Contribution ID: **11** Type: **Poster**

ATLAS Event Index prototype for cataloguing large amounts of data

Description of content and intended audience</br>- the outcome you expect to achieve.

This project consists in the development and deployment of a catalogue of events for experiments with large amounts of data, such as those currently taking data with the LHC accelerator at CERN. The ultimate goal is to have available for our experiments, and make available to other users, a consistent set of software packages that allow fast and efficient searching of events of interest among the billions of events recorded in millions of files scattered in a world-wide distributed computing system.
The most innovative part of this project is the adaptation and application, for the first time, of the NoSQL technology for the cataloguing of data of a large experiment. ATLAS alone, but other experiments have similar numbers, accumulated since 2011, 2 billion real and 6 billion simulated events. This enormous amount of data (many TB) can be managed in Oracle database only by investing in hardware and operating personnel, and dividing the database into parts that for practical and financial reasons are located in different computing centres in different countries. With this work we want to explore the usage of NoSQL technologies for this kind of indexing, and evaluate aspects of performance and scalability that are extremely important to have a product that will be useful to the scientific community.

Printable summary: this is the only </br> section of the abstract that will
be published in the Book of Abstracts.

The Event Index project consists in the development and deployment of a catalogue of events for experiments with large amounts of data, such as those currently taking data like the ATLAS detector with the LHC accelerator at CERN. A database with the references to the files including each event in every stage of processing is necessary in order to later retrieve the selected events from data storage systems, and to be used as a reference index for final users, or for automated tools.
In this poster we present the architecture and the current implementation for the different parts involved in the project, including the data collection and upload to the central Hadoop server, and the designed infrastructure to access the stored information.
Data to be stored in the EventIndex are produced worldwide by all production jobs that run on Tier-0 or the Grid. For every permanent output file a snippet of information, containing the file unique identifier and for each event the relevant attributes is sent. In our first prototype we are using messaging technologies like Stomp protocol and ActiveMQ broker to convey the information. The estimated rate (in May 2013, during the LHC shutdown) is about ~20 Hz of file records containing ~3.4 kHz of event records, summing up a rate of 300GB of event information per day. During data-taking periods these numbers will be doubled; as both the data-taking rate and the Grid processing power are expected to increase by perhaps a factor two by 2015.

Primary author: FERNANDEZ, Alvaro (CSIC)

Co-authors: FAVARETO, Andrea (Università di Genova and INFN, Genova, Italy); BARBERIS, Dario (Università di Genova and INFN, Genova, Italy); MALON, David (Argonne National Laboratory, Argonne, IL, United States); GALLAS, Elizabeth (Oxford University, Oxford, United Kingdom); PROKOSHIN, Fedor (Universidad Técnica Federico Santa María, Valparaíso, Chile); SANTIAGO, GONZALEZ (CSIC); CRANSHAW, Jack (Argonne National Laboratory, Argonne, IL, United States); SÁNCHEZ, Javier (CSIC, Valencia, Spain); SALT, Jose (IFIC (Instituto de Física Corpuscular)); HŘIVNÁČ, Julius (LAL, Université Paris-Sud and CNRS/IN2P3, Orsay, France); NOWAK, Marcin (Brookhaven National Laboratory, Upton, NY, United States); BRIONGOS, Pablo (CSIC); YUAN, Ruijun Justine (LAL, Université Paris-Sud and CNRS/IN2P3, Orsay, France)

Presenter: FERNANDEZ, Alvaro (CSIC)