DARPA Brandeis Case Research and Development

PULSAR-VLDS

Stealth PULSAR Privacy-preserving Technology for Data Sharing and Analytics for Virginia Longitudinal Data System

(In Alphabetical Order)

Dr. George Alter, University of Michigan

Dr. Chongwon Cho, Stealth Software Technologies

Ms. Quinn Grier, Stealth Software Technologies

Dr. Steve Lu (Principal Investigator), Stealth Software Technologies

Mr. Tod Massa, State Council of Higher Education for Virginia

Dr. Rafail Ostrovsky, University of California, Los Angeles

Acknowledgements. This work was supported by DARPA and NIWC Pacific under contract N66001-15-C-4065. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes not withstanding any copyright notation thereon. The views, opinions, and/or findings expressed are those of the author(s) and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

Table of Contents

Overview......2
Project Partners......2
Use Case......3
PULSAR-VLDS Privacy-preserving Approach......4
System Tests with SCHEV Data...... 5
Performance Benchmarks......7
Conclusion and Future Work......9

Overview

The PULSAR-VLDS demonstrates secure, privacy-protecting computation of statistics from sensitive data held in multiple government organizations. PULSAR-VLDS is a partnership between Stealth Software Technologies, a leader in privacy-enhancing technologies (PETs), and the Virginia Longitudinal Data System (VLDS), a pioneering collaboration among state agencies to produce better data for public policy.

PULSAR-VLDS is an innovative application of Secure Multiparty Computation (MPC) to overcome the barriers to sharing data among agencies while protecting private information. Policy-makers frequently require statistics involving data from multiple sources. Although these reports would not compromise individual privacy, sharing data across agencies with varying regulations and legal limitations is expensive, time consuming, and often impossible. MPC uses cryptography to join and aggregate data from separate databases without releasing any private information. Each database owner controls the keys to their own data, and no unencrypted data is ever released. MPC relies on cryptography rather than secret code, and data-privacy guarantees can be demonstrated with open source software.

The PULSAR-VLDS platform uses new architectures and tools to overcome limitations of previous MPC technologies. PULSAR-VLDS provides:

1. *Strong data-privacy* guarantees where unencrypted individual data is never shared or revealed to other agencies during computation process
2. *Accurate* statistical results, because the outcome does not use inaccuracy (e.g., noise) to mask identities
3. *Immediate* answers, because computations do not depend upon deidentifying and collecting data in a single place
4. *Efficiency and scalability*, due to Stealth Software's fast and innovative software

PULSAR-VLDS also includes additional disclosure protections developed by VLDS for aggregate data. PULSAR-VLDS is designed to automate privacy protections in a modular manner that allows manual intervention where necessary.

VLDS anticipates that the new platform will increase the number of state agencies willing to collaborate and will enable new policy-relevant statistics and analyses.

Project Partners

VLDS is a pioneering collaboration for Virginia's future, giving the Commonwealth an unprecedented and cost-effective mechanism for extracting, shaping and analyzing partner agency data in an environment that ensures the highest levels of privacy. VLDS was initiated by the 2009 Statewide Longitudinal Data Systems Grant Program of the United States Department of Education and consists of several component technologies that support authorized research addressing today's top policy and state program questions. VLDS is the result of a shared effort by several Virginia government agencies. VLDS merges, shapes, and shares data across the participating agencies with deidentification hashing process for the privacy of individual datasets in order to enable critical government research. VLDS is designed for onboarding other state agencies in the future.

Stealth Software Technologies, Inc. was established in 2005. The Stealth team consists of a world-class team of experts in cryptography (including multiple university professors), developers, and consultants, including IACR Fellow Dr. Rafail Ostrovsky (who is also an IEEE Fellow). Stealth team members have pioneered several of the most commonly used cryptographic tools and techniques that have helped shape the area of Secure Multiparty Computation and more; to name just a few, concepts such as Oblivious RAM, Private Information Retrieval and its symmetric variant SPIR, efficient Oblivious Transfer Extension, Searchable Public-Key Encryption and Searchable Symmetric Encryption, Batch Codes, Oblivious PRFs, efficient zero-knowledge proofs and secure computation via “MPC in the Head”, and Garbled RAM have all originated from academic and technical works of Stealth team members and have been the topic of active research and implementation efforts. Our team and consultants were/are involved in the following highly relevant projects: IARPA HECTOR; DARPA PROCEED, Cyber Fast Track, Brandeis, SafeWare, RACE, SIEVE and several SBIR projects DARPA Secure Messaging (Phase I & II), MATH MARATHON (Direct-to-Phase II), and LiLaC SALSA (Phase I & II).

In this pilot project PULSAR-VLDS, we demonstrate the possibility to expand the VLDS's current capabilities by using the modern cutting-edge cryptographic system, known as MPC. PULSAR-VLDS demonstrates a promise of an efficient solution system enabling data extraction, shaping, and analysis with stronger data privacy guarantee in practice, without needing participating agencies to ever share even de-identified datasets with VLDS (or other agencies).

Use Case

PULSAR-VLDS can be a game changer for the following use cases:

The first use case is simply to expand participation by Virginia agencies that are not currently participating in VLDS. For example, the State Council of Higher Education for Virginia (SCHEV) has been producing reports on the wage and student debt outcomes of college graduates since 2012. Virginia was one of the first states in the nation to do this, and is the only state to provide the details on student debt and wage outcomes out to 20 years post-completion. The data is limited to participating agencies, namely SCHEV and the Virginia Employment Commission (VEC). The income data held by the VEC exclude 20% or more of the workforce in Virginia, particularly federal employees, including the military. Leveraging tools developed in Stealth's PULSAR project (contracted under the DARPA Brandeis program), VLDS provides a method of participation that is acceptable to other agencies, which could lead to providing more comprehensive earnings reports. This would be a tremendous improvement in what we know about Virginia college graduates working in Virginia.

A second use case is a suite of reports on Agency Intersections. These reports, still under development, link individuals across each agency and provide aggregate counts of the numbers of records in common. For example, we can see how many individuals in SCHEV are also found in VEC and the Virginia Department of Health Professions (DHP), or any other possible combination of agencies. Knowledge about which agencies are serving the same people will provide policy makers with guidance in targeting and streamlining the provision of services. For example, there are nearly 30 million records held across the partner agencies, representing over 19 million unique individuals, six million of which appear in multiple systems. There are agencies that desire an approach to data-sharing that is more restrictive, that VLDS cannot offer without MPC.

PULSAR-VLDS Privacy-preserving Approach

Stealth leverages the core PULSAR technologies known as a Secure Multi-Party Computation (MPC) to realize the platform, architecture, and tools for data-private VLDS computations. Essentially, the PULSAR-VLDS prototype enables secure and data-private computations among three parties such as (1) a data analyst (accessing via the VLDS webpage) who desires to obtain certain statistics that can be computed from two distinct datasets A and B, and (2) two data providing agencies holding private datasets A and B respectively where each agency has their dataset in their own database behind its firewall (See Figure 1).

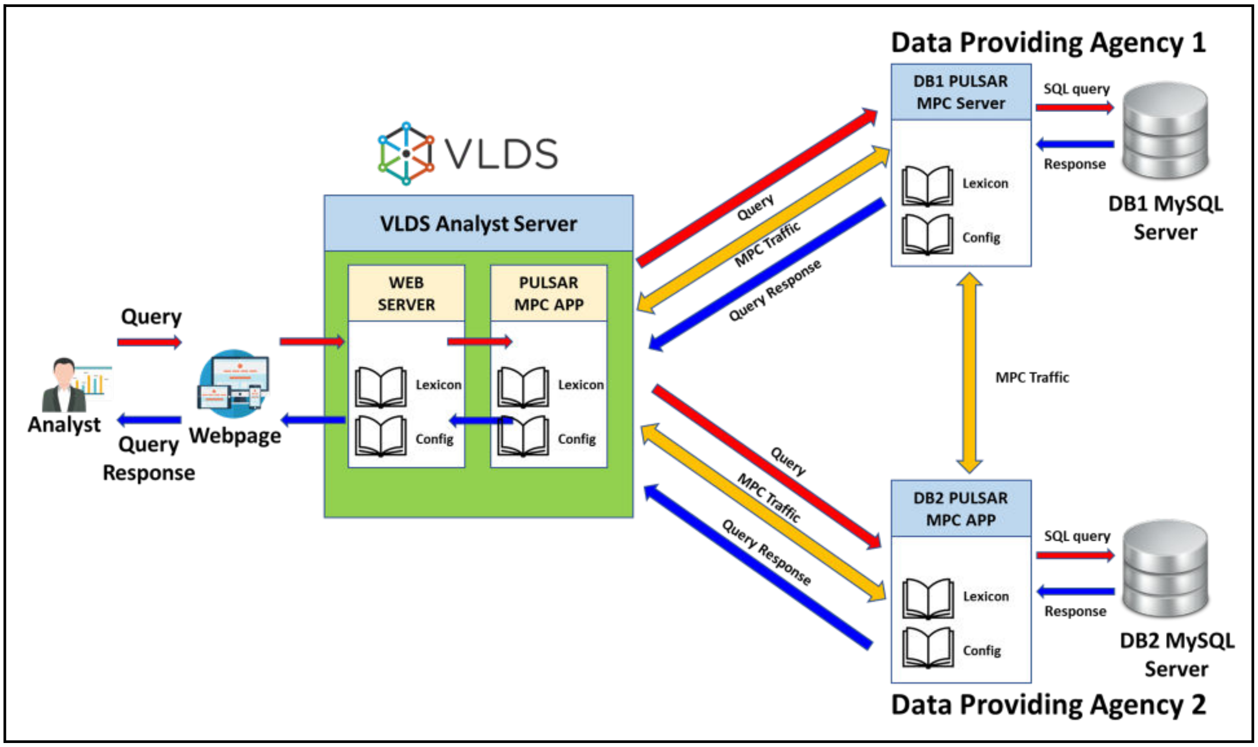


Figure 1. PULSAR-VLDS Architecture and Workflow Overview

PULSAR-VLDS's Innovation The technical crux of the PULSAR-VLDS's MPC engine is that the analyst's desired statistics is computed on encrypted database streams without requiring data providing agencies to share their private data unencrypted. Our design ensures that private data never leaves its data-providing agency's premises unencrypted. PULSAR-VLDS's architecture does not require a trusted data collecting agency that may be potentially a single-point of failure for data-privacy. Furthermore, PULSAR-VLDS is designed to be highly scalable with huge datasets where the size of datasets is potentially over tens of gigabytes. In contrast to the other standard MPC-based solutions which are not scalable with huge input datasets, our novel MPC design ensures that computations over huge encrypted data are performed in a streaming manner so that they do not require a large accumulation of encrypted data at a single point or system with the size of memory usage tunable by the system operators.

PULSAR-VLDS's MPC engine does not rely on technologies adding noise to private datasets or output of statistics for data-privacy. Such noise-based technologies often require mathematically challenging validation of their data-privacy guarantees which depend on types and characteristics of underlying datasets. In addition, the outputs of such noise-based systems are noisy by nature, thus potentially resulting in inaccuracy of analytics. These technical problems become even worse when analytics are computed across two datasets. The PULSAR-VLDS does not suffer from such noise problems. Our system protects individual data using encrypted traffic without adding noise to individual data or analytics output, resulting in accurate analytics.

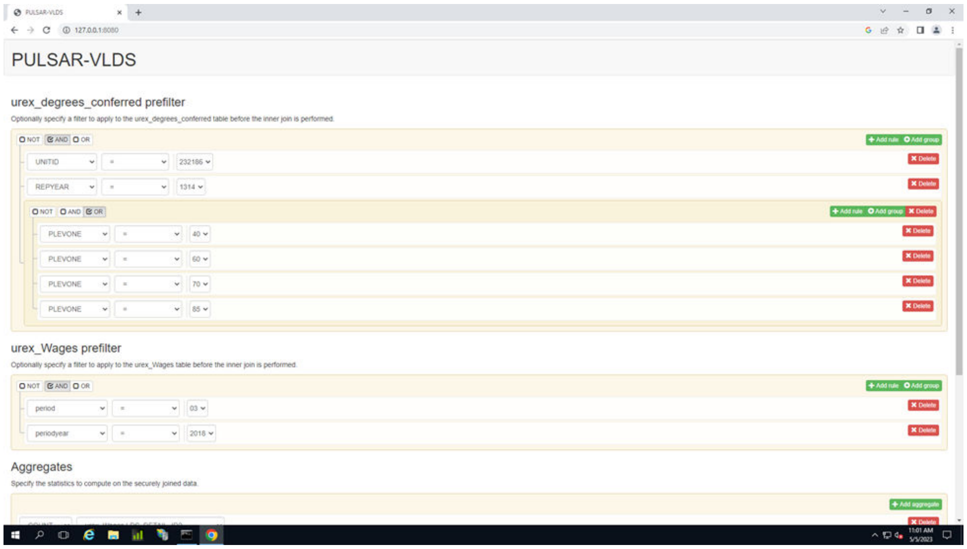
In summary, the PULSAR-VLDS's MPC prototype demonstrates the following innovations:

1. data-private joining between two data sets over a common identifier where all input datasets and the outcome of joining remain fully encrypted from the start to the end,
2. data-private statistics computations over joined data sets, which only reveals the statistical result to the analyst without revealing any information from data contributors to any other party,
3. all computations are efficient and only add a constant overhead to the existing VLDS approach. The MPC engine processes the computation over potentially huge datasets by streaming, so that the amount of memories for each participant remains constant (i.e. independent of the size of data sets) according to a system parameter tunable by system administrator,
4. and importantly, it is easy to deploy and integrate PULSAR-VLDS with the existing database servers
5. using easy-to-learn lexicon and configuration files.

System Tests with SCHEV Data

We tested our PULSAR-VLDS system with SCHEV's test cases. The standard test case is the matching of millions of rows of college graduate data against quarterly wage data. Both datasets possess a social security number that cannot be exposed to the analyst. Using PULSAR-VLDS's secure MPC engine, an analyst may build queries of varying complexity. The following example pulls the graduates from a specific institution in 2013-14 and matches to their 3rd quarter wage in 2018.

An analyst in the PULSAR_VLDS system formulates a query by selecting various combinations of selections for student data, wage data, and target aggregates defined according to the config files. Figure 2 and Figure 3 below show a screenshot of an example analyst query. In Figure 2, the initial filter, “urex_degrees_conferred prefilter” selects the institution and graduation year. The grouped select identifies the specific degree levels for reporting. The next selection, “urex_wages prefilter” identifies the reporting year of the wages and the specific quarter.



PULSAR-VLDS

urex_degrees_conferred prefilter

Optionally specify a filter to apply to the urex_degrees_conferred table before the inner join is performed.

○ NOT ☑ AND ○ OR

UNITID = 232186

REPYEAR = 1314

○ NOT ○ AND ☑ OR

PLEVONE = 40

PLEVONE = 60

PLEVONE = 70

PLEVONE = 85

urex_Wages prefilter

Optionally specify a filter to apply to the urex_Wages table before the inner join is performed.

○ NOT ☑ AND ○ OR

period = 03

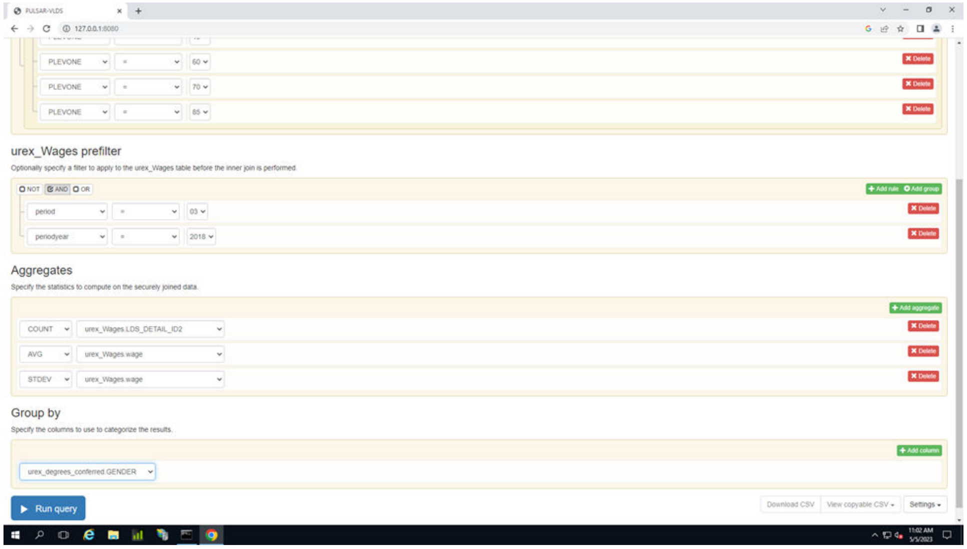
periodyear = 2018

Aggregates

Specify the statistics to compute on the securely joined data.

Figure 2. The screenshot of example query selections for student data (urex_degrees_conferred prefilter) and wage data (urex_Wages prefilter)

Figure 3 below shows the example selection of target aggregate. In the “Aggregates” section, we specify the count of the individuals matched, the average quarterly wage, and standard deviation for output responses.



PLEVONE = 60

PLEVONE = 70

PLEVONE = 85

urex_Wages prefilter

Optionally specify a filter to apply to the urex_Wages table before the inner join is performed.

○ NOT ☑ AND ○ OR

period = 03

periodyear = 2018

Aggregates

Specify the statistics to compute on the securely joined data.

COUNT urex_Wages.LDS_DETAIL_ID2

AVG urex_Wages.wage

STDEV urex_Wages.wage

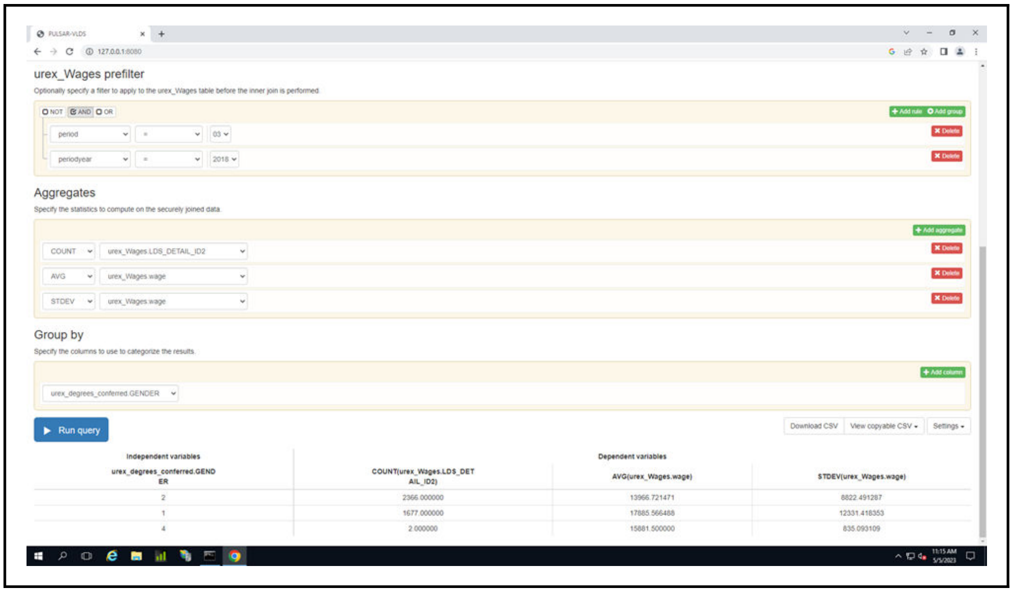
Group by

Specify the columns to use to categorize the results.

urex_degrees_conferred.GENDER

Figure 3. The screenshot of aggregate type selection

Figure 4 shows the selection screen for “Group by” which is chosen to be "gender" in the screenshot. Hence, the query response is grouped by the gender types (specified according to the input config files) and displayed at the bottom of Figure 4 upon the completion of PULSAR-VLDS computation.



urex_Wages prefilter

Optionally specify a filter to apply to the urex_Wages table before the inner join is performed.

period = 03

periodyear = 2018

Aggregates

Specify the statistics to compute on the securely joined data.

COUNT urex_Wages.LDS_DETAIL_ID2

AVG urex_Wages.wage

STDEV urex_Wages.wage

Group by

Specify the columns to use to categorize the results.

urex_degrees_conferred.GENDER

Independent variables	Dependent variables		
urex_degrees_conferred.GENDER	COUNT(urex_Wages.LDS_DETAIL_ID2)	AVG(urex_Wages.wage)	STDEV(urex_Wages.wage)
2	2366.000000	13966.721471	8822.491287
1	1677.000000	17885.566488	12331.418353
4	2.000000	15881.500000	835.093109

Figure 4. The screenshot of Group by selection and the query result

Performance Benchmarks

We tested our PULSAR-VLDS system with two SCHEV's datasets: SchoolDB and WageDB. The SchoolDB is a dataset that contains up to 583,000 rows with columns like ID (randomized individual identifier), degree, gender, school ID (representing universities), and the WageDB contains approximately 1 million rows with columns like ID, wage, year, and quarter. The following SQL query is a representative of the type of query used in benchmarking.

```
SELECT AVG(WageDB.wage), STDEV(WageDB.wage)
FROM SchoolDB INNER JOIN WageDB ON SchoolDB.id = WageDB.id
WHERE SchoolDB.degree = 'Bachelor'
  AND SchoolDB.gender <> '4'
  AND WageDB.year = '2018'
  AND WageDB.quarter = '03'
GROUP BY SchoolDB.UID, SchoolDB.GENDER
```

Table 1. The PULSAR-VLDS Test Query expressed in a SQL query

The above query asks for an analytics report with an average and standard deviation on wages in 2018 who graduated with a bachelor's degree by all genders and institutions. The performance benchmarks are taken by running the query in Table 1 against datasets of 40k, 80k, 120k, 160k, 200k, 240k, and finally 1.58 million records in total. The following figure (Figure 5) provides the benchmarks of end-to-end query-response latency from the time of the above client query submission from a client system to the time of analytics returned to the client system.

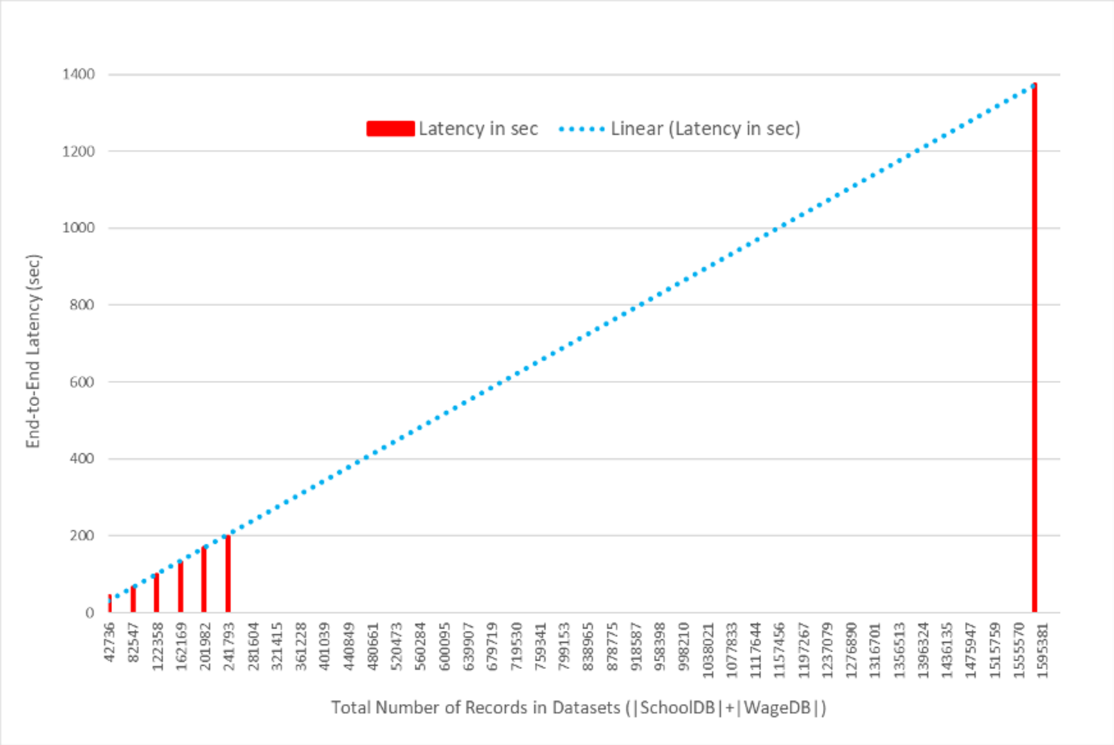


Figure 5. The End-to-End Query-Response Latency of PULSAR-VLDS

The end-to-end latency is linearly proportional to the total size of input datasets. The first experiment with datasets of 40k records in total takes 44 seconds to produce the analytics report. The final experiment with datasets of 1.58 million records in total takes 22 minutes to produce an analytics report.

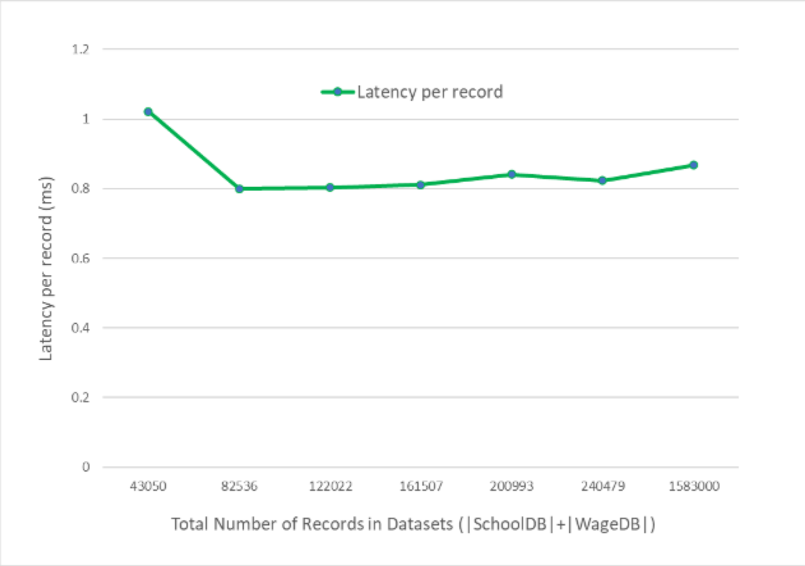


Figure 6. End-to-End Latency per Record

Figure 6 shows the end-to-end latency per record, further providing the evidence that the latency is linear in the total number of records in the input datasets. The average per-record latency is 0.852 milliseconds with a standard deviation of 0.078 milliseconds.

Conclusion and Future Work

The PULSAR-VLDS project demonstrates that secure MPC based privacy-preserving technology can be efficient and increase the accessibility and usability of sensitive data beyond the current data-sharing limitations. The PULSAR-VLDS platform developed new architectures and tools to overcome limitations of previous MPC technologies. PULSAR-VLDS provides:

1. *Strong data-privacy* guarantees where unencrypted individual data is never shared or revealed to other agencies during computation process
2. *Accurate* statistical results, because the outcome does not use inaccuracy (e.g., noise) to mask identities
3. *Immediate* answers, because computations do not depend upon deidentifying and collecting data in a single place
4. *Efficiency and scalability*, due to Stealth's fast and innovative protocol and software

In the future, we envision the strengthening of data-privacy guarantee with developing a data privacy compliance verification mechanism that allows data-provider organizations to determine whether output responses adhere to organizational data-privacy regulations or laws. This will enable us to redact the encrypted outputs before they are released in the (a priori unknown) event of potential privacy impact. Specifically, when an output response occurs to provide an aggregate related to a small number of individuals, it may be policy to redact the cohort due to insufficient number of individuals contributing to it. Therefore, an additional layer could be added to the PULSAR-VLDS prototype for secure post-processing after completion of the query to remove any results that do not meet a specified level of disclosure.

Another future work is to expand further the current prototype to enable data sharing across more than two data providers with more diverse join types. The current prototype supports only the inner join across two datasets. Extensions can include multiparty intersections, or as a first step, extending our prototype to support multiple pairwise intersections.

With the promising outcomes of the project, we are looking forward to future opportunities to extend the PULSAR-VLDS project to other use case workflows requiring the privacy-preserving guarantees.