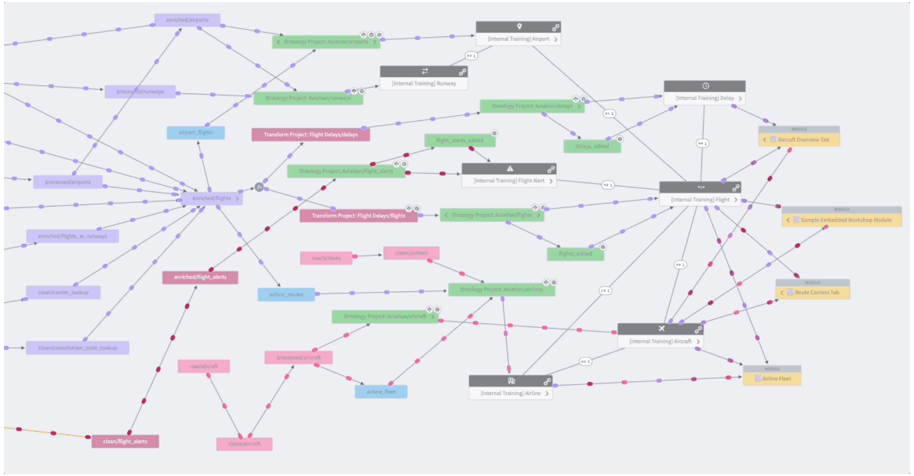
Introduction to Data Lineage

Understanding Data Flow and Provenance in Foundry

- **What is Data Lineage?:** Data Lineage is an interactive Foundry tool that enables users to visualize and trace the complete flow of data—from raw sources through transformations to final datasets—providing transparency and control across the entire data pipeline.

- **Purpose and Importance:** By mapping data relationships and dependencies, Data Lineage helps ensure data quality, auditability, and compliance, while enabling teams to diagnose issues and optimize pipeline performance.

- **Key Benefits:** • Holistic understanding of data flow • Easier discovery of datasets • Faster troubleshooting • Collaborative data exploration and governance.


enriched/airports
Ontology Project: Aviation/airports
[Internal Training] Airport
[Internal Training] Runway
processed/runways
Ontology Project: Aviation/runways
Ontology Project: Aviation/delays
[Internal Training] Delay
airport_flights
Transform Project: Flight Delays/delays
flight_alerts_edited
delays_edited
MODULE
Aircraft Overview Tab
processed/airports
Ontology Project: Aviation/flight_alerts
[Internal Training] Flight Alert
enriched/flights
Transform Project: Flight Delays/flights
Ontology Project: Aviation/flights
[Internal Training] Flight
MODULE
Sample Embedded Workshop Module
enriched/flights_w_runways
raw/airlines
clean/airlines
flights_edited
clean/carrier_lookup
enriched/flight_alerts
airline_routes
Ontology Project: Aviation/airlines
MODULE
Route Carriers Tab
Ontology Project: Aviation/aircraft
[Internal Training] Aircraft
clean/cancellation_code_lookup
raw/aircraft
processed/aircraft
airline_fleet
[Internal Training] Airline
MODULE
Airline Fleet
clean/flight_alerts
clean/aircraft
↔ 1
↔ 2

Core Capabilities of Data Lineage

Powerful Tools for Data Discovery and Pipeline Exploration

- **Dataset Discovery and Search:** Users can quickly find datasets by project, table, or column name. Data Lineage supports browsing through Foundry Projects and filtering via advanced search options for targeted exploration.

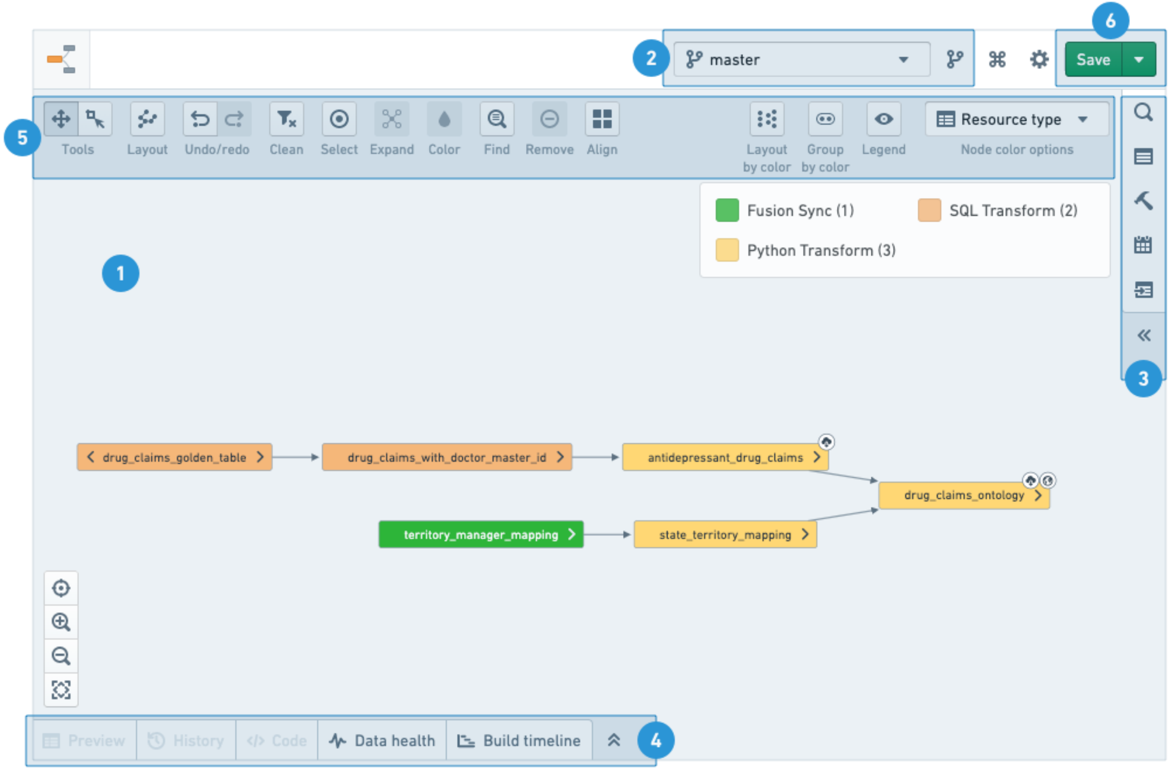
- **Graph-Based Exploration:** The interactive lineage graph provides an intuitive workspace to visualize, expand, and manipulate data relationships, allowing teams to trace dependencies and identify bottlenecks.

- **Pipeline Management:** Users can view and manage data pipelines, inspect schema and code, and color nodes to represent health, build status, or ownership for monitoring and troubleshooting.

- **Collaboration and Sharing:** Create snapshots of lineage graphs to share with teammates, enabling collaborative troubleshooting and documentation of data processes.

Interface Overview

Navigating the Data Lineage Environment



Interface Overview

Navigating the Data Lineage Environment

Lineage Graph Workspace

The graph is the central workspace where nodes representing datasets, artifacts, or object types are visualized. Users can pan, zoom, expand ancestors and descendants, and apply auto or manual layouts.

Side Panel and Search Tools

The side panel allows dataset search, filtering, and browsing of Foundry resources. Users can add resources directly to the graph or apply advanced search filters for granular control.

Graph Tools and Layout Options

Includes tools for node selection, expansion, and layout customization (vertical, hierarchical, or grouped). Color schemes and shortcuts streamline navigation and data inspection.

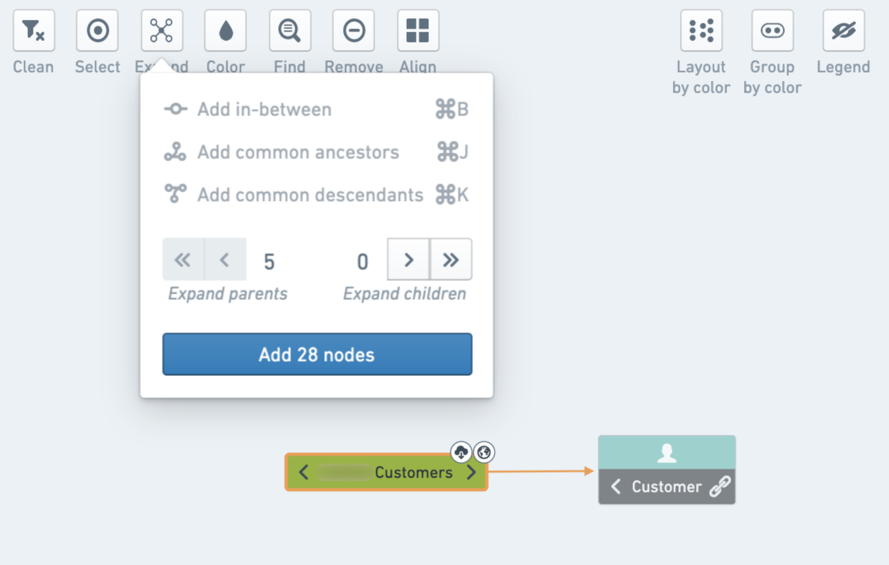
Properties and Histograms

When selecting nodes, users can view detailed attributes, dataset health, and frequency histograms to analyze shared properties and identify outliers.

Exploring Data Pipelines

Tracing Data Flows and Dependencies

- **Visualize and Expand Relationships:** Users can add datasets to the graph and expand to view ancestors or descendants, revealing upstream and downstream dependencies to understand full lineage paths.
- **Drill into Dataset Details:** Each node reveals metadata such as schema, build status, history, and the code that generated it, helping users connect logical and physical transformations.
- **Interactive Exploration:** Tools like Expand, Find, and Selection enable dynamic navigation through complex pipelines, supporting selective exploration and filtering by resource type or attribute.
- **Performance and Best Practices:** To maintain usability, users are advised to limit node expansion, track performance via node count, and focus on relevant datasets for clarity and insight.



Managing Builds and Schedules

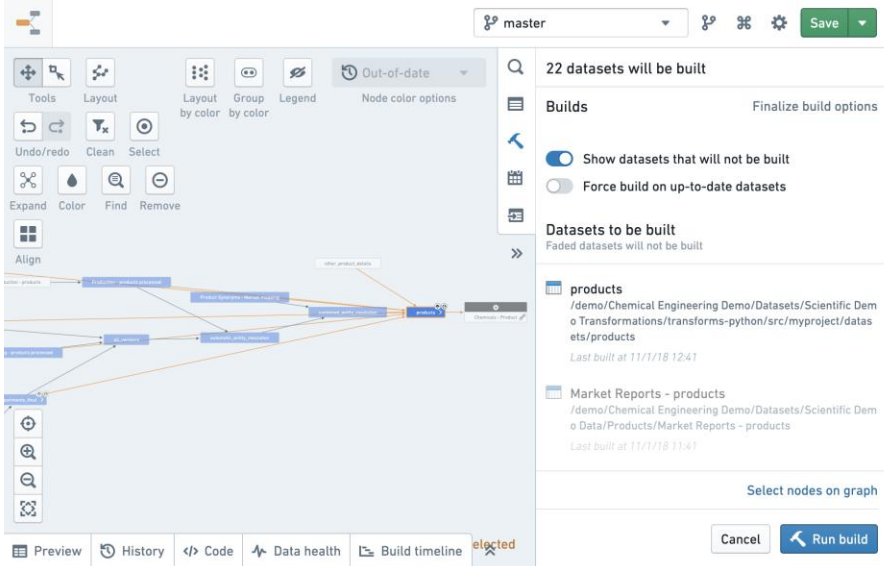
Controlling Data Refresh and Pipeline Automation

- **Build Strategies:** Data Lineage supports multiple build workflows—building selected datasets, all ancestors, or all transforms in between selected datasets—to optimize performance and ensure up-to-date data.

- **Preview and Execution:** Before triggering builds, users can preview the datasets that will be built, select forced rebuilds, and validate dependency integrity for precise execution control.

- **Schedule Management:** Schedules can be configured directly within Data Lineage, defining when and how pipelines should run based on dataset updates, time triggers, or dependencies.

- **Monitoring and Logs:** Users can view latest runs, update timestamps, job details, and build timelines as Gantt charts to evaluate pipeline performance and troubleshoot failures.



22 datasets will be built

Builds

Finalize build options

● Show datasets that will not be built

○ Force build on up-to-date datasets

Datasets to be built

Faded datasets will not be built

products
/demo/Chemical Engineering Demo/Datasets/Scientific Demo Transformations/transforms-python/src/myproject/datasets/products
Last built at 11/1/18 12:41

Market Reports - products
/demo/Chemical Engineering Demo/Datasets/Scientific Demo Data/Products/Market Reports - products
Last built at 11/1/18 11:41

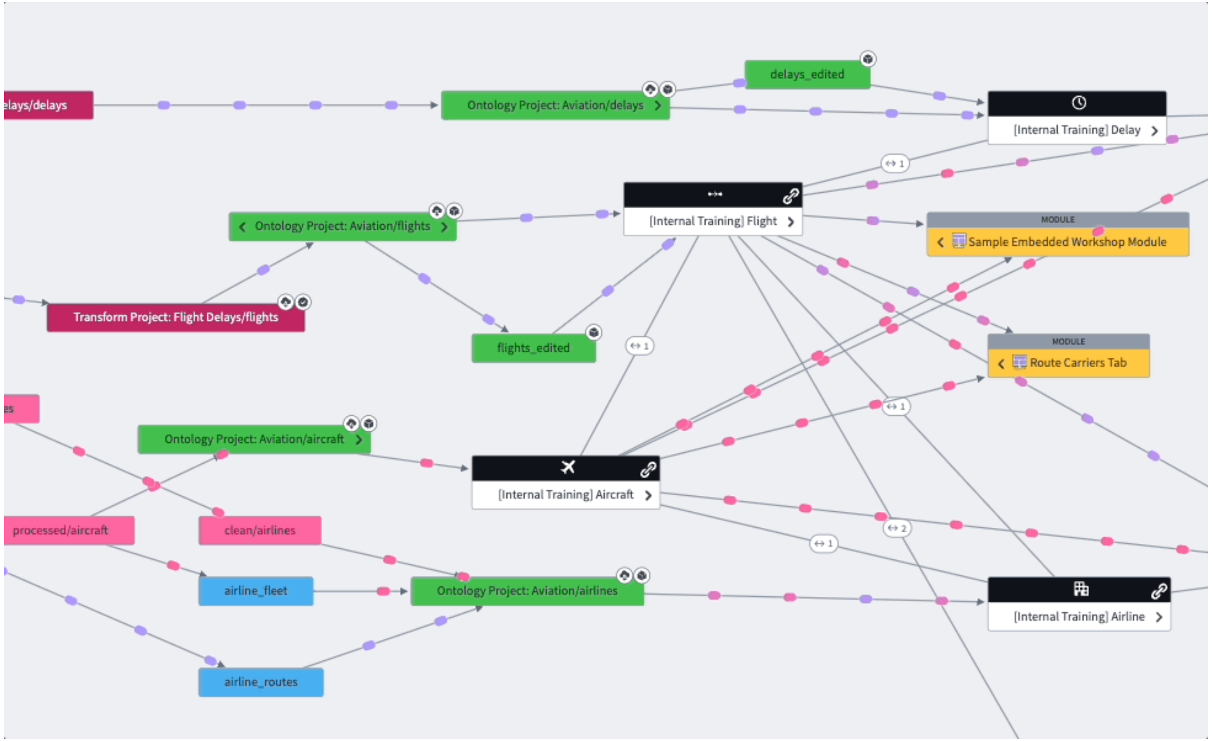
Select nodes on graph

Cancel

Run build

Node Coloring and Visualization

Decoding Data Health, Status, and Structure



elays/delays
Ontology Project: Aviation/delays
delays_edited
[Internal Training] Delay
Ontology Project: Aviation/flights
[Internal Training] Flight
MODULE
Sample Embedded Workshop Module
Transform Project: Flight Delays/flights
flights_edited
MODULE
Route Carriers Tab
es
Ontology Project: Aviation/aircraft
[Internal Training] Aircraft
processed/aircraft
clean/airlines
airline_fleet
Ontology Project: Aviation/airlines
[Internal Training] Airline
airline_routes
↔ 1
↔ 1
↔ 1
↔ 2
↔ 1

Node Coloring and Visualization

Decoding Data Health, Status, and Structure

Purpose of Node Coloring

Node coloring provides instant visual cues about dataset attributes such as build status, health, permissions, and project grouping, improving interpretability of complex graphs.

Coloring Options

Users can color nodes by over twenty metrics, including Resource Type, Build Status, Data Health, Permissions, Project, and Storage. Custom coloring enables tailored visual insights.

Health and Performance Indicators

Health-based coloring highlights datasets with failed checks or outdated builds, enabling rapid issue detection and prioritization for maintenance.

Practical Use Cases

Coloring supports root cause analysis, access auditing, and compliance tracking by visualizing data quality and access status across entire pipelines.

Collaboration and Sharing

Enabling Teamwork and Transparency in Data Lineage

Sharing Lineage Graphs
Users can save, export, or share their lineage graphs using quick share links or SVG exports, providing others with read-only access or interactive collaboration within Foundry.

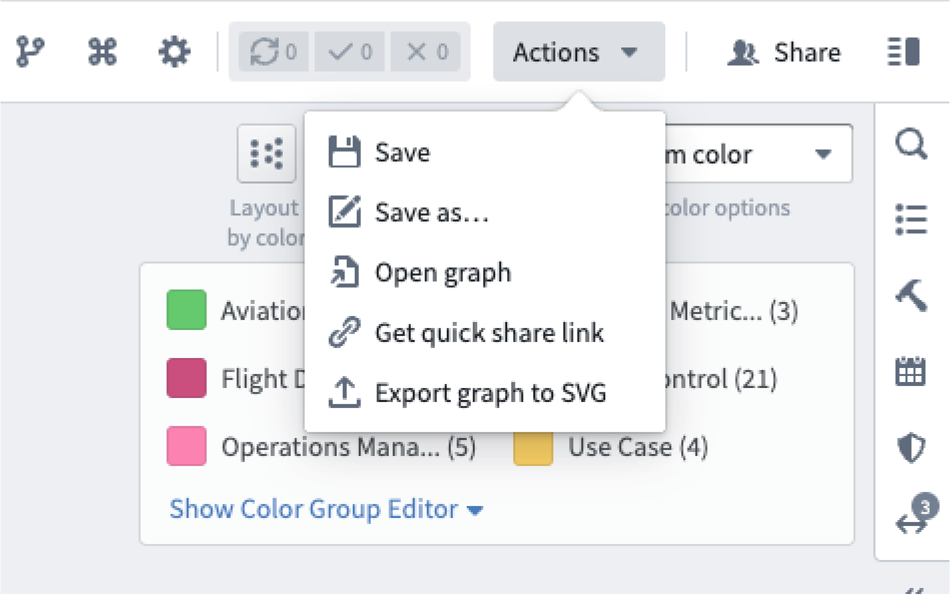
Version Control and Snapshots
Lineage snapshots preserve the current state of a pipeline for review or troubleshooting, allowing teams to document changes and maintain historical visibility.

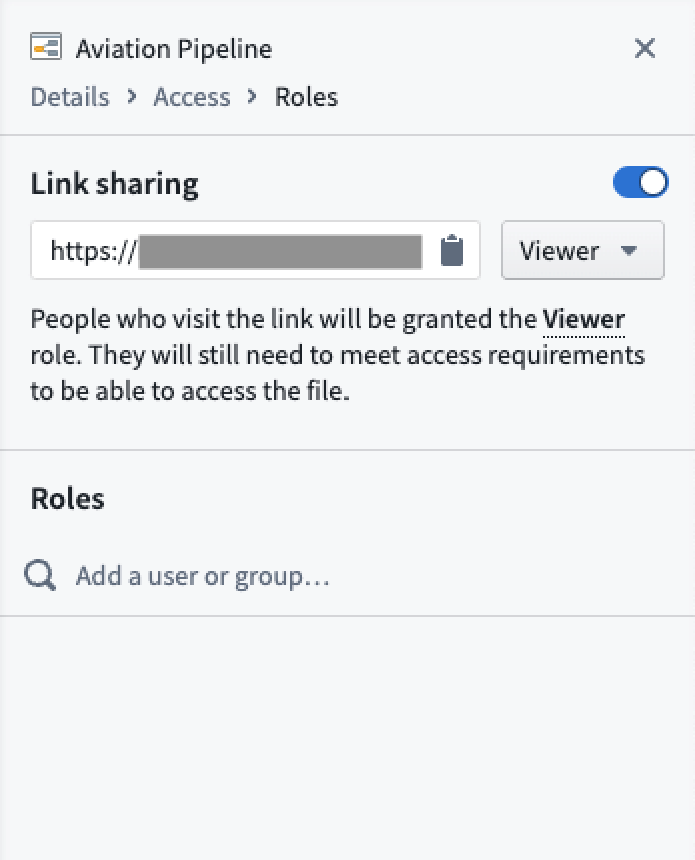
Role-Based Access
Access controls enable fine-grained sharing of lineage assets. Teams can assign viewer, editor, or admin roles to manage permissions across projects and organizations.

Cross-Team Collaboration
Shared graphs facilitate communication among engineers, analysts, and governance teams, turning lineage views into shared knowledge bases for coordinated decision-making.

Collaboration and Sharing

Enabling Teamwork and Transparency in Data Lineage





People who visit the link will be granted the **Viewer** role. They will still need to meet access requirements to be able to access the file.

Pipeline Rollback

Restoring Data Integrity Through Controlled Reversion

Purpose of Rollback

Pipeline rollback enables users to revert datasets and their downstream dependents to earlier, stable versions, maintaining data integrity after logic or input errors.

Rollback Execution

Users select a dataset, choose a branch, and pick a transaction to roll back to. Data Lineage previews affected downstream datasets and unsupported resources before confirmation.

Warnings and Safeguards

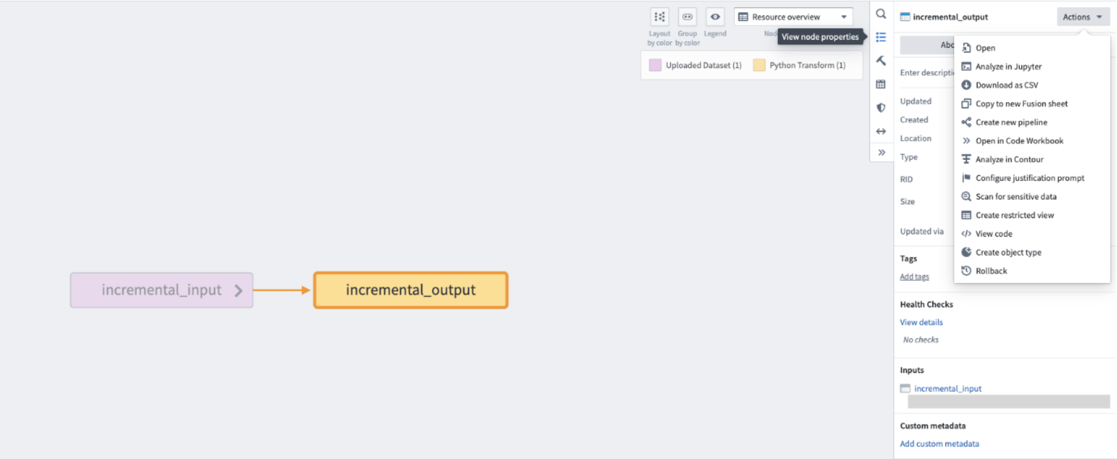
The system highlights potential conflicts, such as logic changes or missing downstream datasets, ensuring informed decision-making and controlled reversion.

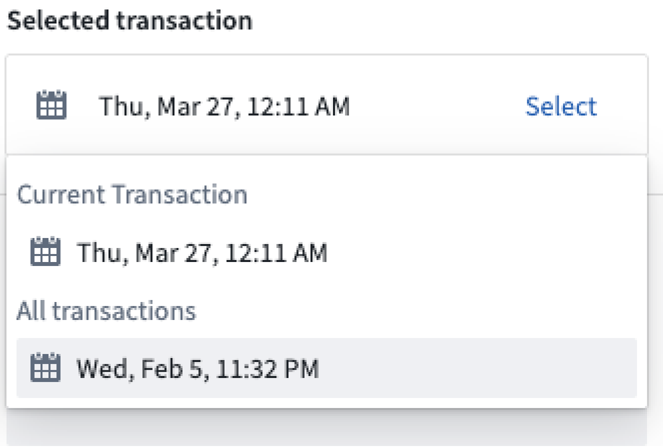
Limitations

Only transactional datasets are supported for rollback; media sets, streaming datasets, or resources without JobSpecs remain unchanged.

Pipeline Rollback

Restoring Data Integrity Through Controlled Reversion




Selected transaction

Thu, Mar 27, 12:11 AM Select

Current Transaction

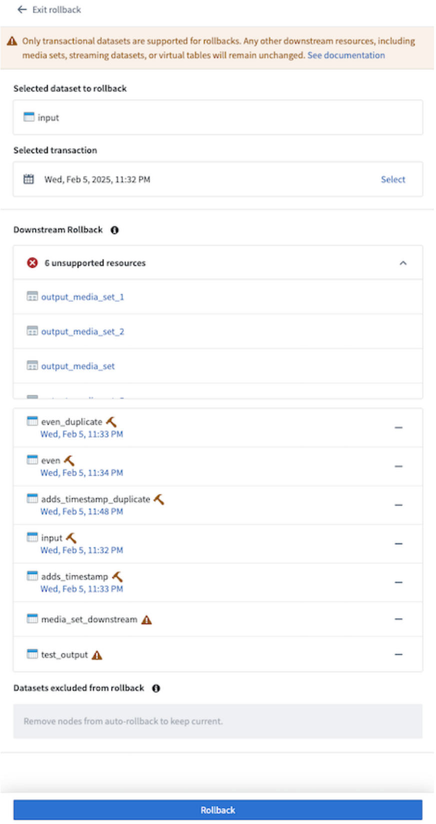
Thu, Mar 27, 12:11 AM

All transactions

Wed, Feb 5, 11:32 PM

Pipeline Rollback

Restoring Data Integrity Through Controlled Reversion



⚠ Only transactional datasets are supported for rollbacks. Any other downstream resources, including media sets, streaming datasets, or virtual tables will remain unchanged. See documentation

Selected dataset to rollback

input

Selected transaction

Wed, Feb 5, 2025, 11:32 PM

Downstream Rollback

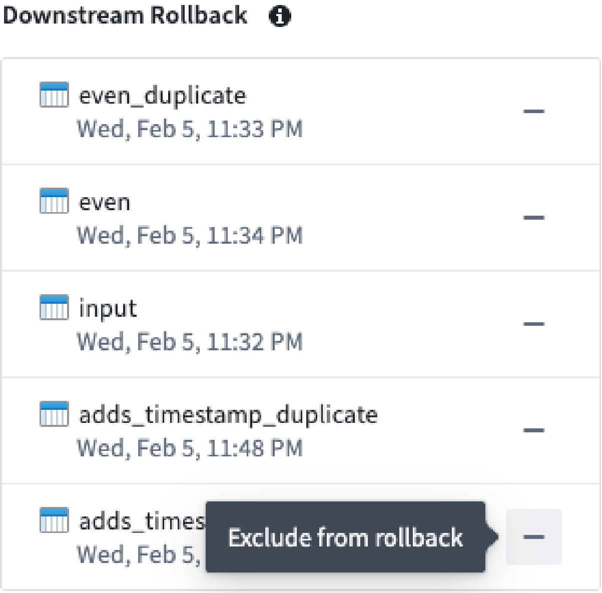
6 unsupported resources

- output_media_set_1
- output_media_set_2
- output_media_set

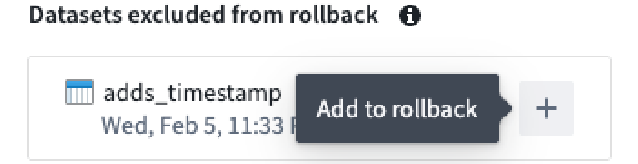
even_duplicate Wed, Feb 5, 11:33 PM	–
even Wed, Feb 5, 11:34 PM	–
adds_timestamp_duplicate Wed, Feb 5, 11:48 PM	–
input Wed, Feb 5, 11:32 PM	–
adds_timestamp Wed, Feb 5, 11:33 PM	–
media_set_downstream ⚠	–
test_output ⚠	–

Datasets excluded from rollback

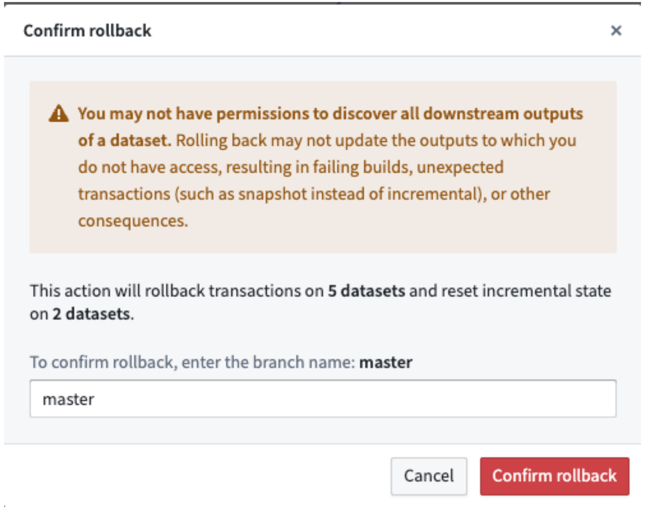
Remove nodes from auto-rollback to keep current.


Downstream Rollback

even_duplicate Wed, Feb 5, 11:33 PM	–
even Wed, Feb 5, 11:34 PM	–
input Wed, Feb 5, 11:32 PM	–
adds_timestamp_duplicate Wed, Feb 5, 11:48 PM	–
adds_times Wed, Feb 5,	–


Datasets excluded from rollback

adds_timestamp
Wed, Feb 5, 11:33


Confirm rollback

⚠ **You may not have permissions to discover all downstream outputs of a dataset.** Rolling back may not update the outputs to which you do not have access, resulting in failing builds, unexpected transactions (such as snapshot instead of incremental), or other consequences.

This action will rollback transactions on **5 datasets** and reset incremental state on **2 datasets**.

To confirm rollback, enter the branch name: **master**

master

Data Permissions and Marking Simulation

Assessing Access, Governance, and Compliance

Permission Visualization

Using the 'Permissions' coloring option, users can visualize dataset access levels across the graph, highlighting who can view or edit data resources.

Access Simulation Mode

Simulation mode enables users to apply or remove Markings and preview how changes affect data access propagation through the lineage graph.

Impact Analysis

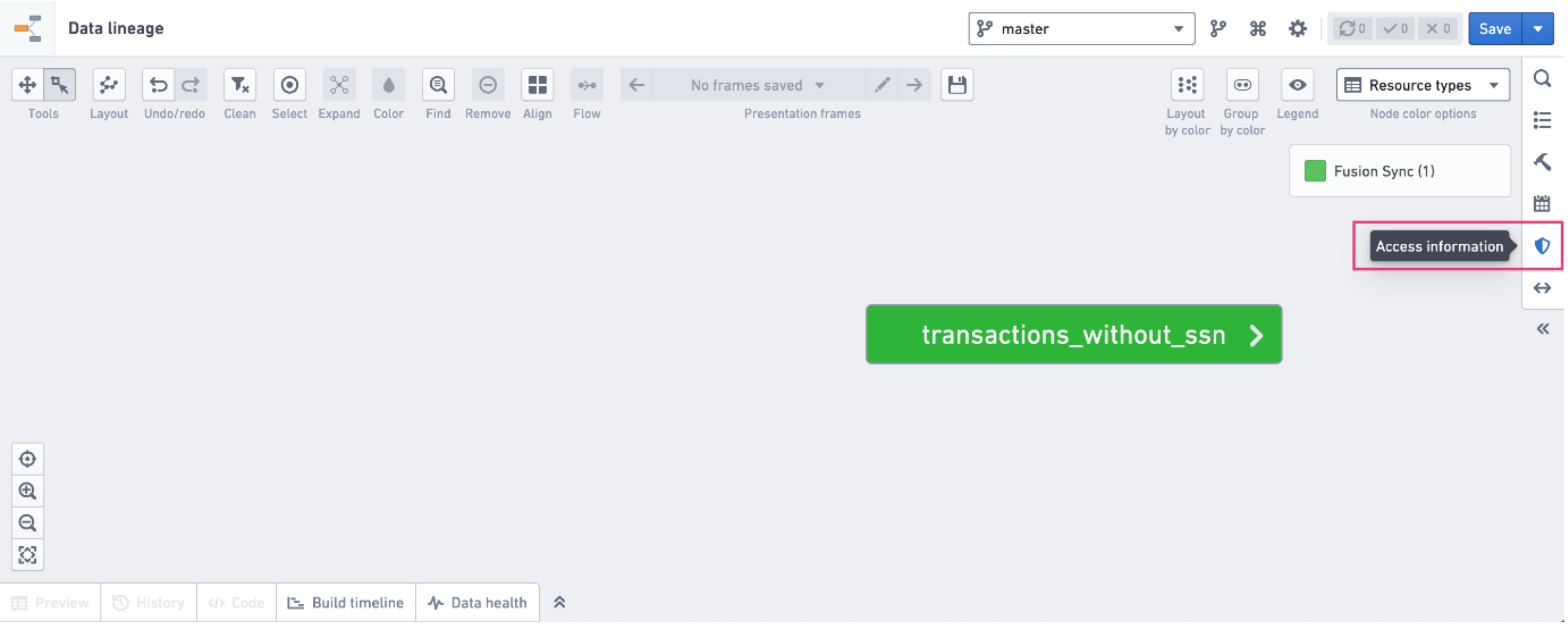
Datasets are color-coded as access affected, unaffected, or unchanged, helping governance teams assess risk and compliance impacts before enforcing changes.

Troubleshooting and Governance

By inspecting permissions and Marking propagation, users can detect inconsistencies, ensure least-privilege access, and maintain compliance with data regulations.

Data Permissions and Marking Simulation

Assessing Access, Governance, and Compliance



Conclusion

Empowering Data Transparency and Trust with Data Lineage

- **Unified Data Understanding:** Data Lineage connects the dots between datasets, transformations, and systems—helping organizations build a shared understanding of their data ecosystem.

- **Operational Efficiency:** Through visual mapping, automation, and rollbacks, teams can reduce downtime, streamline builds, and ensure the accuracy of analytical outputs.

- **Enhanced Governance and Compliance:** By integrating permissions, health monitoring, and Marking simulation, Data Lineage strengthens transparency and data stewardship.

- **Future-Ready Data Management:** As data landscapes grow in complexity, Data Lineage provides the foundation for scalable, compliant, and intelligent data operations.