Real World Education Data Management Performance Testing

Secure, Rapid, High Volume Data Integration with the SIF Infrastructure Specification

Executive Summary

The demand for data in education continues to skyrocket (Data Quality Campaign, 2022). In the real-world, however, constantly churning, time sensitive, high volume record sharing has historically been the Achilles' heel of PK12 data integration efforts. When schools start to capture real time attendance events, the extra volume degrades and often completely overwhelms systems that rely on quick and timely data updates. Because of the extra load, updates to students and staff information could get lost in a sea of attendance events causing them to have to wait hours in line for processing.

To address these "digital ecosystem" challenges, the Access 4 Learning (A4L) Community released the **SIF Infrastructure Specification (global) 3.4.** This release is unique in the technical standards world in that it supports data management by both standardizing data and privacy on the "wire". Previous versions of this blueprint are already in use in marketplace products and in schools across Australia, New Zealand, North America and the United Kingdom but the 3.4 Infrastructure version has a number of enhancements that allow real time information to more efficiently flow through "data pipes".

The purpose of this white paper is to report on a third-party evaluation of these enhancements and present use-cases to demonstrate capabilities for handling high-volume data exchanges via "load testing". The 3.4 Infrastructure Specification was tested to previous versions to compare performance. A data generator and a data subscriber were created to generate and push sample Student Daily and Period Attendance.

Using just the SIF Infrastructure Specification 3.4 created ***20,000 times*** increase in throughput performance when compared to previous SIF 2.x Infrastructure Specification releases. The "hybrid" approach model, using existing SIF Infrastructure 2.x plus SIF Infrastructure 3.4 versions for those ecosystems who want to scale-up technologies, also showed improvements. By upgrading the receiving system, the result was ***286 times increase in throughput performance*** than the traditional SIF integration. These results quantify "on the ground" implementation stories where states have updated

their systems and discovered the increase in performance. These performance improvements enable greater data efficiency and opens the door to new use cases that require timely access to even more data.

The performance feature enhancements in the SIF Infrastructure Specification (global) 3.4 indicate a substantial increase in the "load" and "speed" in processing high volumes of PK12 data. In the future, it is expected that this standardized infrastructure will be used in conjunction with other data models, inside and outside of the education vertical, to allow for greater interoperability performance to support learning.

Overview

While "data" was once just a synonym for standardized test scores, education systems now require high volumes of data to be communicated in real or near-real time exchanged across multiple software systems. The sheer volume of data needed across these systems may cause data processing backlogs that would take over 24 hours to process. With an entire ecosystem stressed, any type of system hiccup or outage would take days or weeks to untangle! Rostering data are shared across the many software systems students and staff use throughout the school day; grades and attendance data are being shared across many applications; timely and accurate attendance data, combined with eligibility for programs such as free and reduced lunch, are being shared between state departments of education and human services for aid, etc. The need for data also is to be shared not just within pK-12 education but across pK-12 and higher education, connected to workforce data, and shared across human resources, health and human services, and many other related systems. This allows for analysis to understand trends and also for the ability to deliver critical services in real time.

Since 1997, the SIF Specifications have been the ONLY technical standards with the most comprehensive pK12 data model AND standardized infrastructure. In the past, the infrastructure of how the data exchange was orchestrated was tightly bound to the SIF Data Model, meaning that *only data within the spectrum of that data model could be exchanged.*

The SIF Specifications are now made up of two components:

- **Infrastructure:** The *'HOW'* - Defines the transport and messaging functionality over the "wire" where payloads are securely exchanged.
- **Data Model:** The *'WHAT'* - A set of XML and JSON schemas that define the payload format of "objects" as they are exchanged between SIF-compliant applications.

In order to understand the gains that have been achieved by SIF Specifications over the last 25 years, we first must understand the legacy infrastructure which earlier versions of the Specification were built. Previously, if a resync of the data was needed, there were bulk requests made to get a "point in time" refresh of all of the data. But what if the daily volume of changes increased from the relatively mild

flow of students and staff updates into a torrent of 10's of thousands or even 100's of thousands of events or more? Our one-at-a-time message-based system needed to be rebuilt so that it could scale, and we chose to utilize the efficiency of REST to build the new SIF APIs.

To address these challenges, the Access 4 Learning (A4L) Community released the **SIF Infrastructure Specification (global) 3.4.** Since 2014, the SIF Specifications have separated the SIF infrastructure from the SIF data model. The resulting separation provided immediate improved capacity and throughput and allows for the SIF infrastructure to be applied to other data models beyond the SIF Specifications. This also allows the SIF Infrastructure to orchestrate data exchange in a way that accommodates the increasing demand for interoperability across systems that may subscribe to different standards, such as higher education, workforce, human resources, and health and human services.

The purpose of this white paper is to report on a third-party evaluation of the enhancements in this Specification, present use-cases to demonstrate capabilities for handling high-volume data exchanges, and to showcase how SIF is designed for the increasing demands of data now and into the future including:

- Parallel Processing
- Endless Queue Enabling
- Data Bundling
- Privacy on the Wire
- Scalability
- Remaining Data Model Agnostic

Testing Methodology

Load testing was conducted by a third-party education integration leader, Cedar Labs. In order to isolate the infrastructure covered by previous versions of the SIF Specification 2.x and SIF Specification 3.4, a "Mock SIS" (data generator) and "Data Digester" (data subscriber) were created using the load testing framework Apache JMeter. The Mock SIS generates and pushes sample StudentDailyAttendance and StudentPeriodAttendance out to a SIF Enterprise Service Bus (ESB). The Data Digester receives the sample data and logs the results. The method for pushing and receiving the messages are driven by the use case - but will follow either the SIF Specification 2.x or SIF Infrastructure Specification 3.4. Each of the tests were run independently for one hour each. Cedar Labs' HostedZone broker was provisioned in the Amazon cloud on a server of t3.medium instances. The broker also has the capability to service "hybrid" implementations, where the SIS uses a legacy SIF 2x adaptor with the consuming applications taking advantage of the SIF Infrastructure Specification 3.4 REST API's.

The SIF Specification is capable of moving a wide range of data both within and outside the SIF data model - attendance was used in this scenario simply because it represents one of many areas where high-volume, real-time transfer is needed.

Load Test Results

Use Case: SIF Infrastructure Specification 2.x

As a baseline control group, the legacy **SIF 2 infrastructure** was used to move as many StudentPeriodAttendance and StudentDailyAttendance messages as possible from the Mock SIS to the Data Digester.

This system was not bottlenecked by hardware - in fact, CPU on the server didn't rise above 8% throughout the entirety of the Data Digester load test. The limiting factor in this baseline legacy test was the Data Digester's requirement to process single-record events one at a time. The Mock SIS's ability to produce events far outstripped the Data Digester's ability to consume. Also note that while the Mock SIS was able to generate a relatively large number of messages, it required a large number of (expensive) parallel processes to support this throughput.

Use Case: Hybrid (SIF 2 Mock SIS + SIF Infrastructure 3.4 Data Digester)

Our final use case is to plug a SIF 2 Mock SIS into a hybrid SIF environment. Because this test used the same SIF 2 Mock SIS as the baseline, there can only be 1 record per event. The motivation for this test is to demonstrate what sort of advantage a legacy implementation could gain almost immediately by simply plugging into a Hybrid SIF implementation. A number of states have already adopted this approach as both a way to scale up the implementation for immediate needs and also set themselves on the path to upgrade to full SIF support.

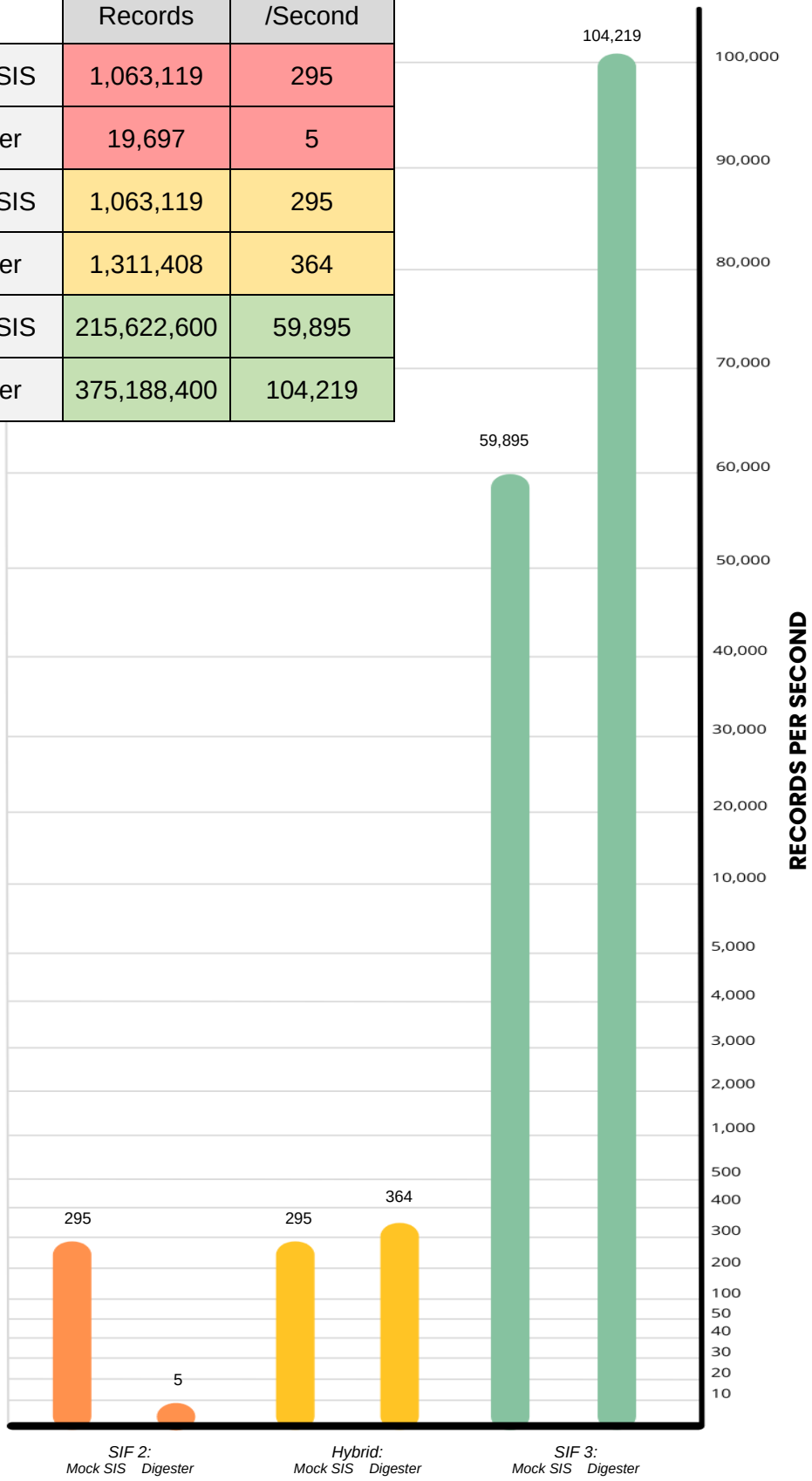
Both the Mock SIS and Data Digester were able to take full advantage of the server resources. If there is a need to scale up to accommodate more than 1 million records in an hour, more servers could be added. While in this test, scaled both the producing and subscribing sides of the equation to over 1 million records/hour.

Use Case: SIF Infrastructure Specification 3.4

Next, the SIF Infrastructure Specification 3.4 was used to move as many StudentPeriodAttendance and StudentDailyAttendance messages as possible from the Mock SIS to the Data Digester. We configure the records per message to 300 and set the number of concurrent queue processors to 100.

This test was able to take full advantage of the server resources. If there was a need to process more than 200 million records in an hour, it would be as simple as adding another server or two, and the system would be able to scale up in a matter of minutes. Using just the SIF Infrastructure 3.4 model created *20,000 times* increase in throughput performance when compared to previous SIF 2.x Infrastructure Specification releases. This approach was more efficient (around 240X) than the hybrid approach.

		Total Records	Records /Second
SIF 2 Infrastructure	Mock SIS	1,063,119	295
	Digester	19,697	5
Hybrid	Mock SIS	1,063,119	295
	Digester	1,311,408	364
SIF Infrastructure 3.4	Mock SIS	215,622,600	59,895
	Digester	375,188,400	104,219



Note that these load test results were achieved using Cedar Labs technology. Real-world results may vary, depending on the capabilities of the technology implemented in your use case.

Impact

Parallel Record Processing

In previous versions of the SIF Infrastructure Specification, and in many other data standards, events needed to be processed one after the other. If there were 10,000 records to be processed, the first record would be received…then processed…then acknowledged…and only after this was completed would the second record be received. ***This meant that no matter how much horsepower you had in the consuming server, you were limited to processing from 5 to 10 records per second***.

The upgraded SIF Infrastructure Specification provides mechanisms to have a configurable number of concurrent processors. In the field, we have seen 100 or more concurrent processors set up to fan out the processing during times of high load. ***This means that we've seen 100X or more increases in throughput from parallel processing alone***.

Refined Queues

In prior versions of the SIF Infrastructure Specification, events from a providing system, such as an SIS, to a consuming application would ***end up in one big queue.*** This would mean that if there were 80K new Attendance events in the queue ahead of a change to a student record that was needed for state reporting, that Student record wouldn't be processed until all 80K attendance records were processed.

The latest SIF Infrastructure Specification allows ***any number of queues to be created*** and provides a mechanism to route events to each queue based on record type. In the Attendance scenario, a queue can be created that only collects Attendance events, allowing an implementation to keep high priority record changes (such as Student, Staff, and Enrolment updates) isolated from the noise and processing quickly and efficiently.

Batch Events

A common limitation of data standards, including the earlier SIF 2.x standard, is that all ***event messages sent from a providing system would contain exactly one record***. In the case of 80,000 new Attendance records, that would be 80,000 separate event messages to process…one after the other.

In the latest SIF Infrastructure Specification, SIF events may be **batched** - which means that an arbitrary number of records can be sent in a **single message**. Those 80,000 Attendance records could now be bundled into messages of 500 each, so the number of messages to process would go from 80,000 down to 160, which means both greatly reduced cost and faster processing for both the providing and consuming systems.

Conclusions

The performance improvements provided for in the SIF Infrastructure Specification 3.4 have a wide range of implications for those education agencies with existing implementations as well as those with new data sharing initiatives. Certain use cases that previously were simply not possible are no longer limited by infrastructure.

- The new SIF Infrastructure Specification 3.4 takes better advantage of the hardware it runs on and can scale up or down to meet the needs of the moment.
- Efficiency gains mean users can do more work with less resources, saving money on implementations while providing a sizeable performance improvement.
- Finally with the hybrid scenario, any existing implementation can immediately start to take advantage of this powerful new architecture without having to embark on a huge multi-year implementation or rip up existing tooling and processes that are vital to operations.

SIF Specification ROI

Built for the Future

The volume, breadth and expectations around data integration needs to continue to grow exponentially. In education, an upgrade can meet current demands such as daily and period attendance and grade pass back, but also to prepare for future demands such as data to track competency-based learning or the extensive data produced by online learning systems. Additionally, the flexibility of the SIF Infrastructure Specification 3.4 provides the ability to leverage this high-performance infrastructure for other data models (not just those defined in the SIF Data Model Specification), allowing organizations to manage data from broad sources such as Medical / Human Services, Human Resources, Higher Education, and K-12 Education all through a common REST API.

More Efficient/Less Expensive to Operate

Those utilizing a SIF implementation benefit from the move to utilizing the SIF Infrastructure 3.4 Specification. The **data provider** (such as a SIS) can take advantage of sending out bundled event messages. This is the single largest performance improvement that was made with the new infrastructure. One of the most “expensive” points of posting an event is the web service handshake - creating the secure connection. By packing up hundreds of objects in a single payload, this cuts down on the chattiness of a system by 2 orders of magnitude, allowing your processes to run leaner (and cheaper).

The **data subscriber** (our Data Digester) is also able to take advantage of receiving bundled payloads. However, the consumer can also take advantage of two other newly introduced innovations:

parallel message processing and refined queues. Parallel message processing allows an arbitrary number of threads to consume messages at the same time, this alone increased the throughput by 100X. Also, by allowing for refined (and special purpose) queues, SIF implementations are able to isolate user-interactive queues (such as a data administrator's high priority queries) from the chaos of attendance or grade events. This level of control makes running a large-scale implementation much easier to understand what is going on under the hood and makes troubleshooting much more reasonable.

Well Worn, Incremental, and Safe Migration Path

This work has clearly established that the SIF Infrastructure Specification unlocks exciting new use cases, while also providing significant performance improvements and cost savings at scale. However, without a known, low-risk path to adoption, the update of mission critical systems can often present too large a burden for software system developers. Fortunately, a number of large-scale implementations have successfully pioneered the "hybrid" approach, providing an incremental adoption path that realizes the performance gains and cost savings, without taking on risk. The Access 4 Learning Community (A4L) and its community-developed SIF Specifications have a history of open source, community-based approach to implementation that goes back decades and provides mechanisms for these organizations to share their experiences and know-how with interested parties. Migrating from a SIF Specification 2.x to "hybrid" has been accomplished in a matter of months, allowing for predictable project planning, without surprise disruptions to existing processes.

SUMMARY

This third-party evaluation details how using the freely available SIF blueprint can increase YOUR data throughput to meet even the most demanding needs, leverage your existing technical investments, and support your privacy data stewardship. All of this allows end-users and marketplace providers to get the right data to the right person at the right time by more easily creating "connected and secure effective learning ecosystems".

For more information on the SIF Specifications, please visit: https://data.A4L.org

---

Since their first implementation in 2013 (which is still going strong), Cedar Labs has been focused on how to simplify the real-time connection of education data across systems. They may be bigger now but haven't forgotten their roots as a customer-focused, solutions-oriented company that puts solving your problems first.

cedarlabs.com