Enterprise Platform and Integration Concepts:
Research Group of Prof. Dr. Hasso Plattner

Real-time Analysis of Genome Data

Motivation

The vision of the human genome project was born in the early 1980s. One decade later, it was officially started in the U.S. in 1990. Another decade later, a first draft of the humane genome was announced in 2000. In the same period costs for computer hardware dropped and capacities of main memory and storage systems underwent an exponential growth. Today, DNA sequencing and genome analysis are turned into reality. For example, malicious tissue from tumor patients is analyzed to derive concrete treatment decisions in course of personalized medicine. Suspects at crime scenes are identified by DNA profiling. Optimized crops are selected based on the results of their genetic analysis to improve harvests in agriculture worldwide. All examples have in common: Genome data is huge and its analysis takes days to weeks. For example, the humane genome consists of ~3.2 billion base pairs (= 3.2 GB) distributed across 23 chromosomes, building 20k-30k genes that code 50k-300k proteins.

Genome data is a specific subset of scientific data. Data management for scientific data comes with various challenges, such as huge storage requirements, traditional scanning algorithms are based on reading sequences of characters from files, processing of operational data in databases is only rarely considered, parallelization of processing, etc.

Goal

Building on our long-lasting experience in applying in-memory technology to selected enterprise challenges, we also focus on processing and analyzing of scientific data sets in real-time. In particular, the applicability of in-memory technology for analysis of genome data will be evaluated. Proof of concept prototypes will be engineered and showed to real-world users in the course of this project.

External Partner

The project team will have frequently contact with experts of our cooperation partners:

- Department of Pathology of the Charité - Universitätsmedizin Berlin, and
- SAP AG, Walldorf.

Thus, trips to the headquarters of either both companies are very likely.

Setting

The project team will work on latest server hardware, in-memory, and multi-core technology provided by the “Enterprise Application Architecture Laboratory” at our group and HPI’s “Future SOC Lab”. The laboratory builds the foundation for HPI’s in-memory technology activities. Due to our cooperations with hardware and software vendors, we are able to access high-end hard- and software before it is available for the public market. For example, SAP’s in-memory database “SAP HANA”, which is optimized for enterprise data management, will be used as technology foundation.

Skills

Our external partners provide the required real-world biological and technical input for this project. Thus, we expect you to work with interdisciplinary experts from our project partners.

Due to the expected intensive work with database technology, a passed exam in at least one database technology or equivalent lecture is favorable. Furthermore, knowledge in working with either or all of the following development languages is helpful: C++, Python, L, R, Bash, SQL.

In the course of the project, you will be equipped with knowledge about the foundation of in-memory technology and biological skills. We also provide introductions to further technologies, such as SQL, SQLScript, L, R, and BFL. However, it might be necessary to investigate additional technologies, which also requires you to deep into new areas by yourself or as team.

You will have frequent contact to experts from Charité and SAP to gain additional insights in their work.

Team Structure and Kickoff

The team will consist of 6-8 students. The project kickoff is scheduled for Mon, Oct 22, 2012.

Contact

Please feel free to contact us at “Hasso Plattner High Tech Park” at August-Bebel-Str. 88 or via e-mail.

- Prof. Dr. Hasso Plattner (office-epic@hpi.uni-potsdam.de), room V-2.13
- Matthieu Schapranow (schapranow@hpi.uni-potsdam.de), room V-0.01