

Prof. Hasso Plattner



A Course in In-Memory Data Management

The Inner Mechanics
of In-Memory Databases

September 4, 2015

This learning material is part of the reading material for Prof. Plattner's online lecture "In-Memory Data Management" taking place at www.openHPI.de. If you have any questions or remarks regarding the online lecture or the reading material, please give us a note at openhpi-imdb@hpi.uni-potsdam.de. We are glad to further improve the material.

Research Group "Enterprise Platform and Integration Concepts",
<http://epic.hpi.uni-potsdam.de>

High Performance In-Memory Genome Data Analysis

Precision medicine aims at treating patients specifically based on individual dispositions, e.g., genetic or environmental factors [Jai09]. For that, researchers and physicians require a holistic view on all relevant patient specifics when making treatment decisions. Thus, the detailed acquisition of medical data is the foundation for personalized therapy decisions. The more fine-grained the available data is, the more specific the gained insights will be, but the complexity of data processing will rise as well. This requires tool support to identify the relevant portion of data out of the increasing amount of acquired diagnostic data [PS14].

Figure 35.1 depicts the genome data workflow in the course of precision medicine. After a sample has been acquired, it is sequenced, which results in short chunks of Deoxyribonucleic Acid (DNA) in digital form. The DNA chunks need to be aligned to reconstruct the whole genome and variants compared to a reference, e.g., normal vs. pathologic tissue, are detected during variant calling. The analysis of genome data builds on the list of detected variants, e.g., to identify driver mutations for a medical finding [B⁺10].

Figure 35.2 provides a comparison of costs for sequencing and main memory modules on a logarithmic scale. Both charts follow a steadily declining trend, which facilitates the increasing use of Next-Generation Sequencing (NGS) for whole genome sequencing and IMDB technology for data analysis. Latest NGS devices enable the processing of whole genome data within hours at reduced costs [Ans09]. The time consumed for sequencing is meanwhile a comparable small portion of the time consumed by the complete workflow. As a result, data processing and its analysis consume a significantly higher portion of the time and accelerating them will affect the overall workflow duration.

We focus on how to optimize the time-consuming data processing and analysis aspects of the workflow by combining latest software and hardware trends to create an integrated software system, which supports life science experts in their daily work.



Fig. 35.1: Data processing steps involved in the analysis of genome data. Sequencing the samples results in chunks of DNA available in digital form. During alignment their position within the whole genome is mapped. Variant calling results in a list of differences compared to a fixed reference. The analysis obtains new insights based on the list of detected variants.

Our High Performance In-memory Genome (HIG) system architecture modeled as Fundamental Modeling Concept (FMC) block diagram is depicted in Figure 35.3 [KGT06]. It combines data from various data sources, such as patient-specific data, genome data, and annotation data, within a single system to enable flexible real-time analysis and combination. In the following, the system layers are described in further detail.

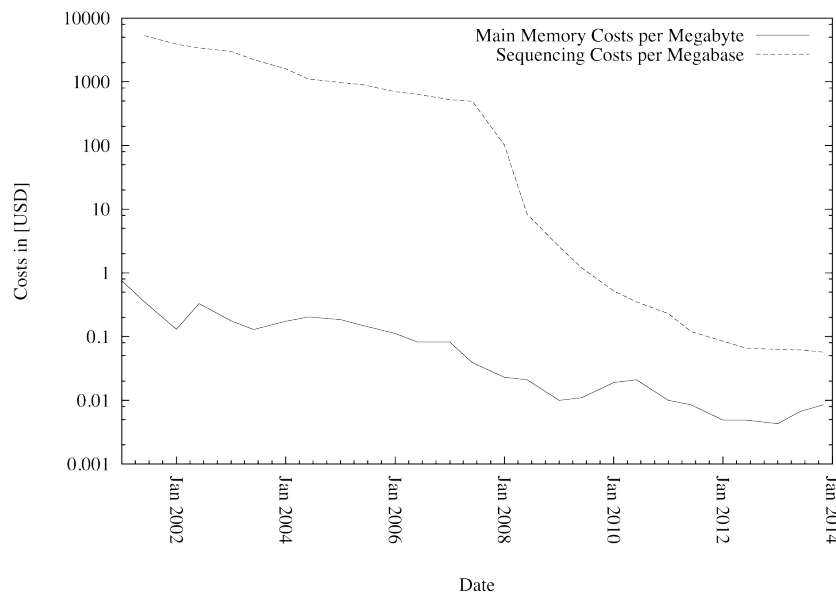


Fig. 35.2: Costs for next-generation sequencing and main memory 2001-2014 adapted from [Nat13, McC13].

35.1 Application Layer

The application layer consists of special purpose applications to answer medical and research questions [PS14]¹. We provide an Application Programming Interface (API) that can be consumed by various kinds of applications, such as web browser or mobile applications. Figure 35.3 depicts the data exchange via asynchronous Ajax calls and JavaScript Object Notation (JSON) [Hol08, Cro06]. As a result, performing specific analyses is no longer limited to a specific location, e.g., the desktop computer of a clinician. Instead, all applications can be accessed via devices connected to the Internet, e.g., laptop, mobile phone, or tablet computer. Thus, having access to relevant data at any time enhances the user's productivity. The end user can access these cloud applications via any Internet browser after registration. Selected cloud applications are our cohort analysis and clinical trials search, which are described in further detail in Section 35.1.2 and Section 35.1.1 respectively.

¹ You can access our cloud services online at <http://www.analyzegenomes.com>

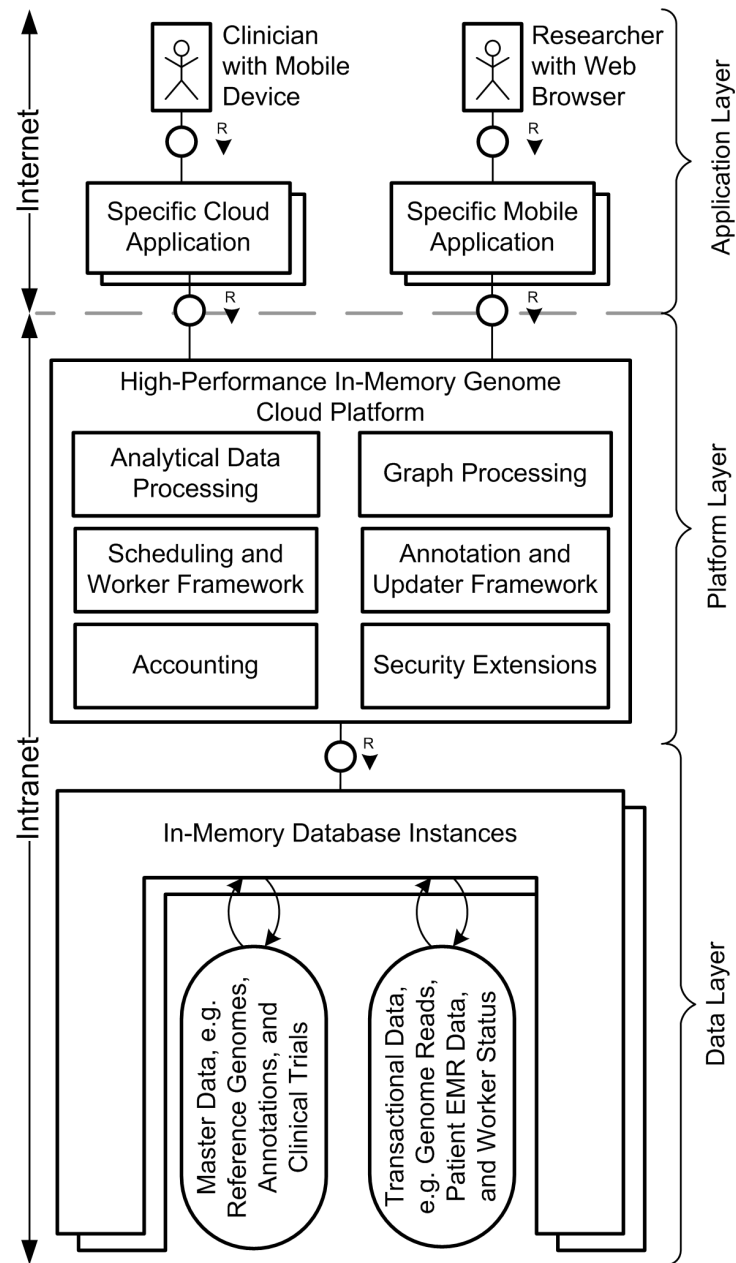


Fig. 35.3: Our system architecture consist of application, platform, and data layer. Analysis and processing of data is performed in the platform layer, which eliminates time-consuming data transfer.

35.1.1 Clinical Trials Application



Fig. 35.4: The patient-specific clinical trials search results based on the individual anamnesis of a patient. Relevant entities from the free-text description of a clinical trial are extracted with the help of specific text-mining rules.

Our clinical trials search assists physicians in finding adequate clinical trials for their patients. It analyses patient data, such as age, gender, pre-conditions and existing mutations, and matches them with clinical trials descriptions from clinicaltrials.gov [U.S13a]. Our analysis incorporates more than 130,000 clinical trial descriptions, which are processed and ranked in real-time accordingly to the personal anamnesis of each individual patient. The ranked results are summarized on a single screen and provided to the researcher as depicted in Figure 35.4.

The clinical trials search incorporates the extraction of entities and features from the textual description. For that, we assembled a set of customized vocabulary. For example, we use a vocabulary set for human gene identifiers with more than 120,000 gene names and synonyms and a vocabulary set for pharmaceutical ingredients with more than 7,000 pharmaceutical ingredients. Our ontologies incorporate a set of standardized vocabularies, e.g., the Metathesaurus Structured Product Labels (MTHSPL) of the Unified Medical Language System (UMLS) [U.S13b].

35.1.2 Patient Cohort Analysis

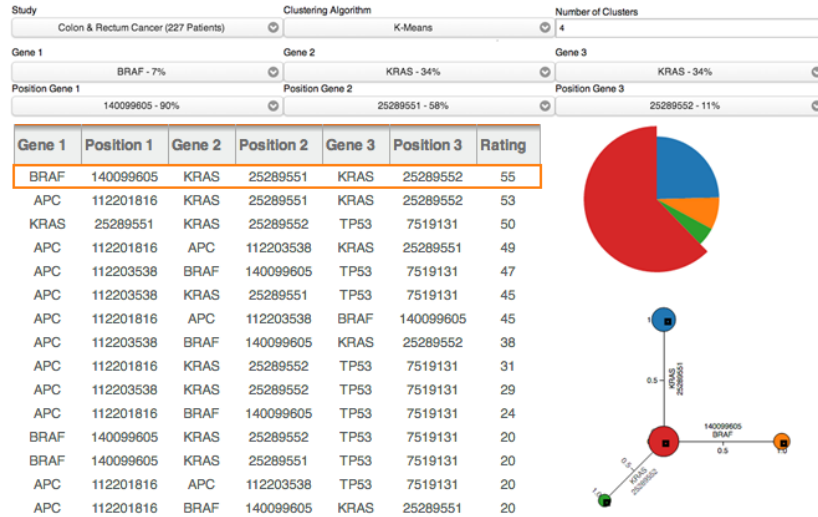


Fig. 35.5: The results of an interactive analysis of a cohort of colon carcinoma patients using k-means clustering. It proposes relevant combinations of genomic loci, such as gene KRAS on chromosome 12 at position 25,289,551, which are present in the majority of cohort members as depicted by the pie chart on the right.

Our cohort analysis application enables researchers and clinicians to perform interactive clustering on the data stored in the IMDB, e.g., k-means and hierarchical clustering [Kra09, Chap. 13]. Thus, researchers and clinicians are able to verify hypotheses by combining patient and genome data in real-time. Therefore, they use patient-specific and genome data loaded into the IMDB and perform the interactive cohort analysis as part of the platform layer as depicted in Figure 35.3.

35.2 Platform Layer

The platform layer holds the complete process logic and consists of the IMDB system for enabling real-time analysis of genomic data. We developed specific extensions for the IMDB system to support genome data processing and its analysis. In the following, selected extensions and their integration in the IMDB system are described in more detail.

35.2.1 Scheduling of Data Processing

We extended the IMDB by a worker framework, which executes tasks asynchronously, e.g., alignment of chunks of genome data. The framework consists of a task scheduler instance and a number of workers controlling dedicated computing resources, e.g., individual computing nodes. Workers retrieve tasks and parameters by the scheduler instance and perform specific tasks, such as workbench preparation, task execution, and maintenance of status information. Thus, all workers are connected to the IMDB to store status information about currently executed tasks.

Furthermore, the scheduler supervises the responsiveness of individual compute resources. If a predefined response behavior is no longer guaranteed, e.g., due to an overloaded compute node or a crashed worker process, workers are marked as unresponsive. As a result, work-in-progress tasks of the unresponsive worker are reassigned to a new worker and this worker is scheduled for a restart.

35.2.2 Updater Framework

We consider the use of latest international research results as enabler for evidence-based therapy decisions [SHP13]. The updater framework is the basis for combining international research results. It periodically checks all registered Internet sources, such as public FTP servers or web sites, for updated and newly added versions of annotations, e.g., database exports as dumps or characteristic file formats, such as CSV, TSV, and VCF. If the online version is newer than the locally available version, the new data is automatically downloaded and imported in the IMDB to extend the knowledge base.

The import of new versions of research databases is performed as a background job without affecting the system's operation. We import new data without any data transformations in advance. Thus, it is available for real-time analysis without any delay [BSP12, FCP⁺12]. For example, the following selected research databases are checked regularly by our updater framework: National Center for Biotechnology Information (NCBI), Sanger's catalogue of somatic mutations in cancer, University of California, Santa Cruz (UCSC) [Nat, F⁺10, M⁺12].

35.3 Data Layer

The data layer holds all required data for performing processing and analyzing of genomic data. The data can be distinguished in the two categories

of master and transactional data [DM11]. For example, human reference genomes and annotation data are referred to as master data, whereas patient-specific NGS data and Electronic Medical Records (EMR) are referred to as transactional data [The, RNK91]. Its analysis is the basis for gathering specific insights, e.g., individual genetic dispositions, and for leveraging personalized treatment decision in the course of personalized medicine [Jai09].

35.4 Summary

Nowadays, a range of time-consuming tasks has to be accomplished before researchers and clinicians can work with analysis results, e.g., to gain new insights. In the following, selected tasks are outlined:

- Genome data of patients has to be analyzed in customized workflows, e.g., to identify genetic variants relevant for a certain disease,
- Researchers and clinicians have to gather and combine relevant information from distributed, heterogenous data sources and link them to patient data, and
- Researchers and clinicians need to apply gained insights on their patient data, e.g., to formulate new hypotheses and to verify them.

Nowadays, almost all aforementioned tasks have to be carried out manually. As a result, it takes a considerable amount of time until an optimal treatment decision can be made.

With our HIG platform, we redefine and accelerate the process of finding treatment decisions. We provide the infrastructure and computational resources to run configurable genome data analysis workflows. Furthermore, we have created a knowledge base by integrating content from various heterogenous data sources, which are updated constantly to always provide latest findings from research. With the capabilities of the applied in-memory technology, we can immediately combine tremendous amounts of data, taking also advantage of the whole spectrum of analysis features provided, e.g., for textual analysis. With our innovative approach, gaining new insights on patient data shifts from a time consuming, manual process to immediate, interactive data exploration.

As a result, researchers and clinicians can spend additional time on formulating hypotheses, their verification, and drawing corresponding conclusions to come up with more accurate treatment decisions.

35.5 References

- [Ans09] Wilhelm J Ansorge. Next-Generation DNA Sequencing Techniques. *New Biotechnology*, 25(4):195–203, 2009.
- [B⁺10] I Bozic et al. Accumulation of Driver and Passenger Mutations during Tumor Progression. *Proceedings of the National Academy of Sciences of the United States of America*, 107(43):18545–50, Oct. 2010.
- [BSP12] Anja Bog, Kai Sachs, and Hasso Plattner. Interactive Performance Monitoring of a Composite OLTP and OLAP Workload. In *Proceedings of the Int'l Conf on Mgmt of Data 2012*, pages 645–648, Scottsdale, AZ, USA, 2012. ACM.
- [Cro06] D. Crockford. RFC4627: The application/json Media Type for JavaScript Object Notation (JSON). <http://www.ietf.org/rfc/rfc4627.txt> [retrieved: Sep3, 2012], July 2006.
- [DM11] Tapan Kumar Das and Manas Ranjan Mishra. A Study on Challenges and Opportunities in Master Data Management. *Int'l Journal of Database Mgmt Syst*, 3(2), May 2011.
- [F⁺10] S. A. Forbes et al. The Catalogue of Somatic Mutations in Cancer: A Resource to Investigate Acquired Mutations in Human Cancer. *Nucl Acids Res*, 38, 2010.
- [FCP⁺12] Franz Färber, Sang Kyun Cha, Jürgen Primsch, Christof Bornhövd, Stefan Sigg, and Wolfgang Lehner. SAP HANA database: Data Management for Modern Business Applications. *SIGMOD Rec.*, 40(4):45–51, January 2012.
- [Hol08] Anthony T. Holdener. *AJAX: The Definitive Guide*. O'Reilly, 1st edition, 2008.
- [Jai09] K.K. Jain. *Textbook of Pers. Medicine*. Springer, 2009.
- [KGT06] A. Knopfel, B. Grone, and P. Tabeling. *Fundamental Modeling Concepts: Effective Communication of IT Systems*. John Wiley & Sons, 2006.
- [Kra09] S.A. Krawetz. *Bioinformatics for Systems Biology*. Humana Press, 2009.
- [M⁺12] L. R. Meyer et al. The UCSC Genome Browser Database: Extensions and Updates 2013. *Nucl Acids Res*, 2012.
- [McC13] John C. McCallum. Memory Prices (1957-2013). <http://www.jcmit.com/memoryprice.htm> [retrieved: Feb 11, 2014], Feb 2013.
- [Nat] National Center for Biotechnology Information. All Resources. <http://www.ncbi.nlm.nih.gov/guide/all/> [retrieved: Jan 5, 2013].
- [Nat13] National Human Genome Research Institute. DNA Sequencing Costs. <http://www.genome.gov/sequencingcosts/> [retrieved: Feb 11, 2014], Apr 2013.
- [PS14] Hasso Plattner and Matthieu-P. Schapranow, editors. *High-Performance In-Memory Genome Data Analysis: How In-Memory*

- Database Technology Accelerates Personalized Medicine*. Springer-Verlag, 2014.
- [RNK91] A.L. Rector, W.A. Nolan, and S. Kay. Foundations for an Electronic Medical Record. *Methods of Information in Medicine*, pages 179–186, 1991.
- [SHP13] Matthieu-P. Schapranow, Franziska Häger, and Hasso Plattner. High-Performance In-Memory Genome Project: A Platform for Integrated Real-Time Genome Data Analysis. In *Proceedings of the 2nd International Conference on Global Health Challenges*, pages 5–10. IARIA, Nov 2013.
- [The] The Genome Reference Consortium. Genome Assemblies. <http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/data.shtml> [retrieved: Jan 5, 2013].
- [U.S13a] U.S. National Institutes of Health. Clinicaltrials.gov. <http://www.clinicaltrials.gov/> [retrieved: Jul 17, 2013], 2013.
- [U.S13b] U.S. National Library of Medicine. Unified Medical Language System (UMLS). <http://www.nlm.nih.gov/research/umls/> [retrieved: Sep 19, 2013], Jul 2013.