Accelerating AI & Machine Learning Adoption at Scale:

Requirements for Success

WHITE PAPER

early adopter research

kinetica

Contents

early adopter research

Introduction

Even at companies with a mature data science team, the number of great ideas that make it out of the lab and into a production machine learning environment is shockingly low.

If AI is a revolution, why are most machine learning and artificial intelligence projects stuck in the pilot phase? Even at companies with a mature data science team, the number of great ideas that make it out of the lab and into a production machine learning environment is shockingly low.

A 2019 survey of 227 data scientists by Dimensional Research found that 78% of all AI projects stall before reaching production. Half of those interviewed had not yet gotten a project into production.

This leads us to ask why, despite ample funding, such a strategic program is nonetheless floundering at most organizations. And like most complex questions, it doesn't have a simple answer. Instead, we have identified seven areas where organizations are experiencing considerable difficulties in adopting AI and ML.

Barriers to AI & Machine Learning Success

There are a variety of areas where AI/ML projects get stuck or slowed down. Challenges include the following:

- **Production deployment difficulty.** It's hard to move AI/ML projects into production. The process demands attention to issues like scale, reliability, reproducibility, response time, and security.
- **Lack of support for new and emerging toolsets.** The rapid evolution of the field itself makes it challenging for enterprises to support the latest and most popular tools. Further, existing solutions quickly become stale as the availability of leading technology outpaces many organizations' development velocities.
- **Data problems.** Data preparation for AI/ML is not linear; it's a stop-and-start process that requires iterations, time, and context switches. Obtaining adequate and clean data for training poses challenges, as does refreshing data for AI and ML in production.
- **Outstanding risks.** Many organizations are understandably cautious about AI and want to ensure they can explain inferences, reproduce results, and minimize risks associated with deployment. Often data scientists admit that this is just not possible given current tooling and practices.

Data scientists are skilled at building and training models, but not typically at provisioning and running them in production, where issues like scale, reliability, response time, and security must be taken into account.

- **Unmonitored model evolution.** Many are familiar with the pressure to release a new version of an application, but few understand how AI/ML models must be monitored and how they can drift or simply stop working.
- **Lack of skilled resources.** AI expertise is rare, from model creation to training to evolution. Putting AI/ML into production requires multiple skillsets beyond data science, such as data engineering and systems engineering, which are equally difficult to staff.
- **Lagging computing power.** Larger datasets, which yield superior results, take longer to process and require more resources. High-performance computing, supported by GPUs, is necessary to speed up AI/ML deployments and test hypotheses, shrinking computing time from weeks or months to days, hours, or even minutes.

These are some of the reasons that AI and ML projects stall. Many models never deploy in production. These challenges hinder adoption of AI and ML.

Requirements for AI & Machine Learning Success

To break the current logjam, we need to take hand-coded, one-off processes and turn them into a pipeline that moves AI/ML seamlessly from the drawing board through training, and where appropriate, into production, scaling, and evolution. An enterprise AI/ML solution should address all of these challenges and lead to AI/ML deployments in production.

Outlined here are the requirements for an enterprise AI/ML solution.

Requirement: Easy ML production deployment

Data scientists are skilled at building and training models, but not typically at provisioning and running them in production, where issues like scale, reliability, response time, and security must be taken into account.

In many cases, putting a model into production requires running it in parallel, distributed across many processors to achieve production-level performance. It also requires managing how much computing power and memory the AI can consume.

An enterprise AI solution must empower data scientists and incorporate support for a range of existing and emerging frameworks.

An enterprise AI solution should offer tools that streamline resource configuration and handle machine learning orchestration. Such capabilities make it easier to serve models in production, scale out, manage upstream and downstream integrations and flows, recover from failures, and bring applications back online automatically.

Such a solution should empower data scientists to monitor their models in a production environment. An enterprise AI/ML solution must make it as easy as possible to move models from training into production, including feature engineering pre-processing steps.

Requirement: Flexible ML toolkit support

AI is a rapidly developing field. Hire a data scientist today and another in three months and they may well use different tools and frameworks—that is how fast the field is evolving.

Data scientists need the ability to use their preferred toolkits and open source frameworks. An enterprise AI solution must empower data scientists and incorporate support for a range of existing and emerging frameworks. Toolkits like RAPIDS are designed to increase model performance without the need to rewrite models in a low-level language like C.

Requirement: Unified data pipeline

In most cases, creating and deploying AI/ML requires many different platforms and context switches. Data scientists assemble and explore data in one environment, perform feature engineering in another, then develop the model, train the model, test the model, and finally hand it off for deployment to production. Each of these steps may involve iteration and necessitate going back to add or modify data and move the process forward again.

With current tools, data scientists find this process frustrating, forced to repeat their work with each iteration or refinement, and beset by delays moving data between platforms.

An enterprise solution demands an end-to-end product that integrates all of these steps, and supports them through the full ML lifecycle.

Opportunity: RAPIDS

What is RAPIDS?

RAPIDS, TensorFlow, Kubeflow, and PyTorch are examples of open source machine learning toolkits.

RAPIDS is a highly parallelized distributed machine learning toolkit that takes advantage of GPU memory and modern GPU chip facilities. GPUs crunch large volumes of data faster and more efficiently than CPUs because they work in parallel, instead of in sequence. RAPIDS uses features on the latest NVIDIA GPU chips, with thousands of processing cores available on a single card, to reduce AI/ML model training time.

RAPIDS is a toolset that lets data scientists run machine learning models on very high performance computing platforms using familiar tools and the same interfaces as existing libraries, such as XGBoost. RAPIDS implements these libraries in a way that takes advantage of high performance GPUs and obviates the need to rewrite models to run them in production.

While RAPIDS intends to be a best-in-class machine learning toolkit and library, it is not a complete enterprise solution. As with all open source frameworks, implementation requires expertise.

Why is RAPIDS important for businesses considering new AI/ML projects?

RAPIDS includes Dask, an execution framework to distribute work across many machines. NVIDIA added support for a number of existing models, including linear regression, logistic regression, classification, PageRank, and others, all re-implemented to run optimally on multiple cards and multiple machines, while also utilizing GPU memory for intermediate data.

Take for example XGBoost, a popular machine learning algorithm. RAPIDS is effective for running XGBoost in a distributed fashion. Companies struggle to run XGBoost on a single machine because it is memory intensive. With RAPIDS, you can distribute XGBoost across many machines, adding machines until you have enough processing capacity for all your data.

Rather than relying on data scientists to apply best practices on their own, the underlying platform should make such practices a given, so data scientists can periodically evaluate their models and determine whether they should evolve or be replaced.

Requirement: Transparent risk and compliance

Questions have been raised and will be raised about whether training data is biased and why models deliver certain outcomes.

An end-to-end solution should allow users to see how models were created, what results came out, and what inputs drove those results. Auditing makes it possible to trace results back to their origins, showing the provenance of decisions made by AI/ML models.

The ability to trace the lineage of the model's decision-making is important not just today, but in the future, when regulation, legal compliance, or litigation may require it. Few platforms offer this essential functionality.

Requirement: Integrated model monitoring

Given that models run on ever-changing data, data scientists must review and revisit models periodically to ensure they are working properly.

An enterprise AI/ML platform should capture results over time and persist them for later review.

Few today persist the results of AI/ML models, and if they do, it is done in an ad hoc manner. Given the need to revisit models, an enterprise solution should store results in a format that is cost-effective and easy to review.

Rather than relying on data scientists to apply best practices on their own, the underlying platform should make such practices a given, so data scientists can periodically evaluate their models and determine whether they should evolve or be replaced.

AI and Ethics: Everyone's Concern

Adopting AI requires careful consideration because applying this technology has significant implications for society as a whole.

AI is fraught with ethical questions. AI misuse can happen as a result of malintent, mistake, or mere mindlessness. Even the unconscious biases of data scientists can be passed on to their models.

We can't avoid the thorny questions AI raises; we must collaborate in responding to them because they have societal implications that go well beyond the boardroom's bottom line.. To that end, Kinetica and other forward-looking companies are participating in the World Economic Forum's initiative to promote AI governance. We invite you to join us.

Kinetica offers an Active Analytics Platform that unifies the ML process and satisfies all of the requirements of an enterprise-grade AI/ML solution.

Requirement: Inclusive user interface

AI can't be a revolution if more people—particularly people with domain knowledge—can't participate. Powerhouse modeling systems such as TensorFlow, RAPIDS, and PyTorch don't come with a user interface (UI); instead, they require coding or embedding to work. An end-to-end platform should democratize access to frameworks like RAPIDS and TensorFlow by offering an accessible UI across the entire machine learning life cycle, from model import to training, deployment to archiving. delivering support frameworks like RAPIDS and TensorFlow.

Requirement: Accelerated AI/ML Processing

More relevant data produces better models. A solid AI/ML solution should offer high-speed data processing, ultimately accelerating model training and delivery.

Kinetica eliminates typical model latency by running models on streaming data instead of subsets of historical data.

Kinetica: Unified AI/ML Solution

The Kinetica Active Analytics Platform brings order to the Wild West that is AI and ML today. Models today are created by passing flat files to a data scientist at a workstation. There's no process in place to then bring the model back to the organization to implement. And once the model is incorporated into data processes, it remains a black box, with no transparency about how it was created or where it's been implemented. How is a CDO supposed to get a handle on today's data science practices? Kinetica offers an Active Analytics Platform that unifies the ML process and satisfies all of the requirements of an enterprise-grade AI/ML solution.

Holistic Overview

With Kinetica, you can see all of your models from a central view. It gives you a transparent overview of your enterprise ML strategy at scale, making model deployment manageable.

Kinetica streamlines deploying AI/ML models in production. It handles all deployment tasks including configuration of resources. Kinetica lets you allocate resources like memory, disk, and GPU or CPU cycles.

Kinetica performs all of the low-level pipeline work as well as machine learning orchestration. Kinetica supports multiple ML toolkits, such as TensorFlow (which uses Kubernetes as its execution framework) and NVIDIA RAPIDS (which uses Dask). Whatever ML toolkit you select, Kinetica manages the orchestration for you.

Ultimately, Kinetica delivers a process that works for hundreds of models at scale and hundreds of applications across the business.

Unparalleled Speed

Older data means outdated results. Kinetica eliminates typical model latency by running models on streaming data instead of subsets of historical data.

Kinetica seamlessly integrates machine learning models and algorithms with your data, for real-time predictive analytics at scale.

Kinetica can do this because our Active Analytics Platform is GPU-accelerated. Inferencing on GPUs is faster and diminishes model latency. Instead of looking at historical subsets of data, Kinetica's GPU-accelerated Active Analytics Platform can run models on live, streaming data. Kinetica can inference in the moment—much more relevant for today's challenges, from dynamic inventory replenishment to targeted advertising to financial risk management. It is thus poised to help organizations take advantage of ML toolkits like NVIDIA RAPIDS that are optimized for GPUs.

Faster Development

Much of the challenge of AI involves data: discovering it, assembling it, training it, and creating a pipeline from the data to the model. By working with a unified platform, data scientists enjoy a seamless experience, from exploration to training to production. Utilizing a single platform removes much of the friction from the machine learning lifecycle, which cuts your development cycle time and increases the number of ideas that you can test.

Kinetica stores the data itself, making it natural to create a pipeline that goes from preparing and assembling data to executing the model in production. This approach accelerates development cycles because, with everything in one platform, data scientists do not need to move data from one tool to another as they develop and iterate on their models. Because the data is stored in Kinetica, if a data scientist decides to go back and do more feature engineering, it does not require starting over or switching contexts.

Since Kinetica is GPU-accelerated, processor-intensive activities like model training do not require data scientists to find computing resources; Kinetica is designed to handle it.

In this way, Kinetica alleviates much of the mundane work machine learning pipelines require, freeing everyone up for value-added activities.

In essence, when you work in the Kinetica platform, Kinetica tees up the next steps for you. As input comes in, it goes through feature discovery and transformation. All of this happens in-line, no matter how complex the data pipeline is, because it takes place directly within the platform.

Many of the challenges hampering AI adoption can be addressed with an integrated platform like Kinetica.

Auditing for Answers

Achieve auditability and traceability by snapshotting key artifacts that capture the training data, the test data, and the data used in production. Keep records of these system inputs in case questions arise about why the model behaved as it did, and evaluate any biases in training data.

This is a critical compliance and risk feature.

Kinetica seamlessly integrates machine learning models and algorithms with your data, for real-time predictive analytics at scale. By unifying the traditionally siloed workflows of analytics and model execution, Kinetica dramatically simplifies data engineering and active analytical application development. Kinetica does this with:

- **Machine learning-powered analytics.** Organizations can embed machine learning and advanced algorithms into their active analytical applications without the headache of complex data engineering, migrating data between disparate systems. Kinetica can also import popular pre-trained models.
- **A "Bring Your Own Algorithm" approach.** Take the model to the data, not the data to the model. Bring existing models and analytics as containers and embed them into your analytical workflows and applications without the heavy lifting of migrating data to and from siloed model execution environments.
- **Automated deployment and data orchestration.** Kinetica automates model deployment on Kubernetes in continuous, on-demand, or batch modes. No need to worry about deployment, network configuration, or scaling. Once deployed, Kinetica automatically orchestrates the full analytical pipeline—from ingest, to database, to model, and back to the database and downstream applications.
- **Push-button distributed training.** Explore data interactively, at scale, across dimensions to find patterns. Then rapidly experiment with built-in support for the most popular TensorFlow templates and fully automated distributed training.
- **Support for model audits.** Kinetica makes it possible to track, govern, and audit data that's part of your analytics and ML workloads. Kinetica tracks the full data lineage, including raw data, feature transformations, and model output. With Kinetica, you can perform a full model audit or search to find a needle in a haystack by auditing a specific inference.

Kinetica speeds the process by providing the following support for ML projects:

- Best-in-class GPU-accelerated OLAP to eliminate data extraction steps, speed up data preparation, and dramatically shorten experimentation and training cycle time.
- Tiered storage that automatically migrates data from cold storage to GPU VRAM, greatly expanding data scientists' ability to work with large-scale data.
- GPU-accelerated 2/3D visualization that lets data scientists rapidly explore and evaluate large-scale data, including geographic and time-series data. The integrated platform allows seamless model evaluation in development and production.
- Through regular ANSI SQL queries, Kinetica intelligently places dataframes in GPU memory to dramatically speed up ML feature engineering and distributed training.
- High-speed parallel streaming data ingest
- Enterprise-grade security and reliability

Many of the challenges hampering AI adoption can be addressed with an integrated platform like Kinetica. Reach out to us to learn more about how Kinetica can help you accelerate AI adoption in your organization.

This paper was written by Early Adopter Research and sponsored by Kinetica

Learn more about the Kinetica Active Analytics Platform or request a live demo.

Connect with us

early adopter research