Contribution ID: **9** Type: **not specified**

Hadoop in Complex Systems Research

Tuesday, September 2, 2014 11:15 AM (25 minutes)

I am planning to shed light onto the theme of 'Metadata Management' in Hadoop. The Hive-Metastore exists for a long time and complementary to to it, there is HCatalog.
With this Pig users and MapReduce developers can access those Metadata as well. But how do we handle time-dependent aspects of Complex Systems that consist of multiple interrelated layers represented as graphs?
To handle such aspects efficiently, a new methodology that uses a semantic Wiki is proposed and demonstrated. The triple store is used as a centralized database and as an automatic system integration layer which works with a SPARQL-like query language.
Researchers and analysts can concentrate on system modeling aspects while developers focus on efficient I/O operations - whereby the content of the data is of minor importance.
I demonstrate the concept with an example using Apache Giraph and Gephi. Such analysis workflows can span numerous distributed clusters and all dependencies are documented in the Semantic Wiki. So we maintain a meta model for an arbitrary analysis-workflow which can be split into separate 'local Oozie workflows.'

Presenter: KÄMPF (CLOUDERA), Mirko

Session Classification: Plenary talks