, allowing different platforms to "speak the same language", making it easier to integrate data from multiple platforms. For example, if one school tracks student performance or course completion differently from another, standards ensure these data points can still be cleaned, reformatted, and aligned to fit this shared structure before meaningful analysis (Obilikwu & Ogbuju, 2020).

Other sectors, such as healthcare, have successfully used common data models to combine large datasets from multiple sources into unified repositories (Obilikwu & Ogbuju, 2020). Education is increasingly moving in the same direction. Several widely used standards support LA:

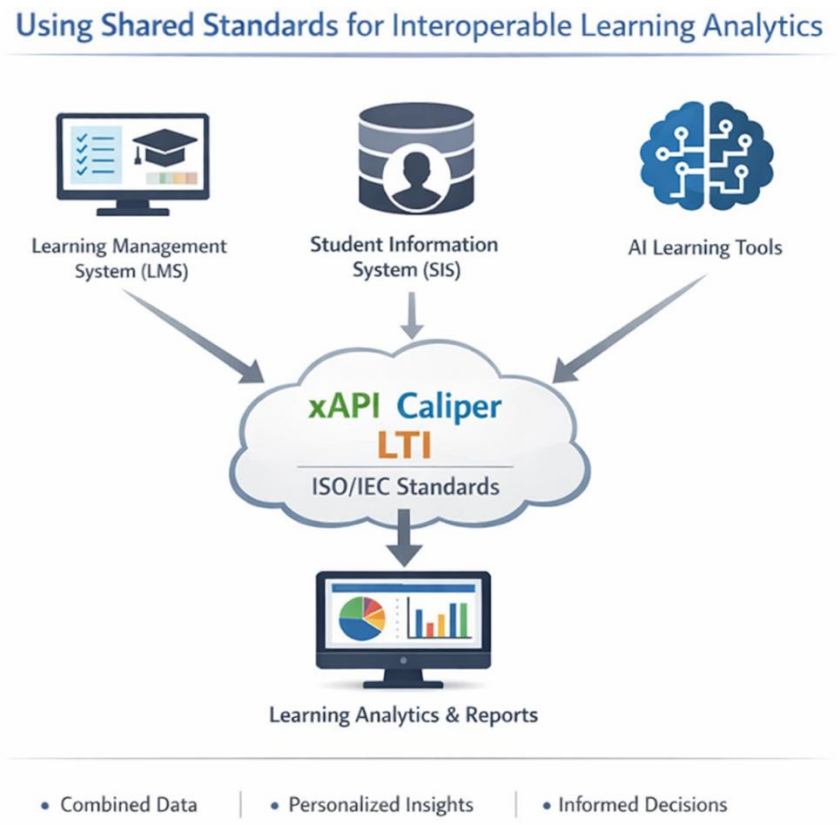
- **IEEE xAPI (Experience API):** Defines a common format for a wide range of learning activities across platforms, including mobile apps, virtual labs, offline activities, and more—not just traditional e-learning, enabling different systems to communicate and exchange information consistently.
- **IMS Caliper and LTI (Learning Tools Interoperability):** Enable seamless data transfer between learning management systems (LMS), digital textbooks, and other tools, making it easier for schools to add new digital tools without complicated custom integration or data silos.

- **ISO/IEC 20748 series**, including 1:2016, 1:2027, 3:2020, and 4:2029: A comprehensive framework for different aspects of LA. It also brings together practical guidance on how relevant standards can be applied effectively in LA systems to ensure system interoperability, data management, and privacy.

These standards provide common definitions of what student information is collected, how learning content is described (metadata), and how systems communicate, so that data can move securely and consistently between systems (Thajchayapong et al., 2025). For example, metadata might describe the skill level a lesson targets, the curriculum goal it addresses, or how long it typically takes to complete. Using shared data standards supports interoperability, reduces duplication, and strengthens the reliability of LA and EDM.

Figure 10 shows how shared data standards (xAPI, Caliper, LTI, ISO/IEC) enable different educational systems—LMS, SIS, and AI tools—to feed into a central LA platform for combined data, personalized insights, and informed decision-making.

Figure 10: Shared data standards enable LMS, SIS, and AI tools to feed harmonized data into a central learning analytics platform (AI-generated image).



Practical benefits for schools and institutions

Adopting shared standards helps:

- Combine data from multiple sources while keeping student identities and roles consistent.
- Ensure teachers, students, and administrators see the correct information and have appropriate access rights.
- Support analytics tools that monitor learning progress, personalise instruction, and inform evidence-based decisions.

While using shared standards makes learning data more reliable, comparable, and actionable, supporting better teaching, learning, and policy decisions across schools and platforms, in practice, establishing shared data standards is challenging because educational platforms and tools are offered by many different vendors. Achieving harmonization requires collaboration between standardization bodies, government agencies, and software providers. The goal is not strict agreement, but a practical consensus on common formats or data structures, such as a digital course template, that can be recognized and used across multiple systems.

Many countries are working to develop shared standards for educational data to improve consistency and enable comparisons across institutions. Organizations like the UK's JISC provide practical examples of interoperable LA architectures that follow xAPI, IMS, and ISO standards. Their approach integrates technical interoperability with ethical guidance and best practices, offering a model for schools and institutions planning their own systems.

In Finland, a national education agency working group is preparing recommendations on how schools should collect and structure data. While not yet formal standards, these guidelines aim to create a common framework for reliable

data use. Moreover, every Finnish student is assigned a unique, long numeric ID that reveals nothing about the individual but allows data from multiple registries to be reliably linked—a major advantage for large-scale research. In Slovenia, a similar unique student ID will be introduced across all educational levels by 2027, simplifying tracking, integrating records across schools, and supporting consistent educational data analysis.

4.4 Applying Data Storage and Harmonization in Educational Data Mining

Predicting Student Performance Using Educational Data

One of the most common applications of EDM is predicting student academic performance. Accurate predictions require access to comprehensive, well-organized, and harmonized data stored in repositories such as an EDW or data marts (Gul et al., 2021; Ramaswami et al., 2019). These repositories consolidate information from multiple sources—such as LMS activity logs, assessment results, attendance records, and survey responses—providing a unified view of each student's learning journey.

Quality and completeness of data available in the storage are critical for these predictions (Ramaswami et al., 2019). For example, student engagement data collected from online quizzes, discussion forums, and self-paced activities can be combined with historical grades to improve the accuracy of predictive models. Harmonized data ensures that information from different sources—such as LMS platforms and student information systems—is consistent and comparable, making it possible to identify patterns and trends across students, courses, and even academic years.

Real-world applications include predicting which students may need additional support or which learning interventions are most effective. For instance, a teacher might use predictive insights to identify students at risk of falling behind and offer targeted feedback or resources. The longitudinal value of harmonized data also allows schools to track performance over time, helping administrators accurately plan curriculum improvements and support strategies based on evidence from multiple cohorts (Ang et al., 2020; Ramaswami et al., 2019).

Using Data to Support Educational Decisions

Educational data storage and harmonization are not only useful for analysis but also for supporting decision-making in schools and institutions. Data-driven decision support systems (DSS) leverage structured repositories, such as educational data marts or warehouses, to provide timely and actionable insights for academic planning and management (Hamoud et al., 2021).

For example, a DSS can consolidate data on student performance, attendance, and engagement from multiple systems and present it through dashboards with Key Performance Indicators (KPIs). These KPIs might include average grades by class, course completion rates, or participation in online learning activities, giving educators and administrators a clear picture of how students and programs are performing.

Evaluation studies have shown that DSS based on harmonized data marts, combined with KPIs and OLAP tools, can effectively support both short-term classroom decisions—such as identifying students who need additional support—and long-term institutional planning, such as curriculum development or resource allocation (Hamoud et al., 2021). By organizing data in a consistent and accessible way, DSS helps schools make evidence-informed decisions while maintaining privacy, security, and data quality standards.

Managing and Using Data in E-Learning Systems

E-learning systems generate large and diverse amounts of data. Every click in a learning management system (LMS), quiz attempt, video view, discussion post, and assignment submission produces digital traces that can support EDM

(Fernández et al., 2014; Liu & Yu, 2022). To make meaningful use of this information, the data must be properly stored and harmonized.

Cloud computing is widely used to support e-learning data storage because it can scale as the number of students, courses, and learning activities grows (Fernández et al., 2014). For example, a university offering online courses to thousands of learners may store LMS data, video analytics, and assess results in cloud-based systems to ensure reliable access and performance.

Many e-learning systems use a two-level storage approach. First, **raw data storage** keeps original, detailed records, such as complete activity logs, for future analysis; Second, processed **data layer** stores cleaned, organized, and harmonized data that is ready for reporting, dashboards, or predictive models (Ang et al., 2020). Data harmonization typically takes place between these two layers, ensuring that data from different sources (e.g., LMS logs, student information systems, and survey tools) use consistent formats and definitions.

This structured storage supports practical applications in teaching and learning. For instance, harmonized LMS data can be used to identify students' learning patterns, preferences, or levels of engagement (Jittawiriyanukoon, 2019). A system might detect that certain students engage more with video content, while others prefer interactive quizzes. These insights can then be stored in a data warehouse or mart and used to personalize learning materials, recommend resources, or adjust instructional strategies.

Comparing Data Across Educational Contexts

One of the most valuable outcomes of data storage and harmonization is the ability to compare data across schools, districts, or universities. When data are organized in consistent formats and use shared definitions, institutions can benchmark performance, identify trends, and inform policy decisions beyond a single educational context (Daniel, 2014; Obilikwu & Ogbuju, 2020).

In practice, many institutions store their data in separate systems that were designed independently, which greatly reduces data comparability and makes cross-population analysis a herculean task (Obilikwu & Ogbuju, 2020). For example, one school may define "course completion" differently from another or use different grading scales. When data are structured inconsistently, meaningful comparison becomes difficult and time-consuming. Harmonization addresses this challenge by aligning data formats, terminology, and indicators across institutions. For instance, if multiple schools agree on common definitions for attendance rates, assessment scores, and course completion, their data can be consolidated into a shared repository using a uniform format. This allows education authorities or networks of schools to compare outcomes, identify high-performing programs, and share effective practices (Obilikwu & Ogbuju, 2020).

Cross-institutional comparability is particularly important for performance comparison in higher education, where universities operate in competitive and rapidly changing environments (Daniel, 2014). Reliable, harmonized data enable system-level analysis, such as evaluating student retention patterns across regions or assessing the impact of national education policies. Looking ahead, the development of shared standards and interoperable systems will further strengthen cross-institutional data harmonization (Buenaño-Fernández et al., 2019; Thajchayapong, 2025). In practical terms, this means creating agreed-upon data definitions and exchange formats so that information can move smoothly between institutions while maintaining privacy and data security. In short, harmonized data storage not only supports individual schools but also enables broader collaboration, benchmarking, and evidence-based decision-making across the education system.

4.5 Key Considerations for Responsible Educational Data Storage and Harmonization

Protecting Student Data: Privacy and Security

Storing and harmonizing educational data requires careful attention to privacy and security. Educational datasets often include sensitive information such as academic records, attendance, behavioral data, and demographic details (Ang et al., 2020). Protecting this information is both a legal obligation and an ethical responsibility for educational institutions.

When data from multiple systems—such as student information systems, learning management systems (LMS), and assessment platforms—are combined into a warehouse or data mart, the potential risks increase. Centralized storage makes analysis more efficient, but it also means that strong safeguards must be in place to prevent unauthorized access or data breaches (Alkhubouli et al., 2024).

In practice, protecting student data involves clear policies and technical measures. For example, access to dashboards may be restricted based on roles (e.g., individual students can only view their own data, teachers can view their own students' data, while administrators can access school-level summaries). Data can be pseudonymized, anonymized, or aggregated before use. Pseudonymization replaces names with codes, anonymization removes identifiers entirely, and aggregation combines data at a group level, protecting individual privacy. Access to raw, pseudonymized, anonymized, or aggregated data should be clearly defined, with only authorized users allowed to see sensitive information. Regular audits, secure login procedures, and encrypted storage help reduce privacy and security risks.

Studies have shown that well-designed systems—such as decision support systems built on structured data marts—can maintain strong privacy and performance standards when properly implemented (Hamoud et al., 2021). However, as educational institutions collect increasing amounts of diverse and real-time data, maintaining privacy and security remains an ongoing challenge (Alkhubouli et al., 2024; Ang et al., 2020).

In short, effective data storage and harmonization must always balance analytical benefits with responsible data governance, ensuring that student information is protected while still supporting meaningful educational insights.

Ensuring Ethical and Fair Use of Educational Data

Educational data storage and harmonization raise ethical questions beyond privacy and security. Key concerns include informed consent, transparency, and fairness in how data are analyzed and used (Ang et al., 2020). Students and families should understand what data are collected and how they inform decisions.

Predictive models, such as those used to identify students at risk, can support timely intervention. However, if based on biased or incomplete historical data, they may unintentionally reinforce existing inequalities (Ang et al., 2020; Gul et al., 2021). For example, a model trained on past performance patterns could unfairly label certain groups of students, influencing expectations or access to opportunities.

Ethical considerations should therefore be built into data systems from the outset. This includes careful data selection, regular review of algorithms for bias, and ensuring that predictive insights are used to support—not penalize—students. Ultimately, harmonized educational data should be used in ways that promote fairness, transparency, and student well-being.

Ensuring High-Quality Educational Data

The value of educational data depends on its quality. Even the most advanced storage systems or analytics tools cannot produce reliable insights if the underlying data are inaccurate, incomplete, inconsistent, or outdated

(Moscoso-Zea et al., 2019; Ramaswami et al., 2019). Ensuring high-quality data to support reliable EDM is therefore, a foundational step in storage and harmonization.

Data quality issues often complicate educational data harmonization. Student names may differ across systems, attendance records may be missing, or grading scales may not align. Different identifiers—such as full account IDs versus email addresses—can prevent automated data matching across datasets, requiring manual correction. Addressing these inconsistencies is essential to ensure reliable analyses, such as predicting student performance or comparing outcomes across courses and institutions.

Improving data quality involves systematic processes such as cleaning incorrect entries, filling in missing values where appropriate, and standardizing formats and definitions (Glover et al., 2010). For instance, harmonization may require converting all grades to a common scale or ensuring that course codes follow the same structure across systems. Investing in data quality management is not simply a technical task—it is a strategic priority. Clean, consistent, and up-to-date data enable trustworthy EDM, more accurate reporting, and better-informed educational decisions.

Establishing Clear Data Governance Practices

Effective data storage and harmonization require more than technical systems—they also require clear governance. Data governance refers to the policies, roles, and procedures that guide how educational data are collected, stored, used, shared, and archived over time (Alkhubouli et al., 2024; Shah et al., 2021). Without clear rules and responsibilities, even well-designed data systems can become inconsistent or unreliable.

In practice, governance frameworks help educational institutions manage data throughout their lifecycle—from initial collection to long-term storage or deletion. For example, a school may establish guidelines on who is allowed to enter or edit student data, in what format data are stored, how often records should be updated, and how long historical data should be retained. Data lifecycle frameworks, such as those proposed by Shah et al. (2021), emphasize structured processes including data collection, preparation, enrichment, and archiving.

Data archiving is particularly important for EDM. Older records, such as past cohorts' performance data, may no longer be needed for daily operations but remain valuable for longitudinal analysis. A tiered storage strategy can move inactive data to secure archival storage while keeping frequently used data readily accessible. This improves system performance and ensures that historical data remains available for future research and evaluation. Clear governance practices ensure that stored and harmonized educational data remains accurate, accessible, and usable over time. By defining responsibilities, retention policies, and access rules, institutions can support sustainable, transparent, and responsible data use.

4.6 Challenges and Future Directions

Scaling Educational Data Storage and Harmonization

As educational data grows, scaling storage and harmonization remains a key challenge (Ang et al., 2020; Koliçi et al., 2014; Moscoso-Zea et al., 2019). Virtual campuses and MOOCs can generate 15–20 GB of student activity logs daily. LA can support monitoring, personalized courses, and adaptive learning—but only if infrastructure can handle these volumes (Koliçi et al., 2014). A key challenge is deciding how much of the same data to store across multiple systems. As datasets grow, storing everything in every system can quickly consume storage and reduce accessibility, while storing only selected parts can save space but may limit future analyses. Clear policies on what to retain, archive, and access are essential. Thoughtful planning ensures educational data storage is scalable, efficient, and reliable, supporting LA without unnecessary duplication or wasted resources.

Traditional storage systems and batch-based harmonization approaches are often insufficient for managing the volume and speed of big educational data (Ang et al., 2020; Moscoso-Zea et al., 2019). This has led to the adoption of distributed storage and processing frameworks, which allow schools and institutions to consolidate heterogeneous data from multiple sources into unified repositories (Al-Yadumi et al., 2021; Buenaño-Fernández et al., 2019).

To keep up with the rapid growth in educational data, scalable storage and harmonization are essential. These systems should support flexible, near-real-time data integration while maintaining quality, consistency, and accessibility. Harmonizing formats and structures can reduce the need to store duplicate data across multiple systems, though perfect harmonization is only possible when data pipelines are designed from the start with harmonization in mind. Future developments in open standards and interoperability frameworks will further enable cross-platform data sharing, supporting research, benchmarking, and evidence-based decision-making (Ang et al., 2020; Thajchayapong et al., 2025).

Ensuring Semantic Consistency Across Educational Data

A major challenge in educational data harmonization is ensuring that data from different sources "mean the same thing," a concept known as semantic interoperability (Buenaño-Fernández et al., 2019; Thajchayapong et al., 2025). Even when data is stored consistently in a data warehouse or cloud repository, differences in terminology, definitions, or measurement scales across systems can prevent meaningful comparison. For example, one school might record "course completion" as finishing all assignments, while another defines it as passing final exams. Without harmonization, combining these data for analysis can lead to misleading insights.

Semantic interoperability goes beyond simply converting heterogeneous data into a consistent format—it requires aligning the meaning of data elements across sources. Techniques from ontology management and AI can help resolve these differences, allowing systems to understand that differently labeled data may represent the same concept (Nica et al., 2011). In practice, this might involve mapping grade scales from multiple schools to a common standard or ensuring that engagement metrics from an LMS and a MOOC platform are comparable.

Applying these approaches within data storage systems, such as data warehouses or cloud repositories, supports automated and consistent harmonization. By integrating semantic knowledge directly into storage and processing pipelines, educational institutions can reduce errors, improve cross-institutional comparability, and make data mining and LA more reliable.

Leveraging Emerging Technologies for Smarter Data Management

Emerging technologies are transforming how educational data is stored, harmonized, and used. Open standards and interoperable data frameworks make it possible to securely combine data from multiple sources, while cloud and distributed systems provide the scalability needed to handle growing volumes of student and course data (Thajchayapong et al., 2025).

Artificial intelligence (AI) plays an increasingly important role in this process. AI-augmented systems can automate many harmonization tasks, such as detecting inconsistencies in terminology, mapping data from different schemas, and validating data quality. For example, AI tools could automatically align grading scales from different schools or reconcile engagement data from an LMS and a MOOC platform, saving time and reducing errors compared to manual harmonization (Nica et al., 2011; Thajchayapong et al., 2025).

One practical example is the Architecture for AI-Augmented Learning (A4L), which integrates data ingestion, preprocessing, organization, analytics, and visualization. This framework allows secure and seamless integration

across systems like Student Information Systems (SIS), LMS platforms, and AI-enabled learning tools, supporting personalized learning and data-driven decision-making (Thajchayapong et al., 2025).

Agile data approaches, such as the MAD (Magnetic, Agile, Deep) philosophy, complement traditional data warehouses by offering flexible and faster integration of new data sources. These methods allow analysts to explore and harmonize dynamic and diverse educational data without being constrained by rigid structures, making it faster and easier to respond to evolving educational needs (Cohen et al., 2009).

In short, emerging technologies provide schools and institutions with smarter, faster, and more flexible ways to store, manage, harmonize, and analyze educational data, supporting personalized learning, institutional planning, and system-wide insights.

Managing the Educational Data Lifecycle

Managing the lifecycle of educational data is a key challenge for data storage and harmonization, covering everything from data collection and cleaning to storage, analysis, archiving, and disposing (Gul et al., 2021; Shah et al., 2021). Effective lifecycle management ensures that data remains accurate, consistent, and usable over time, supporting reliable insights and decision-making.

Frameworks like DaLiF highlight the main stages of the data lifecycle: collection, preparation, enrichment, storage, and archiving (Shah et al., 2021). During the preparation stage, raw data from multiple sources—such as LMS logs, student information systems, and survey responses—is cleaned, filtered, and normalized into a consistent format suitable for analysis. For instance, attendance records from different schools may use different formats or codes, which need to be standardized before comparing student engagement across classes. The enrichment stage further refines the data, combining and harmonizing it to create a mature, reliable dataset. These harmonized datasets can then be used for analyses, such as predicting student performance or identifying trends in course participation or archived for future reference. Proper storage and archiving practices, including cloud-based or distributed storage systems, help ensure that historical educational data remains accessible for longitudinal studies and policy evaluation.

Developing clear, practical data lifecycle management strategies is essential for schools and institutions to maintain high-quality, harmonized educational data and to support robust EDM and LA now and in the future (Shah et al., 2021; Thajchayapong et al., 2025).

References

Alkhubouli, M., Lala, H., AlHabshy, A., & ElDahshan, K. (2024). Enhancing Data Warehouses Security. *International Journal of Advanced Computer Science and Applications, 15*(3). https://doi.org/10.14569/ijacsa.2024.0150358

Al-Yadumi, S., Tan, E., Wei, S., & Boursier, P. (2021). Review on Integrating Geospatial Big Datasets and Open Research Issues. *IEEE Access, 9*, 10604–10620. https://doi.org/10.1109/access.2021.3051084

Ang, L., Ge, F., & Seng, K. (2020). Big Educational Data & Analytics: Survey, Architecture and Challenges. *IEEE Access, 8*, 116392-116414. https://doi.org/10.1109/access.2020.2994561

Buenaño-Fernández, D., Villegas-Ch, W., & Luján-Mora, S. (2019). The use of tools of data mining to decision making in engineering education—A systematic mapping study. *Computer Applications in Engineering Education, 27*(3), 744–758. https://doi.org/10.1002/cae.22100

Cohen, J., Dolan, B., Dunlap, M., Hellerstein, J., & Welton, C. (2009). MAD skills. Proceedings of the VLDB Endowment, 2(2), 1481–1492. https://doi.org/10.14778/1687553.1687576

Daniel, B. (2014). Big Data and analytics in higher education: Opportunities and challenges. *British Journal of Educational Technology, 46*(5), 904–920. https://doi.org/10.1111/bjet.12230

Fernández, A., Peralta, D., Benítez, J., & Herrera, F. (2014). E-learning and educational data mining in cloud computing: an overview. *International Journal of Learning Technology, 9*(1), 25. https://doi.org/10.1504/ijlt.2014.062447

Gul, S., Bano, S., & Shah, T. (2021). Exploring data mining: facets and emerging trends. *Digital Library Perspectives, 37*(4), 429-448. https://doi.org/10.1108/dlp-08-2020-0078

Hamoud, A., Hussein, M., Alhilfi, Z., & Sabr, R. (2021). Implementing data-driven decision support system based on independent educational data mart. *International Journal of Electrical and Computer Engineering (IJECE), 11*(6), 5301. https://doi.org/10.11591/ijece.v11i6.pp5301-5314

Jittawiriyanukoon, C. (2019). Proposed classification for eLearning data analytics with MOA. *International Journal of Electrical and Computer Engineering (IJECE), 9*(5), 3569. https://doi.org/10.11591/ijece.v9i5.pp3569-3575

Koliçi, V., Xhafa, F., Barolli, L., & Lala, A. (2014). Scalability, Memory Issues and Challenges in Mining Large Data Sets. *In Proceedings of International Conference on Intelligent Networking and Collaborative Systems.* Italy, pp. 268–273. https://doi.org/10.1109/incos.2014.50

Liu, M., & Yu, D. (2022). Towards intelligent E-learning systems. *Education and Information Technologies, 28*(7), 7845–7876. https://doi.org/10.1007/s10639-022-11479-6

Moscoso-Zea, O., Castro, J., Paredes-Gualtor, J., & Luján-Mora, S. (2019). A Hybrid Infrastructure of Enterprise Architecture and Business Intelligence & Analytics for Knowledge Management in Education. *IEEE Access, 7*, 38778-38788. https://doi.org/10.1109/access.2019.2906343

Nica, A., Suchanek, F., & Varde, A. (2011). Emerging multidisciplinary research across database management systems. *ACM Sigmod Record, 39*(3), 33–36. https://doi.org/10.1145/1942776.1942786

Obilikwu, P. and Ogbuju, E. (2020). A data model for enhanced data comparability across multiple organizations. *Journal of Big Data, 7*(1). https://doi.org/10.1186/s40537-020-00370-1

Ramaswami, G., Sušnjak, T., Mathrani, A., Lim, J., & García, P. (2019). Using educational data mining techniques to increase the prediction accuracy of student academic performance. *Information and Learning Sciences, 120*(7/8), 451–467. https://doi.org/10.1108/ils-03-2019-0017

Shah, S., Peristeras, V., & Magnisalis, I. (2021). DaLiF: a data lifecycle framework for data-driven governments. *Journal of Big Data, 8*(1). https://doi.org/10.1186/s40537-021-00481-3

Sun, L. (2011). Batch loading in metadata creation: a case study. *The Electronic Library, 29*(4), 538–549. https://doi.org/10.1108/02640471111156786

Thajchayapong, P., Carbonaro, S., Couper, T., Helmick, B., Rugaber, S., & Goel, A. (2025). *Evolution of A4L: A Data Architecture for AI-Augmented Learning.* https://doi.org/10.48550/arxiv.2511.11877