Big Data - Jacek

I'd be happy to supervise one student if we find a good match. Damini Singh brought my attention because of the mention of Hadoop.

Project

We are working on optimizing the performance of Qserv, a distributed query handling system. We intend on running tests and creating empirical models for various aspects of it (disk I/O, memory, cpu, data distribution and hardware configurations etc). To do this, we have a computing cluster in France of ~50 nodes (more in the future) as an environment for our testing purposes. An important part of the process is instrumenting Qserv to log relevant measurement parameters for statistical analysis. A good project to work on will be the harvesting of these log files. In addition, there is a possibility for LSST projects in the database and web-app space too, like standalone MySQL tests. So, enthusiasm about log analysis/tools, LSST/astronomy/physics and databases/web-apps would steer the direction of the project proposal.

what skills do you need?

Skills needed

- Some C++ or Python

what kind of projects/tasks that you have might be suitable?

- It very much depends on the skills…
  - if the student is good with C++, than code tweaks, examples: migrating code to conform to C++ 11, or chasing and fixing a well defined problem
  - if the student is good with python, then running some tests. In our case most tests typically boil down to disk I/O rather than network I/O

any questions we should pose to them?

- It'd be good to see their resumes first. I'd be interested to know what they would like to work on if they had a choice of C++ related project, python related project, mysql related project and testing of a distributed server software

My top 3 choices are Damini, Jahin and Anwesha