Integrating SAP data with machine learning systems

A deep dive into the different approaches and platforms for integrating your SAP data with ML systems

nagarro

Table of contents

Introduction

SAP systems are the backbone of countless enterprises worldwide, facilitating critical business operations across various domains. These systems capture and store vast amounts of structured data, from sales transactions and inventory management to customer interactions and financial records. However, as organizations strive to remain competitive and fuel innovation, the potential contained within this data extends far beyond conventional ERP applications.

Enter machine learning(ML) —a dynamic field of artificial intelligence that empowers systems to learn from data, identify patterns, and make accurate predictions or valuable recommendations. By integrating SAP data with ML systems, companies can tap into insights hidden within their operational data, amplifying returns on their SAP investments.

The fusion of SAP and ML technologies allows companies to supercharge various business processes. Predictive maintenance can help you identify potential equipment failures before they occur, reducing downtime. ML can also help analyze customer behavior to personalize marketing campaigns and improve customer satisfaction. You can use ML-driven demand forecasting and inventory management to optimize supply chains.

Integrating SAP data with ML systems also fosters a culture of continuous improvement. ML algorithms can analyze historical SAP data to identify optimization opportunities, streamline operations, and increase efficiency. Insights gained from ML models can inform decision-making processes, enabling businesses to stay agile, adapt to market trends, and drive sustainable growth.

However, integrating SAP data with ML systems has its challenges. The intricacies of data extraction, transformation, and model deployment require careful consideration to ensure seamless integration and reliable results. This whitepaper addresses these challenges, providing practical guidance and best practices to help organizations successfully bridge the gap between SAP and ML systems.

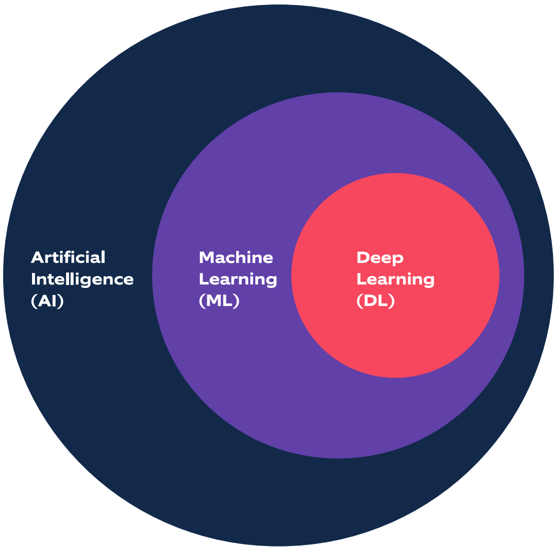
For companies already leveraging SAP's powerful enterprise resource planning (ERP) software, this whitepaper explores the possibilities and benefits of bridging the gap between SAP and ML.

Machine learning: An Overview

Before we begin discussing ways and approaches to integrate your SAP data for your ML systems, let us revisit the concept of machine learning.

ML is a subset of artificial intelligence (AI). While AI encompasses a broader range of applications, ML focuses on learning models that automate tasks, and make advanced analytical predictions using extensive databases and an intricate network of rules often called a Blackbox. These models can require extensive computation depending on the type of ML model.

While some ML applications work well with on-premises computations and require minimal setup and investment for effective training, others require substantial data volumes for success. An understanding of the ML project beforehand helps identify the required resources and compatible processes.



ML can be categorized into several types, including shallow models, deep learning models, and more. Understanding these categories helps you select appropriate models.

Shallow models: Shallow models, or traditional machine learning models rely on statistical algorithms to learn and derive value from given features in a dataset. They typically use regression models, or models that derive insight from statistical observations. It is common to see such models based on decision trees, such as XGBoost.

Most platforms support these models. These models often use large corporate datasets for efficient training and deployment and most training ML libraries are often equipped with these models.

While we recommend cloud computing for shallow models, you can train them on on-premises systems with limited datasets and still derive powerful insights. They are commonly used for prediction (regression) and segmentation (classification and clustering) and are quick to set up, making them suitable for Proof of Value (PoV) scenarios.

Deep learning models:

Deep learning models are a subset of machine learning models that imitate the structure and function of the human brain with the model structure called Neural Networks, often referred to as a Black box. These models automatically learn representations from raw data and excel at handling complex patterns and large datasets.

Transformer models, like Chat GPT, are widely used for powerful tools and generative language models. Implementing deep learning models requires advanced tools and frameworks. These models require substantial computational power and datasets to match to derive their full power.

When you think of deep learning models, think of Image recognition software, generative tools, and powerful predictive models that consider more factors compared to a shallow learning model.

Approaches to integrate your SAP data with your ML systems

While selecting the integration approach, evaluating your underlying architecture, and outlining a plan to leverage SAP in combination with your existing tools and licenses is crucial.

Utilizing the S/4HANA script server to run HANA-ML in your SAP cloud system:

- This is an exceptionally low commitment approach, as it works directly on your SAP system.

It does not require moving any data; you can work directly on your S/4HANA system. While it offers powerful, shallow learning models, it does not allow you to explore deep learning possibilities. It can be used in conjunction with other options, as a powerful tool to keep data entirely within company datacenters.

- **Using Datasphere/BW alongside FEDML (Federated Machine Learning Library) to train models with a Cloud Partner:**

  This is a new approach born out of customer desire, as many companies seek to utilize their ERP (Enterprise Resource Planning) data and train models with the available cloud platforms such as Azure or Google Cloud Platform (GCP).

  Doing ML on these platforms requires moving your data to said platforms, which can involve significant costs depending on the data. However, this approach provides a seamless intermediary system to store and structure your ML approach when working with many cloud providers.

  The Datasphere and FEDML are fully customizable and allow you to explore shallow and deep learning models on both cloud and on-prem platforms.

- **Engaging an SAP partner platform to facilitate your ML projects:**

  While this is a restrictive approach, it is also immensely powerful. It integrates data, updates shareholders on progress, and brings an out-of-the-box and structured approach to machine learning.

  Doing ML on these platforms requires moving your data to said platforms, which can involve significant costs depending on the data. However, this approach provides a seamless intermediary system to store and structure your ML approach when working with many cloud providers.

- **Employing SAP CPI or another SAP integration service to feed data to the cloud provider:**

  SAP Cloud Platform third-party (CPI) is the oldest and most robust option. It acts as the middleware between your SAP system and the third-party applications. It gives you the freedom to explore all types of ML projects.

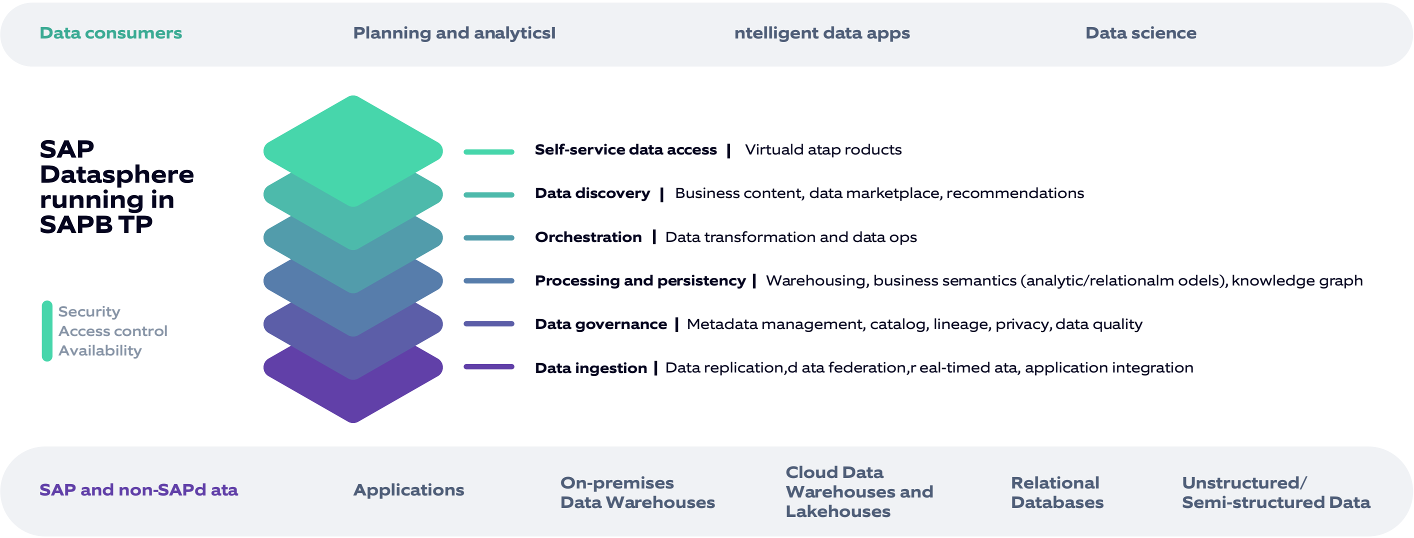
  Bear in mind, it is a pricier option and requires a keen understanding of the existing workflows and other systems.

  When deciding on the approach for your ML project, it is crucial to consider the level of integration required between SAP products. Whether conducting a proof of value (PoV) or establishing an architecture for a fully integrated Machine Learning Operations (MLOps) ecosystem, each approach offers unique advantages for your business.

Datasphere – The cloud data warehouse

For its data warehouse, SAP has shifted its focus to the SAP Datasphere as the successor to SAP BW, with existing BW users getting access to Datasphere for migration. It plays a key role when integrating your SAP data with ML systems.

Before diving deep into the different integration approaches, let us define Datasphere and understand its role in your integration journey.



Data consumers | Planning and analyticsl | ntelligent data apps | Data science

SAP Datasphere running in SAPB TP

Security
Access control
Availability

Self-service data access | Virtuald atap roducts

Data discovery | Business content, data marketplace, recommendations

Orchestration | Data transformation and data ops

Processing and persistency | Warehousing, business semantics (analytic/relationalm odels), knowledge graph

Data governance | Metadata management, catalog, lineage, privacy, data quality

Data ingestion | Data replication,d ata federation,r eal-timed ata, application integration

SAP and non-SAPd ata | Applications | On-premises Data Warehouses | Cloud Data Warehouses and Lakehouses | Relational Databases | Unstructured/ Semi-structured Data

Why use Datasphere?

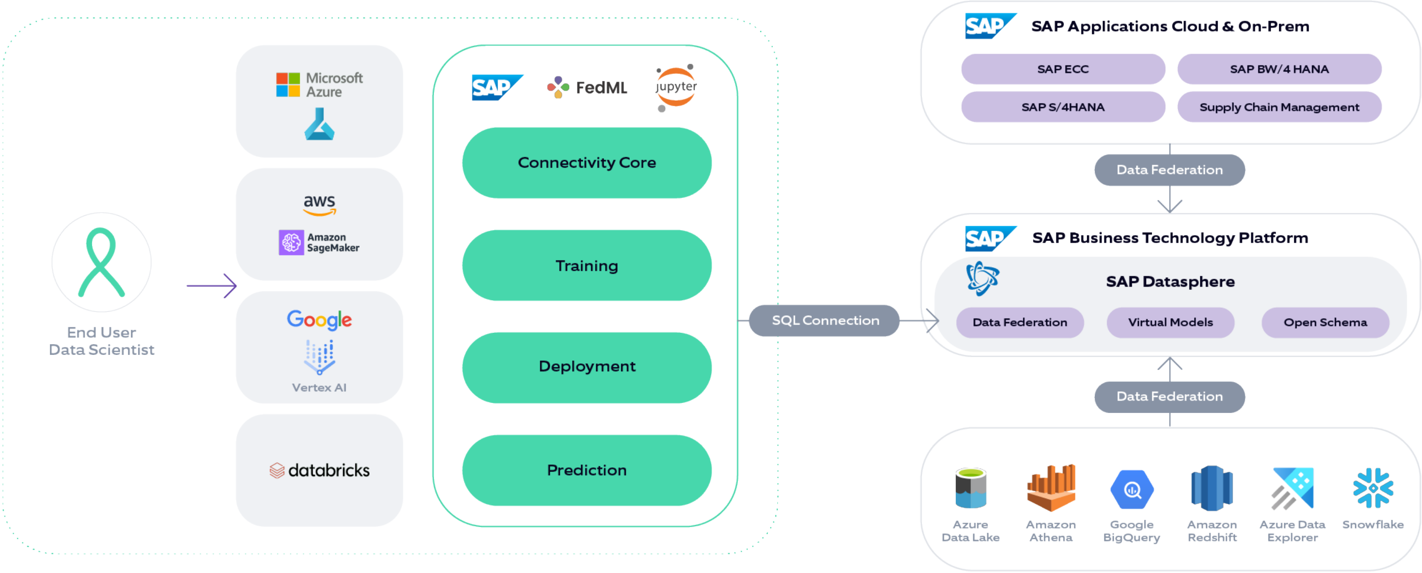
- Its flexibility makes it a viable option for both big and small businesses.
- It serves as a unified front for various industry standards in ML and BI, allowing customizations for specific business needs.
- It integrates seamlessly with popular cloud platforms such as Azure, Google Cloud Platform (GCP), and Amazon Web Services (AWS).
- It helps develop your architecture integrate data from various data sources and transform them into compatible formats.

Eventually, SAP intends to let people tailor Datasphere as a unified front for various industry standards in ML and BI (Business Intelligence) and is scheduled to be the replacement for their BW system.

One thing to remember when working with Datasphere is that it is still in development with no options for version control. Currently the database spaces have no way to communicate internally with one another.

While Datasphere offers the freedom to work in any preferred environment, and cloud provider. If working with a tech partner for your ML project, you can leverage their support and prebuilt tools.

SAP - Federated Machine Learning FEDML



Transferring data into and out of cloud environments incurs sufficient costs due to computation and data quantity fees. SAP's Federated Machine Learning Library (FEDML) offers a cheaper alternative for SAP integration with other cloud platforms.

When used with a cloud platform, FEDML acts as an intermediary between your environment and the cloud provider's environment, hosting your SAP data within the SAP Datasphere.

With multiple integration points for other databases, you can federate your databases for machine learning purposes, eliminating the need for data replication or migration.

This reduces the complexity of managing multiple data pipelines, ensures data consistency, and reduces the computation cost.

After training the model on the chosen cloud platform, the latest update to FEDML provides the flexibility to extract the model and utilize it within the SAP Datasphere and SAP Business Technology Platform (BTP) such as Kyma.

This approach is not only cost-effective, but it also simplifies an otherwise complicated process. It allows you to automate the workflows sparing your teams to focus on the more strategic tasks instead.

It is fairly easy to get started with FEDML, but it will require an S/4HANA and Datasphere instance. You can go through the following steps to set up the SAP Datasphere with FEDML with a connection to GCP:

- Set up a cloud subscription to feed or extract your data. We will demonstrate this with an example considering GCP's Vertex AI as the ML platform.
- Set up a Datasphere tenant, this can be done directly on the datasphere or BTP.
- Use DP agent to connect your HANA database to your Datasphere.
- Create A Vertex Notebook and container registry for model storage.
- You can use platforms like Google Big Query if your data is stored on the cloud or if you wish to migrate it to the cloud.
- Make a key or IAM on the chosen cloud platform to certify your connection in SAP Datasphere.
- Confederate your data in various views on SAP Datasphere.

Since FEDML does not function the same for each cloud provider, it is important to understand its functionality for different platforms.

GCP Vertex AI allows you to meet specific hardware requirements depending on what modules you wish to use, such as TensorFlow and CUDA, and specific graphics processing unit usage.

SAP also provides an excellent blog series on federated machine learning using SAP Data Warehouse Cloud and Azure Machine Learning v2[1] that detail their approach to other platforms.

SAP provides several options to facilitate seamless integration with your chosen cloud platform. However, it is important to note that setting up your partner system's cloud infrastructure, containers, and models requires significant work, regardless of the chosen connection method.

By setting up Datasphere and utilizing FEDML, you can harness the power of machine learning in collaboration with your preferred partner, unlocking valuable insights and driving business outcomes.

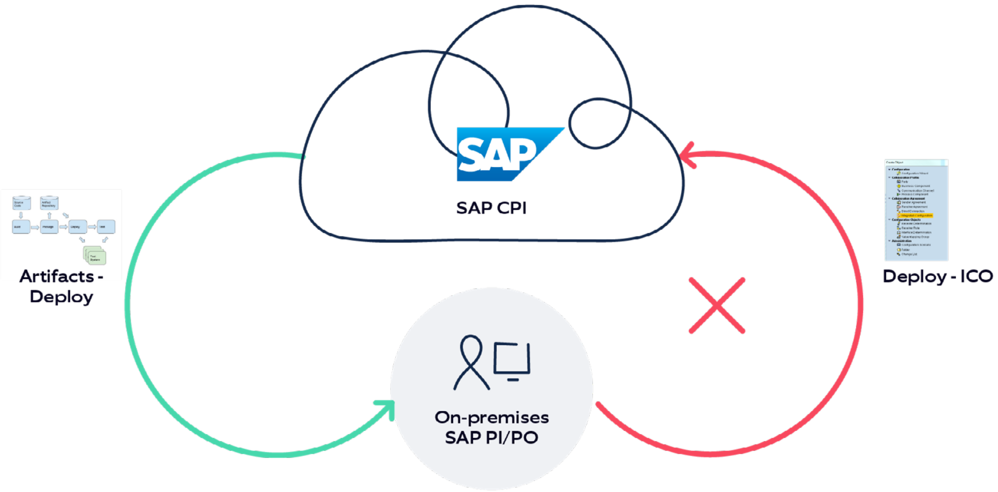
[1] https://blogs.sap.com/2022/06/09/federated-machine-learning-using-sap-data-warehouse-cloud-and-azure-machine-learning-v2/

SAP cloud platform integration – Flexible Integration

Another way to integrate your SAP data with your ML systems is SAP Cloud Platform Integration (CPI). CPI is a powerful tool for enabling seamless integration in various domains. It is pivotal in facilitating and connecting crucial components in machine learning initiatives. It seamlessly facilitates the implementation of intelligent decision support systems.

Suppose you are not interested in extensive platforms and investments to facilitate your ML. In that case, you can instead utilize CPI to connect your various toolsets and processes to SAP BTP or you are on-premises system.

Let us understand how CPI complements and works with other tools and services within an SAP ecosystem to maximize the benefits of machine learning.



Data integration and orchestration:

CPI connects diverse data sources, enabling organizations to create comprehensive, unified datasets for machine learning models. It facilitates the integration of databases, enterprise systems, cloud services, and IoT (Internet of Things) devices, allowing access to crucial data. However, data preprocessing, cleansing, and transformation tasks are performed using specialized tools or frameworks outside CPI.

Real-time data streaming and event-driven processing:

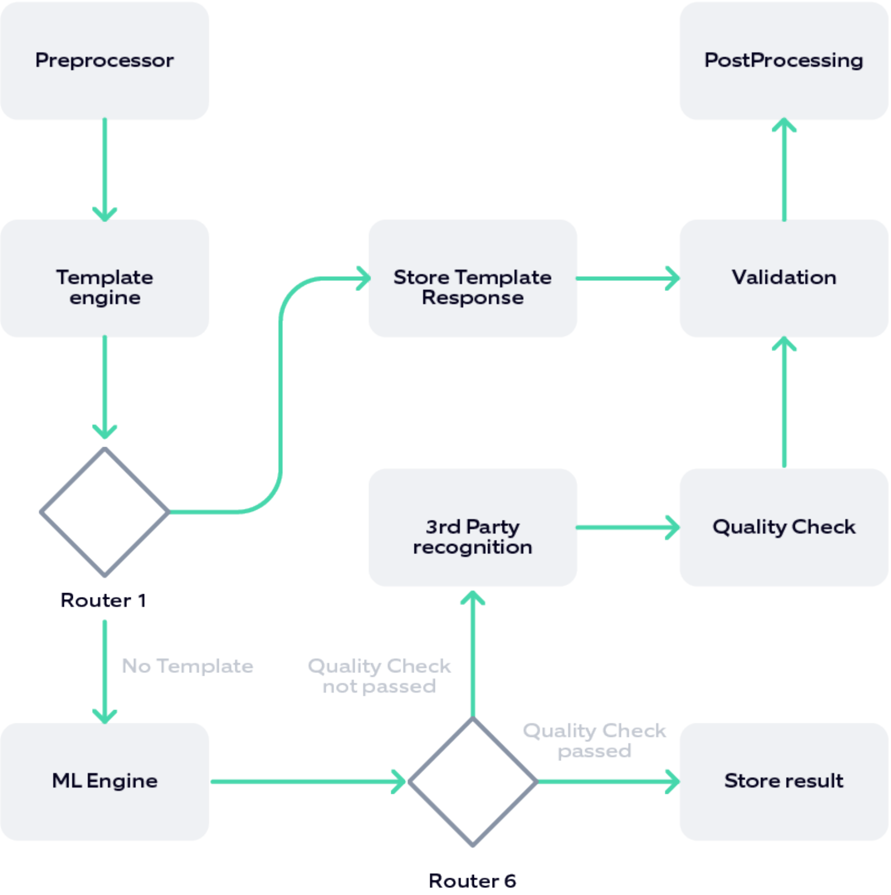
CPI enables ingestion and processing of real-time data streams, which is crucial for time-sensitive machine learning applications. It integrates with messaging, event-based platforms, and streaming services to enable near-instantaneous predictions and automated actions.

While CPI provides the infrastructure for real-time data handling, the actual machine learning models and algorithms are typically implemented and executed using dedicated platforms or frameworks.

Integration with model deployment and management services:

While CPI does not handle model deployment and management independently, it seamlessly integrates with other services within the SAP Cloud platform ecosystem specializing in those areas.

For example, the SAP Cloud Platform Machine Learning service lets organizations upload, deploy, and serve machine learning models as APIs (Application Programming Interfaces) or microservices.



CPI ensures a smooth data flow between systems and the deployed models. It safely delivers the data requested by the applications and transfers it to the third-party system. You can use it to monitor and manage the overall integration processes. This can in turn help set up a powerful middleware application to handle a robust workflow between your applications.

Tools like CPI are crucial if you don't wish to use a dedicated platform to perform your ML projects, as it instead becomes the bridge between your various platforms.

A CPI success story

Nagarro leveraged the CPI method to build an invoice scanner for a client working with SAP. The solution utilizes SAP to scan and extract invoices from customer emails and stores them in your S/4 HANA system. It then uses CPI to enable seamless communication between your cloud-based or on-prem S/4 HANA instance and third-party applications.

We built the invoice scanner, utilizing the microservice architecture by making Containers using Kubernetes. These Containers utilize python as programming language after which they orchestrated by Jenkins and Airflow. Jenkins ensures the automation of building and deployment processes, while Airflow handles the workflow.

CPI handles the workflow between SAP and the client's third-party applications or data. It receives invoices from SAP or other systems and extracts the information from the invoices with the invoice application.

The application consider different patterns and invoice types and utilize different template constructs first to see if the data can be extracted cheaply without ML, scoring the extraction, should these extractions not be satisfactory it will then utilizing a series of ML models until one returns a satisfactory result.

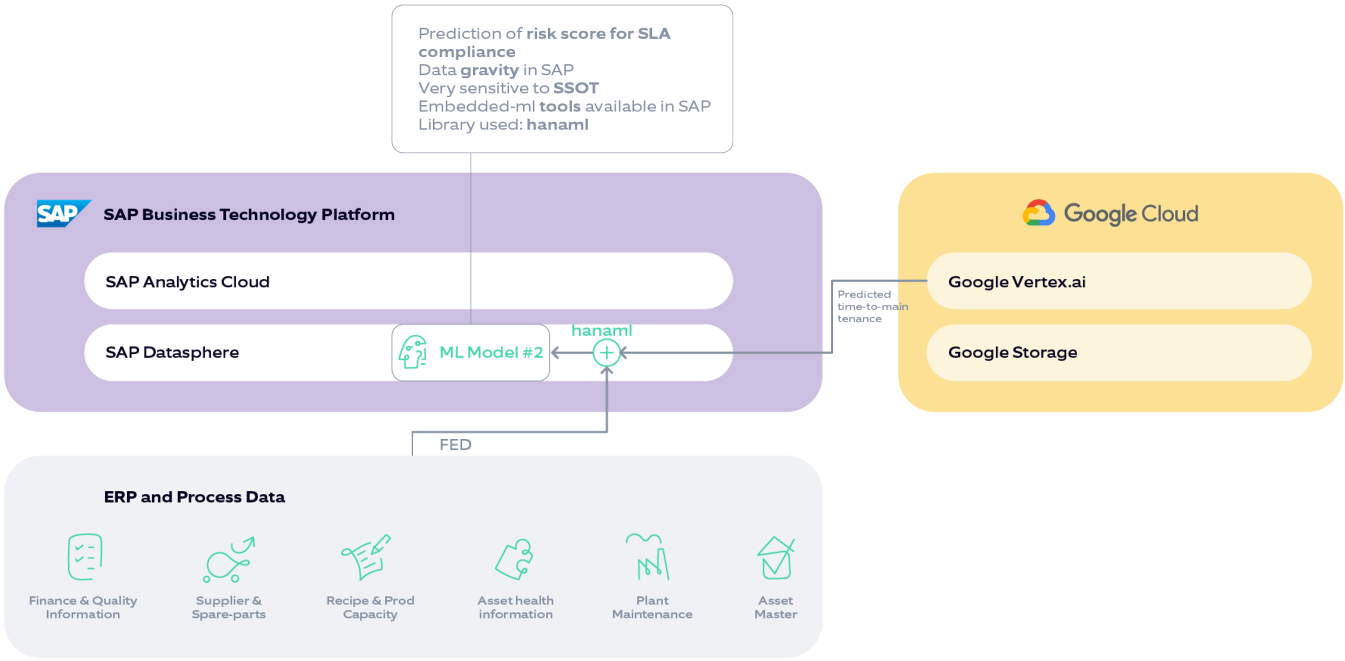
Here's how it works:

- We upload the invoice PDF into the SAP system and the CPI fetches it from there.
- The CPI sends the PDF file to a preprocessor engine to check the page count and the language.
- CPI then sends the PDF file to this template engine. If there is a matching template for this PDF file, then the data is extracted via regular expressions. If this is successful, then the invoice will not be processed by any other engine.
- If the template extraction doesn't succeed, the PDF file is sent to the ML engine. A quality check afterwards ensures quality.
- If the ML engine doesn't throw satisfactory results, the PDF file is sent to an external third-party system for the extraction.
- Finally, CPI performs several validations and summarizes the results to create an invoice object in the SAP system.

S/4HANA-ML & Low Integration Options

Another integration method SAP offers is its proprietary machine learning library called HANA-ML. You can utilize it on SAP Cloud Foundry or the KYMA engine with a script server in your SAP S/4HANA instance.

HANA-ML offers low-code models offering easy development. It is a quick and efficient way to use ML with your existing SAP infrastructure.



An advantage of using HANA-ML is that your data will remain within your HANA server ensuring adherence to the principle of a Single Source of Truth (SSOT). Utilizing HANA-ML can prevent data duplication, as indicated in the workflow above.

In addition, if you are working with sensitive information, such as medical information, it is recommended to utilize a private cloud or have it on-premises, where you can utilize HANA-ML.

This approach may not be suitable if you want to scale your operations extensively. It primarily supports basic data exploration and model creation, lacking capabilities for deep learning.

Additionally, as SAP does not offer information on support duration for HANA-ML, it may not be the best solution for longevity and scalability. In such scenarios, we recommend using open-source solutions instead.

While an on-prem approach is plausible with local versions of ML platforms or synchronization of teams with other platforms, we recommend an on-cloud dedicated platform for better accessibility and scalability.

In either case, it is important to remember that you can still extract SAP data without any of these tools, as discussed in the Dataiku section. HDBCLI or HDBSQL allows you to extract your data on any platform. However, we recommend these systems for easier integration, their ability to connect with other systems, and strengthen workflows.

SAP Partner Platforms (These platforms can work with or without Datasphere)

Let us now look at ML platforms that have sponsored integrations with SAP. These platforms provide structured environments and allow you to inform your shareholders about the workflows and the corresponding results. It should be mentioned that not all these platforms are SAP partners, but all have SAP integration options.

These platforms powerful visualization capabilities and ensure the following:

Accessibility and collaboration: Their visualization capabilities ensure that even those with little data science knowledge can work with facilitating project evaluation and promoting collaboration.

Efficient resource allocation: They allow data scientists to focus on building models rather than investing time in data preprocessing while benefiting from advanced tools and solutions for more complex tasks.

Enhanced performance: They offer advanced functionalities such as model training, development tools, and monitoring, improving performance and scalability for handling large-scale data processing and training.

Effective model management: Dedicated platforms provide robust features like model versioning, deployment automation, A/B testing, and monitoring capabilities, enabling streamlined model management and evaluation.

While these platforms integrate with third-party apps and improve data access, there's a concern that they may not align with your business model or organizational culture, as they require you to give up tech flexibility. It is, therefore, extremely crucial to understand the benefits and commitments for each platform.

Evaluating ML partner platforms for SAP

To help you decide, we have analyzed the pros and cons of a few data platforms for an ML project in an SAP environment.

These platforms help you ensure data compatibility, accuracy, and monitoring, give your data engineers an overview of the current data and allow your data scientists to request changes.

Databricks

Strengths	Weaknesses
• **Scalable and collaborative:** It leverages Apache Spark for distributed computing, allowing users to process large data volumes efficiently, ensuring scalability and collaboration. • **Unified analytics platform:** Databricks offers a unified platform that integrates data engineering, data science, and business analytics. It provides a seamless workflow from data ingestion to model development and deployment. • **Integration with popular tools and libraries:** It supports integration with various data sources, tools, and libraries commonly used in the data science ecosystem like Python, R, SQL, and popular machine learning frameworks like TensorFlow and scikit-learn.	• **Complex systems:** Databricks' advanced capabilities and distributed computing nature can make it challenging for beginners or those without prior experience in big data and distributed systems. • **Cost considerations:** While it offers a powerful platform, it involves higher costs, especially for large-scale deployments or resource-intensive workloads.

DataRobot

Strengths	Weaknesses
• **Automated machine learning:** It specializes in Auto-ML, providing automated tools and workflows for model selection, feature engineering, and hyperparameter optimization. This simplifies and accelerates the machine learning process. • **Ease of use:** Its user-friendly interface enables data scientists and business users to build and deploy machine learning models without extensive programming knowledge. • **Model transparency:** DataRobot emphasizes model transparency and interpretability, providing insights into the factors that influence predictions. This is crucial for regulatory compliance and model Explainability requirements.	• **Limited customization:** It offers limited customization options. Advanced users may find restrictions in implementing complex algorithms or specific requirements. • **Dependency on the DataRobot platform:** Vendor lock-in makes it difficult to make the switch from DataRobot to another platform.

Dataiku

Strengths	Weaknesses
• **Flexible and collaborative platform:** It provides a collaborative environment that enables cross-functional teams to work together. It supports end-to-end data preparation, modeling, deployment, and monitoring. • **Extensive integration capabilities:** Dataiku offers integration with various data sources, tools, and frameworks. It supports a wide range of programming languages, SQL, Apache Hadoop, Spark, and provides connectors to popular databases and cloud platforms. • **Customizable and extensible:** It allows you to customize and extend the platform using Python or R code, enabling advanced data manipulation, feature engineering, and model development.	• **Learning curve:** Dataiku's extensive features and customization options may require a learning curve for new or inexperienced users. • **Performance with large datasets:** While Dataiku can handle large datasets, the performance varies depending on the complexity of operations and data volumes. You may need to optimize it for for resource-intensive tasks.

Various other data science platforms offer varying levels of support for ML projects with SAP. They all have their strengths and weaknesses. And while SAP may not officially support many of these platforms, you can extract data from SAP via the S/4HANA Database Command Line (HDBCLI) or S/4HANA Database SQL (HDBSQL).

Cloud-based or on-premises platforms?

Architecturally you can choose either an on-prem or cloud-based approach. Current industry leans towards the cloud, or at least a hybrid approach, for several reasons:

One key reason is that model training is computationally intensive and often infrequent, making it more cost-efficient to scale your system as needed using a cloud vendor.

Moreover, given these PaaS and SaaS are charged on a usage basis, it makes sense to test them before making large investments.

They allow you a high degree of flexibility in testing various hardware, libraries, and frameworks for your project without making any greater initial investments. The on-cloud PaaS systems promote collaboration work without any glitches.

Cloud providers offer various managed services and automation tools specifically designed for machine learning workflows. These services handle time-consuming tasks such as data preprocessing, model training, and deployment, freeing your data scientists and engineers to focus on higher-value tasks.

While on-premises models also support data integration, they involve a larger initial investment to gain the computational power to train these models. This is combined with a continual cost to maintain the hardware to store and train them.

However, on-premises solutions have their own advantages, like reduced latency and full control of your system. One might choose an on-premises solution due to data and regulations surrounding it, as they offer much higher data security. This is particularly the case for medical and healthcare solutions.

If you are more interested in shallow learning models or AI approaches, it is completely viable to begin with an on-premises approach, as these require much less computation. You can start with a smaller investment and work your way upwards. Before you invest in a cloud platform or an on-premises system, we suggest utilizing Open-Cloud for a trial.

It is important to note that sometimes a hybrid approach is preferred, especially if you are dealing with heavily regulated data or have high data security levels.

How do you use ML within your SAP system?

Given that multiple options are available in the market, making the right choice can often be difficult. While we have tried to help, you make that decision through this paper.

Datasphere currently facilitates migration of your ERP data from SAP to the biggest cloud providers in the industry. While vendor lock-ins is challenging, most cloud providers offer a way around to tackle the lock-in.

SAP also offers CPI as a powerful integration option that allows you to utilize your data on your preferred cloud platform. It connects your systems beyond your ML ecosystem and helps you structure powerful workflows for various systems and applications.

These data science platforms allow you to visualize your workflows and progress for your shareholders, allowing you to demo your systems in a much easier fashion.

Working with a tech partner or an in-house ML expert is key to a successful ML implementation in an SAP environment.

At Nagarro, we emphasize the importance of closely tying the development process to real-world data and leveraging domain expertise. We work closely with your development team, guiding them with domain insights to ensure that you see quick results.

We leverage our extensive knowledge and experience to help you select the right tools and platforms that are well-equipped to facilitate your requirements.

Nagarro's expertise spans various domains, including BI, big data, machine vision, NLP generation, and processing. Our team of consultants has extensive SAP experience and is always ready to assist you in extracting maximum value from your ERP data.

Our host of accelerators enable us to quickly demonstrate the utility of the various integration tools, highlighting their value to your business.

Let's talk and take the first step toward unlocking the potential of your SAP data.

References

[1] https://www.silo.ai/ebooks-reports/nordic-state-of-ai-2022

[2] https://www.nagarro.com/en/blog/why-proof-of-concept-is-vital

[3] https://itrexgroup.com/blog/how-ai-proof-of-concept-helps-you-succeed-in-your-ai-endeavor/

[4] https://www.ittbusiness.at/article/nagarro-logicals-sqeed

[5] https://blogs.sap.com/2022/06/08/fedml-the-federated-machine-learning-libraries-for-hyperscalers-version-2.0/

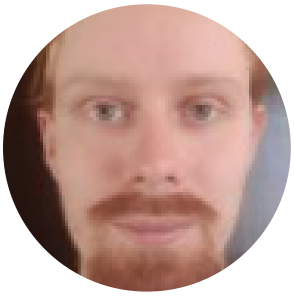
[6] https://help.sap.com/docs/cloud-integration

[7] https://help.sap.com/docs/SAP_DATASPHERE

[8] https://www.nagarro.com/en/whitepapers/ai-testing-ai

[9] https://www.nagarro.com/en/whitepapers/enterprise-mlops

Author bio



Benjamin Toubøl Benjamin Toubøl is a Data scientist with a passion for practical and explainable Machine Learning advocating for incremental adoption of new technologies with small and big businesses alike to assist and enhance business processes for the future of the intelligent enterprise.

About Nagarro

Nagarro is a global digital engineering leader with a full-service offering, including digital product engineering, digital commerce, customer experience, AI and MLbased solutions, cloud, immersive technologies, IoT solutions, and consulting on next-generation ERP. We help our clients become innovative, digital-first companies through our entrepreneurial and agile mindset, and we deliver on our promise of thinking breakthroughs.

Our guiding principles are defined by one word – CARING, denoting a humanistic, people-first way of thinking with a strong emphasis on ethics. Caring guides us as a global company.

We have a broad and long-standing international customer base, primarily in Europe and North America. This includes many global blue-chip companies, leading independent software vendors (ISVs), other market and industry leaders, and public sector clients.

Today, we are over 19,000 experts across 34 countries, forming a Nation of Nagarrians, ready to help our customers succeed.

For more information, visit www.nagarro.com