ASTRONOMER EBOOK

How to Accelerate AI with Apache Airflow

Introduction

In the age of digital transformation, the majority of companies acknowledge that data is their most important asset in driving business and delivering value to customers and stakeholders. They also agree that the modern enterprise has become, at its core, a data machine—and that modern data orchestration should form the central nervous system of that machine.

Most enterprises, however, still struggle to properly harness the vast amount of data being collected. According to one **MIT Technology Review Insights report**, even as 96% of employers believe that generative artificial intelligence (AI) will impact their business, only 9% of them have actually fully deployed an actual AI use case. With the worldwide market size for artificial intelligence expected to grow by **619%** through 2030, the race is on for businesses to enhance their offerings with AI-centric innovations. This growth will demand that modern organizations take certain best practices into consideration to be able to fully operationalize and scale AI and machine learning (ML) initiatives.

Fortunately, opportunities exist for businesses to more effectively orchestrate data, facilitate best practices, and position data in their organizations to drive AI and ML innovation. This presents itself in the form of modern tools like the open-source workflow management platform Apache Airflow. Read on to learn how modern enterprises are successfully orchestrating their data and accelerating their AI innovation with Apache Airflow.

Apache Airflow Overview

Since its inception in 2014, Airflow has risen to become the industry's leading workflow management platform for data pipelines. This is seen especially in recent years with the platform's meteoric rise in adoption, with nearly 166 million downloads in 2023, a 67% year-over-year increase. And of all the organizations running Airflow today, nearly 30% are using it to support AI initiatives.

An ever-growing list of data integrations with the most prominent applications, databases, tools, and cloud services has led to Airflow becoming the de facto standard for data workflows, and the glue that holds together the modern data stack for countless businesses. Additionally, Airflow's sizable and active open community is 31,000+ members strong, ensuring the platform stays up to date with new and existing data sources and providers.

More recently, Airflow has naturally expanded beyond the data engineering team to become the MLOps platform of choice for AI and ML teams. Data scientists and machine learning engineers find it ideally suited to kickstart and scale AI initiatives and standardize best practices.

67% YEAR-OVER-YEAR INCREASE IN DOWNLOADS OF APACHE AIRFLOW

The Advancements and Challenges of ML

The acceleration of AI requires all of the components of traditional data pipelines and more—extract, transform, and load (ETL) processes, data cleansing and feature generation, model training and monitoring, not to mention invocations or fine tuning of large language models (LLM)—and all of these components run via data pipelines. These pipelines are essential to delivering the advancements we've come to associate with AI, like →

- **Personalized recommendations**
- **Content generation and text summarization**
- **Predictive maintenance**
- **Sentiment analysis**
- **Chatbots for customer support**
- **Fraud detection and risk assessment in financial services**
- **Marketing optimization and customer segmentation**
- **Supply chain optimization and demand forecasting**

In particular, as part of the AI race, data engineering and ML teams are being told to build LLM applications as fast as possible, while remaining compliant with ethical, corporate, and legal standards. However, even as these teams realize how critical it is for data and model pipelines to be coordinated and centralized, they often struggle to manage and maintain these pipelines—and effectively implement ML—for the four following reasons →

1. Isolated Development and Deployment

Analytics originated as a research activity rather than an operational discipline, and so, even today, data science and ML tend to be developed in silos. This complicates things when it's time for the notebook of an individual data scientist to make the transition to full-fledged production services. LLMs in particular enjoy an abundance of enthusiasm as prototypes, only to experience a dearth of production-ready practices later. Too often, time and resources are wasted, as far too many AI apps that began life in a hackathon fail to reach production or external use.

2. Lack of Coordination and Standardization

Even when models find their way to production, the teams involved in the implementation of ML come up with their own different processes, approaches, and technologies. Each group has its own set of metrics and features, resulting in inconsistencies and costly duplication of effort. Without a centralized location for these diverse teams to coordinate and collaborate, standardized best practices, testing, and documentation rarely take shape. Instead, the development environment becomes a chaotic free-for-all where outputs are often unreliable, inaccurate, and prone to failure, exposing the company to risk.

At best, the positive impact of AI innovations is lessened by a lack of focus. At worst, low-quality models that reach production can end up resulting in a degraded end user and customer experience or even misleading information.

3. Overwhelming Technology Landscape

Data science and ML teams are inundated with an ever-multiplying swarm of technology options—from experiment tracking to feature storage to model registries, interacting with a thousand ML libraries, running on GPUs and containers, integrated with upstream databases and downstream applications. In part this necessarily reflects the multidisciplinary aspect of machine learning. It's also the result of a still-evolving array of vendors, open-source solutions, and technologies, each of which targets a specific niche.

With the pressure to get the most accurate prediction, data scientists are more likely than software engineers to use whatever tools they can get their hands on — but they are less likely to have the patience or skill to integrate and maintain them. But the lesson of data orchestration is that pipelines must be centralized, standardized, and coordinated.

4. Operational and Compliance Hurdles

Technical challenges aside, data science and ML teams grapple with issues that come with day 2 operations, from provisioning to compute costs to managing failures (i.e., monitoring, alerting, troubleshooting), observability, auditing, access controls, upgrades, and more. As long as teams operate in a siloed, fragmented environment, reining in these challenges becomes virtually impossible.

Equally challenging are the tasks of ensuring data privacy and navigating compliance for predictive analytics. Businesses should always know how each prediction was produced—by which model, on which dataset it was trained, by which transformations it was generated, from which sources it was ingested, and by whom. Unfortunately, getting the answers to these questions grows exponentially more difficult with each new component and technology that teams introduce to the data stack.

Furthermore, they know they need to ensure their platform is compliant with regulations like GDPR and HIPAA. In practice, however, with so many moving parts and diverse technologies involved, data leaders are hard-pressed to actually fulfill these requirements, exposing their business to additional risk.

How to Overcome Barriers to AI Innovation

Fortunately, mature businesses are finding that these four challenges can be overcome—and they can rapidly harness the potential of AI and natural language processing and accelerate innovation—when they commit to three critical strategies →

1. Choose a Centralized Framework for Coordinating all Data Pipelines

The successful delivery of AI innovations depends on the transparency, reliability, and efficiency of well-architected ML Operations—the many steps in the process of data preparation, feature engineering, model training, and monitoring. In order for ML teams to ensure that each step is executed successfully and efficiently, at the right moment in time according to complex dependencies with many points of failure, they need an orchestration framework.

Successful delivery also requires a platform that facilitates hand-in-hand collaboration between data engineering, data science, and ML teams across everything from managing traditional data pipelines to coordinating feature development to building AI applications. Such a platform should also support best practices and promote developer productivity. When businesses successfully invest in and pursue this strategy, AI development is not just accelerated, but also brought to production with few to no errors.

2. Invest in the Right Integrations

The data foundation described above relies heavily on the ability to integrate seamlessly with the various applications, databases, tools, and cloud services that make up the modern data stack. Standardizing on the right integrations and having a platform that makes it easy for developers to utilize in a coordinated way, from development to production, and to cater to evolving business demands is paramount.

An ideal platform should support all the main components (traditional ETL, data quality, feature stores, model training and deployment) and platforms (warehouses, data lakes, vector databases, LLM services) used by data teams. It should always be up to date as those platforms change and new ones arise, and allow for custom integrations.

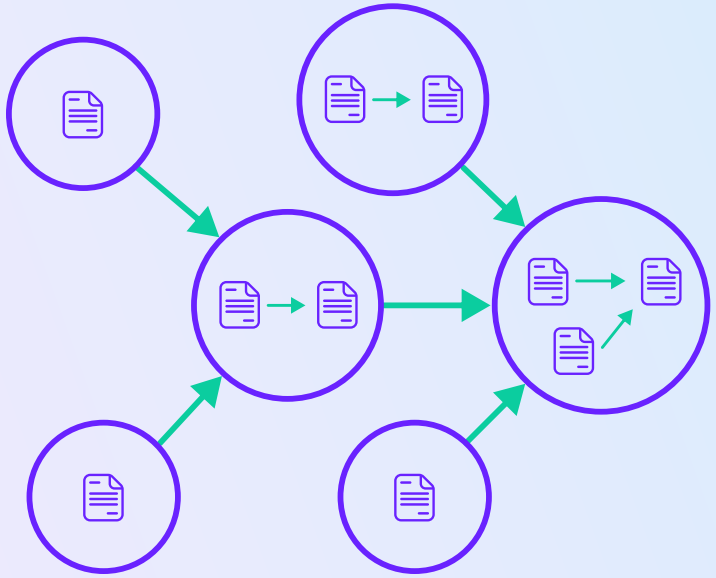
Simultaneously, however, these integrations should balance the orchestration needs of ML engineers, data engineers, and data scientists with the unique requirements and tools each prefers to work in, with seamless interoperability among distributed resources of all kinds, from on-prem resources that exchange data with host-based or client server interfaces to cloud services that use APIs for data exchange.

When businesses invest in robust integrations, they provide their teams with powerful tools to collaborate, orchestrate, and ensure data reliability so they can efficiently and confidently take AI innovations to production again and again.

3. Leverage Data Lineage

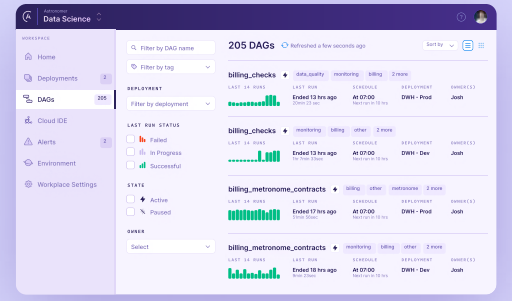
Data reliability in AI applications is only as good as an organization's ability to clearly trace the historical path of the data, starting with data ingestion and the model development process. Data lineage tracks data as it moves from the source system to different forms of persistence and transformations to its consumption by an application or analytics model. Ideally, the business' platform performs this task independently in the background, without relying on developers to do the right thing or add specific code.

With nonexistent or patchy data lineage, troubleshooting, governance, and validation become difficult or impossible. On the other hand, a comprehensive lineage system empowers teams to confidently troubleshoot issues, validate results, maintain governance, and establish reliability and trustworthiness in their AI applications. It's also essential for auditing.



How Astronomer Unlocks the Full Power of Airflow to Accelerate AI Innovation

As mentioned earlier, ML teams are adopting Airflow as an orchestration platform that meets many of the requirements of ML operations, including those listed above. For all its strengths, however, it can be difficult to manage a platform as sophisticated as Airflow, especially once it expands to multiple teams, specifically in regard to scaling AI initiatives.



Astronomer is a unified data platform built on Apache Airflow that enhances efficiency and reach, providing a central place where data and ML engineers can meet to bridge the gap between their teams, collaborate, orchestrate their data, and accelerate their organization's AI initiatives. With these capabilities, businesses can →

Unify and Standardize Development Practices for Production-Ready AI

Astronomer supports the entire AI lifecycle, from prototype to production, with its unified and standardized environment for AI development. The platform also drives collaboration between data and ML engineers across traditional data pipelines, preparing ML for production, and building AI applications on Airflow. Astronomer also provides a common framework and enforces best practices for all teams to unite around and effectively orchestrate the development and deployment via integrated IDE and CI/CD (continuous integration and continuous delivery/deployment) and pluggable compute. Additionally, it includes complete monitoring, alerting, and data lineage, guaranteeing enterprise-grade uptime to minimize the risk of costly AI operation outages.

Support Next-Generation Applications with Unmatched Compute Power

Astronomer offers the largest compute power in the managed-Airflow market, twice as much as the nearest competitor, making it ideal for businesses that are looking to scale up their AI workloads. In addition, other features let organizations stay efficient and cost-effective as they scale. For instance, the worker queue feature enables teams to dedicate larger machine types to their heaviest workloads.

CASE STUDY

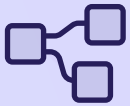
How Anastasia delivers AI-powered insights with Astro

Learn more →

CASE STUDY

Laurel's timekeeping transformation with AI and Airflow

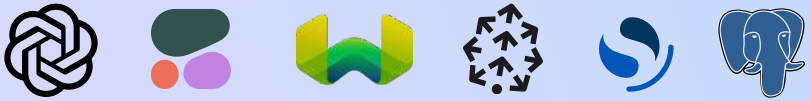
Learn more →



Ensure AI Trustworthiness with Data Lineage

Astronomer automates the extraction and analysis of lineage metadata. Armed with this data, the platform's Lineage view gives data science and ML teams clear visibility into the origins and transformations of data, to improve troubleshooting, aid in the validation of results, and enhance the overall reliability of AI models.

Astronomer's lineage metadata also puts the information into organizations' hands to effectively govern their data. It enables CAOs, CDOs, data stewards, and others to pinpoint silos and safeguard sensitive data, so they can effectively bring practices into compliance with regulatory requirements and ensure data is reused responsibly. Features like Day 2 Ops, comprehensive monitoring and alerting, and lineage guarantee that compliance will continue to be maintained.



Accelerate AI Development with Seamless Integrations

Airflow has hundreds of integrations with almost all components of the modern data stack, and the **Astronomer Registry** provides a curated library of providers with examples and documentation. This includes integrations with tools and platforms that support AI use cases and empower organizations to harness the full potential of LLMs and AI out of the box — examples include OpenAI, Cohere, Weaviate, Pinecone, Pgvector, OpenSearch, and frameworks like SageMaker and AzureML.

Integration with leading providers accelerates the AI development process by easing integration complexity. Data engineers and scientists can effortlessly leverage diverse technologies, focusing on building impactful models and applications without the usual interoperability challenges, making it easier to get AI into production.

ASTRONOMER

Ready to accelerate AI Innovation?

Operationalize and scale AI and ML initiatives, unlock the full power of Airflow to deliver production-ready AI, and accelerate workflow development with Astronomer.

Power your next big AI project.

Try Astro Free for 14 days →