WORKFLOW MANAGEMENT SYSTEMS

IN THE BIG DATA ERA

Andrea Bignamini & the IA2 Team
INAF

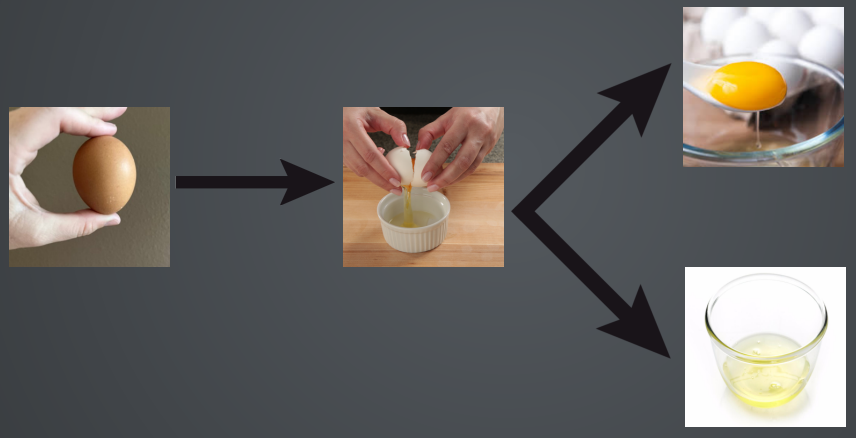
Archives and Data Management Systems in the Big Data Era
Bologna 26-28 February 2025

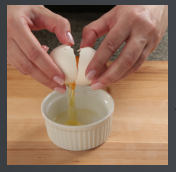
WHAT IS A WORKFLOW MANAGEMENT SYSTEM?

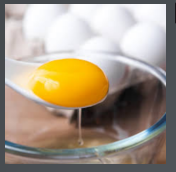
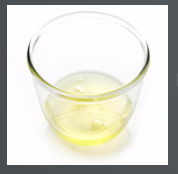
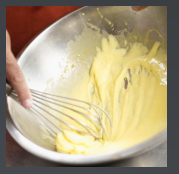
- **Definition**:
  - A *Workflow Management System (WMS)* is a tool that automates, coordinates, and monitors complex processes from start to finish.
- **Key Components**:
  - *Task Automation*: Executes repetitive tasks efficiently.
  - *Dependency Management*: Organizes workflow steps based on logical relationships.
  - *State Monitoring*: Provides visibility into progress and intermediate results.
- **Common Applications**:
  - *Data Analysis Pipelines*: Allows users to create and manage complex data analysis workflows, enabling researchers to process large datasets efficiently.
  - *Reproducibility*: Automates workflows to reduce manual intervention and ensures reproducibility of results by documenting each step in the process.
  - *Access to Computational Resources*: WMS can be configured to run on different computational infrastructures. This means users can leverage powerful computational resources to handle large datasets and complex analyses.
  - *Integration of Software and Tools*: Allows users to integrate numerous tools and software into their workflows. This flexibility enables researchers to select the best tools for their specific analyses.

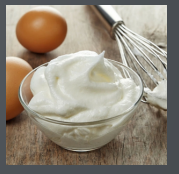
WORKFLOW EXAMPLE 1

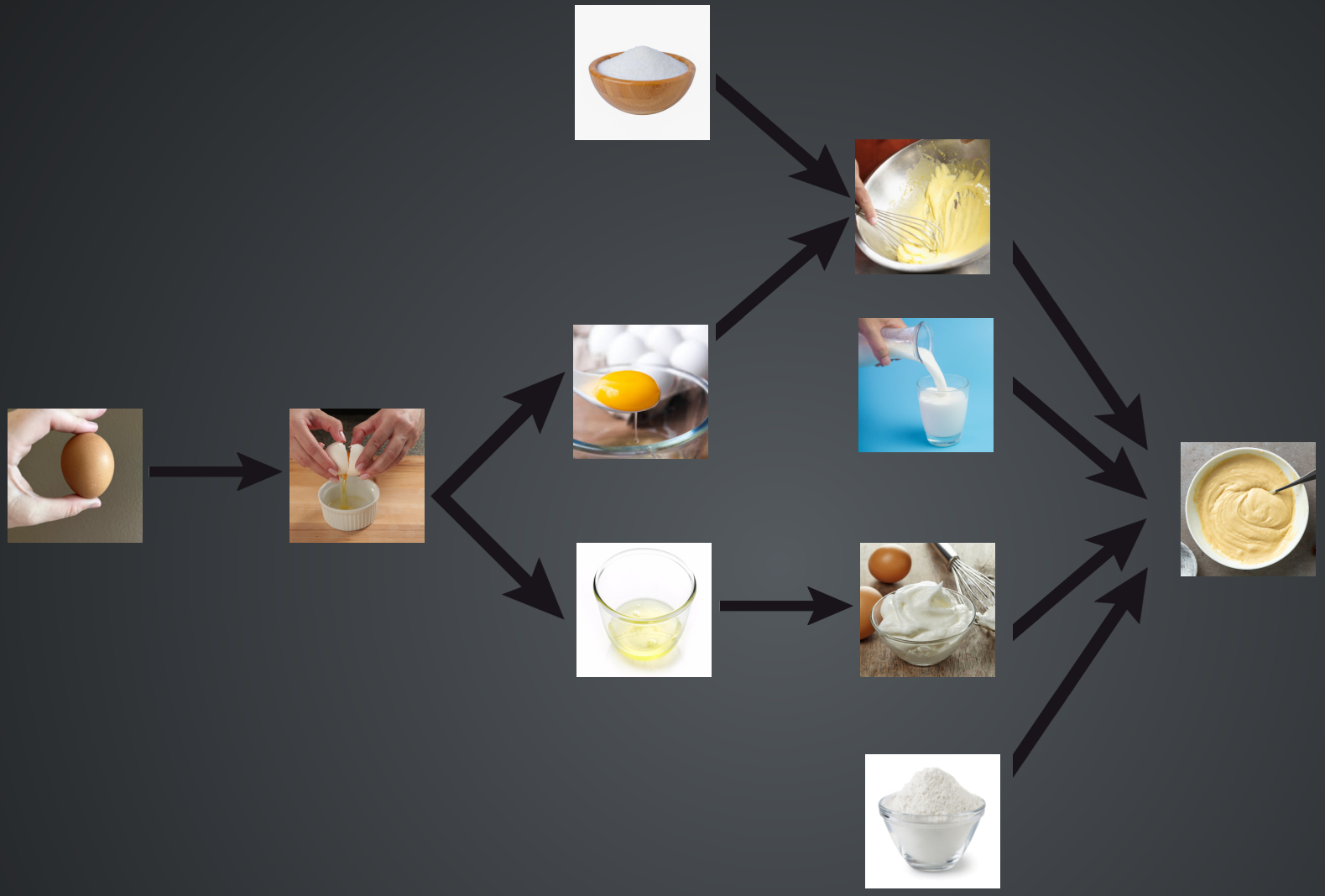
Get an egg



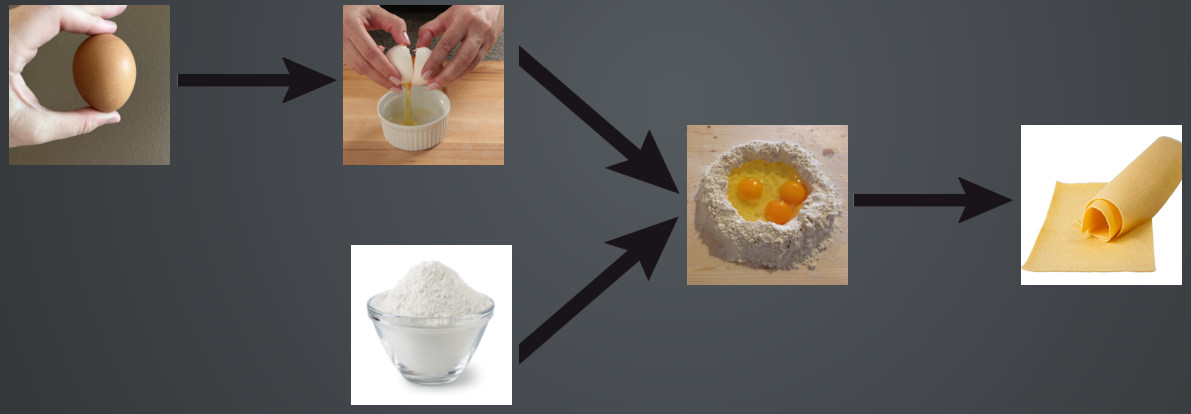


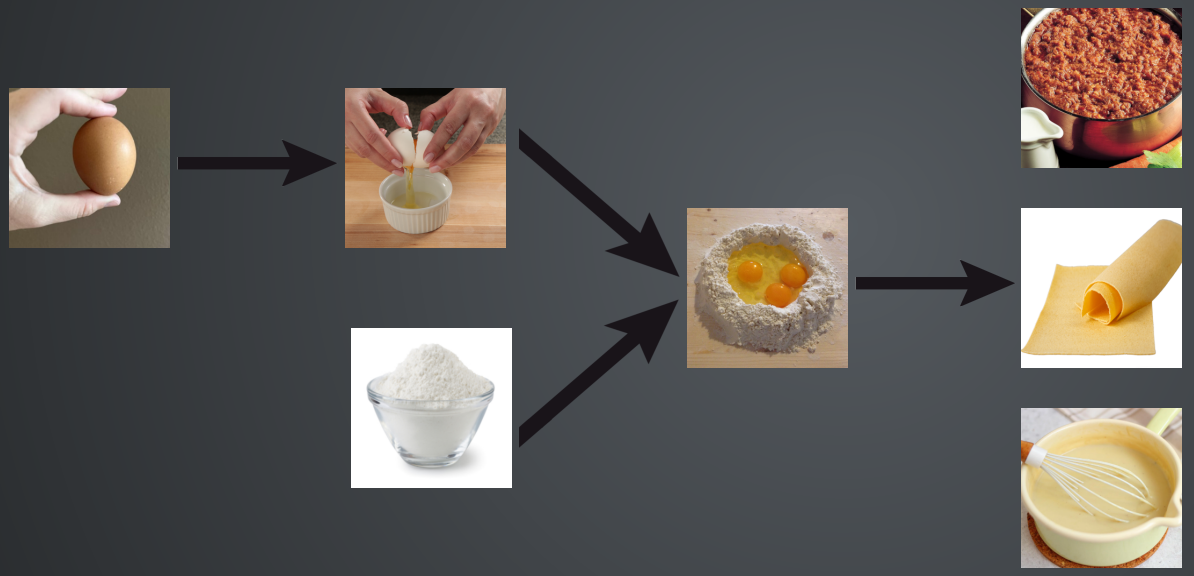



You got pancakes!

WORKFLOW EXAMPLE 2

Get another egg





You got lasagna!

INGREDIENTS OF A WORKFLOW

- A *complex process* is divided into *elementary jobs* (tools).
- Each job receives *inputs* and produces *outputs*.
- Jobs are organized in a *chain* (workflow) where the output of a job is the input of the next job.
- In the workflow execution, the jobs are organized in *dependency*.
- Each tool can be used in the construction of *different workflows*.

INGREDIENTS OF A WMS

- **Platform for workflow design and monitoring**: This platform provides a user-friendly interface for designing the steps of your analysis, and then allows you to monitor the progress of your workflow as it executes.
- **Managing connections**: This is where the magic happens. A WMS acts as a bridge between:
  - **Users**: It allows you to access and manage your workflows, data, and resources.
  - **Data resources**: It helps you manage your data, including storing, accessing, and sharing it securely.
  - **Software**: It provides access to a library of tools and software packages that can be incorporated into your workflows.
  - **Computational resources**: It allows you to utilize computing power, such as servers or clusters, to run your workflows efficiently.

A SUCCESSFUL STORY

SCIENTIFIC MOTIVATION FOR A WMS @ IA2

- **GAPS** (*Global Architecture of Planetary Systems*) is a long-term program for the comprehensive characterization of the architectural properties of planetary systems as a function of the hosts' characteristics (mass, metallicity, environment):
  - more than 80 INAF and associated scientists in Italy, and from foreign institutes
  - more than *20.000 HARPS-N spectra at TNG* since August 2012

- **Request**: customizable data reduction of GAPS private data with appropriate spectral line mask and options
  - HARPS-N reduced data are available through the TNG archive managed by IA2
  - ...but only data reduced with default input parameters
  - ...and HARPS-N DRS (Data Reduction Software) pipeline is not public

A SUCCESSFUL STORY

SCIENTIFIC MOTIVATION FOR A WMS @ IA2

- **GAPS** (*Global Architecture of Planetary Systems*) is a long-term program for the comprehensive characterization of the architectural properties of planetary systems as a function of the hosts' characteristics (mass, metallicity, environment):
  - more than 80 INAF and associated scientists in Italy, and from foreign institutes
  - more than *20.000 HARPS-N spectra at TNG* since August 2012

- **Request**: customizable data reduction of GAPS private data with appropriate spectral line mask and options
  - HARPS-N reduced data are available through the TNG archive managed by IA2
  - ...but only data reduced with default input parameters
  - ...and HARPS-N DRS (Data Reduction Software) pipeline is not public

A SUCCESSFUL STORY

SCIENTIFIC MOTIVATION FOR A WMS @ IA2

- **GAPS** (*Global Architecture of Planetary Systems*) is a long-term program for the comprehensive characterization of the architectural properties of planetary systems as a function of the hosts' characteristics (mass, metallicity, environment):
  - more than 80 INAF and associated scientists in Italy, and from foreign institutes
  - more than *20.000 HARPS-N spectra at TNG* since August 2012

- **Request**: customizable data reduction of GAPS private data with appropriate spectral line mask and options
  - HARPS-N reduced data are available through the TNG archive managed by IA2
  - ...but only data reduced with default input parameters
  - ...and HARPS-N DRS (Data Reduction Software) pipeline is not public

A SUCCESSFUL STORY

SCIENTIFIC MOTIVATION FOR A WMS @ IA2

- **GAPS** (*Global Architecture of Planetary Systems*) is a long-term program for the comprehensive characterization of the architectural properties of planetary systems as a function of the hosts' characteristics (mass, metallicity, environment):
  - more than 80 INAF and associated scientists in Italy, and from foreign institutes
  - more than *20.000 HARPS-N spectra at TNG* since August 2012

- **Request**: customizable data reduction of GAPS private data with appropriate spectral line mask and options
  - HARPS-N reduced data are available through the TNG archive managed by IA2
  - ...but only data reduced with default input parameters
  - ...and HARPS-N DRS (Data Reduction Software) pipeline is not public

- **Solution**: deploy a WMS at IA2 to manage:
  - access to GAPS private data
  - access to HARPS-N DRS usage without distributing the code

YABI

- **Yabi** is a 3-tier application stack to provide users with an intuitive, easy to use, abstraction of compute and data environments. Developed at the Centre for Comparative Genomics and Murdoch University, Yabi has been deployed across a diverse set of scientific disciplines and high performance computing environments
- For **Yabi deployed at IA2** we need to
  - *Divide in steps* the HARPS-N DRS pipeline
  - *Create tools* in Yabi to run these steps
    - Providing entry points to set input custom parameters
  - *Manage user access* to tools and data through fine-grained authorization levels (backends, credentials, toolsets)
- In the Yabi interface directories users can find:
  - Proprietary reduced data subdivided by observation nights and targets
  - Standard tools and masks and any proprietary tools and masks (i.e. tools and masks developed by a user are only available to that user)

YABI STATISTICS

UPDATED TO 2025-02-25

- **Yabi** *has evolved over the years*
- **Yabi** *access extended to all HARPS-N users*
  - 375 total users
  - 157 active users
- **New tools and masks** *developed also by users*
  - 14 tools, 8 different types of workflows
  - 40 custom masks developed by users
  - 14019 workflows executed since March 2014

3 LEVELS OF AUTHORIZATIONS IN YABI

1. **User access** to *data*

    - Input data (i.e. proprietary raw data from archive)
    - Output data

2. **User access** to *computational resources*

3. **User access** to *software* (tools)

    - Open software
    - Licensed software
    - Software developed by users

AUTHORIZATIONS FOR DATA AND COMPUTATIONAL RESOURCES

BACKENDS, CREDENTIALS AND BACKEND CREDENTIALS

- The Yabi **Backend** is a demon that provides execution and file services to the Yabi stack
  - It abstracts away the details and complexity of individual protocols and resources
  - 3 types of Backend: *execution*, *storage* and *null* (for fileselector tools)
  - Several schema and connectors: localex, localfs, PBSPro, SGE, Torque, SSH, SFTP, Slurm, Amazon S3, OpenStack Swift
- Each Yabi **Credential** belongs to a Yabi user to allow that user to have access to Backends (i.e. to execute workflows or to access data)
  - In the Credential table all the fields will be needed, depending on the credential type: e.g. ssh key, certificate or user-passwd
- The Yabi **Backend Credential** is a linking table between a Credential and a Backend
  - To allow a *user* to use a *Backend*, you need to connect the *user Credential* to the *Backend* through a *Backend Credential*
  - A Backend Credential may also define additional rules (e.g. define user's Default Stageout as a single directory where all the user's results will be staged out to)

AUTHORIZATION FOR SOFTWARE

TOOLSETS AND TOOLGROUPS

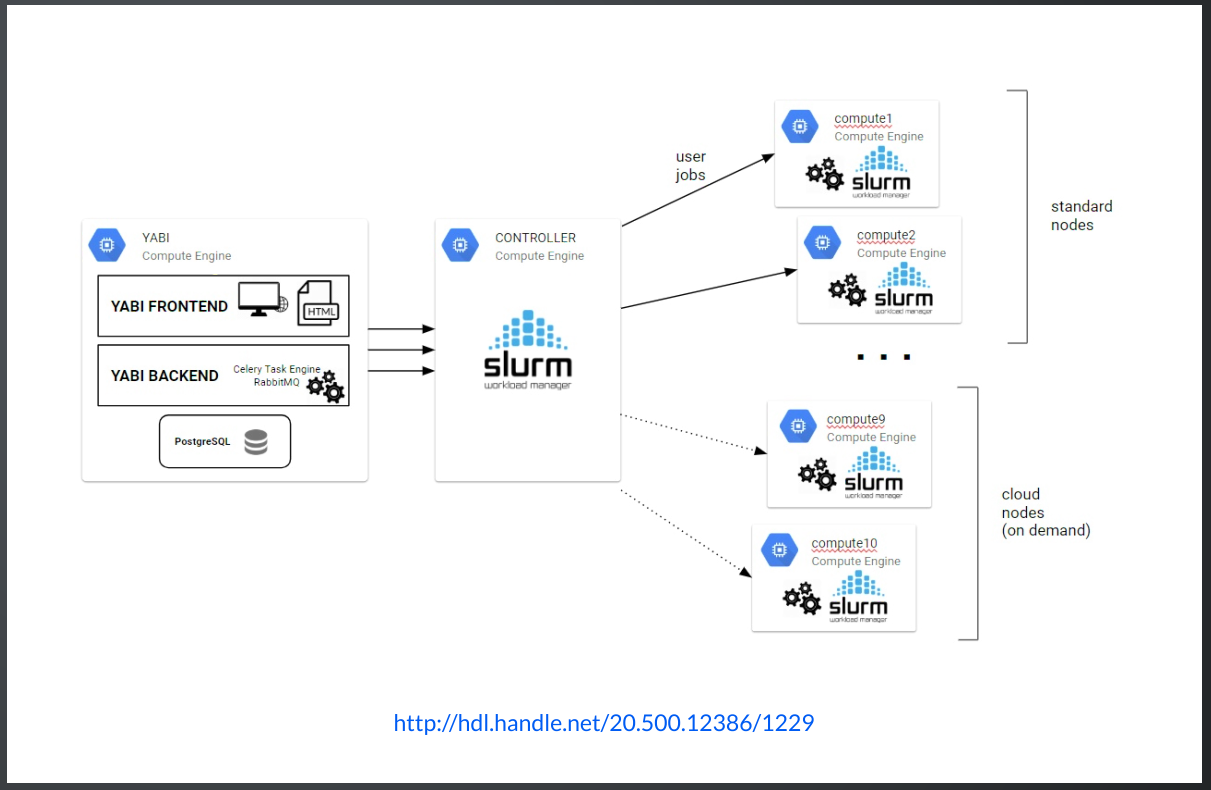
- **Tools** are the elements the user builds the workflow. Each *tool collects only information about itself*:
  - A description of the tool, and the parameters it accepts and it returns
  - Information about how to run the tool on a backend and the file system backend where outputs are saved
- **Toolsets** are *groups of users* that determine which tools they have access to
  - *All users in a toolset share the same privileges* to access tools
  - Any user can belong to more than one toolset
- **Toolgroups** determine:
  - How tools are grouped in the user interface
  - Which users can use which tools
    - *Each tool in a toolgroup is assigned to a toolset*, i.e. to a group of users that have access to that tool

MAIN BENEFITS OF WMS

- **Zero Code Workflow Design**: The final user does not have the bother of software installation or hardware configuration, but he can just focus on scientific analysis
  - *Reduce errors* on recurrent and redundant manual tasks
  - *Reproducibility* of results
- **Remote Data**: No need to retrieve locally huge amount of data from remote archive
- **Remote Computational Resources**: Exploit larger computational resources
  - Export the workflow (e.g. json, YAML) and move it to another data centre to be run
- **Single Software Version**: All users of a collaboration agree on a single software version
  - Remember! ICT GitLab is availabe to all INAF users https://www.ict.inaf.it/gitlab
  - News! There is a fancy Git/GitLab course available here http://gitlab-school.pages.ict.inaf.it/howto-gitlab
- **Private Software**: Hide proprietary software behind curtains of A&A layers

YABI POC ON GOOGLE CLOUD PLATFORM

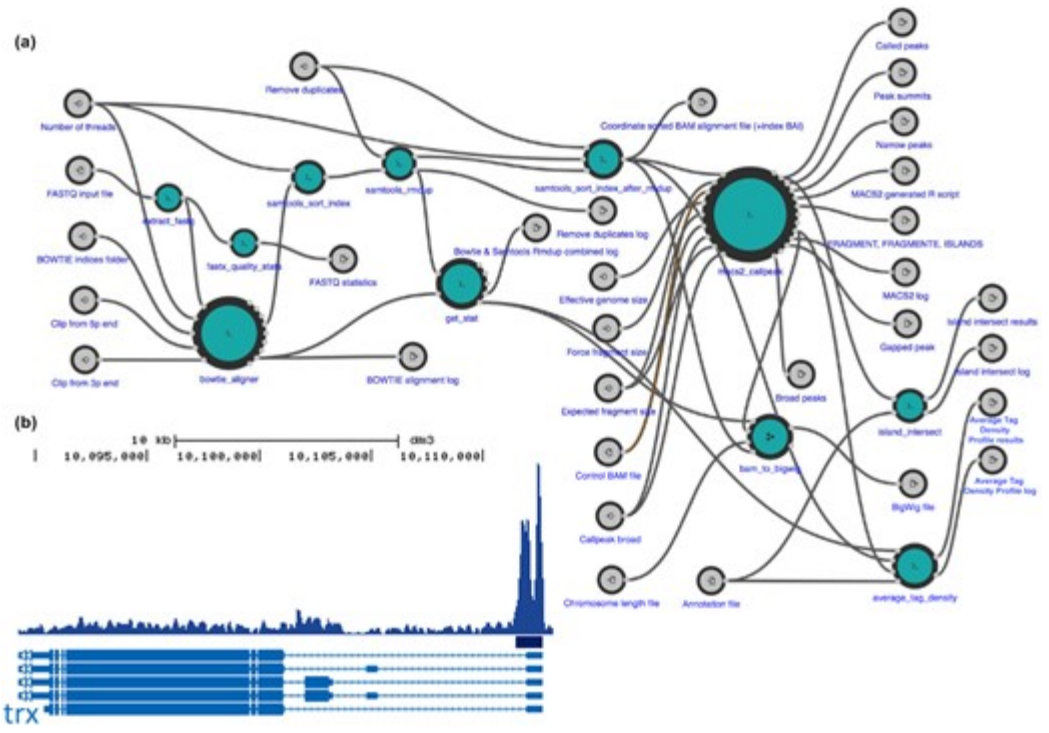
- Proof of Concept “Yabi Workflow Execution on Google Cloud Platform” to run data reduction pipelines on TNG archive data
  - **Goals**: simplify infrastructure management (SaaS/PaaS) and software deployment (Docker), optimizing and balancing the scalability of the service (Slurm and Kubernetes)
  - **Results**: excellent scalability and good costs (total estimated charges of 200 EUR/month to maintain architecture up and running on GCP)
  - **Criticalities**: needs to write data reduction pipeline optimized for containerization



http://hdl.handle.net/20.500.12386/1229

HOW TO GET THE MOST OUT OF YOUR WORKFLOWS IN THE CLOUD

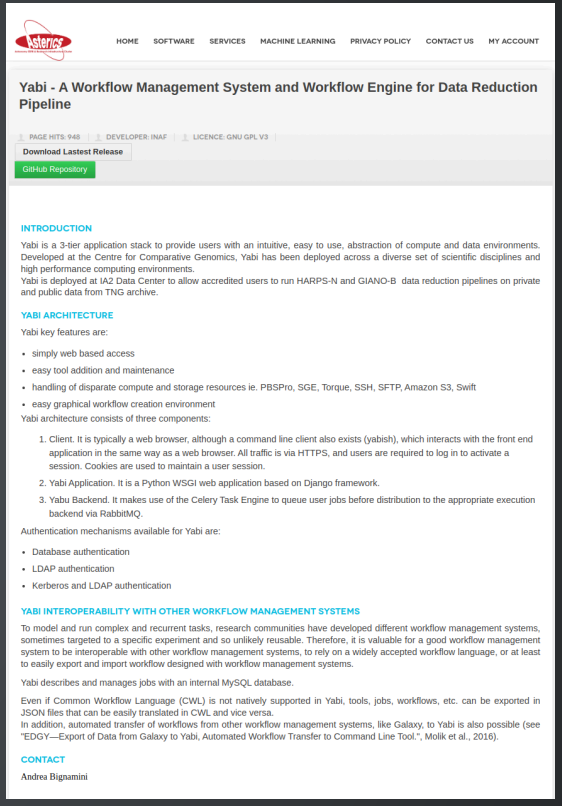
- A **monolithic pipeline** can exploit only scalability at workflow level
- A pipeline splitted into **atomic tools** can exploit scalability at job level
- It needs a **paradigm shift** in pipeline writing (e.g. microservices, microcontainer)



Rabix representation of a workflow (Kotliar et al. 2019) DOI:10.1093/gigascience/giz084

WMS AND FAIRNESS

- Several WMSs are available, targeted to specific experiments or scientific communities (Taverna, Kepler, Galaxy, Pegasus, etc.)
- WMSs may have a low level of interoperability, implying difficulties in terms of reusability and reproducibility of scientific results
- Within **ASTERICS-H2020** Project we worked on a prototype CWL (Common Workflow Language) integration in Yabi
  - CWL is a standard, it can provide an high level of interoperability between WMS and portability across different hardware environments
  - CWL supports natively Docker, which makes it appealing in the cloud paradigm
  - CWL is excellent for jobs that must be run periodically



https://repository.asterics2020.eu/content/yabi-workflow-management-system-and-workflow-engine-data-reduction-pipeline

WMS IN THE BIG DATA ERA

WMS may play a crucial role in addressing the challenges of the Big Data Era: *Volume*, *Velocity*, and *Variety*.

- **Scalability**: WMS handles large-scale data using HPC and cloud infrastructures, dynamically allocating resources as needed.
- **Distributed Processing**: It supports parallel execution of tasks across multiple machines, enabling efficient processing of big data pipelines.
- **Data Integration**: WMS facilitates the integration of diverse and heterogeneous datasets from various sources, ensuring seamless analysis.
- **Real-Time Processing**: It manages time-sensitive workflows, enabling near-real-time analysis for streaming data applications.
- **Adaptability**: WMS can adapt to dynamic workflows where task sequences may change based on intermediate results or external conditions.
- **FAIR Principles**: WMSs should adopt interoperable standard to describe their workflows (i.e. CWL) and they should ensure that workflows adhere to the FAIR principles.

CONCLUSIONS AND DISCUSSION TOPICS

- **Workflow Management Systems are great** ...
  - Reproducibility of results
  - Zero Code Workflow Design
  - Reduce errors
  - Great potential for use in the Cloud
  - Export workflow descriptions and move them to another data centre
- ... *but they can go bad*
  - Need to to write data reduction pipeline optimized for containerization and WMS
  - Need to improve interoperability and standardize workflow description (CWL)
- **Keep data close to computational resources**
  - How to allow access to remote data resources?
  - How to run workflows on data stored in geographically distributed sites?
- How to provide user access to perform **custom data reduction on public data**?
- Which **data policy** applies to output data of custom reduction on public data?
  - Which data quality for data reduced by WMS? Who provides and guarantees it?