Consuming Intelligence:

An approach for making AI models work for enterprises

June 2018

Abstract

This paper discusses an approach to implement artificial intelligence (AI) at the enterprise scale using AI Ops for ML and AI models. The main factors for success are the proper application of big data to model the creation, training, and automation of processes and containerization to deploy models into production. Distributed training is the key to accelerate time to market and realize rapid ROI.

Surya Prabha Vadlamani
prabha@novaiq.io

WHAT IT TAKES TO CONSUME INTELLIGENCE

Artificial Intelligence is no longer something restricted to research labs and academia. Thanks in part to advancements in technology, as well as readily available computing and storage power, AI now exists comfortably in the 'real world,' helping us address critical business questions.

Machine learning and AI models help transform the wealth of data that enterprises have accrued over decades into platforms that can be used to derive useful insights and provide real business value. The models become an integral part of the enterprise ecosystem, where the data is used to generate inferences and insights that are consumed through upstream and downstream applications. Data scientists and AI engineers make models appropriate for different business cases, which, through their specialization, provide accurate predictions and inferences. The models are not the sole part of the solution to the business case. Instead, they are an integral part of the complete solution. Similar to other enterprise platforms, the models progress through the development life cycle, from development sandboxes to the production-grade infrastructure, to consume the intelligence delivered by them.

DevOps is the combination of development philosophies, practices, and tools that increase an organization's ability to deliver applications and services at high velocity. This allows business to evolve and improve products at a faster pace than organizations using traditional software development and infrastructure management processes. Applied to the development and deployment of AI models into production, DevOps is often referred to as 'AI OPS'. It is part of an organization's digital transformation story, which also involves infrastructure engineering, cloud adaptation, API and microservice-driven architecture, continuous integration and continuous deployment (CI-CD), automation, and digitization of processes.

AI OPS FOCUS AREAS

- AI models require data to thrive and operate properly. Different types and foundations of data originating from multiple sources — such as databases, streaming networks, images, audio conversations, images, chat conversations — are collectively called 'big data'. Availability of clean and usable data is the essential factor in training production-ready AI models. Defining the right data strategy flow, (from identification to cleansing and finally to transformation) is the first step for the enterprise to begin their AI journey.

- Ideally, AI models should be trained on exceptionally large volumes of data to be able to generate accurate predictions and insights. Significant infrastructure will be required to train a model, demanding multiple servers equipped with multiple GPUs. Building a high performance training environment and providing easy and collaborative access to data scientists is essential. An enterprise's cloud strategy helps optimize the cost for training infrastructure, which can be used to facilitate automated provisioning of training instances on an as-needed basis.

- Building AI models requires a specialized set of skills, employed by a group known as data scientists. Data scientists build the models based on the available data and the pre-established business requirements. While building the model, they perform the necessary preparatory steps to prepare the data for the models. Externalizing and automating the data preparation routines form the essential component of the AI OPS pipeline.

- To be used at the enterprise level, the generated AI models have to be exposed to the enterprise in usable manner. This is accomplished by packaging the model(s) for consumption in a microservice with an API overlay. This makes them ideal candidates for containerization. A container is a stand-alone, executable suite of software that includes everything needed to run it: code, runtime, system tools, system libraries and settings, to name a few. Containerized software will always run the same, regardless of the environment. Containers isolate software from its surroundings and help reduce conflicts between teams running different software on the same infrastructure.

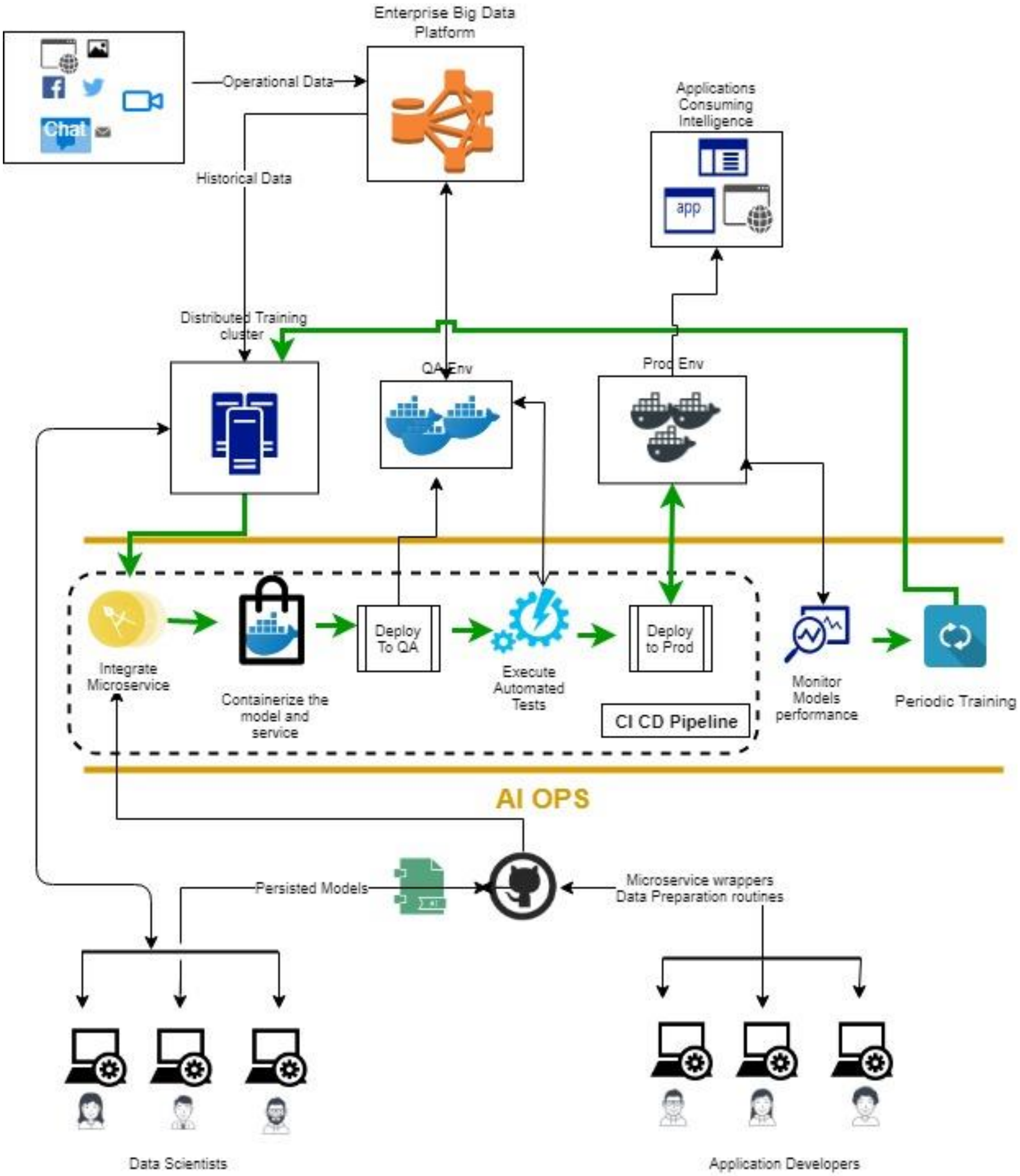
- AI and Machine Learning neural network models are generated after being trained on the initial available data and then deployed into production. The performance of the models can be measured based on the accuracy of the predictions and inferences made. The everyday predictions generated by the model are new data, which can also be used to further train the model, thereby improving the model's accuracy. AI OPS adds value by automating the process of automatically deploying new models, automating ongoing training, and gathering accuracy metrics.

SCALE OUT WHERE IT MATTERS

The most expensive step in building an AI model is the training. Training with very large data sets can be computationally demanding. Training is slow on a single machine because of the data volumes and 'hyperparameters,' which are the individual model parameters that require tuning for optimum performance. In addition, its iterative nature makes training time-consuming, which delays model deployment to production. Distributed computing on a GPU cluster across multiple machines is essential to optimize the training time and reduce duration.

An enterprise-level framework and tool kit for building AI models, coupled with a scalable training infrastructure, accelerates the availability of models for consumption. Open source libraries like Distributed Keras or Deeplearning4j, built on an Apache Spark distributing computing framework, are examples of tools that can train in a distributed environment.

STRINGING ALL TOGETHER



AI OPS Reference Architecture

CONCLUSION

A successful AI journey requires comprehensive AI OPS implementation to deploy the AI models in production. Only then can you truly realize business value. Data scientists and AI engineers should work independently throughout the complex process of building, packaging, and deploying production-grade models. The resulting intelligence should be ready for consumption in a continuous, automated manner and with minimal to no human intervention. The ability to provide a scalable and distributed training infrastructure, as well as execute automated periodic training of the models, forms the basis of continuous product improvement. As a trusted business partner, Nova IQ can enable your enterprise's digital transformation and help you embrace AI at scale.

ABOUT THE AUTHOR

Prabha is the Head of India Innovation Lab at nova IQ, a boutique solutions partner that focuses on solving business problems through disruptive technologies, with particular focus on AI and blockchain. She specializes in architecting and implementing DevOps for disparate technology and domain platforms, and leads AI OPS architecture for our customers. Prabha holds a master's degree in computer applications from Indira Gandhi National Open University.