CWI

Improving Data Minimization through Decentralized Data Architectures

DBDBD 2024 - Amsterdam, the Netherlands

Ilaria Battiston

November 22, 2024

Research Statement

CWI

- Provide a framework for *data privacy*

Motivation

- Increase *data minimization*
  - Maintain **sensitive data control** while preserving *analytical value*

→ provide a **privacy-preserving** cloud alternative

→ use cases: fitness trackers, healthcare applications, advertisements, ...

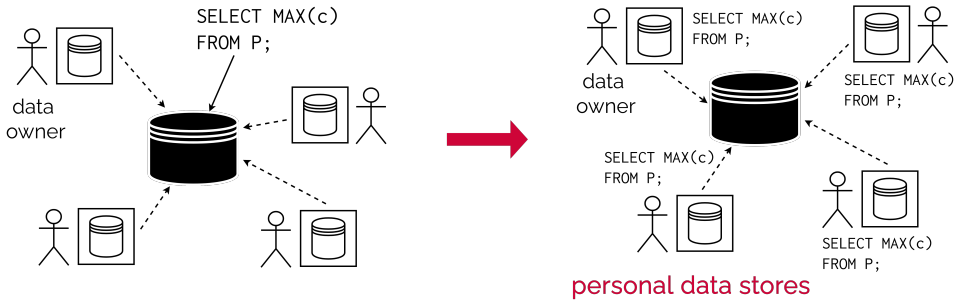
Related Work

- The SOLID Project[1]
- Federated and Distributed Query Processing [1, 2]
- Differential Privacy [3], PINQ [4]
- Encrypted Query Processing [5]

[1] https://solidproject.org/

Our Decentralized Infrastructure

- Local computation of partial analytical aggregations



→ better **data minimization** and **resource consumption**

Research Questions

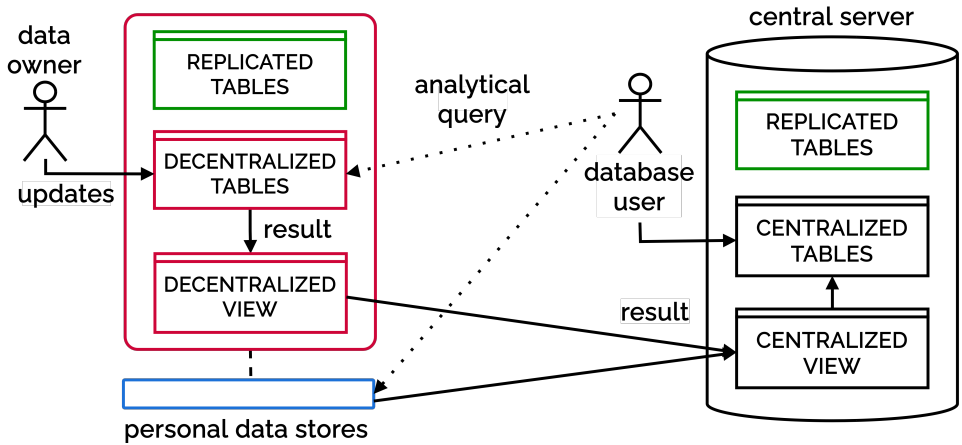
CWI

1) How can we build and specify a **privacy-preserving** decentralized data architecture?

2) How can we enforce **privacy constraints** in an *efficient* and *secure* way?

3) How can we build **trust** in our infrastructure?

Framework


data owner
updates
REPLICATED TABLES
DECENTRALIZED TABLES
result
DECENTRALIZED VIEW
personal data stores
analytical query
database user
central server
REPLICATED TABLES
CENTRALIZED TABLES
result
CENTRALIZED VIEW

Declarative Language

- SQL privacy constraints to be applied at the *parsing* stage or **periodically** within centralized views

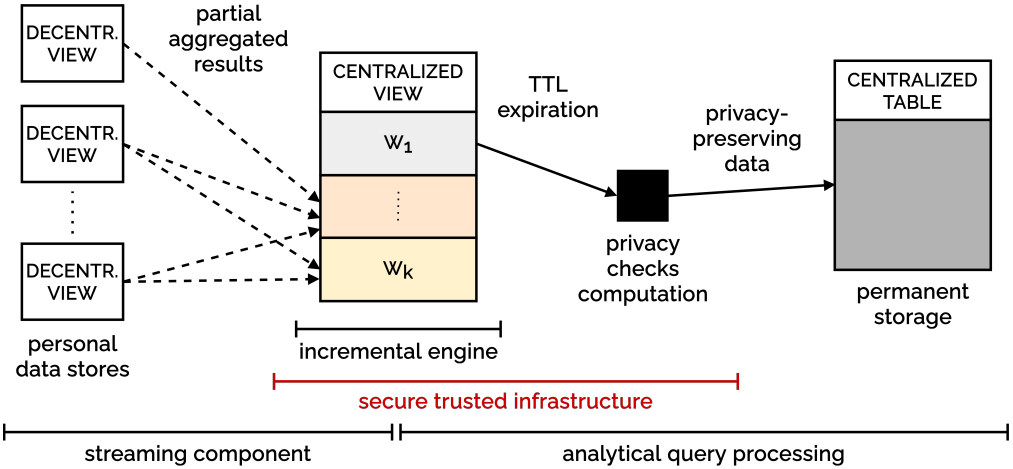
```
CREATE DECENTRALIZED TABLE Workouts (
workout_name VARCHAR(100) NOT NULL,
user_id INT RANDOMIZED,
start_time TIMESTAMP,
duration_minutes INT,
location_id INT MINIMUM AGGREGATION 5,
average_heart_rate INT SENSITIVE);
```

Streaming Semantics

CWI

- Asynchronous, unordered operation mode
  - ➞ window-based streaming aggregates
  - ➞ requirement of abstractions (*Dataflow model* [6])
- Declarative keywords for *expiration time* of tuples and windows, *minimum completeness*
- Need of a **privacy-preserving incremental processing infrastructure**

Privacy-Preserving IVM

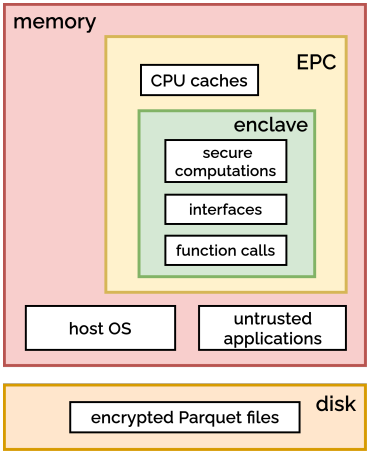


Secure Query Processing

CWI

- Data should be **hidden** while performing privacy-preserving incremental computations
- How to protect *data in use*?
  1) Secure enclaves
  2) Multi-party computation

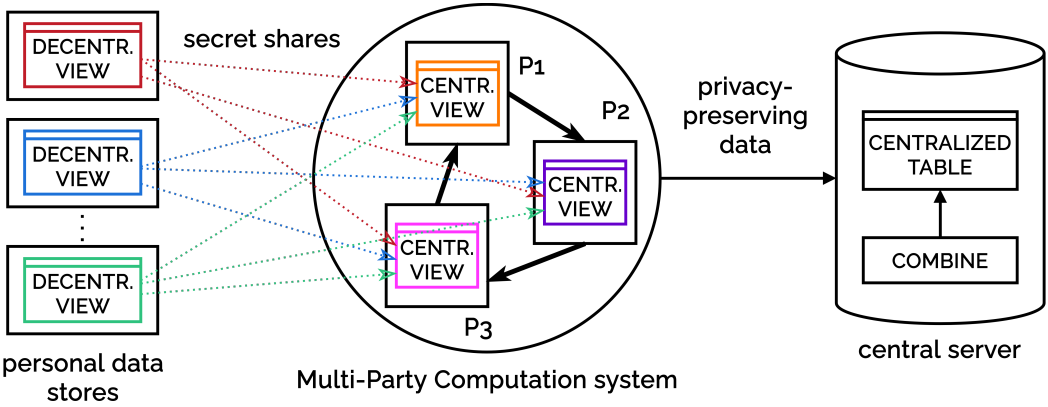
Intel SGX 2



- An **enclave** is a secure container for applications and computations
- Data resides *encrypted on disk*, while the DBMS resides in *encrypted memory*
- 1.2x overhead performing analytical workloads
- More secure hardware solutions to be explored

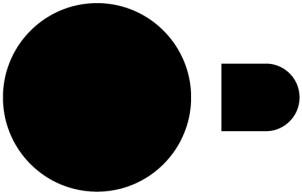
Multi-Party Computation

- We propose **TVA** [7], *MPC system* for secure analytics on time series



Implementation

- DuckDB[1] extension modules:
  - OpenIVM
  - Differential Privacy
  - SQL-to-SQL compiler
- SGX porting



[1] https://github.com/duckdb/duckdb

Future Research

- Query planning, scheduling & optimization
- Semantics and formalisms for data privacy
- Efficient IVM techniques in the MPC framework
- Decentralized architecture testing

References

[1] Hannes Mühleisen. "Architecture-independent distributed query processing". PhD thesis. Free University of Berlin, 2012.

[2] Mark Raasveldt and Hannes Mühleisen. "MonetDBLite: An Embedded Analytical Database". In: *CoRR* abs/1805.08520 (2018). arXiv: 1805.08520. URL: http://arxiv.org/abs/1805.08520.

[3] Cynthia Dwork. "Differential Privacy: A Survey of Results". In: *Theory and Applications of Models of Computation, 5th International Conference, TAMC 2008, Xi'an, China, April 25-29, 2008. Proceedings*. 2008.

[4] Frank McSherry. "Privacy Integrated Queries". In: *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data (SIGMOD)*. Association for Computing Machinery, Inc., June 2009.

[5] Raluca A. Popa et al. "CryptDB: protecting confidentiality with encrypted query processing". In: *Proceedings of the 23rd ACM Symposium on Operating Systems Principles 2011, SOSP 2011, Cascais, Portugal, October 23-26, 2011*. Ed. by Ted Wobber and Peter Druschel. ACM, 2011, pp. 85–100. DOI: 10.1145/2043556.2043566. URL: https://doi.org/10.1145/2043556.2043566.

[6] Tyler Akidau et. al. "The Dataflow Model: A Practical Approach to Balancing Correctness, Latency, and Cost in Massive-Scale, Unbounded, Out-of-Order Data Processing". In: *Proc. VLDB Endow.* 8.12 (2015), pp. 1792–1803.

[7] Muhammad Faisal et al. *TVA: A multi-party computation system for secure and expressive time series analytics*. Cryptology ePrint Archive, Paper 2023/1120. https://eprint.iacr.org/2023/1120. 2023. URL: https://eprint.iacr.org/2023/1120.