21st International Conference on Computing in High Energy and Nuclear Physics **CHEP2015** Okinawa Japan: April 13 - 17, 2015

Contribution ID: **43** Type: **poster presentation**

Scale Out Databases for CERN Use Cases

Data generation rates are expected to grow very fast for some database workloads going into LHC run 2 and beyond. In particular this is expected for data coming from controls, logging and monitoring systems. Storing, administering and accessing big data sets in a relational database system is in certain cases very demanding on the technology and therefore on costs. Notably one of the critical parts in the architecture of Oracle database clusters is the use of shared storage. Therefore there is a high interest in the CERN database community to look for alternative solutions for storing and querying big data volumes with fast and scalable data access time. Scale out database engines are an emerging and rapidly developing area. Recently a technical solution that has attracted attention is Cloudera Impala with columnar storage provided by Parquet on top of Hadoop Distributed File System. This solution has the additional benefit of offering SQL as the main data access interface which makes it easy to integrate with existing client application. In this paper we will describe the architecture of database systems based on Impala Hadoop clusters and we will discuss the results of our tests, including tests of data loading and integration with existing data sources, notably Oracle databases. We will report on query performance tests done with various data sets of interest at CERN, notably the accelerator log database.

Primary author: BARANOWSKI, Zbigniew (CERN)

Co-authors: LANZA GARCIA, Daniel (Univ. Extremadura, Cen. Uni. Merida (ES)); CANALI, Luca (CERN); GRZYBEK, Maciej (Warsaw University of Technology (PL))

Presenter: BARANOWSKI, Zbigniew (CERN)

Track Classification: Track3: Data store and access