Solution Brief

Accelerate Insights with SAP HANA® Machine Learning and Intel® Processors

SAP HANA in-database machine learning capabilities on hardware powered by Intel® Xeon® processors provide a foundation to take machine learning into critical, high-performance production environments

In the data-driven economy, time is money—and latency costs time. So how much is latency costing your company? Not just network or storage latency—the latency between insight and execution, and between prediction and action. It's probably costing you more than you'd like, which is where Intel® Xeon® Scalable processors and SAP HANA® in-database machine learning can help.

Machine learning and predictive analytics improve business processes by making them more intelligent. This intelligence gives companies the ability to anticipate business outcomes. However, in order for machine learning to become part of a business process, it needs to operate with minimal execution latency and in real time, regardless of the size of the data or the complexity of the predictive model. Only once these conditions are met is it possible for users and machines to take action as events happen—not just after they occur.

Running machine learning algorithms where the data resides—in the database—eliminates latency and other delays that can happen when copying data to another server in order to run machine learning algorithms on it. Not only does in-database machine learning help ensure that model training and predictions are fast and decisions are timely, but it also enables you to use data gravity to your advantage. Because in-database machine learning helps avoid the necessity of moving data, it keeps that data near all the apps that build value on top of it. And a powerful way to bring in-database machine learning into production is with the SAP HANA platform and Intel Xeon Scalable processors.

Accelerate Machine Learning with the SAP HANA Platform

The SAP HANA platform provides in-database machine learning capabilities for data scientists, data analysts, and developers. The SAP HANA platform provides:

- Predictive analytics features, including:
  - The Automated Predictive Library (APL), targeted primarily at business and data analysts
  - The Predictive Analysis Library (PAL), an application function library tailored for use by data scientists
  - R* and TensorFlow* integration—enabling integration of models in these popular open source frameworks with the SAP HANA platform
- High-performance machine learning features, including:
  - Real-time scoring
  - SAP HANA streaming analytics—for analytics on event streams

This paper focuses primarily on specific aspects of these technologies. For an overview of machine learning in the SAP HANA platform, refer to *sap.com/products/hana/features/advanced-analytics.html*.

Machine Learning Applications

Machine learning uses statistical relationships in historical data to predict the likelihood of future events. It enables the development of predictive models that do not require explicit coding. Models developed this way capture fundamental decision logic and enable applications to adapt. As models are retrained with new data, the decision logic automatically updates and becomes more current. This adaptability enables new classes of applications to be developed whose predictions can be constantly fine-tuned by adjusting the underlying models, without having to rewrite code.

Machine learning is transformative in a variety of use cases. Intel predicts that the following industry segments will lead machine learning adoption through 2020:[1]

Consumer

- Smart assistants
- Chatbots
- Search personalization
- Augmented reality
- Robots

Health

- Enhanced diagnostics
- Drug discovery
- Patient care
- Research
- Sensory aids

Finance

- Algorithmic trading
- Fraud detection
- Research
- Personal finance
- Risk mitigation

Retail

- Customer support
- Experience
- Marketing
- Merchandising
- Customer loyalty
- Supply chain
- Security

Government

- Defense
- Data insights
- Safety and security
- Resident engagement
- Smarter cities

Energy

- Oil and gas exploration
- Smart grid
- Operational improvement
- Conservation

Transportation

- Autonomous cars
- Automated trucking
- Aerospace
- Shipping
- Search and rescue

Industrial

- Factory automation
- Predictive maintenance
- Precision agriculture
- Field automation

Other

- Advertising
- Education
- Gaming
- Professional and IT services
- Telecommunications/media
- Sports

Predictive Analytics Features in SAP HANA

SAP® Predictive Analytics primarily benefits business and data analysts, enabling them to easily find answers to business questions without requiring extensive training or data-science skills. However, it also has an "expert" mode for users who want a more hands-on predictive experience, such as data scientists.

At the core of SAP Predictive Analytics are proprietary automated machine learning algorithms designed to make predictive technologies more accessible and easier to use. The solution's wizard-driven interface guides users through the predictive modeling process of data preparation, model creation, and model deployment.

SAP Predictive Analytics is optimized for the SAP HANA platform and can support huge data volumes and in-memory processing. Users can connect to the SAP HANA platform in either an automated mode that connects to APL or in expert mode, which connects to PAL. Both the automated and expert modes run within the SAP HANA platform. This means that users do not need to extract the data and perform calculations on their desktops or move data to run predictive workflows; only the predictive results are sent back to a desktop client after processing.

The automated analytics interface within SAP Predictive Analytics provides data analysts and developers with automated machine learning capabilities, and it can be used to create predictive models without requiring data-science experience. SAP Predictive Analytics provides a number of machine learning functions, including:

- Classification
- Regression
- Clustering
- Time series
- Key influencers
- Recommendations
- Link analysis

SAP Predictive Analytics does not require complex predictive models as input. It simply needs to be configured and told what type of data-mining function needs to be applied to the data set; users set the parameters for analysis so that the system can train on an input data set. In addition, SAP Predictive Analytics composes its own models

using sophisticated techniques for automated machine learning; it also creates (and selectively eliminates) metadata as required to create optimal, algorithmically robust models. Users can then apply these models to a target data set and use the results from the predictive calculations in their applications or their analysis. The ability to make use of the predictive modeling power without expertise in mathematics or data science means that predictive capabilities can be put into the hands of more users. This, in turn, increases the value of the data platform as it creates additional predictive insights with minimal effort.

Beyond model creation, the SAP® Predictive Factory within SAP Predictive Analytics provides automatic performance tuning capabilities. This helps ensure that models are tuned to operate at peak performance even as the data used by those models changes with time. Predictive Factory features a browser-based, single-sign-on environment and a user friendly scheduling interface designed for data analysts. As a result, users can try a variety of scenarios, incorporating incremental improvements that they discover back into models in real time. Users can also schedule model refreshes, manage models by exception, and deploy scores instantly to use results in applications and real-time analysis. For more information about SAP Predictive Analytics and Predictive Factory, see *http://help.sap.com/pa*.

SAP HANA for Data Scientists

For data scientists, the SAP HANA platform provides the PAL and integration with R and TensorFlow. Moreover, data scientists can deploy machine learning models created in PAL—not only on data at rest, but also on SAP HANA streaming analytics data—in order to harness real-time insights from live data in motion. And custom logic for processing event streams can also be developed and deployed in the SAP HANA streaming analytics engine.

The SAP HANA platform provides a variety of capabilities for data scientists to implement machine learning in production environments. The SAP HANA PAL contains more than 90 machine learning algorithms in the following categories:

- Classification
- Clustering
- Regression
- Time-series forecasting
- Outlier detection
- Recommender systems

Training models to build with the SAP HANA PAL is straightforward and requires only three primary data structures:

1. An input table containing the data to be used for training
2. A parameter table that contains information specific to the model being trained, such as the number of trees desired in the model, tree depth, or stopping criteria in the case of the random decision trees, and general information such as the number of threads to use
3. Output tables to store the results of each execution, such as model performance

Analytical models developed with the SAP HANA PAL can be consumed in a variety of ways. For example, the models can be invoked from SQL-based applications or client-side code written in Python* and other programming languages. Applications developed on SAP HANA XS advanced (XSA) application server can then consume these machine learning models.

Deploying SAP HANA® APL Models

The Automated Predictive Library (APL) running within the SAP HANA platform provides high performance. Once trained, users can deploy APL models in two ways:

1. The scoring equation of a classification, regression, or clustering model can be generated in the code-type suited for a given production environment, including:
   - SAP HANA SQL statements
   - SAP HANA User-Defined Functions (UDFs)
   - Continuous Computation Language (CCL)
   - Java*

   Models deployed in this manner execute where the data resides.

2. A model is applied in the database to produce prediction outputs that then persist in database tables.

In situations involving partitioned tables, APL models can be applied and run in parallel against a partitioned table that is distributed across multiple nodes. Ensuring the highest level of performance from these libraries is critical for allowing the model training and scoring to happen quickly, in line with business processes, and in real time.

The SAP HANA platform can also use a variety of third-party machine learning frameworks and languages. For example, data scientists might need to use the extensive collections of models and scripts already developed in R, a popular, open source statistical programming language. For situations like this, users can write R code in the SAP HANA platform, and they can mix and match the algorithms with the SAP HANA PAL. The R code is executed externally on a separate Rserve* (Figure 1).

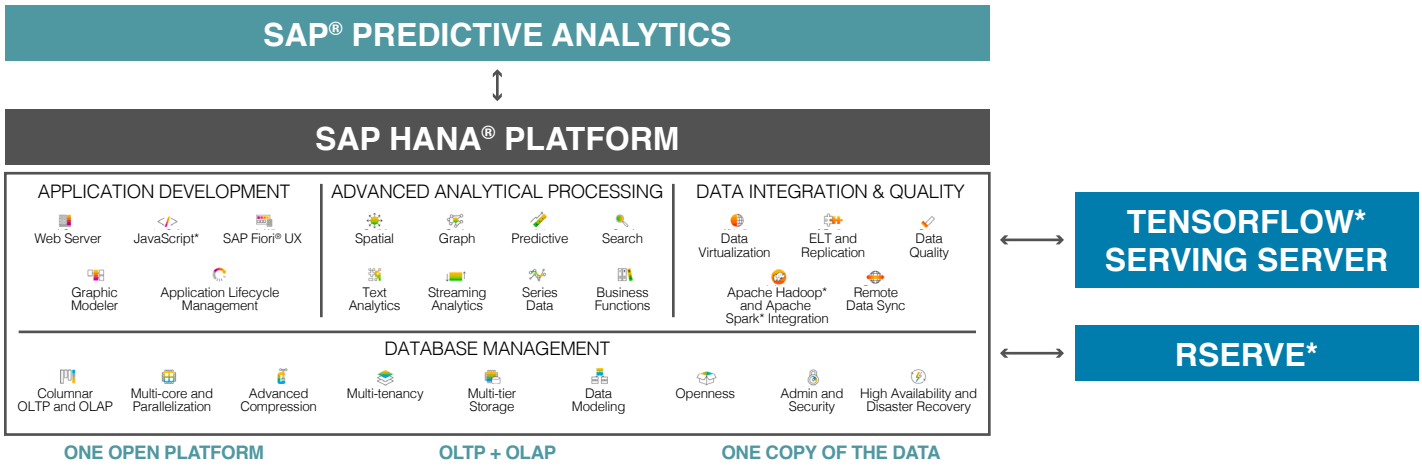

Figure 1. Diagram of the SAP HANA® platform and integration with SAP® Predictive Analytics, Rserve*, and the TensorFlow* Serving server

Users can also develop their models in TensorFlow, a popular, open source software library for machine learning. Models developed in TensorFlow can be loaded into a TensorFlow Serving server and invoked from the SAP HANA platform (Figure 1).

In-Database Machine Learning Performance in the SAP HANA Platform

In-database machine learning in the SAP HANA platform provides high-performance training and scoring for models on data sets both large or small, at rest or in motion.

Scoring Large-scale Data

Consider the situation of a bank that uses a predictive machine learning model to prioritize applications for credit cards. In such a situation, the company's data set is large: hundreds of thousands of applications each with its own long purchase and credit history to analyze. Nevertheless, the bank needs to be able to process the data quickly at any scale. Algorithms in the SAP HANA PAL are optimized to execute in parallel, on partitioned tables distributed across multiple servers. This parallelism helps ensure that the SAP HANA platform can make full use of all the available hardware computational power to deliver fast results. This is essential to quickly delivering timely results to applicants.

That ability of the SAP HANA platform to accelerate scoring on large-scale data is not limited to PAL-based models. For models built in TensorFlow or R, the SAP HANA platform offers load-balancing capabilities. When TensorFlow models are configured to run on multiple servers, they are load balanced using a random round-robin scheme. In the case of R, users can set up multiple machines running Rserve, and the R code will be executed on the machine with the least load. The SAP HANA platform also allows specific database users to be bound to a specific instance of Rserve, so all executions pertaining to a specific database user occur on the specified Rserve. The SAP HANA platform supplies high-performance execution of models from a variety of sources.

Real-Time Scoring

Not all aspects of machine learning involve processing extremely large data sets. Some models need to score a small amount of data repeatedly and with minimal latency. For example, consider a manufacturing firm that uses a decision-tree model for performing predictive quality assurance based on component dimensions that are

gathered at different stages of manufacturing. The company wants to predict if any given component is likely to become defective as manufacturing is underway so that it can perform immediate remediation. This necessitates scoring the machine learning model using new data relating to component dimensions quickly. The SAP HANA platform can meet this challenge because it parses a model once and keeps it in memory for subsequent executions, eliminating the overhead of reloading the model into memory every time. Large, complex models can be scored in real time and inside business processes, thereby enabling quick action and enhanced business agility.

Training and Scoring Models Using Streaming Analytics

Some business processes use streaming data from events such as clickstreams or sensors. For example, consider a retailer that built a model for predicting when to offer a coupon to a customer browsing a website. The company wants to process the clickstream as it is captured and instantly decide whether or not to issue a coupon to the prospect while the user is on the website. The SAP HANA platform has the ability to combine machine learning and streaming analytics processing. The SAP HANA streaming analytics engine scores the machine learning model on incoming clickstreams and allows the application to instantly issue pertinent coupons to website visitors.

Model Training in Multi-Core Environments

The lion's share of time in the machine learning lifecycle is spent training models. Therefore, it is crucial to accelerate the training process in order to minimize overall time to deployment. Machine learning with the SAP HANA platform is engineered to make use of a multi-core server architecture. This feature makes it possible to specify the proportion of threads to use for training a model, thereby benefitting from parallel processing. This thread allocation enables faster model training while also giving users control over the portion of system resources to utilize on a server.

Next Steps

As increasing data needs define the economy, the speed with which organizations can analyze relationships in data and act on those insights will separate successful companies from their competitors. Machine learning is a core component of building intelligent business processes, and machine learning is faster when it is executed where the data resides: in the database. Features and enhancements in the SAP HANA platform provide a powerful foundation for training and scoring machine learning models in production environments, and improvements and optimizations in Intel® Xeon® Platinum processors provide an ideal foundation for running machine learning workloads.

For more information on how the SAP HANA platform and Intel Xeon Scalable processors can accelerate machine learning in your organization, visit:

intel.com/sap
saphana.com

Intel® Xeon® Scalable Processors: Faster Training for Machine Learning on the SAP HANA® Platform

Intel® Xeon® Platinum processors, part of the Intel Xeon processor Scalable family, provide more memory, more cores, and more threads than previous generations of Intel Xeon processors, which can result in faster performance for machine learning. Intel Xeon Platinum processors support up to 1.5 TB of memory, 28 cores, and 56 threads per socket, which can result in faster training of models before taking them to production and faster scoring of models in production.[2] In addition, Intel Xeon Platinum processors provide the following features optimized for computationally demanding SAP HANA workloads like machine learning:

- **Intel® Advanced Vector Extensions 512 (Intel® AVX-512)**, a set of new instructions that can further accelerate machine learning performance by enabling the processing of twice the number of data elements of Intel AVX and Intel AVX2.
- **Intel® Ultra Path Interconnect (Intel® UPI)**, the successor to Intel® QuickPath Interconnect (Intel® QPI), has up to three channels to enable connecting Intel Xeon processors across a high-speed, low-latency path to increase scalability up to eight sockets and to improve bandwidth for input/output (I/O)-intensive workloads.

Looking forward, the Intel® Nervana™ Neural Network Processor (NNP) is a new offering from Intel for artificial-intelligence (AI) workloads. The Intel Nervana NNP is a purpose-built architecture for deep learning designed to further speed up AI workloads. The goal of this new architecture is to provide the needed flexibility to support all deep learning primitives while freeing the Intel Nervana NNP from the limitations imposed by existing hardware that wasn't explicitly designed for AI.

[1] Assessment based on internal Intel research.

[2] Select Intel® Xeon® Platinum processor stock keeping units (SKUs) provide up to 1.5 TB memory capacity. All others provide 768 GB.

Printed in USA 0218/TA/PRW/PDF Please Recycle 336864-001US