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

Genome Data Analysis

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, sequencing and analysis of genome data 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. The humane genome, for example, consists of approx. 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, which is orders of magnitude larger than existing enterprise data. Data management for scientific data comes with various challenges, such as data proximity and a minimum of data transfer. Existing tools for genome data analysis are still based on file operations or single threaded execution.

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 potential end users in the course of this project.

External Partner

The project team will have frequently contact with experts of our cooperation partner SAP AG, Walldorf.

What we expect from you

We are looking for students, who are motivated to adapt to new research area, such as in-memory database technology and genomics. You should be hands-on experienced in using at least one programming/developing language, preferable C++ or Python, and one database query language, preferable SQL. Furthermore, a strong understanding of database concepts is beneficial. We expect you to have strong expertise in applying modeling techniques, such as UML or FMC, to exchange knowledge and design decisions. You should be flexible to work on top of existing tools and software and to extend it with your contributions. In addition, you should have communication abilities to collaborate with team and chair members as well as external cooperation partners.

What you can expect from us

We will provide you extensive introductions to the relevant fields of research, e.g. genomics, and with hands-on experiences, e.g. in in-memory database technology. For that, you will have access to latest server hardware. You will obtain insights in specific software development processes as well as project management and self-organization methods. Furthermore, you will interact with experts and partner in the corresponding fields.

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.

Contact

Please feel free to contact us, if you have any questions. You can reach us either in person at “Hasso Plattner High Tech Park”, August-Bebel-Str. 88 or via e-mail.

Your contact person is:

- Dr. Matthieu-P. Schapranow (schapranow@hpi.uni-potsdam.de), room V-1.01.