Contribution ID: **200**

Type: **Poster**

# A study of data representations in Hadoop to optimize data storage and search performance of the ATLAS EventIndex

*Tuesday 11 October 2016 16:30 (15 minutes)*

This paper reports on the activities aimed at improving the architecture and performance of the ATLAS EventIndex implementation in Hadoop. The EventIndex contains tens of billions event records, each of which consisting of ˜100 bytes, all having the same probability to be searched or counted. Data formats represent one important area for optimizing the performance and storage footprint of applications based on Hadoop. This work reports on the production usage and on tests using several data formats including Map Files, Apache Parquet, Avro, and various compression algorithms.

The query engine plays also a critical role in the architecture. This paper reports on the use of HBase for the EventIndex, focussing on the optimizations performed in production and on the scalability tests. Additional engines that have been tested include Cloudera Impala, in particular for its SQL interface, and the optimizations for data warehouse workloads and reports.

## Secondary Keyword (Optional)

Databases

## Primary Keyword (Mandatory)

Storage systems

## Tertiary Keyword (Optional)

**Authors:** BARBERIS, Dario (Università e INFN Genova (IT)); HRIVNAC, Julius (Universite de Paris-Sud 11 (FR)); CANALI, Luca (CERN); TOEBBICKE, Rainer (CERN); BARANOWSKI, Zbigniew (CERN)

**Presenter:** CANALI, Luca (CERN)

**Session Classification:** Posters A / Break

**Track Classification:** Track 4: Data Handling