Predicts 2021: Operational AI Infrastructure and Enabling AI Orchestration Platforms

Published 2 December 2020 - ID G00735813 - 12 min read

By Analyst(s): Chirag Dekate, Soyeb Barot, Sumit Agarwal

Initiatives: Artificial Intelligence

The use of artificial intelligence in enterprises has tripled during the past two years, requiring IT leaders to reevaluate their core infrastructures and optimize for AI productivity.Data and analytics leaders need to devise AI orchestration platforms to accelerate and sustain AI operationalization.

Additional Perspectives

- Summary Translation + Localization: Predicts 2021: Operational AI Infrastructure and Enabling AI Orchestration Platforms
(26 February 2021)

More on This Topic
This is part of an in-depth collection of research. See the collection:

- Over 100 Data and Analytics Predictions Through 2025

Overview

Key Findings

- Aligning the data, data science and machine learning pipelines alongside the application deployment process is fundamental to the continuous delivery and integration of periodically enhanced ML models in AI-based solutions. This requires leveraging DataOps, MLOps and Platform Ops for AI to scale the AI architecture. Hence, there is an emergence of AI orchestration platforms to operationalize AI architectures.

- With growing AI adoption in their environments, IT leaders are leveraging compute-accelerated AI orchestration platforms via Platform Ops for AI initiatives across hybrid multicloud, edge and Internet of Things environments.

- Open-source development platforms, such Jupyter Notebooks, scikit-learn, TensorFlow, Keras, MLflow and Kubeflow, are AI platform components and foundational building blocks of AI solution development and deployment.

Recommendations

IT leaders responsible for scaling AI initiatives should:

- Secure, monitor and put a governance mechanism in place by leveraging AI orchestration platforms for the development, delivery and management of AI solutions.

- Select accelerators and infrastructure technologies that support the widest range of your data and ML pipeline toolchain by curating accelerated data pipelines, ML pipelines and model deployment environments that leverage compute infrastructures.

- Implement a composable, flexible, reliable, maintainable, scalable and interoperable AI orchestration platform by integrating open-source and commercial components.

Strategic Planning Assumptions

By 2025, 50% of enterprises will have devised artificial intelligence (AI) orchestration platforms to operationalize AI, up from fewer than 10% in 2020.

By 2025, AI will be the top category driving infrastructure decisions, due to the maturation of the AI market, resulting in a tenfold growth in compute requirements.

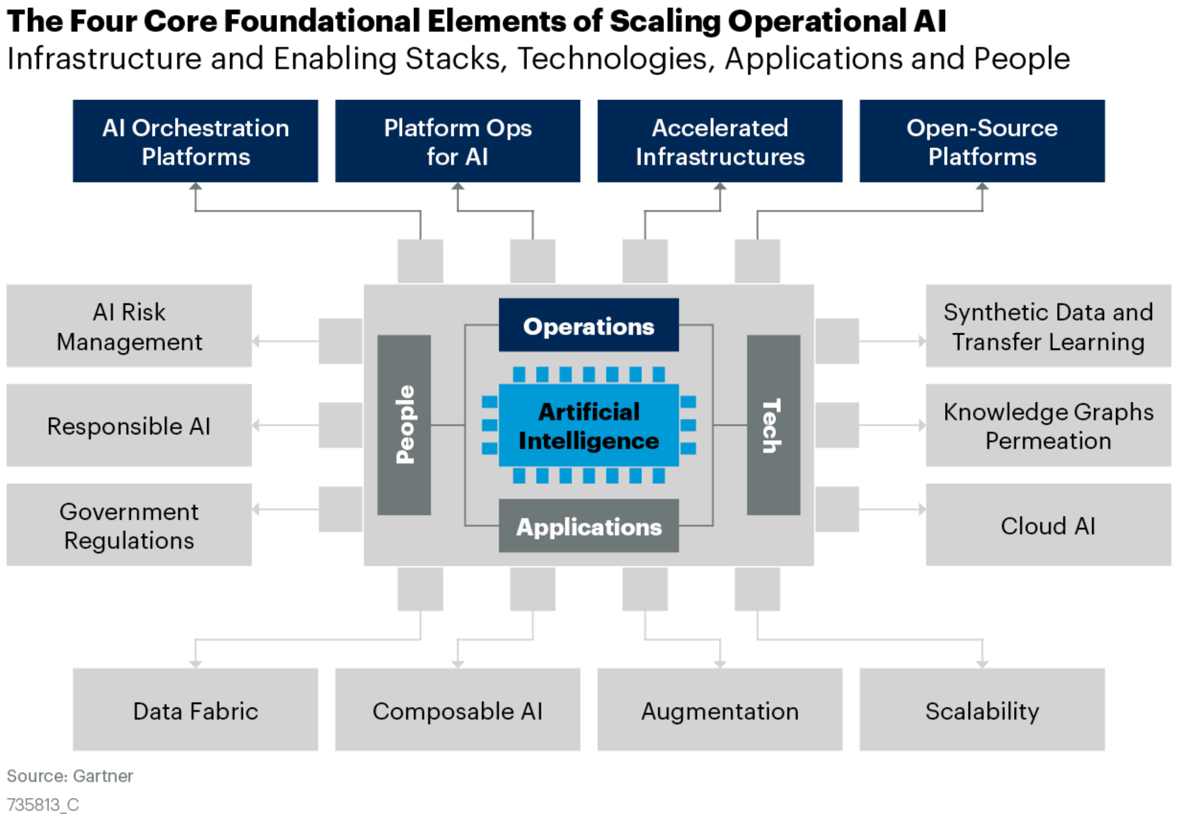
By 2025, 50% of enterprises implementing AI orchestration platforms will use open-source technologies, alongside proprietary vendor offerings, to deliver state-of-the-art AI capabilities.

Analysis

What You Need to Know

Gartner's operational AI infrastructure and enabling stacks predictions for 2021 are focused on the urgency and need for enterprises to devise AI orchestration platforms that accelerate the productionalization of AI (see Figure 1).

Figure 1: The Four Core Foundational Elements of Scaling Operational AI: Infrastructure and Enabling Stacks, Technologies, Applications and People


The Four Core Foundational Elements of Scaling Operational AI
Infrastructure and Enabling Stacks, Technologies, Applications and People
AI Orchestration Platforms
Platform Ops for AI
Accelerated Infrastructures
Open-Source Platforms
AI Risk Management
Responsible AI
Government Regulations
Operations
People
Artificial Intelligence
Tech
Applications
Synthetic Data and Transfer Learning
Knowledge Graphs Permeation
Cloud AI
Data Fabric
Composable AI
Augmentation
Scalability

Source: Gartner
735813_C

Gartner has curated three predictions around the need for enterprises to devise AI orchestration platforms.

First, the successful scaling of AI from pilots into productions, for not just tens of models, but thousands, requires AI orchestration platforms that are adapted to enterprise contexts. Maturation of AI orchestration platforms will create new opportunities for vendors and enterprises resulting in new tools and practices.

Second, the rapid maturation of the production AI will drive significant growth in compute resources across training and deployment ecosystems. Scaling the number of deployed models across hybrid, multicloud, edge and Internet of Things (IoT) contexts will require accelerated compute resources — graphics processing unit (GPU), field-programmable gate array (FPGA), application-specific integrated circuit (ASIC)-augmented — designed to improve productivity.

Third, enterprise-devising, AI orchestration platforms, will leverage open-source technologies to complement proprietary data science platforms. IT leaders should plan early to establish AI orchestration platforms in their enterprises, because these initiatives take time to deliver meaningful results. Gartner inquiries and surveys indicate that enterprises continue to struggle in production AI. Nearly 50% of AI projects never make it into production (see Survey Analysis: Moving AI Projects From Prototype to Production). IT leaders across enterprises, are focused on evolving beyond AI ideation, to operationalizing AI into enterprise-relevant applications. Successful enterprises, stringently measure outcomes, hire and engage diverse skills, and necessitate close alignment across business and IT organizations (see Survey Analysis: Moving AI Projects From Prototype to Production).

As enterprises accelerate production AI, IT leaders will need to devise AI orchestration platforms to scale operational AI. Curating AI orchestration platforms involves nurturing multiple best practices, across data, ML, AI and application development to create an efficient delivery model for AI-based systems. AI orchestration platforms enable enterprises to:

- Architect AI-augmented systems that are resilient and that accept frequent changes in data and model contexts.
- Provide autonomy to business units by enabling discoverable and reusable AI artifacts across the enterprise.
- Scale AI initiatives by modularizing and orchestrating the underlying platforms enabling autonomy to data engineers, data scientists and enterprise architects.

Strategic Planning Assumptions

Strategic Planning Assumption: By 2025, 50% of enterprises will have devised AI orchestration platforms to operationalize AI, up from fewer than 10% in 2020.

Analysis by: *Soyeb Barot, Sumit Agarwal, Chirag Dekate*

Key Findings:

As businesses develop their AI initiatives, AI is becoming a core part of their composable business architectures. IT leaders need to streamline the development, deployment and management of AI solutions and the underlying platforms that help build the end-user applications. Streamlining curation and deployment of AI solutions requires integrating specialized technology, the multiplicity of ML models, AI services, along with the detail and volume of insights the enterprise's data has to offer. Hence, the implementation of an AI operational platform becomes critical to scale AI initiatives:

- The need for operational resiliency across technology and service delivery with AI has never been greater. Organizations that deliver highly digitized business capabilities in a composable and modular manner have the biggest opportunities.
- Aligning the data, data science, and ML pipelines alongside the application deployment process is fundamental to the continuous delivery and continuous integration of periodically enhanced ML models within AI-based solutions. This requires leveraging DataOps, MLOps and Platform Ops for AI to scale the AI architecture. Hence, AI orchestration platforms to operationalize AI are emerging.
- IT leaders need to focus on integrating digital technologies, compute infrastructure and, most notably, the process workflow among the data engineering, data science, ML engineering, and application engineering teams to support the AI orchestration framework.
- Most AI orchestration platforms have been built using a combination of open-source toolkits, such as Kubeflow, Jenkins, Kubernetes, MLflow and Seldon, alongside commercial data science and machine learning (DSML), and MLOps tools.

Market Implications:

The mismatch between processes and tools used across data engineering, model building and application development stages causes significant impedance in the scalability of AI implementations. AI orchestration platforms leverage existing investments and offer a way to tackle siloed DataOps, MLOps, and applications ecosystems, countering data drift and concept drift within the underlying models that drive AI solutions.

The purpose of AI orchestration platforms is to provide an integrated continuous integration/continuous delivery (CI/CD) pipeline across all of the different stages of building AI-based systems — supporting reproducibility, reusability, rollback/rollout, lineage and secure environment. These orchestration platforms are the backbone of some of the leading innovative tech organizations, which have also stepped forth and open-sourced in-house AI orchestration platforms to benefit the broader technology community. This will lead to the mature marketplace of AI platforms and accelerate the delivery and adoption of AI-based solutions.

Recommendations:

IT leaders responsible for scaling AI initiatives should:

- Identify a process and set of common toolkits to support the overall orchestration of the AI platform. Start with evaluating the current process, tools, framework, and operating model used across their data engineering, data science, ML and application development teams. Resist the urge to rely on a single commercial platform that would support the development, deployment and operationalization. The market is still going through a maturity phase, and the most successful implementations seen yet have been a best-of-breed combination of open-source and commercial DSML solutions.
- Not look at AI orchestration platforms merely as a tool or a product. Rather, view it as a framework and operating model that brings the people, process, and technology components together as part of the composable system design principles for scaling AI initiatives.
- Leverage AI orchestration platforms to secure, monitor and put a governance mechanism in place for the development, delivery and management of AI solutions to ensure they continue to deliver value, further the implementation roadmap and meet the overall corporate objective.

Related Research:

- 2021 Planning Guide for Data Analytics and Artificial Intelligence
- A Guidance Framework for Operationalizing Machine Learning
- Solution Path for Building an Effective Technical AI Strategy

Analysis by: *Chirag Dekate, Soyeb Barot, Sumit Agarwal*

Key Findings:

- The adoption of AI continues to accelerate with enterprises leveraging a broad mix of ML and deep learning techniques.
- With growing AI adoption across their environments, enterprises will leverage compute-accelerated AI orchestration platforms via Platform Ops for AI initiatives across hybrid multicloud, edge and IoT environments.
- Accelerated compute resources drive disruptive productivity across training and inference stages by enabling near-real-time training and high-throughput, latency-sensitive inference across hybrid multicloud, edge and IoT environments.
- Enterprises are following a similar path as digital-native organizations (Google, Facebook and Baidu) in scaling training and the deployment of thousands of models using infrastructure as a service (IaaS), SaaS, platform as a service (PaaS) and API strategies. Enterprises are looking to develop AI orchestration platforms that are adapted to their enterprise context (infrastructures, skills, middlewares and applications).

Market Implications:

Primary forms of accelerated computing for AI comprise leveraging hardware systems and technologies across cloud and on-premises environments. The accelerator landscape is evolving with vendors offering capabilities across deep neural network ASICs, GPUs, neuromorphic processors, CPUs and, in some cases, FPGAs.

Enterprises seeking to develop leadership-class capabilities in AI will need to curate AI orchestration platforms that support accelerated scaling from tens of projects to thousands of trained and deployed models. The synergy of computation-intensive (compute, storage, networking, edge and IoT) infrastructure stacks across deployment contexts (on-premises or cloud) will have an accretive effect, resulting in tenfold compute requirement growth in terms of capability through 2025. For data and analytics leaders (and supporting IT stakeholders), this means devising AI orchestration platforms that utilize accelerated infrastructures across data pipelines, ML pipelines and deployment environments.

Recommendations:

IT leaders devising initiatives to operationalize AI should:

- Use accelerated compute and data infrastructures to support multiple data and model pipelines across enterprise.
- Select accelerators and infrastructure technologies that support the widest range of your data and ML pipeline toolchains.
- Accelerate productivity by curating AI orchestration platforms that offer the same set of accelerated compute infrastructures across hybrid multicloud, edge and IoT deployment contexts. This will help minimize risk and technology complexity, while maximizing the productivity of the teams involved.

Related Research:

- Emerging Technologies: Critical Insights on AI Semiconductors for Endpoint and Edge Computing
- Product Managers Developing Edge Systems Will Require Integration of AI Accelerator Chips
- Hype Cycle for Artificial Intelligence, 2020

Analysis by: *Sumit Agarwal, Soyeb Barot, Chirag Dekate*

Key Findings:

- The open-source movement has made several innovative and foundational contributions to the building blocks of AI solution development and deployment. Jupyter Notebooks, scikit-learn, TensorFlow, Keras, MLflow and Kubeflow are examples of AI platform components.
- Hyperscalers, and large technology organizations have existing open-source, in-house capabilities or contributed by providing developers and leadership to several AI open-source projects (e.g., Airbnb, Airflow), Google (e.g., Kubeflow, TFX).
- Pioneering financial services, manufacturing and telecommunications enterprises have successfully implemented AI orchestration platforms, primarily using open-source components.
- Although innovative, open-source AI solutions at an enterprise scale require expert-level developers and product support challenges.

Market Implications:

Open-source projects, such as Kubeflow, Kubernetes and TensorFlow, have volunteers (with deep professional product management experience) bringing a structured approach to feature priority, release process, community contribution standards, DevOps processes, product documentation, training and product evangelism, along with a well-defined roadmap. The professional leadership and collaboration is resulting in AI platform components, which are more aligned with enterprise requirements and quality. This collaboration between industry practitioners, product developers and community contributors is creating the roadmap to develop reliable, stable and easy-to-use components, with a well-defined support model.

Open-source AI tools tend to enable breakthrough capabilities and, in some cases, offer features that are not yet available through proprietary technologies. However, documentation in open-source AI projects is often lacking resulting in implementation challenges. Vendors curating AI orchestration platforms must actively integrate open-source technologies as part of the overall mix.

Recommendations:

IT leaders devising initiatives to operationalize AI should:

- Integrate the open-source and commercial AI components with the goal of a flexible, reliable, maintainable, scalable and interoperable AI orchestration platform implementation.
- Determine the maturity of the open-source AI components, related support models, skills, services, licensing costs and corresponding commercial product offerings to determine the appropriate technical stack.
- Require the future roadmap of product enhancements of open-source and commercial AI products. A well-defined roadmap points toward a clear and stable product vision.

Related Research:

- 2021 Planning Guide for Data Analytics and Artificial Intelligence
- A Guidance Framework for Operationalizing Machine Learning
- Solution Path for Building an Effective Technical AI Strategy

A Look Back

In response to your requests, we are taking a look back at some key predictions from previous years. We have intentionally selected predictions from opposite ends of the scale — one where we were wholly or largely on target, as well as one we missed.

This topic area is too new to have on-target or missed predictions.

Evidence

Gartner's 2019 AI in organizations survey was conducted to uncover the keys to successful AI implementations and the main barriers to the operationalization of AI.

The research was conducted online during November and December 2019 among 607 respondents from organizations in the U.S., Germany and the U.K. Quotas were established for company size and for industries, to ensure the sample as a good representation across industries and company sizes. Organizations were required to have developed AI or intend to deploy AI during the next three years.

Respondents were screened to be part of the organization's corporate leadership or report into corporate leadership roles, have a high level of involvement with at least one AI initiative and have one of the following roles when related to AI in their organizations:

- Determining AI business objectives
- Measuring the value derived from AI initiatives
- Managing AI initiatives development and implementation

Results of this study do not represent global findings or the market as a whole, but reflect the sentiments of the respondents and the companies surveyed.

Recommended by the Authors

Some documents may not be available as part of your current Gartner subscription.

2021 Planning Guide for Data Analytics and Artificial Intelligence

A Guidance Framework for Operationalizing Machine Learning

Solution Path for Building an Effective Technical AI Strategy

Emerging Technologies: Critical Insights on AI Semiconductors for Endpoint and Edge Computing

Product Managers Developing Edge Systems Will Require Integration of AI Accelerator Chips

Hype Cycle for Artificial Intelligence, 2020

Use Startup Mindset to Accelerate Innovation Using AI