Contribution ID: **169** Type: **Sectional**

Metadata curation and integration in High Energy and Nuclear Physics

Friday, September 29, 2017 11:35 AM (15 minutes)

Modern High Energy and Nuclear Physics experiments generate vast volumes of scientific data and metadata, describing scientific goals, the data provenance, conditions of the research environment, and other experiment-specific information. Data Knowledge Base (DKB) R&D project has been initially started in 2016 as a joint project of National Research Center "Kurchatov Institute"and Tomsk Polytechnic University. And later the interest from the ATLAS experiment at LHC guided it to the new area of studies.
Within the project we studied metadata sources in ATLAS. There are many sources of metadata, such as physics topics metadata, papers and conference notes, supporting documents, Twiki pages, google documents and spreadsheets, data sample catalogs, conditions and production analysis system databases. It has been noticed that information between sources is loosely coupled. Therefore, to provide a holistic view on physics topics, including integrated representation of all ATLAS documents and corresponding data samples, scientists need to obtain cross relations among metadata by themselves.
DKB is designed to provide metadata integration and is considered to look for cross references among the metadata from various data sources.
For the end user DKB frontend will be implemented as a graphical user interface, providing convenient integrated metadata representation, navigation, and efficient search - upwards to common metadata (production campaigns, projects, physics groups) and downwards from the specific, fine-grained metadata objects (detector geometry version, software release, conditions tags).
Currently, the data scheme of ATLAS integrated metadata is organized as an ontological model. The backend of the DKB is the OpenLink Virtuoso RDF storage. It is populated with the information from ATLAS publications, supporting documents and underlying data samples. Metadata from unstructured texts were extracted by the PDFAnalyzer utility, developed by the research team. The integration dataflow execution is automated by Apache Kafka Streams.
We observed that Twiki pages are very popular in physics community and they contain the metadata, corresponding to physics topics and production campaigns in semi-structured form. It is natural to expand the DKB functionality by adding the analysis of the Twiki pages. This will allow to have more complete and accurate integrational data model. Implementation of the DKB is closely related to the data samples curation and discovery. To choose the most suitable method, providing a performant look up for data samples by the various combinations of parameters, we should evaluate different technologies, such as graph databases (Neo4j, OrientDB), ElasticSearch, Virtuoso, Oracle JSONs search.
In our report we will summarize the current state of our project, technology evaluation results and the recent prototype of the DKB architecture.

Primary author: Ms GRIGORIEVA, Maria (NRC KI)

Co-authors: Dr KLIMENTOV, Alexei (Brookhaven National Lab); GUBIN, Maksim (Tomsk Polytechnic University); Mrs GOLOSOVA, Marina (National Research Center "Kurchatov Institute"); Prof. WENAUS, Torre (Brookhaven National Laboratory); Mr AULOV, Vasiliy (NRC Kurchatov Institute)

Presenter: Ms GRIGORIEVA, Maria (NRC KI)

Session Classification: Research Data Infrastructures@Computing for Large Scale Facilities

Track Classification: Non-relational Databases and Heterogeneous Repositories