Cool Vendors in Enterprise AI Operationalization and Engineering

Published 11 October 2021 - ID G00734193 - 12 min read

By Analyst(s): Chirag Dekate, Farhan Choudhary, Svetlana Sicular, Sumit Agarwal

Initiatives: Artificial Intelligence

Rapid maturation of enterprise AI initiatives is driving urgency to maximize value capture through productization of AI. Data and analytics leaders must evaluate emerging vendors to build enterprise-grade AI orchestration and automation platforms or solutions to scale enterprise AI initiatives.

Additional Perspectives

- Summary Translation: Cool Vendors in Enterprise AI Operationalization and Engineering
  (26 October 2021)

Overview

Key Findings

- Building an observability framework to enable monitoring, automated root cause analysis and remediation of core functions (from data to deployment) in an artificial intelligence (AI) orchestration platform are essential to scaling enterprise AI initiatives.
- Platforms designed to enable multiple users, teams and AI applications to share hybrid multicloud infrastructures for machine learning (ML) and deep learning are important enablers of training and deployment of models.
- Experiment tracking, dataset versioning and machine learning model management are crucial to improving data scientist productivity and scaling enterprise AI initiatives.
- Automated tools that maximize deep learning and machine learning model performance and enable seamless packaging to rapidly deploy models across on-premises, cloud and edge environments are essential to accelerating production AI.

Recommendations

Data and analytics (D&A) leaders exploring promising emerging techniques and approaches in the market should:

- Evaluate the Arize platform for its ability to deliver observability platforms that aid in monitoring, root cause analysis, performance improvement and problem resolution in AI pipelines.
- Assess OctoML for its ability to automatically improve machine learning model performance across architectures, seamless packaging and deployment of models across hybrid multicloud environments.
- Analyze Spell's platform capabilities around enabling AI teams to orchestrate deep learning training and deployment across GPU environments in a hybrid multicloud context.
- Appraise Weights & Biases (W&B) for experiment tracking, workflow management, and engineering deep learning workloads across diverse frameworks with a collaboration and governance feature set.

Strategic Planning Assumptions

By 2025, 50% of large enterprises will have deployed artificial intelligence 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

This research does not constitute an exhaustive list of vendors in any given technology area, but rather is designed to highlight interesting, new and innovative vendors, products and services. Gartner disclaims all warranties, express or implied, with respect to this research, including any warranties of merchantability or fitness for a particular purpose.

What You Need to Know

Gartner's Cool Vendors in Enterprise AI Operationalization and Engineering are focused on the urgency and the need for enterprises to deploy platforms that accelerate scaling enterprise AI into production.

Scaling enterprise AI initiatives into production and curating orchestration platforms that enable deployment and management of thousands of models in production are some of the critical trends observed in the Hype Cycle for Artificial Intelligence, 2021.

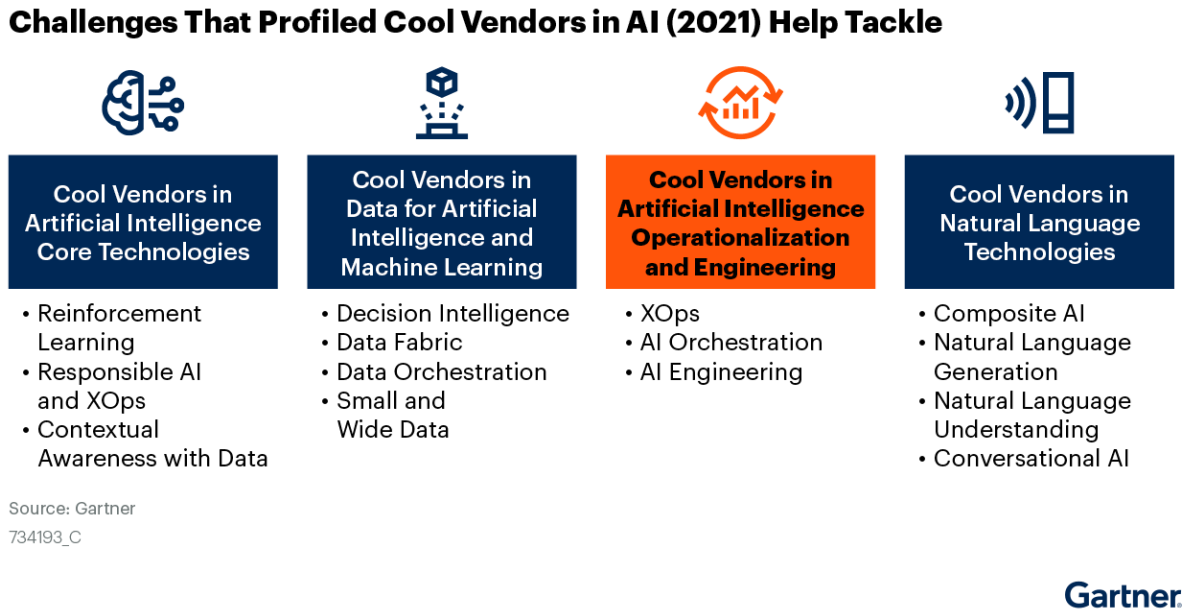
The Cool Vendors in this research seek to address these trends by enabling:

- **AI orchestration,** specifically, performance optimization, packaging and seamless deployment across on-premises, cloud and edge environments.
- **Management and governance** platforms so teams across enterprises can develop, reuse and rapidly innovate on a shared set of machine learning model assets.
- **Observability,** involving monitoring, root cause analysis and effective problem resolution across data and machine learning pipelines.

To get a more holistic view of the Cool Vendors in Enterprise AI Operationalization and Engineering, D&A leaders should also consider reviewing the following (see Figure 1):

- Cool Vendors in Data for Artificial Intelligence and Machine Learning
- Cool Vendors in Conversational and Natural Language Technologies
- Cool Vendors in AI Core Technologies

Figure 1: Cool Vendors in Artificial Intelligence Operationalization and Engineering



Arize

Berkeley, California (https://arize.com)

Analysis by Chirag Dekate and Farhan Choudhary

Why Cool: Arize is cool because its ML observability platform addresses three key challenges inhibiting AI operationalization. It:

- Automatically detects problems such as data quality or drift issues
- Enables faster root cause analysis and problem resolution of ML model
- Continuously improves model performance, interpretability and readiness

With little or no customization, D&A leaders can utilize Arize's SDKs and Python tools to monitor complete AI pipelines including data and model deployment environments. Arize's platform seeks to automatically detect data quality, model drift, model performance, prelaunch model validation, model interpretability, fairness and bias, and data consistency issues — in essence, a unified means of tracking operations of AI pipelines.

Arize's platform also seeks to resolve detected problems through the use of log training, drift cohort analysis, feature-based performance analysis, cross-training and validation comparisons, model input data analysis, explainability, and other means.

Arize's observability platform also seeks to improve models through advanced capabilities including in-the-box cross-pipeline data-feature-model analysis and exploration, concept drift resolution, efficient management, and version control of models and features. It also allows users to "bring their own explainability" for clients looking for more flexibility while being able to perform not just global-level feature importance but also cohort- and individual-level feature importance.

Core to Arize's coolness is its platform-centrism, where it integrates with broader data and MLOps tooling and enables a unified enterprisewide AI pipeline observability experience. Its effective partnership strategy enables Arize to adapt to diverse enterprise implementation contexts and accelerate operationalization of AI pipelines.

Challenges:

- Arize focuses on a core set of observability capabilities that is crucial to operationalization of AI. However, Arize is not designed to provide a comprehensive one-stop enterprise AI orchestration and automation platform. Data and analytics leaders seeking to build an AI orchestration and automation platform will need to complement Arize with a broader set of DataOps and MLOps toolsets.
- Arize's focus on observability also extends into its pricing model. Unlike the rest of the DataOps and MLOps toolsets, the pricing model that Arize uses is based on the number of predictions. Data and analytics leaders will need to partner with Arize to understand this model and reconcile with a broader ecosystem that is often priced differently.

Who Should Care:

- Enterprises seeking to accelerate operationalization of AI pipelines should evaluate Arize for its observability capabilities.
- Enterprises should evaluate Arize for its ability to address production challenges around model and data drift, performance degradation, model interpretability, data quality issues, model readiness, and fairness and bias issues.
- Enterprises seeking to maximize ROI and have visibility into how the model impacts their business's bottom line by increasing organizational focus on model building, improving model productionalization velocity and improving model outcomes should consider Arize.

OctoML

Seattle, Washington (https://octoml.ai)

Analysis by Sumit Agarwal

Why Cool: OctoML is cool because it provides the niche capability of a platform for model optimization, benchmarking and deployment extending the open-source Apache TVM project. The platform is offered as a hosted SaaS solution. The Apache TVM project improves performance of ML models by creating compiled objects optimized for the specific hardware including ARM, Apple M1, AMD, Intel, NVIDIA and Qualcomm. OctoML extends this functionality by providing several enhancements that make the product more usable and powerful. The OctoML platform provides an API and a web UI to improve accessibility. The platform was designed to help OctoML customers get started quickly if they found Apache TVM setup hard and time-consuming.

In addition to TVM optimization methods, OctoML uses historical optimization data to benchmark and improve the performance of new models. The use of historical data enables improvements in optimization time and product improvements to include new operators. OctoML currently uses TVM and ONNX runtime engines for optimization and provides the model with the best performance. The platform is scalable, and OctoML is planning to add additional engines for model optimization in the future. OctoML customers have reported performance improvements of more than 100% on models that were previously optimized in their development framework, as well as optimization time reduction from days to minutes.

Challenges:

- OctoML is based on an open-source project. This provides an opportunity for other organizations to develop competing offerings based on the same open-source project. Amazon SageMaker Neo provides a similar solution.
- Many OctoML customers are leveraging the open-source component instead of the commercially available platform because the open-source code provides functionality that is sufficient to meet their requirements.
- OctoML is supporting such customers by providing paid feature enhancements. However, the risk of open-source components cannibalizing some of the potential demand for the OctoML product remains.

Who Should Care:

- Enterprises seeking to deploy large models (such as computer vision, speech or natural language processing models) that require high execution performance deployed on a variety of devices and computing processors should evaluate OctoML.
- Enterprises looking to use the open-source Apache TVM may consider partnership with OctoML for feature enhancements or product support.
- Enterprises working to establish automated MLOps pipelines should evaluate integration of OctoML within the pipeline.

Spell

New York, New York (https://spell.ml)

Analysis by Chirag Dekate and Farhan Choudhary

Why Cool: Spell is cool because it simplifies deep learning infrastructure and operationalization complexities using a comprehensive platform approach. Spell seeks to transform the entire end-to-end infrastructure-agnostic MLOps platform for experiment orchestration, comparing multiple experiments, hyperparameter optimization, model catalog, deployment and governance of models.

Spell's approach is infrastructure-agnostic, enabling data and analytics leaders to unlock innovation across hybrid multicloud contexts. Spell is currently differentiated from its peers in part because it can operate across any on-premises environment or across key cloud service providers including Amazon Web Services, Google Cloud Platform and Microsoft Azure Cloud.

Spell enables end-to-end AI governance by giving visibility into the full AI pipeline, including data collection, experiment orchestration, model development, model registration, model deployment and monitoring. It provides a complete and auditable lineage for model governance teams.

Spell's platform comprises powerful command-line tools that automate packaging data, and model and orchestrate execution across a hybrid multicloud context. Further, Spell enables data scientists to leverage existing Python notebooks and exposes a notebook-friendly environment that is easy to utilize. Spell offers feature-rich dashboards that enable data and analytics leaders to manage deep learning models across a diversified hybrid multicloud estate for orchestration and productionalization of AI pipelines.

Challenges:

- Data and analytics leaders seeking a complete AI orchestration platform will need to complement Spell's deep learning MLOps platform with an equivalent MLOps platform for machine learning pipelines. Spell's deep learning focus could limit its applicability in situations where data and analytics leaders have a diversified portfolio of machine and deep learning models.
- Spell's approach is currently optimized to maximize utilization of GPUs and expose GPUs as a platform across a diversified IT estate. Spell currently does not support non-GPU environments, such as Amazon Trainium, AWS Inferentia, Google Cloud TPU and the upcoming Intel GPU chips. Spell can address this challenge relatively easily.
- Ensuring a consistent set of features/capabilities in a multicloud context can be challenging. Some of the features available in one cloud provider (for example, automated disk space tracking, currently available only for AWS) might not be available across all providers.

Who Should Care:

- Enterprises seeking to accelerate operationalization of AI pipelines comprising machine learning and deep learning models should evaluate Spell for its capabilities to train and deploy models.
- Enterprises should evaluate Spell for its end-to-end comprehensive set of capabilities to accelerate experimentation to productionalization of deep learning models across a diverse hybrid multicloud IT estate.
- Data and analytics leaders who are looking to build cross-enterprise platforms that accelerate enterprisewide collaborative design and accelerated productionalization without getting bogged down by infrastructure challenges should consider Spell. D&A leaders should actively evaluate Spell's orchestration, automation and powerful management capabilities.

- D&A teams in highly regulated setups for whom the utmost priority is transparency and lineage can also look into Spell's capabilities.

Weights & Biases

San Francisco, California (https://wandb.ai)

Analysis by Svetlana Sicular

Why Cool: W&B is cool because it provides developer tools for machine learning that are easy to use, quick to implement and composable. Customers can mix and match the W&B tools with their own DSML platforms and ML frameworks. This allows them to fill in the gaps in their ML development process, such as ModelOps, MLOps and continuous integration (CI)/continuous deployment (CD). W&B's composable tools provide experiment tracking, dataset versioning, visualization, ML project collaboration and reproducibility. For example, customers can track, compare and visualize ML experiments with five lines of code versus alternatives that require significant refactoring of codebases.

The vendor also made object stores a first-class citizen. Considering that systems are updated on different cadences and training data comes downstream from other models, W&B provides easy-to-understand, informative and intuitive data visualizations. Customers can see data lineage, walk back and, if needed, get directly to a dataset. W&B can equally well integrate with many development frameworks. For ML developer tools, the vendor also has the advantage of composability, so customers use only what they need and complement their current DSML platforms to fill their gaps.

Onboarding is also unique. Customers typically "onboard in a tweet." They start with lightweight onboarding that takes just several lines of code that can fit in a single tweet. Most of them organically expand W&B inside the organization, and some of them run W&B across petabytes of data. Customers like W&B reports that allow them to track various systems' metrics. For example, metrics for GPUs and TPUs enable customers to utilize compute to a complete capacity. W&B also has a substantial GitHub presence and hosts a popular ML podcast called "Gradient Dissent."

Challenges:

- W&B doesn't provide its own ML orchestration capabilities. However, it tightly integrates with GradientCI that specializes in machine learning workflow.

- W&B finds itself in a very competitive space with open-source products such as TensorBoard and some DSML platforms.

Who Should Care:

- Enterprises that have ML models in production, train models frequently, and have four or more practitioners can take advantage of W&B's scale, collaboration and engineering. Such organizations could be in various industries, such as high tech, life sciences, healthcare, gaming, agriculture and insurance.
- Enterprises that need to quickly introduce ModelOps or MLOps capabilities for audit, compliance and transparency of model-building from experimentation to production should evaluate W&B. Capabilities to consider include data transparency, centrally managed access controls and artifact audit logs, and a complete model history that enables traceable model results.
- Enterprises that are looking for best practices for ML teams can benefit from joining the W&B community, which issues technology reports, has vast educational content and examples, conducts weekly seminars, and shares knowledge in GitHub and in the Slack forum.

Recommended by the Authors

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

Applying AI — A Framework for the Enterprise

Applying AI — Techniques and Infrastructure

Quick Answer: How Should CXOs Structure Operating Models to Capitalize on AI Opportunities?

Predicts 2021: Operational AI Infrastructure and Enabling AI Orchestration Platforms

5 Steps to Practically Implement AI Techniques

Demystifying XOps: DataOps, MLOps, ModelOps, AIOps and Platform Ops for AI

A Guidance Framework for Operationalizing Machine Learning

Machine Learning Playbook for Data and Analytics Professionals

Understanding MLOps to Operationalize Machine Learning Projects