Charge on Future of Data Management

M. Kirby, A. Norman, L. Sexton-Kennedy

The long term data management challenges of the HEP community will be defined by both the experiment data volume that has to be stored, processed and analyzed as well as the data rate and time structure with which the data is recorded. The High Luminosity LHC (HL-LHC) will generate data at a rate more than an order of magnitude beyond the current LHC Run 2 rates and with equivalent yearly machine live times. On a similar time scale, the DUNE experiment will also begin collecting data at a rate that is an order of magnitude beyond current neutrino experiments data rates. DUNE has the potential to rival the data rates of the HL-LHC experiments under certain readout conditions (i.e. supernova neutrino burst readouts).

Efforts within the HEP community have generated outlines of the R&D that will be needed in the fields of data storage and data management. These R&D pathways may be challenging and may require significant resources to address them. Funding profiles for the field of HEP will necessitate that collaboration and common tools be explored between the LHC and intensity frontier communities.

The community has also identified that the data rates of future experiments will result in exabyte scale datasets that need to be analyzed by the physics collaborations. The datasets will most likely be hosted at facilities at Fermilab, CERN and other institutions, but the technological landscape, both for data analysis and data storage, may be vastly different than is currently used for the intensity frontier experiments today or the LHC Run II experiments. Whether this landscape involves large distributed grid/cloud resources similar to today's computing environment, high performance computing (HPC) centers, or both, the experimental data and simulation will need to be effectively and efficiently transported, catalogued and managed.

The U.S. computing heads of DUNE, CMS and of the FIFE project recognize these challenges and the need to address the requirements of experiments in relation to the future and the future of the facilities that support our experiments.

We request that a committee be charged to examine and review the needs of the experiments with respect to these large scale data management and data handling challenges. We request that the committee enumerate a set of requirements that reflects those needs with special attention to CMS, DUNE and the FIFE project. The assessment should focus on the commonalities and differences in requirements between the experiments, with the experiments acting as stakeholder in the process. The global assessment should follow a coordinated approach to align the work across the individual assessments of each experiment.

In particular the committee is asked to assess the requirement relating to:

- Archival data storage needs
- Data access models and data distribution models

- I/O Bandwidth needs for DAQ output, production, analysis chains, and other computing activities
- Data/workflow integration for HEP analysis
- Data access/distribution for HPC facilities
- Data access/distribution in highly distributed grid/cloud models
- Needed for changes to authentication/authorization schemes to support large international collaborations

The Atlas collaboration recently switched to a new data management solution [Rucio], and as part of that migration performed a requirements assessment similar to the one we are requesting. We believe that this assessment is quite valid, and should be leveraged in the preparation of this review. In particular we charge the committee to address the current capabilities of the Rucio data handling system, the CMS data handling suite, the SAM/FIFE data management system, and other pertinent data handling systems that may have the ability to address the scales involved in the future of HEP data management. Address how these systems currently can be used to meet future requirements and where these tools are currently deficient. If non-HEP tools or commercial facilities are considered as applicable to this review, they can be included within the report but are not required. Where possible, identify complementary aspects of the tool suites that may address these requirements and areas where features of the tools can be leveraged to meet future requirements.

Provide recommendations on how additional facilities, functionality and effort can be directed to data management solutions and common projects that meet the requirements of this assessment.

The named committee will be composed of experts in Data management, Data Storage, and workflow management. The committee should engage the CMS, ATLAS, DUNE, and FIFE experiments to provide domain specific expertise and serve as stakeholders in the evaluation process.

The committee will draft this set of requirements and outline for the future of data management by Oct. 31, 2017. The report will be presented to FNAL-SCD senior management.

The named committee will be composed of:
Eric Vaandering (CMS)
Robert Illingworth (Data management)
Bo Jayatilaka (Data Storage)