ICFA Data Lifecycle Panel: News

ICFA Data Lifecycle panel meeting - November 5, 2024

Kati Lassila-Perini
Helsinki Institute of Physics - Finland

News

- HSF training pre-CHEP workshop
  - Suggestion for recommendations well received (see ICFA DLCP slides)
  - 14 survey participants have volunteers for editing work
- CHEP
  - Several contributions mentioning FAIR, Reproducibility, Open data, Open Science
    - At different levels: tools, policies, services
  - Networking:
    - In the context of DC24
    - Plenary: Global Networking Challenges in the Coming Decade

Networks and new technologies

Thanks to significant preparation by the experts, the network was not a bottleneck during DC24!

Note that the experiments do not make requests for network capacity

More information on new technologies in the DC24 Final Report

- Network routing
- Flow labelling and packet tagging: Fireflies and SciTags
- Load-balancing between networks: NOTED
- Software Defined Networking in Rucio: SENSE
- IPv6
- TCP congestion protocols: BBRv1 vs CUBIC

Also see CHEP24 talks on Network Analytics with ML and SciTags

→ Significant research ongoing, but difficult to demonstrate effectiveness when network was not congested!

Networking (plenary)

Summary

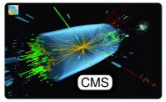
- Wide area networking will continue to deliver quality services for the HEP community into the HL-LHC era.
- But
  - we need to (re)learn how to transfer data efficiently,
  - we need to understand and perhaps manage traffic flows,
  - IPv4 has to go, and
  - life will be more complicated in a multi-science world.

What is PUNCH4NFDI?

Consortium within the **NFDI** - National Research Data Infrastructure in Germany

Particles, **U**niverse, **N**u**C**lei and **H**adrons for the **NFDI**

- From elementary particles to large scale structures
- **Similar challenges** with large data volume
- **Different expertise** in dealing with it



Setup a **federated** and **FAIR** science data platform

- Provide **infrastructures** to process and store **data**
  → Latest news about storage, compute and AAI
- Provide **data portal** to build and re(use) **research products**
  → Integration into analysis platform REANA

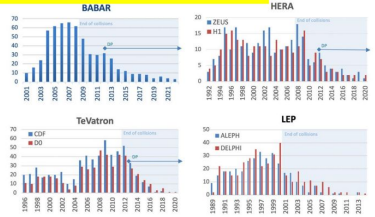
PUNCH4NFDI | Compute4PUNCH | Storage4PUNCH | Access token | REANA | Summary

1/17 Latest Developments of the PUNCH4NFDI Compute and Storage Infrastructures Benoit Roland (KIT)

Conclusions after 10 years: the scientific output

DP is a **cost-effective way of doing fundamental research** by exploiting unique data sets in the light of the increasing theoretical understanding.

- DP leads to
  - a **significant increase in the scientific output** (10% typically)
  - for a minimal investment overhead (0.1%).
  - .... As predicted in 2013



	Data taking stopped	Publications before 2012	Publications 2012-2022	Scientific return increase %
Babar	2008	471	154	33%
H1+ZEUS	2007	436	62	14%

8

FAIR

FAIR principles in High Energy and Nuclear Physics

CHEP 2024

Skills 4 eosc

Already a good practice in many HENP communities

Large experiments adopted FAIR principles for:
- Data Management Plan
  – FAIR access to data and software
- OpenAccess policies

Many leading institutions for OpenScience (CERN, GSI, ...)

What about small communities or individual experiments?
Who can help them?

Project Overview

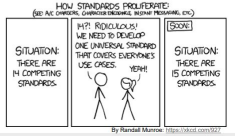
FAIR GSI

Objectives:
- Develop a standardized, adaptable metadata schema.
- Facilitate data input, referencing, management and publication.

Addressing problems:
- Growing unstructured data,
- Diverse data formats and nomenclatures,
- Difficulty in data sharing across institutions,
- No common schema yet between nuclear physics experiments.

NAPMIX

Nuclear, Astro, and Particle Metadata Integration for eXperiments



9

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 | Vilnius University | CHEP 2024

Open Science Wheel
GSI/FAIR OSWG version

GSI

GSI/FAIR "Open Science" wheel

"Open Science and CBM computing tasks"
E.Clerkin, A.Mistry, and V.Friese (2024)
CBM Progress Report 2023

OS Working Group led by Andrew Mistry

GSI/FAIR Open Science Working Group

- Discusses all topics related to Open Science at GSI/FAIR. Interested in formulating policy.
- Representatives from person with different expertise (Lawyers, Admin, Policy, Employee/Project ranking, Scientists).
- Smaller groups dedicated to formulation of Software Production Guidelines, Coding Guidelines, etc.
- Oct. 2023 -> Workshop on Open Science at GSI/FAIR.

Facility for AntiProton and Ion Research in Europe GmbH

The First Release of ATLAS Open Data for Research

CHEP 2024

21 October 2024

Zach Marshall (LBNL) on behalf of the ATLAS Collaboration

A test case for HL-LHC analysis

- The **Analysis Grand Challenge (AGC)** defines a **physics analysis task** with **Open Data** to test **HL-LHC workflows**
  - columnar data extraction from large datasets & data processing into histograms
  - statistical model construction and statistical inference, relevant visualizations
  - ML training & inference

CERN

Cold Storage support on CERN Open Data Portal

Pablo Saiz

21st Oct 2024

Building a Columnar Analysis Demonstrator for ATLAS PHYSLITE Open Data using the Python Ecosystem

KyungEon Choi, Matthew Feickert, Nikolai Hartmann, Lukas Heinrich, Alexander Held, Evangelos Kourlitis, Nils Krumnack, Giordon Stark, Matthias Vigl, Gordon Watts on behalf of the **ATLAS Computing Activity**
(University of Wisconsin-Madison)
matthew.feickert@cern.ch

International Conference on Computing in High Energy and Nuclear Physics (CHEP) 2024
October 21st, 2024

Data Science Institute UNIVERSITY OF WISCONSIN-MADISON

iris hep

ATLAS EXPERIMENT

opendata

LHCb

CERN

LHCb Open Data Ntupling Service: On-demand production and publishing of custom LHCb Open Data

Christine Aidala[1], Dillon Fitzgerald[1], Kai Habermann[2], Ludwig Kramer[2], Adam Morris,[1]Sebastian Neubert[2], **Piet Nogga**[2], Eduardo Rodrigues[3], Marco Donadoni[2], Daan Rosendal[2], Tibor Simko[2]

[1]University of Michigan, [2]University of Bonn, [3]CERN, [4]University of Liverpool

UNIVERSITY OF MICHIGAN

UNIVERSITY OF LIVERPOOL

CHEP 2024

Which ATLAS Open Data?

ATLAS Open data for

Research

See Zach's talk

Education

We talk about this!

CMS Open Data

Leveraging public cloud resources for the processing of CMS open data

CHEP - October 19 - 25, 2024

Kati Lassila-Perini
Helsinki Institute of Physics - Finland
Tom Cordruwisch, Subash Jayawardhana
Lapland University of Applied Sciences - Finland

Open Data

Benchmarking massively-parallel Analysis Grand Challenge workflows using Snakemake and REANA

Marco Donadoni[1] Matthew Feickert[2] Alexander Held[2]
Andrii Povsten[3] Oksana Shadura[4] Tibor Simko[1]
[1]CERN [2]University of Wisconsin Madison (US)
[3]Princeton University (US) [4]University of Nebraska Lincoln (US)

27th Conference on Computing in High Energy and Nuclear Physics
October 21st-25th 2024, Krakow, Poland

Motivation and Facility Overview

- Machine learning/AI has been widely adapted for an array of uses in HEP and beyond.
- However, there are considerations about training these models -
  - *Reproducibility - Can someone later consistently produce the same weights?*
  - *Scalability - How does one scale the training from card->node->cluster scale?*
  - *Efficiency - What are the barriers to using high-speed transports and interconnects?*
- MLTF is a prototype facility that enables training w/these considerations in mind -
  - *Hardware - At this stage, capabilities were chosen over large scale*
  - *Software - Easy-to-use and documented hooks/wrappers for clients to enable functionality*
  - *Infrastructure - Services run at the site to connect the client software to hardware resources*

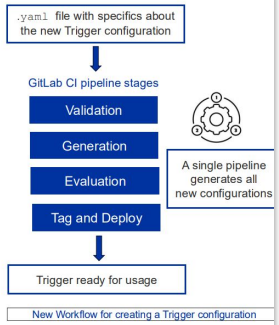
Data discovery, analysis and reproducibility in Virtual Research Environments

Enrique García[1], Giovanni Guerrieri[1], Rubén Pérez[1], Michael Zengel[1], Georgy Skorobogatov[2], Andrés Tanasijczuk[3], Hugo Gonzalez[1] and Xavier Espinal[1]
[1] CERN, [2] ICCUB Barcelona, [3] UCLouvain

CHEP 2024 | Track 9 | 24 October

Summary

- The LHCb Triggers are compiled algorithms that are configured via `python`. These configurations are captured and stored for reference and usage.
- An **automated workflow** via GitLab is now used to generate configurations in a **consistent and reproducible** way, reducing both manual and technical burdens on trigger operators.
  - It provides validations, generation, evaluations and deployment for new trigger configurations.
  - The query-able nature of the trigger configurations allows promising future improvements to aid the LHCb Collaboration's long-term understanding of our Triggers and data-taking.



User sharing of computational workflows in the REANA reproducible analysis platform

M. Donadoni, D. Rosendal, G. Steduto, T. Šimko
(CERN, Geneva, Switzerland)

Facilitating Scientific Reproducibility and Interoperability

Alexandre Boyer , Natthan Pigoux , Luisa Arrabito
1. CERN, EP Department, Geneva, Switzerland
2. Laboratoire Univers et Particules de Montpellier, CNRS/IN2P3, University of Montpellier, France

through in the

CWL Integration Dirac Middleware

Reproducibility

Declarative paradigms for analysis description and implementation

Alberto Annovi, Tommaso Boccali, *Paolo Mastrandrea*, Andrea Rizzi

(INFN & Università di Pisa)

Finanziato dall'Unione europea NextGenerationEU | Ministero dell'Università e della Ricerca | Italiadomani | ICSC

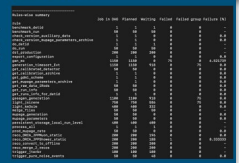
Continuous integration of analysis workflows on a distributed analysis facility

Matteo Bartolini on behalf of the CMS Collaboration

INFN | Università degli Studi Firenze | CMS

Advantages of Snakemake workflow

Systematic logs collection.
Optimized demonstration of failures.
Possibility to restart the workflow from the step it crashed ➟ optimizing resources usage.



Mutualizing resources:
Each run processing is carried out in one place, allowing the use of a local cache directory. If one of the external input is missing, the processing stops before using too much resources.

Containerized software:
Ensuring software version control.
Management of 2nd order dependencies.
Easily integrate processing-ready scripts.

Processing at different computer sites & central data storage: CC-IN2P3, ECAP, Viper, CNAF, Nikhef

PSI | CHEP 2024

Efficient and fast container execution using image snapshotters

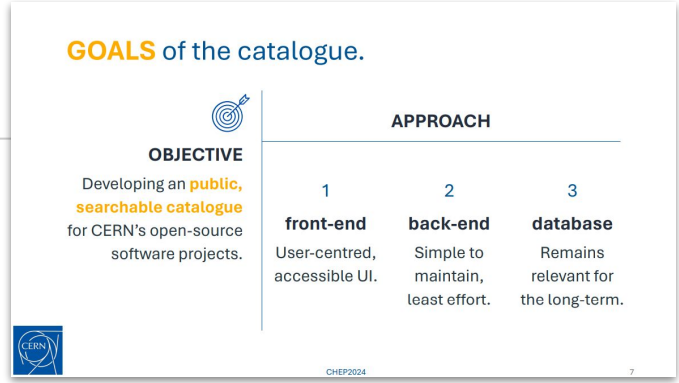
Max Fatouros, Derek Feichtinger, **Clemens Lange (PSI)**
Jakob Blomer, Amal Thundiyil, Valentin Völkl (CERN)
CHEP2024, 22nd October 2024

Designing an Analysis and Grid Facility for DARWIN

KIT

Our Philosophy

- Has to be **easy to deploy** and **easy to use**
- Easily **extendable** for growing collaboration needs
- Rely on **existing and established** tools and experience gained from LHC Computing
- Use **modern**, **open source**, industry **standard** technologies



Open Science

- And – obviously – many other contributions about tools relevant to Data Lifecycle:
  - Root
  - XRootD
  - Rucio
  - …

Outlook

Next meetings on December 3rd, January 21st (then again at 4-week intervals)

Meanwhile, work asynchronously towards the recommendations

Report to ICFA on Thursday