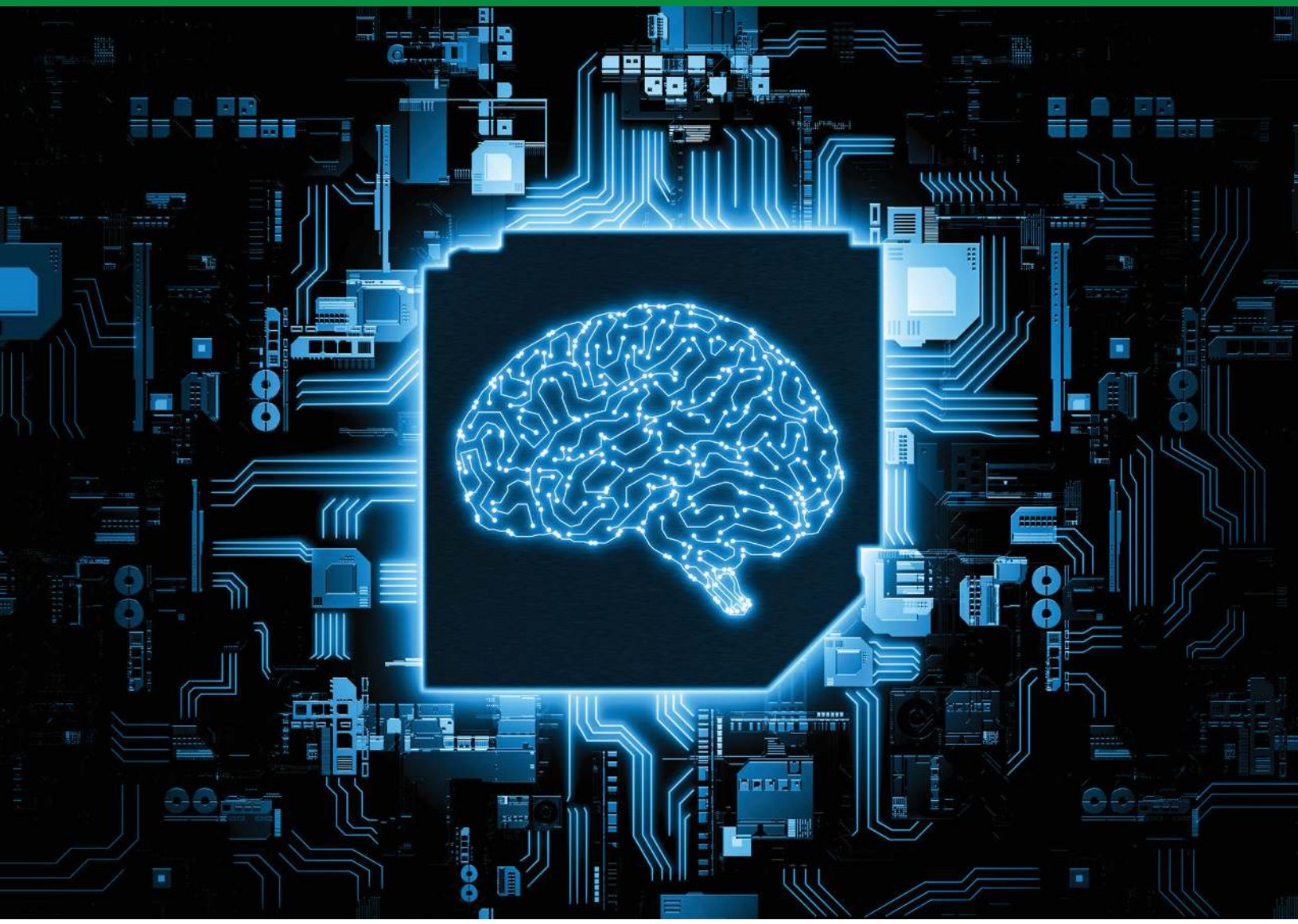
Using Kubernetes as the Core Underpinning of Your End-to-End AI/ML Projects



RTInsights.com
Accelerate Your Business with Real-Time Insights

Red Hat

Introduction

In today's dynamic marketplace, new applications that use data from multiple sources and deliver rapid insights constantly need to be created on very short notice. The challenge is how to have the flexibility to rapidly develop and deploy new applications to meet fast-changing business requirements. The only way to ensure success is to use a dynamic architecture that delivers access to data, processing power, and analytics (including artificial intelligence and machine learning models) on demand.

Traditional development approaches break down. They do not offer the flexibility to easily incorporate new data sources or analytics. Nor do they lend themselves to today's need for continuous changes after applications are deployed. Such problems become unmanageable as AI/ML needs expand throughout an organization.

There are two main approaches to address these problems. Businesses can either adopt an all-encompassing framework from a single AI/ML vendor or leverage the innovation of the AI/ML tools being developed in the open-source community. Both approaches benefit by leveraging a hybrid development and deployment model based on Kubernetes, the open-source system for automating deployment, scaling, and management of containerized applications.

Considerations: shortcomings of maintaining the status quo

Today, every company is a software company. Regardless of the industry, businesses are developing intelligent applications to automate processes, increase operational efficiencies, and improve customer experiences. Unfortunately, traditional approaches to developing such applications do not scale well from several perspectives.

Labor/staffing resources issues

Mainstream adoption of AI/ML is exploding. Rather than being relegated to a small niche of enterprise applications, such intelligence is being embedded into all forms of applications used throughout a business. In 2020, the global AI software market grew by more than 50 percent year-on-year. [1]

As the number of projects grows, organizations find they do not have enough skilled staff to develop or maintain applications. Making matters worse, as more groups within an organization turn to AI/ML, there often is a waste in that many of these efforts duplicate the same basic work.

Most efforts start from scratch, doing the same work. For example, different departments may independently develop ways to access a commonly used dataset, prepare that data for use in their apps, select a suitable analytics model to use on the data, and then train that model.

[1] Artificial intelligence software market growth forecast worldwide 2019-2025, Statista, August 17, 2020 https://www.statista.com/statistics/607960/worldwide-artificial-intelligence-market-growth/

Lifecycle issues – development is just the tip of the iceberg

Modern digital businesses work in real-time. They rely on insights rapidly derived from increasingly larger volumes of data from a growing variety of sources. Legacy approaches often are costly to deploy and maintain, and they do not scale well. New approaches are being driven by the demand to get more value out of data, and that insights must be derived in shorter times.

Beyond development, most modern applications must be continuously updated throughout their lifecycle. Typical factors that must be considered include:

Support new/more data sources and types: Most businesses seek to use new data types as they become available. Manufacturers want to be more proactive using data from smart sensors and the Internet of Things devices deployed throughout their factories and supply chains. Retailers and online merchants want to take customer engagements to a level higher with text and sentiment analysis of social media streams. Financial services institutions want to use new customer data sources to hone decisions about credit limits, risk, and more.

Use of new algorithms or models: Typically, algorithm and model choice are a dynamic process. Data is used to train a model, and the model is then tested. As data changes, new models might need to be considered to better reflect the true state of matters. For example, many retailers and logistics companies had to rethink their models to overcome the disruptions brought on by the pandemic. A similar need for change occurs regularly to address normal market variations over time.

Modifications and addition of new features: Change is the only constant for modern applications. Users, whether employees or customers, expect a continuous stream of enhancements. And they are used to the consumer apps approach where their input is of value and frequently considered. They expect each next version of an application to address their concerns. This new paradigm's practical consequences are a continuous process where applications are deployed, modified, retested, and deployed anew.

Infrastructure issues

When intelligent applications are developed for a single group or have a narrow use case, most things can be kept local. For instance, a business unit can use on-premises servers and storage. Developing applications in this way limits access and scalability. Businesses today must incorporate AI/ML into all phases of their operations and across all business units.

Datasets used in intelligent applications are large and change rapidly change in size. Required storage and processing resources may start small when an application is being developed, and a model is trained but explode when the application is fully deployed. The dynamic nature of this process means businesses must avoid being locked into a specific architecture. They need the flexibility to run applications and store data on-premises, on private clouds, or on public clouds. Hybrid cloud options help avoid lock-in while giving the needed scalability and flexibility.

What's needed

Increasingly, cloud-native is the architecture of choice to build and deploy AI/ML-embedded applications. A cloud-native approach offers benefits to both the business and developers. With a cloud-native approach, applications or services are loosely coupled. Applications and processes are run in software containers as isolated units. Operations are managed by central orchestration processes to improve resource usage and reduce maintenance costs. These attributes enable a highly dynamic system composed of independent processes that work together to provide business value.

Fundamentally, a cloud-native architecture uses microservices and containers that use cloud-based platforms as the preferred deployment infrastructure.

Microservices provide the loosely coupled application architecture, which enables deployment in highly distributed patterns. Additionally, microservices support a growing ecosystem of solutions that can complement or extend a cloud platform. A cloud-native approach to intelligent applications uses containers to provide the underlying infrastructure and tools to use a microservices architecture.

Containers and Kubernetes

Developing, deploying, and maintaining AI/ML applications requires a lot of ongoing work. Containers offer a way for processes and applications to be bundled and run. They are portable and easy to scale. They can be used throughout an application's lifecycle from development to test to production. They also allow large applications to be broken into smaller components and presented to other applications as microservices.

Containers can use Kubernetes to automate the deployment and management of containerized applications. Specifically, Kubernetes provides service discovery and load balancing, storage orchestration, self-healing, automated rollouts and rollbacks, and more.

The industry has embraced Kubernetes as the dominant solution for container orchestration. Many consider it the de facto standard for container orchestration. Groups like the Cloud Native Computing Foundation (CNCF), which is backed by Google, AWS, Microsoft, IBM, Intel, Cisco, and Red Hat, have been Kubernetes proponents for years.

Benefits of a microservices approach

Microservices-based, cloud-native architectures also let businesses modify part of an existing app (e.g., switch from one analytics technique to another or make use of a new data source) without modifying other parts of the app. Additionally, businesses can move particular parts of an app, such as the ingestion and streaming analytics engine, from on-premises to the cloud or vice versa, depending on performance needs, costs, and the efficient use of resources.

Such an architecture lets businesses develop, deploy, move, and scale AI/ML workloads to match the application's requirements. A suitably selected cloud-native architecture based on microservices lets businesses deliver several capabilities, including:

- **Support different deployment options (on-premises, cloud, and edge):** For example, an IoT application might train a machine learning model using public cloud services, deploy that model to the edge device for real-time analysis of data as it is generated to take immediate actions, and look for trends and root causes of problems by analyzing historical data with on-premises systems.

- **Allow workloads to easily move around to match requirements:** For example, a business might want to use public cloud compute services to develop and test an AI-based analysis application but then run the application on-premises to meet regulatory requirements for data privacy and protection.

- **Speed the time to insights:** For example, an AI/ML-based object recognition application that performs image analysis might shorten the time from data ingestion to object identification by moving the application from a general-purpose compute system to one with highly optimized GPUs for faster image analysis.

Transition to real-time

Most business applications worked in batch mode. Data was collected and stored, analysis was applied, results were generated. The data seldom changed. Intelligent applications are radically different. They use real-time data that needs real-time analysis. New data sources and new models are used to compensate for disruptions that obviate the underlying data and assumptions used to create them.

These changes have significant implications on data pipelines. Data pipelines cover all the steps that data can go through throughout its life cycle. That is, everything from ingestion to transformation, including processing, storing, and archiving. Most pipelines involve different sources being merged or split into different destinations with multiple steps of transformations during these processes.

They must now become automated. Automation brings scalability and the ability to use different models for different business cases. Automation is necessary as the number of intelligent applications grows. Businesses do not have the resources for a data engineer or team of engineers to manually manage and prep data for every new application under development.

Characteristics and benefits of a production-class AI/ML environment

Cloud-native approaches based on microservices and Kubernetes offer great flexibility when developing, deploying, and maintaining applications with embedded AI/ML. What's often needed are enterprise-class services to ensure production-level quality. That means ensuring that the governance, security, and reliability properties and features are baked into enterprise Kubernetes.

Many organizations use Red Hat OpenShift® as this container platform because it offers the essential enterprise-class features and support for production applications. Specifically, Red Hat OpenShift is an enterprise-ready Kubernetes container platform with full-stack automated operations to manage hybrid cloud, multicloud, and edge deployments. Red Hat OpenShift is optimized to improve developer productivity and promote innovation.

Companies that use the vast array of commercial AI/ML solutions available on the market can leverage Red Hat OpenShift as an underlying platform to build distributed applications. Such an approach offers benefits to every group (developer, data engineers, data scientists, DevOps, and the business units) at every stage of an application's lifecycle.

For organizations that want to take an open-source approach, using the many AI/ML tools developed by the open-source community, more is needed. Specifically, businesses can greatly benefit by using an architecture that builds in the needed technologies for an end-to-end workflow, taking into account the lifecycle issues. That is the goal of Open Data Hub (ODH), a community effort that seeks to develop and share blueprints and frameworks that show how to use common AI / ML tools (e.g., Jupyter, TensorFlow, Seldon, and more) on top of a Kubernetes environment. Essentially, ODH lets businesses experiment and develop intelligent applications without incurring high costs and mastering AI and ML software stacks' complexity.

ODH offers common tools and helps with the integration of Kubernetes Operators that allow those tools and others to be easily used in an end-to-end intelligent application. ODH helps with deployment and lifecycle management in a hybrid cloud / container world. And there is an AI Library, which is an open-source collection of AI components and machine learning algorithms used for common use cases to allow rapid prototyping. There are pre-existing notebooks in the AI library for:

- Anomaly Detection
- Association Rule Learning
- Correlation Analysis
- Regression
- Flake Analysis
- Duplicate Bug Detection
- Fraud Detection
- Topic Modeling
- Matrix Factorization
- Sentiment Analysis

Additionally, ODH offers a collection of open-source tools and services that natively run on OpenShift and Red Hat products such as Red Hat Ceph Storage® and Red Hat AMQ Streams.

A typical end-to-end intelligent application workflow might start with data stored using Ceph. That data is then transformed, and models are created and trained. These processes might use Spark, TensorFlow, and Jupyter notebooks. The application could then use Kubeflow, which is dedicated to making deployments of machine learning (ML) workflows on Kubernetes simple, portable, and scalable. The models can then be deployed as a service.

Conclusion

Open Data Hub is an open-source community project that implements end-to-end workflows from data ingestion to transformation to model training and serving for AI and ML with containers on Kubernetes on OpenShift. It is a reference implementation on how to build an open AI/ML-as-a-service solution based on OpenShift with open-source tools, including TensorFlow, JupyterHub, Spark, and Kafka.

Using ODH helps address issues that arise in every stage of an intelligent application's lifecycle. It provides the data engineer with access to different data sources enabling the enforcement of data controls and governance. It gives the data scientist the freedom to access a variety of cutting-edge open-source frameworks and tools for AI/ML. And it provides DevOps with an easy way to manage an intelligent application's lifecycle, open-source components, and technologies.

To learn more about how ODH can help speed the development and testing of AI/ML embedded applications and ultimately make their lifecycle management easier, visit openshift.com/ai-ml.

RTInsights is an independent, expert-driven web resource for senior business and IT enterprise professionals in vertical industries. We help our readers understand how they can transform their businesses to higher-value outcomes and new business models with AI, real-time analytics, and IoT. We provide clarity and direction amid the often confusing array of approaches and vendor solutions. We provide our partners with a unique combination of services and deep domain expertise to improve their product marketing, lead generation, and thought leadership activity.

Red Hat is the world's leading provider of enterprise open source software solutions, using a community-powered approach to deliver reliable and high-performing Linux, hybrid cloud, container, eventing, and Kubernetes technologies. Red Hat helps customers develop cloud-native applications, integrate existing and new IT applications, and automate and manage complex environments. A trusted adviser to the Fortune 500, Red Hat provides award- winning support, training, and consulting services that bring the benefits of open innovation to any industry. Red Hat is a connective hub in a global network of enterprises, partners, and communities, helping organizations grow, transform, and prepare for the digital future.