20th International Conference on Computing in High Energy and Nuclear Physics (CHEP2013)

Contribution ID: **357** Type: **Poster presentation**

dCache Billing data analysis with Hadoop

Monday, 14 October 2013 15:00 (45 minutes)

The dCache storage system writes billing data into flat files or a relational database.
For a midsize dCache installation there are one million entries - representing 300 MByte - per day.
Gathering accounting information for a longer time interval about transfer rates per group, per file type or per user results in increasing load on the servers holding the billing information.
Speeding up these requests renders new approaches to performance optimization worthwhile.

Hadoop is a framework for distributed processing of large data using multiple computer nodes.
The essential point in our context is the scalability for big data.
Data is distributed over many nodes in the Hadoop Distributed File System (HDFS).
Queries are processed in parallel on every node to extract the information and combine it in another step.
This is called a MapReduce algorithm.
As the dCache storage is distributed over many storage nodes combining both on every node is obvious.

The transformation of the billing information into the HDFS structure is done by a small script.
The MapReduce functions to create the results to the most important queries are implemented for each request.
We will present the system's setup and performance comparisons of the created queries using Postgresql, flat files and Hadoop.
The overall gain in performance and its dependence on both the amount of analysed data and available machines for paralleling the request will be demonstrated.

Primary author: LEFFHALM, Kai (Deutsches Elektronen-Synchrotron (DE))

Co-author: Mr KNOEPKE, Andreas (DESY)

Presenter: LEFFHALM, Kai (Deutsches Elektronen-Synchrotron (DE))

Session Classification: Poster presentations

Track Classification: Data Stores, Data Bases, and Storage Systems