Contribution ID: **69** Type: **Poster**

Quasi interactive analysis of High Energy Physics big data with high throughput

Wednesday 13 March 2024 16:15 (30 minutes)

The need to interject, process and analyze large datasets in an as-short-as-possible amount of time is typical of big data use cases. The data analysis in High Energy Physics at CERN in particular will require, ahead of the next phase of high-luminosity at LHC, access to big amounts of data (order of 100 PB/year). However, thanks to continuous developments on resource handling and software, it is possible to offer users a more flexible and dynamic data access as well as access to open-source industry standards like Jupyter, Dask and HTCondor. This paves the way for innovative approaches: from a batch-based approach to an interactive high throughput platform, based on a parallel and geographically distributed back-end and leveraging the "High-Performance Computing, Big Data e Quantum Computing Research Centre"Italian National Center (ICSC) DataLake model.
This contribution will report the effort of porting multiple data analysis applications - from different collaborations and covering a wide range of physics processes - from a legacy approach to an interactive approach based on declarative solutions, like ROOT RDataFrame. These applications are then executed on the above-mentioned cloud infrastructure, splitting the workflow on multiple worker nodes and outputting the results on a single interface. A performance evaluation, therefore, will also be provided: tentative metrics will be identified, with speed-up benchmarks by upscaling to distributed resources. This will allow to find bottlenecks and/or drawbacks of the proposed high-throughput interactive approach, and eventually help developers committed to its deployment in the Italian National Center.

Significance

This presentation inherits the state of the art of the INFN high throughput infrastructure (presented in the past, although in a more experiment-specific context, see References [1]), but then will cover the novel upscaling on a national level (within the Spoke-2 of the ICSC Italian National Center: https://www.supercomputing-icsc.it/en/spoke-2-fundamental-research-space-economy-en/), benchmarking multiple physics applications which cover different HEP experiments and different demands in terms of computing resources.

References

[1] https://indico.jlab.org/event/459/contributions/11593/ (CHEP 2023)
What is ICSC: https://indico.jlab.org/event/459/contributions/11805/ (CHEP 2023)

Experiment context, if any

Not a specific experiment per se, the physics applications considered are coming from different collaborations (e.g. CMS, ATLAS, FCC,...)

Primary authors: TARASIO, Alessandro (Universita della Calabria e INFN (IT)); CAGNOTTA, Antimo (Universita Federico II e INFN Sezione di Napoli (IT)); SPISSO, Bernardino (Universita Federico II e INFN Sezione di Napoli (IT)); SIMONE, Federica Maria (Universita e INFN, Bari (IT)); GRAVILI, Francesco Giuseppe (INFN Lecce e Universita del Salento (IT)); SABELLA, Gianluca; BARTOLINI, Matteo (Universita e INFN, Firenze (IT)); ANWAR, Muhammad Numan (Universita e INFN, Bari (IT)); MASTRANDREA, Paolo (Universita & INFN Pisa (IT)); DIOTALEVI, Tommaso (Universita e INFN, Bologna (IT)); TEDESCHI, Tommaso (Universita e INFN, Perugia (IT))

Presenter: TEDESCHI, Tommaso (Universita e INFN, Perugia (IT))

Session Classification: Poster session with coffee break

Track Classification: Track 1: Computing Technology for Physics Research