

FLOWViZ

Framework for phylogenetic processing

Miguel Luís ^{1,2} e **Cátia Vaz** ^{1,2}

¹ INESC-ID Lisboa

² Instituto Superior de Engenharia de Lisboa, Instituto Politécnico de Lisboa

Contexto

A Filogenia é o estudo da evolução entre grupos de organismos e das suas características.

Figura 1 - Um exemplo de uma árvore filogenética. **Fonte:** https://www.researchgate.net/figure/Phylogeny-of-major-groups-of-land-plants-Based-on-13151920-Approximate-numbers-of_fig1_267043482

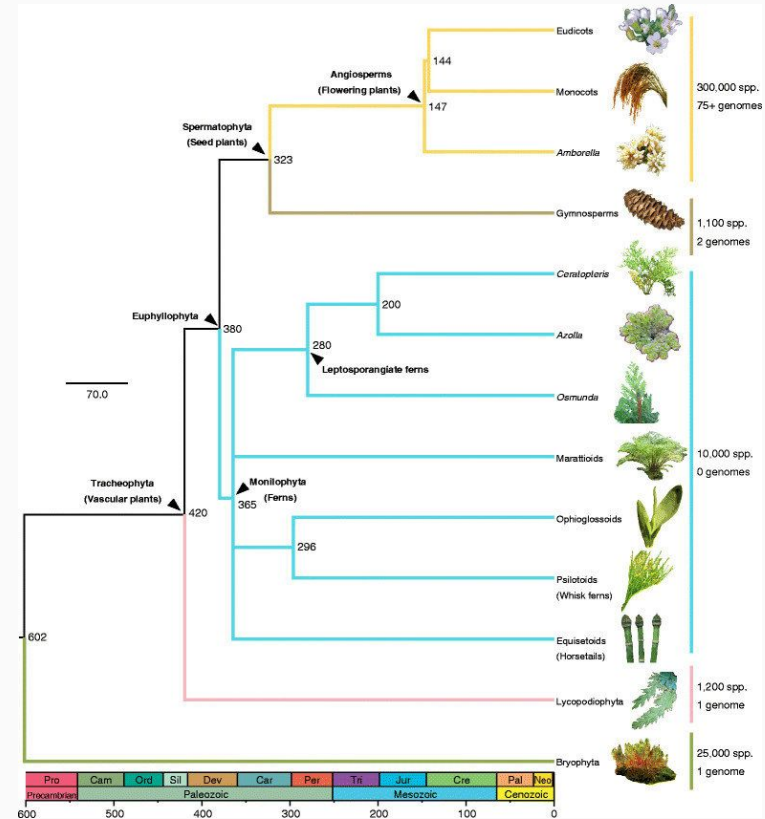


Figura 1

- **Processos complexos**, compostos por **vários passos ou tarefas**;
- Fluxogramas, denominados por ***workflows*** ou ***work pipelines***;
- **Vários *inputs/outputs* por passo** ou tarefa;

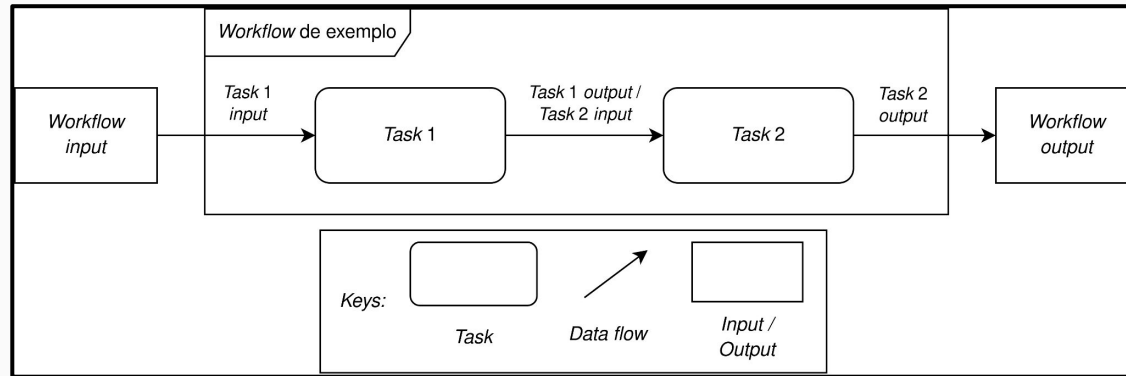
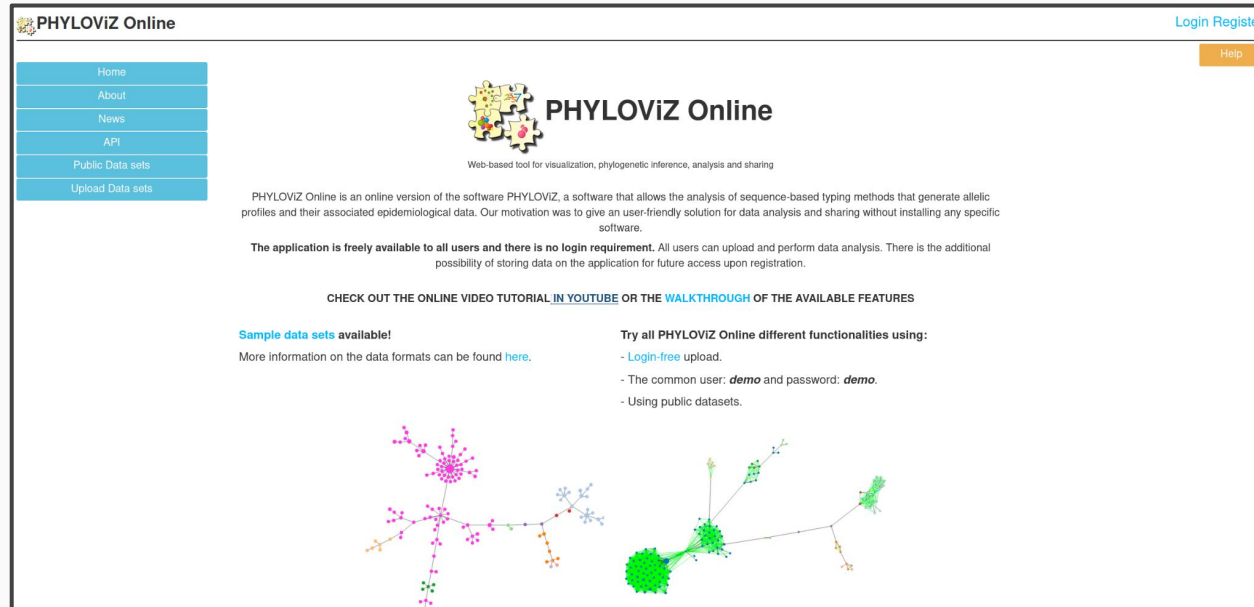


Figura 2 - Estrutura geral de um *workflow*

Ferramentas filogenéticas existentes



The screenshot shows the homepage of PHYLOViZ Online. At the top left is the logo and name 'PHYLOViZ Online'. At the top right are links for 'Login Register' and a 'Help' button. A left sidebar contains a menu with links: Home, About, News, API, Public Data sets, and Upload Data sets. The main content area features the PHYLOViZ Online logo (a cluster of colorful puzzle pieces) and the text 'Web-based tool for visualization, phylogenetic inference, analysis and sharing'. Below this is a paragraph describing the tool as an online version of PHYLOViZ software for sequence-based typing methods. It states that the application is freely available to all users with no login requirement. A link to a YouTube tutorial and a walkthrough of features are provided. There is also a section for 'Sample data sets available!' with a link to more information. A list of functionalities includes login-free upload, a demo user and password, and the use of public datasets. At the bottom, there are two phylogenetic tree visualizations: one with pink and blue nodes and another with green and orange nodes.

PHYLOViZ Online

Login Register

Help

Home
About
News
API
Public Data sets
Upload Data sets

PHYLOViZ Online

Web-based tool for visualization, phylogenetic inference, analysis and sharing

PHYLOViZ Online is an online version of the software PHYLOViZ, a software that allows the analysis of sequence-based typing methods that generate allelic profiles and their associated epidemiological data. Our motivation was to give a user-friendly solution for data analysis and sharing without installing any specific software.

The application is freely available to all users and there is no login requirement. All users can upload and perform data analysis. There is the additional possibility of storing data on the application for future access upon registration.

CHECK OUT THE ONLINE VIDEO TUTORIAL [IN YOUTUBE](#) OR THE [WALKTHROUGH](#) OF THE AVAILABLE FEATURES

Sample data sets available!
More information on the data formats can be found [here](#).

Try all PHYLOViZ Online different functionalities using:

- [Login-free](#) upload.
- The common user: **demo** and password: **demo**.
- Using public datasets.

Phylogenetic tree visualizations showing clusters of nodes.

Figura 3 - Página principal da ferramenta filogenética *PHYLOViZ*.
Fonte: phyloviz.net

Problema

- Processos podem demorar **horas** ou **dias**;
- Processos que envolvem **grandes quantidades de informação**;
- **Configuração manual de workflows.**

Problema

Figura 4 - Exemplo de um *workflow* complexo. **Fonte:** https://www.researchgate.net/figure/A-complex-DAG-structure_fig4_7885204

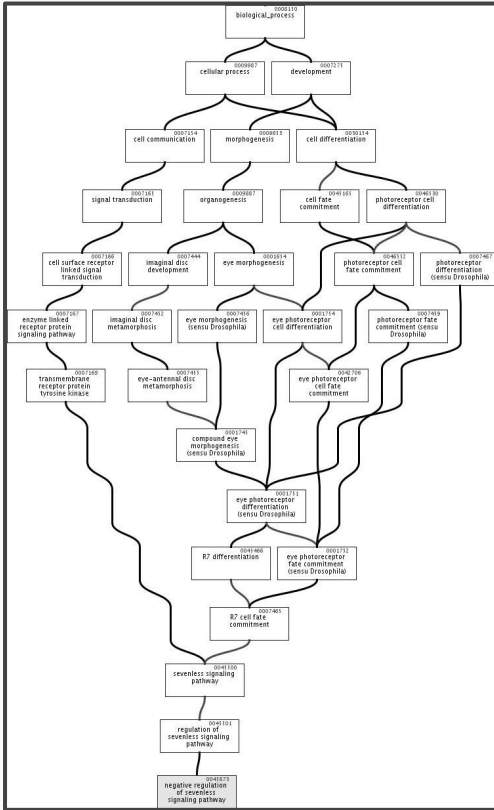


Figura 4

Sistemas de *workflow*

- Fornecem uma **linguagem própria** para construir workflows - ***Domain-Specific Language*** (DSL);
- Permitem **automatizar** processos;
- Recorrem a **computação distribuída** para executar *workflows* e permitir escalabilidade de recursos;
- **Problema:** Várias soluções existentes, mas **pouca integração** entre sistemas de workflow e ferramentas filogenéticas.

Ferramenta filogenética que integrou um sistema de *workflow*



Figura 5.1 - Página principal da ferramenta *NGPhylogeny*
Fonte: ngphylogeny.fr



TITLE NGPhylogeny Analyse - Workflow_generated_bn\QLxTc (imported from API)			
Tool	Step	File Name	Status
PhyML	7.	PhyML Newick tree: BMGE Cleaned sequences Phyml.nhx.svg	...
	6.	PhyML Newick tree: BMGE Cleaned sequences Phyml.nhx	...
	5.	PhyML statistic: BMGE Cleaned sequences Phyml.stats.txt	...
BMGE	4.	PhyML log: BMGE Cleaned sequences Phyml.log	...
	3.	BMGE Cleaned sequences Phyml	...
MAFFT	2.	MAFFT on data 1	...
Upload File	1.	1.phy	✓ .phyip

Figura 5.2 - Lista de tarefas num *workflow*
Fonte: ngphylogeny.fr

FLOWViZ

- Uma *framework* que **integra ferramentas filogenéticas** com o **sistema de workflow - Apache Airflow**;
- A integração de ferramentas, que correspondem aos passos de execução de um *workflow*, é feita a partir de **contratos**;
- Permite aos utilizadores **adicionarem ferramentas filogenéticas** e **executarem workflows** com as mesmas.

Contratos

- **Especificação de regras** entre entidades;
- Relações de natureza *loosely-coupled*:
 - **Adaptabilidade**;
 - **Flexibilidade**.

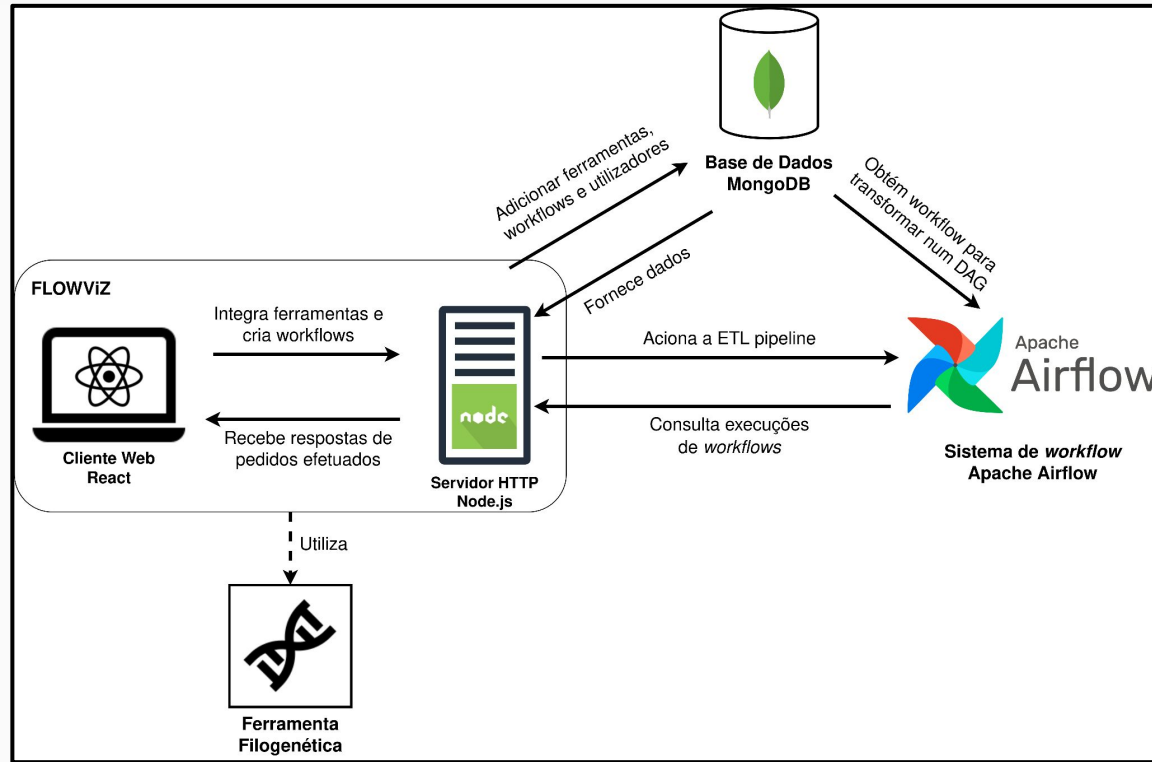


Figura 6 - Arquitetura do FLOWViZ e integração com ferramenta filogenética

Caso de uso - adicionar ferramenta

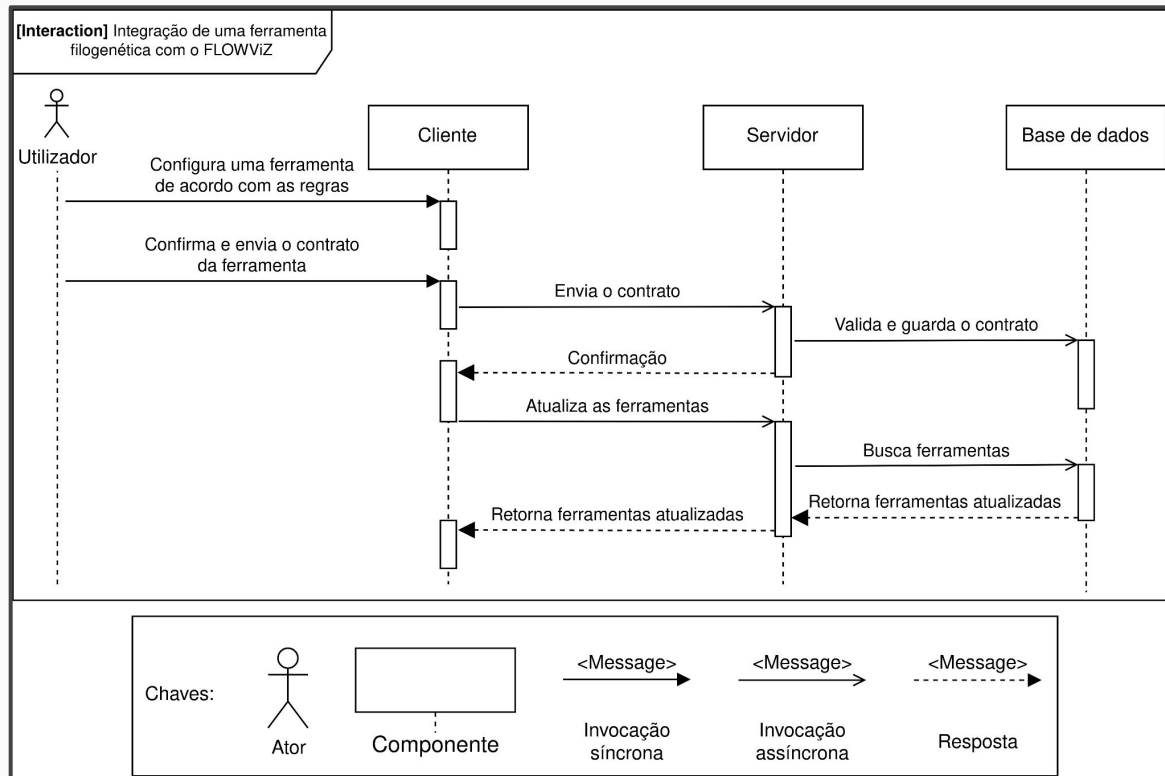


Figura 7 - Diagrama de interação: integrar uma ferramenta filogenética

Demonstração aplicacional - adicionar ferramenta

Add tool

General

Access

Rules

General

Tool name *

Phylolib

Tool description *

PhyloLib is a library.

← PREVIOUS

NEXT →

Figura 8 - Adicionar ferramenta -
informação geral

Demonstração aplicacional - adicionar ferramenta

Add tool

General

Access

Rules

Choose your configuration method

☐ API ☒ Library

Access

Tool address *

localhost

Tool port

☒ Container

Docker image *

luanab/phyloblib

Docker URL *

unix:///var/run/docker.sock

Docker container

phyloblib

Volume source

Volume target

+

/opt/.phyloblibVol /phyloblib

/opt/.phyloblibVol

← PREVIOUS

NEXT →

Figura 9 - Adicionar ferramenta -
informação de acesso à ferramenta

Demonstração aplicacional - adicionar ferramenta

Add tool

General Access Rules

Number of commands groups 2

Name Arguments

Invocation -args

order 0

☐ Allow command repetition

Number of commands 2

Commands

Figura 10.1 - Adicionar ferramenta - Especificação de grupos de comandos

Number of commands 2

Commands

Name distance

Invocation distance

distance

Values kimura

hamming grapetree kimura

Allowed sub-commands

Allowed sub-command sets

Options

Options

Figura 10.2 - Adicionar ferramenta - Especificação de um comando

Demonstração aplicacional - adicionar ferramenta

Phylolib

Type: library

Description: PhyloLib is a library.

Library

Usage

Phylolib [Arguments] [Options]

Arguments

help : <help>
distance : <distance> (hamming,grapetree,kimura) [Options]
correction : <correction> (jukesantor) [Options]
algorithm : <algorithm> (goeburst,edmonds,sl,cl,upgma,upgmc,wpgma,wpgmc,saitounei,studierkepler,unj) [Options]
optimization : <optimization> (lbr) [Options]

Options

File Output : <-o,-out> (file) -> Output file as <format>:<location> with format being (asymmetric|symmetric|newick|nexus)
Dataset Input : <-d,-dataset> (file) -> Input dataset file as <format>:<location> with format being (fasta|ml|snp)
Distance Matrix Input : <-m,-matrix> (file) -> Input distance matrix file as <format>:<location> with format being (asymmetric|symmetric)
Phylogenetic Tree Input : <-m,-matrix> (file) -> Input phylogenetic tree file as <format>:<location> with format being (newick|nexus)
Limit of focus variants : <-l,-lvs> (file) -> Limit of locus variants to consider using goeBURST algorithm [default: 3]

Figura 11 - Documentação da ferramenta adicionada.

Caso de uso - construção e execução de um *workflow*

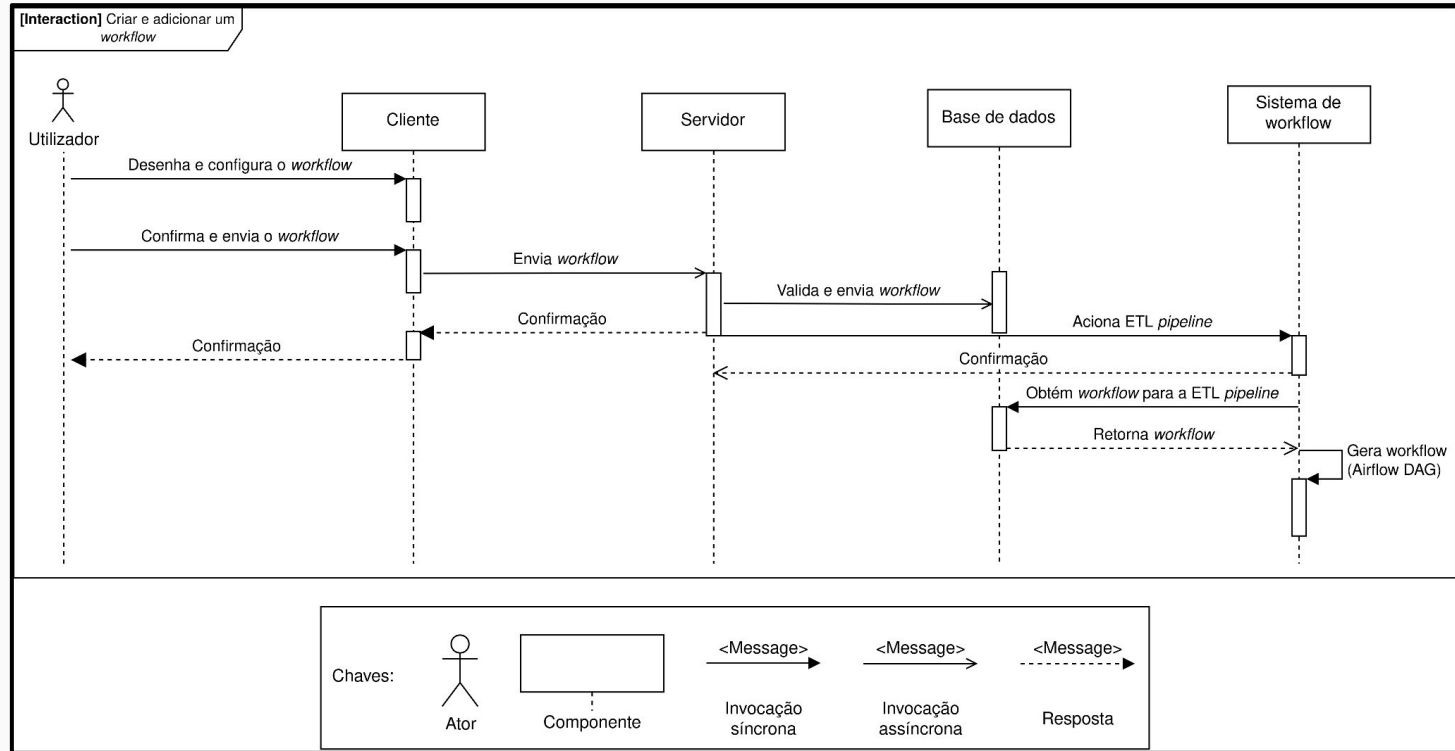


Figura 12 - Diagrama de interação: criar e adicionar um *workflow*

Demonstração aplicacional - construção e execução de um *workflow*

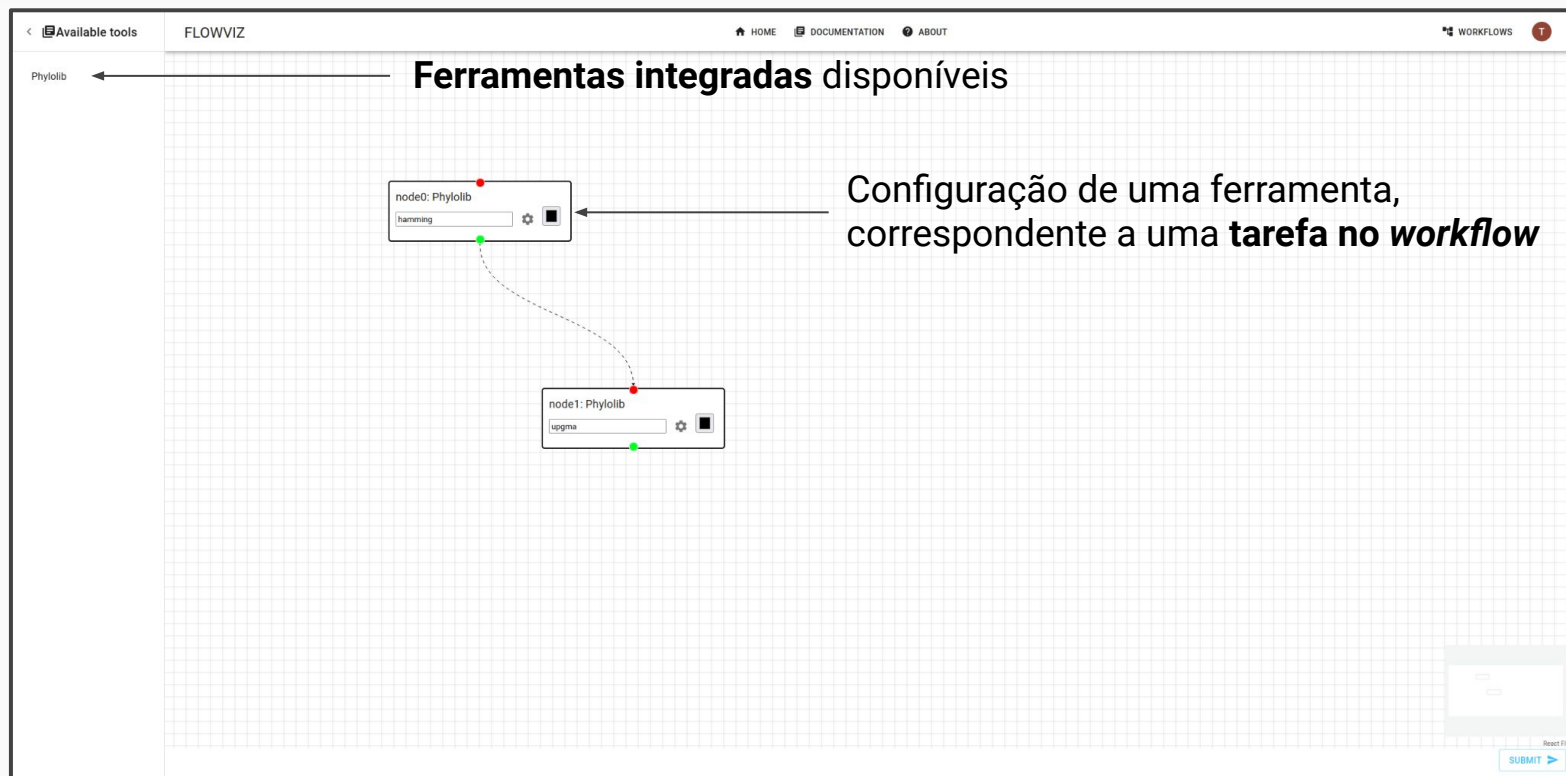


Figura 13 - *Whiteboard* - Desenho e criação de *workflows*

Demonstração aplicacional - construção e execução de um *workflow*

The screenshot shows a 'Task Setup' window for a tool named 'PhyloLib'. The interface is divided into several sections:

- Tool information:** Contains metadata for the tool, including its name, address, Docker image, and container settings.
- I/O variables:** Defines the input and output for the task.
- Command preview:** Shows the command line that will be executed based on the inputs.
- Arguments and Options:** A table-like structure for configuring specific parameters of the command.

Tool information

name: PhyloLib	description: PhyloLib is a library.
address: localhost	dockerUrl: unix://var/run/docker.sock
dockerImage: luanab/phylolib	dockerContainer: phylolib
dockerAutoRemove:	dockerNetworkMode: bridge
dockerApiVersion: auto	

I/O variables

Input	Output
Input key: dataset	Output key: out
Input value: datasets/10.txt	Output value: phylolib/out.txt

Command preview

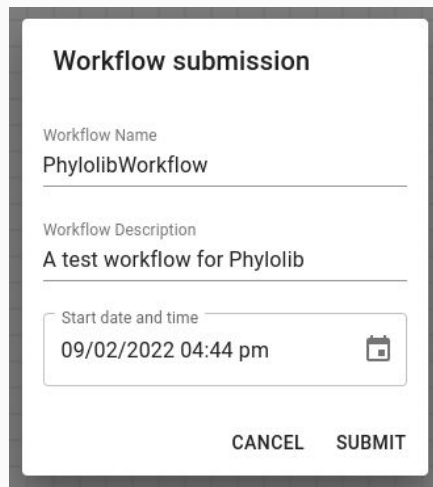
```
phylolib distance grapetree -d dataset -o out
```

Arguments	Options
distance	Dataset Input
grapetree	File Output

ADD COMMAND

CANCEL APPLY

Figura 14 - Configuração individual de uma tarefa do *workflow*



The image shows a web form titled "Workflow submission". It contains three input fields: "Workflow Name" with the value "PhylolibWorkflow", "Workflow Description" with the value "A test workflow for Phylolib", and "Start date and time" with the value "09/02/2022 04:44 pm" and a calendar icon. At the bottom, there are two buttons: "CANCEL" and "SUBMIT".

Workflow submission

Workflow Name
PhylolibWorkflow

Workflow Description
A test workflow for Phylolib

Start date and time
09/02/2022 04:44 pm

CANCEL SUBMIT

Figura 15 - Configuração de informação geral do *workflow* (submissão)

Apache Airflow - Construção dinâmica de DAGs

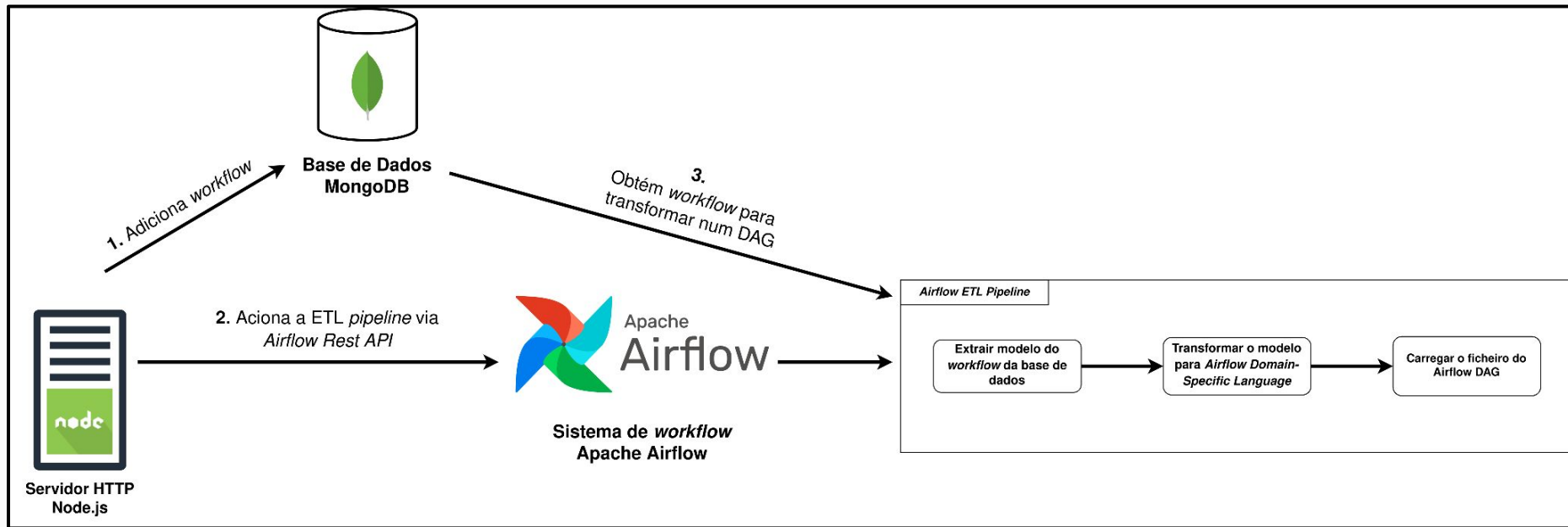


Figura 16 - Fluxo da construção dinâmica de um *Airflow* DAG

Apache Airflow - Construção dinâmica de DAGs

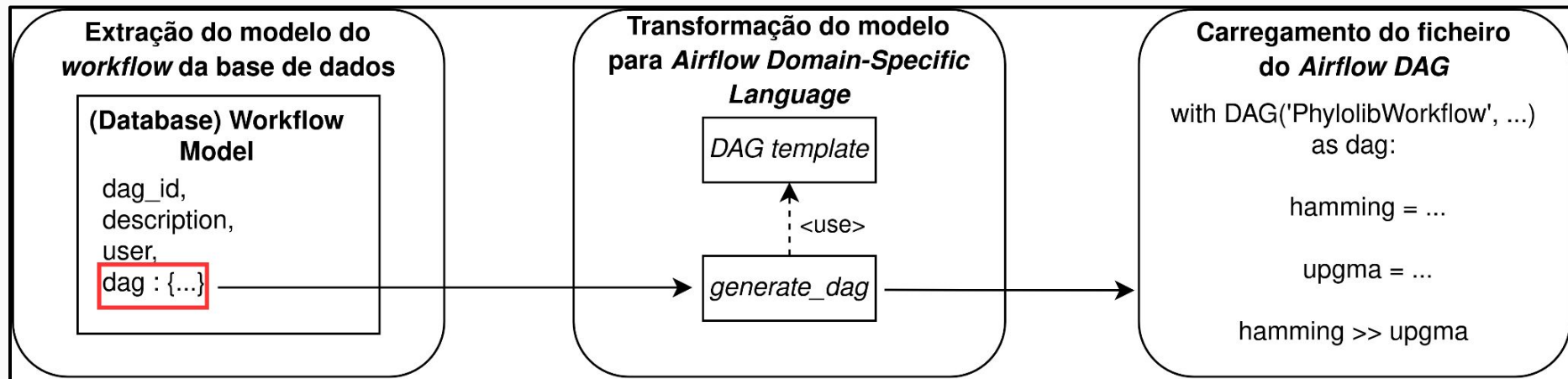


Figura 17 - Detalhe da *ETL pipeline*, desempenhado pelo DAG *dag_generator.py*

Caso de uso - obtenção de resultados

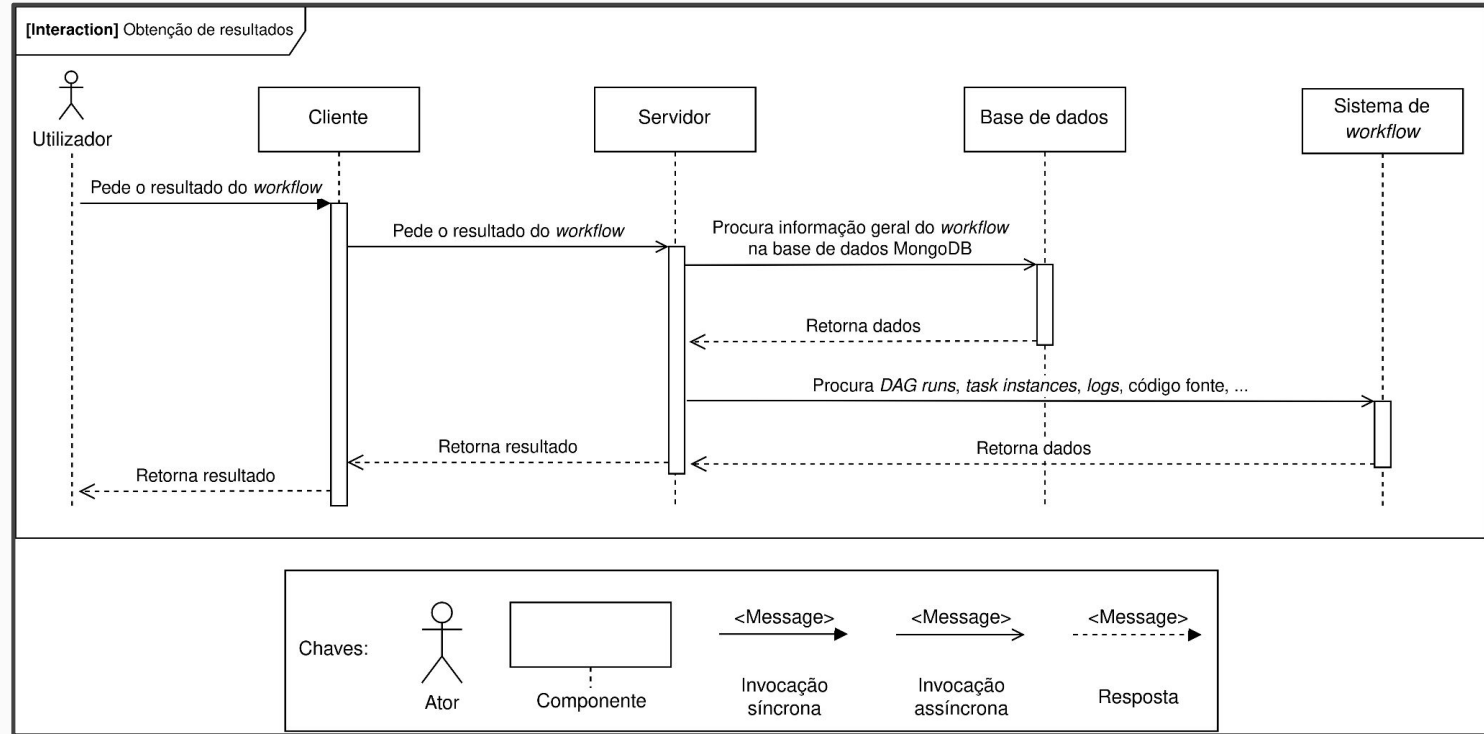


Figura 18 - Diagrama de interação:
Obtenção de resultados

Demonstração aplicacional - Detalhe de um *workflow* do utilizador

PhylolibWorkflow

Description: A test workflow for Phylolib

Number of runs: 1

Run details: manual__2022-09-02T15:48:26.056359+00:00

Run date: 2022-09-02T15:48:26.056359+00:00

Run state: success

Task ID: hamming

Task: hamming

Run date: 2022-09-02T15:48:26.056359+00:00

Run state: success

Log number: 1

```
[('a214ee09c6c6', '**** Reading local file: /opt/airflow/logs/PhylolibWorkflow/hamming/2022-09-02T15:48:26.056359+00:00
/1.log\n[2022-09-02 15:48:26,767] (taskinstance.py:1035) INFO - Dependencies all met for <TaskInstance:
PhylolibWorkflow.hamming manual__2022-09-02T15:48:26.056359+00:00 [queued]>\n[2022-09-02 15:48:26,780]
(taskinstance.py:1035) INFO - Dependencies all met for <TaskInstance: PhylolibWorkflow.hamming
manual__2022-09-02T15:48:26.056359+00:00 [queued]>\n[2022-09-02 15:48:26,780] (taskinstance.py:1241) INFO -
\n[2022-09-02 15:48:26,780]
(taskinstance.py:1242) INFO - Starting attempt 1 of 1\n[2022-09-02 15:48:26,781] (taskinstance.py:1243) INFO -
\n[2022-09-02 15:48:26,789]
(taskinstance.py:1262) INFO - Executing <Task(DockerOperator): hamming> on 2022-09-02
15:48:26.056359+00:00\n[2022-09-02 15:48:26,792] (standard_task_runner.py:52) INFO - Started process 660 to run
task\n[2022-09-02 15:48:26,794] (standard_task_runner.py:76) INFO - Running: [****, 'tasks', 'run',
'PhylolibWorkflow', 'hamming', 'manual__2022-09-02T15:48:26.056359+00:00', '--job-id', '1508', '--raw', '--subdir',
'DAGS_FOLDER/PhylolibWorkflow.py', '--cfg-path', '/tmp/tppgmqcl15', '--error-file', '/tmp/tmpstlul_1']\n[2022-09-02
15:48:26,795] (standard_task_runner.py:77) INFO - Job 1508: Subtask hamming\n[2022-09-02 15:48:26,823]
(logging_mixin.py:109) INFO - Running <TaskInstance: PhylolibWorkflow.hamming manual__2022-09-02T15:48:26.056359+00:00
[running]> on host a214ee09c6c6\n[2022-09-02 15:48:26,862] (taskinstance.py:1429) INFO - Exporting the following env
vars:\nAIRFLOW_CTX_DAG_OWNER=***\nAIRFLOW_CTX_DAG_ID=PhylolibWorkflow\nAIRFLOW_CTX_TASK_ID=hamming
\nAIRFLOW_CTX_EXECUTION_DATE=2022-09-02T15:48:26.056359+00:00
\nAIRFLOW_CTX_DAG_RUN_ID=manual__2022-09-02T15:48:26.056359+00:00\n[2022-09-02 15:48:26,886] (docker.py:258) INFO -
Starting docker container from image luanab/phylolib\n[2022-09-02 15:48:26,891] (docker.py:269) WARNING - Using remote
engine or docker-in-docker and mounting temporary volume from host is not supported. Falling back to
'mount_tmp_dir=False' mode. You can set 'mount_tmp_dir' parameter to False to disable mounting and remove the
warning\n[2022-09-02 15:48:27,923] (docker.py:320) INFO - INFO: Started running command 'distance' with type
'hamming'\n[2022-09-02 15:48:27,936] (docker.py:320) INFO - INFO: Started reading file '/phylolib/data/datasets
/10.txt'\n[2022-09-02 15:48:27,987] (docker.py:320) INFO - WARNING: Ignored invalid profile 'ST'\n[2022-09-02
15:48:27,990] (docker.py:320) INFO - INFO: Finished reading file '/phylolib/data/datasets/10.txt'\n[2022-09-02
15:48:28,003] (docker.py:320) INFO - INFO: Started writing file '/phylolib/out.txt'\n[2022-09-02 15:48:28,010]
(docker.py:320) INFO - INFO: Finished writing file '/phylolib/out.txt'\n[2022-09-02 15:48:28,010] (docker.py:320) INFO
```

Figura 19.1 - Log da execução do *workflow* submetido

Airflow script

```
from airflow import DAG
from datetime import datetime
from airflow.providers.docker.operators.docker import DockerOperator
from docker.types import Mount

default_args = {
    'owner':          'airflow',
    'description':    'A test workflow for Phylolib',
    'start_date':     datetime.today(),
}

with DAG('PhylolibWorkflow', schedule_interval=None, default_args=default_args) as dag:

    hamming = DockerOperator(task_id='hamming',
                             image='luanab/phylolib',
                             api_version='auto',
                             mounts=[Mount(target='/phylolib', source='/opt/.phylolibVol', type='bind')],
                             command='distance hamming --dataset=ml:/phylolib/data/datasets/10.txt --out=symmetric:/phylolib/out.txt',
                             auto_remove='true',
                             docker_url='unix://var/run/docker.sock',
                             network_mode='bridge',)

    upgma = DockerOperator(task_id='upgma',
                           image='luanab/phylolib',
                           api_version='auto',
                           mounts=[Mount(target='/phylolib', source='/opt/.phylolibVol', type='bind')],
                           command='algorithm upgma --out=newick:/phylolib/tree.txt --matrix=symmetric:/phylolib/out.txt',
                           auto_remove='true',
                           docker_url='unix://var/run/docker.sock',
                           network_mode='bridge',)

    hamming >> upgma
```

Figura 19.2 - Código fonte do DAG, dinamicamente gerado no Apache Airflow

Obrigado