20140703 Meeting between UFRJ and SLAC

Luiza has set up a small project in the UFRJ Reference center to provide big data analysis/mining of PingER multidimensional data

Luiza has proposed three approaches:

1. Conventional. Utilization of Pentaho environment to handle big multidimensional data, which enables utilization of enhanced user interfaces.
2. Linked Data. Benchmarking of more sophisticated Triple Stores than the one we use today at PingER LOD (Sesame). Preferably, we should analyze parallel and distributed solutions. CumulusRDF is an example.
    a. Renan is investigating an alternative to Hadoop, which utilizes a Scientific Workflow Management System and makes use of Map/Reduce paradigm to help both querying and provenance of the Linked Data (RDF) data.
    b. Ibrahim is investigating an approach that utilizes Hadoop Map/Reduce in a Key/Value store with PingER data in RDF.
3. Utilization of Greenplum (http://en.wikipedia.org/wiki/Greenplum). This is an intensive high performance database from EMC with many features such as caching. It is partly from the EMC acquisition of Pivotal. There is also a DBMS called Grindplan that explores lots of features using Pivotal.

Les will make available via FTP examples of PingER data. There are two types:

1. Raw data as gathered daily from all the monitoring hosts. This data is ie measured at 30 minute intervals and is quite dirty.
2. Analyzed data by metric. This has been cleaned up. Les recomends UFRJ uses the cleaned up data.

The instructions for the data will also be sent to Luiza

Les will also send Luiza information on how PingER data has been used.