Parallel Workflow Support for StarVZ using Drake

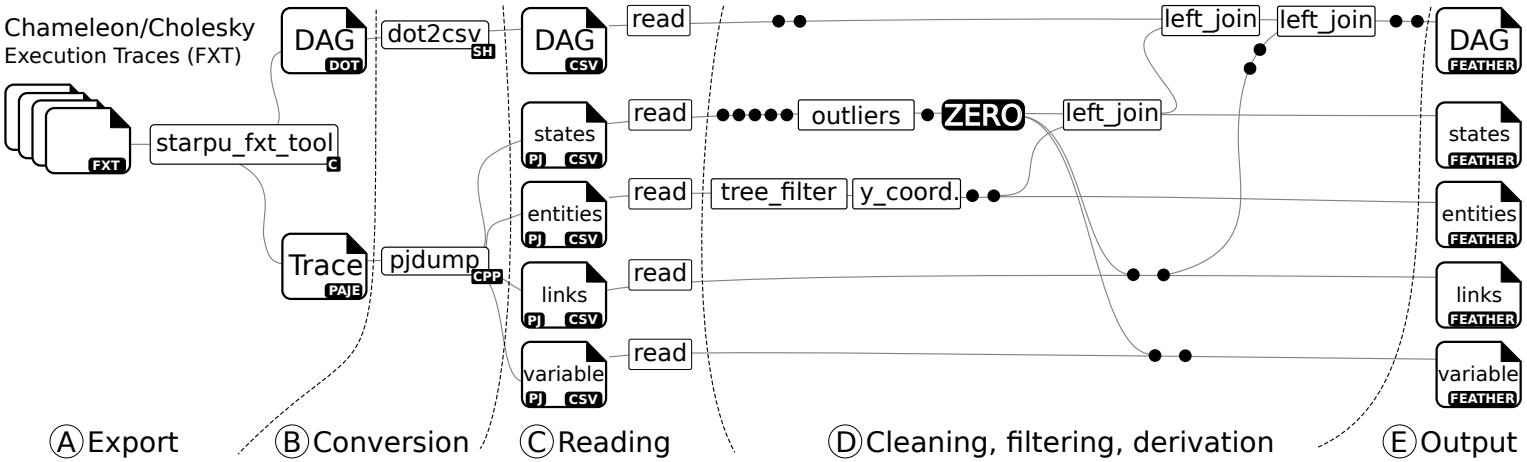
Guilherme Rezende Alles, Lucas Mello Schnorr
{gralles, schnorr}@inf.ufrgs.br

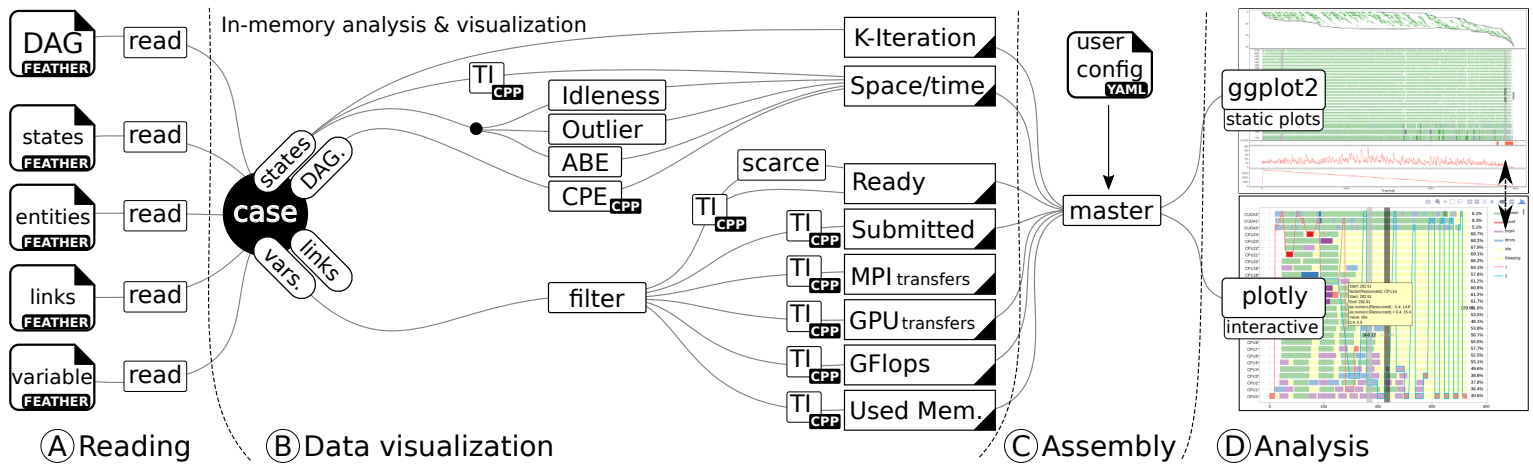
WSPPD, Instituto de Informática
Universidade Federal do Rio Grande do Sul

Porto Alegre, September 5th, 2018

Outline

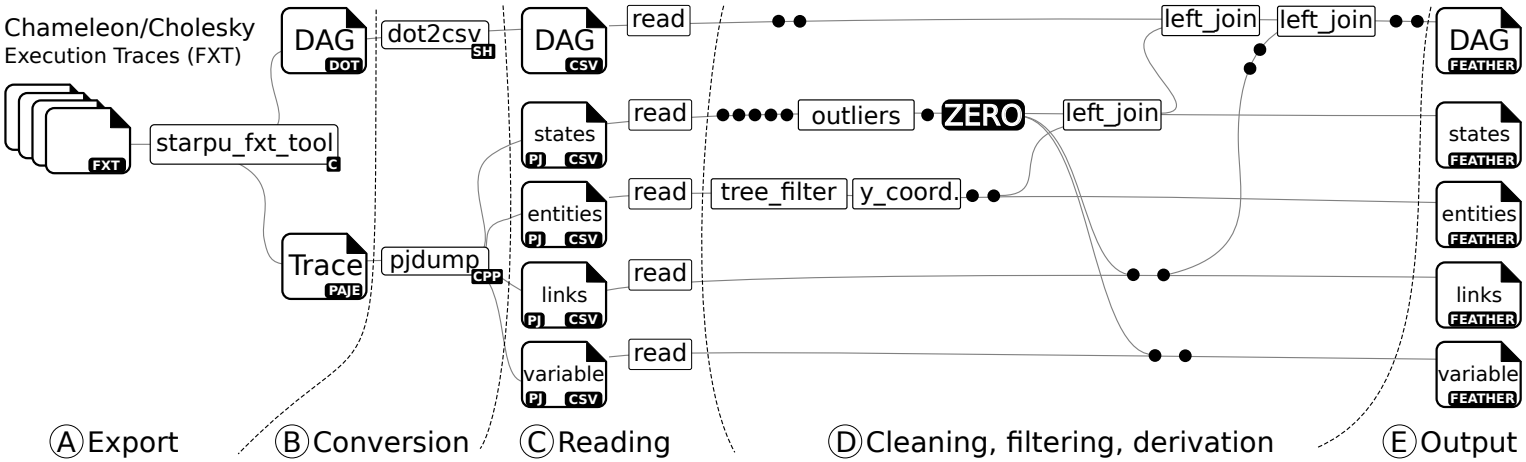
StarVZ: Data analysis tool for StarPU HPC applications

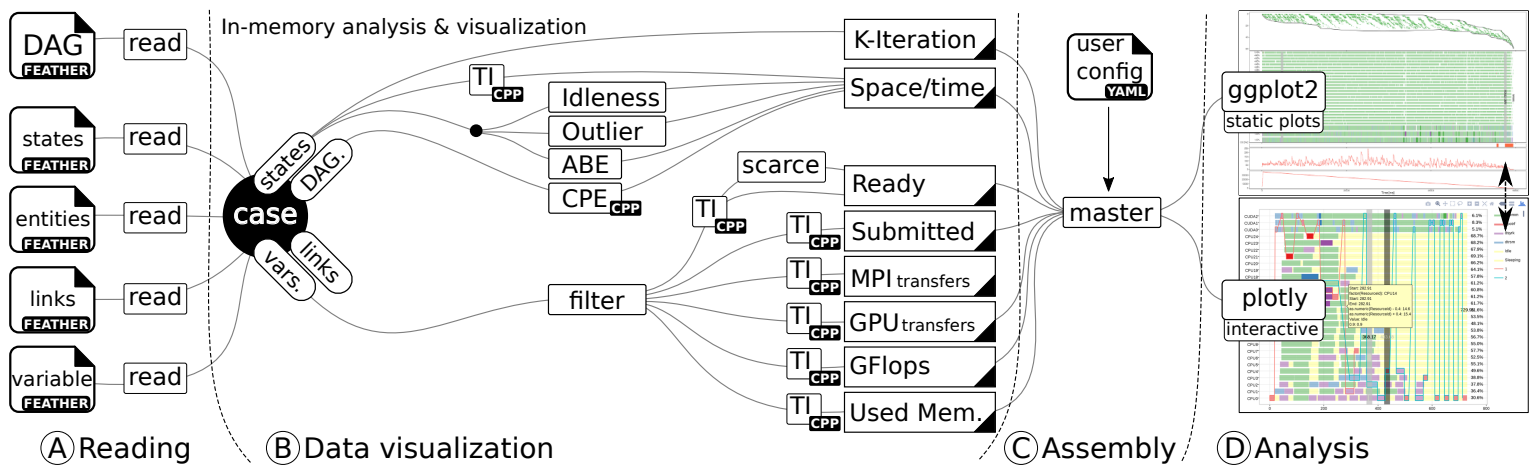




INSTITUTO DE INFORMÁTICA UFRGS

StarVZ: Data analysis tool for StarPU HPC applications





Pre-processing phase is a typical data science case

Many operations / large data frames / many left-joins

Problem Definition and Objectives

Problem Definition

- **StarVZ is sequential**
- Import data, manipulate, plot, analyze, repeat
- Limitation on the amount of data that can be analyzed

Problem Definition and Objectives

Problem Definition

- **StarVZ is sequential**
- Import data, manipulate, plot, analyze, repeat
- Limitation on the amount of data that can be analyzed

Objective

Explore a **parallel workflow** alternative for StarVZ

Approach

- Task parallelism expressed by DAGs
- Expectation: increased performance and scalability
- Setting: the drake R package

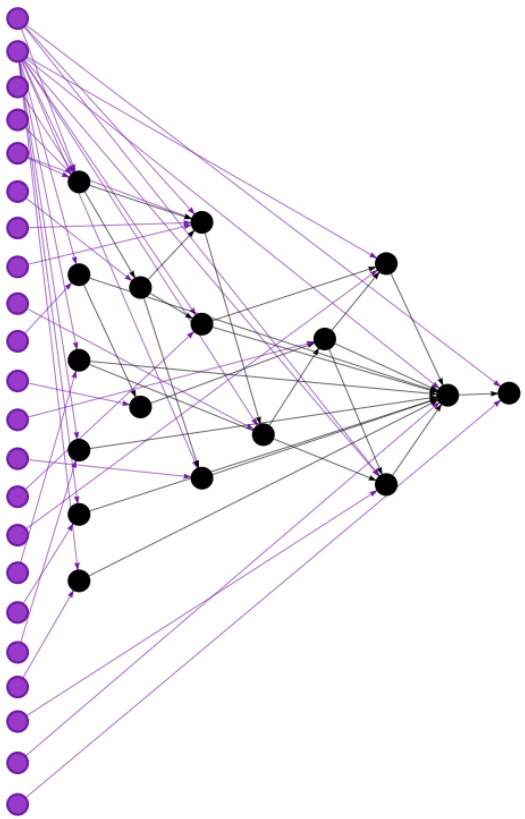
Drake

- R package for managing workflows
- Tasks described as a recipe with dependencies
  - Similar to a Makefile
- Provides
  - Graphical visualization of the workflow in the form of a DAG
  - Parallelism support for independent tasks
  - Caching of previous execution results
    - Enforces the usage of pure functions in recipes

Implementation

Objective: extract data dependencies, make them explicit

1. Separate I/O operations (data imports) from data manipulation
2. Rearrange operations to make tasks as independent as possible
3. Create a Drake recipe to build StarVZ targets



StarVZ with Drake

Speedup benefits of using Drake with StarVZ were not observable

- Drake's caching system writes every target to disk
- Many small tasks that output large data frames as targets
  - I/O operations quickly become a bottleneck
- R's core is single threaded, so parallel Drake workers cannot share memory
  - Hard disk is used for inter-process communication

Conclusions

Drake for data science workflows

- DAG visualization allows for drawing insights on data flow and data transformations
- Explicit data dependencies, facilitate task parallelism
- Caching is also possible

For the case of StarVZ, speedup related features were not applicable

Future work

- Leverage workflow insights to port StarVZ to a data science engine such as Apache Spark

Thank you!

Guilherme Rezende Alles - gralles@inf.ufrgs.br
Lucas Mello Schnorr - schnorr@inf.ufrgs.br

Questions?