acceldata

Increase Your Snowflake ROI with Data Quality, Resource Efficiency, and Spend Forecasting

Why Data Quality is So Important

- Data is the lifeblood of the modern, data-driven enterprise.
- High-quality data — data that is accurate, complete, consistent and timely — is paramount.
- Yet, even the most advanced businesses can suffer from data that is inaccurate, has duplicate or missing records, inconsistent structure/schema, opaque lineage, and more.
- Poor-quality data creates a whole host of problems.

Bad data in, garbage info out. You'll get wrong numbers in your executive reports, wonky charts in your dashboards, or malfunctioning analytics-fed operations. And you won't find out until that all-caps email or Slack message arrives at 11 pm, demanding to know why a mission-critical tool has gone off the rails or the quarterly figures in the boss's presentation don't add up.

Poor data quality is the number one obstacle for organizations to generate actionable business insights, according to 42 percent of executives surveyed by the Harvard Business Review.

Constant firefighting, lost agility. For organizations that choose to not address data quality head-on, that decision almost always backfires. Their data teams end up getting pulled in every direction fixing broken dashboards, applications, and pipelines.

Invariably they end up spending more time and money fixing problems than if they had prioritized data quality. It also leaves their data engineers overworked, fatigued from constant alerts, and demoralized by their inability to focus on value-creating projects.

Disintegration of your data-driven culture. When workers don't trust data, they refuse to reuse existing data pipelines and applications and demand whole new ones instead. Building these is time-consuming for your data engineers, and drives up your storage costs. It also leads to a proliferation of data silos and pools of dark data that exacerbate the data quality problem.

Even worse than the lack of efficient data reuse is the effect of untrusted data on decision making. Business leaders start ignoring solid quantitatively-backed insights in favor of instincts and anecdotes. The data-driven culture your business has been carefully nurturing begins to fall apart.

Bad for the bottom line. Data quality has a huge financial impact. According to a University of Texas study, Fortune 1000 companies that improve their data quality and usability by ten percent reap an extra $2 billion in annual revenue on average. For companies that fail to improve their data quality, this is a huge missed opportunity.

“As organizations accelerate their digital [transformation] efforts, poor data quality is a major contributor to a crisis in information trust and business value, negatively impacting financial performance,” says Ted Friedman, Gartner analyst.

What Causes Data Quality Problems?

- **Data quality can degrade for many reasons**.
- **Schema changes can break processes that feed analytical applications and dashboards.**
- **API calls can fail, interrupting the flow of data**.
- **Manual, one-off data retrievals can create errors and hidden pools of duplicate data.**

Data can be duplicated for good reasons, such as to improve query performance. Without strong data governance, though, this can eventually lead to a confusing overabundance of expensive data silos.

Migrations from on-premises infrastructures to the cloud can also create a new set of data quality and management challenges. The lack of a unified view of the entire data lifecycle can also create inconsistencies that drag down your data quality.

Finally, there’s one problem that virtually all enterprises face today: scale. The amount of data that enterprises are collecting and storing is growing at an incredible rate — a whopping 63 percent growth per month, according to an IDG survey. The number of data sources is also huge: 400 for the average company, 1,000+ sources for 20 percent of firms.

There is also a tidal wave of data tools in every layer of the modern data stack. Companies have no shortage of choices, from event and CDC streaming platforms, ETL/ELT tools, reverse ETL tools that push insights to business apps, data API and visualization tools, real-time analytics databases, and more.

Many of these data tools are point solutions, early entries in the market. Though each has their merits, trying to cobble a stack from these unintegrated tools helps create fragmented, unreliable, and broken data environments.

Why Legacy Data Quality Strategies Fail

Companies have tried to solve data quality for years, typically by manually creating **data quality policies and rules**, often managed and enforced by master data management (MDM) or data governance software.

MDM vendors like Informatica, Oracle, SAP, SAS and others have been around for many decades. Their solutions were born and matured long before the cloud or big data existed.

Unsurprisingly, these antiquated software and strategies can't scale for today's much larger data volumes and ever-changing data structures. Scripts and rules must be created and updated by human data engineers one by one. And when alerts are sounded, your data engineers will also need to manually check anomalies, debug data errors, and clean datasets. That's time-consuming and exhausting.

A good example of the failings of the legacy approach to data quality are **manual ETL validation scripts**. These have long been used by data engineers to clean and validate recently-ingested data. Applied to data-at-rest, ETL validation scripts are easy to create and flexible, as they can be written in most programming languages and support any technology, data system or process.

However, manual ETL validation scripts are often poorly suited for the volume, velocity and dynamic nature of today's enterprise data environments. Take streaming data. Event and messaging streams can be too high volume, too dynamic (with constantly-changing schemas) and too real-time for ETL validation scripts to work. These scripts can only process data in batches and must be manually edited with every change to the data structure.

This results in significant validation latency. And this delay is unacceptable for companies undergoing digital transformation, as it rules out use cases such as real-time customer personalization, data-driven logistics, fraud detection and other internal operations, live user leaderboards, etc.

Beyond real-time data, manual ETL validation scripts have other problems. Any change to your data architecture, systems, schemas or processes will force you to update an existing script or create a new one. Fail to keep them updated and you can transform and map data wrong and inadvertently create data quality problems.

To prevent this, organizations need to constantly check if their ETL validation scripts have become outdated, and then have their data engineers spend hours writing and rewriting repetitive ETL validation scripts. This requires significant ongoing engineering time and effort. And it pulls your data engineers away from more-valuable activities such as building new solutions for the business.

Also, if and when your data engineers leave the organization, they take specific knowledge around your ETL validation scripts with them. This creates a steep learning curve for every replacement data engineer.

To handle today's fast-growing, constantly-changing data environments, data ops teams need a modern platform that leverages machine learning to automate data quality monitoring at whatever scale is required.

Snowflake: "It Just Works" Scale and Agility

Snowflake is one of the most popular cloud data warehouses today. In just one decade, the company has grown to nearly 6,000 enterprise customers and $1.2 billion in annual revenue, more than double its prior year.

Like other cloud data warehouses, including Databricks, Amazon RedShift, and Google BigQuery, Snowflake boasts an attractive combination of low start up costs, constant innovation, and "it just works" manageability. And judging by its breakneck growth and enthusiastic customers, 100 percent of whom recommend Snowflake, Snowflake delivers on these features better than its rivals, especially in these two areas:

1. **High availability with near-zero administration.** Guaranteed uptime with minimal operational hassle, so you don't need a big team of DBAs, data engineers, etc.

2. **Infinite infrastructure deployed instantly.** While most public cloud databases separate compute from storage on an architectural level to enable them to grow or shrink independently, Snowflake is particularly elastic, supporting instant, automatic scale-up and scale-down that smoothly handles planned or unplanned bursts of ingested data or analytical jobs. No need to order hardware, configure clusters, or get your data engineers' approval.

Law of Unintended Consequences

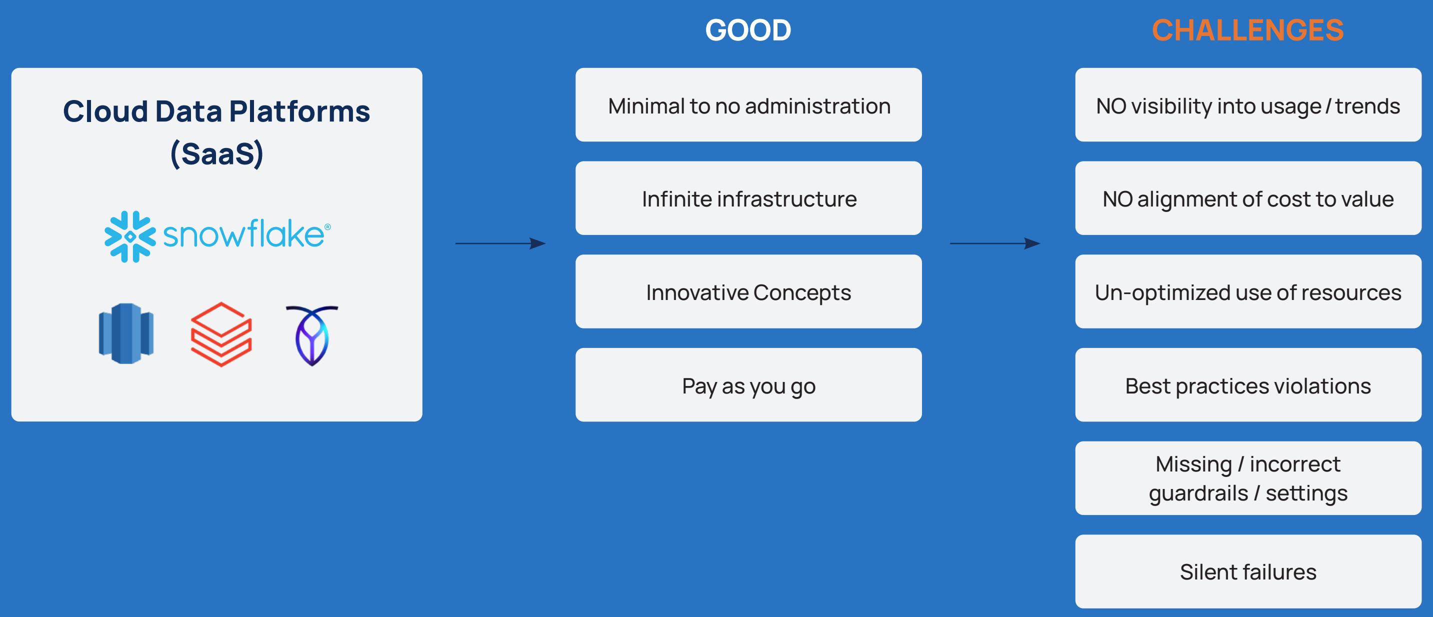
No company would say no to such easy agility. At the same time, operating and scaling Snowflake is such a dream that it lulls many companies into ignoring some important best practices.

Others try to cut corners by importing legacy data quality processes from their on-premises databases and data warehouses, which are often unsuitable and fail.

Without a strong data ops team following a modern set of best practices, this can lead to chaos: uncontrolled growth of data repositories and data pipelines and rampant duplication of data without proper lineage and tracking.

Data errors creep in and silently compound themselves, unchecked. Data pipelines fail and analytical dashboards produce false results. Without strong observability, these issues may continue for weeks or months before users notice and complain. By that point, there could be many terabytes or petabytes of data that need to be rescanned for data quality issues and backfilled throughout the organization, depending on the number of dependencies and how complex your data pipelines are. Your Snowflake Data Cloud has turned into a data swamp. And Snowflake's low ops has officially entered the realm of being *Too Much of a Good Thing.*

Cloud Data Platforms: The good and the challenges



Cloud Data Platforms (SaaS)

GOOD

- Minimal to no administration
- Infinite infrastructure
- Innovative Concepts
- Pay as you go

CHALLENGES

- NO visibility into usage / trends
- NO alignment of cost to value
- Un-optimized use of resources
- Best practices violations
- Missing / incorrect guardrails / settings
- Silent failures

But surely there are other Snowflake users like you that prioritize monitoring and preserving data quality? Yes, but it turns out that this is difficult to do in Snowflake.

Snowflake's primary management interface is a web dashboard called SnowSight. With SnowSight, users can monitor query performance and copy history, create and manage databases and warehouses, and little else. Visibility and control is limited compared to other database management consoles. Users cannot even configure Snowsight to push out real-time alerts or status reports. For data engineers used to continuous visibility and control over their data, this can be jarring.

Most Snowflake users today use ETL scripts to validate and cleanse incoming data as it is processed and stored by Snowflake. If they don't use SnowSight, they tend to use SQL-based reporting applications, such as Tableau, Looker, and Microsoft Power BI, which can ingest and display the operational data that Snowflake exposes. However, the batch-based design of these visualization applications makes them best-suited for executive reporting, not real-time observability or day-to-day management.

Like SnowSight, they lack the alerts, predictive capabilities, and fine-tuned control that data engineers need in order to monitor, fix and prevent data issues in general. And specifically in regards to data quality, Tableau and its ilk lack out-of-the-box templates and scorecards that would enable them to ingest, analyze and display the data quality metrics generated by your ETL scripts. **Meaning these popular reporting tools are wholly unsuited as Snowflake operational dashboards, especially for monitoring and optimizing your data quality.**

Here's a concrete example of how Snowflake's well-meaning attempt at low ops can inadvertently create a data quality issue. Snowflake's data warehouse does not require users to manage partitions or indexes. Instead, Snowflake automatically divides large tables into micro-partitions and calculates statistics about the value ranges contained in each column of data. These statistics help determine which subsets of your data are needed to run queries, hastening query speeds.

The problem? Data migrated from a traditional partition-and-index database into Snowflake will need to be transformed as it is loaded, creating possible data and schema errors. Even a minor issue, such as the case sensitivity of Snowflake SQL code, can lead to broken applications and data pipelines. And these are unlikely to be flagged by Snowflake, or noticed by your data engineers.

Reliant on Snowflake's metrics, traditional MDM and data governance tools are not equipped to catch SQL syntax errors and other more subtle data quality issues. Meanwhile, data profiling tools such as Snowflake Data Profiler can only paint a high-level overview of your data, not perform actual checks that ferret out data quality problems. For reporting tools such as Looker, they can only spot-check inconsistencies and test for data quality at single points in time, such as when data is ingested. Without continuous data quality validation and testing, they won't notice when data errors crop up later.

Solving Data Quality with Data Observability

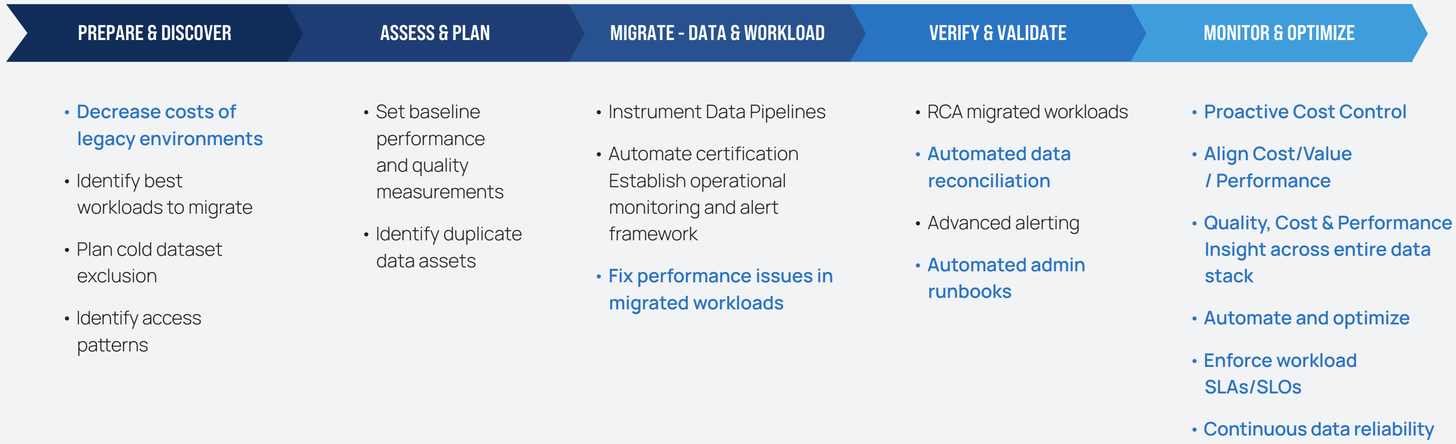
Snowflake users are turning to modern data observability applications to help them automate the continuous data validation and testing required to create organization-wide trust in their data.

- However, not all data observability applications are created equal.
- Some focus on a single dimension — data quality — but lack any insight into data performance, aka compute observability, or how to optimize the price-performance of your data.
- Others rely too much on Snowflake's provided metrics and metadata, limiting the scope of their insights and their predictive power.
- Still others only offer macro level views in lieu of the ability to slice and dice data that can more accurately and cost-effectively determine root causes of data errors.
- Finally, some are actually Application Performance Management (APM) solutions — think DataDog or New Relic — in disguise, trying to pass off their application-level observability as true data observability (ummm, not so).

What companies need is a multi-dimensional cloud data observability platform that draws upon Snowflake's metrics as well as continually gathers its own statistics and metadata around Snowflake data quality.

It then combines this data together to generate its own original data quality profiles and insights. It uses these more sophisticated baselines to automatically profile and validate your data as it is ingested into Snowflake. And it continues to validate and test your data continuously, to account for how your data evolves and your business needs change.

Data Observability for Data Migration



PREPARE & DISCOVER

- **Decrease costs of legacy environments**
- Identify best workloads to migrate
- Plan cold dataset exclusion
- Identify access patterns

ASSESS & PLAN

- Set baseline performance and quality measurements
- Identify duplicate data assets

MIGRATE - DATA & WORKLOAD

- Instrument Data Pipelines
- Automate certification Establish operational monitoring and alert framework
- **Fix performance issues in migrated workloads**

VERIFY & VALIDATE

- RCA migrated workloads
- **Automated data reconciliation**
- Advanced alerting
- **Automated admin runbooks**

MONITOR & OPTIMIZE

- **Proactive Cost Control**
- **Align Cost/Value / Performance**
- **Quality, Cost & Performance Insight across entire data stack**
- **Automate and optimize**
- **Enforce workload SLAs/SLOs**
- **Continuous data reliability**

Acceldata provides such a multi-dimensional data observability platform that can make solving for data quality a realistic target, not an infeasible pie-in-the-sky goal.

On the following pages, we list the seven ways this is achieved:

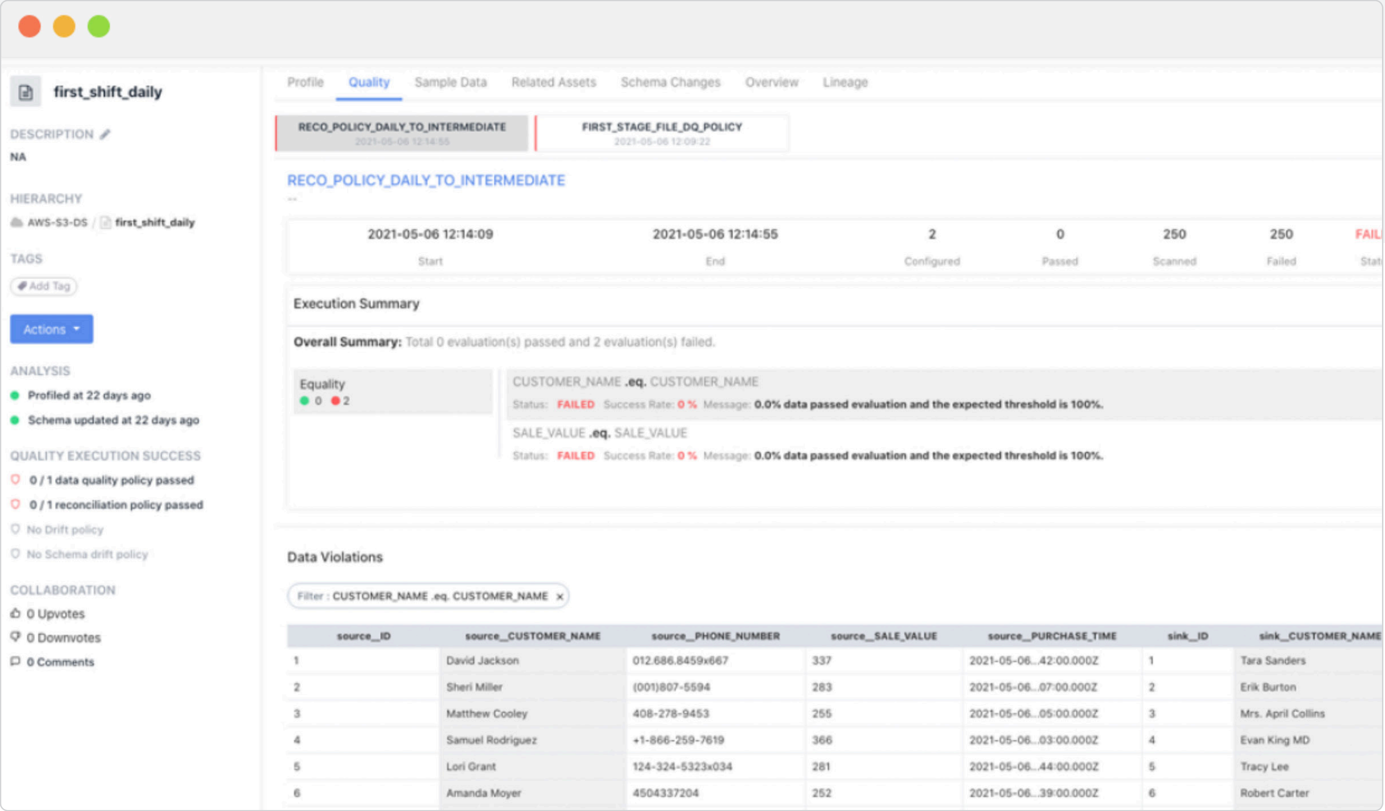
1. Dramatically-Eased Data Migration

At each phase of your migration into Snowflake, Acceldata helps automate steps to help you maximize your data quality.

Proof of Concept: the Acceldata Data Observability Cloud helps you identify essential workloads to migrate, what unused datasets can be excluded, and what data pipelines must be rebuilt in Snowflake.

Preparation: Acceldata automatically creates a data catalog so you can identify duplicate data assets that you can exclude from migration. It also configures your Snowflake account and data layout using best practices, so your Snowflake Data Cloud is secure, high-performance, and cost-efficient.

Ingestion: Acceldata provides deep visibility into the data ingestion process, whether you use Snowpipe, COPY, or other route. Acceldata also validates the data by comparing the source and target datasets. It also does Root Cause Analysis (RCA) for migrated workloads that are not working as expected.



first_shift_daily

DESCRIPTION
NA

HIERARCHY
AWS-S3-DS / first_shift_daily

TAGS

ANALYSIS
● Profiled at 22 days ago
● Schema updated at 22 days ago

QUALITY EXECUTION SUCCESS
0 / 1 data quality policy passed
0 / 1 reconciliation policy passed
No Drift policy
No Schema drift policy

COLLABORATION
0 Upvotes
0 Downvotes
0 Comments

RECO_POLICY_DAILY_TO_INTERMEDIATE 2021-05-06 12:14:55
FIRST_STAGE_FILE_DQ_POLICY 2021-05-06 12:09:22

RECO_POLICY_DAILY_TO_INTERMEDIATE
--

Start	End	Configured	Passed	Scanned	Failed	Stat
2021-05-06 12:14:09	2021-05-06 12:14:55	2	0	250	250	FAIL

Execution Summary

Overall Summary: Total 0 evaluation(s) passed and 2 evaluation(s) failed.

Equality
● 0 ● 2

CUSTOMER_NAME **.eq.** CUSTOMER_NAME
Status: FAILED Success Rate: 0 % Message: **0.0% data passed evaluation and the expected threshold is 100%.**

SALE_VALUE **.eq.** SALE_VALUE
Status: FAILED Success Rate: 0 % Message: **0.0% data passed evaluation and the expected threshold is 100%.**

Data Violations

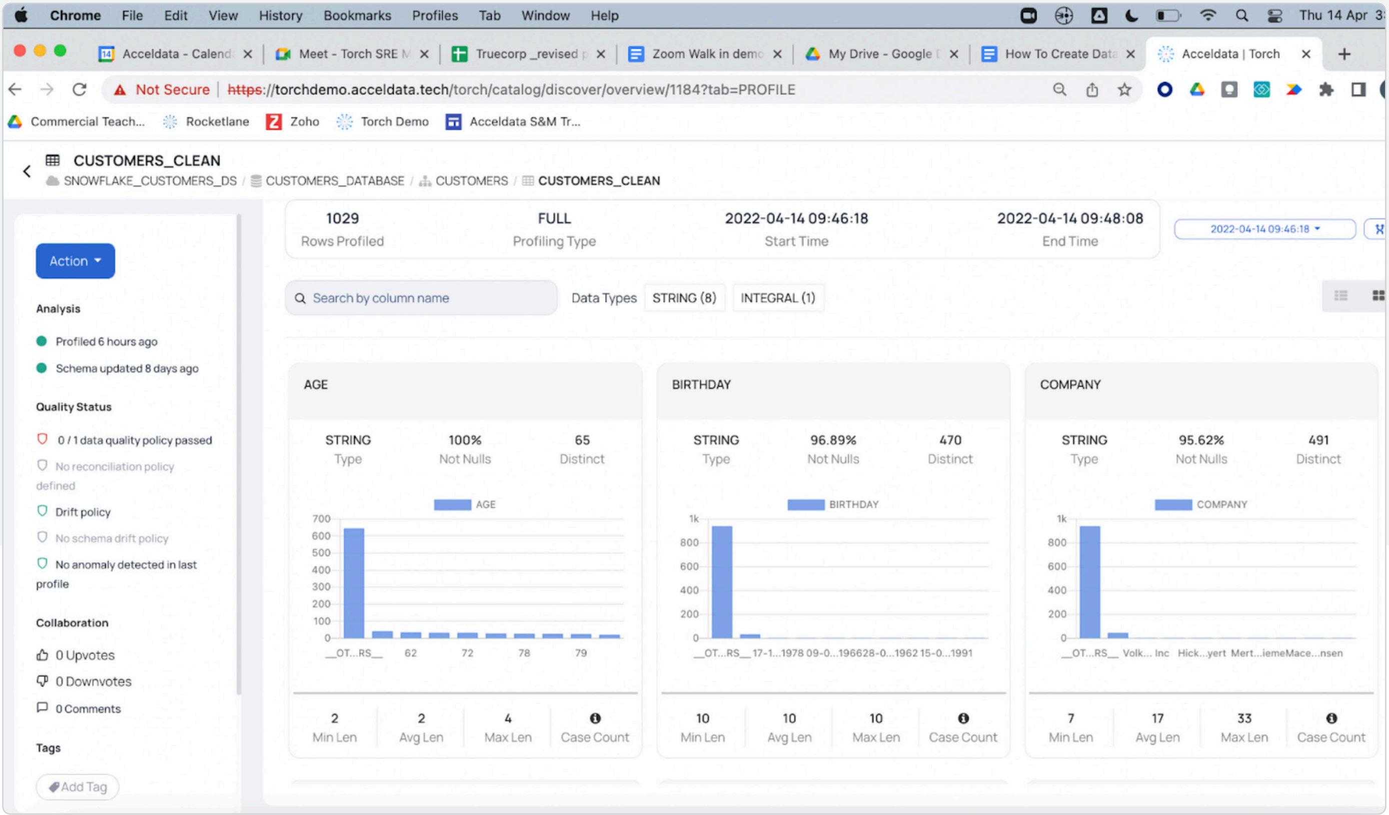
Filter : CUSTOMER_NAME .eq. CUSTOMER_NAME ×

source__ID	source__CUSTOMER_NAME	source__PHONE_NUMBER	source__SALE_VALUE	source__PURCHASE_TIME	sink__ID	sink__CUSTOMER_NAME
1	David Jackson	012.686.8459x667	337	2021-05-06...42:00.000Z	1	Tara Sanders
2	Sheri Miller	(001)807-5594	283	2021-05-06...07:00.000Z	2	Erik Burton
3	Matthew Cooley	408-278-9453	255	2021-05-06...05:00.000Z	3	Mrs. April Collins
4	Samuel Rodriguez	+1-866-259-7619	366	2021-05-06...03:00.000Z	4	Evan King MD
5	Lori Grant	124-324-5323x034	281	2021-05-06...44:00.000Z	5	Tracy Lee
6	Amanda Moyer	4504337204	252	2021-05-06...39:00.000Z	6	Robert Carter

2. Automated Profiling of your Snowflake Data Cloud

Acceldata Torch, the data reliability layer of our platform, also automatically discovers all your datasets and creates a profile of all your data, including their structure, metadata, and relationships, including dependencies and lineages.

This is key. Without this context, it's near-impossible to create data quality rules. The intuitive and interactive charts and graphs in Acceldata Torch provide much-needed insights into your data quality.



CUSTOMERS_CLEAN

SNOWFLAKE_CUSTOMERS_DS / CUSTOMERS_DATABASE / CUSTOMERS / CUSTOMERS_CLEAN

1029	FULL	2022-04-14 09:46:18	2022-04-14 09:48:08
Rows Profiled	Profiling Type	Start Time	End Time

2022-04-14 09:46:18

Data Types STRING (8) INTEGRAL (1)

Analysis

- Profiled 6 hours ago
- Schema updated 8 days ago

Quality Status

- 0 / 1 data quality policy passed
- No reconciliation policy defined
- Drift policy
- No schema drift policy
- No anomaly detected in last profile

Collaboration

- 0 Upvotes
- 0 Downvotes
- 0 Comments

Tags

AGE

STRING	100%	65
Type	Not Nulls	Distinct

2	2	4	ℹ
Min Len	Avg Len	Max Len	Case Count

BIRTHDAY

STRING	96.89%	470
Type	Not Nulls	Distinct

10	10	10	ℹ
Min Len	Avg Len	Max Len	Case Count

COMPANY

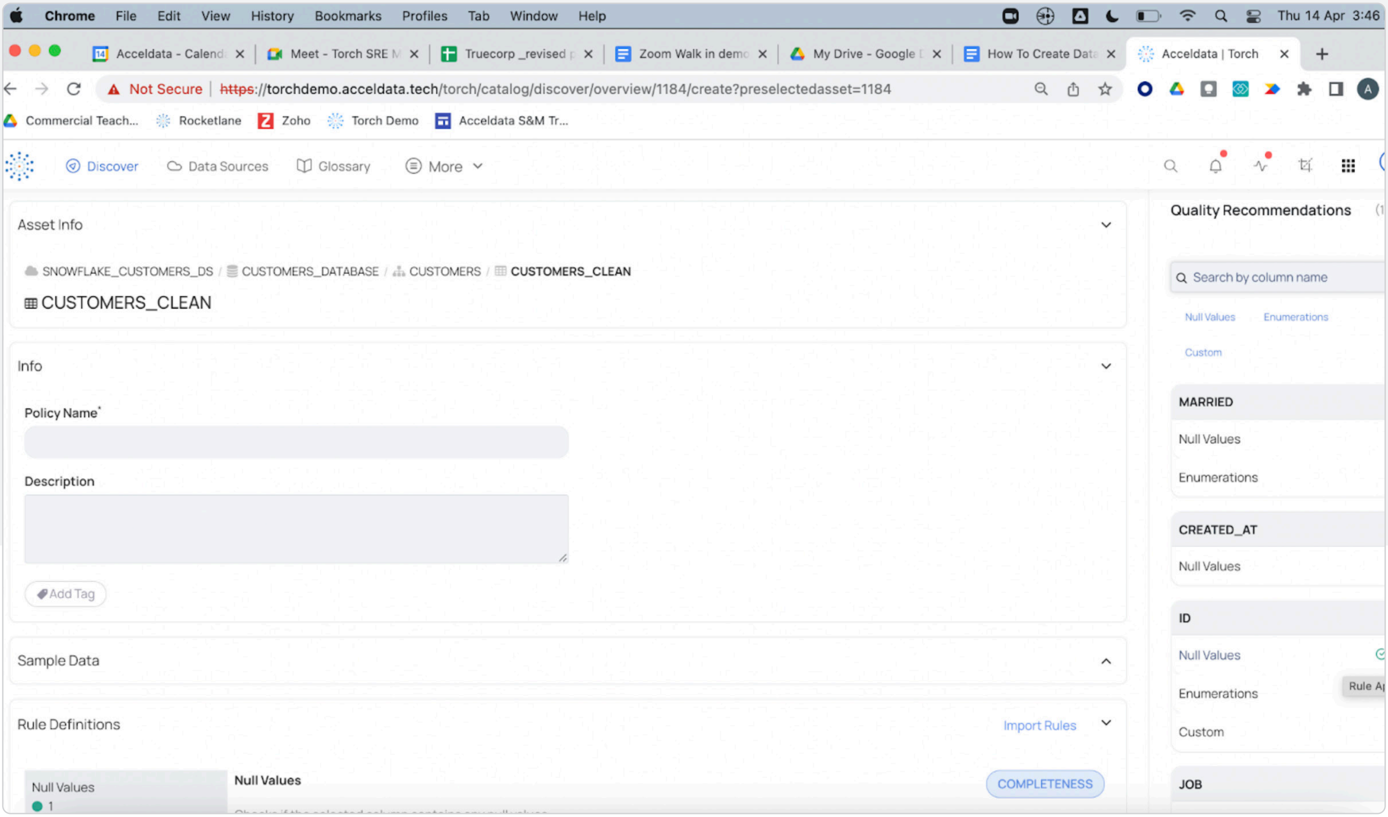
STRING	95.62%	491
Type	Not Nulls	Distinct

7	17	33	ℹ
Min Len	Avg Len	Max Len	Case Count

3. ML-Powered Data Quality Recommendations

Torch goes further. After profiling your data, Acceldata Torch starts offering ML-powered recommendations to streamline the creation of your data quality policies and rules.

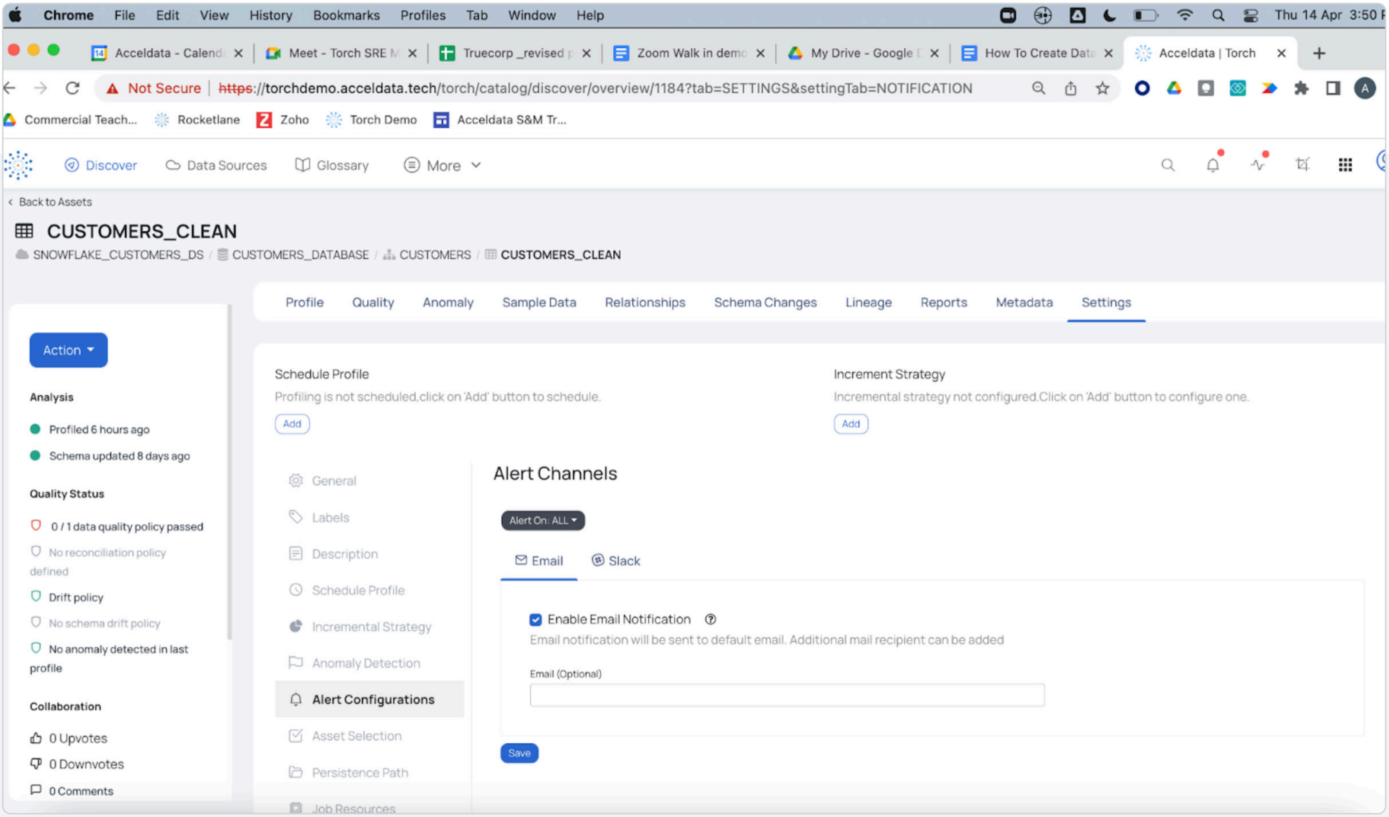
In the following example, Acceldata Torch recognizes that the data in the column should be binary ("yes" or "no") and free from null values. Just click on the recommended rule to add it to your data quality policy.



Null values are just one example. Acceldata Torch can make many other data quality recommendations, including:

- Enumeration checks
- Duplicate checks
- Uniqueness
- Pattern matching
- Range validation
- Schema checks
- Custom rules (formulas)

What used to take hours of painstaking effort can now be finished in mere minutes with a few clicks. Configuring the schedule for your data quality rules to run is also a snap. So is viewing, editing, and deleting your data quality policies. And when rules are executed, your data ops team can receive email or Slack notifications so you can stay informed.



CUSTOMERS_CLEAN

SNOWFLAKE_CUSTOMERS_DS / CUSTOMERS_DATABASE / CUSTOMERS / CUSTOMERS_CLEAN

Analysis

- Profiled 6 hours ago
- Schema updated 8 days ago

Quality Status

- 0 / 1 data quality policy passed
- No reconciliation policy defined
- Drift policy
- No schema drift policy
- No anomaly detected in last profile

Collaboration

- 0 Upvotes
- 0 Downvotes
- 0 Comments

Schedule Profile

Profiling is not scheduled,click on 'Add' button to schedule.

Add

Increment Strategy

Incremental strategy not configured.Click on 'Add' button to configure one.

Add

Alert Channels

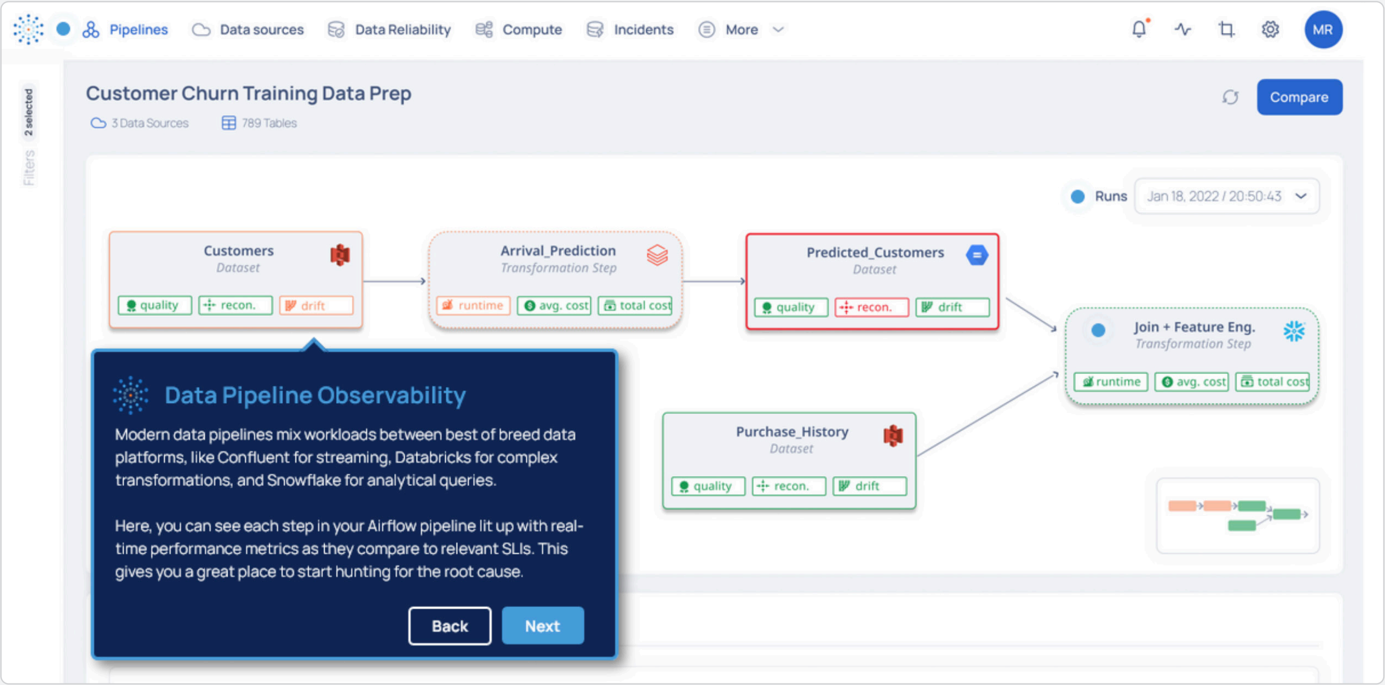
☑ Enable Email Notification

Email notification will be sent to default email. Additional mail recipient can be added

Email (Optional)

4. Ongoing Data Quality Monitoring

Setting up baselines and initial rules is useless without ongoing data quality monitoring. Acceldata provides a single unified view of your entire data pipeline through its entire lifecycle. Torch will continuously monitor and measure your Snowflake Data Cloud for data accuracy, completeness, validity, uniqueness, timeliness, schema/model drift, and other quality characteristics. This helps you maintain reliability even as data is transformed multiple times by different technologies.



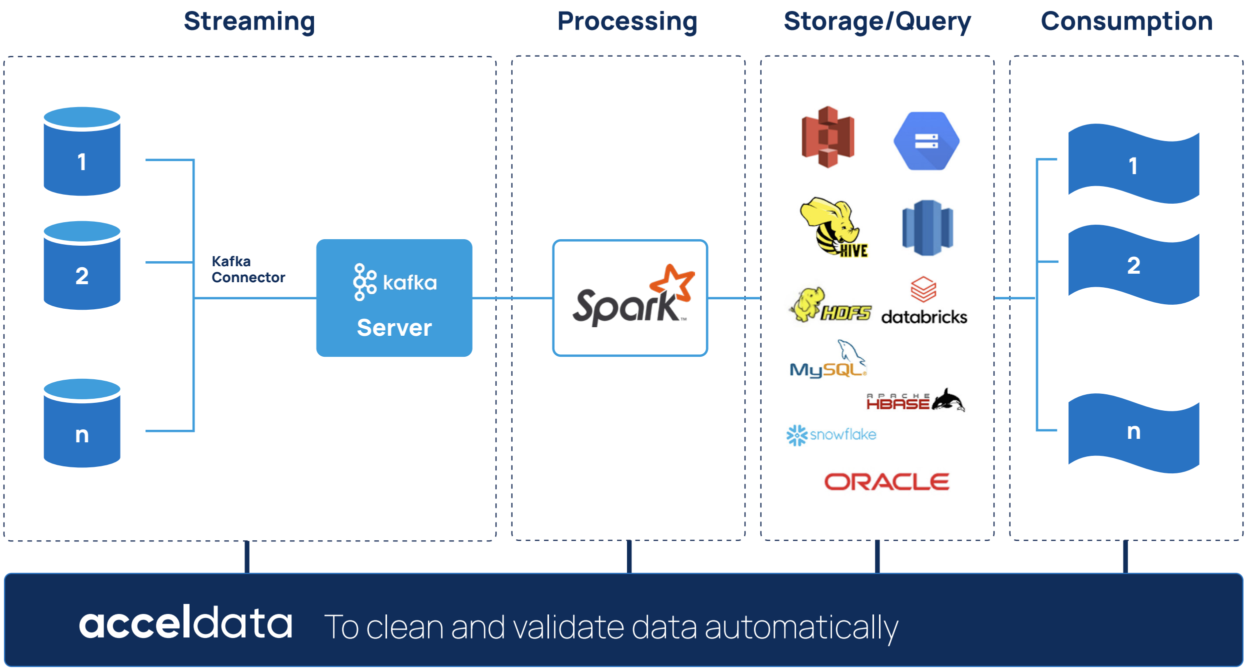
Data Pipeline Observability

Modern data pipelines mix workloads between best of breed data platforms, like Confluent for streaming, Databricks for complex transformations, and Snowflake for analytical queries.

Here, you can see each step in your Airflow pipeline lit up with real-time performance metrics as they compare to relevant SLIs. This gives you a great place to start hunting for the root cause.

Acceldata Torch also allows data teams to easily detect unexpected changes to data structure, or so-called “schema drift”, which can break data pipelines or create data quality issues. Torch can detect anomalies and trends in data (“data drift”) that might pass data quality checks but are a concern, nonetheless. For example, entire records or groups of records that are missing could be detected. Data drift may require retraining ML models to keep them accurate. Moreover, stakeholders may want to be alerted to data trends and anomalies that represent a business opportunity or threat that might not be visible on existing reports and dashboards. Torch also helps reconcile data from source to target to ensure data fidelity. Together, these capabilities help you avoid broken data pipelines, poor data quality, and inaccurate ML and AI models.

Torch can also automatically classify, cluster, and provide associations to raw uncategorized data. This helps data teams make sense of large data sets. This provides a context for how each data record is associated with other records.



5. Granular and Cost-Effective Data Quality Analysis

We listened to feedback from Snowflake users who said that being forced to analyze an entire massive database table was often wasteful, too expensive, and slow. Using Torch's powerful but easy UI, data engineers can define segments of data they want to explore for data quality or to which they wish to apply data quality rules.

With segment analysis, users can limit data quality checks to potentially-problematic rows or columns, new data or high-priority data. Torch even lets users limit data quality analysis by the content of the fields themselves, such as values where N>1000, or when the text is "females" only. Torch also allows you to compare the data quality and health of different segments.

6. Automatically Find Anomalies and Root Cause Problems

Acceldata Torch uses machine learning to analyze historical trends of your CPU, memory, costs, and compute resources and discover anomalies potentially indicative of a data quality problem. It continuously runs tests by making assertions about your data that can be validated — or not.

Torch can also automatically identify root causes of unexpected behavior by comparing application logs, query runtimes, or queue utilization statistics. This lets teams avoid manually sifting through large datasets to debug data quality problems. It also helps them quickly identify which downstream data users are most affected. Besides saving your team valuable time, it also reduces storage and processing costs while boosting performance.

7. Automatically Clean and Validate Real-Time Data Streams

Companies are increasingly deploying real-time customer personalization, data-driven logistics, fraud detection and other internal operations, live user leaderboards, etc. Such disruptive use cases rely on real-time data in the form of event and message streams, change data capture (CDC) updates, and more. For that data to be useful for real-time analytics, it needs to be instantly cleaned and validated first.

Acceldata helps you monitor and automate the cleaning and validation of real-time data sources such as Apache Kafka connected to your Snowflake Data Cloud. Acceldata first analyzes the data stored in your Kafka cluster and monitors the events for faster throughput and better stability.

Acceldata then automatically flags incomplete, incorrect and inaccurate data in real time without requiring manual interventions from your team. This keeps data flowing, and reduces data downtime to a minimum.

Acceldata is flexible. Besides Kafka, Acceldata integrates with processing engines such as Spark, and other cloud storage and querying platforms like Amazon S3, Hive, HBase, Redshift and Databricks, as well as legacy systems such as MySQL, PostgreSQL and Oracle.
Automatically cleaning and validating your real-time data streams frees up your data team to innovate. It is just one of seven features of Acceldata's unified multi-dimensional data observability platform, which provides simple, full traceability of your data as it travels and transforms through its entire lifecycle.

Other solutions lack that multilayered visibility or provide an incomplete view. This forces you to cobble together individual solutions, each of which provides a different, incomplete view. This creates data fragmentation, as data teams can no longer observe data end-to-end. And this causes broken data pipelines, inexplicable data quality issues, and unexpected data outages, which in turn requires data teams to manually debug these problems.

Getting Started

Low ops ≠ No ops. Companies like yours realize that even with Snowflake, you must invest time and effort into building out a continuous data quality cycle in order to test, validate, and ameliorate data errors as they appear.

And the best technical partner to help you create and automate this continuous data quality cycle is a data observability platform like Acceldata that provides the single unified view of your data throughout its journey in your systems.

Learn more about how Acceldata can help you maximize the return on your Snowflake investment with insight into performance, quality, cost, and much more.