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 for making data meaningful.

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

Background

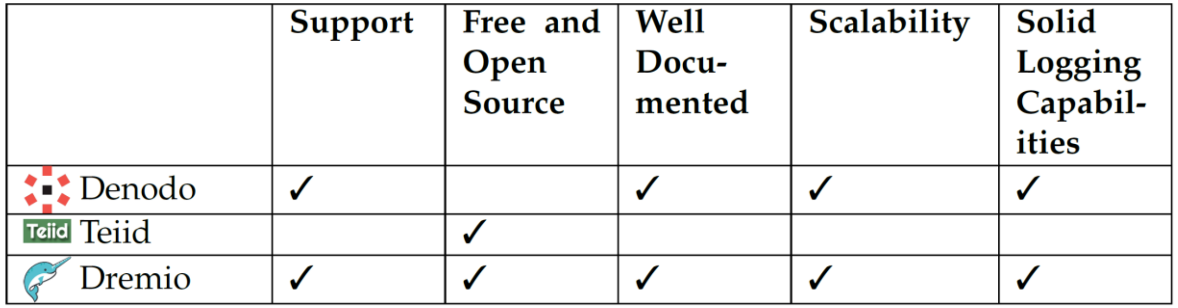
Methodologies:

1. Data **Federation**
2. Data **Virtualization**
3. Semantic Data **Integration** (OBDA)

4. Knowledge Graphs, Ontologies & RDF

Technologies:

1. SPARQL query language;
2. Data **Virtualization** solutions: Denodo, Dremio, Teiid
3. **OBDA** Solutions: Mastro, Ontop



	Support	Free and Open Source	Well Documented	Scalability	Solid Logging Capabilities
Denodo	✓		✓	✓	✓
Teiid		✓			
Dremio	✓	✓	✓	✓	✓

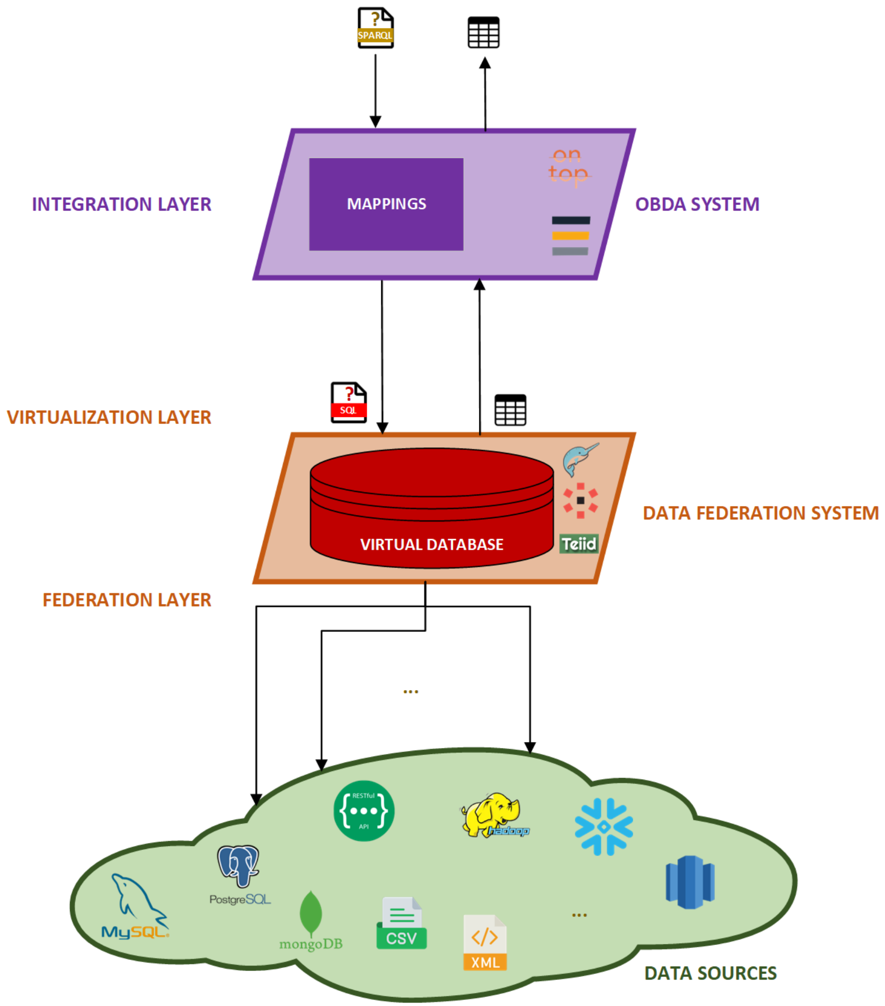
Proposed Solution

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.

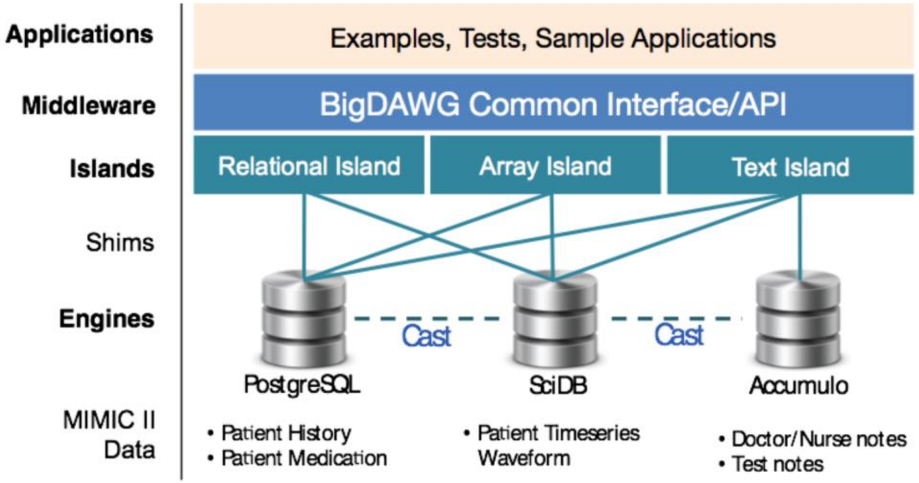
This approach reduces complexities in biomedical data management and accelerates research.



Related Polystore Solutions

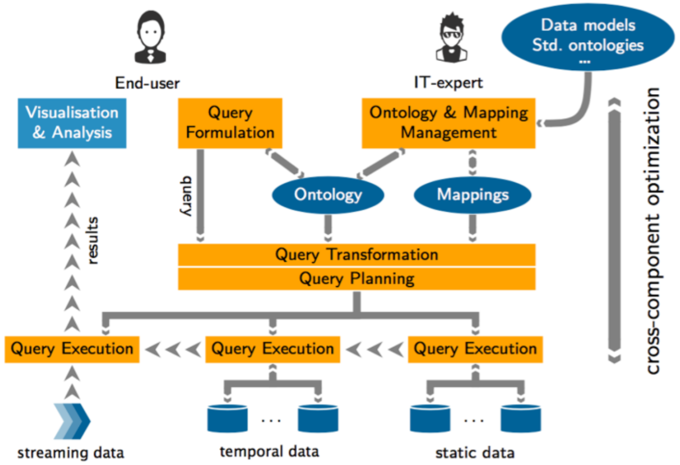
BigDAWG

- It is structured in 4 layers, from the data stores layer up to the application layer.
- A crucial role is played by the "**Island layer**", a layer of autonomous components that abstracts data sources.



Optique

- Initially conceived as an EU funded project, now implemented and maintained by Siemens.
- 3-layers architecture, **OBDA based**, optimized for streaming data. Provides a visual query tool to ease its use.



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) Outcomes:

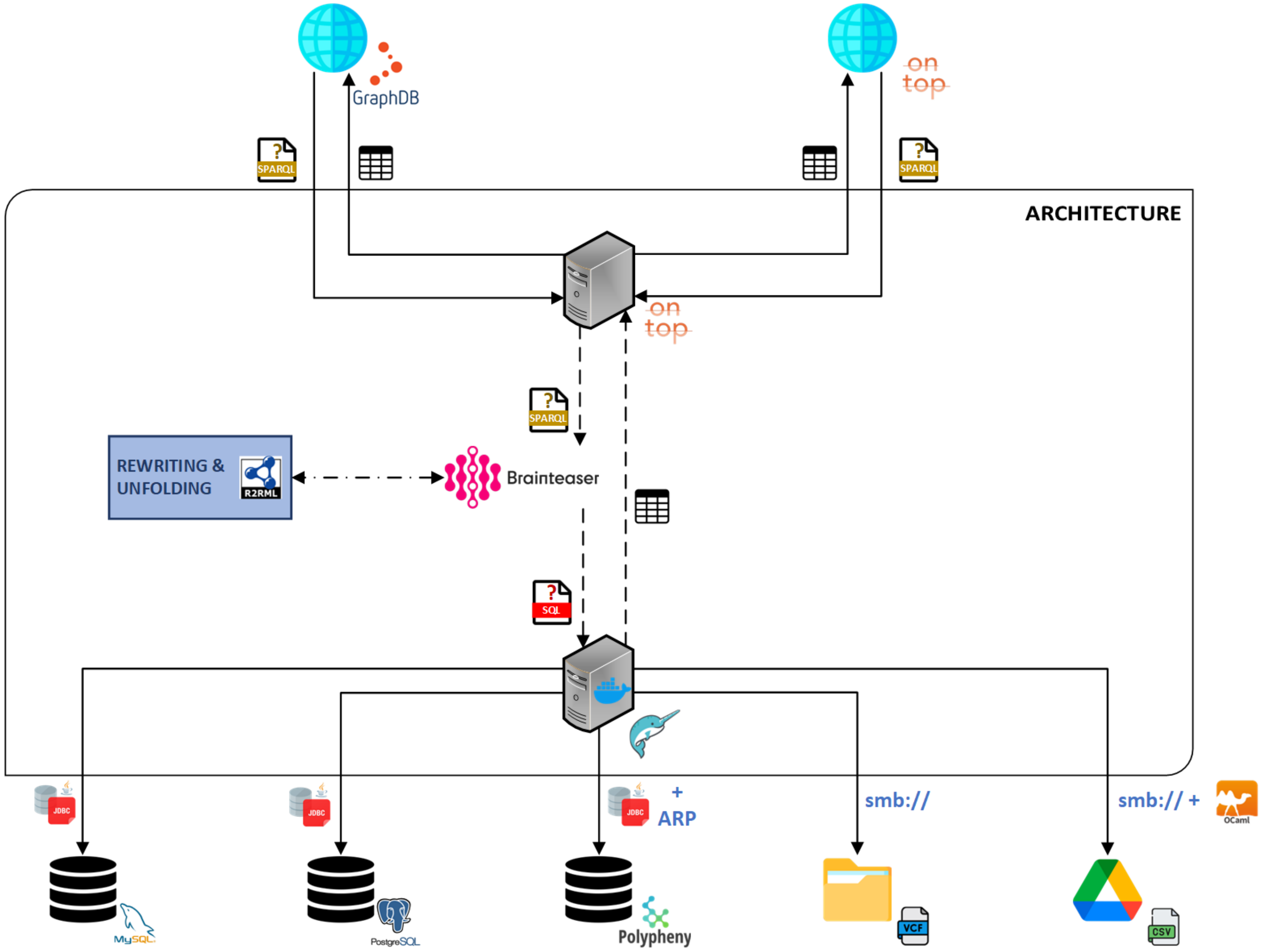
A unified DBMS that supports advanced querying capabilities across different data sources and that allows to gather insightful analytics.

Funded by
the European Union

Led by the Department of
Information Engineering (UNIPD)

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

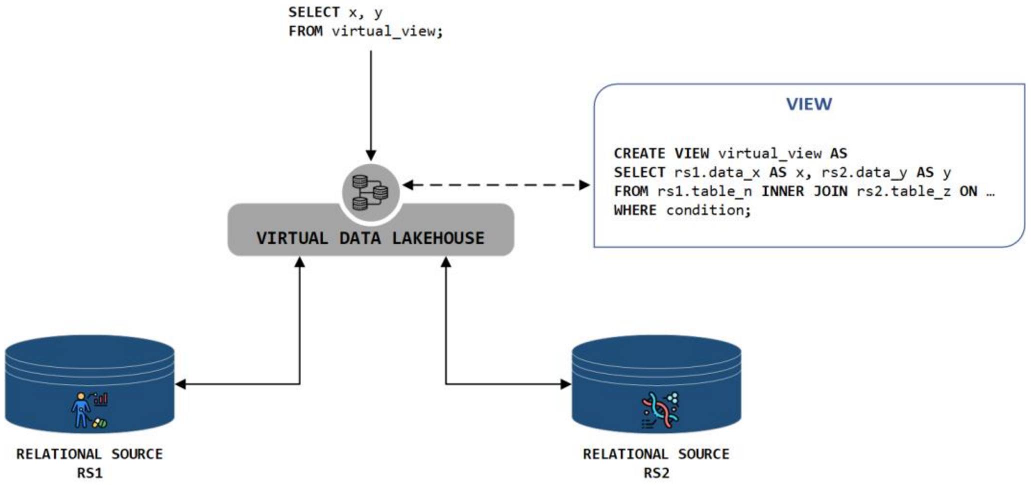
Virtualization Layer:

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.



```
SELECT x, y
FROM virtual_view;
```

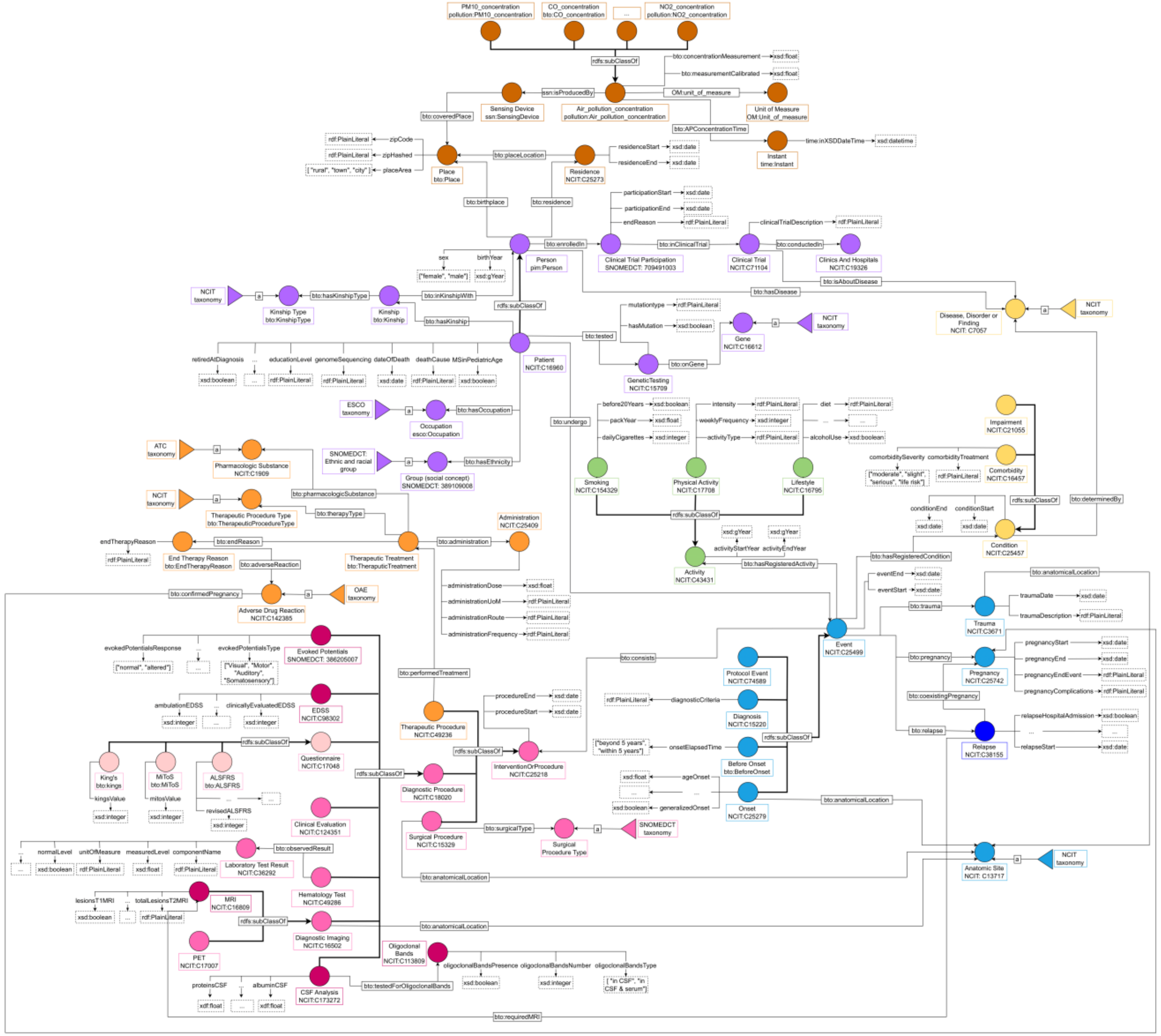
```
CREATE VIEW virtual_view AS
SELECT rs1.data_x AS x, rs2.data_y AS y
FROM rs1.table_n INNER JOIN rs2.table_z ON …
WHERE condition;
```

Ontology Integration

The **BRAINTEASER** Ontology is central to the system

Features:

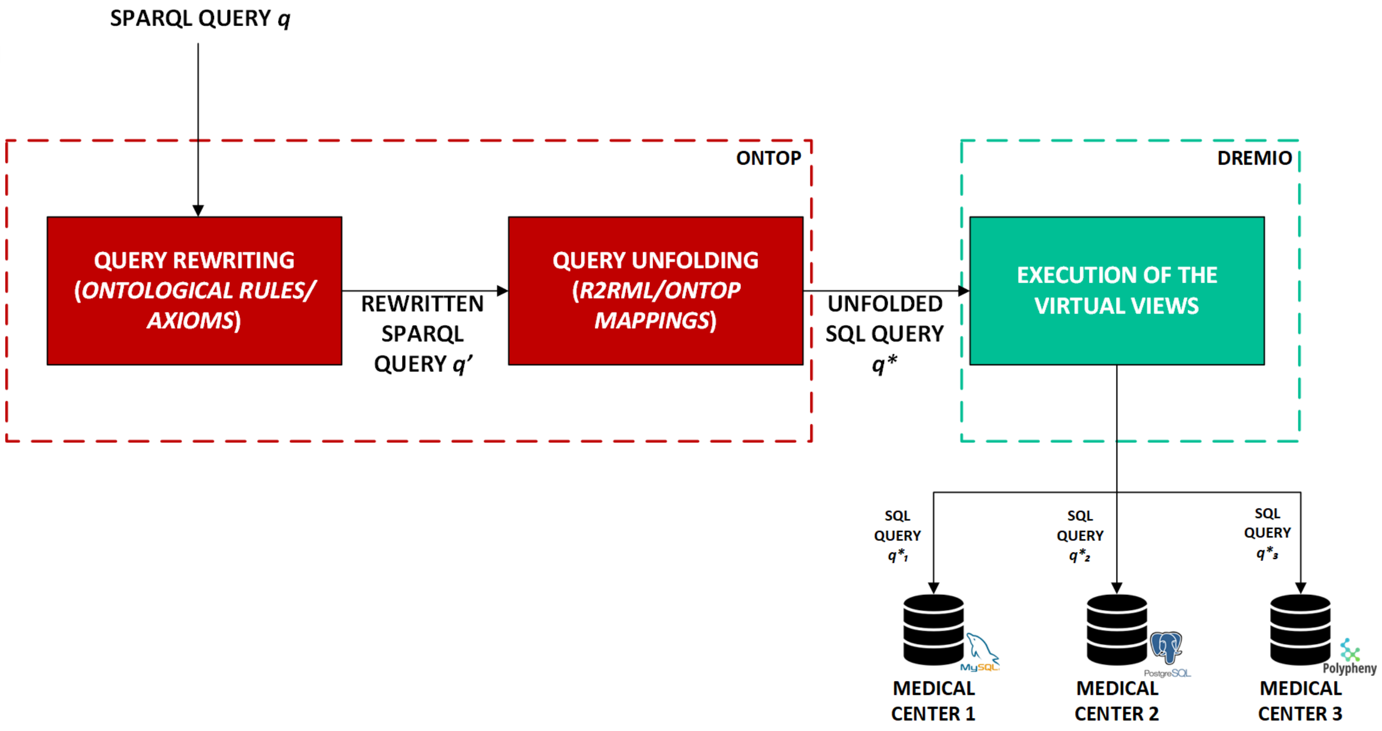
- Models relationships between clinical and genomics data;
- Enhances query capabilities through semantic enrichment;
- Ensures interoperability and scalability.

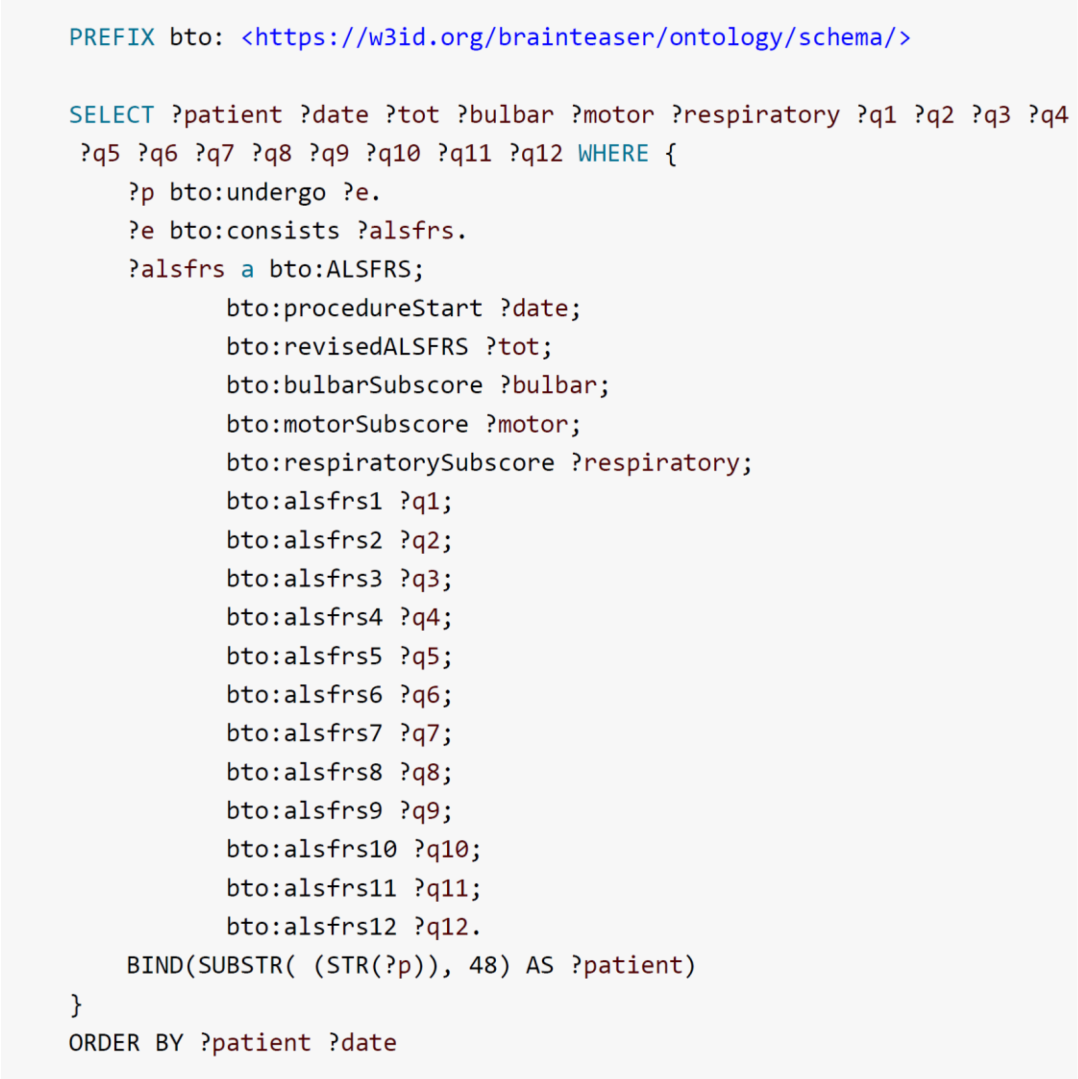

PM10_concentration pollution:PM10_concentration
CO_concentration bto:CO_concentration
NO2_concentration pollution:NO2_concentration
Sensing Device ssn:SensingDevice
Air_pollution_concentration pollution:Air_pollution_concentration
Unit of Measure OM:Unit_of_measure
Place bto:Place
Residence NCIT:C25273
Instant time:Instant
Person pim:Person
Clinical Trial Participation SNOMEDCT: 709491003
Clinical Trial NCIT:C71104
Clinics And Hospitals NCIT:C19326
Disease, Disorder or Finding NCIT: C7057
NCIT taxonomy
Kinship Type bto:KinshipType
Kinship bto:Kinship
Patient NCIT:C16960
Gene NCIT:C16612
GeneticTesting NCIT:C15709
ESCO taxonomy
Occupation esco:Occupation
SNOMEDCT Ethnic and racial group
Group (social concept) SNOMEDCT: 389109008
ATC taxonomy
Pharmacologic Substance NCIT:C1909
Therapeutic Procedure Type bto:TherapeuticProcedureType
End Therapy Reason bto:EndTherapyReason
Therapeutic Treatment bto:TherapeuticTreatment
Administration NCIT:C25409
Adverse Drug Reaction NCIT:C142385
OAE taxonomy
Smoking NCIT:C154329
Physical Activity NCIT:C17708
Lifestyle NCIT:C16795
Activity NCIT:C43431
Impairment NCIT:C21055
Comorbidity NCIT:C16457
Condition NCIT:C25457
Event NCIT:C25499
Trauma NCIT:C3671
Pregnancy NCIT:C25742
Relapse NCIT:C38155
Protocol Event NCIT:C74589
Diagnosis NCIT:C15220
Before Onset bto:BeforeOnset
Onset NCIT:C25279
Evoked Potentials SNOMEDCT: 386205007
EDSS NCIT:C98302
Questionnaire NCIT:C17048
King's bto:kings
MiToS bto:MiToS
ALSFRS bto:ALSFRS
Clinical Evaluation NCIT:C124351
Therapeutic Procedure NCIT:C49236
InterventionOrProcedure NCIT:C25218
Diagnostic Procedure NCIT:C18020
Surgical Procedure NCIT:C15329
Surgical Procedure Type
SNOMEDCT taxonomy
Laboratory Test Result NCIT:C36292
Hematology Test NCIT:C49286
MRI NCIT:C16809
Diagnostic Imaging NCIT:C16502
PET NCIT:C17007
Oligoclonal Bands NCIT:C113809
CSF Analysis NCIT:C173272
Anatomic Site NCIT: C13717

Use Case: ALSFRS Data Querying

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

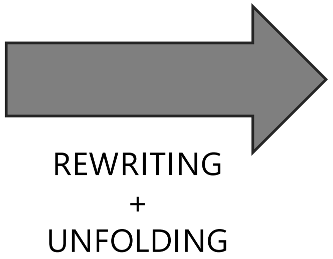
- **SPARQL Query**: Retrieves comprehensive patient data;
- **Process**:
  - Query Rewriting;
  - Unfolding into SQL;
  - Execution in Dremio.



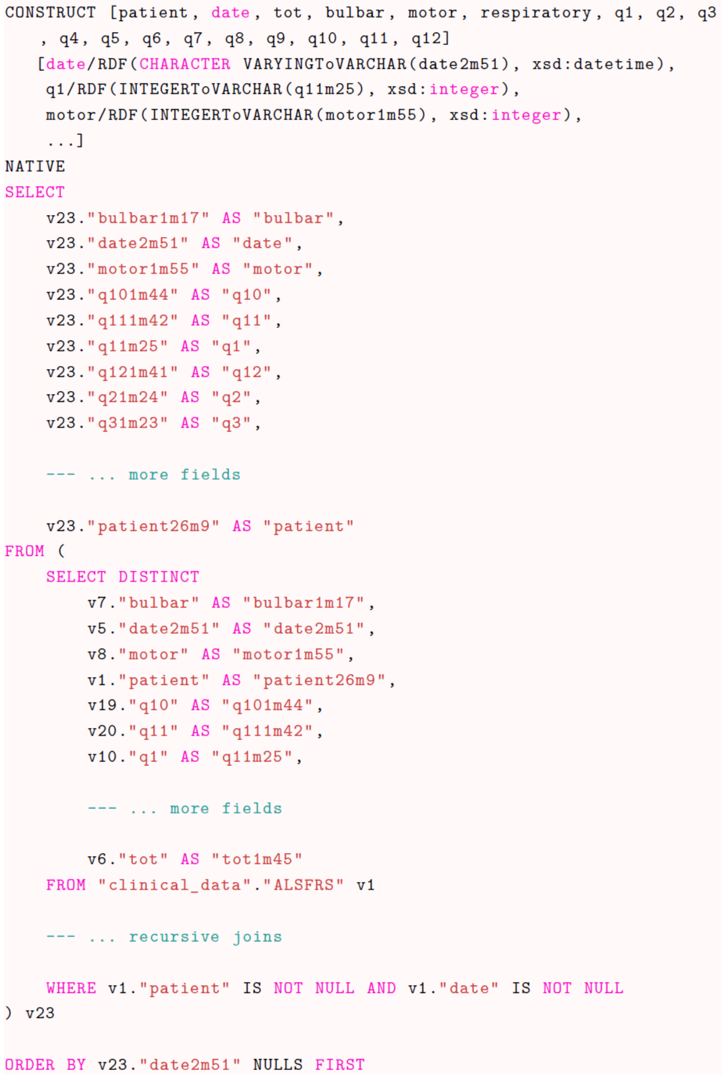


```
PREFIX bto: <https://w3id.org/brainteaser/ontology/schema/>

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



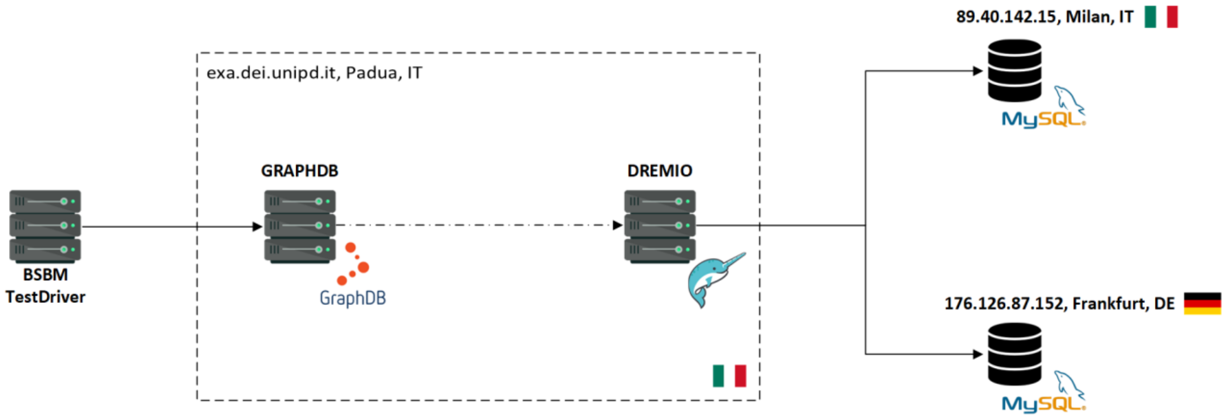
REWRITING
+
UNFOLDING



```
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
```

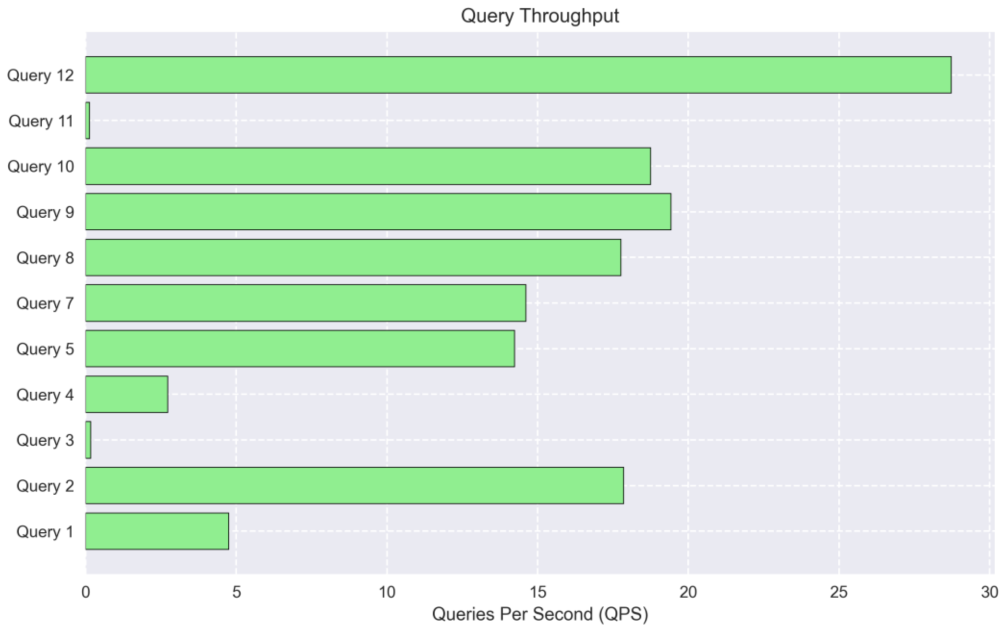
Unfolded SQL Query

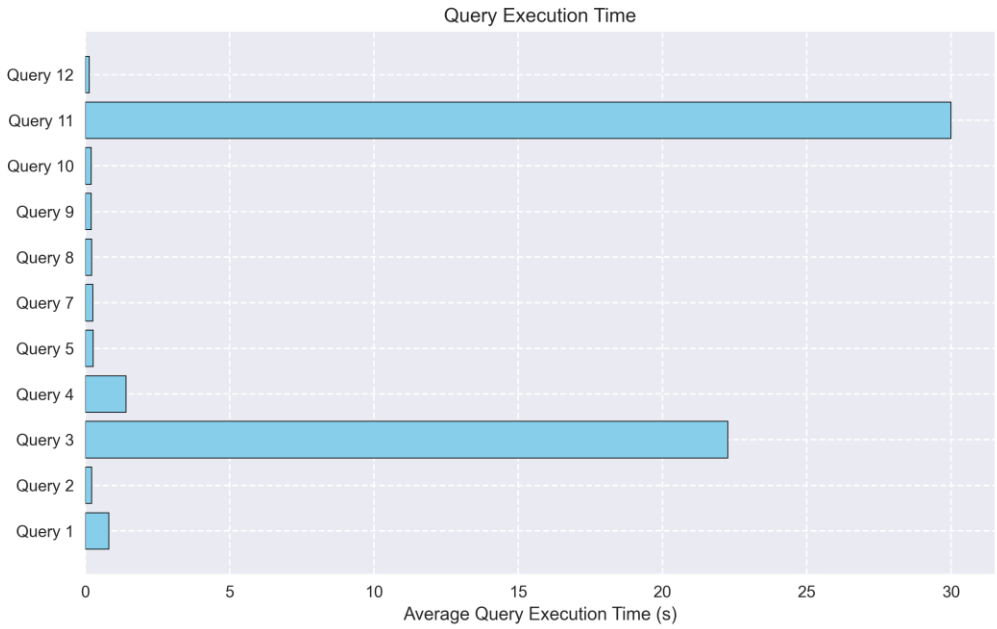
Benchmark 1: BSBM



BSBM Benchmark

Query performances evaluation over a synthetic dataset across diverse sources.

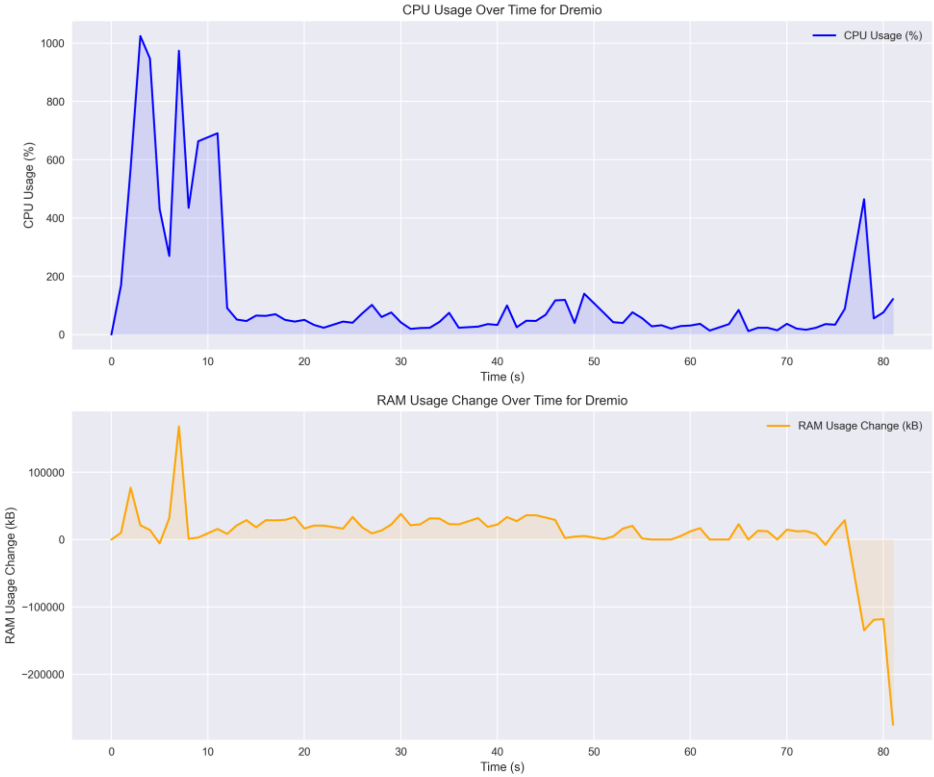




Benchmark 2: SEASHELL

SEASHELL Benchmark

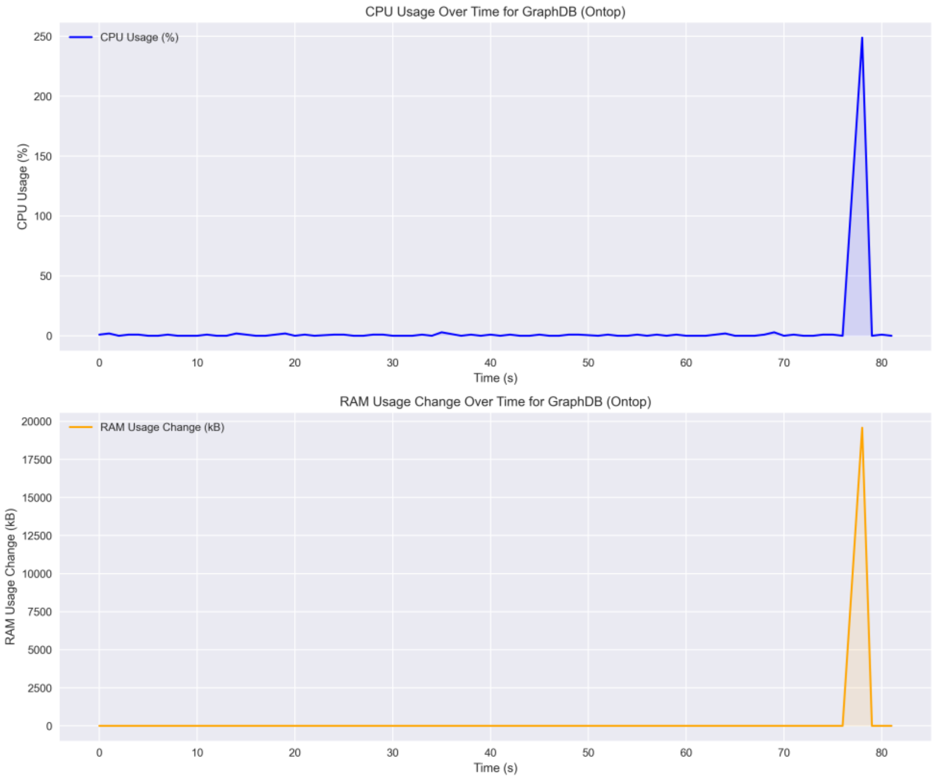
Real-world clinical data highlights system efficiency in handling complex queries.



CPU Usage Over Time for Dremio

RAM Usage Change Over Time for Dremio

Dremio



CPU Usage Over Time for GraphDB (Ontop)

RAM Usage Change Over Time for GraphDB (Ontop)

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?

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

UNIVERSITÀ
DEGLI STUDI
DI PADOVA