Cool Vendors in Enterprise AI Operationalization and Engineering

Published 27 April 2021 - ID G00746832 - 15 min read

By Analysts Chirag Dekate, Farhan Choudhary, Soyeb Barot, Erick Brethenoux, Arun Chandrasekaran, Robert Thanaraj, Georgia O'Callaghan

Initiatives: Artificial Intelligence

More than half of successful artificial intelligence pilots never make it to the deployment stage; thus, data and analytics leaders struggle to capture value from AI investments. Data and analytics leaders must evaluate emerging vendors to build enterprise-grade AI platforms or solutions.

Overview

Key Findings

- Building a data foundation that enables data and analytics leaders to unify existing data foundations and integrate new ones is at the core of any successful artificial intelligence (AI) strategy.
- AI model-monitoring environments that track metrics, data tiers, drifts, outliers and distribution are critical foundations for advanced capabilities including diagnostics, alerting and automated root cause analysis.
- Platforms architected to enable multiple users, teams and AI applications to share infrastructure (on-premises or cloud) are important enablers of training and deployment of models.
- Data stores, feature stores and model stores are core technologies to engineer AI platforms that promote reusability, reproducibility, reliability, retraining and rollback.

Recommendations

Data and analytics leaders exploring promising techniques and methods that are emerging in the market should:

- Evaluate the Ascend.io platform for its ability to integrate with existing data infrastructure including databases, Hadoop stacks and more, and automate curation of DataOps pipelines.
- Appraise Mona for its differentiated contextual monitoring of models in real time that enables diagnostics, alerting and automation of root cause analysis.

- Analyze Run:AI platform capabilities around unifying compute intensive GPU resources, integrating them with machine learning operationalization (MLOps) and data science platforms, and provisioning them across teams and AI workloads.
- Assess Tecton feature store capabilities around self-service and the ability to automate feature discovery and reuse across their enterprise context.

Strategic Planning Assumptions

By 2025, 50% of enterprises will have devised 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.

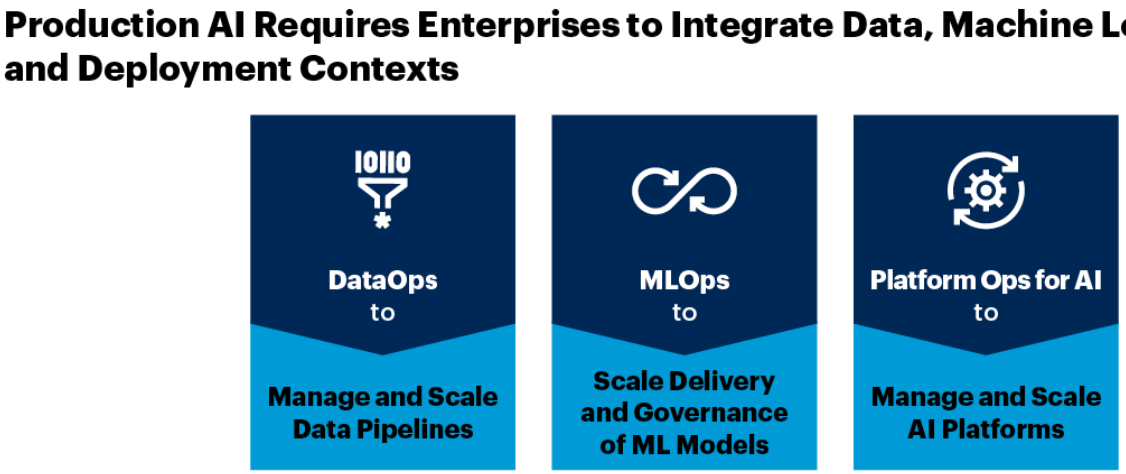
Analysis

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

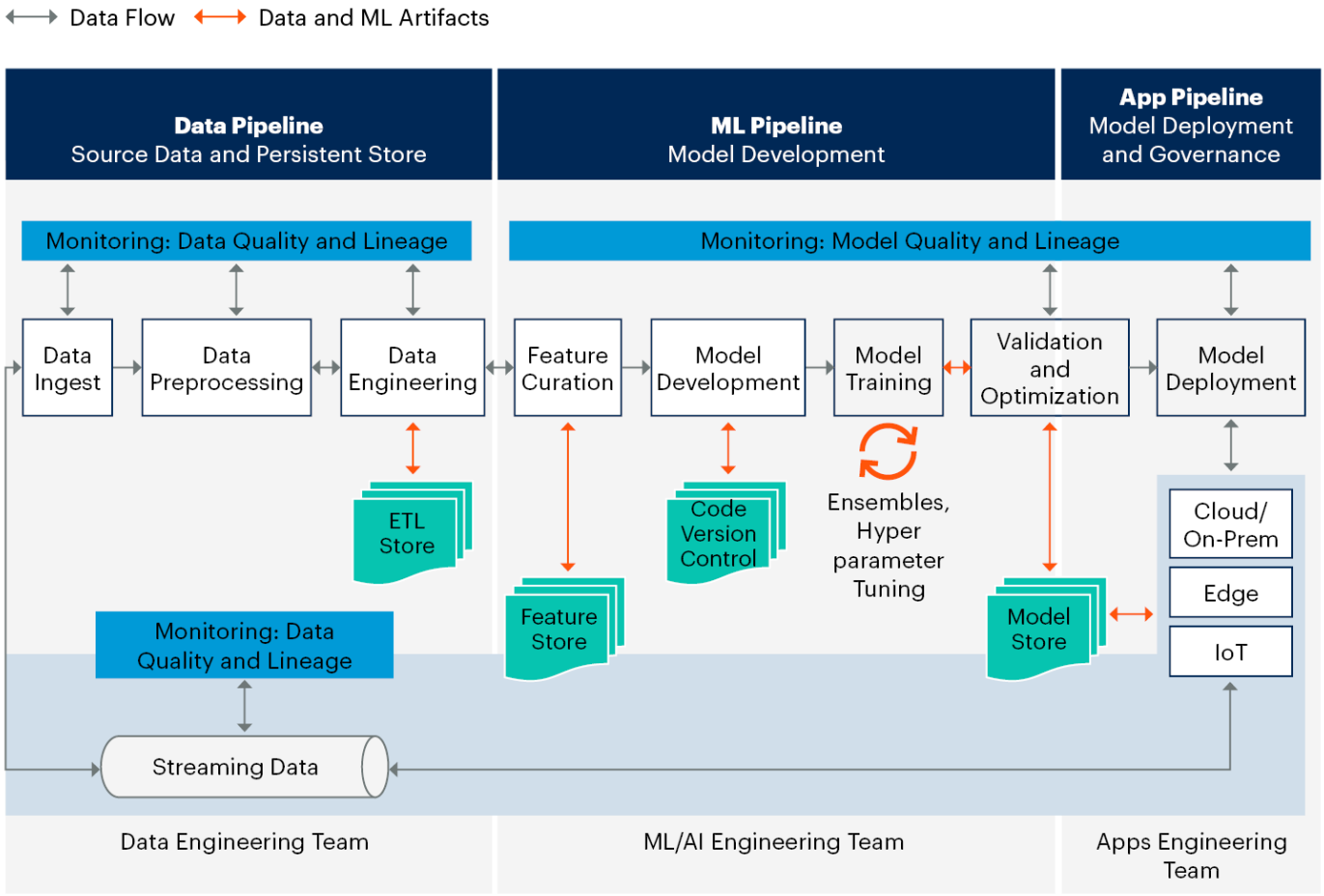
Gartner's Cool Vendors in Enterprise AI Operationalization and Engineering are focused on the urgency and the need for enterprises to develop platforms that accelerate productizing AI (see Figure 1).

Figure 1: Production AI Requires Enterprises to Integrate Data, Machine Learning and Deployment Contexts



Gartner has selected four vendors that specialize in various parts of the data-to-deploy AI life cycle (see Figure 2). These vendors offer differentiated capabilities that are core for enterprises seeking to devise enterprise AI platforms.

Figure 2: Enterprise AI Platform to Accelerate Productizing AI



Source: Gartner
737833_C

To accelerate production AI, IT leaders will need to devise enterprise AI platforms to provide standardization, manageability and governance of data, machine learning (ML) and deployment pipelines.

Curating enterprise AI platforms involves nurturing multiple best practices across data, ML, AI and application development to create an efficient delivery model for AI-based systems. Enterprise AI platforms enable IT leaders to:

- Leverage existing data foundations and adapt agile practices to building and operating data pipelines that enable fusion of newer data sources to expose a unified data interface to data scientists.
- Monitor environments that enable trackability of key metrics across the AI life cycles and, most importantly, also provide advanced automated contextual capabilities.

- Enable shareable platforms for high-value resources including GPU accelerated compute environments.
- Provide autonomy to business units by enabling discoverable and reusable AI artifacts, including data transformation patterns, features and machine learning models across the enterprise.
- Scale value capture from AI investments by delivering environments that enable monitoring, governance and auditing of any AI model in production.

Ascend.io

Palo Alto, California, U.S. (www.ascend.io)

Analysis by Robert Thanaraj, Chirag Dekate and Soyeb Barot

Why Cool:

Ascend.io is cool because it revolutionizes the speed at which data engineers can build and operationalize data pipelines across a distributed data and analytics ecosystem. Its Unified Data Engineering Platform allows data teams to launch Spark-based pipelines in hours, allowing them sufficient time to analyze and improve business logic instead of spending a huge amount of time managing the platform infrastructure.

It uses a declarative programming style to deliver its promise of speed, where the logic of computation is expressed (the what) without having to describe the underlying control flow (the how). In addition to faster development cycles, its customers also benefit from a platform that offers low-code and low-maintenance of data pipelines, while also offering higher-code options for advanced use cases.

At the core of its platform, Ascend.io runs a patented component called “DataAware” that tracks every action performed against data within the platform. It ensures data pipelines run at optimal efficiency and adhere to governance requirements.

The Ascend platform is available as a managed service or a self-managed service on the public cloud or virtual private cloud (VPC).

Challenges:

Ascend claims to self-optimize the data pipeline alongside the underlying compute infrastructure. Its customers are able to realize these operational improvements. However, with any usage-based cloud pricing model, throwing more computational resources (i.e., risk of increased operational cost) is not always ideal over building optimized data pipelines. Price/performance should be a transparent metric to avoid spiking operational costs in the long run.

Organizations starting with a new “greenfield” data engineering platform can benefit from a partnership with Ascend. However, those organizations that have an existing platform will need to factor in the challenges (both from technical and costs perspectives) when migrating their existing codebase into Ascend.

Who Should Care:

Data and analytics leaders seeking to improve data engineering productivity for their data teams should consider Ascend. The Ascend platform can provide an end-to-end visibility of data pipelines in an enterprise, and it can empower data analysts and scientists to quickly build data pipelines feeding their analytical models.

Mona

Atlanta, Georgia, U.S. (www.monalabs.io/)

Analysis by Soyeb Barot and Farhan Choudhary

Why Cool:

What makes Mona cool is its focus on the contextual monitoring and observability of AI models in production. The platform goes beyond just the model metrics and takes into account the data used for training and inference data captured within the retraining pipeline within the context of business key performance indicators (KPIs). One of the biggest challenges that businesses and data scientists face is bridging the gap between model metrics such as gain and lift charts, under the curve (AUC), receiver operating characteristic (ROC), and F-measure with business KPIs like percent of customer churn and business value captured. In addition, enterprises want to gain transparency into the behavior and performance of these models over time. The concerns are around data integrity, concept drift and other performance blind spots. This is where Mona supports exporting of inference data directly from within the model code at the time of inference and assists with the end-to-end monitoring of the model performance.

Data scientists and ML engineers can leverage Mona’s extraction, transformation and loading (ETL) capabilities to create context classes (data table description), schemas, fields and tables to track new metrics derived from the raw data. Mona uses arrays of metric building functions, such as mathematical and logical operators, to leverage the data used within transformations without losing their context within source systems. The model outputs, such as classification results, sentiment scores, and confidence intervals, can be tracked as leading indicators of model performance, and not just precision and recall, thereby proactively detecting anomalous behavior before business KPIs are negatively impacted.

A unique feature with the platform is its anomaly detection engine, called Insights Generator, using the verse object that contains instructions to detect a single anomaly type within the monitoring context.

These instructions specify data segmentation configurations, the metrics to measure and compare, and other specifications regarding assessing whether the behavior is anomalous.

In summary, Mona's value-add lies in its ability to:

- Track and aggregate metrics at the model and data tiers, bringing together business indicators and technical features.
- Detect drifts, outliers and distribution tracking for behavior modeling while the models are in production.
- Diagnose and alert while performing automated root cause analysis via "Insights Generator" using the algorithms and analytics database.

Challenges:

- Mona does not provide a model development environment, and comes out as a niche model monitoring and observability solution for AI models in production.
- The configuration of the data points requires technical expertise, with business domain experts having limited ability to tweak/change parameters to effect changes of the model behavior.
- Mona will need to dedicate substantial effort and resources to educate and build on maintaining the niche of monitoring and observability, since other DSML and MLOps solutions focus and build these capabilities within their individual platforms.

Who Should Care:

- Organizations within specific verticals that require contextual monitoring of models in real time to maintain their competitive edge in the marketplace can leverage Mona's platform to get insights into model behavior.
- The Insights Generator capability allows both technical and business users to drill down to the data element and feature variable levels to identify anomalies and make course corrections.
- Enterprises closely tracking global and region-specific regulatory requirements usually require a robust post-production monitoring system to meet compliance requirements and reduce risks with AI implementations.

Run:AI

Tel Aviv, Israel (www.run.ai)

Analysis by Chirag Dekate

Why Cool:

Run:AI is cool because of its differentiated AI orchestration runtime platform. Current approaches require enterprises to use GPU infrastructures for deep learning in an ad hoc manner often organized by project silos. Further, current solutions either do not enable pooling of GPUs across cloud or on-premises environments or they do not enable fractional access to GPUs. This results in overprovisioning of expensive infrastructures and underutilization of the core resources. Run:AI addresses all these issues and enables enterprise IT leaders to develop a platform strategy for orchestrating their AI workloads across their compute-intensive assets, whether on-premises or in the cloud.

The Run:AI platform should not be confused with an MLOps platform. The Run:AI platform sits between the MLOps middleware (for example Kubeflow, MLflow or other data science tools) and the underlying GPU-accelerated compute infrastructure stacks.

Run:AI enables enterprises to decouple their data science models from the underlying hardware. Rather than manually managing AI workloads on compute resources including GPU-accelerated clusters, the Run:AI platform enables enterprises to share compute resources across AI workloads and teams. Run:AI pools together all the GPUs in an organization and exposes them to multiple users, teams and applications as a unified computing environment. Further, the Run:AI platform also integrates with common MLOps environments including Kubeflow, MLflow and data science tools.

With Run:AI, multiple users, teams and workloads can share a unified compute platform to elastically allocate GPUs according to a customizable policy engine designed to maximize utilization of GPUs across the full environment and alignment to business priorities.

What differentiates the Run:AI platform is:

- Its ability to expose fractional access to GPUs (where users can select just a fraction of a GPU or multiple GPUs) to maximize utilization of compute-intensive environments. A key innovation here is how Run:AI enables multiple containers to run on a single GPU or scale to as many as needed.
- A Kubernetes core makes it easy for the Run:AI platform to be deployed on-premises or via any of the cloud environments.
- Integrated hyperparameter optimization enables customers to run a large array of experiments in parallel while leveraging a shared environment.

Challenges:

- Run:AI focuses on addressing the orchestration problem. Enterprises seeking a fully featured end-to-end stack from data ingestion, development, orchestration and deployment might need to complement Run:AI with relevant platform tooling.

- Prospective clients need to have a relatively high degree of maturity and workloads that require GPUs for training and/or inference.

Who Should Care:

- IT leaders seeking to devise a platform strategy that enables multiple users, teams and workloads to share a common platform for AI orchestration and deployment should consider Run:AI.
- IT leaders looking to maximize utilization and value for AI training and inference from their GPU investments across on-premises or any major cloud provider may benefit from Run:AI.
- IT leaders and data scientists struggling with manual scheduling and management of resources for AI workloads across on-premises and any major cloud provider should actively evaluate Run:IO's automation features and capabilities.

Tecton

San Francisco, California, U.S. (www.tecton.ai)

Analysis by Arun Chandrasekaran, Georgia O'Callaghan and Farhan Choudhary

Why Cool:

Tecton provides an enterprise data feature store that enables data scientists to build, reuse and monitor features for batch, streaming and real-time analytics through a centralized repository to accelerate machine learning model development and deliver reliable data to models in production. Rather than just offering a "store," Tecton provides end-to-end management and orchestration of the feature life cycle, from feature creation and evaluation to deployment, sharing, security, governance and monitoring.

If the data is not agile enough and can't be guaranteed with high fidelity and quick availability, its value significantly diminishes for AI pipelines. A feature store enables several critical capabilities to ease these challenges:

- It allows clients to execute data pipelines that transform raw data into feature values.
- It stores and enables reuse of the features.
- It enables tracking feature versions, data lineage and securing that data easier.
- It serves the feature data consistently and with low latency for training and inference purposes.

The Tecton product enables clients to:

- Connect the feature store to a variety of on-premises and cloud data sources such as S3, Redshift, Snowflake and Kafka for batch and real-time analytics.
- Define, create and register new features through policy as code in an automated fashion.
- Generate feature value and serve those features to AI models in a consistent manner during both the training and inferencing phases.
- Architect feature deployments through self-managed stores or fully managed feature stores (currently in AWS with upcoming support for Microsoft Azure and Google Cloud).
- Leverage preintegration of Tecton with MLOps products such as Amazon SageMaker, Databricks and Kubeflow to accelerate the machine learning pipeline.

Challenges:

- While Tecton is one of the early startups in this space, several data science platform vendors such as AWS, Databricks and IBM have built feature stores today as part of a platform play.
- There are open-source alternatives such as Feast that may satisfy the needs of do-it-yourself (DIY) customers (although Tecton is a core contributor to Feast as well), obviating the need for Tecton.
- Prospective clients need to have a high degree of AI maturity to consider the usage of feature stores, which can raise the barrier to entry for a small startup such as Tecton.

Who Should Care:

- Data scientists looking to accelerate the MLOps pipeline through feature discovery and reuse should consider the Tecton feature store.
- Data and analytics leaders who want to enable self-service, organizationwide sharing and reuse of feature data should shortlist Tecton as a potential vendor in their feature store evaluation.

Where Are They Now?

ModelOp (Formerly, One Data Group)

Chicago, Illinois, U.S. (www.modelop.com)

Analysis by Farhan Choudhary, Soyeb Barot and Erick Brethenoux

Profiled in Cool Vendors in Data Science, 2014

Why Cool Then: Open Data Group's forte was dealing with very large datasets and providing outsourced analytics services, analytic resources and management consulting. It enabled organizations to analyze

data and build predictive analytics models (such as risk and response modeling) across a variety of platforms working with customer data, supplier data and third-party data, as well as data from internal business processes. As its name suggests, Open Data Group has deep expertise in open-source analytics among other analytics infrastructure and strategy services. Open Data Group was also a key contributor to the Predictive Model Markup Language (PMML) standard and released an open-source predictive analytics solution called “Augustus” for data mining and statistical modeling.

Where They Are Now: Open Data Group officially became ModelOp in October 2019. ModelOp is cool because it addresses one of the major roadblocks that clients face when it comes to value realization of AI projects — putting AI models into production. The ModelOp Center by ModelOp helps to monitor, govern and audit any type of AI model in production (e.g., analytical, machine learning, linguistic, heuristic) while being cloud, runtime and dev-environment agnostic. It enables ModelOps at scale while leveraging existing enterprise IT investments and reducing technical debt and complexity of AI solution integration, keeping the organizations at bay to truly realize the value of their AI projects.

ModelOp helps clients build resilient production-ready systems by offering thorough model governance with standard and custom compliance rule enforcement, providing a central model life cycle operations control panel, audit capabilities and model risk management integrations. The ModelOp Center sits between the Model Factories and the Enterprise IT Stack. This allows AI architects and other technical audiences to have a holistic view of the production pipelines. Further, the ModelOps Center enables users to get notified on the basis of concept, data or mathematical drifts, noncompliance, and nonadherence to business rules by integrations with ticketing platforms such as ServiceNow and Jira. This gives AI teams the ability to be agile and respond to changes as they happen.

Apart from newer capabilities that help manage the risks of AI models in production, ModelOp also assists in ModelOps, model-agnostic deployment, execution, monitoring and governance, which includes model packaging, model execution, model monitoring, model retraining and champion/challenger automation. The ModelOp Center also ensures that the business KPIs are duly mapped to AI projects, which empowers business users to make strategic decisions on the basis of outcomes from the models.

Who Should Care:

- Organizations struggling with putting their AI models into production while addressing concerns around governance and compliance should consider ModelOp.
- Organizations that want to have a robust AI model management, monitoring and governance capability but don’t know where to start or are overwhelmed by the complexity of the subject may benefit from ModelOp.
- ModelOp may be a good fit for organizations that want to structure their AI initiatives better and want to have a more holistic view of models in production while adding least technical debt and maximizing ROI.

Recommended by the Authors

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

Innovation Insight for ModelOps

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

Operational AI Requires Data Engineering, DataOps and Data-AI Role Alignment

Predicts 2021: Operational AI Infrastructure and Enabling AI Orchestration Platforms