Instituto Superior de Engenharia de Lisboa
BSc in Computer Science and Computer Engineering
Project and Seminary

PHYLOViZ Web Platform

A Modular and Web-Based Tool for Phylogenetic Analysis

Supervisors:
- Cátia Vaz, ISEL
- Alexandre P. Francisco, IST

48089 André Páscoa
48280 André Jesus
48287 Nyckollas Brandão

July 2023

Table of contents

01 Introduction

02 Motivation

03 Architecture

04 Implementation

05 Software Stack

06 Testing

07 Demonstration

08 Conclusion

01

Introduction

What's **Phylogenetic Analysis**?

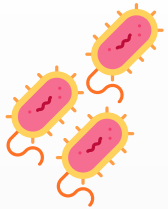
Introduction

Phylogenetic Analysis is a field of biomedical research which allow to understand the evolution of bacterial and viral epidemics.

Phylogenetic Analysis

1. Sequence Alignment
2. Typing Methodology
3. Inference Algorithms
4. Visualization Algorithms

Pipeline



Bacteria

DNA

```
> Sequence 1
GAAGCGAGTGACTTGGCAGAAACAGTGGCC
GCGCGCTTGATTGGTGCGGTTAATACGGTT
> Sequence 2
GAACCGAGTGACTTGGCAGAAACAGTGGCC
GCGCGCTTGATTGGTGCGGTTAATACGGTT
```

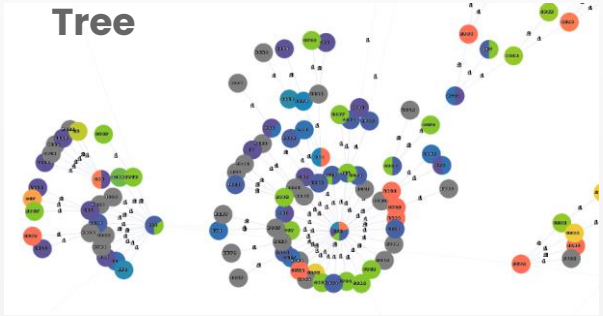
DNA Sequences

ST	*nusA*	*rpoB*	*eno*	*gltB*	*lepA*	*nuoL*	*aroe*
1	1	26	2	2	59	8	1
2	1	26	2	4	59	2	1
3	1	26	2	2	62	8	2
4	1	26	7	2	59	3	2
5	1	27	1	1	62	9	1

Typing Data

$$D = \begin{bmatrix} 0 & 2 & 2 & 3 & 5 \\ 2 & 0 & 4 & 4 & 5 \\ 2 & 4 & 0 & 3 & 5 \\ 3 & 4 & 3 & 0 & 6 \\ 5 & 5 & 5 & 6 & 0 \end{bmatrix}$$

Distance Matrix



Isolate Data

id	*ST*	*isolate*	*species*	*country*	*continent*
1	1	*AU2523*	*A.denitrificans*	*USA*	*NorthAmerica*
2	2	*AU8059*	*A.denitrificans*	*Unknown*	
3	3	*AU8060*	*A.denitrificans*	*France*	*Europe*
4	1	*AU8080*	*A.insolitus*	*USA*	*NorthAmerica*
5	5	*ACH26*	*A.insolitus*	*USA*	*NorthAmerica*

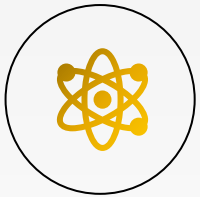
02

Motivation

Related Tools

PHYLOViZ

Desktop application for Phylogenetic Analysis



PHYLOViZ Online

Web application for Phylogenetic Analysis

FLOWViZ

Workflow Manager

PhyloDB

Graph-Oriented Database

Phylolib

Library of Phylogenetic Analysis Algorithms

Problem

- **No future support** for frameworks used by PHYLOViZ Desktop
- Features in both versions of PHYLOViZ are **not the same**
- PHYLOViZ Online is **not modular**, which makes it difficult to extend and maintain
- Current solutions **do not scale** for large data analysis and visualization
- Results and optimizations are not stored for reuse

Solution: PHYLOViZ Web Platform

- **Modular Architecture**
- **Web Based**
- **Advanced Data Management**
- **Data Centric Workflows**

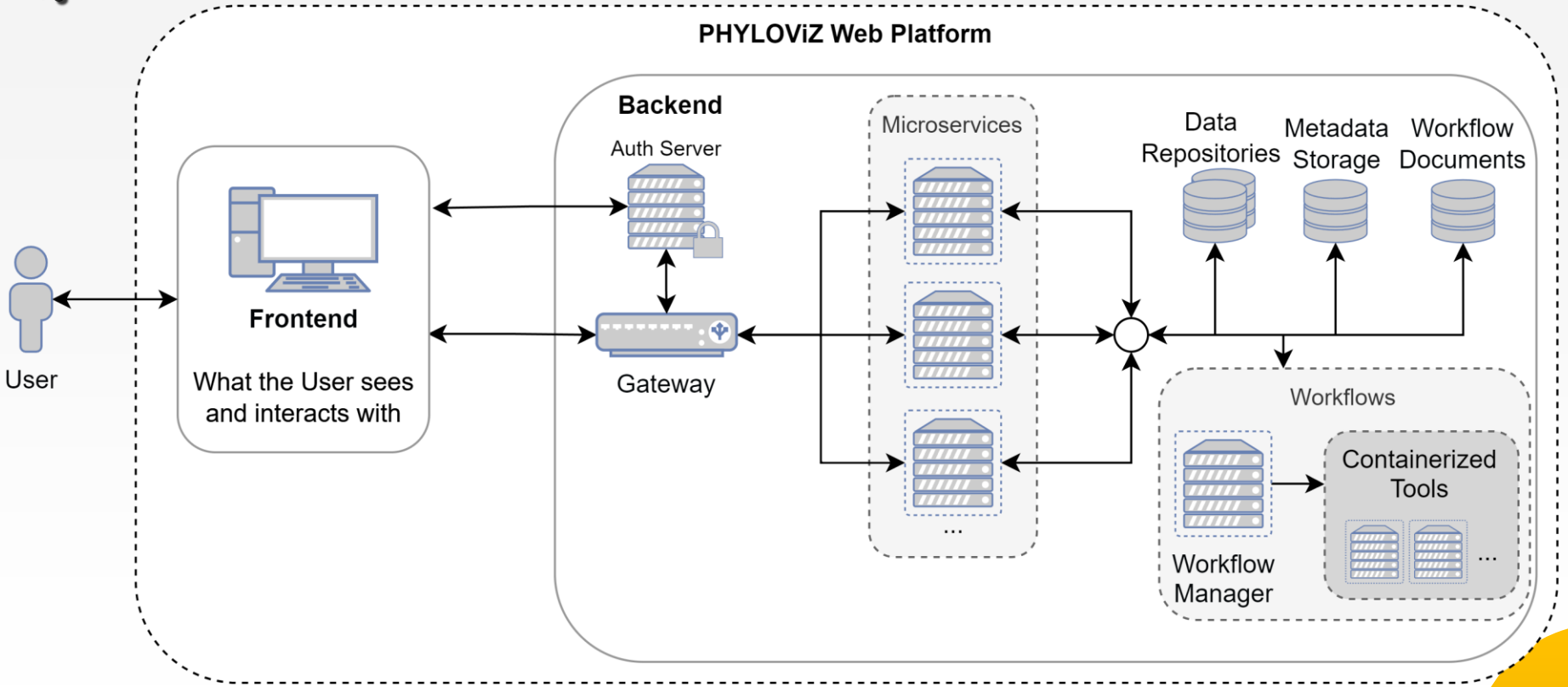
Why Modular?

Allows seamless integration of new modules/tools, creating/editing **workflows**

Customization of **data repositories**, providing flexibility in storing and managing phylogenetic data

03

Architecture


PHYLOViZ Web Platform
Backend
Auth Server
Gateway
Microservices
...
Data Repositories
Metadata Storage
Workflow Documents
Workflows
Workflow Manager
Containerized Tools
...
Frontend
What the User sees and interacts with
User

Data Model

Phylogenetic Data

PhyloDB

Model and Database used for first deployment. Uses the graph database **neo4j**.

OpenStack S3

Object storage of OpenStack used for file storage.

Others

Other repositories can be easily used, even at the same time, because of **metadata storage**.

Data Model

Metadata

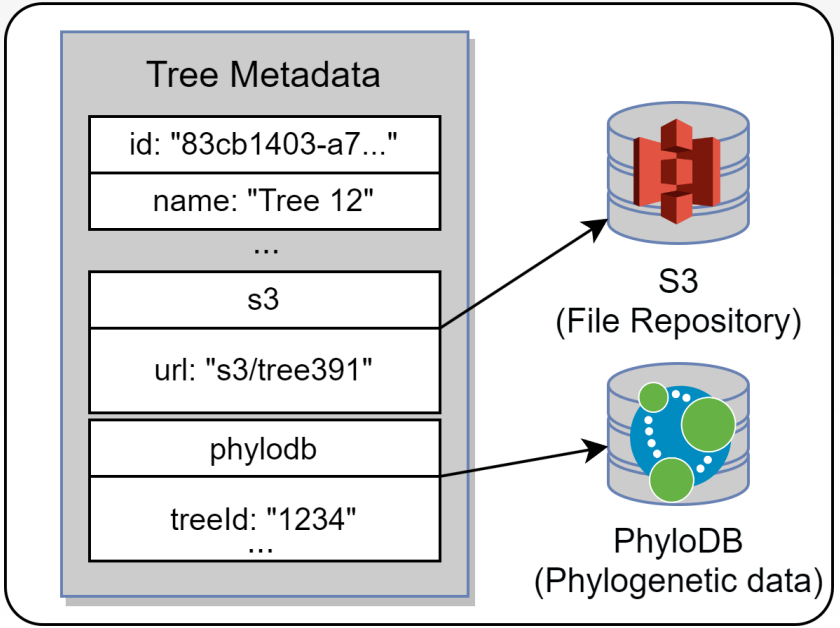
MongoDB

- **Abstraction** between application and data repositories.
- Document in MongoDB.
- Stores **information of each resource** – no need to change existing databases to fit the data model of our application.
- Stores **access information** for each data repository – multiple data representations.

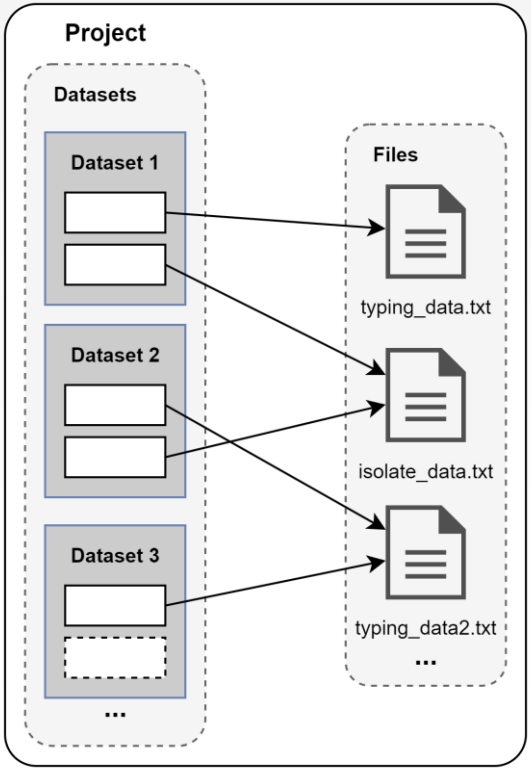
Data Model

Multiple data representations

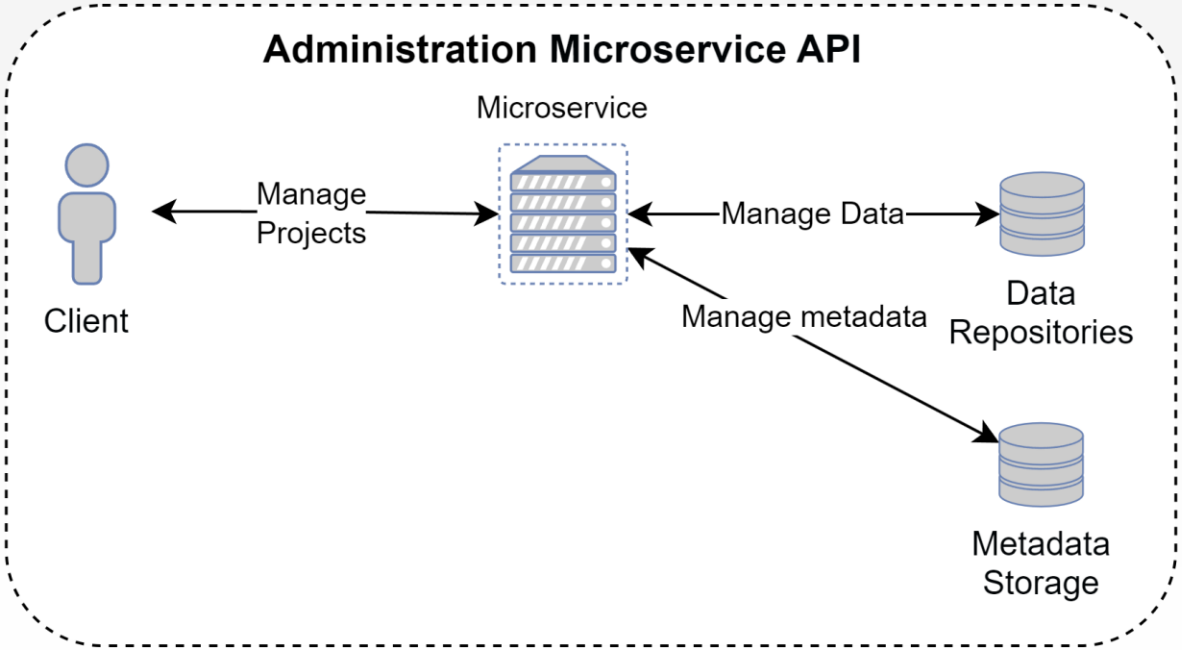
- Integration with **multiple** data repositories
- Different data representations to be used in different operations - **flexibility**.
- Data **replicated** across the repositories.



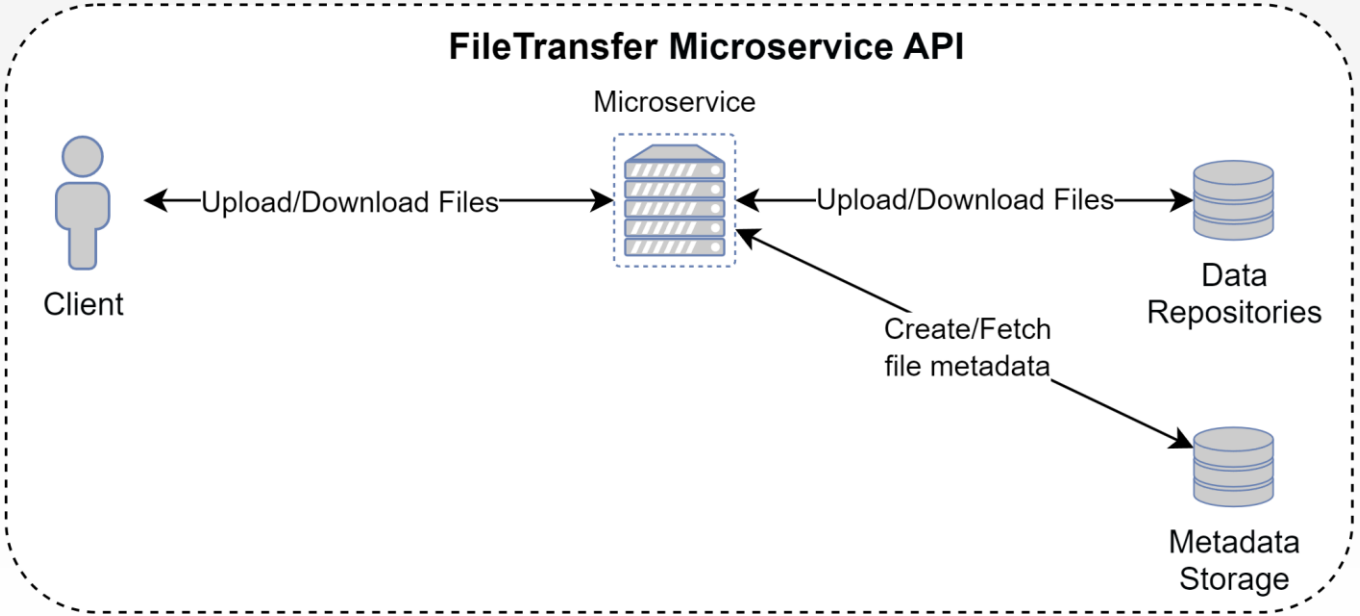
Projects, Datasets and Files



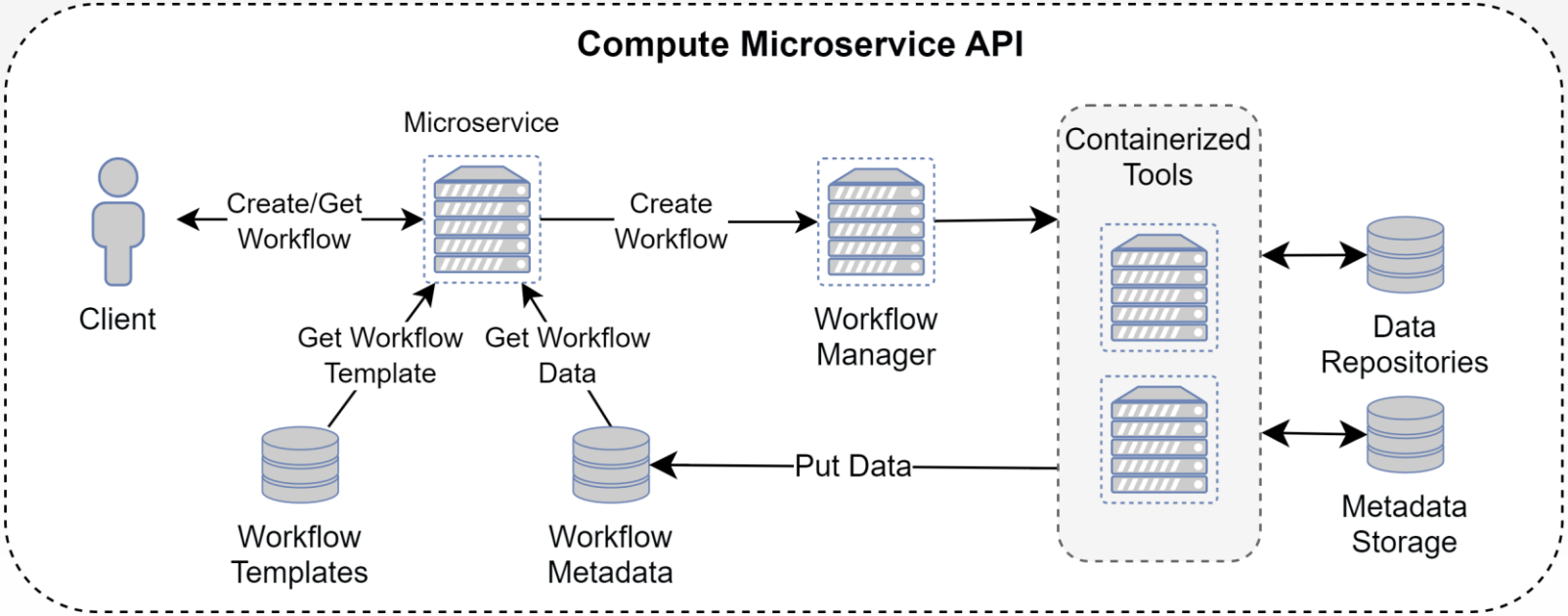
Administration Microservice



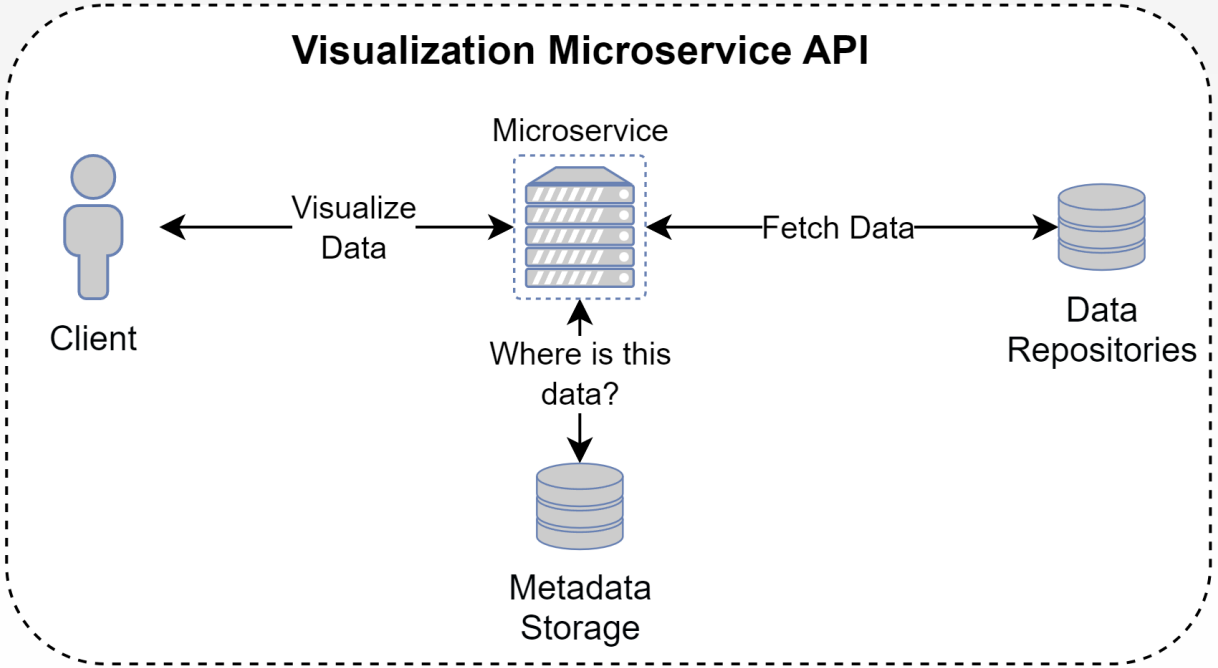
FileTransfer Microservice



Compute Microservice



Visualization Microservice



04

Implementation

Metadata Documents

Project

```
_id: ObjectId('6442f8da3903160faccddf54')
name: "Project1"
description: ""
ownerId: "914cc356-ac86-4ab4-909c-bd02d3776a7b"
```

Dataset

```
_id: ObjectId('6442fb1d3903160faccddf62')
projectId: "6442f8da3903160faccddf54"
name: "Dataset1"
description: "My first dataset."
typingDataId: "7a01d824-e9a7-49cc-8d9d-7dacdcb2e92c"
```

Typing Data

```
_id: ObjectId('645d60621199246130dc94ea')
name: "allele_profiles.txt"
typingDataId: "026dcfaa-127d-4863-9208-55c1c539b983"
projectId: "645d60417f92b75799a8c86d"
▾ repositorySpecificData: Object
  ▾ s3: Object
      url: "http://localhost:9444/phyloviz-web-platform/645d6041
      originalFilename: "allele_profiles.txt"
  ▾ phylodb: Object
      projectId: "01686wcfa29"
type: "ML"
```

Isolate Data

```
_id: ObjectId('645d60c91199246130dc94f0')
▾ repositorySpecificData: Object
  ▾ s3: Object
      url: "http://localhost:9444/phyloviz-web-platform/645d60417f9
      originalFilename: "isolates.txt"
projectId: "645d60417f92b75799a8c86d"
name: "isolates.txt"
isolateDataId: "ac4fd7cb-6a00-4755-a771-66b8e30ee027"
```

Metadata Documents

Distance Matrix


```
_id: ObjectId('645dfcf8ba62c99ef7a314cf')
projectId: "645d60417f92b75799a8c86d"
datasetId: "645d60a27f92b75799a8c86e"
distanceMatrixId: "bf26bbd0-adcf-4162-8973-bd65c5ab1ed4"
name: "Distance Matrix bf26bbd0-adcf-4162-8973-bd65c5ab1ed4"
sourceType: "function"
▾ source: Object
    function: "kimura"
▾ repositorySpecificData: Object
  ▾ s3: Object
      url: "http://localhost:9444/phyloviz-web-platform/645d60417f92
```


Tree

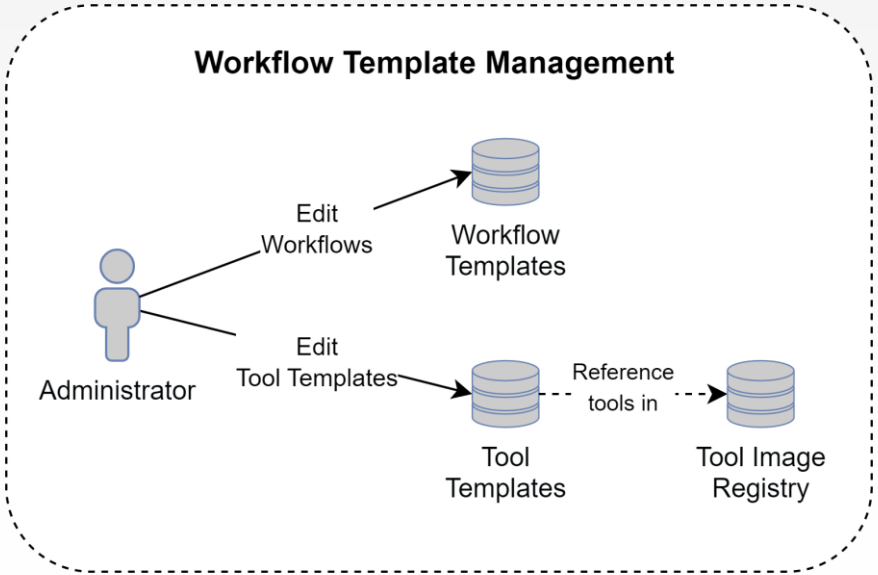

```
_id: ObjectId('645dfcf8ba62c99ef7a314c1')
projectId: "645d60417f92b75799a8c86d"
datasetId: "645d60a27f92b75799a8c86e"
treeId: "a13123bd0-a33f-4262-8973-bd65c52d4"
name: "Tree a13123bd0-a33f-4262-8973-bd65c52d4"
sourceType: "algorithm_distance_matrix"
▸ source: Object
▾ repositorySpecificData: Object
  ▾ phylodb: Object
      projectId: "2j9dk461"
      datasetId: "3j18ddk2"
      inferenceId: "a863-a33f-4oi2-89g5-13vsees"
```


Tree View

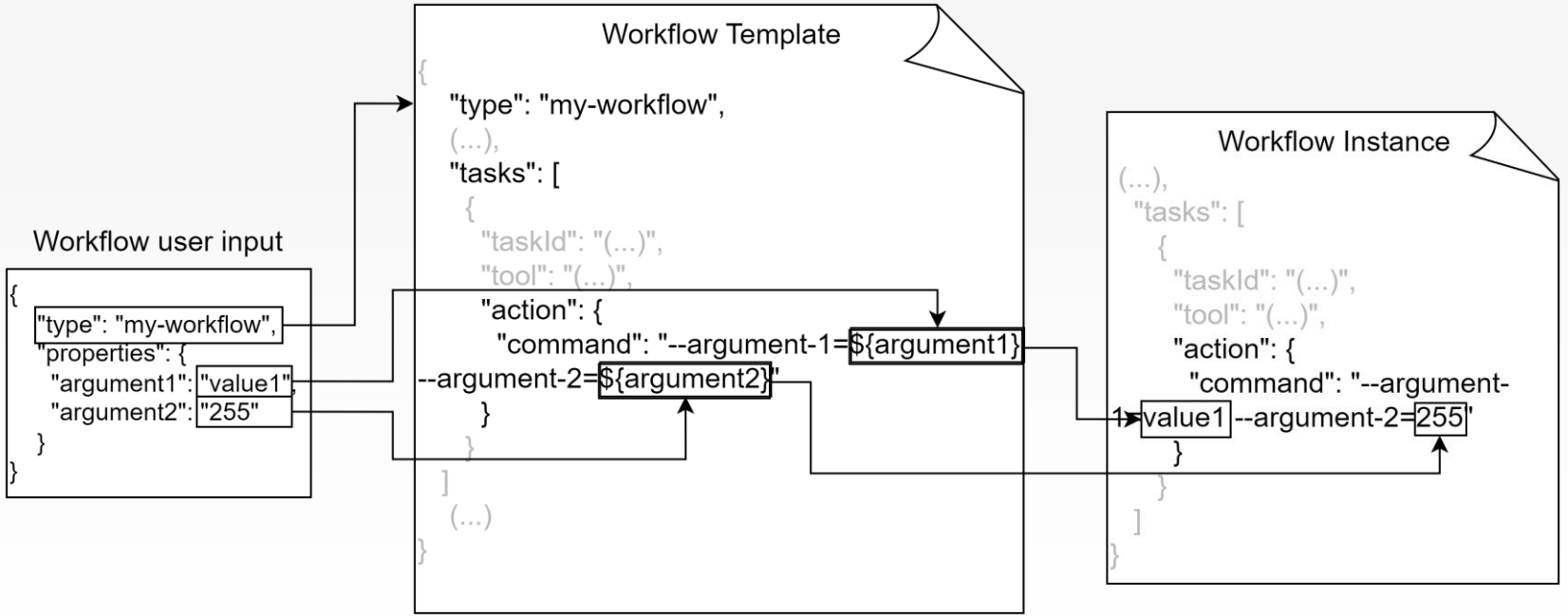

```
_id: ObjectId('64a309ee43b9d415e11290aa')
▸ source: Object
layout: "force-directed"
datasetId: "6488bbf42b06c76a283e2c79"
treeViewId: "1cf29267-af66-457c-9f8e-7285c3dacf79"
projectId: "6488a5122b06c76a283e2c70"
name: "Tree View 1 - force-directed"
▾ transformations: Object
    gravity: 0.01
    repulsion: 0.2
    nodeSize: 4
    nodeLabel: true
    nodeLabelSize: 0
        ...
▾ repositorySpecificData: Object
```

Workflows

- **Workflow** and **tool templates** are managed by the **system administrator** and stored in a document database.
- Ideally tool images are previously uploaded to the custom **Docker image registry**.
- Workflows and tools can be added, edited and removed during **runtime**.

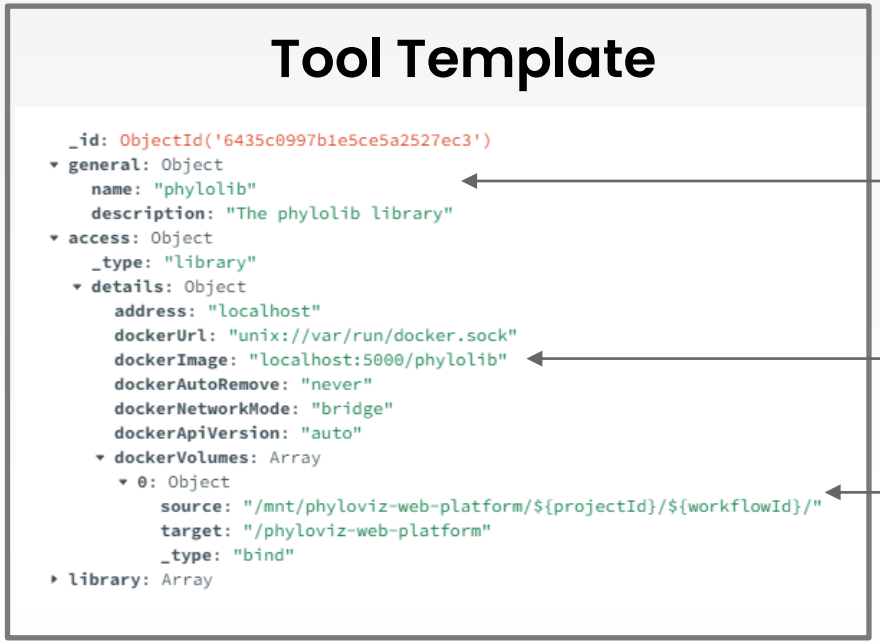


Creating a new Workflow Instance



Tool Document

Tool Template



Tree Visualization

- Cosmos (cosmograph) used as a base, to provide the **force directed layout**, showing the tree as a graph and running the **simulation**.

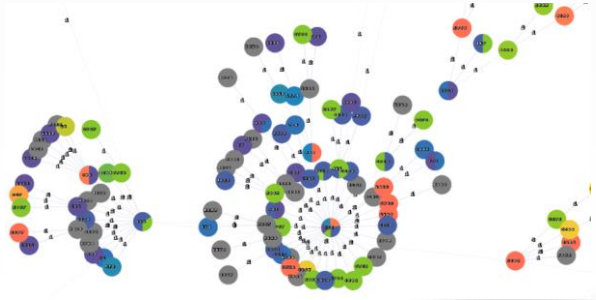
Cosmos

- Cosmos uses **WebGL** to run the simulation on the **GPU**, making use of its **parallel computation capabilities**.

WebGL™

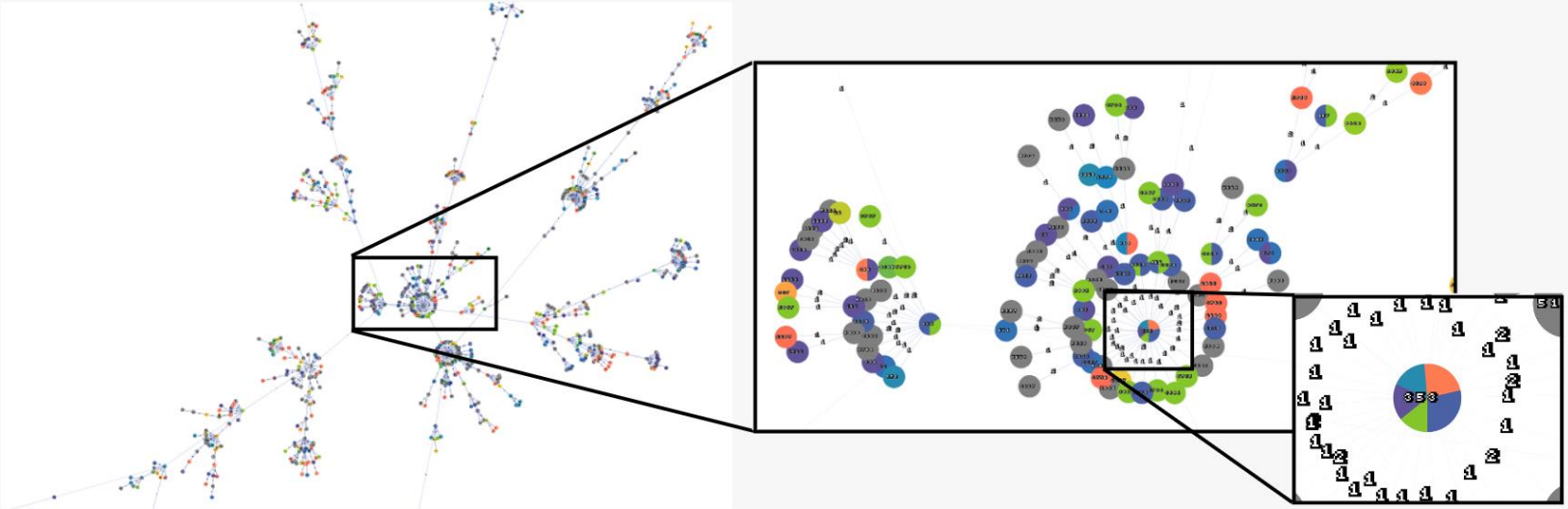
Several additions to cosmos were made:

- **Labels** on nodes and edges;
- **Draggable** nodes;
- **Pie-charts** on nodes to represent ancillary data from isolates.



Tree Visualization

- High scalability, up to **15000 nodes** per cluster visualization.
- Such scalability is not possible in PHYLOViZ Online.



Deployment

- Deployed at **web.phyloviz.net**
- Authentication at **auth.phyloviz.net**
- Infrastructure **Biodata**, within private clouds hosted by Instituto Superior Técnico, using **OpenStack**
- **Two VMs**: one for Apache Airflow and other for the platform itself

05

Software Stack

Backend Technologies

Spring

Gateway and Microservices

Java

Programming Language

Keycloak

Authorization Server

Apache Airflow

Workflow Management Platform

Backend Technologies

OpenStack

S3 Buckets for Big Data and VMs

MongoDB

Metadata Storage

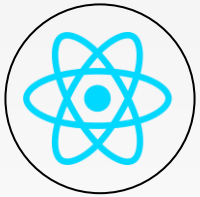
Docker

Containerization Platform

Python

Language used for some starter compute tools

Frontend Technologies



React

User Interface

Material UI

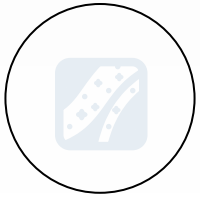
React Library

TypeScript

Programming Language

Webpack

Module bundling



Cosmos

Tree Visualization

06

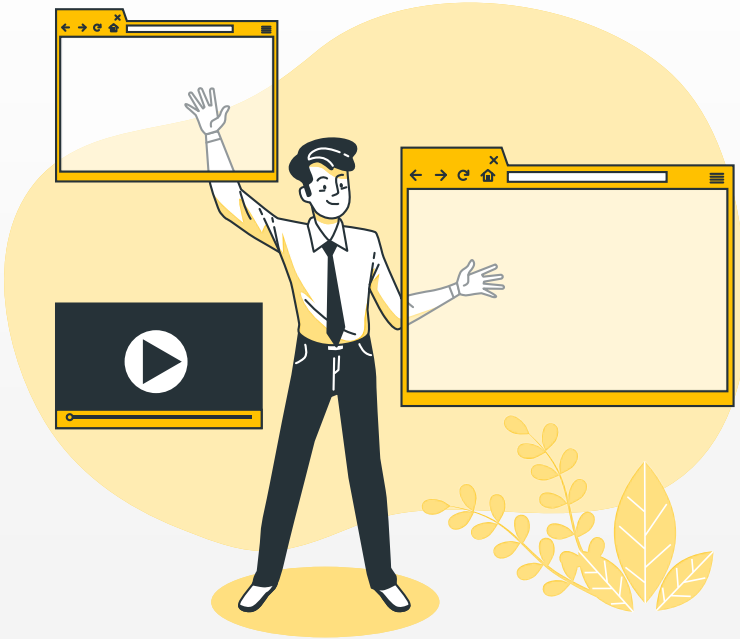
Testing

Testing

- **Manual Testing:** Service mocks in the frontend and Postman for API testing
- **Automatic Testing:** Unitary tests and Integration Tests
- **User Acceptance Testing:** to be conducted by a dedicated team from Institute of Molecular Medicine (iMM), in the future

07

Demonstration



08

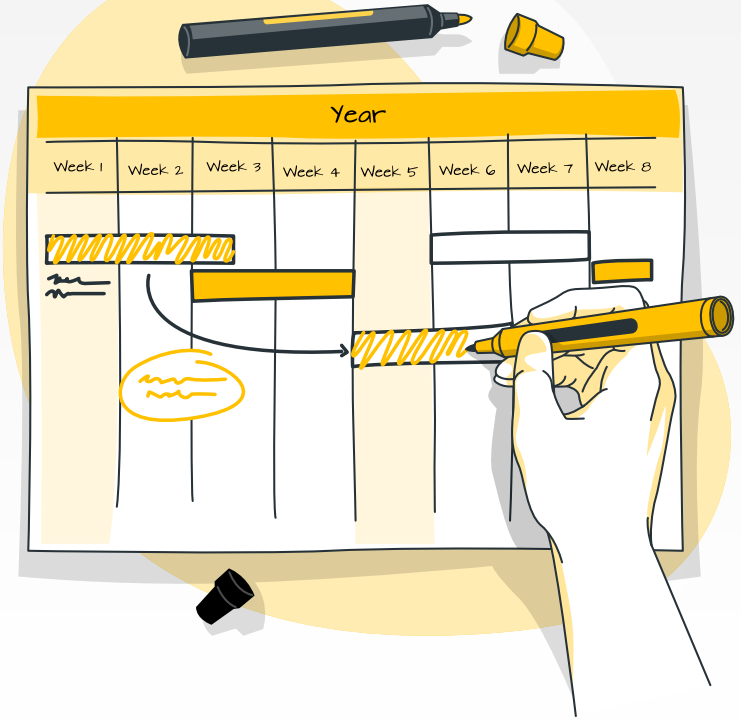
Conclusion

Conclusions

- The PHYLOViZ Web Platform is a **modular** and **extensible** tool for phylogenetic analysis.
- It supports large-scale analyses through parallel computations on **cloud** and HPC environments.
- The platform integrates existing tools and provides **efficient data management** and analysis capabilities.
- It offers a **user-friendly interface** and advanced features for phylogenetic analysis.

Future Work

- Enhanced Visualization, including Collapsible Tree Visualization
- Dynamic Workflows
- Community Engagement and Collaboration
- Integration with PubMLST or other databases



Resources

- Project Proposal
- Progress Presentation
- GitHub Repository
- Project Wiki
- Project Report
- Poster
- Final Presentation

Thanks!

Do you have any questions?

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon** and infographics & images by **Freepik**