AI done right:

With a strong foundation, everyone can map a successful journey—including you

How the IBM Data-Train-Inference AI model positions your enterprise for long-term success

Overview

What if you discovered that a data science background isn't necessary to understand, interpret and act on the most complicated aspects of enterprise AI?

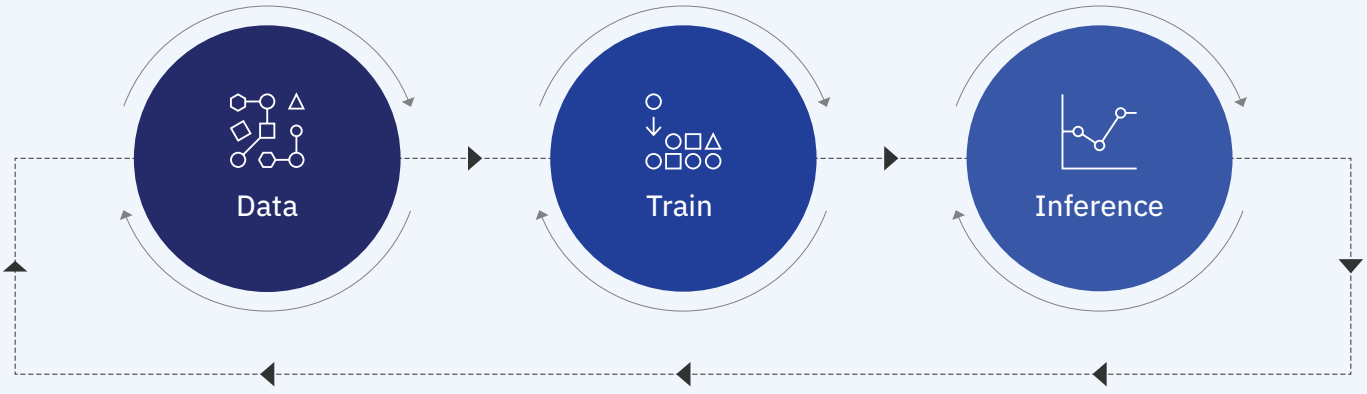
How would that knowledge impact the business processes and applications your organization relies on to remain competitive?

Enterprise AI has traditionally been the focus of analytics experts with a deep understanding of model building and training. But that's all changing as AI-driven initiatives extend across the business. And at the forefront of this shift is a value-based framework for AI efforts. It's called the Data-Train-Inference (DTI) AI model, and explaining it is the goal of this paper.

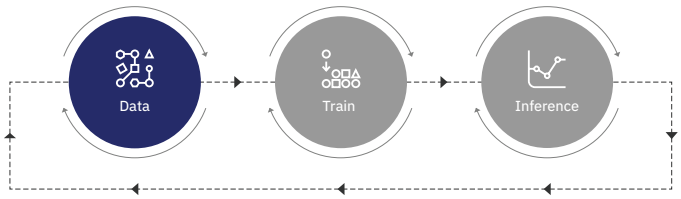
Before diving into specifics, it is important to know that the DTI model is not a linear workflow. It is instead a continuous loop consisting of three stages that interact at all times. And because the process is ongoing, the insights extracted are richer and more valuable.

The result? Key stakeholders can make smarter decisions faster—and with more confidence.

The Data-Train-Inference (DTI) AI model



Phase 1: Data



AI experts point to the data phase as the most time-consuming of the three phases. The amount of work required to prepare existing data for ingestion—the common term for loading data into an AI model for training—is intensive.

As the old saying goes, garbage in means garbage out. This is truer than ever in the realm of AI. Beginning with data that is questionable in quality, veracity or even quantity will create questionable AI models that generate dubious results for the business. Not starting with a solid data foundation will send your AI projects down the wrong path before they've even started.

> "Data that is questionable in quality, veracity or even quantity will create questionable AI models, generating dubious results."

Did you know?

The vast swaths of data that permeate our world, paired with the rapidly increasing capabilities of our server infrastructure, has ignited the AI revolution in the last decade.

Data that are used for AI training come from all sorts of disparate sources. These may be familiar and well-understood sources, such as previous sales or customer numbers from the existing enterprise data warehouse. They may also come in real time from sources like Internet of Things devices at the edge or other Internet data streams, such as Twitter.

The four truths about data

First, data comes from many different sources. They could be any unprocessed format such as text, image, sound or raw numeric values. Data science teams spend tremendous amounts of time gathering data, then cleaning up these data into correct formats to be consumed by frameworks and brought into the environment. These are important steps. Analyzing and identifying specific features of data sets are important to models, which influence the results that drive business value.

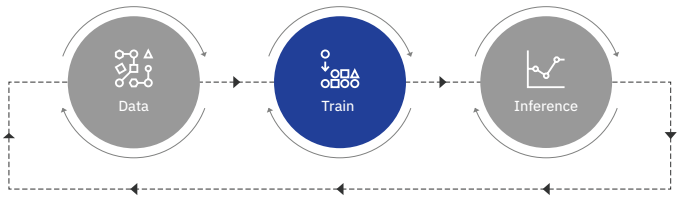
Second, a lot of data can be a blessing and a curse. Data in the enterprise is often fragmented or exists in several places. Accurate scientific conclusions often come from the sum of several different sources, but duplication can lead to unexpected and negative outcomes.

Third, timeliness of data is paramount. With ground conditions potentially changing rapidly, AI models need to be fed up-to-date data to continue driving business value. Without fresh data, the value of the model will suffer. Models are only as “fresh” as the day their underlying data were brought into production and need to be continuously updated with new data. This points to the cyclical nature of this step itself and the entire Data-Train-Inference model. In order to maintain relevancy, a plan must be in place to continuously refresh underlying data sets for training models.

Finally, moving massive amounts of data through the data center requires unique compute infrastructure. Without the proper infrastructure for this task, your AI workflow suffers slowdowns in throughput and performance before you’ve even reached the most demanding step—training.

Bottom line? Clean, relevant and fresh data are a key for unlocking valuable insights. All of this work comes before you even spin up your first training workload.

Phase 2: Train



Training is the phase of the AI workflow that most non-data scientists consider when they hear about AI workloads. That is not an entirely inaccurate view. Training is where the magic of artificial intelligence occurs—where data become AI models.

Without diving deep into the theory behind machine learning, deep learning and artificial intelligence, training can be summarized as an iterative process in which the data from the previous step are used to create models. And those models make future predictions on similar data out in the real world. Only in the last 10 years has it been possible to solve problems in this manner, thanks to the advent of graphics processing units, or GPUs.

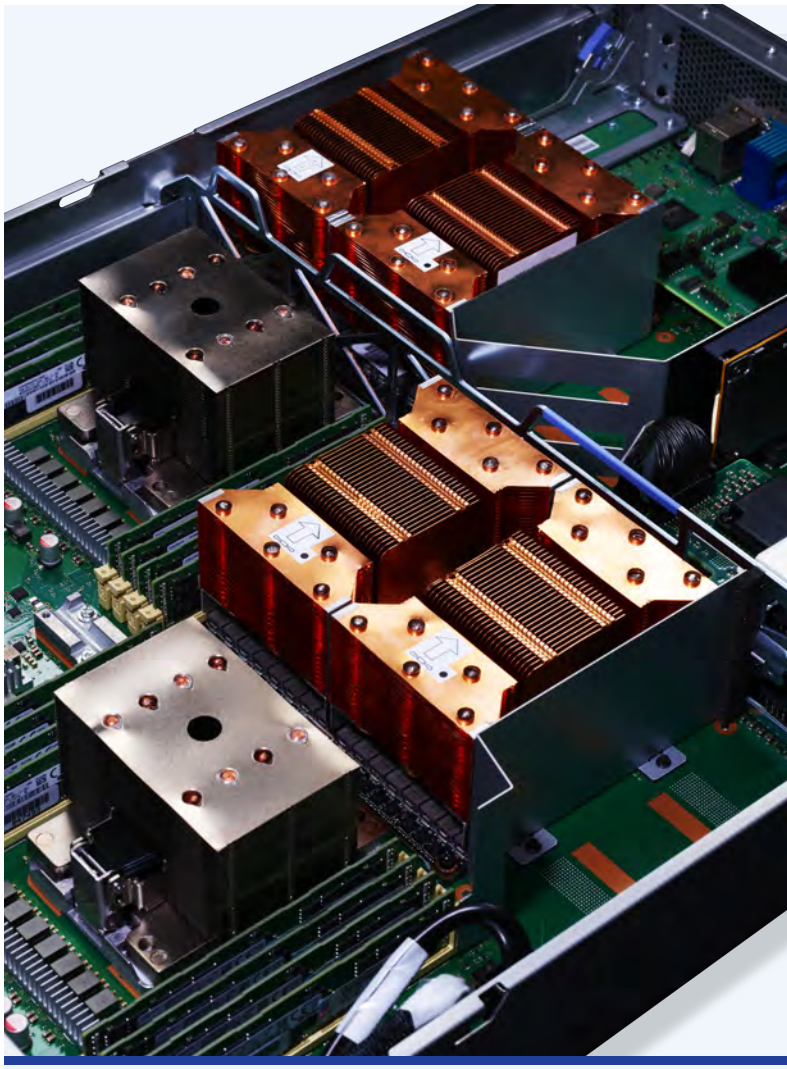
Servers that leverage both central processing units (CPUs) and these new GPUs are considered "accelerated." As traditionally CPU-centric data centers are taking on more AI workloads, it is necessary to augment the data center with accelerator-based servers. This more powerful compute also carries a higher resource cost. That's why correctly allocating these expensive resources is critical. Incorrect allocation can quickly doom your AI project.

"Training is where the magic of artificial intelligence occurs—where data become AI models."

The need for speed (and accuracy)

Even under ideal conditions, training on just one model could take days, weeks or even months. Additionally, the average models are typically trained five to six times before being deployed to production. Accelerating performance for training a model is invaluable, but you need a combination of speed and accuracy to drive key value for time to market.

One of the most time-consuming tasks in the training flow is setting and resetting hyperparameters for models. Hyperparameters are values the data scientist chooses for the model before training begins. Modern models can have hundreds of them. The iterative setting and resetting process can take hours, even with model runs using sample data sets. Automating hyperparameter searches, and running these searches in parallel, can save your data scientists weeks or even months—and shorten time to result and time to accuracy.



"Accelerating performance for training a model is invaluable, but you need a combination of speed and accuracy to drive key value for time to market."

Did you know?

Days, weeks or months could be wasted if the data scientist on the model is unable to determine the early success of the model parameters.

Tools such as training visualization allow data scientists to see training progress and provide alerts that training is not converging. Data teams can stop the job, re-adjust parameters and restart the job within the first few hours of the training, instead of waiting until the end to see the poor result.

‘An age-old conflict’

Unlike traditional code, AI models drift from the underlying data over time if they are not re-trained on fresh data. Therefore, any existing model must consistently be retrained in order to maintain its relevancy and usefulness. However, you must also be able to quickly bring up and promote new models into production.

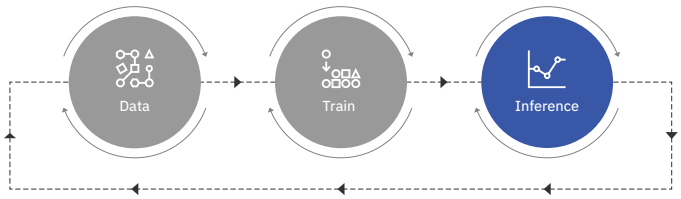
This places IT leaders at the center of an age-old conflict around resource allocation. Some workloads or tenants are more important than others. Or, they’re subject to more aggressive service level agreements (SLAs). This should be reflected in the resource scheduling layer. A faster and more efficient scheduler helps you achieve business-relevant levels of model accuracy more quickly.

Without it, chaos ensues. Each data scientist or project is confined to a box, ultimately creating a silo of compute. This limits the user to a single machine that he cannot flex beyond. It also wastes resources when that single user is not using the machine.

Fair sharing and priority-based scheduling supports dynamic sharing of GPU resources among multiple training jobs. It also enables pre-emption and reclamation of GPUs without killing any jobs. This keeps your expensive data science teams productive; no one will be blocked or starved from GPU resource. Instead, you can continue to drive utilization to maximize the usage of those costly resources because they have elasticity to swap GPUs from workload to workload.

Mindfulness of the high-value nature of training resources and the effort necessary to ensure success in this phase will yield the ultimate outcome: A model that derives value for the business. However, the job is not done once the model is complete. Now it must be deployed.

Phase 3: Inference



In AI, deployment to production is the stage at which you can derive insights from your model. This is known as inference. (Some also call it scoring.) This is where the value of deep learning comes to life and where more advanced AI concepts, like explainability and fairness metrics, can be examined.

The inference phase is really the sum of all prior parts. If your data was bad or the training was inaccurate, inference will suffer. Without proper inference, all prior efforts are for naught.

The primary challenge in this phase is different than the training phase. While training can take place over many cycles, consuming days or weeks of project time, inference is often a sub-second process that demands quick, accurate insight.

“The inference phase is really the sum of all prior parts. Without proper inference, all prior efforts are for naught.”

An example of inference in action

Consider a payment processor that has trained a model to detect fraud occurring in consumer transactions on the platform. Customers will not tolerate undue delay in processing their payments. Therefore, the processor must place a sub-second SLA on the AI model's insight to ensure every transaction progresses smoothly and customer experience is not affected by the fraudulent check.

To overcome those challenges, the underlying hardware must also be different. While training is done centrally in data centers, inference is most commonly performed at the edge, on devices such as smartphones, or near-edge.

An example of a near-edge scenario is a small server rack running in a retail store. More specifically, the server is generating real-time insights on customer transactions or video feeds in the store.

With proper resource scheduling and resource allocation, you can more quickly extract insight from models. With a seamless scale-up strategy, you can swiftly scale up inference needs on-premises or at the edge to handle demand. Similar to the training stage, the ability to elastically move inference tasks in a common resource pool can assist in meeting aggressive SLAs.

As indicated by the looping arrow in the DTI framework, data gathered in the real world via inferencing is fed back into the workflow at the Data phase. This looping action continuously improves model accuracy because deeper and fresher underlying data is applied. And thus, the cycle starts again.

Bringing the model together

At each phase of the AI workflow, it is necessary to have a combination of the right people, processes, and infrastructure (both hardware and software) to be successful. These are the critical components to building a strong foundation—the key to deploying AI across your entire business.

IBM has AI infrastructure that adapts to your changing business priorities, so your organization achieves its goals at every phase of the AI journey. And IBM Power Systems provides industry-leading, purpose-built enterprise AI infrastructure for machine learning, deep learning, and inference. With it, you can:

- Fuel new thinking and capabilities across your organization.
- Drive greater confidence in business decisions at scale.
- Make the best use of people, processes and processors, with a solution designed to grow with your organization.
- Find meaningful results faster with the industry's highest data throughput and IBM research, keeping you on the cutting edge of AI technology.

All of these benefits are provided on top of the proven security of Power Systems, which seamlessly integrates open-source frameworks secured by IBM.

AI done right pays enormous dividends. Now that you know what it takes to consistently chart a successful AI course, only one question remains:

Are you ready to get started?

Visit:

ibm.biz/EnterpriseAI

Produced in the United States of America
July 2019

74027174USEN-00