dbt Labs™

dbt Cloud:
The control plane for data collaboration at scale

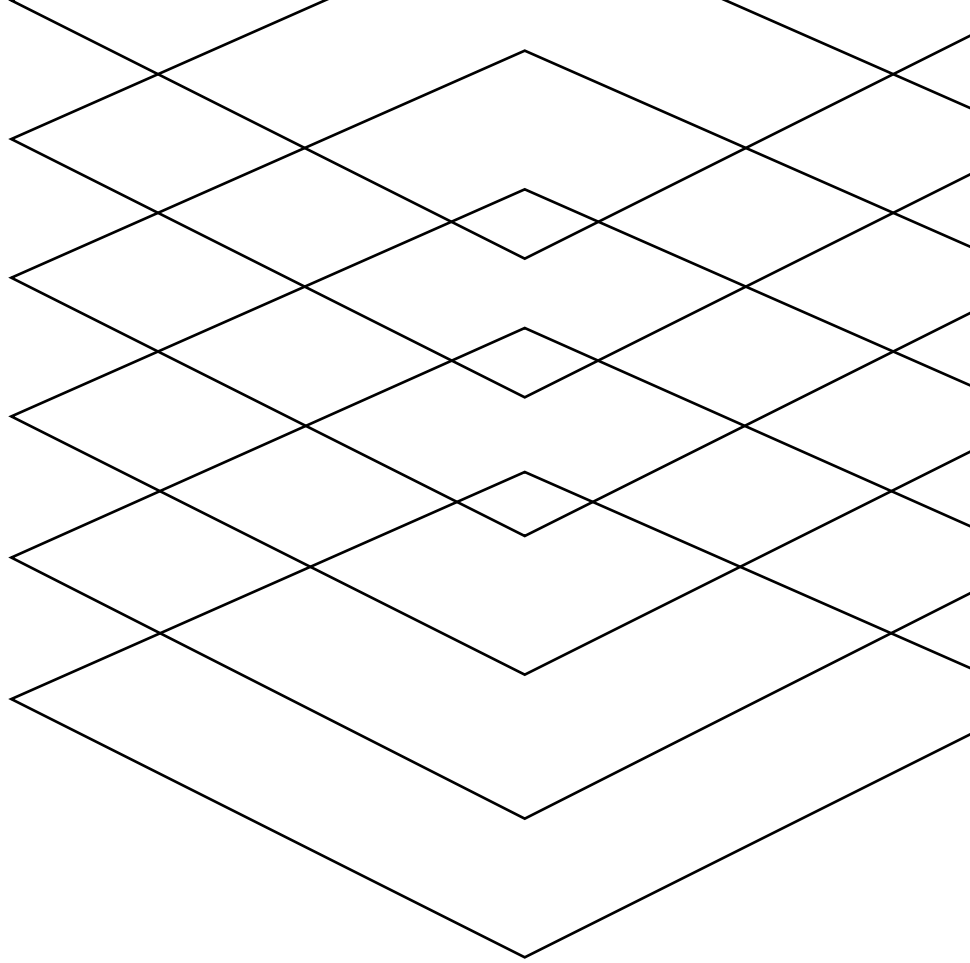Whitepaper

Table of Contents

![dbt Labs logo]

dbt Cloud:
The control plane for data collaboration at scale

In today's digital world, organizations are capturing more customer, product, and operational data than ever before. In the past decade, the rise of cloud data platforms and, importantly, the separation of storage and compute costs have paved the way for a more efficient integration and transformation paradigm: ELT—Extract, Load, Transform.

Transforming raw data *after* it has been loaded in a data platform improves flexibility and agility for how organizations glean insights from data. As organizations standardize on cloud data platforms, they're also modernizing their approach to data integration and transformation so they can effectively use data across the business to build lasting competitive advantage.

More data, more problems?

Access to more data hasn't always resulted in more answers. In many cases, it's actually just led to more chaos. Data transformations are still being done as they were in on-prem environments: they're built and deployed in silos, they're error prone, and they're expensive to maintain. With more and more downstream teams looking to embrace data-driven decision making, data teams are drowning in one-off requests and spend the majority of their time in reactive mode. Until recently, they have lacked the tools needed to be proactive and build on their work.

All of this amounts to a chaotic data workflow that results in a lack of data trust, hindered organizational agility, inefficient spend across the data stack, and wasted resources. The stakes are high to get this right, especially now, as macroeconomic headwinds require companies to do more with less and prove out the value of investments they've already made.

The challenges of managing data at scale

These challenges can be summarized in three categories:

1. **Data teams don't have a standardized approach to build and maintain data products at scale.** Ask any data professional about their workday, and you'll hear a similar story: they spend their precious time debugging code, managing brittle pipelines, and reacting to one-off, urgent requests. Analytics engineers and data analysts have inherited clunky analytics code and bloated pipelines they're now responsible for managing. Meanwhile, data engineers have specialized expertise in the data tools that are required to initiate any downstream data work. No one can move fast enough for their stakeholders, bottlenecks abound, and frustrated stakeholders seek out their own methods of deriving data insights. Without a standardized way to scale, document, and modularly improve on their work, data teams are perpetually stuck in reactive mode, fielding a never-ending queue of requests.

2. **There's a lack of organizational trust in data.** Data analytics are built in service of some larger business objective, or to answer a particular question...but far too often, metrics are inconsistent, incomplete, or incorrect, which impairs trust between data consumers and data producers. Data pipelines are perceived as a black box and business stakeholders don't have confidence in how data is derived. This not only impacts their ability to make the right decisions, but also stymies organizational efforts to become more data-driven.

3. **Data costs spiral out of control and ROI becomes increasingly difficult to prove.** The first two challenges create an unsustainable culture of reactive workflows and debatable data analytics. If the organization doesn't have an effective system for documenting and discovering existing data products, employees will inevitably build their own, unaware that they are recreating existing work. Not only does this lead to wasted time and easily avoidable data platform spend on duplicative code, but it also creates confusion about "what's the source of truth" and begets imprecise decision making.

The sum total of all of this?
Frustrated business stakeholders, burned out data teams, and forgone customer value.

Manage complexity at scale with dbt Cloud

dbt Cloud gives data teams a standardized way of developing, testing, and deploying data products at scale. dbt has quickly become the industry standard for data transformation, and dbt Cloud delivers dbt as a service with the critical security, governance, automation, and collaboration features required for managing data complexity at scale.

The old way
Before dbt came along, data teams followed a monolithic approach to data modeling: they'd build (and inevitably rebuild) 10,000+ line SQL files from raw data or stored procedures, with no way to discover, understand, or reuse similar work that had been created earlier. New models would be rewritten from raw source data, and model dependencies visualized in a DAG revealed a lot of overlapping source data usage.
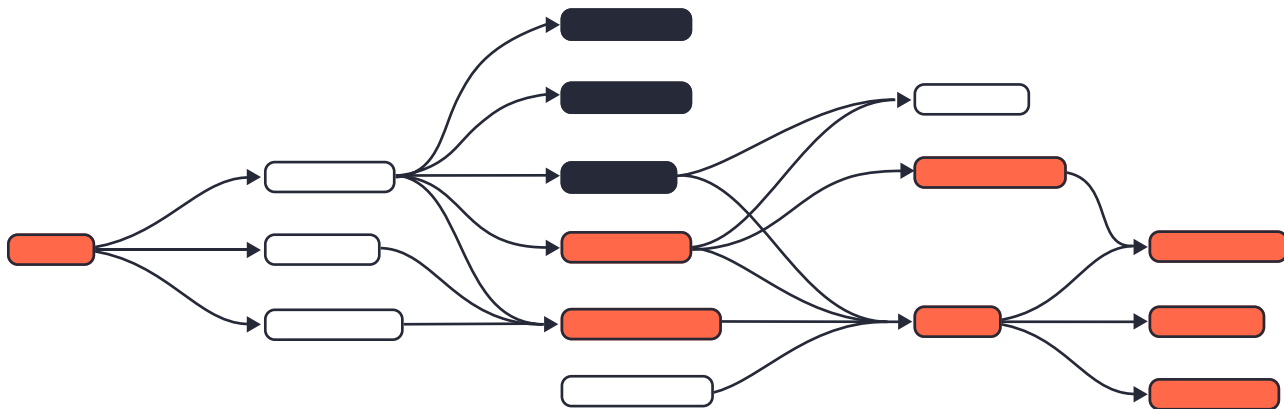
The dbt way
dbt promotes model reuse and discoverability and helps break down silos by fostering a modular approach to data modeling. This means that data models are bite-sized components that data producers and consumers can reference, improve, and build upon. Suddenly, it becomes a lot more straightforward for data teams to understand the foundational modeling work that has already been done and how to build upon it for new use cases. This improves data quality and helps data teams build and ship pipelines faster.

With a modular approach to data modeling using dbt Cloud, entire organizations can collaboratively harmonize their approach to data: data engineers can maintain control of data pipelines and governance while avoiding reactive and toilsome workflows, analysts can participate in data modeling using just SQL, and business stakeholders have the confidence and agility to make business decisions with accurate, trustworthy data. This dynamic fosters a self-fulfilling flywheel between business stakeholders and data producers to seamlessly integrate actionable data into decision-making processes.

dbt Cloud platform features

dbt Cloud is a managed service that sits on top of a user's cloud data platform and enables a workflow where data teams can write analytics code in SQL or Python and leverage software engineering best practices to build, test, document, and deploy data models for downstream consumption. Over 3,500+ organizations across the globe rely on dbt Cloud to manage the complexity of their modern data environment. Using dbt Cloud, organizations can:

- **Foster collaboration at scale:** dbt Cloud offers built-in features that make data transformation a truly "multi-player" sport. Anyone with SQL knowledge can participate in transformation workflows and write code from whatever environment they're most comfortable in (whether the command line or a hosted IDE). With native BI integrations, less technical stakeholders can also self-serve into data models and metric definitions.

- **Understand and improve data pipelines:** dbt Cloud is built on a stateful metadata platform and offers intuitive Uls so teams can visualize lineage, understand dependencies, discover run status, and grasp other critical context across the DAG to explore and ultimately improve their dbt estate.

- **Embrace software engineering best practices:** dbt Cloud makes it easy for data teams to adopt software engineering best practices at every phase of the development and deployment lifecycle with built-in orchestration, Cl, and observability.

- **Get started fast:** dbt Cloud delivers dbt as a managed service and provides critical security and governance features that allow users to build and automate data transformations across various environments.
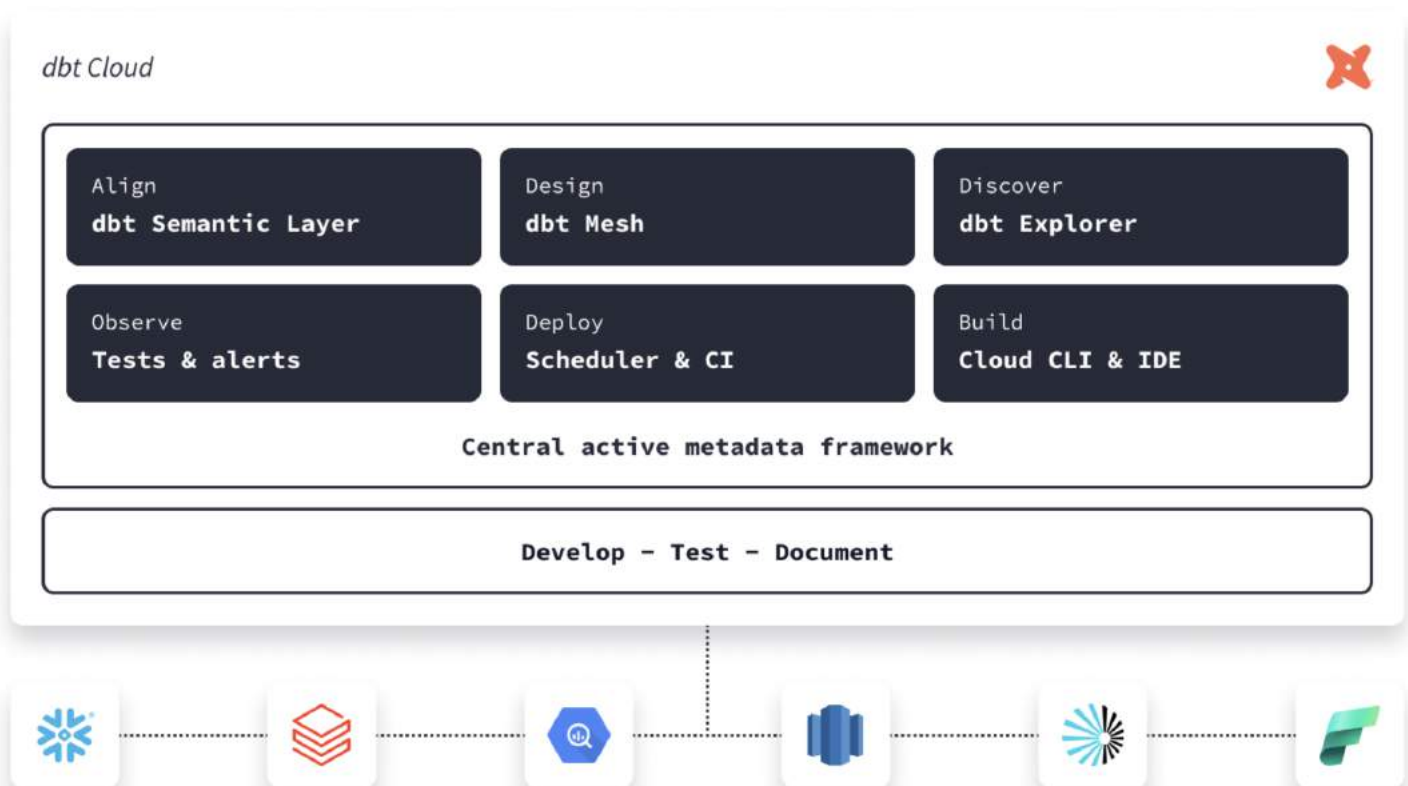
## dbt Semantic Layer: Define metrics centrally and access them everywhere

Organizations that can make better data-driven decisions win. But what happens when the data you're using is inconsistent across teams? This happens more often than not, especially when you consider the number of tools and teams that are leveraging an evolving and growing dataset. Answering a question as seemingly straightforward as "what was our revenue last month?" can turn up a handful of varying answers, which leads to frustration, diminished trust, and ineffective decision-making.

Use the dbt Semantic Layer to centrally define your metrics right alongside your dbt models, and query them across a number of BI platforms. You can use leading analytics tools like Tableau, Google Sheets, Mode (ThoughtSpot), Hex, and Lightdash to query your data and get the same answers everywhere, every time.

The dbt Semantic Layer is powered by MetricFlow, which enables complex metric definition and querying to be done efficiently at any scale. With dynamic join support, MetricFlow can infer the appropriate traversal paths between multiple tables when generating a metric, making all valid dimensions available for your metrics on the fly. The dbt Semantic Layer also supports an array of metrics types, from simple measures to more complex aggregates, cumulative metrics, and ratios.

Defining and querying metrics with the dbt Semantic Layer helps teams:

- Trust their numbers: With consistent metrics delivered across analytics tools, teams can have confidence in data-driven decisions

- Improve data access: Non-technical users can seamlessly query data from the comfort of their own BI tool

- Embrace flexibility: With logic defined once and housed centrally in your dbt project, organizations have the flexibility to adopt new or varying downstream tools without needing to rewrite business logic
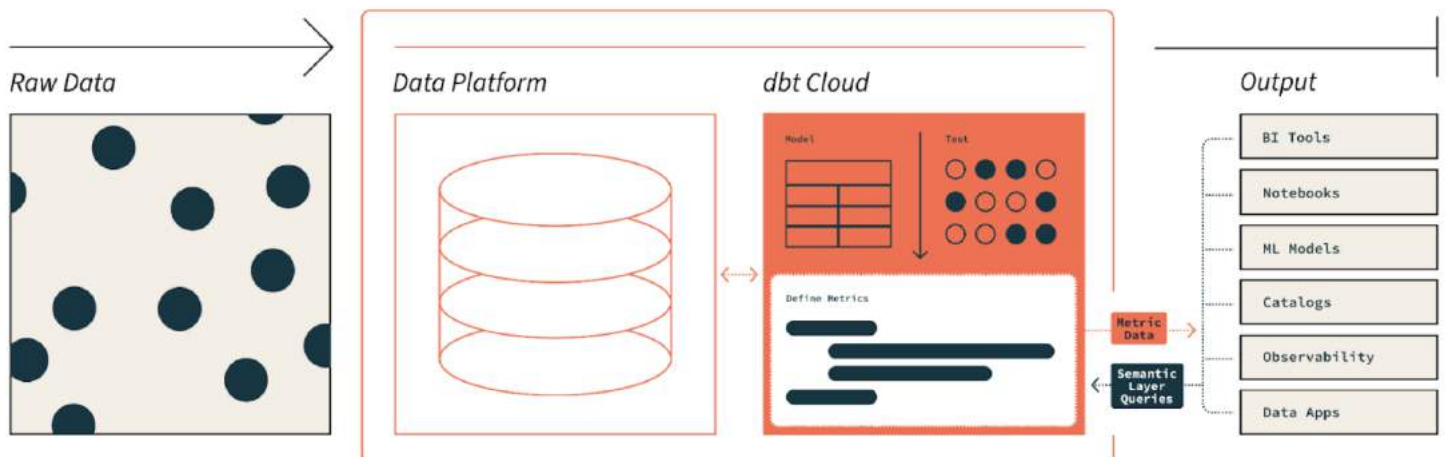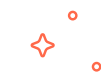
> *"dbt Mesh enables us to make data mesh a reality by offering a simple, cohesive way to integrate and manage data pipelines & products across the enterprise using a single platform."*
>
> *—Marc Johnson,*
> *Data Strategy & Architecture,*
> **Fifth Third Bank**

dbt Mesh: Centralized, secure, and scalable governance

As use cases for data grow, so too does data complexity. What may have started as a manageable dbt project with a few dozen models built and referenced by a small subset of people, can eventually spiral into an unwieldy monolith that's difficult to govern, administer, and optimize. This hinders data velocity, impairs data quality, and results in higher maintenance and platform costs.

That's why dbt Cloud now supports a data mesh paradigm—an architecture design we call dbt Mesh. Rather than relying on a single, monolithic project that becomes a bottleneck as teams try to build and deploy analytics code at scale, domain teams can use dbt Mesh to enforce an architectural design that allows various domain teams to have purview over their own dbt projects and nothing more. Teams can reference resources across projects, and maintainers of reusable data models can enforce proper governance with built-in contracts, versions, and access controls.

This puts domain teams in control of their own data pipelines (within reasonable guardrails), which means they only manage the parts of the dbt lineage that are directly relevant to their work. They can easily reference shared dbt models from other teams, bolstered by the confidence that nothing will break unexpectedly for them, and they can do the same for their downstream consumers. Meanwhile, a central data team can control the overall estate, set global standards for governance, and maintain visibility on end-to-end dbt lineage.
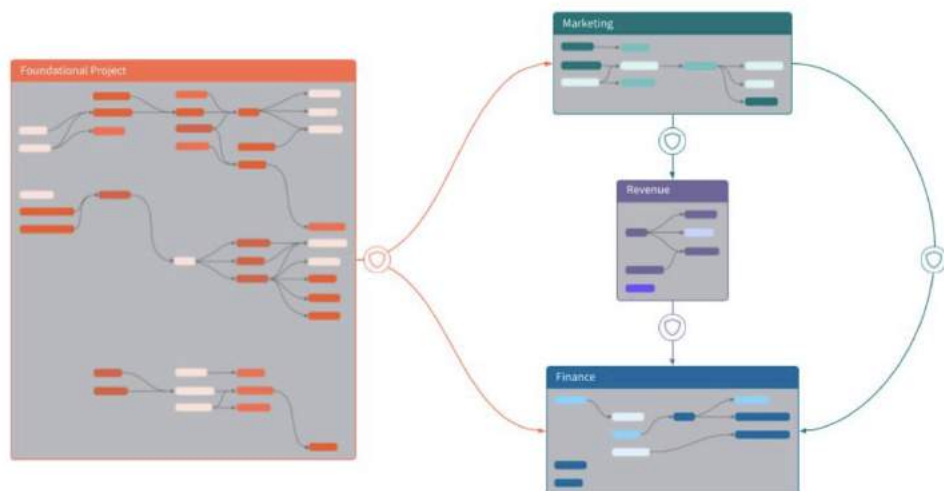
dbt Mesh helps organizations:

- Empower teams: Downstream teams can simplify their estate by building and maintaining data pipelines that are relevant to their domains. They can build and ship faster with full ownership of business logic.
- Accelerate data development: Data developers can work quickly and autonomously without being bottlenecked by a central data team. Users can discover and reuse existing analytics assets to streamline and optimize workflows.
- Standardize governance: Central data teams can customize global development standards and permissions and maintain visibility on end-to-end lineage.

dbt Explorer: Intuitively navigate, understand, and improve dbt Cloud projects

It's difficult to improve your data estate if you can't visualize it. dbt Explorer offers a birds'-eye view of your end-to-end data pipelines so you can quickly understand your dbt projects and dependencies and orient to available resources for development or analysis. dbt Explorer provides documentation enriched after each dbt Cloud run and uses the Discovery API as its robust source of truth for production metadata.

Using dbt Explorer, you can visualize your dbt lineage and zoom in to interact with specific nodes in your lineage graph. When you click on a node, you'll immediately get served relevant context like run status, description, freshness, and more. You can also use global search to find specific resources and understand their upstream and downstream dependencies.

The following features are native to dbt Explorer and help teams not only manage, but actually improve and optimize their dbt deployments:

- Model performance improvements: Get insights into historical model execution performance and identify ways to reduce costs, improve reliability, and save time.

- Recommendations: Monitor test and documentation coverage and receive specific recommendations for project code, modeling, and config optimizations.

- Column-level lineage: Explore column-level lineage across sources and models to diagnose data quality issues, perform impact analysis, and develop faster.

dbt Explorer also natively supports multi-project lineage, making it an essential tool for organizations adopting a dbt Mesh architecture. It is the hub that provides the central visibility essential for distributed collaboration at scale.
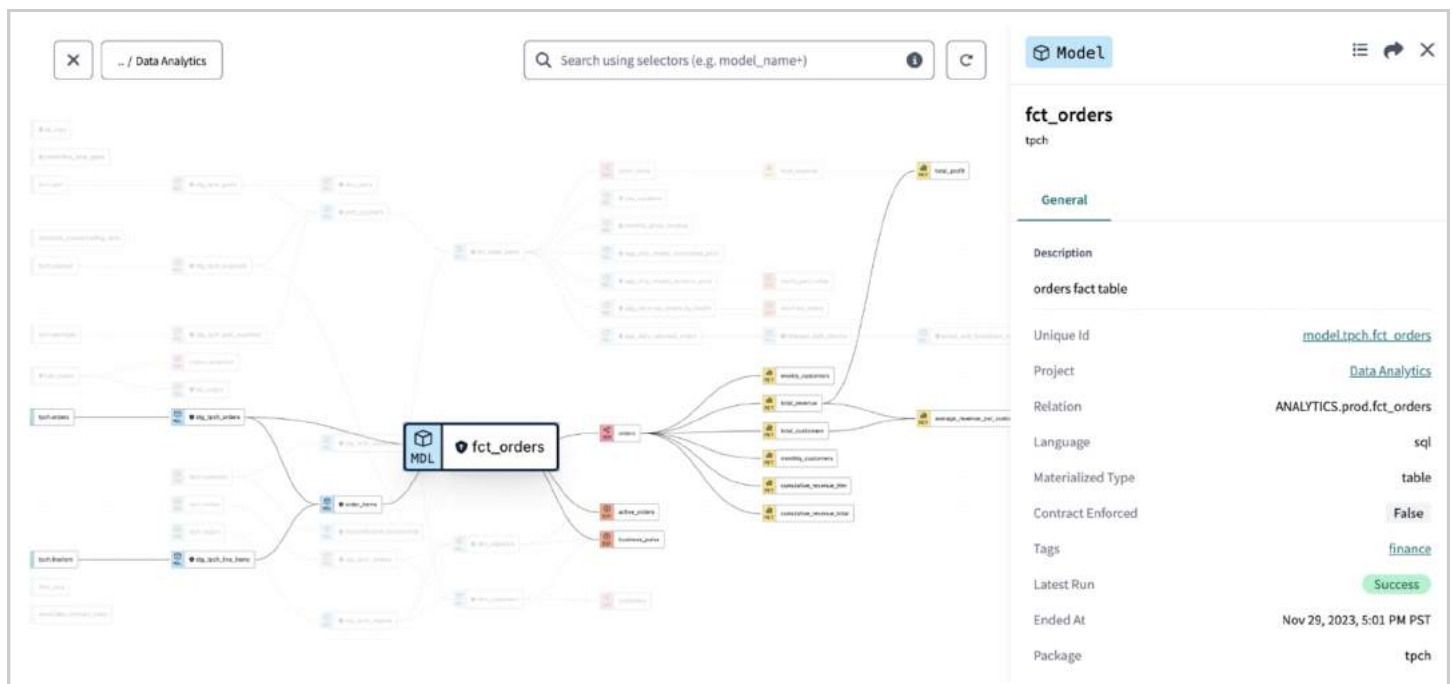
Tests and alerts: Proactively enforce data quality and debug issues fast

Building trust in data products is essential to fostering a data-driven culture. But when relying on outdated approaches, data transformations and model dependencies can be a black box. Without built-in observability, determining performance, quality, and reliability across environments is a guessing game, and there's no way to understand the impact of changes on downstream assets or teams. When pipelines inevitably break, data teams are left scrambling to find and fix the problem before trust is irrevocably compromised.

With dbt, data teams can proactively define assertions—called tests—about their data models. Adding a built-in test is as easy as adding a single line, with the word `unique`; or you can write any `select` query that suits your needs. When you run `dbt build`, downstream models will be skipped if an upstream test fails. If a test fails, dbt Cloud helps you find the SQL that failed so you can debug it. Defining tests is a great way to confirm that your code is working as expected, and helps prevent regressions when your code changes.

"dbt tests add a ton of value with the clarity and quality they enable. With one or two lines of code, I can make sure my data is quality before it even gets to the stakeholder, which is huge."

—*Jared Stout,*
Head of data management,
Sunrun
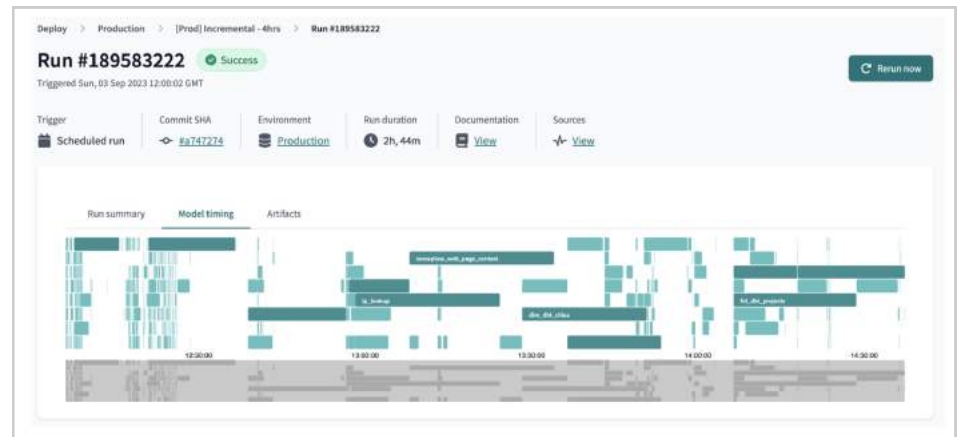
In addition to setting up tests to proactively catch issues, it's easy to monitor your production dbt jobs and alert the right people when something goes wrong:

- Run history: View run history and the model timing visualization to identify where to make improvements to jobs.

- Logs: Parse through in-progress and historical logs for your dbt runs to debug errors quickly.

- Model timing: Visualize the composition, order, and execution time taken by each model in a job run to identify bottlenecks.

- Source freshness: Quickly grok data source freshness in dbt Explorer to ensure you're meeting SLAs.

- Notifications: Configure Slack or email notifications to alert the right teams about job failures, or use outbound webhooks to power custom notifications.

- Dashboard status tiles: Embed tiles into your BI platforms so downstream stakeholders can easily understand whether or not data used in reporting has passed quality and freshness checks.

Scheduler and continuous integration:
Automate deploys confidently

Scheduling

dbt Cloud includes an in-app job scheduler to automate how and when you execute dbt jobs. This frees teams from having to build and maintain their own infrastructure, and ensures the timeliness and reliability of data transformations. The scheduler powers running dbt in staging and production environments, bringing automation, ease, and confidence to continuous integration (CI) workflows and enables observability and governance for deploying dbt at scale.

Deferral to production

dbt Cloud makes "deferral to production" as easy as flipping a switch, both in development and in deployment. dbt Cloud keeps track of the latest state of every model across environments, meaning that when a developer wants to run and test changes to a single model, she can build only that changed model. This holds even if the model references dozens of upstream dependencies. Those references will automatically "defer" to the production locations of the data, thereby shortening development feedback loops and reducing data platform consumption. This workflow is supported in both the dbt Cloud IDE and the dbt Cloud CLI.

Without deferral, it would be necessary to first materialize all of the (unchanged) upstream models into a development schema, just to validate the changes to the one model that needs to be changed. In deployment, dbt Cloud leverages deferral to offer "Slim CI" (build and test just the models that a PR has changed) and rerunning failed jobs (restarted from point of failure) at the click of a button. The result? No wasted developer time or data platform compute spent re-building models unnecessarily.

Cl checks

You can also set up automation that tests code changes by running Cl jobs before merging to production. dbt Cloud tracks the state of what's running in your production environment so that when you run a Cl job, only the modified data assets in your pull request (PR) and their downstream dependencies are built and tested in a staging schema. You can view Cl check status directly from within the PR and enable settings in your Git provider that allow only PRs with successful CI checks to be approved to merge.

Integrate Cl workflows into your dbt Cloud deployments to:

- Provide increased confidence and assurances that project changes will work as expected in production.

- Reduce the time it takes to push code changes to production.

- Make code changes in a standardized and governed way that ensure code quality without sacrificing speed.

Cloud CLI and Cloud IDE: Freedom to develop anywhere

A key principle of dbt is to empower more people to unlock data across more use cases. That means meeting data producers in their preferred development environment. dbt Cloud enables development in either your browser using an integrated development environment (IDE) or in a dbt Cloud-powered command line interface (CLI).

Cloud IDE
The Cloud IDE is a single web-based interface for building, testing, running, and version-controlling dbt projects. It compiles dbt code into SQL and executes it directly on your database. The Cloud IDE includes editing features—including syntax highlighting, auto-completion, code formatting and linting—and version control that create a powerful environment for efficient SQL coding.

Cloud CLI
dbt Cloud also supports a dedicated Cloud CLI, giving more advanced practitioners the flexibility to contribute via any terminal or IDE software of their choosing (such as VS Code, Sublime Text, or Vim).

The dbt Cloud CLI allows you to run dbt commands against your dbt Cloud development environment from your local command line. You can run dbt commands against dbt Cloud's infrastructure and benefit from:

- Secure credential storage in the dbt Cloud platform
- Automatic deferral of build artifacts to your project's production environment
- Speedier, lower-cost builds
- Support for dbt Mesh and cross-project refs

This gives data practitioners the best of both worlds: they can eliminate many of the hassles of local development—such as the complexities of manual configuration, authentication, and version upgrades—while enjoying the benefits of a hosted solution, all from the comfort of their preferred development environment.
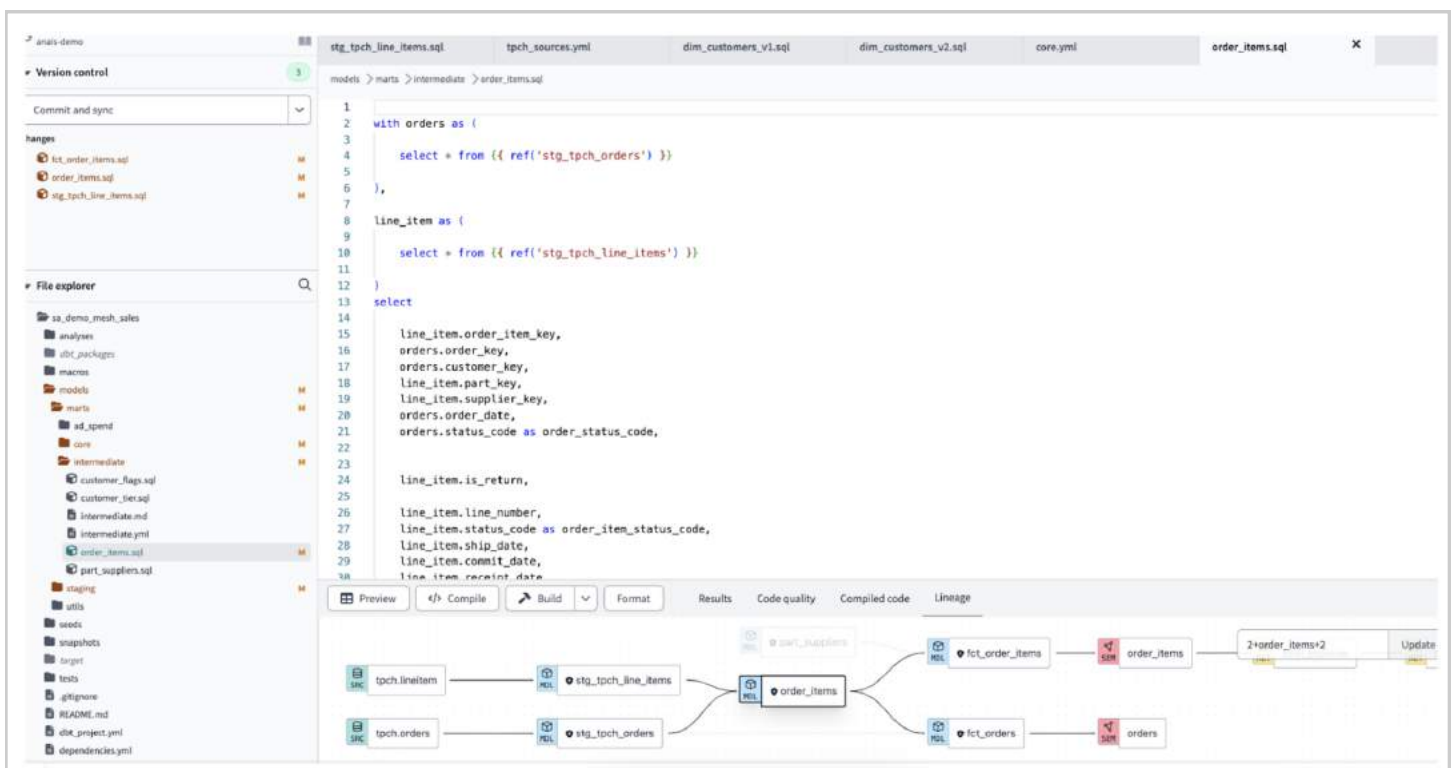
dbt Cloud as the control plane

dbt Cloud helps modern organizations manage data complexity at scale so they can turn insights into competitive advantage. All of the features described above come together to create a platform that serves as a control plane for your dbt assets. With stateful metadata on each model build and dbt run, intuitive Uls to explore and improve your dbt projects, enterprise-grade governance and security, and native integrations across the modern data stack, dbt Cloud helps teams control the complexity of their data estate so they can deliver trusted data products, faster.

Using dbt Cloud:

- **Data teams** have a standardized approach to develop, test, document, and deploy data at scale, with software development best practices like version control, Cl, and modularity built into every part of the data development lifecycle

- This gives **data consumers** a high degree of confidence in the accuracy, freshness, and relevance of the data they use to make decisions

- With data pipelines humming, **entire organizations** are able to embrace a culture of data-driven decision making and experience improved data ROl and enhanced cross-org collaboration along the way.

🔒

A note on security

dbt Labs is dedicated to upholding industry standards for Cloud security and GDPR compliance.

Our compliance certifications include the following
- SOC2 Type Il (including HIPAA)
- ISO27001:2013
- GDPR

Learn more at getdbt.com/security

Learn more about dbt Cloud

Talk to an expert, book a demo, or set up a free trial at **getdbt.com**.

Get started