DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Federated Data Analytics for Genomics Data

Student: **Mirco CAZZARO** – mat. **2076745**

Supervisor: Prof. **Gianmaria SILVELLO**

Master Degree Course in **Computer Engineering** – **Web Information and Data Engineering** (WIDE)

A. A. **2023/2024**

The Challenge

Biomedical data is becoming increasingly complex. The rapid evolution of data storage systems has created a landscape where:

- Diverse data models exist (relational, hierarchical, graph-based);
- Integration of these models is crucial making data meaningful.

Understanding genetic diseases and advancing personalized treatments relies on integrating and analyzing this diverse data effectively.

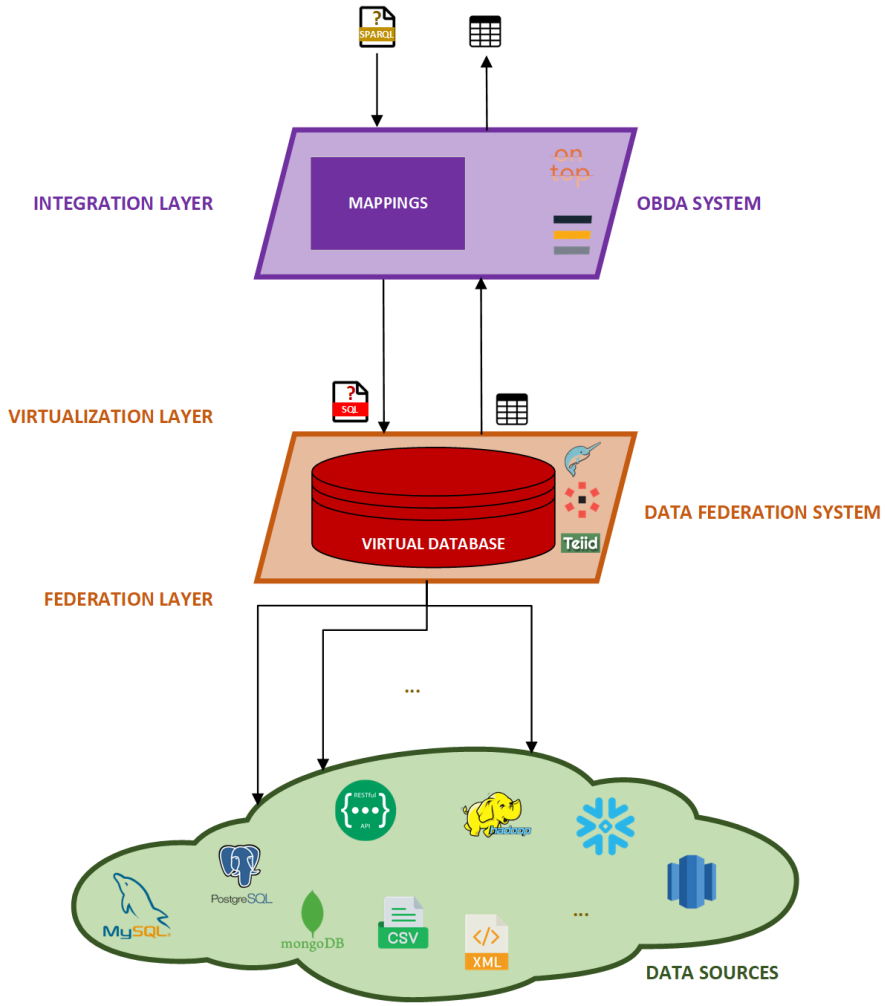
Proposed Solution

Federated Data Analytics System

A system designed to integrate and analyze clinical and genomics data seamlessly.

- Objective: Enable researchers to perform complex queries across multiple datasets without extensive preprocessing.
- Key Components: OBDA, Data Virtualization, Ontology.

This approach reduces complexities in biomedical data management and accelerates research.



EU Project HEREDITARY

Objective:

Transforming our understanding of brain diseases through integrated multimodal data analysis.

Focus Areas:

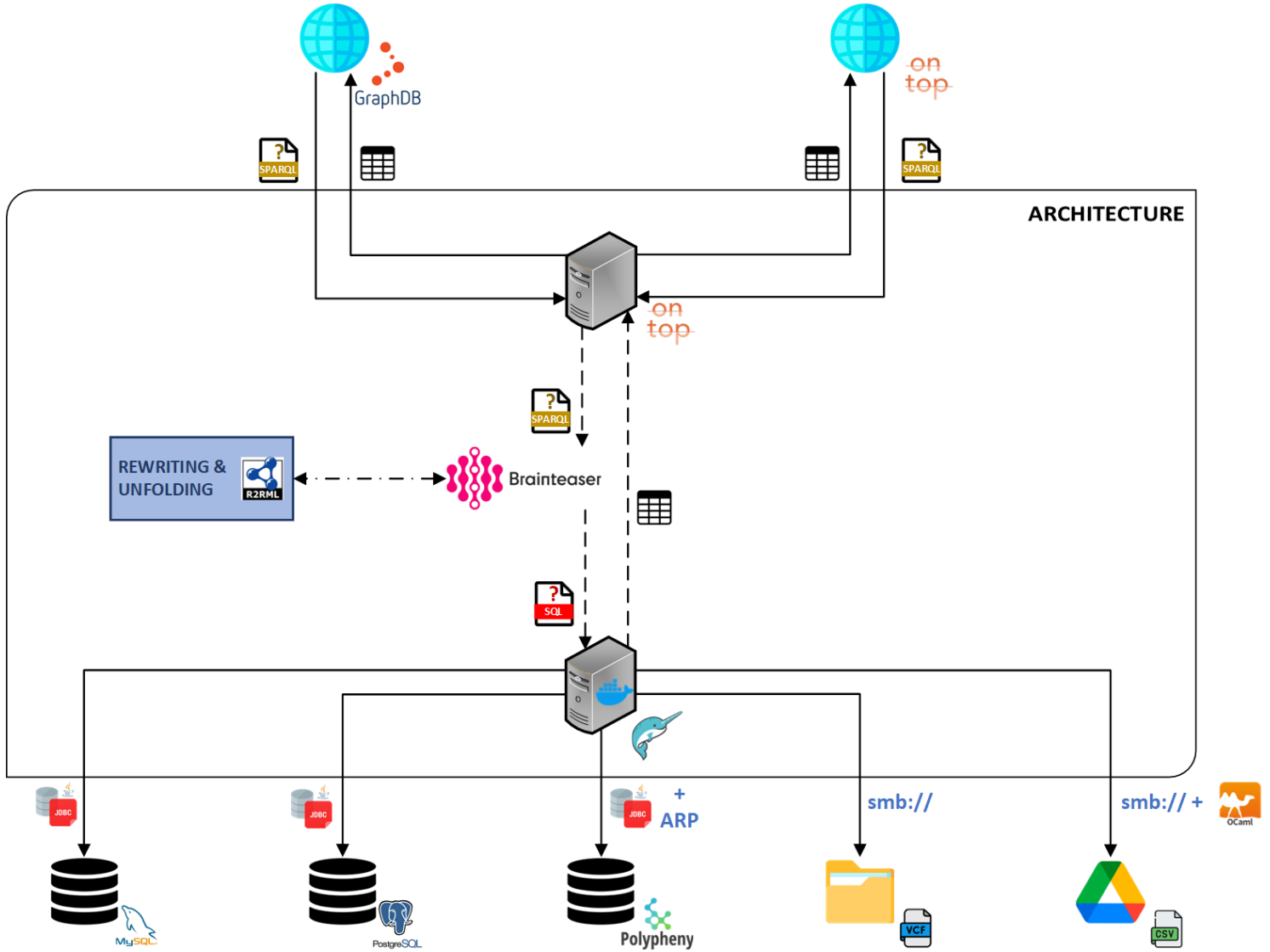
1. Integrating genomic and clinical data from various European stakeholders;
2. Addressing the challenges of data heterogeneity;
3. Ensuring privacy compliance and data security.

(one of the) Results:

A unified DBMS that supports advanced querying capabilities across different data sources and that allows to gather insightful analytics.

System Architecture

- Ontop: Semantic Data Integration Layer.
- Dremio: Data Federation and Virtualization Layer.
- BRAINTEASER Ontology: Provides the vocabulary structure for querying.



System Architecture – Data Sources

- MySQL: Stores structured clinical data.
- PostgreSQL: Manages additional clinical datasets.
- Polypheny: Supports various data models, enhancing flexibility.
- NAS Shared Folders: Hosts genomics data in VCF files.
- Google Drive: Integrates a cloud-based storage use case, hosting CSV files.

Each data source is connected via Dremio's federation layer, enabling a unified view across diverse data repositories.

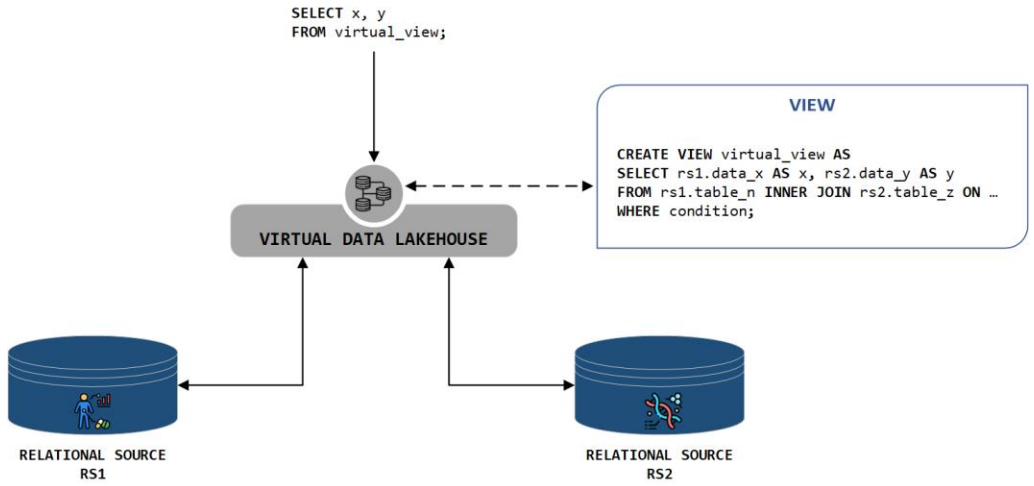
Federation and Virtualization

Virtualization Layer:

Creates virtual views without materializing data, optimizing performance and enabling complex queries across the federated system.

Federation Layer:

Integrates data across various sources, ensuring seamless access and querying.



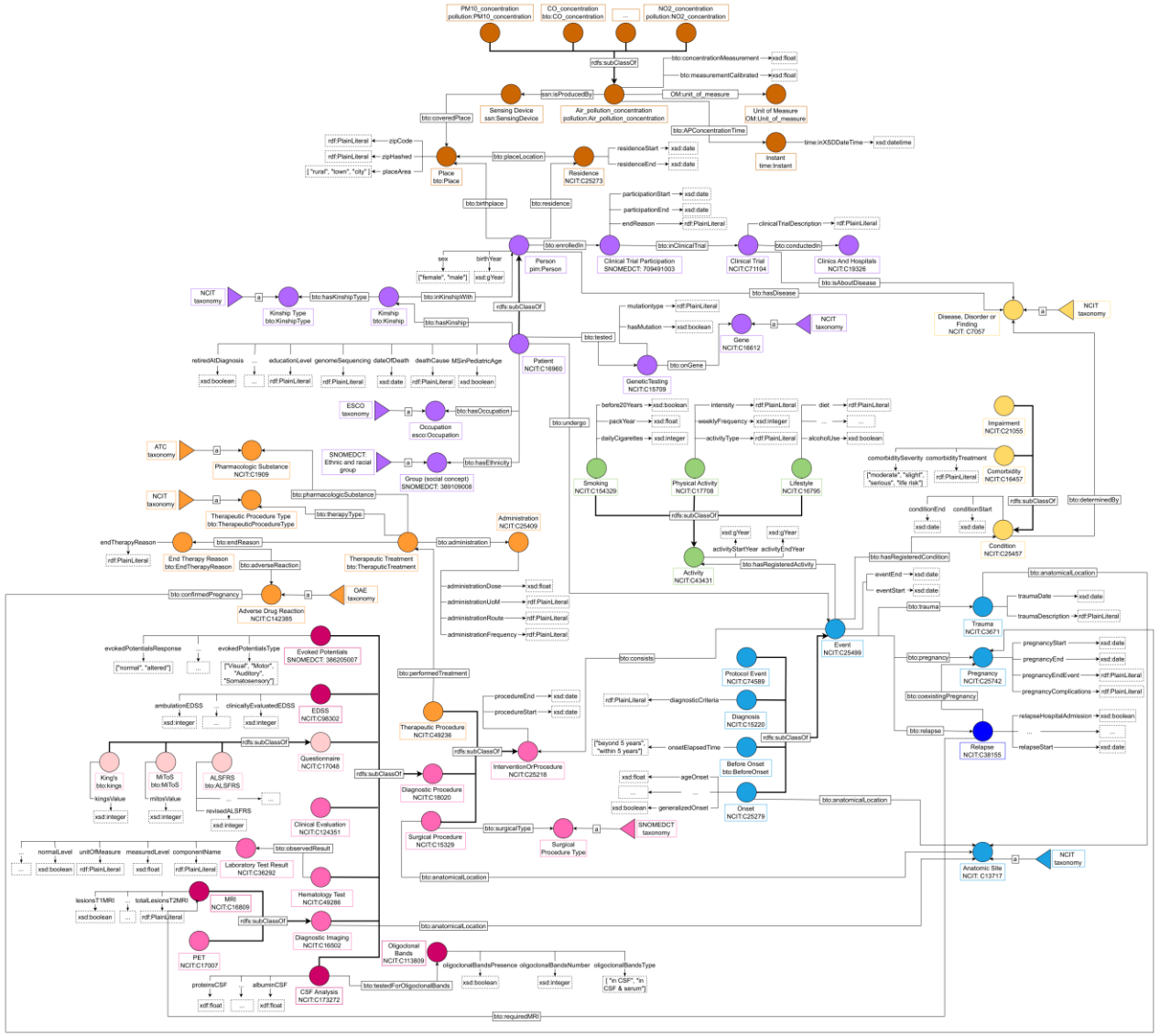
Dremio serves as the core engine for both layers, providing a robust platform for handling large-scale data integration tasks.

Ontology Integration

The BRAINTEASER Ontology is central to the system

Features:

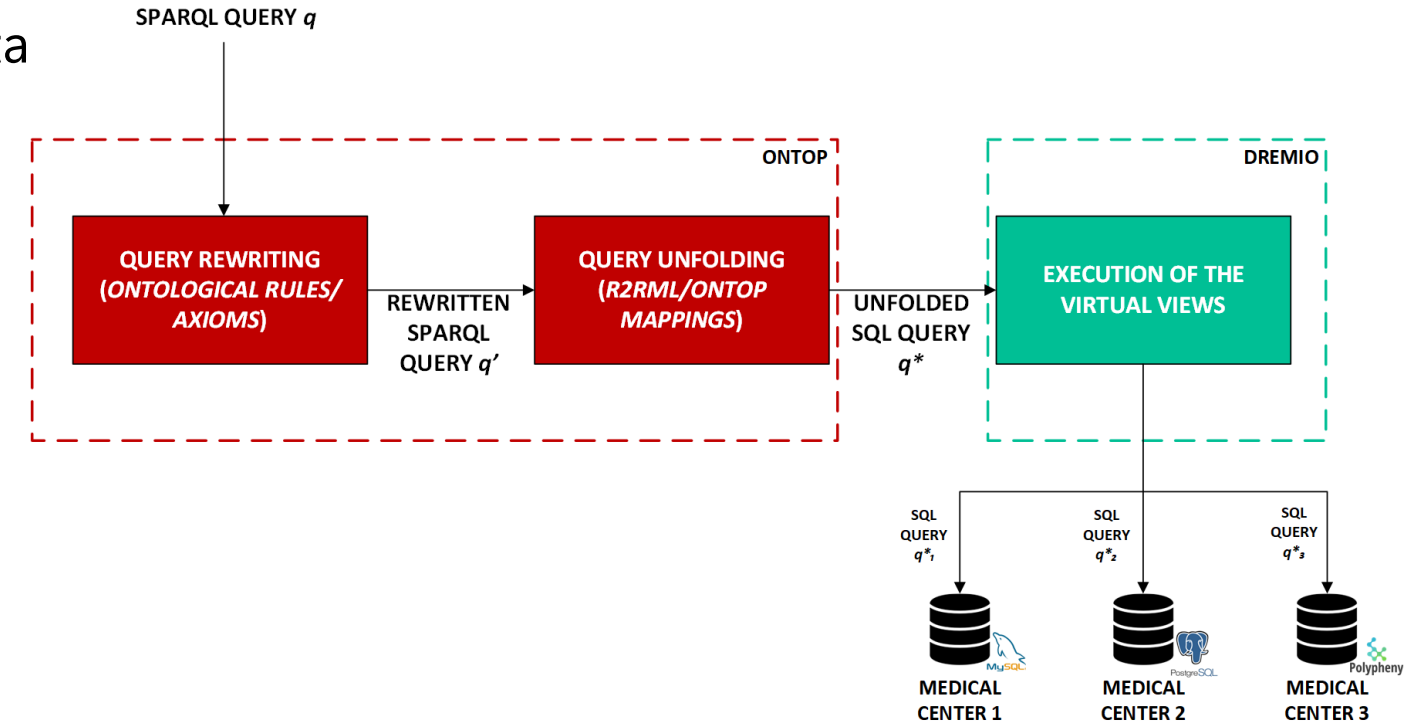
- Models relationships between clinical and genomics data;
- Enhances query capabilities through semantic enrichment;
- Ensures interoperability and scalability.



Use Case: ALSFRS Data Querying

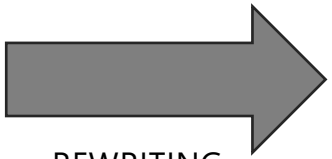
Use Case: Querying ALSFRS (Amyotrophic Lateral Sclerosis Functional Rating Scale) data across multiple sources.

- SPARQL Query: Retrieves comprehensive patient data;
- Process:
  - Query Rewriting;
  - Unfolding into SQL;
  - Execution in Dremio.



```
PREFIX bto: <https://w3id.org/brainteaser/ontology/schema/>

SELECT ?patient ?date ?tot ?bulbar ?motor ?respiratory ?q1 ?q2 ?q3 ?q4
 ?q5 ?q6 ?q7 ?q8 ?q9 ?q10 ?q11 ?q12 WHERE {
    ?p bto:undergo ?e.
    ?e bto:consists ?alsfrs.
    ?alsfrs a bto:ALSFRS;
           bto:procedureStart ?date;
           bto:revisedALSFRS ?tot;
           bto:bulbarSubscore ?bulbar;
           bto:motorSubscore ?motor;
           bto:respiratorySubscore ?respiratory;
           bto:alsfrs1 ?q1;
           bto:alsfrs2 ?q2;
           bto:alsfrs3 ?q3;
           bto:alsfrs4 ?q4;
           bto:alsfrs5 ?q5;
           bto:alsfrs6 ?q6;
           bto:alsfrs7 ?q7;
           bto:alsfrs8 ?q8;
           bto:alsfrs9 ?q9;
           bto:alsfrs10 ?q10;
           bto:alsfrs11 ?q11;
           bto:alsfrs12 ?q12.
    BIND(SUBSTR( (STR(?p)), 48) AS ?patient)
}
ORDER BY ?patient ?date
```

Original SPARQL Query



REWRITING
+
UNFOLDING

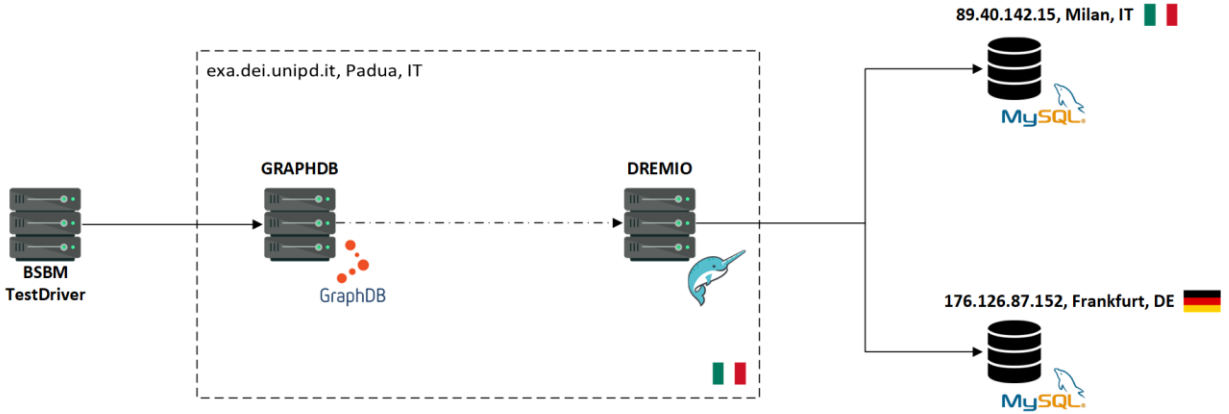
```
CONSTRUCT [patient, date, tot, bulbar, motor, respiratory, q1, q2, q3
    , q4, q5, q6, q7, q8, q9, q10, q11, q12]
   [date/RDF(CHARACTER VARYINGToVARCHAR(date2m51), xsd:datetime),
    q1/RDF(INTEGERToVARCHAR(q11m25), xsd:integer),
    motor/RDF(INTEGERToVARCHAR(motor1m55), xsd:integer),
    ...]
NATIVE
SELECT
    v23."bulbar1m17" AS "bulbar",
    v23."date2m51" AS "date",
    v23."motor1m55" AS "motor",
    v23."q101m44" AS "q10",
    v23."q111m42" AS "q11",
    v23."q11m25" AS "q1",
    v23."q121m41" AS "q12",
    v23."q21m24" AS "q2",
    v23."q31m23" AS "q3",

    --- ... more fields

    v23."patient26m9" AS "patient"
FROM (
    SELECT DISTINCT
        v7."bulbar" AS "bulbar1m17",
        v5."date2m51" AS "date2m51",
        v8."motor" AS "motor1m55",
        v1."patient" AS "patient26m9",
        v19."q10" AS "q101m44",
        v20."q11" AS "q111m42",
        v10."q1" AS "q11m25",

        --- ... more fields

        v6."tot" AS "tot1m45"
    FROM "clinical_data"."ALSFRS" v1

    --- ... recursive joins

    WHERE v1."patient" IS NOT NULL AND v1."date" IS NOT NULL
) v23

ORDER BY v23."date2m51" NULLS FIRST
```

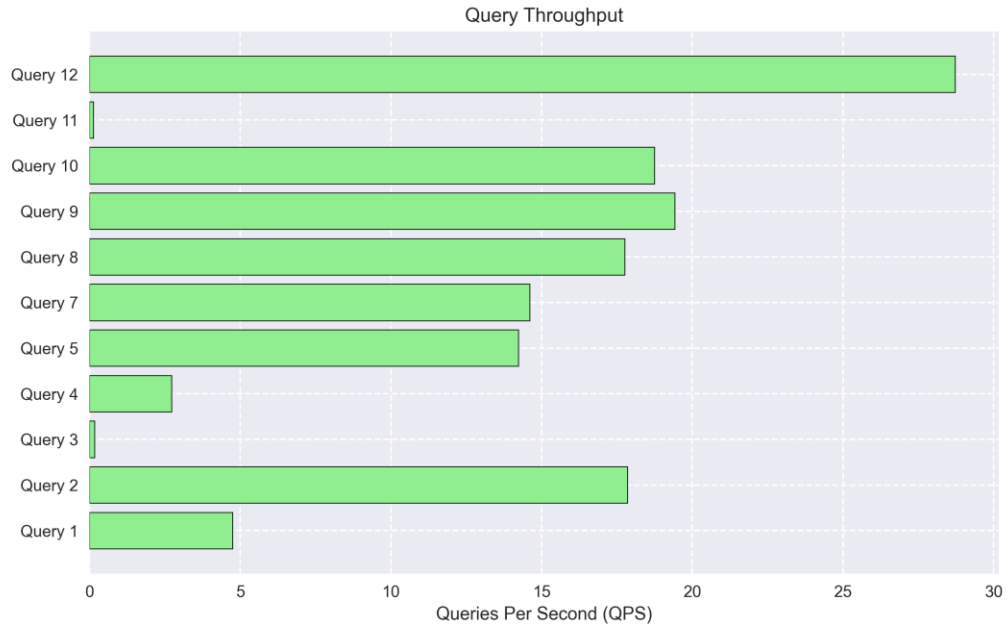
Unfolded SQL Query

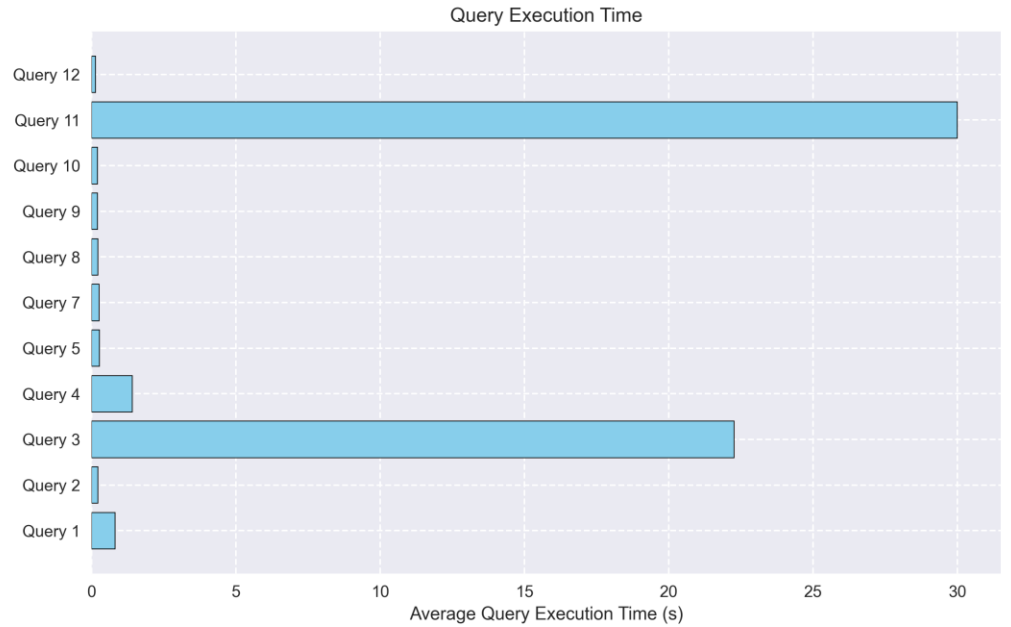
Benchmark 1: BSBM

BSBM Benchmark

Query performances evaluation over a synthetic dataset across diverse sources.

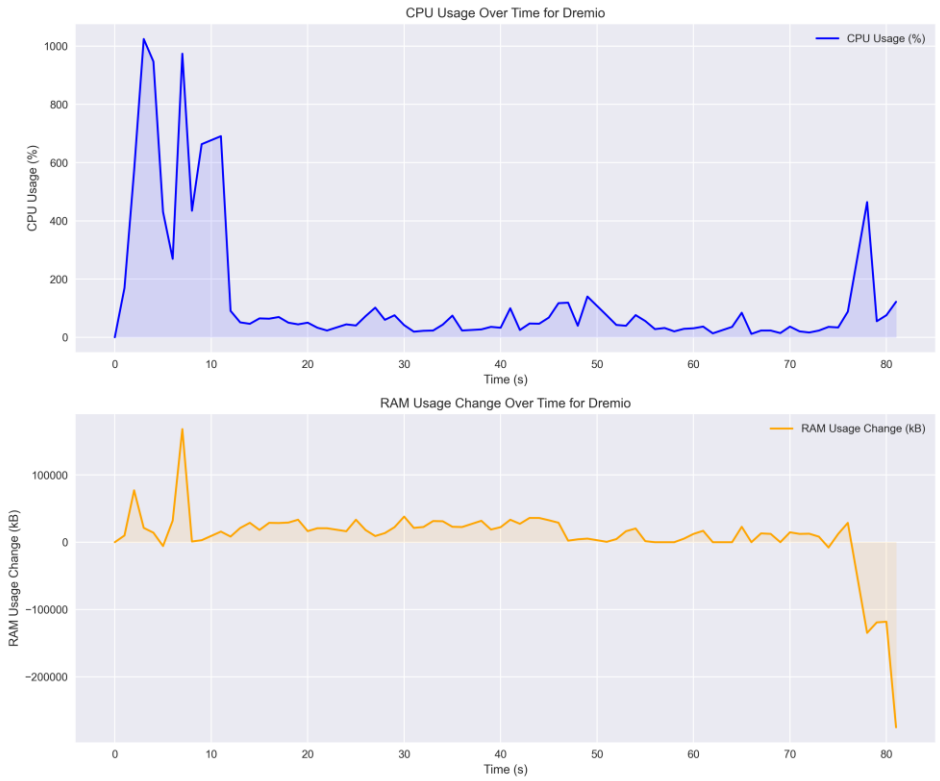




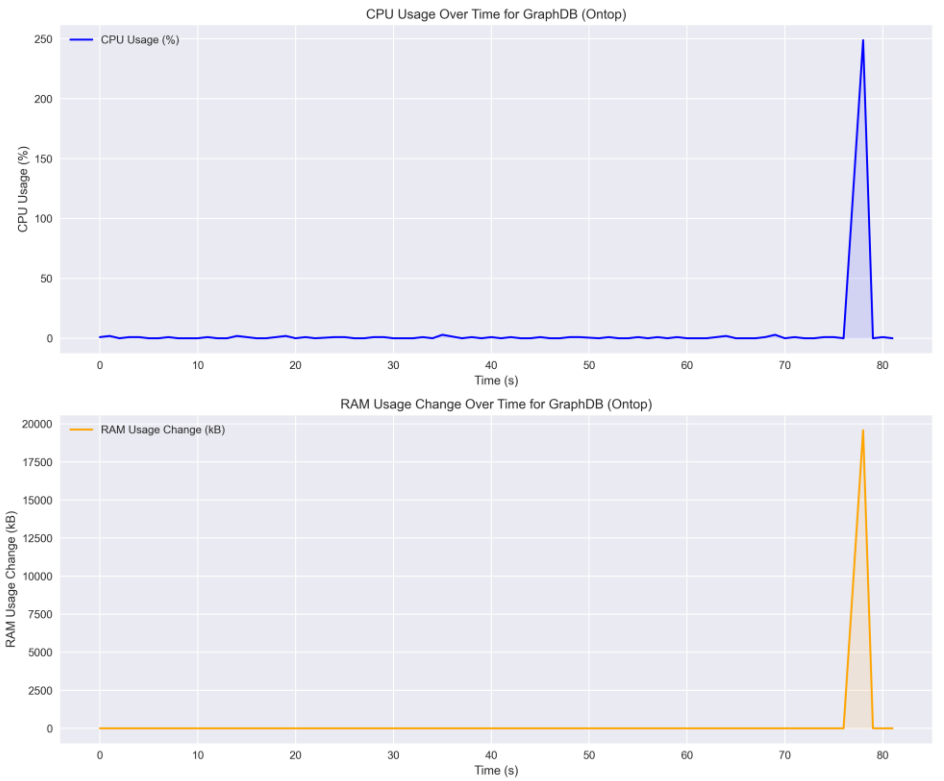


Benchmark 2: SEASHELL

SEASHELL Benchmark Real-world clinical data highlights system efficiency in handling complex queries.



Dremio



GraphDB

Conclusions and Future Works

The federated data analytics system developed in this thesis offers a powerful tool for integrating and analyzing heterogeneous biomedical data.

Future Directions:

1. Optimization: Improve architecture performances, considering SEASHELL monitoring data as an entry point;
2. Privacy: Strengthen data security and compliance with regulations, such as GDPR;
3. Usability: Simplify both the architecture usage and deployment, so to reach a larger audience.

This system lays the groundwork for future advancements in biomedical data research, contributing to better healthcare outcomes.

Thank You!

Do you have any question?

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Master Degree Course in **Computer Engineering – Web Information and Data Engineering** (WIDE)

A. A. **2023/2024**