Copyright: Andreas Heddergott/TUM

Business Challenge

Complex diseases like cancer require highly targeted treatments. How could ProteomicsDB help researchers better understand biochemical processes and gain new therapeutic insights?

Transformation

To deliver faster access to more-detailed data, the team migrated ProteomicsDB, which is based on SAP HANA®, to an IBM® Power Systems™ server, hosted at the SAP® University Competence Center at TUM.

Business benefits:

>1,000
published academic research papers reference ProteomicsDB

243
cancer drugs analyzed to enable personalized treatment in the future

15,721
proteins and 80% of the human proteome quantified and made accessible online

ProteomicsDB

Technical University of Munich generates new protein insights to help fight disease

The ProteomicsDB is the flagship bioinformatics project of the Chair of Proteomics and Bioanalytics at the TUM School of Life Sciences Weihenstephan. The leading interdisciplinary research group at Technical University of Munich (TUM) is made up of around 30 people with backgrounds in cell biology, chemistry, biochemistry, mass spectrometry and bioinformatics. The TUM has 14 departments, more than 40,000 students, and is one of Europe's top universities as well as one of the first universities in Germany to be named a University of Excellence.

"We are proud that a publication based on the ProteomicsDB powered by SAP HANA running on IBM Power Systems has been recommended by the prestigious F1000Prime service."

Prof. Dr. Bernhard Küster, Head of the

Chair of Proteomics and Bioanalytics, TUM

School of Life Sciences Weihenstephan,

Technical University of Munich

Share this

Supporting breakthroughs in advanced medical research

Improving global health is the ambitious goal of life sciences research. One of the prime areas for attention is deeper insight into biochemical processes within the human body, looking at the structure and functional activities of proteins. By collecting and analyzing complex data, researchers and pharmaceuticals companies can uncover new insights from existing experimental results, and potentially transform our understanding of these vital molecules.

The interdisciplinary team at the Chair of Proteomics and Bioanalytics at the Technical University of Munich (TUM) School of Life Sciences Weihenstephan is unique. This team comprises leading international researchers from a wide range of fields, including cell biology, chemistry, biochemistry, mass spectrometry and bioinformatics.

To support the protein research area, the Chair of Proteomics and Bioanalytics launched its ProteomicsDB, a free online database designed to enable researchers to search and visualize data about proteins. In addition, the Chair is involved in the Excellence Cluster Centre for Integrated Protein Science Munich (CIPSM) and the German Cancer Consortium DKTK, which is part of the German Cancer Research Center (DKFZ).

Prof. Dr. Bernhard Küster, Head of the Chair of Proteomics and Bioanalytics at the TUM School of Life Sciences

Weihenstephan, explains: "To treat complex diseases such as cancer, we need personalized medication and therapies. The problem is that we often do not yet have the level of understanding that would enable us to customize treatments to make them more effective. In our research, we often use just one percent of the data we generate. With our ProteomicsDB, we want to increase that percentage substantially, thereby allowing our researchers to make better use of the data they produce in countless experiments.

"Our challenge is to map out all proteins in the human body, and create the 'human proteome' in the same way that researchers have mapped out the human genome. Proteins can be found in almost 250 different cell types, and we have already identified and qualitatively

and quantitatively examined around 80 percent of the 20,000 proteins in the human body. Once we get a better idea of the relationships between the proteins and how they interact with various drugs, our research could pave the way to precision medicine that is tailored to individual patients."

Dr. Mathias Wilhelm, Group Leader Bioinformatics at the Chair of Proteomics and Bioanalytics at the TUM School of Life Sciences Weihenstephan, adds: "Proteome data is highly complex and multidimensional, and the data structures need to reflect the interactions we observe in the human body. To make all this information usable, we want to allow researchers easy and fast access to these large amounts of data."

Challenges of scale and complexity

In many ways, the ProteomicsDB team faces typical business challenges: handling very large data sets and looking for patterns that might provide new insights. To make discoveries, researchers must be able to freely explore and analyze relationships without constraints. However, to achieve high-speed analysis, traditional database designs require pre-prepared data aggregations, which are based on assumptions about which data to include.

That's the core of the problem: how can you deliver high-speed free-form analysis without building and maintaining aggregates that make assumptions about what researchers are looking for?

To create a unique platform to store, analyze and visualize the proteomics data without constraints, the Chair of Proteomics and Bioanalytics worked closely with SAP, exploring the capabilities of the SAP HANA database. SAP HANA in-memory technology allows the ProteomicsDB team to implement a simpler database design without precomputed data aggregations. Thanks to its architecture, SAP HANA offers researchers quick insights and flexibility, which they need to advance understanding of biochemical processes and drive innovation.

Prof. Dr. Bernhard Küster continues, "Without an in-memory database like SAP HANA, we would probably not have focused on proteome analytics to build a public database. SAP HANA offers the robust, versatile, high-speed processing capabilities that allows the team to use standard relational database features as well as graph database modelling to incorporate many different types of data and develop a wide variety of features.

"Working together with the SAP University Competence Center at TUM opened up new research opportunities. Gaining full control over every aspect of the ProteomicsDB would enable closer, interdisciplinary research between my team and our fellow researchers at the Chair for Information Systems."

Dr. Mathias Wilhelm adds: "With SAP HANA, we found the ideal tool to build up our protein-centric database. The SAP HANA in-memory technology enables scientists around the world to interact efficiently with large collections of proteomics data."

Combining SAP HANA and IBM Power Systems to advance data-driven insights

Initially, the SAP team operated the ProteomicsDB on an x86 server cluster with two nodes. As the data volume and complexity grew, the existing infrastructure could no longer keep up with the increasing performance requirements, and the team considered its future platform options. Knowing that data volume would continue to grow, being able to scale the system continuously was a major priority.

The ProteomicsDB team evaluated different options and joined forces with the SAP University Competence Center (SAP UCC) at TUM. The SAP UCC at TUM is one of six SAP UCCs worldwide and run by the Chair for Information Systems at the Department of Informatics at TUM. The goal of the SAP UCC at TUM is to support teaching of students and pupils, to enhance SAP skills of lecturers and to support research in the form of PhD theses, master theses, bachelor theses, seminars, and projects using SAP software.

Funding to expand the research capabilities was granted by the German Research Foundation DFG, and the team planned the transition to a new IBM Power Systems server and to IBM Storwize® storage solutions hosted at the SAP UCC at TUM.

The research team, IBM and IBM Business Partner Axians deployed the new infrastructure. Then the TUM team and IBM migrated the ProteomicsDB to the latest release of SAP HANA and moved the database to the newly installed, dedicated IBM Power Systems server. The IBM Power Systems E870C contains 40 POWER8 processor cores and 6 TB main memory, and runs the SUSE Linux Enterprise Server for SAP Applications on IBM Power Systems operating system.

Data storage for the ProteomicsDB is provided by four IBM Storwize V5020 solutions. The storage systems are equipped with super-fast flash storage modules, and the integrated IBM Spectrum Virtualize™ software streamlines storage management and provides the flexibility to expand capacity easily and rapidly when needed.

Benefits in detail

- Increased system flexibility through use of advanced virtualization

- Faster response to research enquiries based on enhanced system performance

- Manages 8.85 TB of proteome data in the SAP HANA database

- Provides access to more than 43m records of quantitative mass spectrometry-based proteomics data

- Reduces downtime and provides continuous availability to researchers

- Improves performance, enabling more rapid data analytics and visualization with 75% fewer processors

- Larger data analysis capacity with memory capacity increased from 2 TB to 6 TB

- Serves more than 800 registered users with data access via API

- Improves response times and accelerates development cycles by using virtualization to optimize system capacity

The fully virtualized environment takes advantage of IBM PowerVM® technology to provide separate SAP HANA instances for production, quality assurance and development. In addition, the team uses SAP HANA multi-tenant database containers to use its resources as efficiently as possible.

To support highly specialized statistics and analytics workloads written in the open source R programing language, the combined IBM and TUM team integrated the analytics processing seamlessly with the ProteomicsDB SAP HANA environment.

Dr. Harald Kienegger, Managing Director of the SAP UCC at TUM, explains: "The SAP University Competence Center has worked closely with SAP and IBM for many years, building deep experience of running all kinds of educational SAP systems reliably and at scale for universities all around the world. Our highly automated infrastructure management based on IBM Power Systems put us in the perfect position to operate the SAP HANA business data platform for the ProteomicsDB.

"By hosting the ProteomicsDB at the SAP UCC, we also improved the reliability of the solution. At any time, we can fail over from the dedicated IBM Power Systems server to our shared IBM Power Systems infrastructure and keep the ProteomicsDB available for researchers at all times."

"For us, SAP HANA on IBM Power Systems delivers more control and achieves greater than 99.9 percent system availability."

Dr. Harald Kienegger

Managing Director of the SAP University Competence Center, Technical University of Munich

Advancing science through technology

Today, the ProteomicsDB has almost 800 registered users able to access the complex data via a web-based API. In addition, most of the 8.85 TB of data stored in the SAP HANA database is available openly to the public, without registration. The ProteomicsDB already offers data on 15,721 proteins, approximately 80 percent of the human proteome.

The increase in memory and performance provided by the IBM Power Systems server allows researchers around the world to quickly and flexibly explore and test new hypotheses without delay, and continuously discover new aspects of the data.

The migration from a scale-out x86 cluster to an IBM Power Systems scale-up configuration has also unlocked additional technical and practical benefits, with improved availability and more-rapid data analytics with 75 percent fewer processors.

Dr. Harald Kienegger states: "For us, SAP HANA on IBM Power Systems delivers more control and achieves greater than 99.9 percent system availability. At the same time, hosting the systems at TUM enables us to ramp up research resources for the development of the ProteomicsDB. For example, we previously could not give bachelor and master students direct access to the underlying systems. The IBM Power Systems platform allows us to include those research requests, which will help us offer even better services and advance the capabilities of this unique research platform."

Key components

Applications: SAP HANA®

Software: IBM® PowerVM®, IBM Spectrum Virtualize™, SUSE Linux Enterprise Server for SAP Applications

Hardware: IBM Storwize® V5020, IBM Power Systems™ E870C

Services: IBM Business Partner Axians, IBM Services

Dr. Mathias Wilhelm remarks: "The ProteomicsDB now includes more than 43 million quantitative mass spectrometry-based proteomics records. We are very satisfied with the performance – our users always have access to the data with fast response times. Furthermore, the new solution helps us speed up development cycles and implement new features for our users much faster than before.

"By moving to SAP HANA on IBM Power Systems we increased the available memory by a factor of three, from 2 TB to 6 TB. This gives us the headroom we need to add more data to the ProteomicsDB – our current plan is to double the data volume this year. As we extend beyond human proteome data to include flora and fauna, the increase in memory will allow us to run complex analytics which could uncover connections between species, further supporting fundamental proteomics research.

"The ProteomicsDB definitely is faster since we moved to SAP HANA on IBM Power Systems. The rich user interface enables scientists to tap into a wealth of data and gain new insights by adjusting different parameters and evaluating the changes immediately. Since it first went online, the ProteomicsDB has established itself as highly valued and reliable data source for academic research.

"Today, already more than 1,000 academic publications have reference or used ProteomicsDB information. This includes some important discoveries, such as a recent report that described the use of machine learning to analyze anonymized patient information and specific cell lines representative of specific tumor subtypes. The results identified completely new correlations that enable predictions of drug sensitivity for cell lines and patients."

Improving healthcare with evidence-based insights

The new insights gained from the ProteomicsDB also are being used in cancer treatments in hospitals across Munich. In regular meetings with leading doctors, the team advises on targeted, individualized treatments in a 'Molecular Tumor Board'. Based on the data available in the ProteomicsDB, the team can recommend the most suitable treatment options and provide evidence for the different alternatives and their effectiveness under given conditions. In the future, the Chair of Proteomics and Bioanalytics would like to offer this healthcare methodology to all cancer patients in Germany, to strengthen the application of truly evidence-based treatments across the country.

Prof. Dr. Bernhard Küster expands: "In one of our latest studies, published in the highly-respected 'Science' journal, we illustrated how the information in our ProteomicsDB can be used in basic research, drug discovery, and clinical decision-making. We are just at the beginning of the development of personalized medicine and molecularly targeted drugs that can revolutionize the treatment of cancer.

"We are proud that this publication based on the ProteomicsDB powered by SAP HANA running on IBM Power Systems has been recommended by the prestigious F1000Prime service, which focuses on leading biology and medical research publications and working with scientists and clinicians from around the world."

"With our ProteomicsDB based on SAP HANA running on IBM Power Systems, we have the perfect platform to help medical scientists around the world to create more-effective treatment options – to save lives in the future."

Prof. Dr. Bernhard Küster

Head of the Chair of Proteomics and Bioanalytics TUM School of Life Sciences Weihenstephan, Technical University of Munich

Focusing on AI to find more life-saving answers

The TUM team also focuses on technical innovation to boost medical research, and plans to integrate the ProteomicsDB with NVIDIA accelerator cards, using NVIDIA NVLink Technology. The IBM POWER architecture allows the addition of Graphics Processing Units (GPUs) resources, eliminating the barriers between CPU and GPU, which will help to speed up big data processing and accelerate advanced analytics such as machine learning and artificial intelligence (AI).

As a next step, the team is evaluating the use of IBM POWER9™ processor cores to boost performance further, in combination with NVIDIA NVLink.

To achieve additional performance improvements, the team will migrate existing data analytics processing to the SAP HANA Predictive Analysis Library (PAL). Moving machine learning workloads directly into SAP HANA will help to reduce overhead and to streamline and accelerate data processing.

In the longer term, the team is considering replicating its success, by encouraging other research groups to use the same technical platform for their own data projects, leading to new insights beyond the scope of ProteomicsDB.

Dr. Harald Kienegger explains, "We are developing ideas and concepts to offer 'ProteomicsDB as a Service' run by the SAP UCC at TUM, enabled by the IBM Power Systems platform.

"For example, researchers in fields such as other animals and plants could use the platform capabilities to analyze their specific data sets. If we then consider the opportunities to add deep machine learning capabilities and artificial intelligence to make new connections, the opportunities are huge. Modern machine learning algorithms enable discovery of entirely new insights that no human could possibly come up with. This cutting-edge technology will also enable us to analyze and optimize existing models to increase the efficiency of medical research.

"Finally, adding 'ProteomicsDB as a Service' could help us to expand our business model, and the revenue and shared expertise would help us to provide even better service."

Prof. Dr. Bernhard Küster concludes: "We believe that the next years in life sciences research will be all about data analytics. Following the boom in genome data, activity in proteome research is increasing quickly. Only a truly professional IT solution can help us advance this kind of data-driven, world-class research.

"With our ProteomicsDB based on SAP HANA running on IBM Power Systems, we have the perfect platform to help medical scientists around the world to understand the mechanics of cancer drugs and create more-effective treatment options – to save lives in the future."

Learn more, connect with IBM [Twitter] [YouTube] **and SAP** [Twitter] [YouTube]