Optimizing Order Fulfillment Using Machine Learning

Supply-chain reliability is more critical than ever during pandemic conditions. SAP worked with Inspired Intellect to develop an order-fulfillment-optimization solution, built on Red Hat and Intel® technologies, with advanced analytics and artificial intelligence (AI) to help prevent supply shortages and delays.

The growing challenge of supply-chain reliability

Businesses are continuously challenged to improve customer satisfaction and service. This is particularly true for enterprises that have a strong dependency on a supply-chain network. Delays in shipments reaching their customers or target destinations translate to poor customer experiences and the loss of future revenue through customer attrition.

These scenarios were front and center on the world stage in 2020, as nations and businesses around the globe responded to the threat of COVID-19. Hospitals, healthcare facilities, pharmacies, and grocery stores have all experienced pronounced product-delivery delays and supply shortages during the pandemic. Such delays can exact a real cost in terms of lower customer trust, increased customer churn, and loss of business. With the mounting strains on global supply chains, companies are now investing heavily in technologies and processes that enhance adaptability and resiliency in their supply chains. Of course, these same types of supply-related delays and shortages occur in any number of natural-disaster situations, such as hurricanes, earthquakes, and regional flooding.

As a leader in supply-chain logistics, SAP wanted to address the problems of unexpected delays and shortages. To that end, it worked with Inspired Intellect—a leading data analytics consultancy—to build a better supply-chain solution based on advanced analytics and machine learning (ML). Intel® Xeon® Scalable processors, the Red Hat OpenShift Container Platform, and Red Hat OpenShift Container Storage are key components in this solution.

Data analytics with ML to avoid delivery delays

This customer story highlights the benefits of a data-driven solution and the value of AI and ML in mitigating the risks associated with order and shipment delays. It also describes how the technologies involved can deliver a scalable enterprise-grade offering.

The objective for this solution was to proactively predict likely order-fulfillment delays and prescribe alternative options to mitigate the risk of such delays. To achieve this objective, SAP and Inspired Intellect sought to identify the root causes of delivery delays by using advanced analytics, and to then architect a modernized order-management system that does the following:

- Improves customer experiences and customer satisfaction
- Mitigates the risk of customer or revenue attrition due to recurring service-level agreement (SLA) defaults
- Bolsters revenue-recognition potential and reduces penalties associated with SLA defaults

This supply-chain solution incorporates key components from SAP, Red Hat, and Intel.

Powerful 3rd Generation Intel Xeon Scalable processors

3rd Generation Intel Xeon Scalable processors evolve Intel's 2-, 4-, and 8-socket processor foundation for AI-infused, data-intensive workloads. They offer built-in AI-acceleration features that further support real-time AI on software platforms. These features include the latest evolution of Intel Deep Learning Boost (Intel DL Boost), making 3rd Generation Intel Xeon Scalable processors the first general-purpose CPUs to offer built-in Brain Floating Point (bfloat16) instructions. These features also build on existing support for Vector Neural Network Instructions (VNNI). The addition of bfloat16 instructions makes mainstream AI training more widely deployable for workloads such as image classification, speech recognition, and language modelling. It also helps make these general-purpose CPUs a more cost-effective option than purchasing expensive graphics processing units (GPUs) specifically for AI workloads.

Inspired Intellect

Inspired Intellect is an end-to-end service provider of data management, analytics, and application development. Its portfolio of offerings ranges from strategic advisory and design services to development and deployment services to sustained operations and managed services. Inspired Intellect fosters clients' mastery of advanced analytics, which helps them transform data into insights and drive meaningful business outcomes. This is achieved by deploying the latest AI and ML techniques, supported by the latest ML operations, procedures, and automation.

Red Hat OpenShift Container Storage for enterprise-scale Kubernetes

Red Hat OpenShift Container Storage is persistent software-defined storage integrated with and optimized for the Red Hat OpenShift Container Platform. It runs anywhere Red Hat OpenShift runs, whether that's on-premises or in the public cloud. Built on Red Hat Ceph Storage, the platform offers tightly integrated, persistent data services for Red Hat OpenShift and hybrid multicloud environments. Dynamic, stateful, and highly available container-native storage can be provisioned and deprovisioned on demand as an integral part of the Red Hat OpenShift administrator console.

SAP Data Warehouse Cloud solution and SAP Data Intelligence solution

This hybrid supply-chain solution employs two complementary products developed by SAP, as shown in Figure 1. The SAP Data Warehouse Cloud solution serves as the analytics data repository in the cloud, and the SAP Data Intelligence solution serves as the data-science and ML platform on-premises.

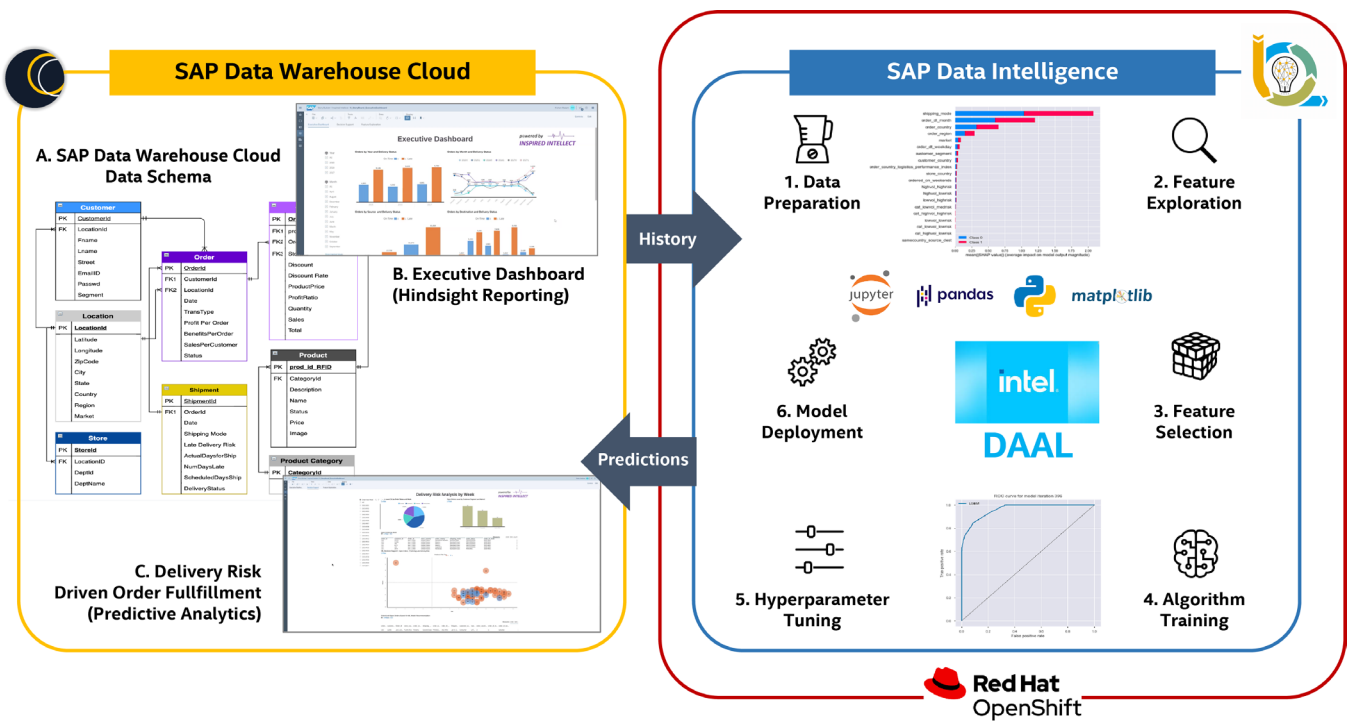


Figure 1. The SAP Data Warehouse Cloud solution warehouses the data, and the SAP Data Intelligence solution uses ML to make predictions (image courtesy of Inspired Intellect)

In order to achieve optimal ML performance and enterprise-grade scalability, the SAP Data Intelligence component employs two key technologies:

- A Red Hat OpenShift containerized environment establishes the platform's security and DevOps capabilities.
- The Intel Data Analytics Acceleration Library (Intel DAAL) provides performance scalability and predictive value-add.

Intel DAAL

Intel DAAL is the library of Intel architecture-optimized building blocks covering all data-analytics stages: data acquisition from a data source, pre-processing, transformation, data mining, modeling, validation, and decision making. To achieve the best performance on a range of Intel processors, Intel DAAL uses optimized algorithms from the Intel Math Kernel Library (Intel MKL) and Intel Integrated Performance Primitives (Intel IPP).

Intel DAAL supports the concept of end-to-end analytics in situations where some data-analytics stages are performed on edge devices—that is, devices that are close to where the data is generated and where it is finally consumed. Specifically, Intel DAAL APIs are agnostic about cross-device communication technologies, and therefore they can be used in different end-to-end analytics frameworks.

Advantages of Red Hat OpenShift Container Platform and Red Hat OpenShift Container Storage

Red Hat's hybrid cloud and container foundation simplifies IT operations to increase consistency, enhance security, and streamline management.

Simplicity

Red Hat OpenShift gives you cloud-like simplicity with full-stack automated operations on a consistent foundation across on-site, cloud, and hybrid infrastructures. With Red Hat Enterprise Linux CoreOS, OpenShift offers automated installation, upgrades, and life-cycle management for every part of your container stack across your business—the operating system, Kubernetes and cluster services, and applications. The included Operator Lifecycle Manager (OLM) further simplifies management of applications native to Kubernetes with built-in operational knowledge. It also extends the Kubernetes API to complex, stateful applications, allowing you to deploy, manage, and maintain them within Red Hat OpenShift.

Consistency

Red Hat OpenShift lets you deploy containers and Kubernetes services consistently across on-site, private cloud, and public cloud infrastructures. It includes the components needed to run containers in production: an enterprise-grade Linux operating system, container runtime, networking, monitoring, registry, and authentication and authorization solutions. These components are tested together for unified operations on a complete platform spanning all clouds. Red Hat Enterprise Linux CoreOS also provides an immutable, container-optimized Linux host that can be deployed on any infrastructure. Finally, Red Hat OpenShift integrates seamlessly with your existing investments to provide a consistent operations, development, and management experience across multiple infrastructures and teams.

Security

Red Hat OpenShift is designed for full-stack, continuous security, from the operating system to the application and throughout the software life cycle. In addition to the security provided by Red Hat Enterprise Linux and the containers themselves, Red Hat OpenShift adds built-in authentication and authorization, secrets management, auditing, logging, and an integrated container registry for granular control over resources and user permissions. Red Hat Enterprise Linux CoreOS provides added automation to keep large deployments up-to-date more easily and to quickly identify security vulnerabilities within your container images.

Flexibility

OpenShift Container Storage runs as a Kubernetes service within Red Hat OpenShift and can also be decoupled and managed as a separate data store, delivering file, block, and object data for one or many OpenShift Container Platform clusters. Tightly bundled offerings including Red Hat OpenShift, Red Hat Middleware, and Red Hat OpenShift Container Storage can also help organizations automate data pipelines, enabling real-time data processing from edge to hybrid and multicloud platforms. This level of integration removes the guesswork from running Red Hat OpenShift across multiple platforms and provides the data-storage functionality, data services, and data protection that enterprises require.

In addition, Red Hat supports MLOps, enabling multiple patterns for AI/ML management. You can train your models in SAP Data Intelligence and kick off an MLOps process to move your models into production. Additionally, you can integrate your edge device with an image repository, such as Quay, and run inferencing at the edge with the latest available models.

Intel technologies in use

Intel Advanced Vector Extensions 512 (Intel AVX-512) is a set of instructions for accelerating complex in-memory workloads like analytics and AI to help improve performance.

Intel Key Protection Technology (Intel KPT) is a security feature of Intel Xeon Scalable processors. Intel KPT is used to protect sensitive encryption keys at runtime, adding an additional layer of protection for the platform.

Intel QuickAssist Technology (Intel QAT) is an accelerator for compressing and decompressing data, enabling faster processing and analysis.

CRISP-DM methodology

Inspired Intellect employed the well-known cross-industry standard process for data mining (CRISP-DM) model to identify advanced analytics opportunities and develop the supporting ML models that would lead to a better understanding of the root causes of order delivery delays. These insights will then facilitate recommendations for corrective actions aimed at improving the order-fulfillment decision process and proactively managing customer expectations in situations involving inevitable order-delivery delays.

Data discovery

This solution employs relevant data from a variety of SAP solutions, such as SAP Customer Relationship Management (SAP CRM), the SAP Material Master object model, the SAP sales and distribution monitor, the SAP Retail industry solution, SAP S/4HANA Sales, and SAP S/4HANA Cloud. It also taps into relevant third-party data sources, such as international customs/trade regulations, weather data, and traffic data.

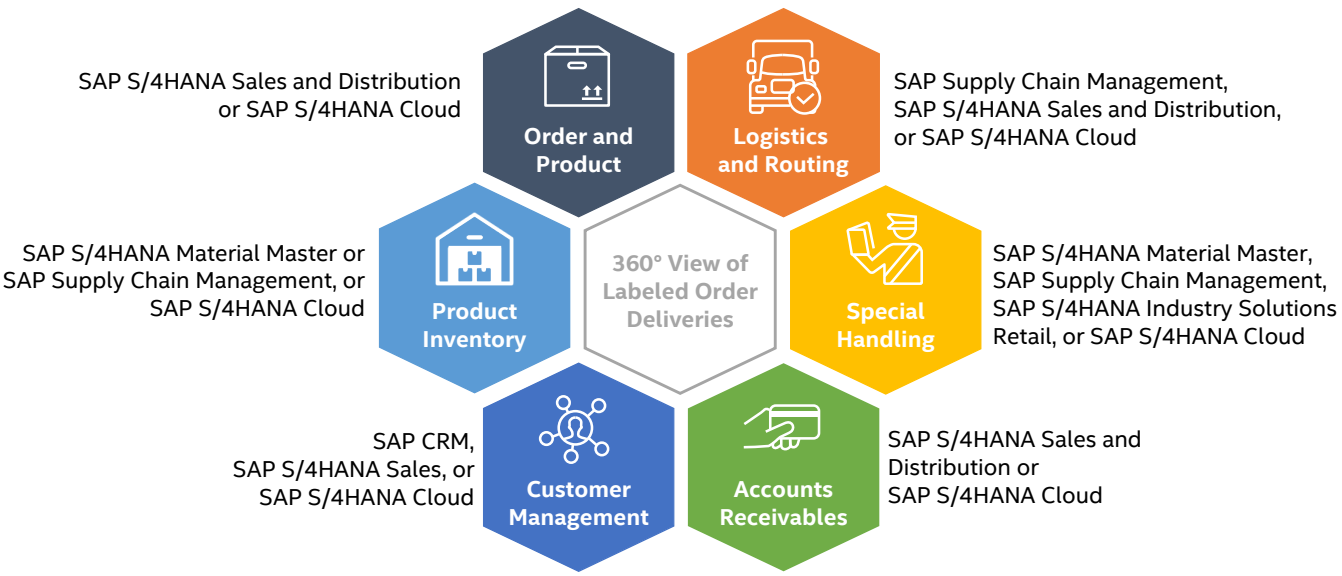


Figure 2. Sourcing all the relevant data from SAP systems (image courtesy of Inspired Intellect)

As prescribed by CRISP-DM, development began with a data-discovery exercise. Employing the data wrangling, data-transformation, and data-visualization capabilities available in SAP Data Warehouse Cloud, the team replicated a typical enterprise executive dashboard that relies on hindsight reporting to determine how best to prioritize incoming orders through the fulfillment and delivery funnel. The team used the dashboard and reports to showcase the deficiencies that are inherent in such hindsight-based decision making, a practice that leaves considerable business opportunities on the table in the form of inflated supply-chain costs and/or lost revenues. In addition, the data discovery also surfaced a set of hypotheses to be validated during model development.

Model development

The findings from the data discovery and the rich vein of available data led to the team choosing a pair of complementary ML models to form the cornerstone of the proactive order-fulfillment decision process. Specifically, these two models were designed to:

1. Predict the likelihood that an incoming order will be delayed, based on the order characteristics
2. Estimate the expected delay (in days) for the incoming order, to better manage customer communications and expectations when faced with potential shipment delays

Application developers used the Intel oneAPI Deep Neural Network Library (Intel oneDNN) to improve the productivity and enhance the performance of the deep learning (DL) framework. This included taking advantage of the 3rd Generation Intel Xeon Scalable processors' built-in AI-acceleration capabilities, like bfloat16 support, to speed up training for AI models.

The SAP Data Intelligence solution provides data scientists and ML engineers with typical productivity and development tools, such as an integrated Jupyter Notebooks integrated development environment (IDE), a Python distribution, and data science and ML libraries such as pandas, NumPy, Matplotlib, Seaborn, and scikit-learn. The SAP Data Intelligence solution also supports automated ML capabilities that collectively simplify the ML algorithm development while making use of the power of an integrated SAP solution suite.

Decision-process transformation

Incoming orders can be scored against each of the two predictive ML models to transform the order-fulfillment process, as illustrated in Figure 3:

1. An incoming order triggers two ML models based on historical order delivery data.
2. One model predicts the likelihood of delay, and multiplies it by the revenue (or margin) of the incoming order to yield the expected revenue (or margin) at risk.
3. Another model predicts the expected length of the delay, in days.
4. The order is prioritized for fulfillment based on expected revenue at risk and expected delivery delay.

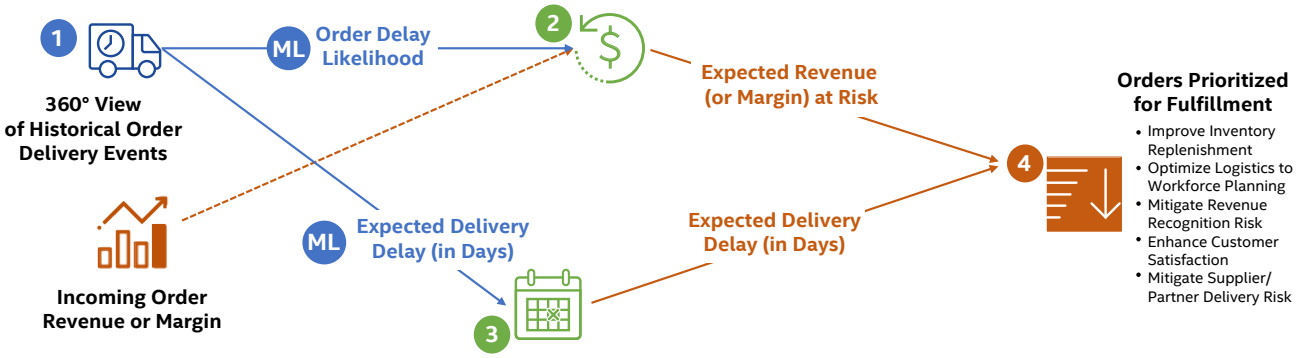


Figure 3. Calculating an expected delivery delay and the revenue (or margin) at risk

Cross-tabulating the risk estimate against the expected delay (in days) from the second model yields an intuitive 2x2 matrix to determine order-fulfillment priorities, as shown in Figure 4.

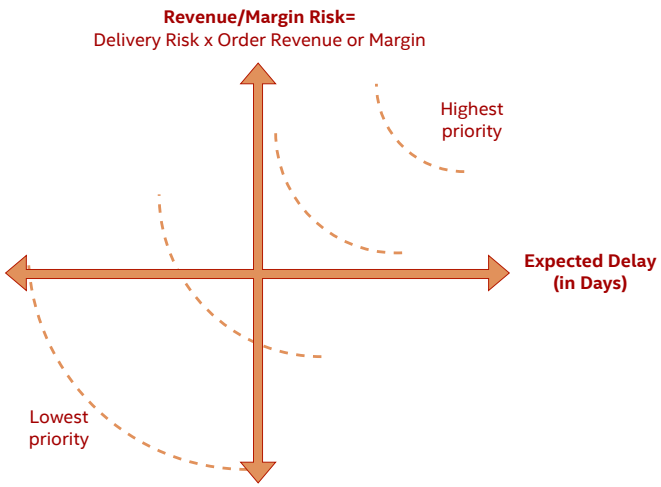


Figure 4. Order-fulfillment priorities matrix

The hindsight-based executive dashboard discussed earlier can now be supplemented with a predictive delivery risk score and a days-delayed forecast estimate. This empowers order-fulfillment and distribution center supervisors with data-driven insights and enables them to take corrective action prior to an order-delivery delay.

The benefits of enterprise readiness

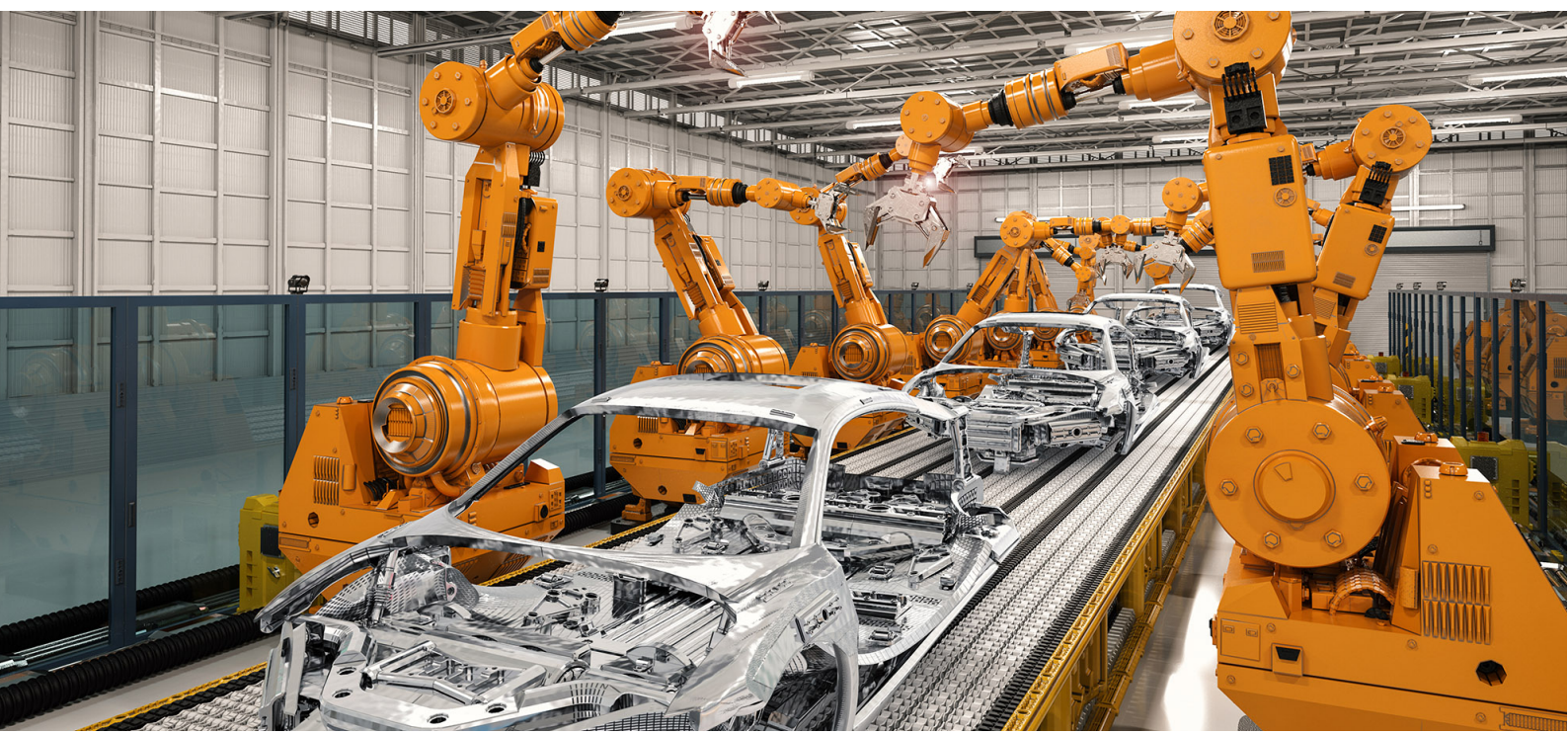
To validate the enterprise readiness of the order-fulfillment optimization offering, the testers evaluated two criteria: performance scalability and DevOps scalability.

Performance scalability

To demonstrate the performance-scalability opportunities, the traditional open source ML libraries were supplemented with Intel DAAL–optimized libraries. These libraries use the latest available vector-instructions on each architecture, enabling them at the hardware level. Intel DAAL has been demonstrated in the scientific community to scale to 1,024 nodes with more than 88-percent efficiency.[1]

Research on this solution confirmed significant compute performance gains using Intel's libraries for compute-intensive algorithms such as gradient-boosted trees and methods such as grid-search for hyperparameter optimization.

In addition, the value-add of Intel's libraries was also evident to varying degrees across the end-to-end ML pipeline. Research is continuing into the benefits of Intel's libraries by replicating typical enterprise data science/ML workloads to include multiple use cases, a breadth of data sources (across batch and streaming data for structured and unstructured data), distributed compute requirements, concurrent users, and a host of other parameters that could influence scalability. This will foster a broader performance comparison of Intel's libraries against the capabilities of open source ML libraries beyond the context of this use case.



DevOps scalability

DevOps scalability is a natural next step beyond solution-performance scalability. The SAP Data Intelligence solution was deployed within a Red Hat OpenShift containerized environment in order to address the DevOps scalability priorities of enterprise IT organizations. This also brings to bear a microservice-focused application architecture, and it provides a consistent platform for hybrid cloud deployments.

In addition, Red Hat OpenShift addresses a number of challenges that dominate operational costs within a typical enterprise IT organization. The following are some of the most frequently encountered challenges that are readily addressed by the Red Hat OpenShift Container Platform and Red Hat OpenShift Container Storage solution:

- It takes weeks for technology environments to be deployed.
- Environment deployments are manual, painful, and infrequent.
- Applications behave differently in production than in test environments.
- A proliferation of technologies must be supported and maintained.
- Each environment has a different set of configurations to manage.
- There are persistent storage requirements (at both the file and data block levels) for containerized applications.

Red Hat OpenShift eases the orchestration of these and many more technology and DevOps challenges currently facing enterprise IT functions by simplifying Kubernetes serverless containerized operations.

Bottom-line benefits

Organizations that adopt a data-driven approach can significantly enhance their competitive positioning in the market. The benefits of such an approach include:

- Improved customer experience, reduced customer churn, and improved customer satisfaction
- Improved operational efficiencies resulting from optimized logistics and distribution planning through improved management of inventory, fulfillment processes, and resource plans
- Enhanced management of strategic relationships with suppliers and logistics/distribution partners to maintain desired customer service levels
- Reduced SLA penalties incurred and reduced margin erosion from customer discounts offered to compensate for delayed order deliveries
- Additional benefits delivered through enterprise-grade scalability of the offering, resulting from employing Intel DAAL–optimized ML libraries and Red Hat OpenShift containerization

Conclusion

The risks associated with supply-chain delays have never been higher. Inspired Intellect assembled the best-in-class partnership of SAP, Red Hat, and Intel to deliver an integrated and scalable supply-chain solution that uses advanced data analytics with AI and ML to mitigate those risks. Two components—the SAP Data Warehouse Cloud solution in the cloud and the SAP Data Intelligence solution on-premises—provide the core supply-chain functionality for this solution, while the Intel DAAL library is used to optimize performance and scalability.

The solution employs ML models created by Inspired Intellect to improve supply-chain operational efficiencies by isolating the factors that contribute to higher operational costs, margin erosion, future revenue risk, or customer lifetime value. The solution transforms the order-fulfillment decision making process by providing decision makers with a more complete understanding of what factors influence order-delivery delays. Additionally, it uses these insights to present corrective recommendations that will minimize the risk of delivery delays or SLA penalties.

Companies adopting this solution can incorporate a wide range of data assets to be considered by the AI/ML models, including inventory/out-of-stock data, routing information, and specialized handling requirements (for example, climate control and customs/regulatory controls), among others. Moreover, this solution is not limited to the supply-chain context, but is a reusable and scalable framework that can be applied to many other types of use cases and business challenges in a wide variety of verticals.

The insights and recommendations presented in this case study can be enhanced by complementing internal enterprise data assets with external data and influential factors. For example, visibility to heatmaps of natural disaster epicenter migrations (for example pandemics, wildfires, earthquakes, hurricanes, floods, or tornados)—in addition to social-media-driven insights into trending product shortages and first-responder containment strategies and needs—can prepare and empower enterprises to proactively manage their supply chains no matter what the operating environment calls for.

The performance and scalability of the SAP Data Intelligence solution, which delivers the predictive ML component, are enhanced by Red Hat and Intel technologies. Red Hat OpenShift delivers DevOps scalability in a containerized environment with simplicity, consistency, and security. Intel technologies accelerate the ML process with the latest 3rd Generation Intel Xeon Scalable processors and Intel DAAL–optimized libraries.

The bottom line is that the SAP Data Intelligence solution facilitates holistic data-driven decision making by anticipating the dynamic demand patterns that are challenging supply-chain-dependent businesses operating during the current pandemic.

Learn more

For more information on how Intel and SAP are working together to build a more intelligent enterprise, see intel.com/sap.

[1] Intel DAAL K-means fit, week scaling results (87.44 GB/node, 84 features, 8 clusters, 100 iterations, float32). Source: https://arxiv.org/abs/1909.11822.

Configuration: Testing by Intel as of 09/25/2019. 7 x m5.2xlarge AWS instances, Intel DAAL 2019.3; Intel Xeon processor E5-2698 v3 at 2.3 GHz, 2 sockets, 16 cores per socket, MPI4Py (3.0.0), Intel Distribution of Python (IDP) 3.6.8, float, Source: https://arxiv.org/abs/1909.11822.

Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice. Notice revision #20110804.

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.

Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

Printed in USA 0121/VS/PRW/PDF Please Recycle 345606 -001US