Astro: Modern Data Orchestration Managed by Astronomer

Navid Aghdaie
Airflow Summit, May 2022
Astronomer at a Glance

The Numbers

2018
Year Founded

300
Employees covering six continents

150 +
Engineering Team

Investors

SIERRA VENTURES
venrock
SUTTER HILL VENTURES
INSIGHT PARTNERS
salesforce
J.P. Morgan
Bain Capital
MERITECH
The de facto standard and a vibrant community

Huge Supporters and Contributors to Airflow, Commercial Provider

How to enable the modern data orchestration platform
This Session’s Focus: Astro
astronomer.io/product
About Me

UCLA
Computer Science PhD
Fault-Tolerant Distributed Systems Web Services

Ask
VP Technology
Data Systems Web/News Search

EA
Sr Director Engineering
Data & AI Platform and Services Games

Astronomer
VP Engineering
Cloud-native Managed Services Data Orchestration
The Growing Need for Managed Orchestration

Industry Trends

Success of Data Driven Businesses
- Criticality of Data
- Rise of Cloud Services
- Proliferation of Data and Sources

⇒ Data Industry Segment Growth
- Diversification of Data Ecosystem
- Increased Complexity in Data Pipelines and Dependencies
- Scarce Engineering Talent

⇒ New Ideologies
- Data Democratization
- AI Everywhere

⇒ Difficult challenges for data pipeline and platform developers
  SLAs, Visibility, Data Quality, Issue Remediations (e.g., reprocessing), Compliance, Operational Overhead...

⇒ Need Managed Orchestration that Just Works!
The data ecosystem is powerful...but complex to Integrate and Operate.

Source: https://mattturck.com/data2020/
Data Never Stops

**Build**
Simplify the development of data pipelines with an integrated workflow.

**Run**
Harness the power of the cloud with a secure and performant Airflow runtime.

**Observe**
Gain visibility and insight into how your data ecosystem is performing.
Astro Cloud Architecture

Control Plane
- Registry
- Cloud UI
- Astro CLI
- Account
- Academy

Data Plane
- Runtime

Data Services
- AWS
- A
- Google Cloud
- Google Sheets
- Google Drive
- Cloudflare
- Apache Airflow
Taking a Deeper Look

Secure Access via SSO Integration

Control Plane
End-to-end visibility and control in Astronomer’s Cloud

Manage Environments
Deploy Projects
Control Users/Access
View Metrics

Deploy Airflow Projects / Access Airflow UI
Pull Images
Send Metrics/Alerts

Data Plane
Airflow Running in Your Cloud

Astro Cluster
company-east-1

Astro Cluster
company-west-2

Private & Secure Connectivity

External Data Systems

Databases
Data Lakes & Warehouses
APIs
SaaS Products
Data Applications
Data Science & ML Tools
Three Common Examples of Pain Point Areas

- Onboarding & Productivity
- Scalability
- Observability
Onboarding & Productivity

Problem:

- Setup experience is burdensome
- Maintenance overhead requires engineering time
  - Airflow version upgrades, patches
  - Availability and robustness issues due to failures at various pipeline stages

With Astro:

- **Optimized provisioning** of infrastructure on multiple Clouds and Regions (Data Plane)
  - All you need is a Cloud account
  - Spin up isolated environments in minutes - eg dev, test, prod
  - Astro CLI enables integration with your CI/CD, Develop locally and push to production
- **Runtime versioning** simplifies upgrades by handling applicable adjustments to Data Plane
- System and Pipelines **monitored** and issues remediated **by Astronomer**
  - Eg transient network/compute failures on public cloud, resource availability, etc
Scalability

Problem:

- Scaling usage is complex, costly, and adds operational overhead
  - More users, teams, use-cases, pipelines, ...

With Astro:

- **Control Plane** designed with **scalability** in mind
  - IAM/RBAC, SSO, Workspaces help you manage users and teams
  - Manage 100s of Airflow deployments easily, on a single pane of glass
  - Clusters on different VPCs, Regions, Clouds

- **Runtime** engineered for Cloud with **Auto-Scaling**
  - Auto-Scaling with configurable resource limits
    - Resource Monitoring per Deployment (eg tasks running, queued)
    - Scale workers up or down
  - Leverage efficiency gains of Airflow advancements
    - eg Deferrable Tasks
Observability

Problem:

- Lack of visibility into:
  - Overall end-to-end health
  - Lineage
  - Dependencies

⇒ Extended outages and poor data quality.
Working to advance Apache Airflow

The principal drivers of OpenLineage
**Enhanced Data Observability**

**Single View of Pipelines**
Distributed pipelines come together with central observability across deployments.

**Operational Lineage Explorer**
Pinpoint root cause, with a full understanding of upstream and downstream impact.

**Task-Level Resource Visibility**
Granular worker consumption visibility helps you unblock starved tasks and identify opportunities for optimization.
Real-Time, Operational Lineage Unlocks Shared Understanding

Resolve Data Outages Faster
identify root causes, determine impacts, and remediate issues that cause data downtime with less effort

Make Sense of Cross-Team Dependencies
explore and understand complex dependencies across pipelines, environments, and clouds

Visualize Quality and Performance Over Time
pinpoint bad data and bottlenecks sooner, and quickly remediate impacts throughout your data ecosystem
Add Lineage Support with No Code Changes

Operators facilitate the exchange of instruction and information with external data services

The certified OpenLineage Package in Astro Runtime includes extractors, that attach to supported operators automatically at task runtime

The extractor receives lineage and performance information from the external data service at execution time
Lineage in Astro

Data Plane

Operator-Based Integrations

Control Plane

Capture Lineage Events with OpenLineage Extractors in Astro Runtime

OpenLineage Metadata

Trace Full Upstream and Downstream Paths in the Lineage Explorer

Understand Trends with Data Quality Monitoring

Direct Integrations
Astro
The Orchestration Cloud

The modern data orchestration platform, powered by Apache Airflow, that empowers the entire data team to build, run, and observe data pipelines.
Keep Your Data Flowing with Astro

Get a demo that’s customized around your unique data orchestration workflows and pain points.
We are hiring! Join us!

https://www.astronomer.io/careers/
Thank you