A History of Distributed LOFAR Workflows

Alexandar Mechev, Leiden University

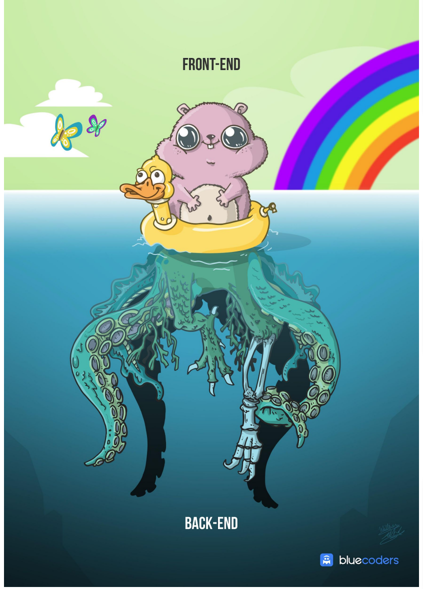
AENEAS Oct 2018

Overview

- Challenges
- Implementations
- Successes
- Lessons

State of LOFAR

- Can’t mass-process at University
- Multiple Science cases
- Multiple Archive locations
- Evolving Software
- Complex (but parallel) pipelines



Pipelines

- Can be parallelized
- Distributed
- Single run vs Automated
- Not versioned
- Fast moving

HTC->HPC

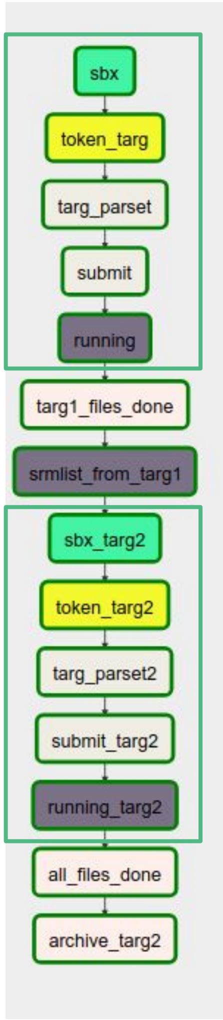
- LTA Locations (HTC):
    - Data transfer
    - Parallelization -> Speed
    - Optimize for Science Cases
- Track progress remotely
- Imaging on HPC

Implementation

1. Run jobs on Amsterdam GRID cluster
    a. Job DB ⇔ Run anywhere
2. Scripts vs LOFAR S/W
3. Submitting
        4. Intermediate Data (proxy required)
5. Workflow Orchestration

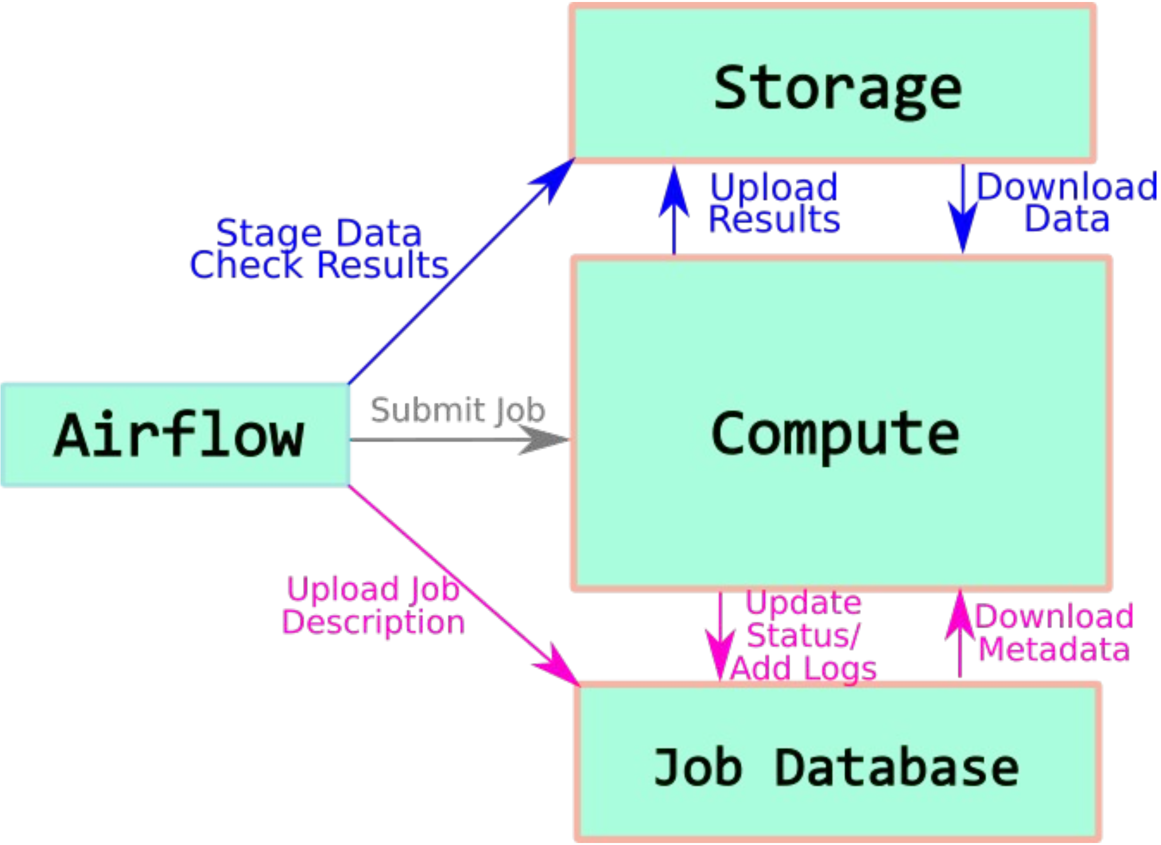
Orchestration Details

1. Apache Airflow
   a. Custom Operators/Sensors
   b. Running on login node
   c. Integrated with middleware
2. Abstract Orchestration, Processing
3. Use git to track pipelines (≈versioning)
   a. Reproducible



DI Target flag/ Averaging

DI Target Calibration



Workflow Orchestration

Successes @ Amsterdam

1. 500+ Datasets @ 2/day
2. Well integrated with LTA
3. High Throughput (~4h/obs)
4. Storage woes
5. Software versioning

Successes @ Juelich

1. 200+ Datasets
2. Local implementation
3. Integrated with workflow
4. Processing woes
5. LTA woes

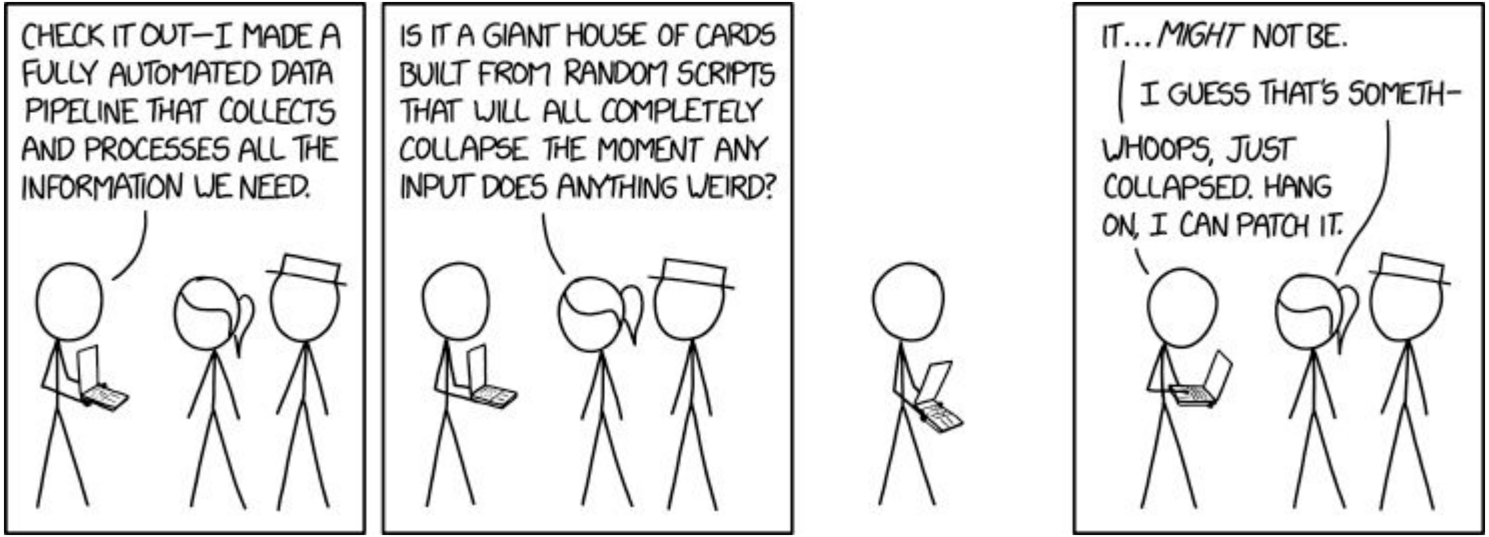
Lessons so far

1. Need High Throughput Computing
2. Need Workflow Orchestration
3. Mapping Credentials non-trivial
   a. Needed for storage access
4. Integration tests (!!)
5. Communicate between scientists

Future

1. Create front-end service for LOFAR
   a. REST
2. Make testing/iterating of pipelines easy!
3. Resolve credentials
4. Offer(LOFAR) as a service
   a. Parameters, auth, rate-limit

Thanks!


CHECK IT OUT—I MADE A FULLY AUTOMATED DATA PIPELINE THAT COLLECTS AND PROCESSES ALL THE INFORMATION WE NEED.
IS IT A GIANT HOUSE OF CARDS BUILT FROM RANDOM SCRIPTS THAT WILL ALL COMPLETELY COLLAPSE THE MOMENT ANY INPUT DOES ANYTHING WEIRD?
IT... MIGHT NOT BE.
I GUESS THAT'S SOMETH-
WHOOPS, JUST COLLAPSED. HANG ON, I CAN PATCH IT.

“

> *The most amazing achievement of the software industry is its continuing cancellation of the steady and staggering gains made by the hardware industry.*