Leveraging Workflow Engines and Computing Frameworks for Physics Analysis Scalability and Reproducibility

Conference on Computing in High Energy and Nuclear Physics (CHEP 2024)

Dr. Mindaugas Šarpis

Vilnius University

October 22, 2024

CERN

LHCb

VILNIAUS UNIVERSITETAS · 1579 · UNIVERSITAS VILNENSIS

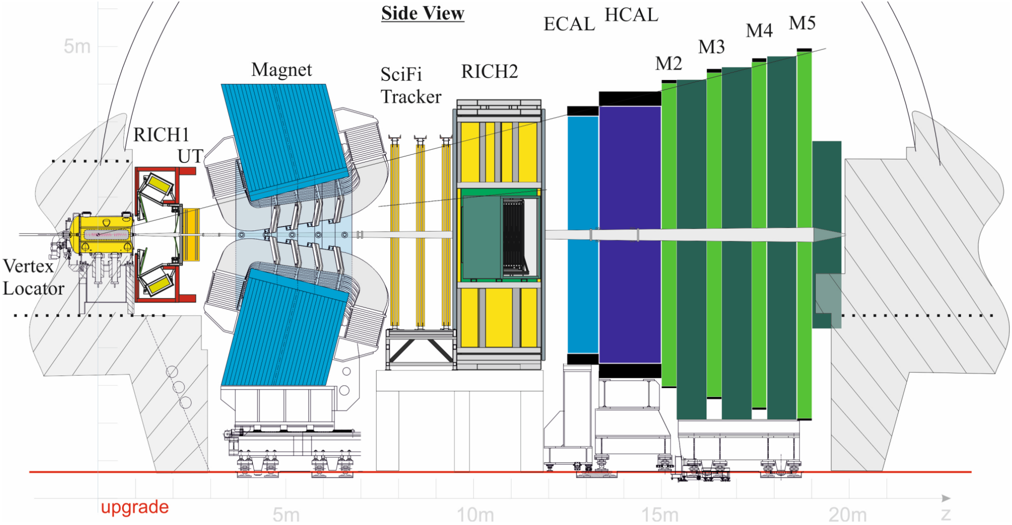
Vilnius University

October 19 - 25, 2024

CHEP 2024

Conference on Computing in High Energy and Nuclear Physics

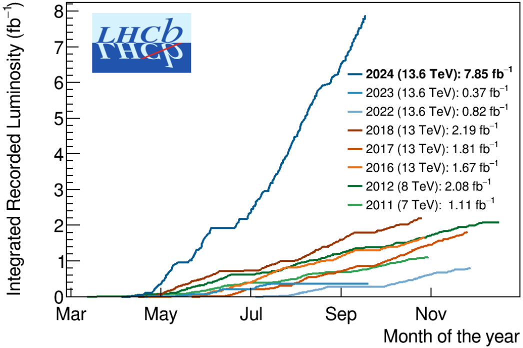
Quick Look at LHCb - Example of a Large Experiment



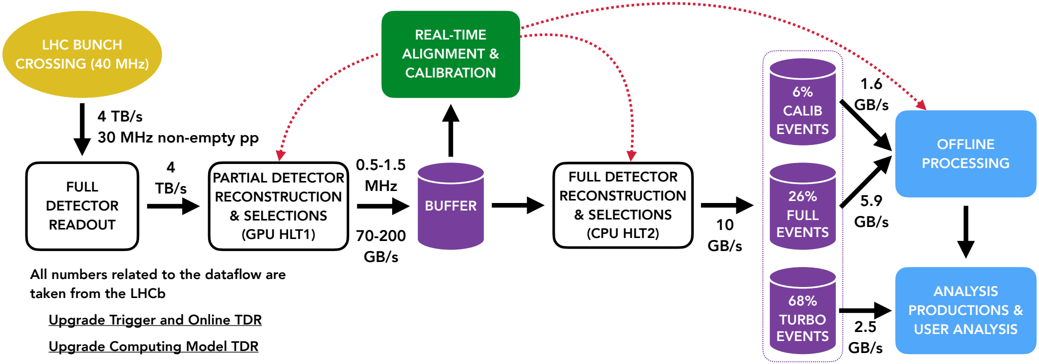
Amount of Data in Run3 and Beyond

- The amount of data collected by the LHC and other large experiments is exploding
- In 2024, LHCb already collected more pp collision data than in all the previous years combined

	ALICE	ATLAS	CMS	LHCb
Run 2:	2 PB	0.5 PB	2 PB	10 PB*
Run 3:	4 PB	1.0 PB	4 PB	45 PB
Total:	**6 PB**	**1.5 PB**	**6 PB**	**55 PB**



Example of Data Flow in HEP

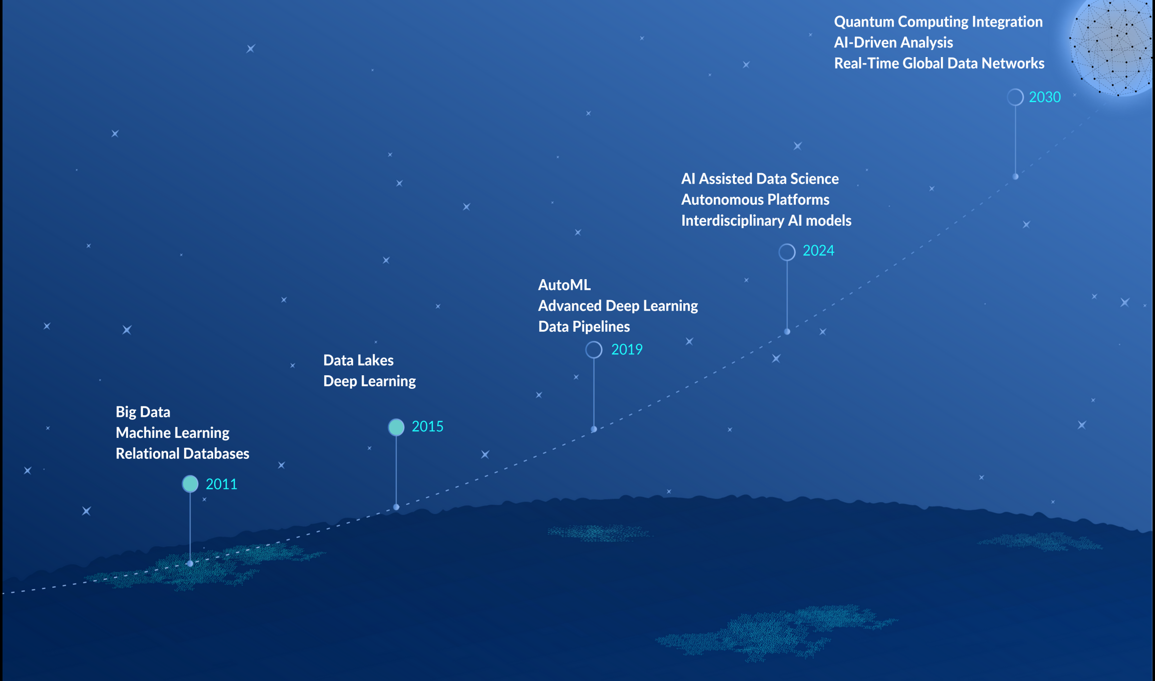

LHC BUNCH CROSSING (40 MHz)
4 TB/s
30 MHz non-empty pp
FULL DETECTOR READOUT
4 TB/s
PARTIAL DETECTOR RECONSTRUCTION & SELECTIONS (GPU HLT1)
0.5-1.5 MHz
70-200 GB/s
BUFFER
REAL-TIME ALIGNMENT & CALIBRATION
FULL DETECTOR RECONSTRUCTION & SELECTIONS (CPU HLT2)
10 GB/s
6% CALIB EVENTS
26% FULL EVENTS
68% TURBO EVENTS
1.6 GB/s
5.9 GB/s
2.5 GB/s
OFFLINE PROCESSING
ANALYSIS PRODUCTIONS & USER ANALYSIS

All numbers related to the dataflow are taken from the LHCb

Upgrade Trigger and Online TDR

Upgrade Computing Model TDR

regression-models
uncertainty-quantification
b-tagging
background-suppression
data-reduction
event-filtering
luminosity
alignment
systematics
track-fitting
monte-carlo
neural-networks
simulation
event-reconstruction
calibration
kinematics
p-value
machine-learning
classifiers
selection
particle-identification
energy-deposition
cross-section
artificial-intelligence
bayes-theorem
deep-learning
vertexing
jet-clustering
signal-extraction
mva
grid-computing
decision-trees
hyperparameter-tuning



2011

Big Data
Machine Learning
Relational Databases

2015

Data Lakes
Deep Learning

2019

AutoML
Advanced Deep Learning
Data Pipelines

2024

AI Assisted Data Science
Autonomous Platforms
Interdisciplinary AI models

2030

Quantum Computing Integration
AI-Driven Analysis
Real-Time Global Data Networks

Findable

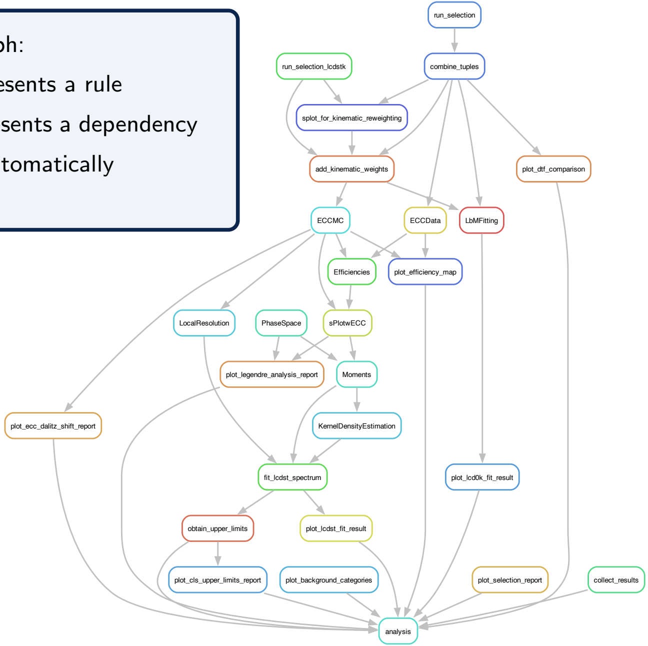
Accessible

Interoperable

Reusable

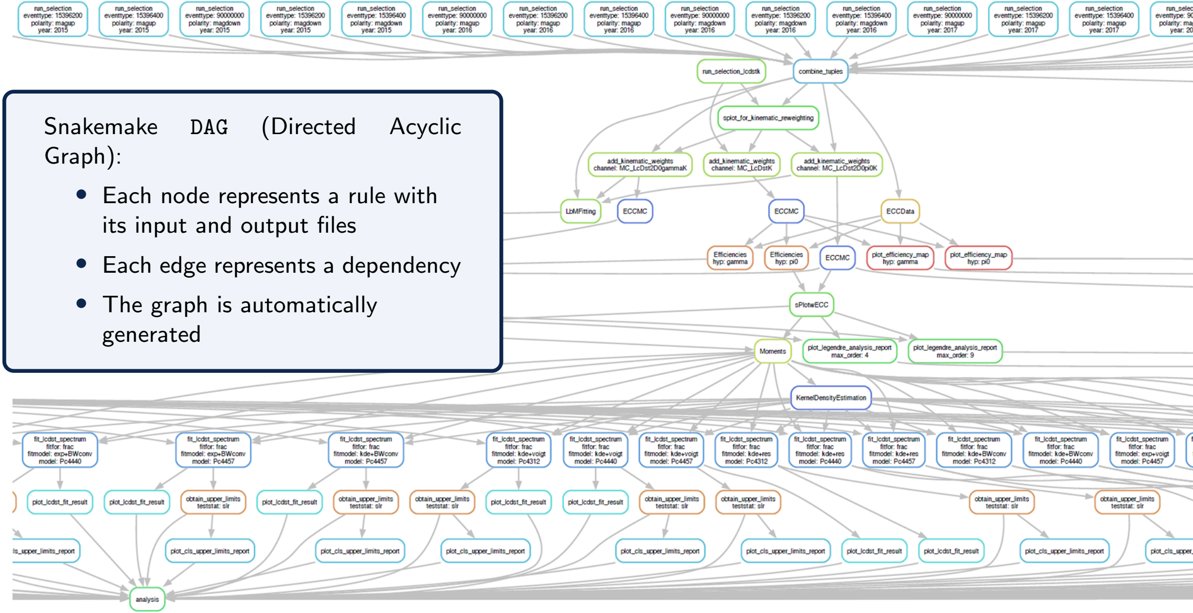
Snakemake Rule Graph:

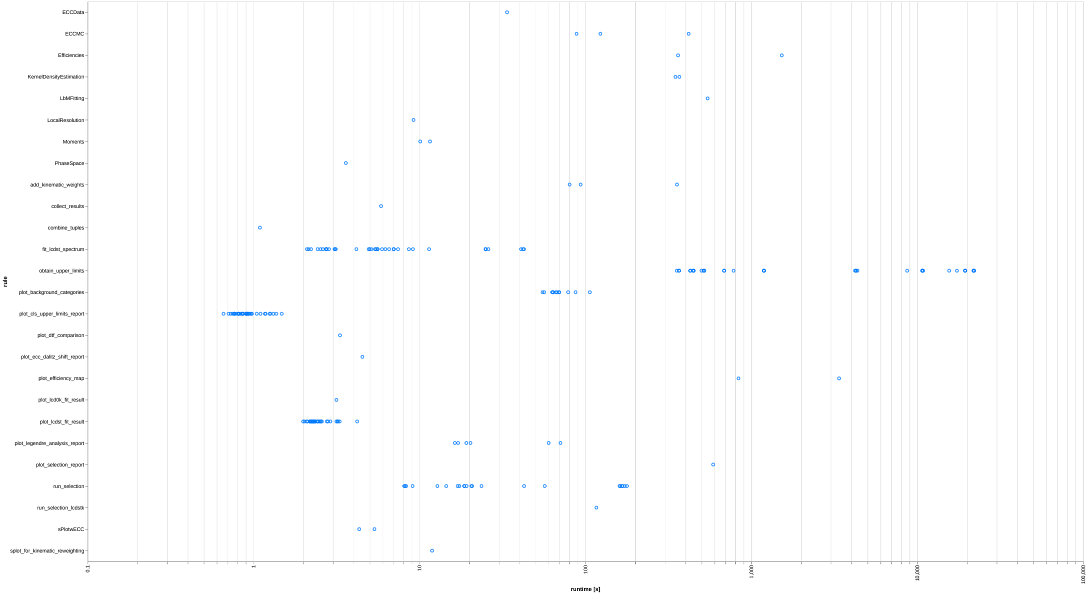
- Each node represents a rule
- Each edge represents a dependency
- The graph is automatically generated



Snakemake DAG (Directed Acyclic Graph):

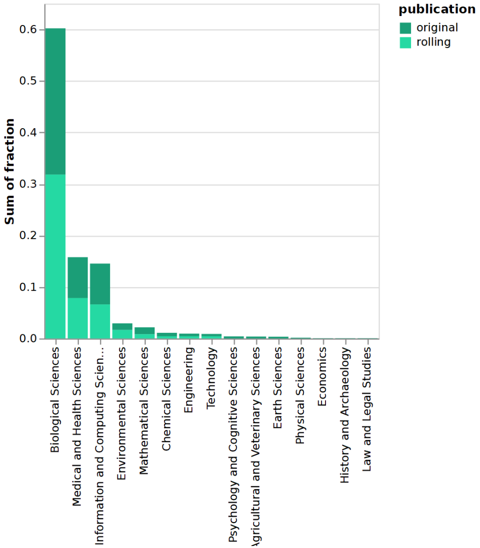
- Each node represents a rule with its input and output files
- Each edge represents a dependency
- The graph is automatically generated


run_selection eventtype: 15396200 polarity: magup year: 2015
run_selection eventtype: 15396400 polarity: magup year: 2015
run_selection eventtype: 90000000 polarity: magdown year: 2015
run_selection eventtype: 15396200 polarity: magdown year: 2015
run_selection eventtype: 15396400 polarity: magdown year: 2015
run_selection eventtype: 90000000 polarity: magup year: 2016
run_selection eventtype: 15396200 polarity: magup year: 2016
run_selection eventtype: 15396400 polarity: magup year: 2016
run_selection eventtype: 90000000 polarity: magdown year: 2016
run_selection eventtype: 15396200 polarity: magdown year: 2016
run_selection eventtype: 15396400 polarity: magdown year: 2016
run_selection eventtype: 90000000 polarity: magup year: 2017
run_selection eventtype: 15396200 polarity: magup year: 2017
run_selection eventtype: 15396400 polarity: magup year: 2017
run_selection_lcdstk
combine_tuples
splot_for_kinematic_reweighting
add_kinematic_weights channel: MC_LcDst2D0gammaK
add_kinematic_weights channel: MC_LcDstK
add_kinematic_weights channel: MC_LcDst2D0pi0K
LbMFitting
ECCMC
ECCMC
ECCData
Efficiencies hyp: gamma
Efficiencies hyp: pi0
ECCMC
plot_efficiency_map hyp: gamma
plot_efficiency_map hyp: pi0
sPlotwECC
Moments
plot_legendre_analysis_report max_order: 4
plot_legendre_analysis_report max_order: 9
KernelDensityEstimation
fit_lcdst_spectrum fitfor: frac fitmodel: exp+BWconv model: Pc4440
fit_lcdst_spectrum fitfor: frac fitmodel: exp+BWconv model: Pc4457
fit_lcdst_spectrum fitfor: frac fitmodel: kde+BWconv model: Pc4457
fit_lcdst_spectrum fitfor: frac fitmodel: kde+voigt model: Pc4312
fit_lcdst_spectrum fitfor: frac fitmodel: kde+voigt model: Pc4440
fit_lcdst_spectrum fitfor: frac fitmodel: kde+voigt model: Pc4457
fit_lcdst_spectrum fitfor: frac fitmodel: kde+res model: Pc4312
fit_lcdst_spectrum fitfor: frac fitmodel: kde+res model: Pc4440
fit_lcdst_spectrum fitfor: frac fitmodel: kde+res model: Pc4457
fit_lcdst_spectrum fitfor: frac fitmodel: kde+BWconv model: Pc4312
fit_lcdst_spectrum fitfor: frac fitmodel: kde+BWconv model: Pc4440
fit_lcdst_spectrum fitfor: frac fitmodel: exp+voigt model: Pc4457
plot_lcdst_fit_result
plot_lcdst_fit_result
obtain_upper_limits teststat: slr
plot_lcdst_fit_result
obtain_upper_limits teststat: slr
obtain_upper_limits teststat: slr
plot_lcdst_fit_result
plot_lcdst_fit_result
obtain_upper_limits teststat: slr
obtain_upper_limits teststat: slr
obtain_upper_limits teststat: slr
obtain_upper_limits teststat: slr
ls_upper_limits_report
plot_cls_upper_limits_report
plot_cls_upper_limits_report
plot_cls_upper_limits_report
plot_cls_upper_limits_report
plot_cls_upper_limits_report
plot_lcdst_fit_result
plot_lcdst_fit_result
plot_cls_upper_limits_report
plot_cls_upper
analysis

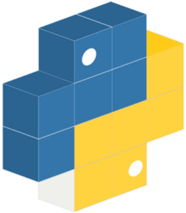

ECCData
ECCMC
Efficiencies
KernelDensityEstimation
LbMFitting
LocalResolution
Moments
PhaseSpace
add_kinematic_weights
collect_results
combine_tuples
fit_lcdst_spectrum
obtain_upper_limits
plot_background_categories
plot_cls_upper_limits_report
plot_dtf_comparison
plot_ecc_dalitz_shift_report
plot_efficiency_map
plot_lcd0k_fit_result
plot_lcdst_fit_result
plot_legendre_analysis_report
plot_selection_report
run_selection
run_selection_lcdstk
sPlotwECC
splot_for_kinematic_reweighting
rule
0.1
1
10
100
1,000
10,000
100,000
runtime [s]

Adoption of Snakemake

- Snakemake is a great example of a tool serving interdisciplinary research
- It helps with the reproducibility of the analysis
- There are a number of great features enhancing the analysis Workflow
- It is still growing in popularity in HEP community



snakemake



reana

GitLab CI

Conclusions

- A modern HEP (or any larger scale) data analysis is becoming impossible without proper workflow management
- There are a number of tools available to ensure analysis reproducibility and scalability
- Workflow engines like Snakemake facilitate the process of efficient analysis
- On the other hand, with the same resources and effort a large scale analysis can be undertaken if using modern workflow paradigms

Thank you for your attention!