Operationalizing Data Governance for Accountability and Transparency

About Us

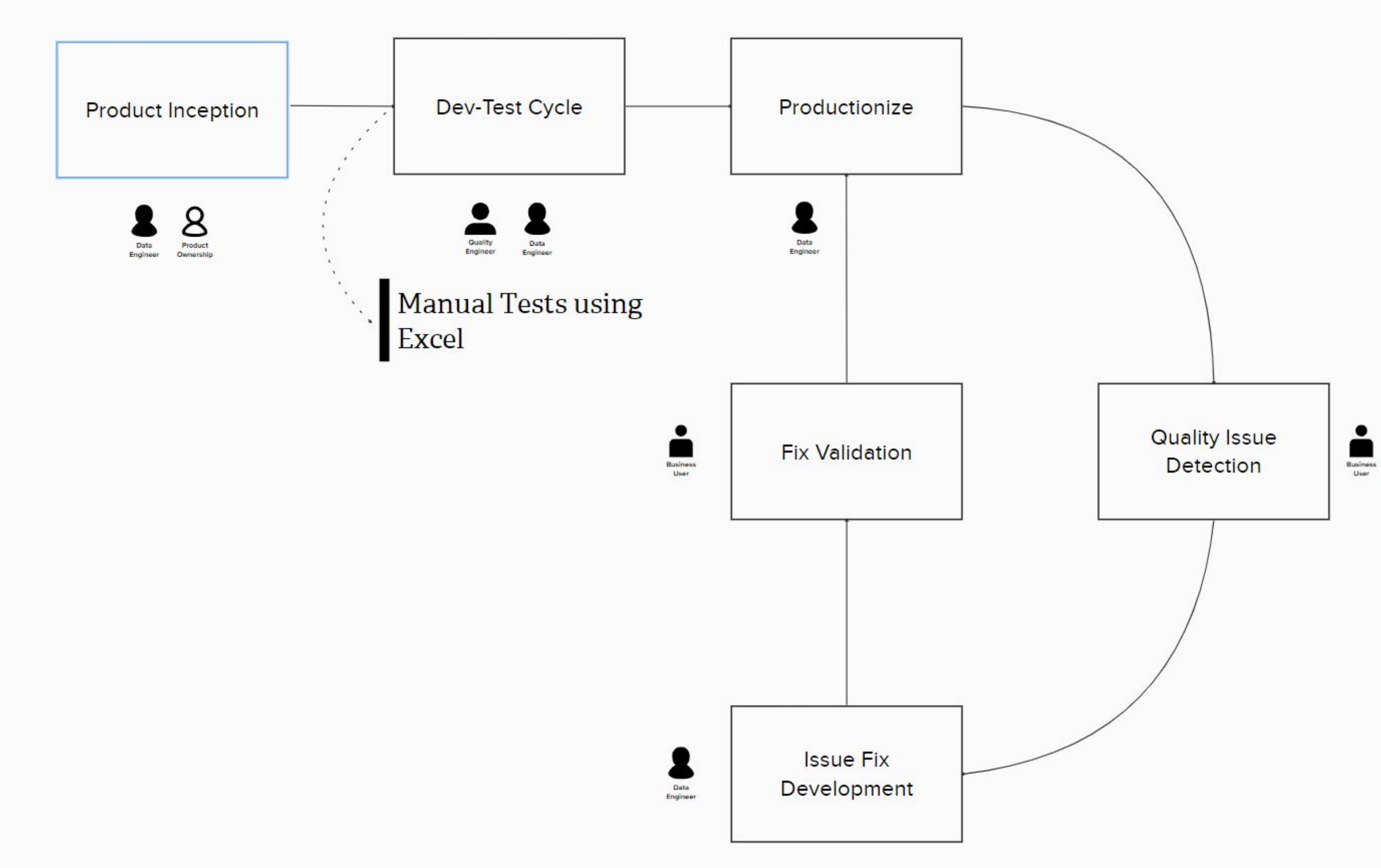
Anand Sethukumar
Staff Engineer - Current
Data Engineer - 15 years

Latravia White
Data Governance Manager- Current
Data Professional - 10 years

AGENDA

- Problem Statement
- Proposed Solution
- Contracts & Expectations
- Proactive Data Quality
- Soda Platform
- Inspiration
- Data Governance Strategy
- Q&A

Reactive Data Quality



To summarize..

Manual Tests

Tests written and executed in excel sheets

Regression Test

Little to no regression testing capabilities

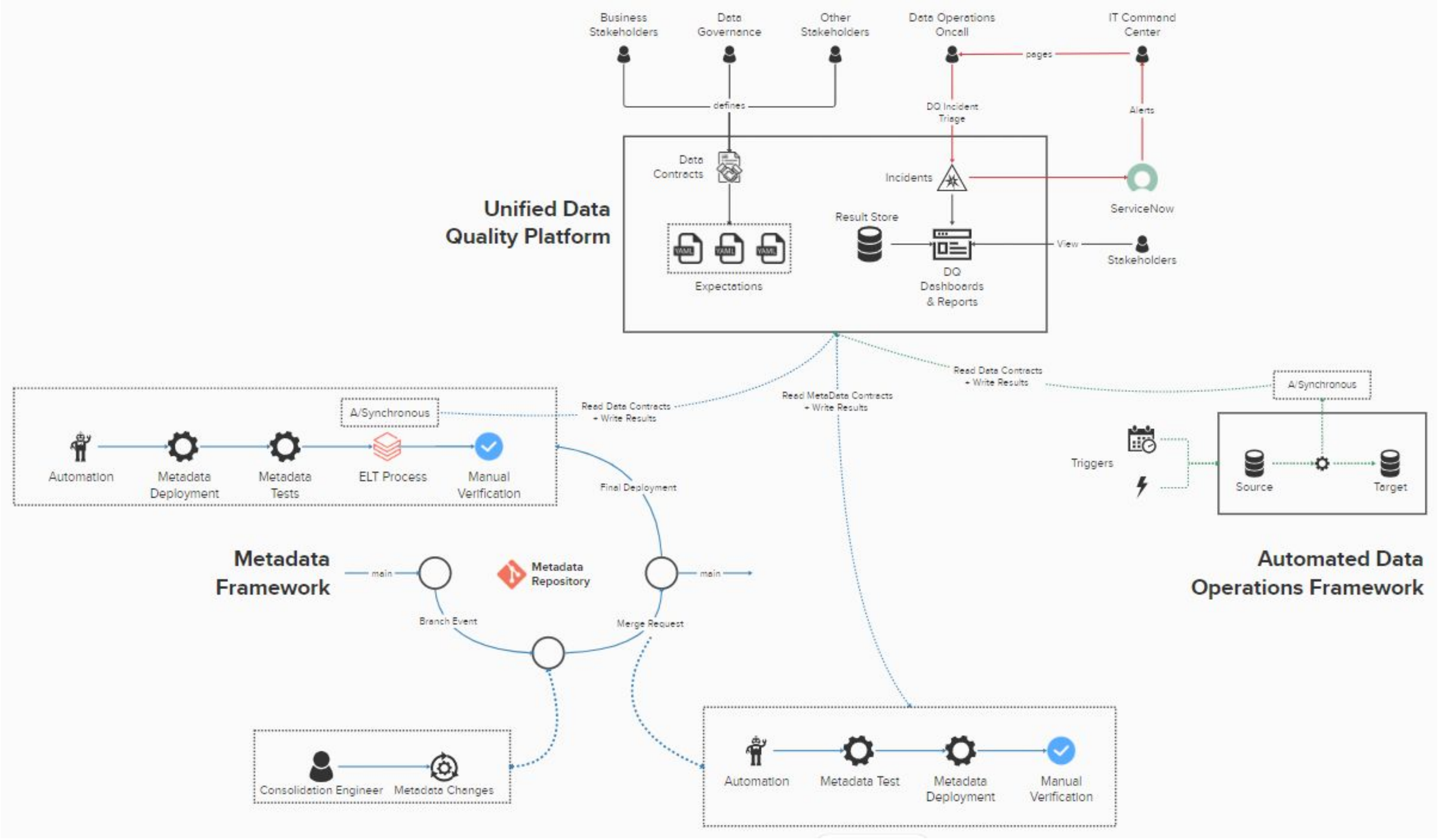
Tribal Knowledge

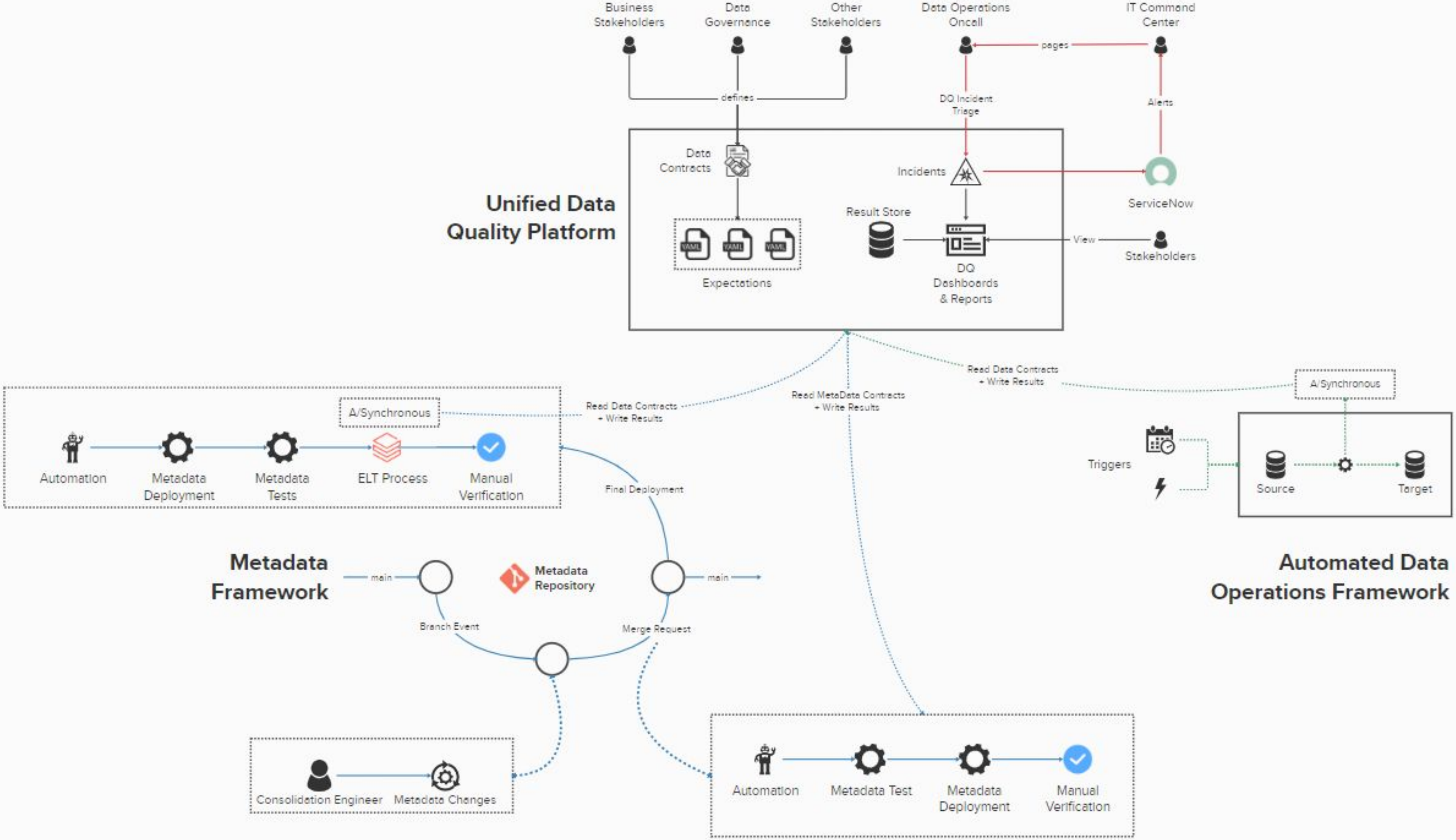
Knowledge of data expectations often confined to small groups of individuals

Expensive

Reactive data quality increases operational expenses

Data Observability





Unified Data Quality Platform

Business Stakeholders
Data Governance
Other Stakeholders
Data Operations Oncall
IT Command Center
defines
pages
DQ Incident Triage
Alerts
Data Contracts
Incidents
ServiceNow
Result Store
Expectations
DQ Dashboards & Reports
View
Stakeholders
Read Data Contracts + Write Results
Read MetaData Contracts + Write Results
Read Data Contracts + Write Results
A/Synchronous
Automation
Metadata Deployment
Metadata Tests
ELT Process
Manual Verification
Final Deployment

Metadata Framework

main
Metadata Repository
main
Branch Event
Merge Request
Consolidation Engineer
Metadata Changes
Automation
Metadata Test
Metadata Deployment
Manual Verification
A/Synchronous
Triggers
Source
Target

Automated Data Operations Framework

Objectives

Empowered Governance

Governance contributes directly to operations

Data Observability

A one-stop-shop view of the health of all data assets

Increased Participation

Data owners play a greater role in curating their domain's data

Engineering Discipline

Implementing contemporary engineering practices

Vendor Selection Scorecard

User-Friendliness

To enable non-technical folks to contribute to Data Quality

Connectivity

Pre-Built connectors to the data lake + data sources

Automation Patterns

CI/CD integration , ELT pipeline embedding & Orchestration

Incident Management

Integration with our incident management platform

No Data shuffling

Checks are translated to appropriate dialect and “pushed-down” to the target system

IaM & Security

Integration with our active directory through SAML and MFA. VPC Deployments

Custom Expectations

Domain Knowledge

Custom Logic

Regression Tests

SQL-Based Checks

Types of Proactivity

Data Supply Chain

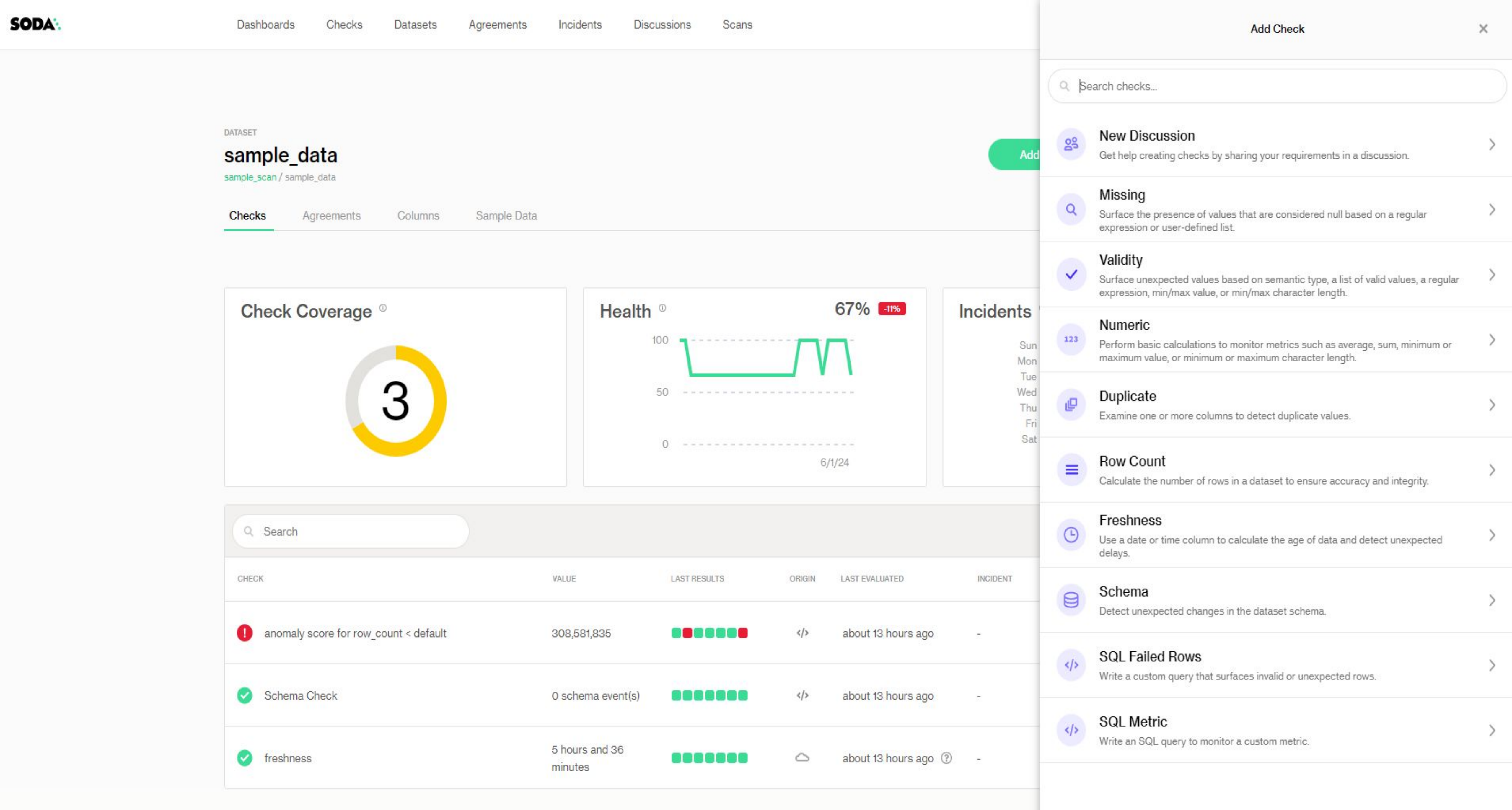
Where are checks executed in the data supply chain ?

Point of Materialization

At which point of data materialization (in memory , after materialization , after batch)

SDLC

TDD , CI/CD pipelines



DATASET

sample_data

sample_scan / sample_data

Checks | Agreements | Columns | Sample Data

CHECK	VALUE	LAST RESULTS	ORIGIN	LAST EVALUATED	INCIDENT
anomaly score for row_count < default	308,581,835	🟩🟥🟩🟩🟩🟩🟥	</>	about 13 hours ago	-
Schema Check	0 schema event(s)	🟩🟩🟩🟩🟩🟩🟩	</>	about 13 hours ago	-
freshness	5 hours and 36 minutes	🟩🟩🟩🟩🟩🟩🟩	☁	about 13 hours ago	-

Add Check

Search checks...

New Discussion
Get help creating checks by sharing your requirements in a discussion.

Missing
Surface the presence of values that are considered null based on a regular expression or user-defined list.

Validity
Surface unexpected values based on semantic type, a list of valid values, a regular expression, min/max value, or min/max character length.

Numeric
Perform basic calculations to monitor metrics such as average, sum, minimum or maximum value, or minimum or maximum character length.

Duplicate
Examine one or more columns to detect duplicate values.

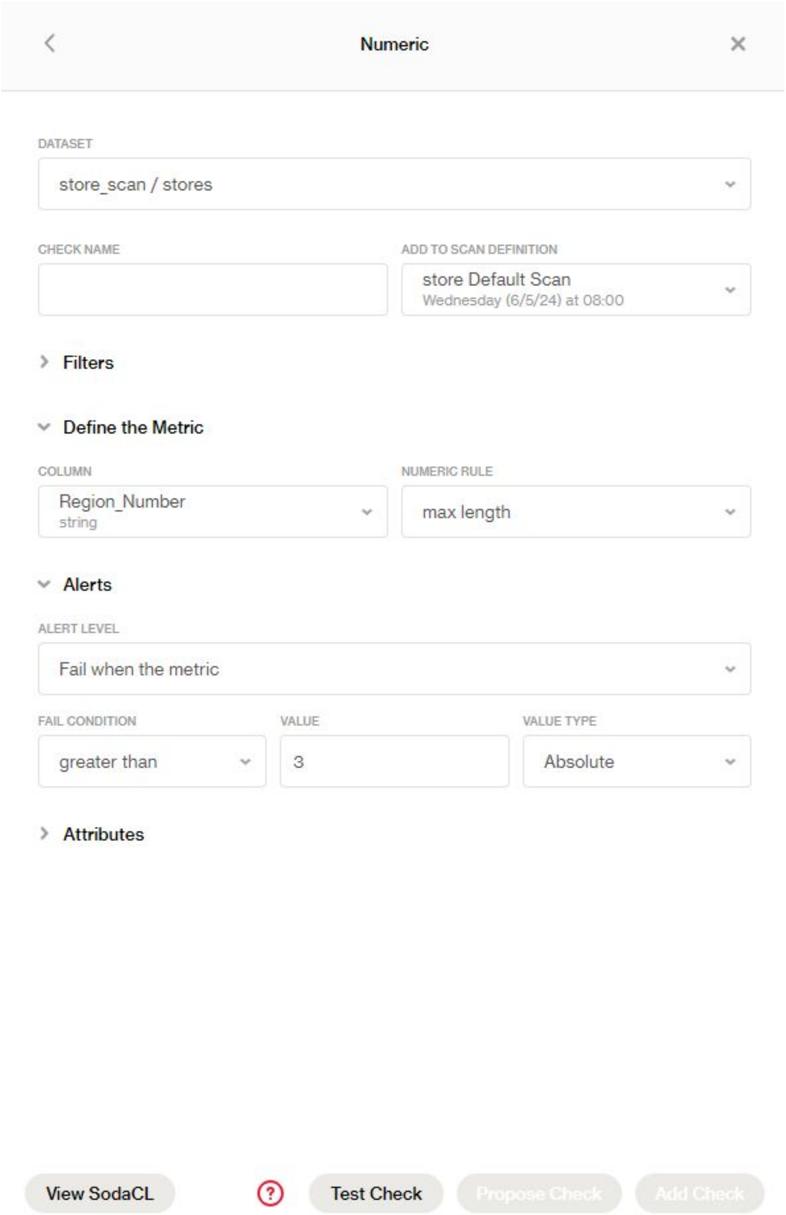
Row Count
Calculate the number of rows in a dataset to ensure accuracy and integrity.

Freshness
Use a date or time column to calculate the age of data and detect unexpected delays.

Schema
Detect unexpected changes in the dataset schema.

SQL Failed Rows
Write a custom query that surfaces invalid or unexpected rows.

SQL Metric
Write an SQL query to monitor a custom metric.



Numeric

DATASET
store_scan / stores

CHECK NAME

ADD TO SCAN DEFINITION
store Default Scan
Wednesday (6/5/24) at 08:00

Filters

Define the Metric

COLUMN
Region_Number
string

NUMERIC RULE
max length

Alerts

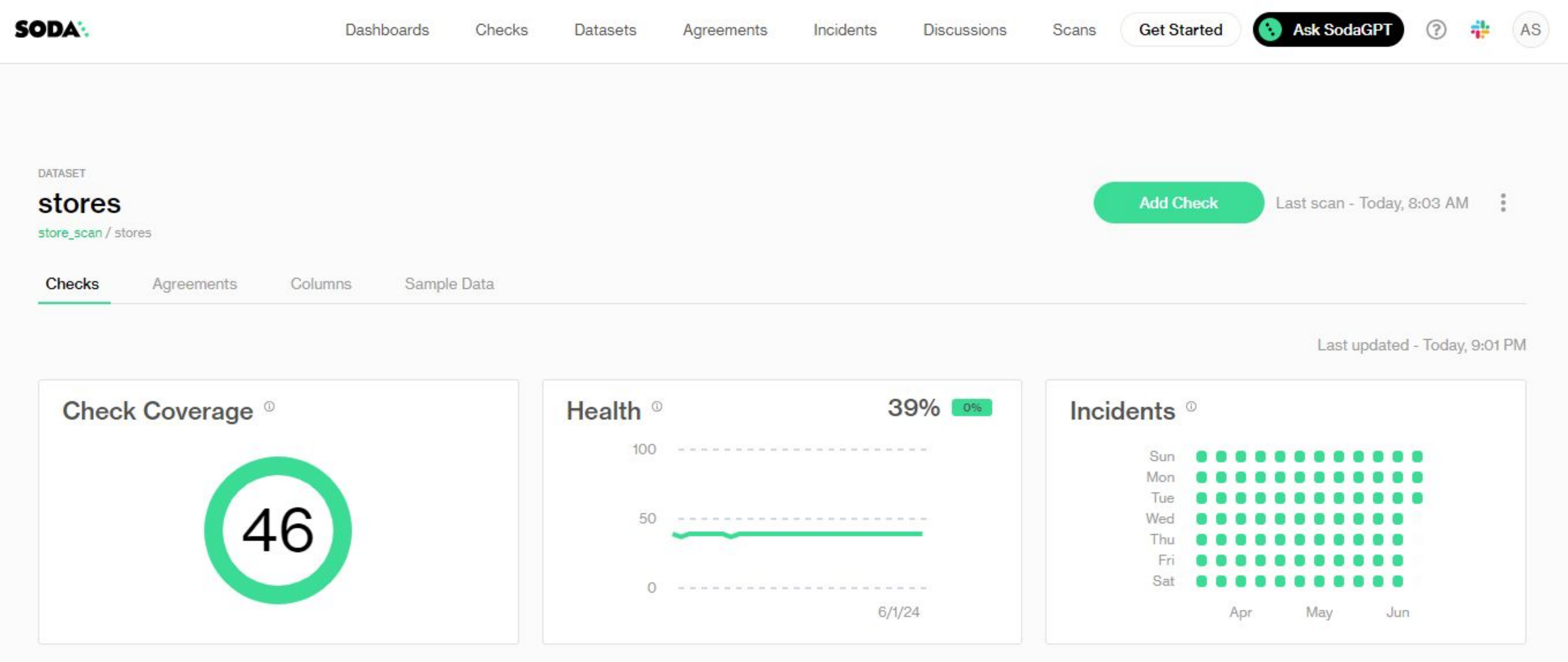
ALERT LEVEL
Fail when the metric

FAIL CONDITION
greater than

VALUE
3

VALUE TYPE
Absolute

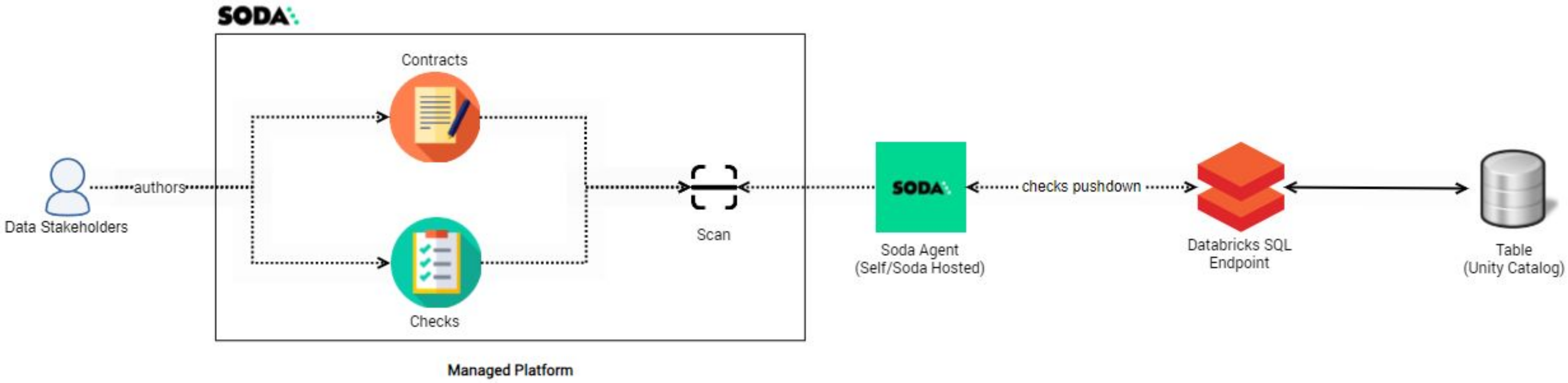
Attributes


SODA
Dashboards
Checks
Datasets
Agreements
Incidents
Discussions
Scans
Get Started
Ask SodaGPT
AS
DATASET
stores
store_scan / stores
Add Check
Last scan - Today, 8:03 AM
Checks
Agreements
Columns
Sample Data
Last updated - Today, 9:01 PM
Check Coverage
46
Health
39%
0%
100
50
0
6/1/24
Incidents
Sun
Mon
Tue
Wed
Thu
Fri
Sat
Apr
May
Jun

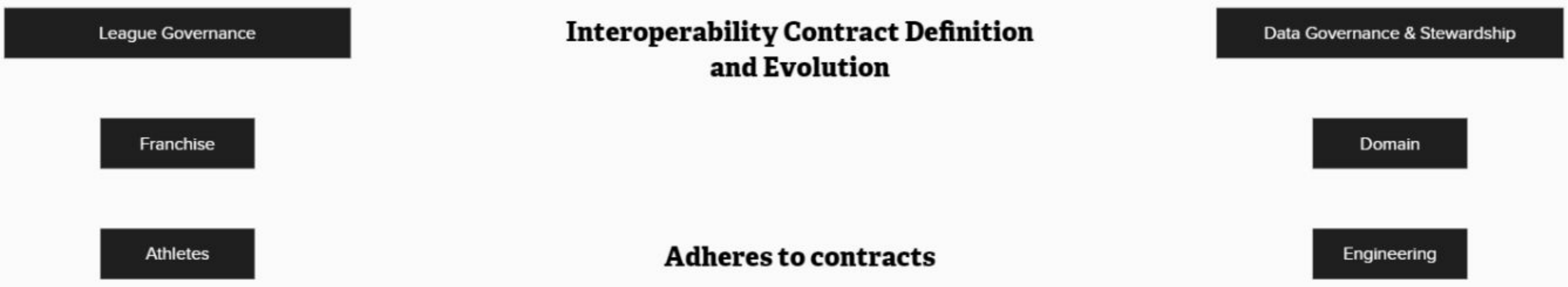
sodadata/sodacore

docs.soda.io

Booth 16


SODA
Contracts
Data Stakeholders
authors
Scan
Checks
Managed Platform
SODA
Soda Agent
(Self/Soda Hosted)
checks pushdown
Databricks SQL
Endpoint
Table
(Unity Catalog)

Inspiration



League Governance

Interoperability Contract Definition and Evolution

Data Governance & Stewardship

Franchise

Domain

Athletes

Adheres to contracts

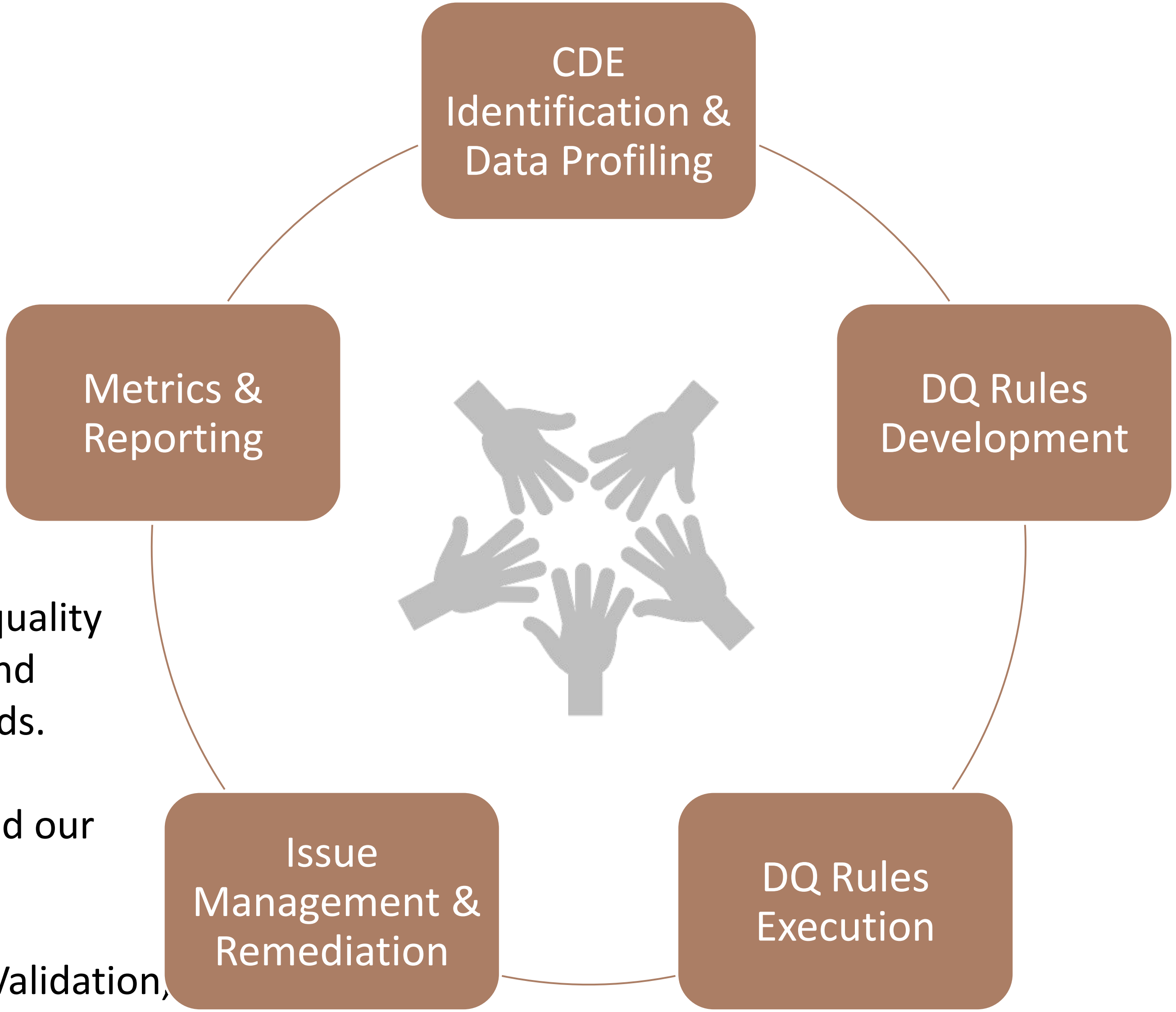
Engineering

Data Quality Strategy

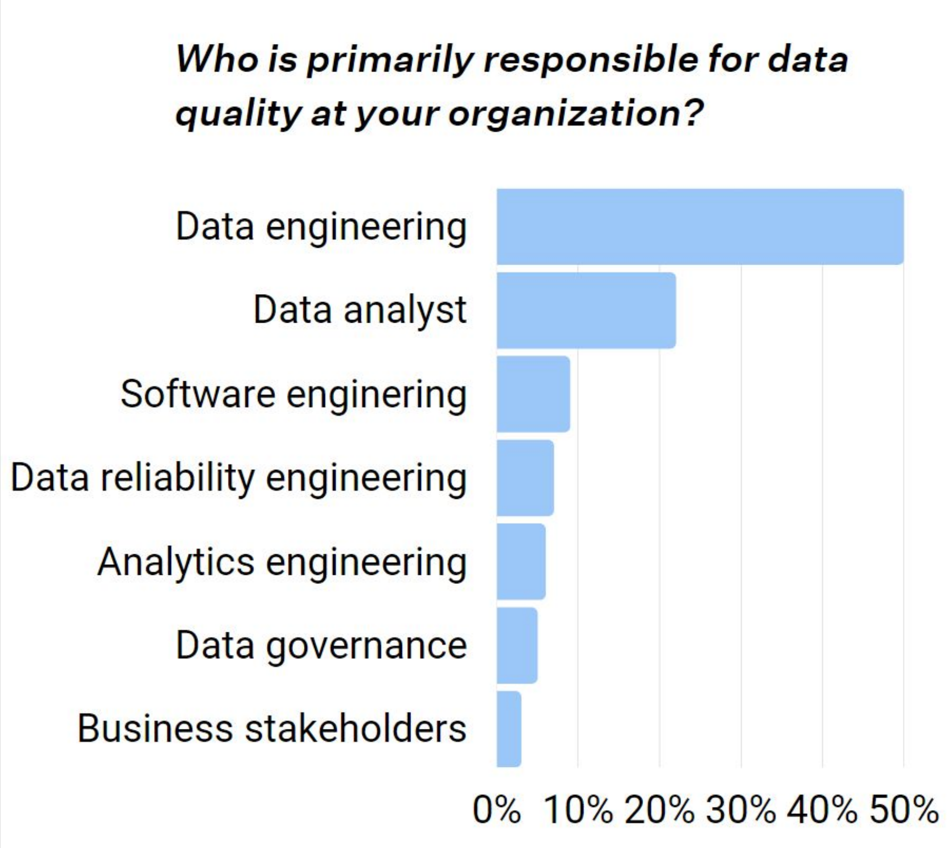
Governance plays a crucial role in defining a data quality strategy by establishing the framework, policies, and accountability needed to ensure high data standards.

Bridges the gap between Technology, Processes and our Stakeholders.

Pros of a Data Quality Strategy: Automated Data Validation, Increased Efficiency, Faster Decision Making, Improved Compliance, Enhanced Customer Satisfaction



CDE Identification & Data Profiling
DQ Rules Development
DQ Rules Execution
Issue Management & Remediation
Metrics & Reporting



Who is primarily responsible for data quality at your organization?

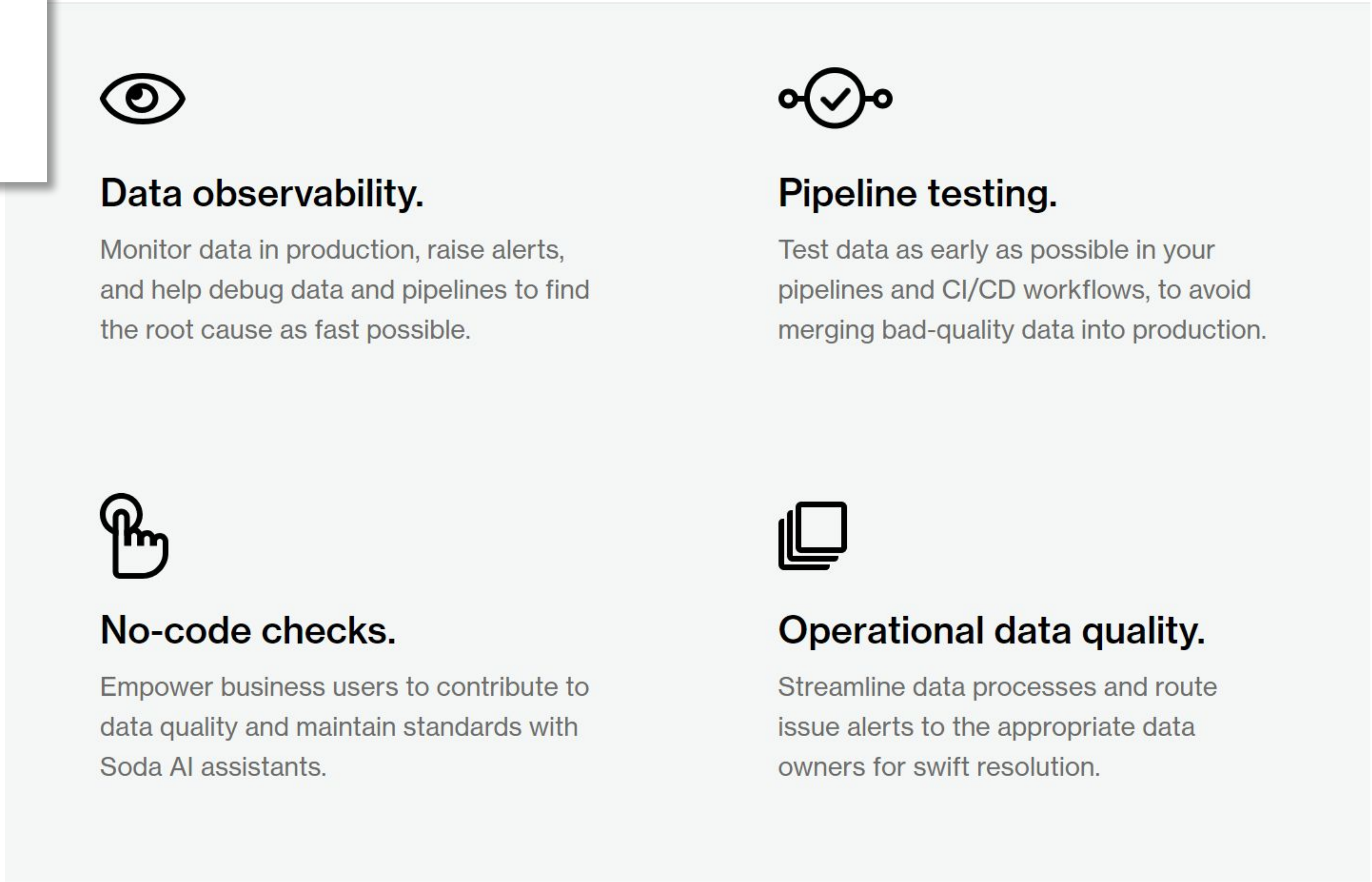
It is reported, 50% of Data Quality issues are caused by Engineers.

Governance Teams must break down the silos!

How do we do this?!

Share

Sharing data quality tools equips them to proactively detect and resolve issues, ensuring that data remains accurate and reliable throughout development. This collaboration enhances system performance and fosters accountability for data integrity



Data observability.

Monitor data in production, raise alerts, and help debug data and pipelines to find the root cause as fast possible.

Pipeline testing.

Test data as early as possible in your pipelines and CI/CD workflows, to avoid merging bad-quality data into production.

No-code checks.

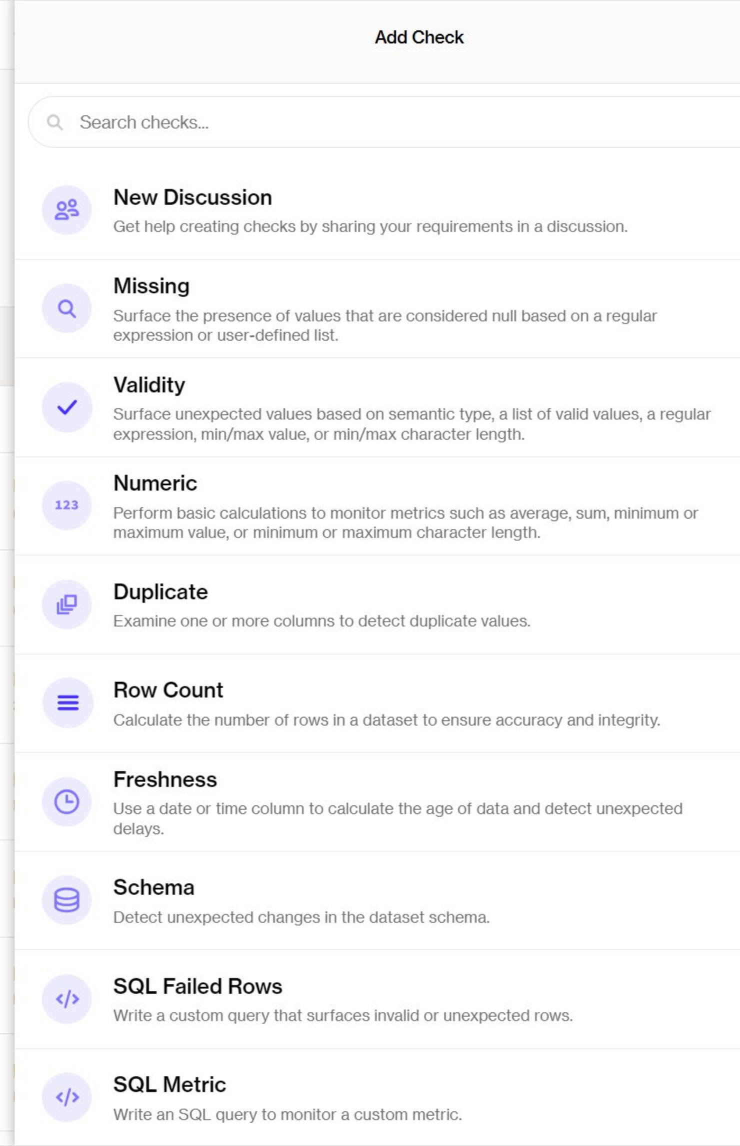
Empower business users to contribute to data quality and maintain standards with Soda AI assistants.

Operational data quality.

Streamline data processes and route issue alerts to the appropriate data owners for swift resolution.

Empower

Empowering engineers with data quality ensures they have the tools and knowledge to build reliable, high-performance systems that drive accurate decision-making. By embedding data quality best practices into their workflows, engineers can proactively prevent errors and enhance the overall trust in the organization's data assets.



Add Check

Search checks...

New Discussion
Get help creating checks by sharing your requirements in a discussion.

Missing
Surface the presence of values that are considered null based on a regular expression or user-defined list.

Validity
Surface unexpected values based on semantic type, a list of valid values, a regular expression, min/max value, or min/max character length.

Numeric
Perform basic calculations to monitor metrics such as average, sum, minimum or maximum value, or minimum or maximum character length.

Duplicate
Examine one or more columns to detect duplicate values.

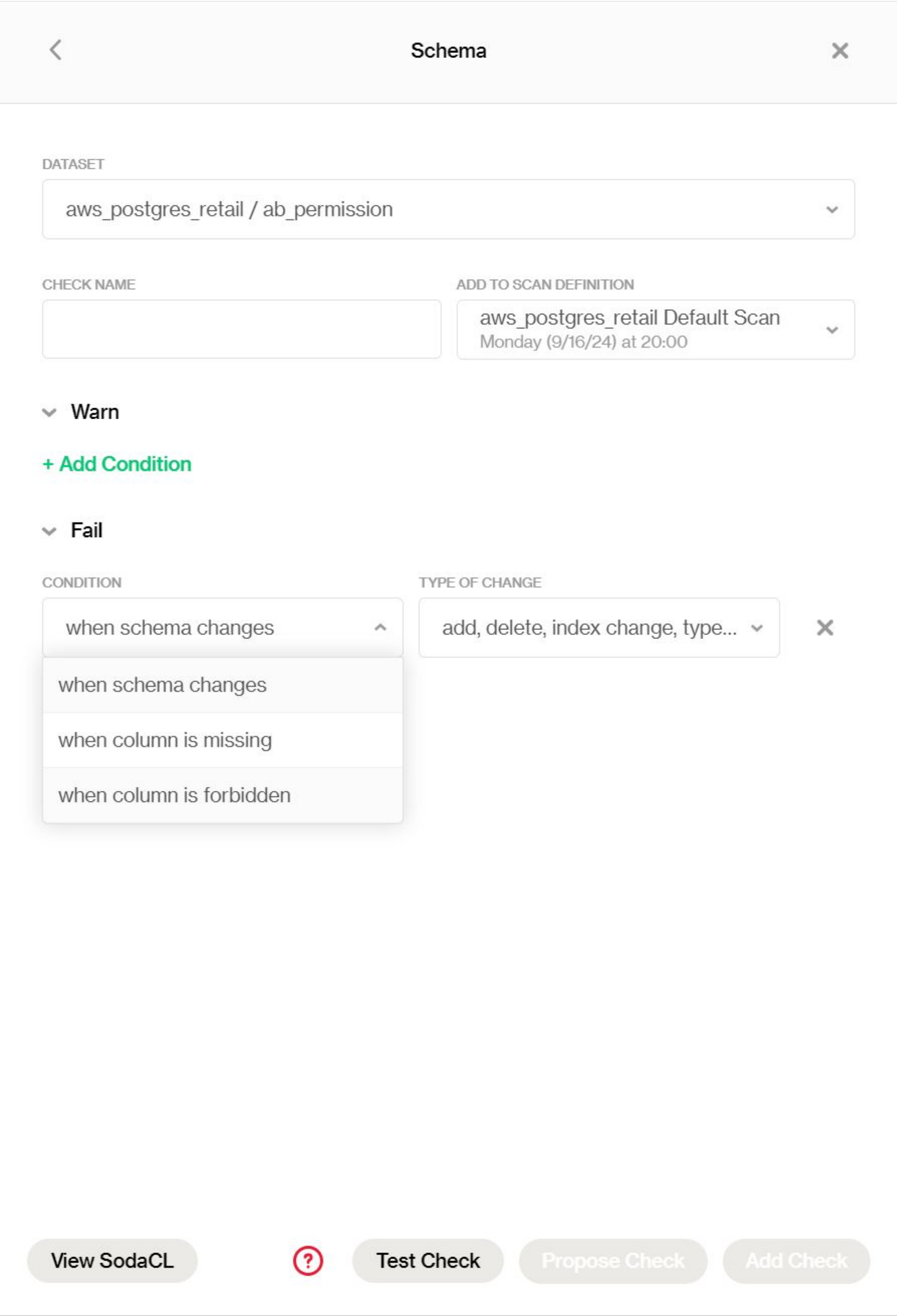
Row Count
Calculate the number of rows in a dataset to ensure accuracy and integrity.

Freshness
Use a date or time column to calculate the age of data and detect unexpected delays.

Schema
Detect unexpected changes in the dataset schema.

SQL Failed Rows
Write a custom query that surfaces invalid or unexpected rows.

SQL Metric
Write an SQL query to monitor a custom metric.



Schema

DATASET: aws_postgres_retail / ab_permission

CHECK NAME:

ADD TO SCAN DEFINITION: aws_postgres_retail Default Scan
Monday (9/16/24) at 20:00

Warn

+ Add Condition

Fail

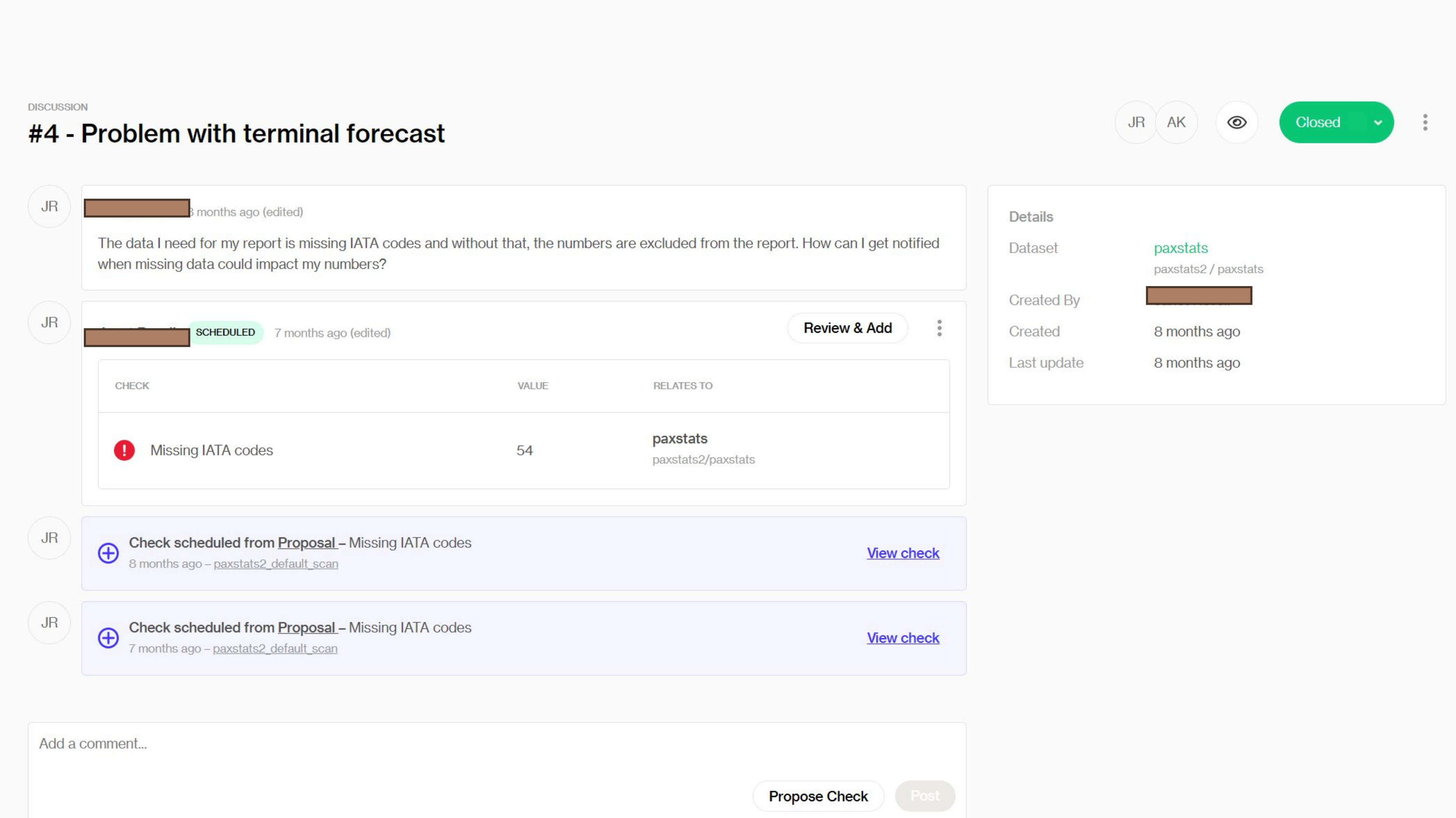
CONDITION: when schema changes

TYPE OF CHANGE: add, delete, index change, type...

- when schema changes
- when column is missing
- when column is forbidden

COLLABORATION

Fosters a shared responsibility for maintaining clean, accurate data across all systems and processes. This teamwork ensures that best practices are consistently applied, enabling smoother data integration and improved overall performance.



DISCUSSION

#4 - Problem with terminal forecast

JR [redacted] months ago (edited)

The data I need for my report is missing IATA codes and without that, the numbers are excluded from the report. How can I get notified when missing data could impact my numbers?

JR [redacted] SCHEDULED 7 months ago (edited)

CHECK	VALUE	RELATES TO
Missing IATA codes	54	paxstats paxstats2/paxstats

JR Check scheduled from Proposal – Missing IATA codes
8 months ago – paxstats2_default_scan

JR Check scheduled from Proposal – Missing IATA codes
7 months ago – paxstats2_default_scan

Add a comment...

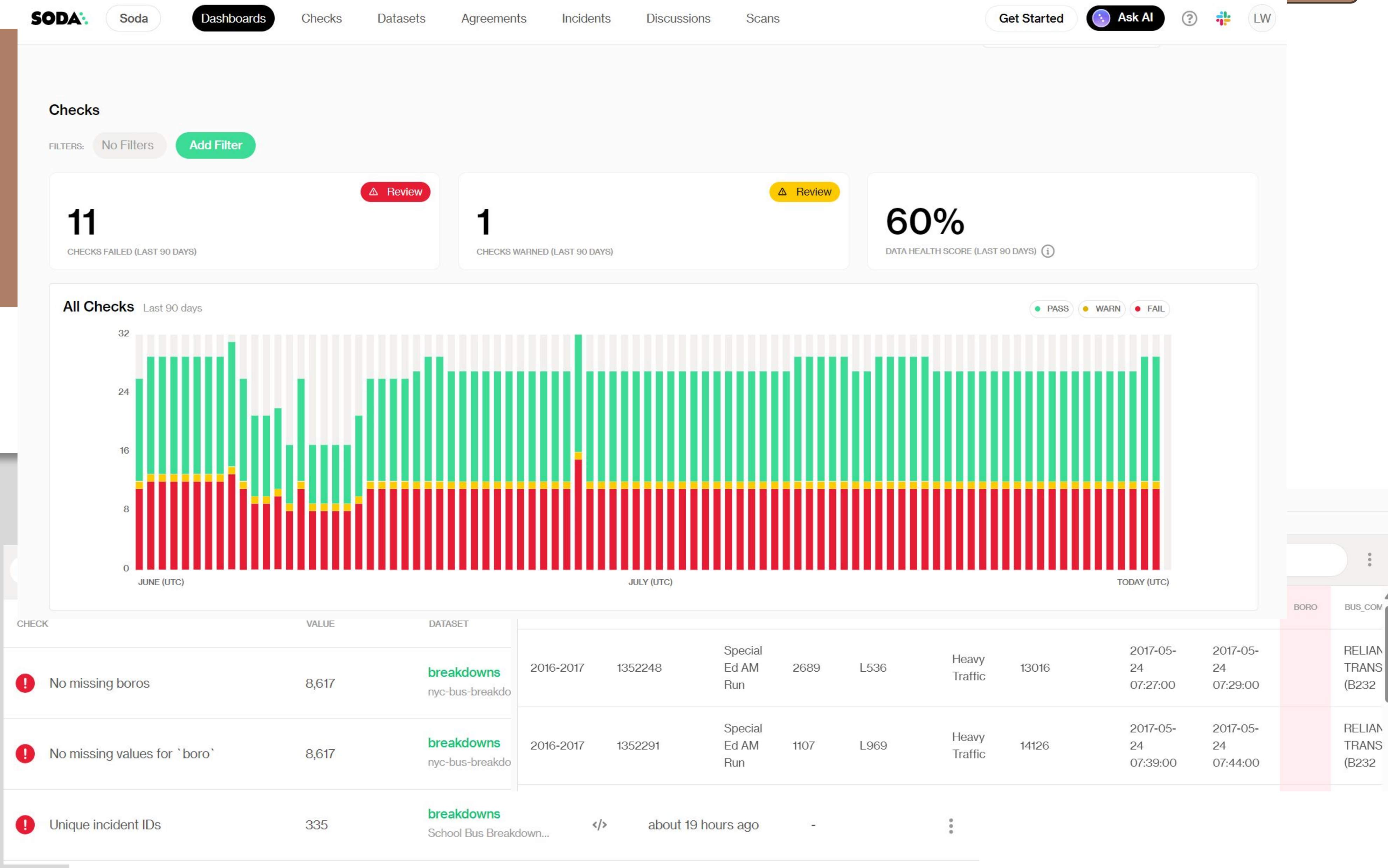
Details

Dataset	paxstats paxstats2 / paxstats
Created By	[redacted]
Created	8 months ago
Last update	8 months ago

MONITORING

Dashboards are vital for monitoring data quality in real time, providing clear visual insights into key metrics like accuracy, completeness, and consistency. They enable quick identification of issues, allowing engineers to address data quality problems proactively and maintain system reliability.

Sample Results



Checks

FILTERS: No Filters Add Filter

Review
11
CHECKS FAILED (LAST 90 DAYS)

Review
1
CHECKS WARNED (LAST 90 DAYS)

60%
DATA HEALTH SCORE (LAST 90 DAYS)

CHECK	VALUE	DATASET			
No missing boros	8,617	breakdowns nyc-bus-breakdo			
No missing values for `boro`	8,617	breakdowns nyc-bus-breakdo			
Unique incident IDs	335	breakdowns School Bus Breakdown...	</>	about 19 hours ago	-

									BORO	BUS_COM
2016-2017	1352248	Special Ed AM Run	2689	L536	Heavy Traffic	13016	2017-05-24 07:27:00	2017-05-24 07:29:00		RELIAN TRANS (B232
2016-2017	1352291	Special Ed AM Run	1107	L969	Heavy Traffic	14126	2017-05-24 07:39:00	2017-05-24 07:44:00		RELIAN TRANS (B232

Q & A

Feel Free to connect with us on LinkedIn!

linkedin.com/in/latraviawhite/ & linkedin.com/in/anandsethukumar/