Project title: Data Analytics of CernVM Server Logs
Supervisor: René Meusel, Gerardo Ganis

Project description:

CernVM [1] is a virtual appliance that is used by the four LHC experiments and the CERN theory group in order to run simulation and data processing applications in the Cloud. Every CernVM instance, be it a worker node in the cloud or a virtual machine on an end-user's laptop, requests configuration information from a central configuration service. The log data collected at this configuration service provide valuable information to us, for instance about used software versions, geographical distribution of virtual machines, virtual machine lifetime and potential problems.

In this project, the student develops a web application for log file analysis. The student is supposed to design and implement data collection and storage, feature extraction, and presentation of the log data. The student is supposed to familiarize himself/herself with modern log data analysis tools, such as the Hadoop file system, Hadoop-based query languages, or geospatial web services.

[1] http://cernvm.cern.ch/

Student profile: Computer scientist
Project content: 100% computing
Training value: The student will familiarize himself/herself with modern Big Data analysis tools. She/he will get hands-on experience in the development of a web-based data retrieval application. The student will learn how to extract meaningful features from log files.
Computing skills: Working knowledge of Linux and networking. Programming experience in C++, Java, Perl/Python. Experience with Hadoop and Hadoop-based tools is a plus.