Big Data Tools and Cloud Services for High Energy Physics Analysis in TOTEM Experiment

Valentina Avati†, Milosz Blaszkiewicz†, Enrico Bocchi*, Luca Canali*, Diogo Castro*, Javier Cervantes*, Leszek Grzanka†, Enrico Guiraud*, Jan Kaspar*, Prasanth Kothuri*, Massimo Lamanna*, Maciej Malawski†, Aleksandra Mnich†, Jakub Moscicki*, Shravan Murali*, Danilo Piparo*, Enric Tejedor*
†AGH Universtity of Science and Technology, Krakow, Poland, Email: {grzanka,malawski}@agh.edu.pl
* CERN, CH-1211 Geneva 23, Switzerland, Email: {first.last}@cern.ch

***Abstract*—The High Energy Physics community has been developing dedicated solutions for processing experiment data over decades. However, with recent advancements in Big Data and Cloud Services, a question of application of such technologies in the domain of physics data analysis becomes relevant. In this paper, we present our initial experience with a system that combines the use of public cloud infrastructure (Helix Nebula Science Cloud), storage and processing services developed by CERN, and off-the-shelf Big Data frameworks. The system is completely decoupled from CERN main computing facilities and provides an interactive web-based interface based on Jupyter Notebooks as the main entry-point for the users. We run a sample analysis on 4.7 TB of data from the TOTEM experiment, rewriting the analysis code to leverage the PyRoot and RDataFrame model and to take full advantage of the parallel processing capabilities offered by Apache Spark. We report on the experience collected by embracing this new analysis model: preliminary scalability results show the processing time of our dataset can be reduced from 13 hrs on a single core to 7 mins on 248 cores.**

I. Introduction

Researches and scientists in the field of High Energy Physics (HEP) typically perform massive data analysis in two stages: first, they aggressively reduce the amount of information to be processed by filtering the source dataset; second, they run the final steps of the analysis on the reduced dataset using local computing resources, e.g., their laptop or a computing cluster. The goal of this work is to allow scientists to perform analysis on much bigger datasets with interactive or quasi-interactive response times so to let them use different filtering parameters and enable the exploration of new physics.

To achieve this, we explore the use of modern Big Data tools and we deploy them on the elastically-provisioned infrastructure of the Helix Nebula Science Cloud (HNSciCloud) [1], which is decoupled from existing HEP computing centers. We also explore new parallelisation techniques at the application level by combining the following software ingredients:

- Analysis code from the TOTEM experiment;
- ROOT Data Analysis Framework [2];
- Science Box [3] services – EOS, CERNBox and SWAN;
- Apache Spark task distribution layer.

We compare the performance of this environment with the environment where TOTEM data analysis takes place currently. In this paper we evaluate data processing and parallelisation aspects of the system. Evaluation of other aspects, such as cloud infrastructure, storage services and integration with end-user work environment, is left for future publications.

II. Related work

Offline data analysis is currently performed using a highly-optimized ROOT framework [2], which is customized for the needs of HEP physicists. As a comprehensive and tailored solution, it has become the de-facto tool in the CERN physics community. Nonetheless, there are limitations that constrain the way the analysis is performed. Both the lack of parallelization capabilities and the deficiencies imposed by the execution on a single machine lead to the inability of working with large datasets in an interactive way. New additions to the ROOT framework such as the RDataFrame, scrutinized as a part of this ongoing effort, solve part of the former problem.

Apache Spark is a popular, open-source framework for the distributed computing. Its Spark DAG and task schedulers features allow the deployment of map-reduce-type operations in a scalable way on large clusters. In this work, we leverage these features to parallelize and scale ROOT jobs across a cluster. A different approach consists of reading ROOT files into Spark using the Spark-Root connector developed by the CMS Big Data project [4]. This approach, however, requires data processing jobs to be written using native Spark Dataframe APIs, as opposed to using ROOT APIs, as in the rest of this work.

III. Architecture

The architecture of our deployment on the HNSciCloud is shown in Fig. 1 and it involves the following components:

• EOS [5] is a distributed storage system used to host all physics data at CERN. A dedicated EOS instance storing a subset of the TOTEM experiment data is deployed on the HNSciCloud, allowing for fast data access from the computing nodes operating in the HNSciCloud.

• SWAN [6] is a web-based platform to perform interactive data analysis. It inherits the Jupyter notebook interface and integrates the ROOT analysis framework with the dedicated ROOT C++ kernel. In addition, it is capable of offloading massive computations to a Spark cluster, it uses CVMFS to access scientific software packages, and provides access to EOS emulating a local filesystem access.

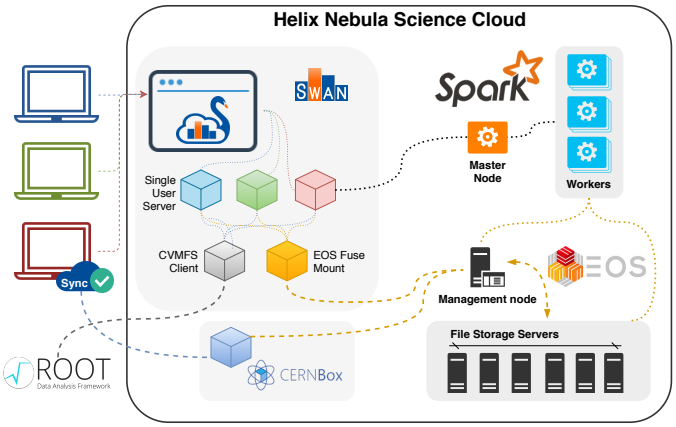

Fig. 1. Architecture and deployment of the distributed analysis components

• A dedicated Spark cluster is deployed on the HNSciCloud and is integrated with SWAN. Specifically, we make use of RDataFrame: a new interface of ROOT that allows the partitioning of the dataset in the form of DataFrame and uses Spark for spreading the computations across worker nodes.

• CERNBox [7] offers a web interface to manage files stored on EOS and allows for easy sharing of SWAN notebooks between users as well as for synchronization of selected folders with personal laptops.

All the Science Box services run in Docker containers orchestrated by Kubernetes, while Spark runs on plain VMs and is configured via Cloudera Manager. The overall deployment consists of 25 VMs, 388 CPUs, 1,450 GB of memory, and 21.5 TB of storage. 288 CPUs and 1,088 GB of memory are reserved for Spark, while 16.4 TB is the physical storage space available for EOS (the actual space available is 8.2 TB due to a `replica 2` layout of the stored files).

IV. Evaluation

The typical analysis scenario consists of three main steps:

1) Source datasets are copied from the storage infrastructure at CERN to EOS on the HNSciCloud;
2) Physics analysis is performed using the web-based interface of SWAN and the computing resources of Spark;
3) The produced results can be viewed interactively with SWAN, synchronized with personal laptops or shared with colleagues via CERNBox.

The physics analysis we used in evaluation is the analysis of the elastic scattering data gathered by TOTEM experiment in 2015 during a special LHC run with optics parameter β^* adjusted to 90 m. The dataset comprises 1153 files summing to 4.7 TB of data in ROOT Ntuple format, and stores 2.8 Billion events representing proton-proton collisions.

The original analysis was written using the ROOT framework and comprises 2 stages: 1. Data reduction, and 2. Filtering based on physics cuts. It followed a traditional approach of implementing a main processing loop which accesses individual events. The output is a set of histograms representing distributions of interesting observables.

The analysis was re-implemented using ROOT's RDataFrame [8], a new high level interface. It preserves the flexibility in the actions that can be performed inside the event-loop, while offering a concise declarative syntax. Implicit multi-threading and other low-level optimisations allow exploiting all the computing resources available transparently, while the data can be partitioned and distributed to multiple nodes by the Spark task distribution layer.

We performed preliminary experiments using a Spark cluster with up to 248 cores allocated to Spark workers. The collected results show that is is possible to substantially reduce the processing time of our 4.7 TB dataset from about 13 hours on a single core to less than 7 minutes on 248 cores.

V. Conclusions and Future Work

In the first phase of pilot deployment we focused on validation of the results at the application level: we assured that the physics results obtained in the new system are correct and correspond to the known and validated results provided by the TOTEM collaboration. In the second phase we focused on optimization of processing: improving the RDataFrame implementation, fine-tuning the Spark cluster and understanding task distribution strategies.

The speedup observed in preliminary tests is promising: if further reduction of time is achieved, it should be possible to use the proposed approach for quasi-interactive analysis of the whole dataset, instead of multi-step batch processing.

Once the preliminary testing is completed, we plan to run more large-scale experiments to evaluate possible performance gains and to better understand the impact of parameters such as data partitioning, distribution or possible caching strategies. The results of this study will be of interest to wider physics community exploring the use of modern Big Data tools.

Acknowledgments: This work was supported in part by the Polish Ministry of Science and Higher Education.

References

[1] M. Gasthuber, H. Meinhard, and R. Jones, "HNSciCloud - Overview and technical challenges," *J. Phys. : Conf. Ser.*, vol. 898, no. 5, p. 052040. 5 p, 2017. [Online]. Available: http://cds.cern.ch/record/2297173

[2] I. Antcheva *et al.*, "ROOT: A C++ framework for petabyte data storage, statistical analysis and visualization," *Comput. Phys. Commun.*, vol. 180, pp. 2499–2512, 2009.

[3] CERN. (2018) Science Box. [Online]. Available: https://sciencebox.web.cern.ch

[4] M. Cremonesi *et al.* Using big data technologies for hep analysis. CERN. [Online]. Available: https://indico.cern.ch/event/587955/contributions/2937521/attachments/1684310/2707721/chep_bigdata.pdf

[5] A. Peters, E. Sindrilaru, and G. Adde, "EOS as the present and future solution for data storage at CERN," *J. Phys.: Conf. Ser.*, vol. 664, no. 4, p. 042042. 7 p, 2015. [Online]. Available: http://cds.cern.ch/record/2134573

[6] D. Piparo, E. Tejedor, P. Mato, L. Mascetti, J. Moscicki, and M. Lamanna, "SWAN: a Service for Interactive Analysis in the Cloud," *Future Gener. Comput. Syst.*, vol. 78, no. CERN-OPEN-2016-005, pp. 1071–1078. 17 p, Jun 2016. [Online]. Available: http://cds.cern.ch/record/2158559

[7] L. Mascetti, H. G. Labrador, M. Lamanna, J. Moscicki, and A. Peters, "CERNBox + EOS: end-user storage for science," *J. Phys.: Conf. Ser.*, vol. 664, no. 6, p. 062037. 6 p, 2015.

[8] E. Guiraud, A. Naumann, and D. Piparo, "TDataFrame: functional chains for ROOT data analyses," Jan. 2017. [Online]. Available: https://doi.org/10.5281/zenodo.260230