Mission Control for Learning and Performance

The enterprise learning ecosystem in the age of advanced streaming data architectures and real-time learning analytics

Written by Shelly Blake-Plock and Will Hoyt and published September 26th, 2018 by Yet Analytics, Inc.

Objectives

When designing and implementing a scalable learning ecosystem, the most important design considerations are those of data logistics, system architecture, and data design. When these considerations are not at the forefront of the design process, the ecosystem will most likely experience unpatchable issues which hinder its potential to provide meaningful learning experiences. As in any system, the foundation on which a learning ecosystem is built determines the capabilities and the extensibility of the learning ecosystem.

The demands of scale put a significant amount of pressure on the variety of technologies that comprise the modern learning stack. Take for example a system that implements a recommendation engine to provide the learner with suggestions on how to optimize their learning experience via the presentation of relevant and impactful content. In order for the recommendation engine to be responsive and operate at near real time speeds, it must have timely access to massive amounts of data. Each consumed data point must contain sufficient but not excessive contextual information and the data must be stored in a scalable database capable of meeting the needs of the recommendation engine as well as handling the massive amount of data produced by a modern learning ecosystem. If those conditions are not met, a bottleneck ensues. This bottleneck reduces the speed at which the recommendation engine can function. It does not matter if the bottleneck is due to input data retrieval latency, prolonged computational time due to processing excessively large data points, or the input data source failing to provide any data at all. The result is the same: degradation in system performance culminating in a substandard user experience for the learner by increasing wait time for the

201 E. Baltimore St. Ste. 630, Baltimore, MD 21202 USA
info@yetanalytics.com

recommendation engine to issue the suggestions which facilitate an optimal and personalized learning experience.

This is a scale problem. Because scale is such a daunting challenge, it is necessary to look to other technology domains, specifically those in which scale has long been a crucial part of business, in order to determine the ideal methodologies for handling scale — technologies where success is considered the accurate creation of a user profile based on the user's activity across a system — namely, the businesses of social media and advertising technology. In those domains, streaming data architectures have become the rule. Most visible is Apache Kafka, a stream processing technology that was originally built within LinkedIn and then open sourced. By the end of 2017, over a third of Fortune 500 companies had implemented Kafka and over the course of that year, LinkedIn, Microsoft, and Netflix had each passed over one trillion data statements over Kafka.[1] The reason? Kafka was built to help large organizations manage the scale of streaming data. And prior solutions — even Big Data solutions like Hadoop — were too slow.

Scale has often been a second thought in learning technologies because the capabilities and therefore the related data capture of this range of technologies has been relatively meek when compared to the use of data across other business concerns. With the introduction of robust, high resolution data capabilities within the learning space, this is now changing. In some learning scenarios, such as hybrid simulations that collect copious amounts of biometric data or cyber defense training where granular micro-behaviors are tracked, the magnitude of activity data being produced can only be handled efficiently by streaming architectures and is comparable to the amount of data generated and utilized by sales or marketing teams in the enterprise. In fact, bolstered by the development of technologies designed to leverage activity streams in the learning space — developments such as the Experience API (xAPI) which was developed by the Advanced Distributed Learning Initiative and is being leveraged at scale in deployments of learning solutions by the United States Office of Personnel Management as well as in Fortune 500 corporations — learning organizations will soon find themselves in a position where the ability to handle scale is not an option. Rather, it is the most crucial technological requirement.

When working at scale with data streams, one of the key findings that businesses have made is that there is no such thing as one data model to rule them all; multiple models are necessary to accurately represent and capture the complexity of and nuances within the underlying data generation processes. For the learning organization this means that it is neither viable to forgo standardization in the description of learning activity data nor is it scalable to attempt to coerce all data about a learner into a single learning activity data model. Rather, a more sophisticated approach to data design is required to ensure that the data produced by a learning ecosystem is

[1] https://www.techrepublic.com/article/an-inside-look-at-why-apache-kafka-adoption-is-exploding/

actionable. The architectural decisions made concerning data will have a first order effect on the digital transformation of the learning organization.

Digital Transformation

And so it is that for digital transformation to take hold in the field of learning — something that is essential if learning technologies are to serve the growth needs of global organizations — it is necessary not only to observe and gain insight from the architectural decisions and data logistic strategies of successful businesses, specifically those whose success is dependent on processing and handling data at massive scale, but also consider data design and modeling from the onset of any learning technology project so that the success and longevity of said project will not be crippled by the inability of other systems to extract, interpret, utilize, and report on the data produced by the learning technology.

xAPI: the Specification for Learning Activity Data

Standardized formats such as xAPI should be used to record event based, micro-level, learning activity data. These types of data formats are inherently meaningful, readable and understandable to human analysts while also being machine readable. xAPI Profiles, a companion specification to xAPI which serves to describe a linked data approach to xAPI data design, validation and modeling, allow for a human domain expert to encode their knowledge in a way which is useful to both human-based and machine-based processes. An xAPI Profile serves as a methodology for defining macro-level patterns of behaviors by the micro-level events which compose them. These defined patterns of behaviors can then serve as the basis of business and organizational processes such as the automation of key performance indicators and alerts for interventions among teams and departments.

Alignment of Alternate Data Models and Non-Standard Data

Other types of data produced within learning ecosystems, such as anonymized aggregates about a piece of learning content, are not typically recorded as standalone xAPI statements but can still be utilized either alongside or within xAPI statements. The way in which they are utilized can be defined within an xAPI Profile, for instance as a result or context extension, such that the xAPI Profile not only models the event-level user data but also models how those individual data points relate back to top level processes and metrics such as the calculation of content popularity or effectiveness. Due to these properties, business data and information gathered in alternate forms should be gathered alongside the event-level xAPI data such that the two can be transported, processed, and utilized within the same data streaming architecture. Non-xAPI data formats do not hinder xAPI in any way, instead, they provide additional context which helps

to make learning profiles more meaningful than those based either on only xAPI data or on only non-xAPI data.

Metadata

Metadata should be leveraged to bolster connections and improve the management of identities, activities, and outcomes across the system. In the end, structured data will mingle in a data stream among semi and unstructured data such that microservices feeding on the stream will use each type of data for the purposes defined by the functions and parameters of each individual service. An additional advantage of the streaming architecture is the ability to push macro and attribute-oriented data such as economic, workflow, or productivity data to the same real-time dashboard as the activity stream data. The ability to make all types of data easily available allows for the creation of dashboards which present a holistic view of a learner by utilizing all aspects of their digital footprint, regardless of how it was recorded. When that digital footprint contains well crafted xAPI statements purpose built for learning analytics, dashboards can be utilized for informing and supporting data driven decisions. One of the most attractive features of streaming architectures is how they actually make all types of data easily accessible. In a streaming architecture, applications and services consume data from the stream independently of each other such that adding a new data consumer or removing an existing one does not affect the other consumers. This property results in the ability to easily iterate on a learning ecosystem such that new features can be quickly implemented, without generating technical debt, as the needs of the learning organization and learning ecosystem evolve over time.

State of the Art

From the strategic perspective, when the objective is to train up new recruits, advance skills in an existing labor force, and increase the speed at which a large organization is able to take on complex tasks, it is imperative that the learning ecosystem — comprised of content, technology, and instructional design all working together to fulfill the promise of research-validated learning methodologies — be designed to serve the goals of the organization.

But as a tactical matter, as an organization adds applications for the purpose of serving the needs of learners and related constituents, the learning ecosystem itself becomes increasingly convoluted and points of failure increase in number each time a new application is added. The existing model of point-to-point technology integration across a server-oriented infrastructure is endemic of the type of solution that does nothing to fix this problem. Unfortunately, this model is still used fairly consistently within domains in which it is no longer a viable solution, typically resulting in forward thinking projects not reaching their full potential. This approach is

inconsistent with best practices in the application of data streaming in Cloud architectures and is not a suitable fit for learning activity data that by design will outgrow the capabilities and exacerbate the latencies inherent in a point-to-point fix.

Streaming data architectures and the supporting data logistics and analytics engines necessary to support the real-time learning ecosystem are found in technologies including Apache Kafka, Apache NiFi, and the variety of microservice processors that serve the data streams of systems designed in accordance with the Kappa Architecture — an approach to data architecture that treats everything as data in a stream. The modern real-time learning ecosystem should be built on these streaming data architectures which have been tested and proven effective at web-scale in social media and advertising technology.

The Components Required to Create Mission Control for Learning and Performance

Mission Control for learning and performance will be comprised of a variety of data and information technology assets working across an ecosystem to provide end-users with access to real-time data about learners.

The core components of the new enterprise learning ecosystem are:

Data Stores

An example of a data store is the Learning Record Store (LRS) which both stores and validates the authentic conformance of xAPI data across a learning ecosystem. It is a requirement that the LRS successfully pass the xAPI LRS Test Suite hosted by the Advanced Distributed Learning Initiative (ADL).[2] Similarly the conformant LRS should be used to validate the conformance of all data sources intending to produce xAPI data statements to be used within the learning ecosystem. Data sources which are found to not produce valid xAPI data should either be reconsidered, replaced, or augmented via xAPI translations applied to the activity API output of the data source.

Data Sources

Data sources include Learning Record Providers such as a learning management system (LMS), a learner experience platform (LXP), mobile learning applications, computer-based and

[2] https://lrstest.adlnet.gov/

hybrid simulations, serious games, virtual reality and augmented reality experiences, wearables delivering biometric and location-based data during training exercises, and Internet of Things sensors and devices capturing physical and environmental information about the learner and the learning environment. Some learning record providers will be native xAPI while others will need to be translated in xAPI. Data sources also include those business systems such as Human Resources Information Systems (HRIS) and databases housing personnel and business records. These data sources rarely support native xAPI and design decisions can be made as to what degree of integration should be done in any particular implementation in a learning ecosystem. Data sources also include contextual and “3rd party” data including that of the web and social media as well as data gathered in the aggregate from open data and open government data projects. In the case of activity data collected from these sources, such as activity generated by common business web services, it may be appropriate to transform into xAPI. For data which is generally tabular in nature, this is not uniformly necessary. It is therefore important that expertise be applied in choosing which data streams to coerce into which data models to best serve the scalability of the learning ecosystem.

Data Analytics

Data analytics serve two roles — providing insight to human analysts and other end users and providing computational functions and parameters for the automated work of machines. In the modern learning ecosystem, it is especially important that the data analytics engines employed are able to process and deliver real-time data. Lesser-capability data analytics resources can cause a significant bottleneck in a learning ecosystem. While there is no hard-and-fast rule, in order to serve the needs of all of the types of data which may flow through the data stream, it is recommended that the data analytics engines used be able to work seamlessly with NoSQL databases and that they are able to support batched, scheduled, persistent, and live data feeds.

Data Logistics

Identity, integration, and processing are all served by data logistics. This is the construct providing the ability for data to flow through the system in the proper way. In a streaming architecture, it is necessary to provide an automated identity management system at point of ingress to the stream. The stream itself leverages either Apache Kafka or a similar commercial stream processor such as AWS Kinesis. The data management and connectors between processes are best served in a streaming architecture by the application of logistics technologies meant for massive scale such as Apache NiFi. In data logistics, it is especially important to leverage open source resource where possible so as to cut down on the potential for technical debt and vendor lock-in.

Data Processors

The data processors use the assets delivered via the logistic schema to make use of identity, data collection, and data transformation for the purpose of powering alignment of activity with competency frameworks and recommendation engines. In addition to Kafka as mentioned above, the processors in the system include recommendation and competency alignment technologies. To the extent possible, recommendation engines should be built open source so as to make auditable the manner in which recommendations are made. Purely proprietary recommendation systems may inhibit the ability for a user to understand how or why a recommendation was made — an issue with significant implications when people's careers and future success are at stake such as is the case with a learner. Competency alignment should be based on published and openly available competency frameworks.

Data Models

The data models include the specified and/or standardized definitions and resources governing the design of the data, the vocabularies used to support that design, the data transport mechanisms, and the method of storing and validating data. To account for the variety of activity moving across the enterprise learning ecosystem, xAPI should be implemented as the data model. Additionally, the companion specification to xAPI — the xAPI Profiles specification[3] — provides clear guidance on the use of xAPI within the context of semantic and linked data. It is crucial that xAPI data designed to work within a learning ecosystem be properly modelled to the specifications of xAPI Profiles. In a stream processing system such as described here, there is flexibility in the ability to incorporate a variety of data models, however wherever possible complexity should be limited through the alignment of data with known models or applied processes for dealing with data variety. In the end it is a matter of "junk in, junk out", meaning not even the most complex learning ecosystem will be able to make use of data that was designed poorly from the outset. For that reason, data modeling is a key first step in the design of any learning ecosystem.

An Illustrated Framework

The following illustration portrays a modern learning ecosystem featuring all of the necessary components as described above. In this model, contextual and business system data may flow into the same stream as learning activity data. All of the data passing through the stream ends up in a data lake. The applications and data stores adjacent to the stream subscribe to data passing through the stream via the custom microservices portrayed as arrows. Where early in the days of xAPI, there was a push to use the LRS as the source of record for everything in the

[3] https://github.com/adlnet/xapi-profiles

learning ecosystem, in practice that has proven to be a naive and unscalable implementation. The scalable method is that portrayed below where the data is managed at the point of ingress into the stream and is then subscribed to by the various consumers of that data. Highlighted in green here are the components which are the primary users of xAPI data. The red line represents the flow of data processed through the LRS back into the application layer of learning record consumers. The mission control dashboards themselves will render data visualizations based on the xAPI data, data pushed by other business intelligence systems, and select queryable data captured in the data lake.

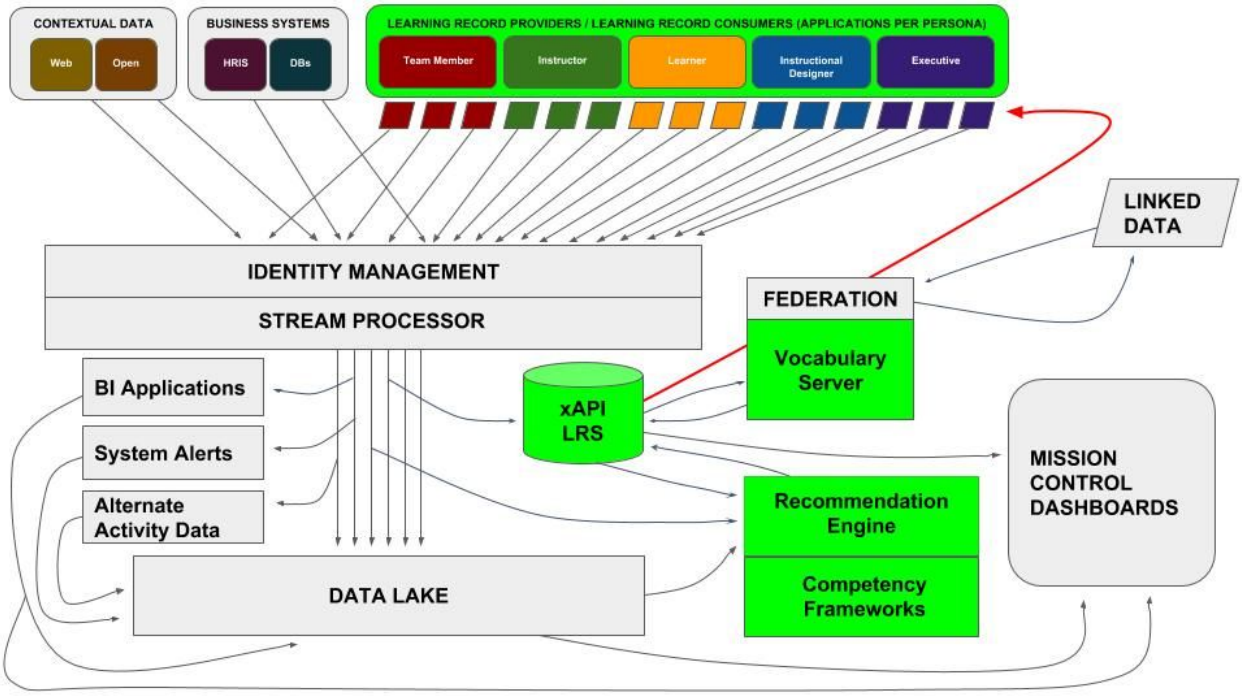

Figure 1. Illustration of a Streaming Data Architecture serving Multiple Data Models in the Modern Learning Ecosystem. Copyright 2018 Yet Analytics, Inc.

As demonstrated in this illustration, the components are affiliated through a shared architecture. The Advanced Distributed Learning Initiative refers to this model of ecosystem behavior as the Total Learning Architecture (TLA) and is currently building specs based around the application of xAPI to an ecosystem of learning technologies. As ADL states: “The TLA is not a particular training device or educational tool. Rather, it is the “glue” that connects all other learning technologies into an integrated, coherent system.”[4] The attention paid to alignment and integration is key. Alignment and integration with these data models — especially xAPI, the core ADL-tested data model for learning and performance activity data — and other learning and

[4] https://adlnet.gov/projects

business process data standards will help to ensure a more interoperable and cost-efficient learning ecosystem.

Why This Matters

The architecture described above will support the real-time ability to generate and maintain a learner profile. This learner profile is the verifiable record of the learner's experience, prior-knowledge, training success, and indication of readiness for any specific job.

The ability to capture, extract, analyze, and visualize learning activity data opens up the opportunity to support formative assessment, promote ongoing and real-time growth and development, and provide personalized and mission-relevant learning experiences on-demand and at scale. In any organization facing massive growth, consistently high attrition, or a retiring generation of leaders and knowledge gatekeepers, implementing a modern enterprise learning ecosystem is table stakes. In high stakes environments where learning or not learning means the difference between winning or losing, a poorly designed learning ecosystem implementation can have ruinous results.

Key Concerns

When fully leveraging technologies accessible and in implementation right now — as opposed to in the future (it should be stressed that nothing discussed in this paper is science fiction) — the learning organization will be able to stand up a data command center for education and training. The user interface will be a mission control for learning and performance featuring automated real-time data visualizations and reports designed to auto-generate based on the ability to track performance and progress towards competencies, achievements, and credentials. But the dashboards that comprise the organization's mission control will only be as good as the data that serves it and the data architecture that provides the connectivity between all aspects of the system. Therefore, it is crucial that care be give to data design and the architecture that will support it. Starting with analytics and the end-metrics that will be made accessible by the dashboard visualizations in mind will produce the best results. This is a design process as much as anything else.

The enterprise learning ecosystem — built in the context of a digital paradigm shift across business towards streaming data architectures — must be designed to meet the demands of massive scale and will be engineered to support on-demand global learning, assessment, and reporting. But just as scale-mindfulness is crucial to the successful implementation of learning technologies, likewise neither the designers of learning experiences nor the learners themselves should be hindered by artificial barriers established to support scale. Therefore, the enterprise

learning ecosystem must be extensible and include the components necessary to support competency tracking, business model and organizational career pathing alignment, and whatever digital and offline experiences instructional designers and the needs of a program may demand of it. While nothing can be future-proofed, the technology will at a minimum be built to support open extensibility in an effort to provide a foundation for future developments rather than be a hindering technical debt to the next generation of innovation.

Implementation

The implementation of a learning ecosystem requires six key elements:

- Data assessment and needs analysis
- Data design and architecture blueprints
- Initial key integrations and prototype implementation
- Testing and iteration
- Full integration
- Quality assurance, production implementation, and delivery of full documentation

Depending on the complexity of the implementation, the length of time from initial assessment to production could range from 24 to 48 months.

Staff needed to carry out the implementation include:

- A principal investigator who is a subject matter expert in xAPI and the Total Learning Architecture
- A consulting subject matter expert who will control the data assessment and needs analysis
- A lead software engineer with experience implementing xAPI; building, testing and deploying full scale LRS solutions; and setting up and deploying enterprise streaming architectures
- A data engineer with experience designing xAPI Profiles and modeling activity data for the purpose of producing learner profiles aligned to competency frameworks
- A small team of three to four software engineers to develop integrations and microservices
- A dev ops engineer to maintain and support the architecture and learning ecosystem
- A project manager to control the schedule and scope of implementation following a standardized practice of product management

Depending on the outcome of the data assessment and needs analysis, some configuration of the following technologies will be required to support the implementation of the learning ecosystem:

- Data Stores: xAPI Learning Record Store, Data Lake, Databases
- Data Sources: Content Platform Integration, Registration and Tagging Interface, Content-Competency Alignment, Content and Learner Experience Web Portal, User and Administrator Management
- Data Analytics: Analytics Engine, Neural Networks and Deep Learning Models, Persona-facing Customizable Dashboards and Reporting Infrastructure, Auditable Data and Security Layers
- Data Logistics: Kappa Implementation, NiFi Processors, Kafka Streams Ingress Layer and Processor, Microservices
- Data Processors: Recommendation Engine, Competency Framework, (potentially) a Credentialing Service
- Data Model: xAPI Profile Server, xAPI Profiles, Vocabulary Registry, Semantic Alignment, Alignment of alternate standards and non-activity data to the requirements of the learner profile

Take Away

The implementation of an enterprise learning ecosystem capable of supporting mission control for learning and performance requires expertise in data logistics, system architectures, and data design. Because the data sources feeding into the learner profiles will be of a heterogenous nature — both in terms of data model and data availability — it is necessary to implement an advanced streaming data architecture in order to scale the learning ecosystem. This foundation will provide a basis for future extensibility and iteration in a way that is most scalable, cost-effective (because it is not necessary to rebuild the whole system each time a future data source or data consumer capability is added), and secure (because all authentication and permissioning happens at the point of ingress to the stream). Based directly on commercial infrastructures designed to support real-time data in business, this learning ecosystem will support the demands of real-time learning and the development of authentic and auditable learner profiles.