Journal of Universal Computer Science, vol. 25, no. 13 (2019), 1668-1686
submitted: 6/7/18, accepted: 20/11/19, appeared: 28/12/19

Adopting Trust in Learning Analytics Infrastructure: A Structured Literature Review

George-Petru Ciordas-Hertel
(DIPF | Leibniz Institute for Research and Information in Education
Frankfurt am Main, Germany
orcid.org/0000-0003-1589-7845
ciordas@dipf.de)

Jan Schneider
(DIPF | Leibniz Institute for Research and Information in Education
Frankfurt am Main, Germany
orcid.org/0000-0002-1229-2579
schneider.jan@dipf.de)

Stefaan Ternier
(Open Universiteit, Heerlen, Netherlands
orcid.org/0000-0002-1598-9558
stefaan.ternier@ou.nl)

Hendrik Drachsler
(DIPF | Leibniz Institute for Research and Information in Education
Frankfurt am Main, Germany
Johann Wolfgang Goethe-Universität, Frankfurt am Main, Germany
Open Universiteit, Heerlen, Netherlands
orcid.org/0000-0001-8407-5314
drachsler@dipf.de)

Abstract: One key factor for the successful outcome of a Learning Analytics (LA) infrastructure is the ability to decide which software architecture concept is necessary. Big Data can be used to face the challenges LA holds. Additional challenges on privacy rights are introduced to the Europeans by the General Data Protection Regulation (GDPR). Beyond that, the challenge of how to gain the trust of the users remains. We found diverse architectural concepts in the domain of LA. Selecting an appropriate solution is not straightforward. Therefore, we conducted a structured literature review to assess the state-of-the-art and provide an overview of Big Data architectures used in LA. Based on the examination of the results, we identify common architectural components and technologies and present them in the form of a mind map. Linking the findings, we are proposing an initial approach towards a Trusted and Interoperable Learning Analytics Infrastructure (TIILA).

Key Words: Learning Analytics, Software Architecture, Infrastructure, Big Data, Trust, Data Protection, Privacy, GDPR, Education

Category: D.2.11, D.2.12, L.3.6, K.3

1 Introduction

In recent decades a rising number of educational institutions have been investigating in learning management systems (LMSs) such as Moodle[1] to promote students with online learning facilities. We count such LMSs as virtual learning environments (VLEs). In addition to course content such as texts, videos, and quizzes, VLEs offer a variety of tools. These tools are communication tools such as forums and messengers, as well as digital repositories and social networking tools. These tools capture data in a natural way to provide their functionality, e.g., for the management of learning content or synchronous and asynchronous interaction of participants. To a lesser extent, they capture the history of activities or versions of artifacts, unless it is necessary for supporting their user stories. Most VLEs do not have suitable tools for analyzing the data, nor can they integrate other VLEs. To address these issues, Learning Analytics (LA) arose with the aim of better understanding the learning processes and environments in order to improve teaching by collecting, measuring, analyzing, and reporting data [Gasevic et al., 2017].

With this mindset, we are looking for software architecture to enroll LA at our institute. Conceptional frameworks such as from Greller & Drachsler [Greller and Drachsler, 2012] inform about soft barriers and limitations of LA. Brief research revealed projects mostly targeting small data exports. We plan to serve possibly thousands of users in the roles of students, teachers, and administrators with a responsive on-premise solution. Responsive solution refers in this context to the responsiveness of the user interface. As well, it refers to the calculation and publication of the analytic processes results. This type of LA system holds multiple challenges to software designers. Such challenges include the huge and rapidly arriving amounts of data. This can be the case for LA data arriving from multiple VLEs and multimodal sources such as sensors [Schneider et al., 2015]. Another challenge is that data from such diverse sources is usually fragmented, duplicated, differently identified, and represented in a non-standard manner. In addition, users expect a responsive user interface with processed information provided in real-time in a comprehensible way [van Merriënboer et al., 2002].

Such challenges need to be faced by a Big Data software architecture [Taylor and Munguia, 2018]. Big Data can be described by examining its characteristics [Katal et al., 2013]. The three main characteristics are volume, velocity, and variety. Volume refers to the exponentially growing amount of data. Velocity refers to how fast data arrives from different sources and flows within the system. Variety refers to the different categories of data. These categories are unstructured, semi-structured, and structured data. Recent research extends those three characteristics by four additional characteristics. These characteristics are veracity,

[1] https://moodle.org/

value, variability, and visualization. Veracity concerns the accuracy and trustworthiness of data. The value corresponds to the usefulness of data. Variability denotes the problem of constantly changing the meaning. Lastly, visualization refers to the presentation of the data.

As LA processes personal data, and even sensitive personal data, privacy concerns have to be taken into account. The General Data Protection Regulation (GDPR) [European Parliament, 2016] came into effect on May 25th, 2018, and fundamentally changed the European privacy rights. Although some national data protection legislation such as the BDSG in Germany already contained some of these rights, the overall concept is new to many European nations.

The GDPR introduces the *privacy by design* and the *privacy by default* framework. *Privacy by design* states that organizations need to consider data protection and privacy from the beginning of the software design throughout the complete development process of personal data processing products. Additionally, *privacy by default* means that only personal data that is necessary for each specific purpose of the processing shall be processed. Furthermore, the GDPR grants extended rights to the user as the data subject. Each data subject has first of all the *right to be informed* (RtbInfo) in a sufficiently enough form on how the software works and how personal data is processed. The *right to access* (RtAcc) instructs the data controller to provide a copy of the personal data, free of charge, in an electronic format. Moving along the *right for data portability* (RfDPort) demands a structured, commonly used, and machine-readable format. Given this at hand, the data subject can enforce its *right to rectification* (RtRect) of personal data. With the *right to object* (RtObj) to processing of its data, the data subject can at any time stop processing on illegitimate grounds. Similarly, applying its *right to restrict processing* (RtRProc) personal data may, except for storage, only be processed with the data subject's consent. Finally, the *right to erasure* (RtEras) entitles the data subjects to have the data controller erase their data. These regulations, among others, are hard requirements that an LA system designer needs to take into account in order to be compliant with law [Hoel et al., 2017].

Initiatives such as Apereo[2] and Jisc[3] provide reasonable LA solutions, but these are not suitable as a Big Data, on-premise solution. An extensive framework for Open LA is the Open Learning Analytics Platform (OpenLAP)[4]. It is mainly focused on the generation of personalized indicators and not on Big Data processing or privacy[Muslim et al., 2018].

Within the *Trusted Learning Analytics* research project, that is initiated by the DIPF | Leibniz Institute for Research and Information in Education, the

[2] https://www.apereo.org
[3] https://www.jisc.ac.uk/
[4] https://www.uni-due.de/soco/research/projects/openlap.php

University of Frankfurt, and the Open University of the Netherlands, we aim to enroll a modern Big Data software architecture which is not only in conformance with GDPR but enforces user-guided privacy control. The ability of the users to control the processing and collection of their data in an informed and, therefore, consent way is our approach to gain trust in LA systems. By seeking the trust of the users instead of evading consent by hiding behind legitimate interest, we hope to raise commitment and engagement with LA.

To conceptualize a suitable architecture that addresses these needs, we conducted a structured literature review [Fink, 2013] guided by the following research questions:

RQ1 Which are the Big Data architectures currently used in the Learning Analytics domain?

RQ2 How is privacy currently handled in Big Data architectures in the Learning Analytics domain?

This introduction follows the review protocol of the structured literature review explaining the search strategy and the refinement process. Next, the results of the conducted comprehensive overview of LA software architectures are shared. This encloses the description of the results found for research questions 1 and 2. Concluding from these results, an expanding structured meta literature review was conducted to gain a broader overview of Big Data architectures and Big Data privacy concerns. From all results, we present a mind map of LA software architecture suitable Big Data technologies. After that, we conclude an initial approach towards a trusted LA Big Data architecture follows. We finalized this publication, stating some of the limitations and future steps of this work as well provide a conclusion.

2 Review Protocol

We took different steps in order to identify suitable literature in the context of a structured literature review [Fink, 2013].

We started by pre-selecting possible literature databases out of the field of computer science by searching for a high number of possibly relevant literature. Possibly relevant literature was identified using the search query "big data AND education". We based the search on title, abstract, and author keywords. Based on the result of the search query, we selected the four databases IEEE, ScienceDirect, SpringerLink, and ACM. In order to reduce the number of irrelevant literature, we only took publications from edited books, journals, or proceedings into account. As the technology around Big Data develops so fast, we focused only on recent publications from the years 2015 to 2019. We performed the search on April 29, 2019. The first column of table 1, shows the number of documents resulting from the query.

To finally identify the relevant literature, we applied two additional refinement phases. In each phase, we checked if the publication tackles at least one of the research questions. In the first phase, we analyzed titles and abstracts. In the second phase, we considered the complete content. Table 1 shows the results of the first and the second phase.

Database	Initial Query	First Phase	Second Phase
IEEE	265	58	11
ScienceDirect	111	13	3
SpringerLink	421	24	5
ACM	250	29	5
Total	1047	124	24

Table 1: Number of publications identified in each step of the SLR

3 Results

This section presents the answers to our research questions based on the identified publications. Only a representative set was chosen to give an overview.

3.1 RQ1 - Which are the Big Data architectures currently used in the Learning Analytics domain

As shown in table 2, we took into account a total number of 20 different publications for answering our research questions. The selected publications offered a variety of approaches to implement an LA system at their institution. The publications differ quite strongly from each other in the richness of detail and realization state. In order to get an overview, we classified each publication based on what its architectural concept is. We identified four architectural concepts. These architectural concepts are *experimental*, *generic*, *manual* and *automatic*. The technologies mentioned hereafter are in detail explained in section 5.

Six of the publications describe *experimental* setups to explore a possible solution for their institutions. The authors mostly collected data from different Massive Open Online Courses (MOOCs) such as edX[5] [Santur et al., 2016, Gómez-Berbís and Lagares-Lemos, 2016, Tang et al., 2015]. As an example to start with

[5] https://www.edx.org

Architectural Concept	Number	Specific Publications
Experimental	9	[Santur et al., 2016, Gómez-Berbís and Lagares-Lemos, 2016, Huang et al., 2016, Swathi et al., 2017, Santoso and Yulia, 2017, Tang et al., 2015, Hu et al., 2019, Dahdouh et al., 2018, Vagliano et al., 2018]
Generic	5	[Petrova-Antonova and Ilieva, 2019, Matsebula and Mnkandla, 2017, Jiangbo Shu et al., 2017, Zhang et al., 2016, Srinivasan et al., 2015]
Manual	2	[Laveti et al., 2017, Furukawa et al., 2017]
Automatic	8	[Rabelo et al., 2015, Li et al., 2017, Chen et al., 2017, Chaffai et al., 2017, Zhao et al., 2017, Yang and Huang, 2016, Logica and Magdalena, 2015, Lopez et al., 2017]

Table 2: Synopsis of the results for RQ1 and RQ2

data processing a classical machine learning workflow to get an idea of how to classify dropout students designed with and run in WEKA[6] is suitable [Tang et al., 2015]. As it comes to Big Data platforms the teams where experimenting with Apache Hadoop[7] [Huang et al., 2016, Swathi et al., 2017] and Apache Spark[8] [Santur et al., 2016]. These platforms were, for example, used to cluster the students [Gómez-Berbís and Lagares-Lemos, 2016].

Four publications are about rather abstract technology-independent architectural concepts for LA. They described *generic* architectural components [Matsebula and Mnkandla, 2017, Jiangbo Shu et al., 2017, Zhang et al., 2016] and cloud architectures [Srinivasan et al., 2015]. These publications provide an overview of the different architectural components an LA system requires to be effective. Abstracting from the different terms used in the *generic* publications, we identified six architectural components of a typical data processing workflow: (a) Collection, (b) Processing, (c) Transmission, (d) Storage, (e) Analytics, and (f) Visualization.

The authors of the last ten publications designed concrete systems based on specific technologies the designers had chosen. We classified these systems by the terms manual and automatic. Core of the distinction between *manual* and *automatic* is how the analytics results are created. Two institutional

[6] https://www.cs.waikato.ac.nz/ml/weka
[7] https://hadoop.apache.org
[8] https://spark.apache.org

LA systems used workflows where the analytics is done *manually* [Furukawa et al., 2017, Laveti et al., 2017]. By using the term *manual*, we refer to a non-automatic process where every step is performed independently by somebody. One workflow [Furukawa et al., 2017], for example, starts with the collection (a) of Moodle database records from an SQL[9] dump. They are then processed (b) to pseudonymize and transform them into a CSV[10] file. That CSV file is further processed (b) into xAPI[11] Statements and stored (d) in an instance of Learning Locker[12]. The analytics (e) of those xAPI Statements is carried out using R and visualized (f) with a web-based dashboard.

The remainder of the publications proposes automatic LA systems. These systems consist of a continual or continuous process triggered by either a scheduler or incoming events. Each trigger incites the analytics to reevaluate. We, therefore, grouped these LA systems with the term *automatic*. An exciting approach proposed a system consisting of a variety of VLEs, storage, and multiple analytics services [Rabelo et al., 2015]. Collected (a) xAPI Statements from all of those VLEs are transmitted (c) continuously to storage via a message bus. To verify (b) and enhance the meaningfulness of the stored (d) data, they used a custom software called OntoLAK to provide an ontology for a Titan[13] database. The authors did the analytics (e) with SparQL[14] and a variety of algorithms getting data via a REST service provided by their storage implementation. A custom dashboard visualizes (f) the results. A different approach is describing a real-time analytics system based on Apache Spark [Chaffai et al., 2017]. Using Moodle as a VLE, they are collecting (a) and transmitting (c) static data such as course content data periodically with Apache Sqoop[15] and real-time events such as clickstreams with a combination of Apache Flume[16] and Apache Kafka[17]. Processing (b) and analytics (e) are done with Apache Spark jobs, which load the static tables from the Apache HBase[18] storage (d) and registers to one or multiple Apache Kafka streams for the real-time events. Apache Thrift[19] is used to extract the analytics results from the data storage. The dashboard visualization (f) is finally done with D3.js[20].

We can identify the previously identified architectural components (a-f) in

[9] https://www.w3schools.com/sql
[10] comma-separated values
[11] https://xapi.com
[12] https://learninglocker.net
[13] https://titan.thinkaurelius.com/
[14] https://www.w3.org/TR/rdf-sparql-query
[15] https://sqoop.apache.org
[16] https://flume.apache.org
[17] https://kafka.apache.org
[18] https://hbase.apache.org
[19] https://thrift.apache.org
[20] https://d3js.org

all the publications describing LA systems. We, therefore, consider them to be suitable.

At last, we should mention a unique cloud-based approach [Lopez et al., 2017]. Those authors state that over time, the amount of their data became too big for their on-premise LA system to be responsive any more. Aggregating data took them in some cases, longer than 48 hours. Instead of further expanding their on-premise setup, they decided to transform their solution into a Big Data cloud solution. They developed a cloud solution called edx2bigquery[21] to use Google services such as BigQuery[22]. This is an interesting finding that demonstrates the dilemma between having a responsive system and the need to protect the data of EU students, according to GDPR. Although there should be no further issues with data protection when data processing providers comply with GDPR, data subjects might not trust third party data processors. For such a solution, it is essential to consider whether an institution has as much local expertise in data security as a professional cloud service provider with a mature security framework.

3.2 RQ2 - How is privacy currently handled in Big Data architectures in the Learning Analytics domain?

None of the authors of the publications we found in this literature review designed any specific privacy functionality as required for the GDPR in their LA solution. This, although some have been introduced to some national law even before. Only Furukawa et al. [Furukawa et al., 2017] pseudonymized the data before further processing it. In general, we can state that *privacy by design* was not an issue of the identified previous efforts and the new information rights and that to the best of our knowledge, concluding functionality for the data subject coming from the GDPR up until now has not been considered.

4 Meta Literature Review

As a consequence of the diversity among the few publications explaining their LA architecture and less mentioning privacy, we decided to extend the literature study to identify commonalities, best practices, and methods to deal with trust, to transfer these insights to the LA community and for the Trusted Learning Analytics project. Therefore, we formed two additional research questions.

RQ3 What comprises current common Big Data architectures?

RQ4 How are privacy requirements addressed in Big Data architectures?

[21] https://github.com/mitodl/edx2bigquery
[22] https://cloud.google.com/bigquery

We addressed these research questions by conducting a structured meta literature review.

4.1 Review Protocol

Since we are still focusing on the area of Big Data, the same conditions as explained in section 2 hold.

To narrow down the number of results, we chose the well-established computer science meta database DBLP[23]. An initial search for the term *big data* resulted in 9612 matches. In order to find publications giving us a general overview, a more specific search query was necessary. For the specific search, the initial query "big data AND ('literature review' OR survey)" was used. This initial query revealed 151 publications.

To finally identify the relevant literature, we applied a refinement process consisting of two phases. In each phase, we checked if the publication tackles at least one of the research questions. In the first phase, we analyzed titles and abstracts. In the second phase, we considered the complete content. Applying these two phases narrowed the relevant literature down to a total of 12 publications.

4.2 Results

Research Question	Initial Query	Refinement	Specific Publications
RQ3	151	4	[Salvador et al., 2017, Volk et al., 2017, Chen et al., 2016, Liu et al., 2018]
RQ4		9	[Salleh and Janczewski, 2016, Ye et al., 2016, Nelson and Olovsson, 2016, Chen and Yan, 2016, Victor et al., 2016, Grover and Aulakh, 2017, Thangaraj and Balamurugan, 2017, Miloslavskaya and Makhmudova, 2016, Chen et al., 2016]

Table 3: Synopsis of the results for RQ3 and RQ4

Table 3 shows the results from the search for publications to answer RQ3 and RQ4. We described only the most important findings.

[23] https://dblp.uni-trier.de

4.2.1 RQ3 - What comprises current common Big Data architectures?

The four publications provided us with a good overview of the requirements, architectural components, and technologies of a Big Data system. The availability and reliability of data access are significantly impacted by how Big Data is stored and indexed. In-memory data storage, management, and manipulation perform much faster, requires significantly less memory and CPU support [Chen et al., 2016]. Big Data computing can be either done by firstly storing it and then analyzing it or continuously on the flow of data. The first is called batch and the latter stream processing [Chen et al., 2016]. Stream processing is used when freshness is of importance. As doing analytics on data with machine learning is resourceful, different tools and scaling methods should be taken into account. Salvador et al. [Salvador et al., 2017] listed multiple of such Big Data platforms with criteria such as scaling and storage. Finally, each Big Data system needs to be operated on some infrastructure. Using a cloud container technology can be a suitable method. Liu et al. [Liu et al., 2018] are proposing such a cloud scheme for all kinds of heterogeneous architectures while naming explicit platforms such as OpenStack[24].

4.2.2 RQ4 - How are privacy requirements addressed in Big Data architectures?

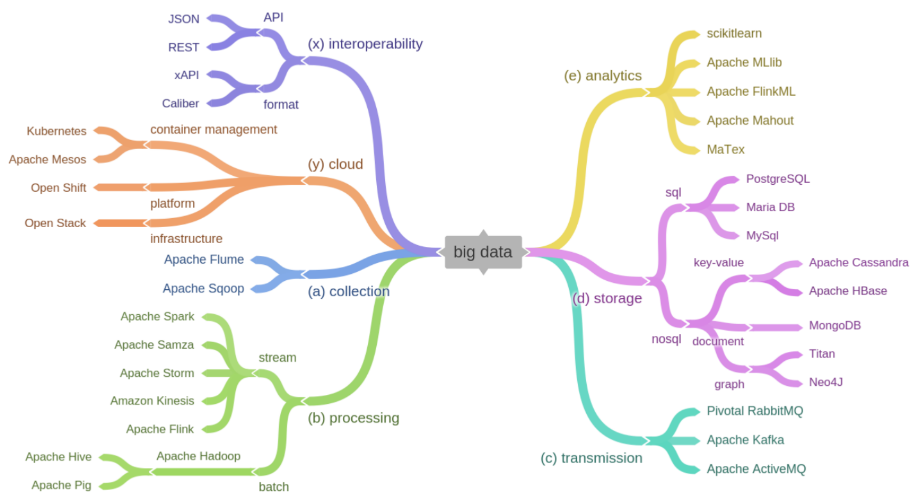
We found a total of nine publications, which offer a variety of approaches to deal with privacy concerns. Security issues and privacy concerns can be classified under three major contexts: technological, organizational, and environmental [Salleh and Janczewski, 2016]. These contexts can help organizations in their Big Data adoption process. According to Chen et al. [Chen and Yan, 2016], a Big Data system workflow contains the phases data collection, data transmission, data processing, and data storage with each phase having requirements such as confidentiality, efficiency, authenticity, availability, and integrity. They state that technologies such as homomorphic encryption and secure data storage schemes are suitable to address those requirements. Ye et al. [Ye et al., 2016] used a slightly different approach by discussing the challenges of each of the characteristics of Big Data. Unlike the predominant general approaches, Victor et al. [Victor et al., 2016] focused on data sharing and publishing scenarios of Big Data systems.

5 Discussion

During the examination of the results, we identified the following architectural components: (a) Collection, (b) Processing, (c) Transmission, (d) Storage, (e)

[24] https://www.openstack.org

Figure 1: This mind map shows the Big Data technologies identified in the reviewed publications



Analytics, and (f) Visualization. As revealed by the results, these architectural components can be implemented using a diversity of technologies. We present all the results from the publications as a mind map in figure 1.

The collection (a) of data can be done by an application such as Apache Flume, which is designed to fetch streaming data such as tweets[25] or log files from Moodle. In some cases, it might be necessary to develop their Learning Activity Sensor (LAS) to collect data from closed source applications. In order to transmit (c) the events and facts collected, a message broker such as Pivotal RabbitMQ[26] can be useful. Such a message broker helps, among other things to decouple, supervise and orchestrate microservices as well as transform messages and synchronize data. As it goes for the interoperability (x) of the microservices RESTful[27] web services with JSON[28] as the data-interchange format became a standard. The data collected from clickstreams are in the majority of times processed in its original or by some custom format. In a few cases, the authors used a standard such as xAPI. xAPI can be a powerful format that allows enriching data [Berg et al., 2016]. More so since users are able to create their own

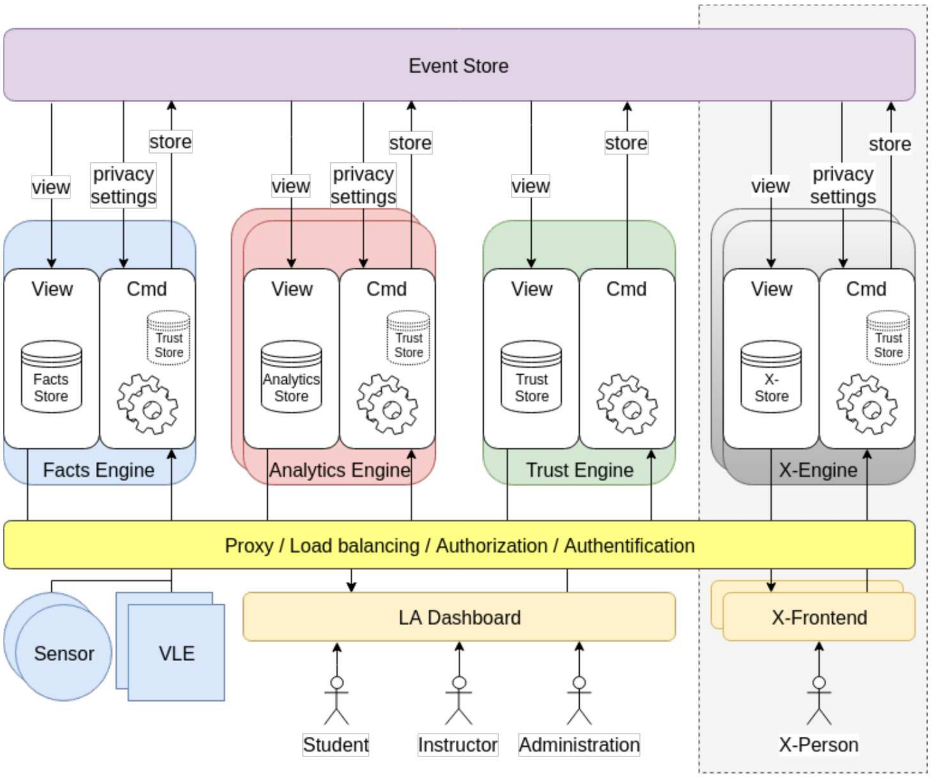
[25] https://twitter.com
[26] https://www.rabbitmq.com
[27] https://www.restapitutorial.com
[28] https://www.json.org

recipes[29] and extensions[30] for custom learning scenarios. The authors used batch processing (b) frameworks such as the well known Apache Hadoop for calculations where access to a complete set of records is required. For instance, when calculating averages and totals. When operations on individual data entering a system are needed, stream processing frameworks such as Apache Storm[31] come into place. These are event-based and make results immediately available. A processing framework is usually coupled with an analytics (e) framework providing the user, among others, with the ability to filter, classify, or cluster data. Traditional SQL storage (d) is for Big Data characteristics such as variety not suited. Therefore, projects with Big Data requirements use NoSQL databases such as MongoDB[32]. In case the project uses xAPI statements, the stores for learning data are commonly referred to as the Learning Record Store. As such, Learning Locker is processing xAPI and using the NoSQL database MongoDB for storage. The figure does not show the visualization (f) technologies since those are subject to constant change and preference. Since the deployment into either a public or private cloud (y) was subject to some publications, we extended the mind map by cloud technologies. The shown technologies can be used by an institute to deploy their LA system on-premise by using, for example, the Docker[33] container technology. Container technologies will allow for portability and scalability if the software developers design the software architecture accordingly.

As it comes to privacy challenges, multiple techniques can be used to address those. Data anonymization, for example, is mostly associated with the challenge of data sharing and publishing. However, it can make it possible to process data if a data subject gave no consent. The *right to be erased* challenges the operators of an LA system to delete all personal data of users if they have the right to and wish so. In this case, legal requirements for data storage must, of course, be observed, for example, in the case of grades. Personal data is not only useful for analytics regarding a particular user but also universal predictions. Hence, instead of deleting personal data, anonymization techniques could be used to keep de-identified data[Victor et al., 2016]. Such techniques are, for example, generalization, suppression, anatomization, permutation, or perturbation. When it comes to cluster computation in untrusted environments such as public clouds, techniques such as homomorphic encryption can help to keep data private while not influencing its advantages[Chen and Yan, 2016]. Not each privacy concern can be tackled by some technique. Some concerns need to be faced by *privacy by design* [Le Métayer, 2016].

[29] https://xapi.com/recipes
[30] https://xapi.com/blog/deep-dive-extensions
[31] https://storm.apache.org
[32] https://www.mongodb.com/
[33] https://www.docker.com

Figure 2: System diagram of the designed Trusted and Interoperable Infrastructure for Learning Analytics (TIILA)



6 Towards a Trusted and Interoperable Infrastructure for Learning Analytics

As described in section 1, the GDPR imposes multiple challenges upon an LA architecture. Within this section, we want to conclude the findings of the two-step literature review and propose an architectural design for a Trusted and Interoperable Infrastructure for Learning Analytics (TIILA), as shown in figure 2. This design intends to install a user-guided privacy control into an interoperable infrastructure.

We based the architecture on the software architecture patterns Event-driven Architecture, Event-sourcing, and Command Query Responsibility Segregation (CQRS). The core consists of the four parts Event Store, Facts Engine, Analytics Engine, and Trust Engine. We will explain the elements in the grey area at the end of this section. The LA Dashboard is a web-based application, while all engines are independent microservices.

All communication within the infrastructure is channeled through a message broker. This message broker contains a dynamic amount of structured queues. These queues sequentially persist all the messages send to them. Each engine interested in a particular type of message or data source subscribes to a specific queue or dynamic group of queues. A dynamic group of queues would be identified by regex on the queue identifiers. This method allows subscribing to all future data sources if their names fit the same regex. Not only external data sources but also internal microservices would publish their data to the queues. The level of indirection provided by using the Publish and Subscribe pattern of a message broker allows for the application of the Event-driven Architecture pattern. All microservices are essentially listening to events triggering their routines. This pattern enables scalability and extendability. Every published message in the infrastructure is considered an immutable event. By using the message broker with this concept of storage, it becomes a type of database called an Event Store. In order to be able to comprehend and audit what happens in the running system, the software architecture pattern is applied to all microservices. This pattern is called Event Sourcing. It allows for a particular extended amount of transparency. Correspondingly to Event Sourcing, an additional effort is taken by making use of the CQRS pattern. The CQRS pattern allows for scalability and interoperability by segregating the command (cmd) from the view routines. This segregation supports multiple denormalized views that are scalable and performant. Scalability and performance are a necessity for architecture with Big Data requirements.

In a running system, multiple data sources would send their data to the infrastructure. This data is sent in its original data format by the usage of pre-built or custom-made Learning Activity Sensors. Using the original data format is a benefit for edge devices with limited processing power. By using the original data format, we also make sure no pre-interpretation has taken place. Nevertheless, we are aware that additional effort needs to take place in order to provide context information to the event. This context information is useful or even necessary to understand the event and ease the analytics workflow. By using xAPI to store the results of all internally generated interpretations, we generate a layer of human-understandable statements. Consequently, a data subject can on request be provided by the Facts Engine with its data in a meaningful standard format (RtPort, RtAcc). Along, if even interested more, a data subject can also be provided with its original data formatted in JSON. Thus, being transparent and enforcing it to verify its data (RtRect). All incoming data needs to be preprocessed in order to apply any privacy settings such as validation, filtering (RtObj), or anonymization (RtObj) to it. The Facts Engine would do this preprocessing.

The Analytics Engine and the Facts Engine would be in sync with the privacy settings to restrict specific processing of users' stored data or filter out user-specific incoming data (RtRProc, RtObj). In case the users want to make use of their *right to be erased*, the Facts Engine would anonymize stored data, and the Analytics Engine would delete any personalized user data where any aggregated data would persist. Modification or deletion in an immutable Event Store is a challenge that is still under investigation by researchers. Solutions include the creation of cleaned queue copies and user-specific encryption of events.

A universal dashboard guides all the interactions of the users. Three user roles are anticipated. Administration, instructors, and students are provided with role constrained widgets. Administrators are customizing those widgets to visualize role-specific content. When first using the dashboard, the users will be welcomed with a wizard explaining them a privacy consent and guiding them to customize the privacy settings to their needs (RtbInfo). Those setting processed in the Trust Engine and stored the event store from where they are propagated to local trust stores in all microservices by the publish-subscribe mechanism of the message broker. Since the privacy preferences of users might change over time, the users shall always be able to adjust their privacy settings and make the system adapt. Additionally, as it comes to algorithmic transparency, automated notifications based on metadata provided by the algorithm, designers shall inform users on updates about processing and data usage (RtbInfo).

Since LA seems to be an ever-innovating field of research, an infrastructure needs to be flexible enough to adapt to changing use cases as well as research personnel. Because of its design based on the three previously explained patterns, the infrastructure allows for a more natural adaption of new analytics use cases. The grey area in figure 2 shows a possible extension by some X-Engine with its X-Frontend. Such an extension could be written in any programming language suitable to a researcher as long as it implements an adapter to the message broker. Since all data is persisted in the event store, it can potentially reprocess data from any data source and any period. The additional workload would only be put on the Event Store, which is easily horizontally scalable. This method allows researchers to create their interpretations of the original data. By subscribing to the events of the Trust Engine, a local Trust Store will be available. Therefore, privacy settings can be applied since individual message brokers allow to restrict access to specific queues X-Engines can even be restricted in access to data.

7 Limitations and Future Work

The main limitation of this publication consists of the number of publications analyzed for the literature review. Within the scope of this study, it was not feasible to cover all publications within the topic.

We took only publications of the last four years into account for our analysis. However, we do not consider this as a significant issue because technology development around Big Data technologies for LA purposes is evolving fast, and our selection provides us with a state-of-the-art overview.

The research questions are also solely focused on the term Big Data where some authors might not frame their architecture as Big Data. For future research, we intend to expand our knowledge by exploring the use of Big Data in domains such as medicine, where information is also personal and sensitive. At the moment, it is vital to acknowledge that the suggested Trusted Learning Analytics infrastructure is yet not fully implemented and operated. We plan to deploy the proposed solution among all partners of the Trusted Learning Analytics project. A long term operation of it will reveal its weaknesses and strengths.

8 Conclusion

The overarching goal of this study was to gain an overview of the state-of-the-art of LA architectures and to examine how data protection is handled in relation to GDPR in such environments. In order to achieve this goal, we conducted a structured literature review. We found diverse designs of LA architecture while answering RQ1. By combining our results with those of RQ3, we identified several technologies that are connected to the different architectural components of an LA system. The mind map shown in figure 1 exposes those connections. At the same time, the investigation of RQ2 revealed little evidence of privacy concerns. However, the results of RQ4 showed several techniques that are suitable to meet the challenges of GDPR. Linking the results of all research questions, we propose a first approach towards a Trusted and Interoperable LA Infrastructure (TIILA). This approach can serve the Learning Analytics community in Europe as a basis for future research and development with the specific requirements the GDPR. Moreover, it may also be relevant to the global LA community or even other data-driven communities from the perspective of recent data scandals[Rosenberg and Dance, 2018].

References

[Berg et al., 2016] Berg, A., Scheffel, M., Drachsler, H., Ternier, S., and Specht, M. (2016). Dutch Cooking with xAPI Recipes: The Good, the Bad, and the Consistent. In *2016 IEEE 16th International Conference on Advanced Learning Technologies (ICALT)*, pages 234–236. IEEE.

[Chaffai et al., 2017] Chaffai, A., Hassouni, L., and Anoun, H. (2017). E-Learning Real Time Analysis Using Large Scale Infrastructure. In *Proceedings of the 2nd international Conference on Big Data, Cloud and Applications - BDCA'17*, pages 1–6, New York, New York, USA. ACM Press.

[Chen and Yan, 2016] Chen, H. and Yan, Z. (2016). Security and Privacy in Big Data Lifetime: A Review. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10067 LNCS, pages 3–15. Springer, Cham.

[Chen et al., 2017] Chen, J., Tang, J., Jiang, Q., Wang, Y., Tao, C., Zhang, X., and Liao, J. (2017). Research on Architecture of Education Big Data Analysis System. In *2017 IEEE 2nd International Conference on Big Data Analysis, ICBDA 2017*, pages 601–605. IEEE.

[Chen et al., 2016] Chen, Y., Chen, H., Gorkhali, A., Lu, Y., Ma, Y., and Li, L. (2016). Big Data Analytics and Big Data Science: A Survey. *Journal of Management Analytics*, 3(1):1–42.

[Dahdouh et al., 2018] Dahdouh, K., Dakkak, A., Oughdir, L., and Messaoudi, F. (2018). Big data for online learning systems. *Education and Information Technologies*, 23(6):2783–2800.

[European Parliament, 2016] European Parliament, C. o. t. E. U. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Da. *Official Journal of the European Union*, L119:1–88.

[Fink, 2013] Fink, A. (2013). *Conducting Research Literature Reviews : From the Internet to Paper*. SAGE Publications Inc, Thousand Oaks, United States.

[Furukawa et al., 2017] Furukawa, M., Yamaji, K., Yaginuma, Y., and Yamada, T. (2017). Development of Learning Analytics Platform for OUJ Online Courses. In *2017 IEEE 6th Global Conference on Consumer Electronics (GCCE)*, pages 1–2. IEEE.

[Gasevic et al., 2017] Gasevic, D., Siemens, G., and Rose, C. P. (2017). Guest Editorial: Special Section on Learning Analytics. *IEEE Transactions on Learning Technologies*, 10(1):3–5.

[Gómez-Berbís and Lagares-Lemos, 2016] Gómez-Berbís, J. M. and Lagares-Lemos, Á. (2016). ADL-MOOC: Adaptive Learning Through Big Data Analytics and Data Mining Algorithms for MOOCs. In *Technologies and Innovation*, pages 269–280. Springer, Cham.

[Greller and Drachsler, 2012] Greller, W. and Drachsler, H. (2012). Translating Learning into Numbers: A Generic Framework for Learning Analytics. *Journal of Educational Technology & Society*, 15:42–57.

[Grover and Aulakh, 2017] Grover, C. and Aulakh, M. K. (2017). A Literature Review: Big Data Privacy and Security. *3rd International Conference on Emerging Trends in Engineering and Management Research*, pages 277–283.

[Hoel et al., 2017] Hoel, T., Griffiths, D., and Chen, W. (2017). The Influence of Data Protection and Privacy Frameworks on the Design of Learning Analytics Systems. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference on - LAK '17*, pages 243–252, New York, New York, USA. ACM Press.

[Hu et al., 2019] Hu, H., Zhang, G., Gao, W., and Wang, M. (2019). Big data analytics for MOOC video watching behavior based on Spark. *Neural Computing and Applications*, 0123456789.

[Huang et al., 2016] Huang, X., Ge, W., and Liu, Y. (2016). Design and Implementation of E-Training Decision-Making System. In *2015 International Conference of Educational Innovation Through Technology, EITT 2015*, pages 24–28. IEEE.

[Jiangbo Shu et al., 2017] Jiangbo Shu, Xu Wang, Li Wang, Zhaoli Zhang, Hai Liu, Qianqian Hu, and Min Zhi (2017). Exploration on College Education Big Data Open Service Platform. In *2017 IEEE 2nd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, pages 161–165. IEEE.

[Katal et al., 2013] Katal, A., Wazid, M., and Goudar, R. H. (2013). Big Data: Issues, Challenges, Tools and Good Practices. In *2013 Sixth International Conference on Contemporary Computing (IC3)*, pages 404–409. IEEE.

[Laveti et al., 2017] Laveti, R. N., Kuppili, S., Ch, J., Pal, S. N., and Babu, N. S. C. (2017). Implementation of Learning Analytics Framework for MOOCs using State-of-The-Art In-memory Computing. In *2017 5th National Conference on E-Learning and E-Learning Technologies, ELELTECH 2017*, pages 1–6. IEEE.

[Le Métayer, 2016] Le Métayer, D. (2016). Whom to Trust? Using Technology to Enforce Privacy. In *Enforcing Privacy*, volume 395, pages 395–437. Springer, Cham.

[Li et al., 2017] Li, Y., Li, P., Zhu, F., and Wang, R. (2017). Design of Higher Education Quality Monitoring and Evaluation Platform Based on Big Data. In *2017 12th International Conference on Computer Science and Education (ICCSE)*, pages 337–342. IEEE.

[Liu et al., 2018] Liu, W., Fan, W., Li, P., and Li, L. (2018). Survey of Big Data Platform Based on Cloud Computing Container Technology. In *CISIS 2017: Complex, Intelligent, and Software Intensive Systems*, pages 954–963. Springer, Cham.

[Logica and Magdalena, 2015] Logica, B. and Magdalena, R. (2015). Using Big Data in the Academic Environment. *Procedia Economics and Finance*, 33:277–286.

[Lopez et al., 2017] Lopez, G., Seaton, D. T., Ang, A., Tingley, D., and Chuang, I. (2017). Google BigQuery for Education. In *Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale - L@S '17*, pages 181–184, New York, New York, USA. ACM Press.

[Matsebula and Mnkandla, 2017] Matsebula, F. and Mnkandla, E. (2017). A Big Data Architecture for Learning Analytics in Higher Education. In *IEEE Africon 2017 Proceedings*.

[Miloslavskaya and Makhmudova, 2016] Miloslavskaya, N. and Makhmudova, A. (2016). Survey of Big Data Information Security. In *2016 IEEE 4th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW)*, pages 133–138. IEEE.

[Muslim et al., 2018] Muslim, A., Chatti, M. A., Bashir, M. B., Varela, O. E. B., and Schroeder, U. (2018). A Modular and Extensible Framework for Open Learning Analytics. *Journal of Learning Analytics*, 5(1):92–100.

[Nelson and Olovsson, 2016] Nelson, B. and Olovsson, T. (2016). Security and Privacy for Big Data: A Systematic Literature Review. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 3693–3702. IEEE.

[Petrova-Antonova and Ilieva, 2019] Petrova-Antonova, D. and Ilieva, S. (2019). Using Big Data Value Chain to Create Government Education Policies. pages 42–49.

[Rabelo et al., 2015] Rabelo, T., Lama, M., Amorim, R. R., and Vidal, J. C. (2015). SmartLAK: A Big Data Architecture for Supporting Learning Analytics Services. In *Proceedings - Frontiers in Education Conference, FIE*, volume 2014, pages 1–5. IEEE.

[Rosenberg and Dance, 2018] Rosenberg, M. and Dance, G. J. (2018). 'You Are the Product': Targeted by Cambridge Analytica on Facebook.

[Salleh and Janczewski, 2016] Salleh, K. A. and Janczewski, L. (2016). Technological, Organizational and Environmental Security and Privacy Issues of Big Data: A Literature Review. *Procedia Computer Science*, 100:19–28.

[Salvador et al., 2017] Salvador, J., Ruiz, Z., and Garcia-Rodriguez, J. (2017). Big Data Infrastructure: A Survey. In *Biomedical Applications Based on Natural and Artificial Computing*, pages 249–258. Springer International Publishing.

[Santoso and Yulia, 2017] Santoso, L. W. and Yulia (2017). Data Warehouse with Big Data Technology for Higher Education. *Procedia Computer Science*, 124(2017):93–99.

[Santur et al., 2016] Santur, Y., Karakose, M., and Akin, E. (2016). Improving of Personal Educational Content Using Big Data Approach for Mooc in Higher Education. In *2016 15th International Conference on Information Technology Based Higher Education and Training (ITHET)*, pages 1–4. IEEE.

[Schneider et al., 2015] Schneider, J., Börner, D., Van Rosmalen, P., and Specht, M. (2015). Augmenting the senses: A review on sensor-based learning support. *Sensors*

(Switzerland), 15(2):4097–4133.

[Srinivasan et al., 2015] Srinivasan, A., Abdul, Q. M., and Vijayakumar, V. (2015). Hybrid Cloud for Educational Sector. *Procedia Computer Science*, 50:37–41.

[Swathi et al., 2017] Swathi, R., Kumar, N. P., KiranKranth, L., Madhav, L. S., and Seshadri, R. (2017). Systematic Approach on Big Data Analytics in Education Systems. In *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 420–423. IEEE.

[Tang et al., 2015] Tang, J. K., Xie, H., and Wong, T. L. (2015). A Big Data Framework for Early Identification of Dropout Students in MOOC. In *Communications in Computer and Information Science*, volume 559, pages 127–132. Springer, Berlin, Heidelberg.

[Taylor and Munguia, 2018] Taylor, S. and Munguia, P. (2018). Towards a data archiving solution for learning analytics. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge - LAK '18*, pages 260–264, New York, New York, USA. ACM Press.

[Thangaraj and Balamurugan, 2017] Thangaraj, M. and Balamurugan, S. (2017). Survey on Big Data Security Framework. In *Knowledge Management in Organizations*, pages 470–481. Springer, Cham.

[Vagliano et al., 2018] Vagliano, I., Gunther, F., Heinz, M., Apaolaza, A., Bienia, I., Breitfuss, G., Blume, T., Collyda, C., Fessl, A., Gottfried, S., Hasitschka, P., Kellermann, J., Kohler, T., Maas, A., Mezaris, V., Saleh, A., Skulimowski, A. M., Thalmann, S., Vigo, M., Wertner, A., Wiese, M., and Scherp, A. (2018). Open Innovation in the Big Data Era with the MOVING Platform. *IEEE Multimedia*, 25(3):8–21.

[van Merriënboer et al., 2002] van Merriënboer, J. J. G., Clark, R. E., and de Croock, M. B. M. (2002). Blueprints for complex learning: The 4C/ID-model. *Educational Technology Research and Development*, 50(2):39–61.

[Victor et al., 2016] Victor, N., Lopez, D., and Abawajy, J. H. (2016). Privacy Models for Big Data: A Survey. *International Journal of Big Data Intelligence*, 3(1):61.

[Volk et al., 2017] Volk, M., Bosse, S., and Turowski, K. (2017). Providing Clarity on Big Data Technologies: A Structured Literature Review. In *2017 IEEE 19th Conference on Business Informatics (CBI)*, pages 388–397. IEEE.

[Yang and Huang, 2016] Yang, S. J. and Huang, C. S. (2016). Taiwan's Digital Learning Initiative and Big Data Analytics in Education Cloud. In *2016 5th IIAI International Congress on Advanced Applied Informatics*, pages 366–370. IEEE.

[Ye et al., 2016] Ye, H., Cheng, X., Yuan, M., Xu, L., Gao, J., and Cheng, C. (2016). A Survey of Security and Privacy in Big Data. In *2016 16th International Symposium on Communications and Information Technologies (ISCIT)*, pages 268–272. IEEE.

[Zhang et al., 2016] Zhang, G., Yang, Y., Zhai, X., Yao, Q., and Wang, J. (2016). Online Education Big Data Platform. In *2016 11th International Conference on Computer Science & Education (ICCSE)*, pages 58–63. IEEE.

[Zhao et al., 2017] Zhao, Z., Wu, Q., Chen, H., and Wan, C. (2017). Learning Quality Evaluation of MOOC Based on Big Data Analysis. In Qiu, M., editor, *Smart Computing and Communication*, pages 277–286. Springer International Publishing.