Sharing Analytics: Developing a standard to employ high impact analytics without stressing networks

Darryl W. Roberts Ph.D.[*], Victoria Spina[†], John Griffith[†], Joseph L. Ronzio D.H.Sc.[‡], and Cj Rieser Ph.D.[†]

Affiliations: *General Dynamics Information Technology, †The MITRE Corporation, ‡Department of Veterans Affairs

The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs, Veterans Health Administration, the Office of Health Informatics or the United States federal government.

COVID19 and the need for rapid, low bandwidth analytics

Since the recent outbreak of COVID-19, clinical researchers, drug manufacturers, economists, social scientists, and other experts have sought access to data for analytics. Much of the data resides in disparate networks and within organizational boundaries. It takes considerable time and effort to obtain permission to acquire this data, to gain authorization from ethics committees or institutional review boards (IRBs) to use it, and to transfer it across overstressed networks. Most researchers would prefer to spend the precious time consumed by these rather tedious tasks on analysis. A more efficient approach would be to have a directory that says what data is where and what it contains, and then develop analytics to leverage the data in a way that preserves privacy and minimizes latency.

In support of this approach, volunteers formed an IEEE Working Group to develop P2795 – *A Standard for Shared Analytics Across Secure and Unsecured Networks*. As proposed, this standard will outline and support easy transmission of valid, reliable, and precise analytic objects with integrated tests of object integrity, veracity, and privacy, both pre- and post-distribution. Additionally, the standard would make data analysis bandwidth-liberating, instead of bandwidth-depleting, leading to overall faster communication and processing. During the COVID-19 pandemic, the extraordinary need for data has coincided with an unprecedented overload of the Internet and data resources. These circumstances make the development of a standard for sharing analytics all the more valuable. While COVID-19 presents a prime example of the need to share analytics without having to move data across congested networks, it is certainly not the only use case. There are many other scenarios where shared analytics could be used locally, such as in deciding the best treatment for a patient by having access to abundant data. The standard will also preserve data privacy, allowing transmission across both *secured* and *unsecured networks*. The standard would allow analysis at the source of the data and return results without revealing personally identifiable information (PII) or patient health information (PHI).

Evolution from simple to complex

As proposed, P2795 has initial and aspirational capabilities. Growth from the former to the latter will conform to the successful implementation of Gall's Law, which states, "A complex system that works is invariably found to have evolved from a simple system that worked."[1] Therefore, the standard will not attempt to capture all potential future capabilities, but rather serve as a flexible, fundamental, simple system. This simple initial standard will facilitate and inspire collaboration and research, which will, in turn, stimulate and improve personal investment. The P2795 Working Group is beginning to identify and distinguish use cases for both the baseline and future generations of the standard. Ultimately, the Working Group will develop a roadmap that prioritizes an evolution in capabilities.

The initial standard will incorporate several important basic features: a data directory, a querying ability, data traffic monitoring, and scalable encryption technology. The standard will establish a four-part handshake between an analytic requestor and a data owner.

> **Part 1:** The analytic requestor sends a description of the kind of analysis it wants to perform to a directory service.
>
> **Part 2:** A data owner, via the directory service, provides metadata, i.e., a description of the relevant kind of data they have and any conditions on their data's use.
>
> **Part 3:** The requestor constructs and transmits specific analytic requests based on that metadata.

[1] https://en.wikipedia.org/wiki/John_Gall_(author)#Gall.27s_law

Part 4: The data owner verifies that the request conforms to the conditions and returns the results of the analysis.

All parts of the protocol use FIPS 140-2 compliant[2] encryption to provide scalable, modular, and secure transactions. Later generations of the standard will support additional features, such as:

- Data configuration and query management to track changes in data between queries
- Federated machine learning over heterogeneous, distributed data
- Near real time analytics to support time-sensitive reporting
- Unique identifiers to support detection of common data elements, such as individuals or items, across data systems and organizations

Potential Use Cases

The P2795 Working Group believes that the initial and subsequent capabilities will support several use cases. Here are some examples.

- A research team must identify results of patient COVID-19 testing conducted across multiple systems within a given period without revealing PII, PHI, or the systems that conducted the tests. In this use case, researchers could program the analytic by entering metadata to show its intended purpose, and then present the analytic and its respective metadata to an IRB for speedy review and approval. Once approved, they could release the analytic and quickly obtain the results to determine the specificity and sensitivity of multiple types of testing programs.
- Operations analysts need a real-time, source-agnostic program to support the efficiency of the personal protective equipment (PPE) supply chain. In this use case, the analytic could track and report item counts and locations to facilitate immediate distribution. This is of vital importance to support immediate recall of faulty items developed during the rush orders from the initial COVID-19 wave.
- Future use cases might include the ability to flag PPE or other essential items (e.g., disinfectant wipes) in point-of-sale (POS) systems to support distribution restrictions or limitations to healthcare providers during an outbreak.

[2] https://csrc.nist.gov/publications/detail/fips/140/2/final

- Different patients need modified treatment options specific to their conditions. Access to a large quantity of patient health data would support stratification along patient demographics, health conditions, and treatment outcomes to determine personalized treatment paths for each patient.

Call to action – Make P2795 *your* standard

Low bandwidth, high utility, privacy preserving analytics based on a reliable standard would promote research and operations for multiple use cases in health care, as well as in other industries. Development of this standard will inspire innovation and subsequent improvement as it moves from a flexible, simple baseline standard to a complex, mature, and widely accepted standard. The best way to move from baseline to maturity is to engage potential users, thought leaders, and technical experts to share their non-proprietary insights, best practices, knowledge, and unique perspectives. Involvement in the development of a standard is one of the best ways to maximize its value to *you*. The more you share about your needs, the more likely it will be that the standard will help you achieve those needs. People working together in a strong community with a shared goal and a common purpose create a powerful force. In short, *your support and insights can make P2795 your standard.*

The authors would like would like to acknowledge the ongoing contributions of the IEEE P2795 Working Group members. For information on joining the P2795 Working Group, see https://sagroups.ieee.org/2795.