gable

How Glassdoor Built Data Trust at Petabyte Scale with Gable

Table Of Content

Company Overview

Glassdoor operates one of the largest job and company review platforms, processing petabyte-scale data across millions of job postings, user reviews, salaries, and analytics queries.

As its data infrastructure evolves, Glassdoor continues to refine how it ensures data quality, governance, and reliability while maintaining the speed and agility needed for innovation.

Chapter 1: The Challenge: Managing Data Quality at Scale

Glassdoor's data team faces ongoing challenges in maintaining data integrity across a growing, complex ecosystem:

Standardizing Processes for Managing Data Changes

- Data-producing application code is frequently updated, leading to unintended changes in schemas, transformations, and field definitions.
- Existing monitoring tools historically focused on database schema drift, missing critical changes in event structures, serialization formats, and null handling.
- Example: A timestamp formatting update previously misaligned event data across time zones, creating discrepancies in reporting.

Ensuring Visibility into How Data Is Used Across Teams

- Data producers and consumers work across multiple teams, making it essential to understand how schema or logic changes impact downstream systems.
- Traditional tools primarily detect failures after they occur, whereas Glassdoor's team needs early warnings before risky changes are deployed.
- Example: An update to event attributes affected feature engineering pipelines, leading to degraded model performance. A proactive approach was needed to flag such risks before deployment.

Scaling Data Governance Without Slowing Engineering Down

- Relying on manual anomaly detection and informal communication between teams was no longer sustainable.
- Engineers needed an automated way to validate and enforce data expectations before deployment.
- Example: A field used in business-critical dashboards was removed without proper coordination, leading to last-minute rollbacks and urgent fixes.

Chapter 2: How Glassdoor Uses Gable

Glassdoor implemented Gable's shift-left approach to data governance, making data quality enforcement part of the development process rather than a reactive fix.

Detecting Data-Impacting Changes in Application Code

- Gable scans source code and transformation logic to detect modifications in data structures, event schemas, and serialization formats.
- Engineers receive automated validation in CI/CD pipelines, flagging changes that could cause downstream failures.
- Example: A restructuring of API response fields was detected before deployment, allowing time to coordinate changes with consuming teams.

Enforcing Data Contracts to Align Data Producers & Consumers

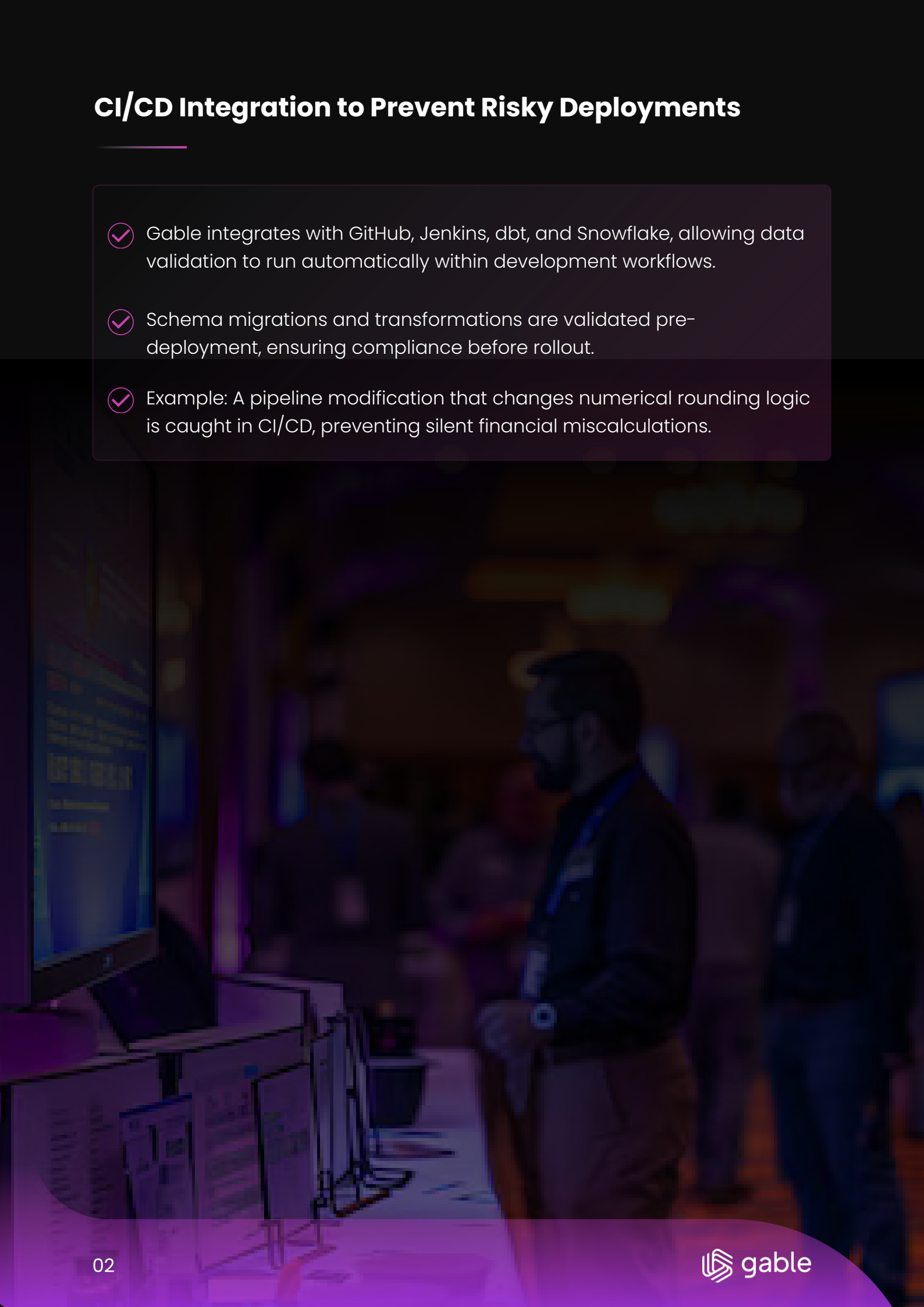
- Glassdoor uses data contracts to define schema structures, data types, and required attributes, ensuring consistency across teams.
- These contracts are validated pre-deployment, ensuring unexpected changes are reviewed before they impact production.
- Example: A change modifying required fields in a key dataset is flagged in CI/CD, allowing teams to make necessary adjustments before the update goes live.

Automating Impact Analysis & Team Collaboration

- Gable maps dependencies across data pipelines, allowing teams to understand the downstream effects of proposed changes before they are deployed.
- Engineers, analysts, and data scientists receive early notifications about schema modifications, enabling proactive adjustments.
- Example: A planned schema update is flagged for potential impact on dashboards, giving the analytics team time to adjust queries in advance.

CI/CD Integration to Prevent Risky Deployments

- Gable integrates with GitHub, Jenkins, dbt, and Snowflake, allowing data validation to run automatically within development workflows.
- Schema migrations and transformations are validated pre-deployment, ensuring compliance before rollout.
- Example: A pipeline modification that changes numerical rounding logic is caught in CI/CD, preventing silent financial miscalculations.



Chapter 3: What You Can Achieve with Gable

Organizations that adopt shift-left data governance and automated contract enforcement can expect:

Fewer Data-Related Incidents
Catching schema drift and transformation errors before deployment reduces production issues.

Faster Time-to-Resolution for Data Breakages
Automated impact analysis makes it significantly easier to diagnose and resolve data issues quickly.

Greater Trust in Business Analytics & ML Models
Validated data ensures that dashboards and reports are reliable, improving confidence in decision-making.

Reduced Engineering Time Spent on Data Quality Issues
Automated enforcement eliminates the need for manual schema coordination and debugging.

Key Takeaways for Data Teams

- Shift-left governance prevents costly downstream data failures
- Data contracts ensure consistency, keeping data producers and consumers aligned
- CI/CD integration embeds data quality enforcement into existing development workflows

What Glassdoor Engineering Says

"Gable has helped us bring trust to petabyte-scale data by ensuring governance at the source. We no longer spend engineering cycles fixing broken data downstream."

Learn More: Download "Gable: The Easy Way to Do for Data What DevOps Did for Code"