PIPELET
SCIENTIFIC DATA PROCESSING FRAMEWORK .

MAUDE LE JEUNE - LEJEUNE@APC.UNIV-PARIS7.FR
MARC BETOULE - BETOULE@LPNHE.IN2P3.FR

HTTP://GITORIOUS.ORG/PIPELET

CONTEXT

Scientific data analysis typically faces two challenges:

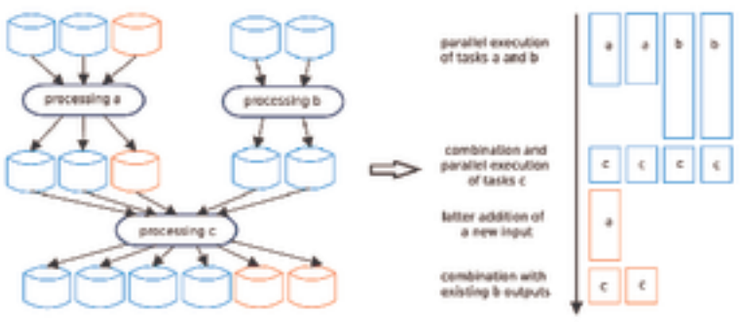
- Large data sets and/or large processing CPU time
- Complex processing with multiple interdependent steps and parameters to tweak

New data sets tend to become not only extremely large but also complex. As a consequence, data analysts have to deal with extended computation **and development** costs.
The *Pipelet* framework helps to reduce both costs. It enables developers to:

1. write and manipulate pipelines with **complex dependency scheme**;
2. **keep track** of the different **versions** of processing and parameters; provide easy data access;
3. carry the same source code **from development on laptop to production on clusters**.

1. PIPELINE SCHEME

The **pipeline scheme** reflects how a complex data processing can be cut into elementary **segments**. *Pipelet* handles any **directed acyclic graph**, expressed in the `graphviz dot` language, which makes the dependency relations between segments easy to define and display.



A segment is a part of the processing that follows the **Single Processing-Multiple Data** paradigm. *Pipelet* exploits this property to:

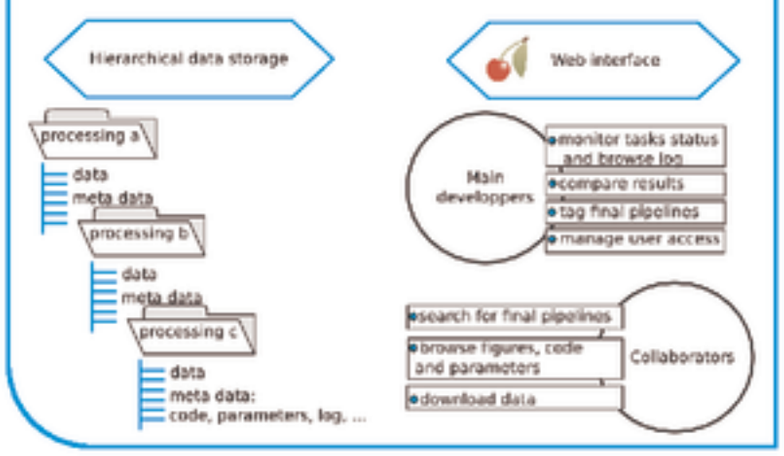
- provide automated parallelization of the processing,
- fully handle further completion of the processed dataset with minimal recomputation.

2. PERMANENCE

Each segment has a **unique identifier** computed from the processing code and parameters. This guaranties the **traceability** of the processing. Any change in the code, or parameter values triggers the automatic recomputation of the segment outputs and its downstream dependencies.

The comparison and sharing of the results are eased by a **web interface** which offers:

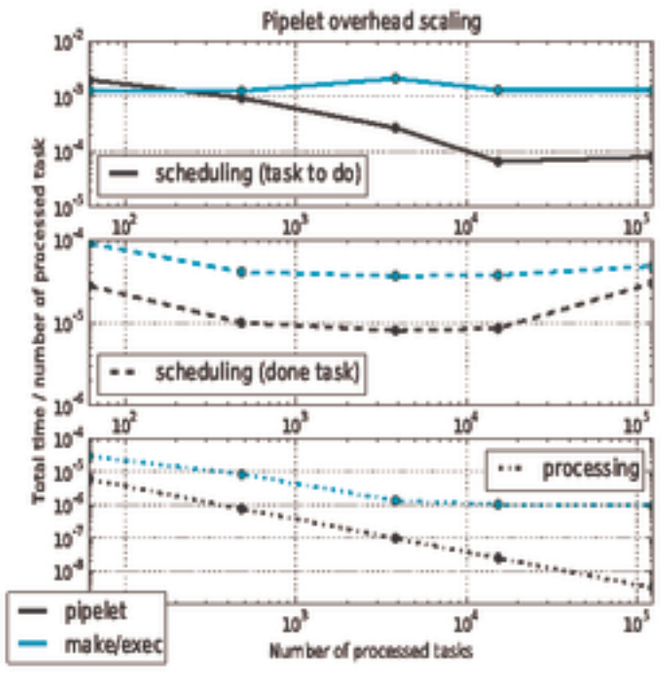
- access to data and figures via the web;
- tagging and search facilities;
- monitoring of the processing (running, failed and successful task), and access to the log.



3. SCALABILITY

Design: *Pipelet* follows the **scheduler** design pattern. It converts a whole pipeline into unit tasks with their sequence of execution. This task queue is then emptied by a pool of workers. **Interactive** workers are provided to ease fine-tuning and debug, while **batch** workers are able to take advantages of SMP and cluster machines. This design allows a totally costless transition between development and production phases of a study.

Performances: *Pipelet* relies on a **sqlite** database to store tasks status and dependencies. This leads to remarkable performances when compared to other pipelining tools. The plot below shows the scheduling speed as a function of the number of tasks for *Pipelet* (in black) and `GNU make` (in cyan) on a non trivial dependency scheme. On the processing side, the overhead is negligible compared to any non-trivial tasks (always lower than the unix `exec` call).



COSMIC MICROWAVE BACKGROUND

Pipelet is currently used in two CMB ongoing experiments: **Planck**, an ESA satellite mission currently observing the microwave sky with an unprecedented precision, and **Polarbear**, a ground-based observatory dedicated to B-mode polarization of this emission.

The latter will make use of Pipelet for its daily light analysis performed at the **NERSC computing center** in Berkeley, California. Pipelet web interface has been set up on the science gateways which have been made available to access HPC computers and storage systems of the center.

LARGE IMAGING SURVEYS

Pipelet powers the analysis pipeline of the calibration dataset for the Supernovae Legacy Survey (**SNLS**). The SNLS is a large photometric survey conducted during 5 years with MegaCam, the 400 megapixels wide-field imager at Canada-France-Hawaii Telescope. The calibration dataset alone represents rougly **1TB** of **raw** data.

Following this sucessfull experience, *Pipelet* is now considered for the offline analysis of the upcoming **SkyMapper** supernovae survey, the south analog of the Sloan Digital Sky Survey, sited at Siding Spring Observatory, in Australia.

Project wiki and `Git` URLs at : **HTTP://GITORIOUS.ORG/PIPELET**