Contribution ID: **427** Type: **Poster**

Last development of the Long Term Data Preservation project for CDF at INFN CNAF

Thursday, 13 October 2016 16:30 (15 minutes)

The long term preservation and sharing of scientific data is becoming nowadays an integral part of any new scientific project. In High Energy Physics experiments (HEP) this is particularly challenging, given the large amount of data to be preserved and the fact that each experiment has its own specific computing model. In the case of HEP experiments that have already concluded the data taking phase, additional difficulties are the preservation of software versions that are not supported anymore and the protection of the knowledge about data and analysis framework.

The INFN Tier-1, located at CNAF, is one of the reference sites for data storage and computing in the LHC community but it also offers resources to many other HEP and non-HEP collaborations. In particular the CDF experiment has used the INFN Tier-1 resources for many years and, after the end of data taking in 2011, it faced the challenge to both preserve the large amount of data produced during several years and the ability to access and reuse the whole amount of it in the future. According to this task the CDF Italian collaboration, together with the INFN CNAF and Fermilab (FNAL), has implemented a long term data preservation project. The tasks of the collaboration comprises the copy of all CDF raw data and user level ntuples (about 4 PB) at the INFN CNAF site and the setup of a dedicated framework which allows to access and analyze the data in the long term future. The full sample of CDF data was successfully copied from FNAL to the INFN CNAF tape library backend and a new method for data access has been set up. Moreover a system for doing regular integrity check of data has been developed: it ensures that all the data are accessible and in case of problems it can automatically retrieve an identical copy of the file from FNAL. In addition to this data access and integrity system, a data analysis framework has been implemented in order to run the complete CDF analysis chain in the long term future. It is also included a feasibility study for reading the first CDF RUN-1 dataset now stored in old Exabyte tapes.

As an integral part of the project, detailed documentation for users and administrations, has been produced in order to analyse data and maintain the whole system.

In this paper we will illustrate the different aspects of the project: from the difficulties and the technical solutions adopted to copy, store and maintain CDF data to the analysis framework and documentation web pages. We will also discuss the learned lesson from the CDF case which can be used when designing new data preservation projects for other experiments.

Tertiary Keyword (Optional)

Storage systems

Secondary Keyword (Optional)

Data processing workflows and frameworks/pipelines

Primary Keyword (Mandatory)

Preservation of analysis and data

Primary authors: FATTIBENE, Enrico (INFN - National Institute for Nuclear Physics); DELL'AGNELLO, Luca (INFN-CNAF); RICCI, Pier Paolo (INFN CNAF); AMERIO, Silvia (University of Padova & INFN); PEZZI, michele (Infn-cnaf)

Presenter: PEZZI, michele (Infn-cnaf)

Session Classification: Posters B / Break

Track Classification: Track 8: Security, Policy and Outreach