PESC
Programa de Engenharia de Sistemas e Computação

COPPE UFRJ
Instituto Alberto Luiz Coimbra de Pós-Graduação e Pesquisa de Engenharia

Using Provenance Analyzers to Improve the Performance of Scientific Workflows in Cloud Environments

João Carlos Gonçalves, Daniel de Oliveira, Kary Ocaña,
Eduardo Ogasawara, Jonas Dias, Marta Mattoso

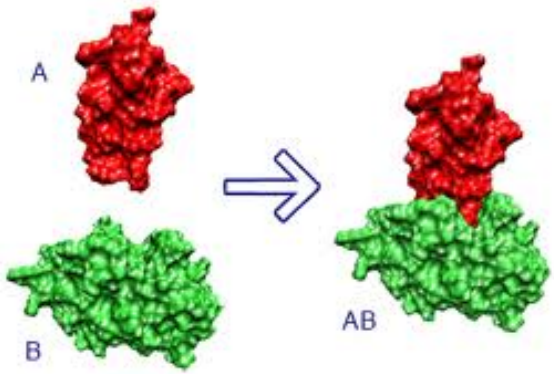
SBBD2012
SIMPÓSIO BRASILEIRO DE BANCOS DE DADOS

Large Scale Scientific Experiments

- Process a large amount of data
- Modeled as scientific workflows
- Assisted by scientific workflow management systems (SWfMS)
- Must gather provenance data

Exploratory Workflows

- Scientists have to explore the behavior of their model under different inputs
  - ✓ This occurs in many areas such as bioinformatics, computational fluid dynamics, uncertainty quantification, dark energy analysis



- These data-centric workflows are computationally intensive and they may run for hours/days
- Demand distributed and parallel processing

Challenges in Exploratory Workflows

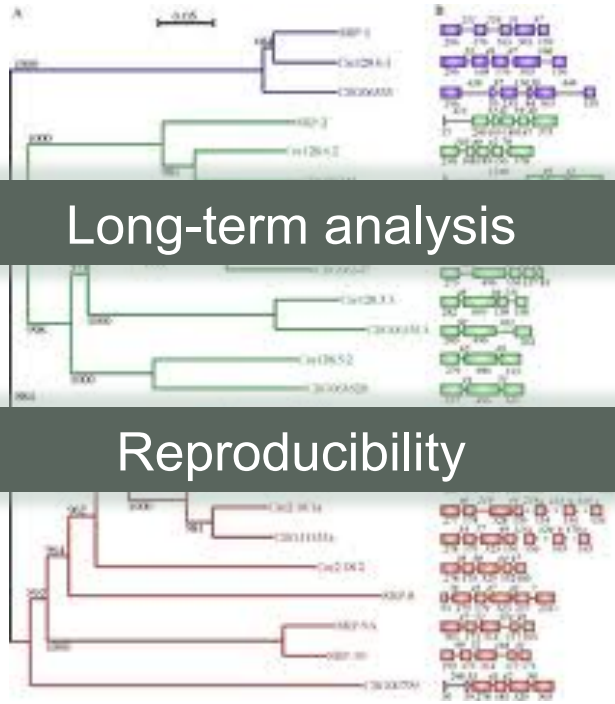
- Several executions of the workflow produce unnecessary data
  - ✓ Does not comply with a specific quality criteria
  - ✓ Results are inconsistent
  - ✓ Executions did not produce data files

- If this data is not propagated for subsequent activities of the workflow, can we improve performance?

Research Question

- Problem
  - ✓ Given a scientific workflow that typically handles a vast amount of data, how to improve its performance
- Solution
  - ✓ Development of an approach to analyze data quality during the workflow execution and enable the filtering of data based on provenance and domain-specific data.

Expanding provenance to improve performance



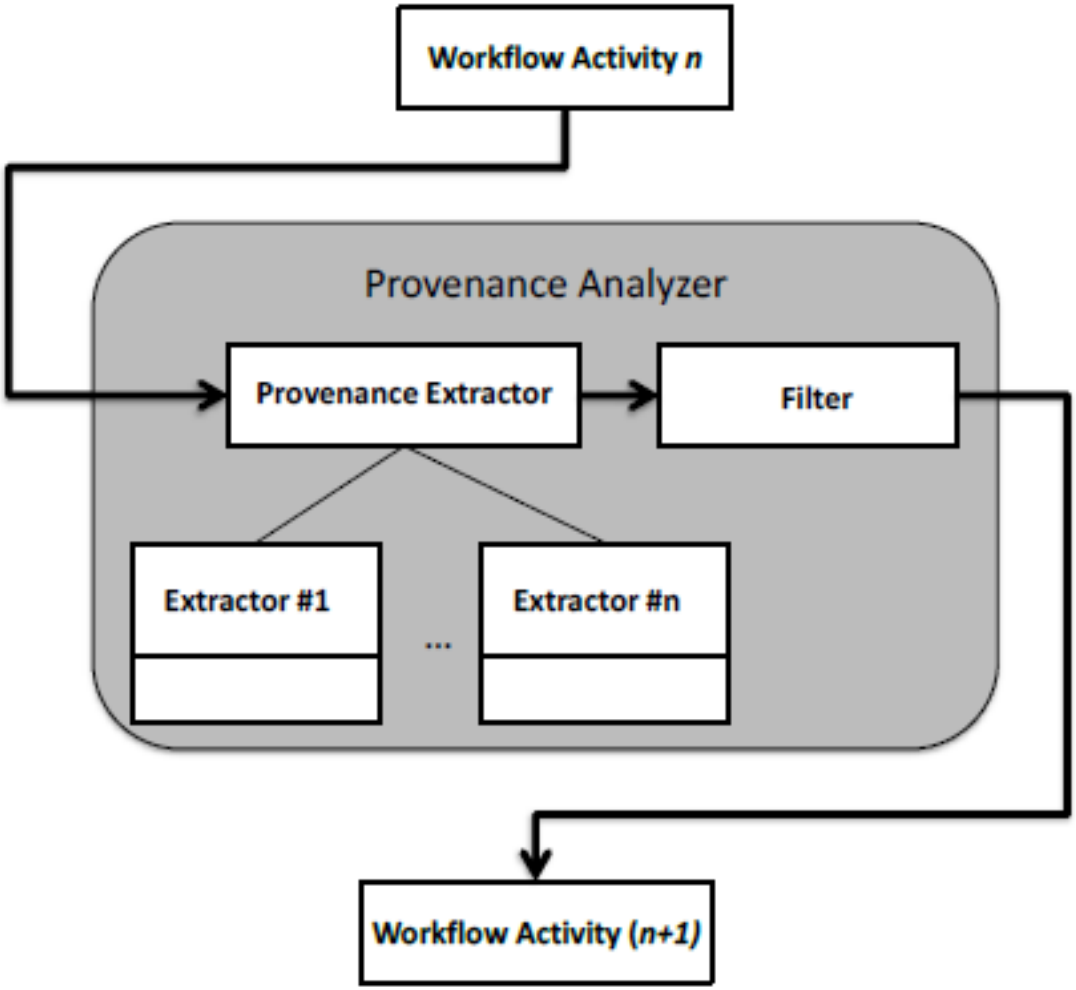
Long-term analysis

Reproducibility

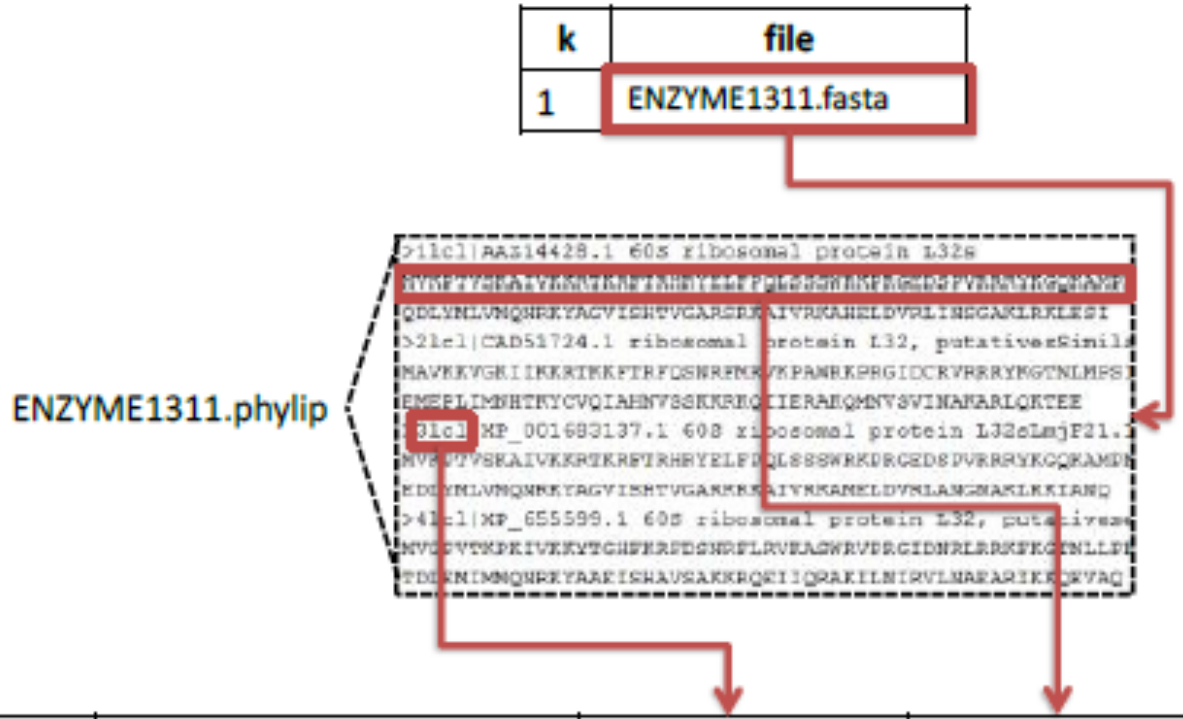
Run-time analysis

Data quality

Provenance Analyzers



Provenance Analyzers



k	file	fasta_origin_ID	fasta_seq	is_valid
1	ENZYME1311.fasta	3lcl	MVKPTVSKAIVKK RTKFTRHRYELFFP QLSSSWRKPRGED SPVRRRYKGQKAM	TRUE

Final Remarks

- Performance results show benefits of up to 36% when using PA when filtering is about 23% of the produced data.
  - ✓ The approach show the potential of analyzing provenance data during execution time to improve the performance
- Data quality criteria are generic enough to represent the scientists' goals during the experiment.

PESC
Programa de Engenharia de Sistemas e Computação

COPPE UFRJ
Instituto Alberto Luiz Coimbra de Pós-Graduação e Pesquisa de Engenharia

Using Provenance Analyzers to Improve the Performance of Scientific Workflows in Cloud Environments

João Carlos Gonçalves, Daniel de Oliveira, Kary Ocaña,
Eduardo Ogasawara, Jonas Dias, Marta Mattoso

SBBD2012
SIMPÓSIO BRASILEIRO DE BANCOS DE DADOS