Usability TRL4 Report: DATASIM

Data and Training Analytics Simulated Input Modeler and Technology Readiness Analysis

31 March 2020

This work was supported by the U.S. Advanced Distributed Learning (ADL) Initiative (HQ0034-19-C-0061). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the ADL Initiative or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes.

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)* 04-30-2020	**2. REPORT TYPE** Research Paper	**3. DATES COVERED** *(From - To)* 31-03-2019 to 31-03-2020

4. TITLE AND SUBTITLE

DATASIM TRL 4 Usability Report

5a. CONTRACT NUMBER

HQ0034-19-C-0061

5b. GRANT NUMBER

N/A

5c. PROGRAM ELEMENT NUMBER

0603769D8Z

6. AUTHOR(S)

Shelly Blake-Plock

5d. PROJECT NUMBER

UPR_030603769D8Z776

5e. TASK NUMBER

20_0400_0603769D8Z

5f. WORK UNIT NUMBER

HQ0642922724

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Yet Analytics
201 E. Baltimore St., Ste. 630
Baltimore, Maryland 21202

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

OUSD Personnel & Readiness
Advanced Distributed Learning Initiative
13501 Ingenuity Drive, Suite 248
Orlando, Florida 32826

10. SPONSOR/MONITOR'S ACRONYM(S)

OUSD/P&R/FE&T/ADLI

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Distribution A

13. SUPPLEMENTARY NOTES

14. ABSTRACT

DATASIM is an open-source application that will provide a valid means of producing the data sets necessary to benchmark and stress test the Total Learning Architecture (TLA) and distributed learning acquisitions. Additionally, DATASIM can help learning scientists, ISDs, IT and technology acquisition, and decision-making stakeholders to determine the effectiveness of xAPI data design and implementation across the TLA. This document analyzes the current state of the DATASIM project from the perspective of a DoD Technology Readiness Assessment.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Nick Armendariz
U	U	U	UU	19	**19b. TELEPHONE NUMBER** *(Include area code)* 407-381-5550

Reset

Usability TRL4 Report: DATASIM

Data and Training Analytics Simulated Input Modeler and Technology Readiness Analysis

Prepared by Shelly Blake-Plock, PI, Yet Analytics, Inc.

Submitted to ADL on March 31, 2020

1. Purpose of this Document

The purpose of this document is to analyze the current state of the Data and Training Analytics Simulated Input Modeler (DATASIM) project from the perspective of a DoD Technology Readiness Assessment.

2. Project Description

DATASIM is a software project capable of generating simulations resulting in realistic xAPI data. The environment uses xAPI Profiles to model behavior for a cohort of simulated actors. These datasets may be used to benchmark and stress-test components of the Total Learning Architecture (TLA) and other distributed learning projects.

Additionally, DATASIM helps learning scientists, engineers, ISDs, IT staff and decision-making stakeholders to determine the effectiveness of xAPI data design and implementation across the TLA.

DATASIM is open source under the Apache License Version 2 and is available on GitHub here:

Backend <https://github.com/yetanalytics/datasim>
Frontend <https://github.com/yetanalytics/datasim-ui>

3. Technology Readiness Level 4 Criteria

A TRL 4 Software Prototype is classified as a stand-alone prototype solving a synthetic full-scale problem, or stand-alone prototype processing fully representative data sets.

3.1 Architecture

At Readiness Level 4 for Software Technology, as defined by the DoD, the architecture of the system may still be relatively primitive with regard to efficiency and robustness. The primary goal of the architecture of a software project at this stage is to integrate the individual components in order to fully establish that they will work together.

While aspects of the prototype may not be fully explored, all software architecture at this stage should begin to address interoperability, reliability, maintainability, extensibility, scalability, and security issues. While it is unlikely that the prototype will be fully matured in any of these aspects at this stage, it is important that the system is architected with explicit room for enhancement in all of those categories. It is ideal if the prototype at this stage serves as a framework to mature these components at a later maturity level, rather than exist as a one-off demonstration without room for expansion.

3.2 Testing

At this stage of Technology Readiness the prototype software should be able to be tested either as a standalone with a synthetic full-scale problem, or with fully representative datasets.

The results of testing at this stage should demonstrate the software’s capabilities to accomplish its stated purpose, and should serve as proof of the efficacy of the solution such that it can be comfortably adapted to solve real world use cases in the next phase. The results of testing at this phase may still show room for growth in areas of efficiency and robustness, though testing at this stage should establish efficiency benchmarks and known limitations.

3.3 Interfaces and User Interactions

The state of the user interaction aspects of a software technology project expected at Readiness Level 4 may vary depending on the nature of the project. If the nature of the project is fundamentally tied to an improvement in user interaction, then this aspect may be highly matured in order to demonstrate the solution’s efficacy.

If the software project instead represents a prototype in a problem space other than user interaction, it is likely that the user interaction aspects will be more primitive and focused on function over form. It may be common at this stage to demonstrate a software product using a command line interface (CLI) or basic UI. What is most important at this stage is that via defined interaction with the product, the demonstrated outcomes are reproducible and reliable. To this end, extensive and complete documentation of usage patterns is essential.

4. DATASIM Alpha

4.1 Component Integration

The individual aspects of the DATASIM Project have all been fully integrated into a complete solution. This includes statement generation, all input types, xAPI Profile and other input validation, user interface (CLI and web application), LRS integration, and software bundling for deployment. The prototype has a JVM compatible backend and a web application frontend with broad browser compatibility. The only notable separation of components is that the frontend and backend are two separate code repositories, and while they seamlessly integrate in a deployment they are built and deployed as separate packages. Rather than this being a reflection of an incomplete state of prototype integration, it instead reflects a design choice consistent with current industry best practices in application architecture.

4.2 Simulation Generation

The xAPI generation engine is fully implemented and working as expected. It uses a deterministic random time series generator to provide time-based “motivation” to cohort activities. It implements full traversal of valid xAPI Profiles and bases statement generation accurately on Profile Patterns and Templates. It has a sufficiently high throughput to generate millions of statements in minutes and at this time can be scaled vertically with additional hardware.

4.3 Inputs and Validation

At this stage of development, DATASIM is capable of accepting the four primary JSON inputs (Profiles, Actors, Alignments, Sim Parameters) or alternatively the full Simulation Specification via multiple methods. They can each be input the following ways:

- By Filename via CLI
- By “Upload Dialog” via provided UI
- By URL via provided UI

At this time, all inputs are validated and errors are presented to the user. The most mature (and sophisticated) validation is currently performed on the xAPI Profile input type, wherein it offers deep specification-based validation on provided Profiles. The remaining specifications are at this time primarily checking format and protocol as the ruleset is much smaller than for xAPI Profiles.

4.4 Interoperability

At this stage of the prototype, the only relevant interoperability is with Learning Record Stores. DATASIM's LRS API was built around the xAPI Communications protocol and should be interoperable with every ADL LRS Test Suite conformant LRS. This LRS interoperability feature is integrated with the prototype and can be simply demonstrated in the current state of the user interface.

4.5 User Interface

DATASIM has a fully functional user interface which can be used to interact with the entirety of DATASIM's current featureset. The user interface was designed with a focus on simplicity and consists of only one page, and a few dialogs. This was an intentional choice as at this stage of the product it is more important to demonstrate core functionality and integration than user experience. That being said, the design team went through multiple iterations and created a very user friendly interface that should aid independent demonstration and experimentation with the product.

4.6 Security

In its current state, DATASIM does not have its own persistence, and is therefore by definition incapable of storing sensitive data. Each user session, whether it's with the CLI or UI, only contains the inputs the user chooses during that session, and those inputs remain with the user not on any central storage. Because of this, the current prototype UI does not require credentials to access at this time.

Despite this, there does exist within the prototype a potential for abuse. The current state of the prototype, having accomplished the goal of large volume throughput of xAPI statements in a simulation, can be leveraged to potentially overload a small LRS installation.

To remedy this, two sets of credentials are required to do anything meaningful with DATASIM (entered in the UI). The first set of credentials is between the client and server installation, and is a rotated set of Basic Authentication credentials to prevent generation by a malicious third party. The second set of required credentials is to the LRS itself. This enforces that not only would a user need to have knowledge of LRS credentials in order to abuse it, but would also have to be authorized to use a particular DATASIM installation.

Neither of these security measures are intended as proper enterprise security for the application, and the project is open source and can be installed and used by anyone for any purpose on their own infrastructure. These security measures serve only as a means to prevent abuse of a given deployed and hosted prototype by a malicious third party. As the product matures beyond TRL 4, appropriate Authentication and Authorization measures should be implemented.

4.7 Fault Tolerance

DATASIM has the capability to recover from a defined scope of previously encountered errors and issues in runtime. The error messages for these scenarios are communicated to the user and in most cases with enough information for appropriate remediation. This area should be expanded to include much more verbose validation warnings, however, to accurately inform users of any issues which might be impacting their simulations.

In terms of failover and redundancy, the system is currently architected as a vertical scaling model, and a server-side failure does not have a software-based recovery model. Multiple instances of the server-side model can, however, be placed easily behind a load balancer or proxy with health checks (and endpoint for which exists in the application). Because all operations are stateless, failover should be seamless to an end-user in the event of server failure.

4.8 Documentation

At the current stage of the prototype, documentation is up to date on all intended usage scenarios of the application, for both the frontend and backend. Build and installation instructions as well as user instructions are available on Github at the following locations:

Backend: https://github.com/yetanalytics/datasim
Frontend: https://github.com/yetanalytics/datasim-ui

5. Conclusion

The current state of the DATASIM prototype meets or exceeds the expectations of a Readiness Level 4 Software Application in all areas of evaluation. In some areas, such as user interface, efficiency, and the packaging / deployment, the prototype may even be suitable for a TRL 5 stage prototype and testing with real-world stakeholders. This means that the next phase of maturity for the product can focus more narrowly on the most important remaining areas.

6. Next Steps

The DATASIM Prototype would benefit from enhancement across a number of categories in order to fully mature to the next phase of Technology Readiness.

6.1 Scale

The current DATASIM prototype can be scaled vertically, via access to hardware or additional computing resources. It leverages a streaming architecture so it does not have a defined upper limit on number of statements generated, it will just keep running until it is finished. It can also be duplicated horizontally to deliver higher total throughput to a single location for volume and load testing purposes.

What DATASIM currently cannot do is scale horizontally in a coordinated multi-node generation. What this means is that a user can not use the current state of the product to coordinate a singly-defined simulation across *n* independent-but-collaborating nodes to reliably produce a deterministic simulation. In future use cases this will be instrumental to coordinated load testing and also to the fast reproduction of high throughput simulated scenarios (e.g. high resolution instrumented simulators and real-world exercises). This improvement would represent adding a clustered architecture to the DATASIM backend enabling accurate and reliable orchestrated xAPI Statement generation.

6.2 Security, Authentication and Authorization

DATASIM currently does not have proper user accounts, permission schemes, or user-specific persistence. It will be important as the product matures and features are added to implement a standards-based Authentication, Authorization and Audit framework. Aside from curbing potential misuse of the system, and creating accountability for all system actions, this will enable features like user preferences, saving simulations by version, and storing LRS details to reduce the labor to an Analyst or Data Scientist using this tool in the real world.

6.3 User Interface

To match the addition of Security components to the application, there will be an overhaul of the user interface to account for those features. This includes account management, administrative functionality, LRS management, and Simulation management.

Additionally, as the product approaches the next phase of Technology Readiness, attention should be paid to usability issues that arise from this phase. The product should be iteratively revamped based on stakeholder feedback and a general user experience review.

6.4 Validation

One of the most important aspects of the DATASIM Project is its capability to validate inputs. Chiefly among these is its capacity as a workbench for xAPI Profile validation. This capability should be expanded and validation issues should be communicated clearly and with remediation suggestions. This will help DATASIM become an even more invaluable tool to the DoD and xAPI community as a whole as an application central to conformance testing across the Total Learning Architecture.