From Flat to Sparkling

Monitoring Data Quality with Soda in Microsoft Fabric

Soda + Microsoft Fabric

Data Platform Next Step · Copenhagen · June 2, 2026



Sam Debruyn

Freelance Data Platform Architect
Microsoft Data Platform MVP · dbt Champion
Contributed the Soda Fabric connector

sam@debruyn.dev · debruyn.dev

Premium sponsors

VENZO_ INSPARI

valantic companies

itm8®

Eddytor™

dataon

Power your data

Standard sponsors

Backstage

Columbus

TheVIT

DATA - ANALYTICS - AI

twoday

Why Data Quality Matters

The silent cost of bad data

Sam Debruyn

Why Proactive Data Quality Monitoring?

You don't want to appear in the following list...

Catch issues early

Before they reach reports, dashboards, or downstream consumers.

Build trust

Stakeholders trust data they know is being actively monitored.

Reduce firefighting

Automated checks replace manual spot-checking and reactive debugging.

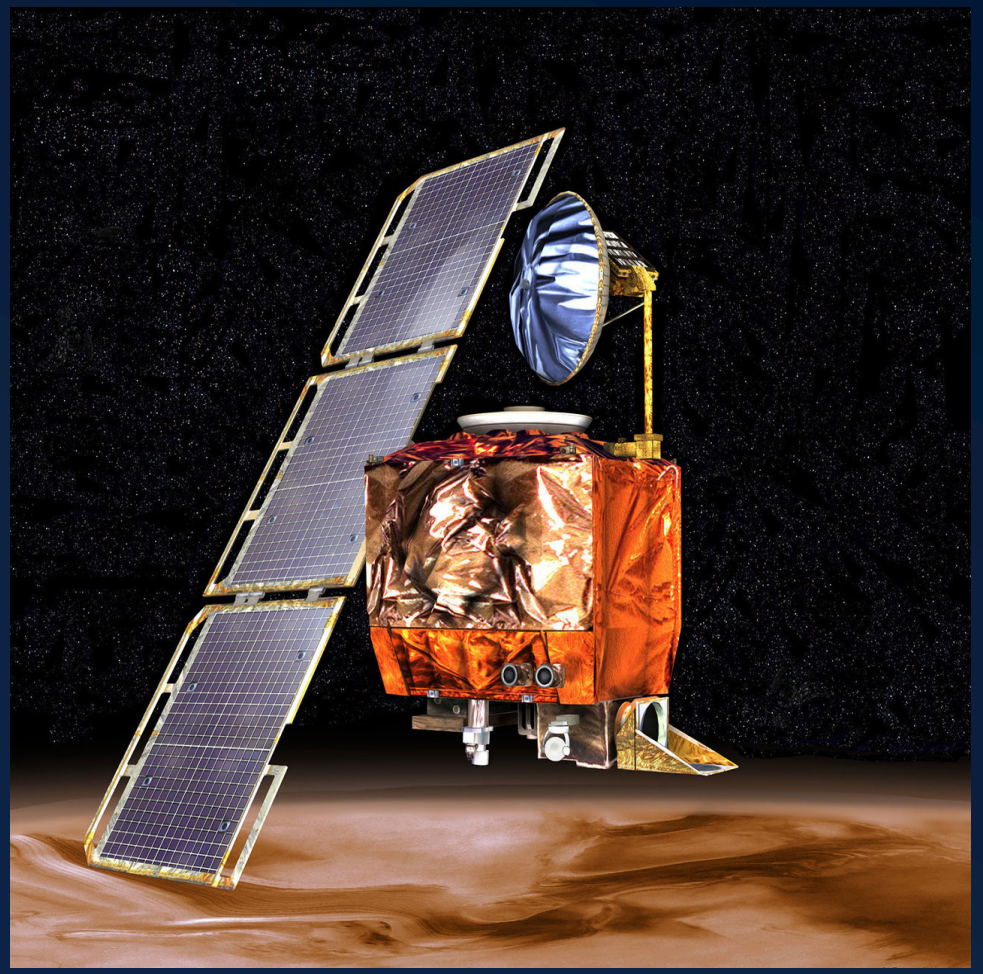
Bad data doesn't announce itself — you have to go looking for it.

NASA (1999)

The Mars Climate Orbiter burned up in space because Imperial units were used instead of Metric in navigation calculations.

Cost: $327.6 million

A unit conversion data quality issue



American Airlines (2017)

A glitch in the scheduling system allowed too many pilots to schedule vacations at the same time, grounding 15,000+ flights during the holiday season.

15,000+ flights grounded

A consistency data quality issue



Samsung Securities (2018)

A data entry error led to the distribution of 30 times more shares than available to employees as part of a stock dividend.



Market value loss: $300M

A data validation issue

Citigroup (2025)

A backup payments screen pre-populated the amount field with 15 zeros. The operator didn't clear them. Two reviewers approved. A $280 transaction posted as $81 trillion.

Near-miss: $81,000,000,000,000

Caught after 90 minutes. A UI default-value issue.



British Post Office / Fujitsu (1999–2015)

Accounting software Horizon, developed by Fujitsu, reported incorrect financial data for over 15 years.



Hundreds of postmasters wrongfully convicted of fraud

Destroyed livelihoods, imprisonments, suicides.

The largest miscarriage of justice in UK legal history.

Spanish Navy (2013)

Due to a decimal point error, a newly built submarine was more than 75 tons overweight and would not float.

Cost: €2.2 billion

A precision data quality issue

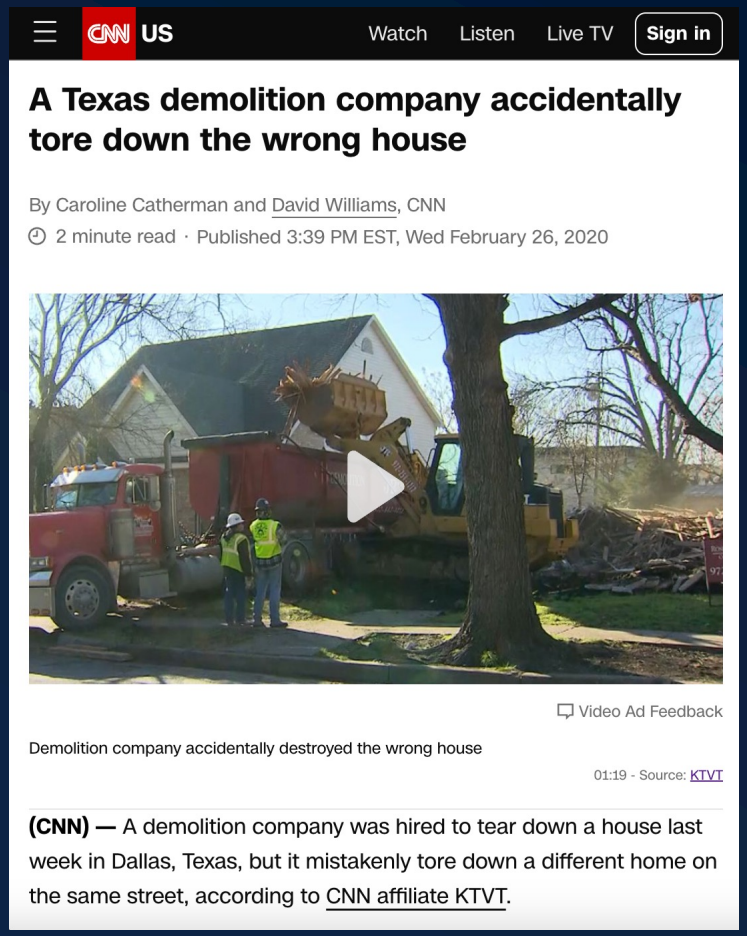


JR's Demolition (2020)

It seems to happen quite often that demolition companies tear down the wrong house.

Wrong address, wrong building

A referential integrity data quality issue



CNN US Watch Listen Live TV Sign in

A Texas demolition company accidentally tore down the wrong house

By Caroline Catherman and David Williams, CNN

2 minute read · Published 3:39 PM EST, Wed February 26, 2020

Video Ad Feedback

Demolition company accidentally destroyed the wrong house

01:19 - Source: KTVT

(CNN) — A demolition company was hired to tear down a house last week in Dallas, Texas, but it mistakenly tore down a different home on the same street, according to CNN affiliate KTVT.

Qantas QF520 (2024)

An employee mistyped the aircraft code as a Boeing 717 instead of a 737. The system silently removed 51 passengers from the loadsheet. The plane took off 4.29 tonnes heavier than the pilots believed.



4,291 kg weight discrepancy

A cascading master-data issue

Data Quality Dimensions

What to check — and what Soda can monitor

Accuracy

Equal to the real-life value

Completeness

All mandatory values present

Consistency

Values align across data sets

Timeliness

Data available when needed

Uniqueness

No unintended duplicates

Validity

Values within allowable bounds

"… the planning, implementation, and control of activities that apply quality management techniques to data, in order to assure it is fit for consumption and meet the needs of data consumers."

— DAMA-DMBOK (Data Management Body of Knowledge)

Data Quality Monitoring in Fabric

A growing landscape of tools and approaches

Purview DQ

Rules engine in Microsoft Purview

Separate portal
UI-only, no code

Paid: per governed asset + DGPU

dbt tests

SQL assertions in dbt models

Requires dbt adoption
No built-in results UI

Free (OSS)

Soda

YAML checks and data contracts

Native Fabric package
CLI + CI/CD ready

Free (ELv2)

GX

Python validation checkpoints

Heavy Python setup
No Fabric package

Free (OSS)

Today's deep-dive → Soda: code-first · Fabric-native · contract-driven

Meet Soda

Data quality testing, monitoring, and contracts

What is Soda?

Data quality testing, monitoring, and contracts

Open-source data quality engine

Write declarative checks in YAML
Run scans against any data source
Monitor metrics over time
Define and verify data contracts
Integrate in CI/CD and pipelines

Founded in Brussels, Belgium
Soda Core v4 · Elastic License 2.0

Quick start

```
pip install soda-fabric
soda contract verify -ds my_fabric
```

A data contract

```
dataset: dwh/dbo/orders
columns:
  - name: order_id
    data_type: int
    checks:
      - missing:
      - duplicate:
```

Core Concepts

The building blocks of Soda

Data Source

A data store: Fabric DWH,
Lakehouse, Postgres, …

Dataset

A table, view, or file
in a data source

Check

A test that measures a metric
and evaluates a threshold

Contract

Schema + checks for a dataset.
The v4 primary artifact.

Scan

A run validating one or more
contracts against a source

Check outcomes

Pass: Within threshold

Warn: Soft threshold exceeded

Fail: Hard threshold exceeded

Error: Soda could not execute

Data Contracts

The foundation of Soda v4

What changed in Soda Core 4?

Contracts replace SodaCL checks as the default.

A contract defines the full schema of a dataset plus all the checks you want to run on it.

The old SodaCL syntax (checks for dataset:) is no longer actively developed. A migration tool is available to convert existing checks.

Nine check keys, lots of options.

Contract-first workflow

1. **Define contract YAML**
   Schema + checks for each dataset
2. **soda contract verify**
   Run checks against your data source
3. **Integrate in CI/CD**
   Fail pipelines on contract violations
4. **Publish to Cloud**
   Track metrics and share with team

Anatomy of a Data Contract

Schema, checks, and thresholds in one file

```yaml
dataset: dwh/dbo/fact_orders
columns:
  - name: order_id
    data_type: int
    checks:
      - missing:
      - duplicate:

  - name: status
    data_type: varchar
    checks:
      - invalid:
          valid_values: ['PAID', 'PENDING']

checks:              # dataset-level
  - schema:
      allow_extra_columns: false
  - freshness:
      column: order_date
```

dataset

Fully-qualified table path in your source

columns

Full schema definition with types

checks

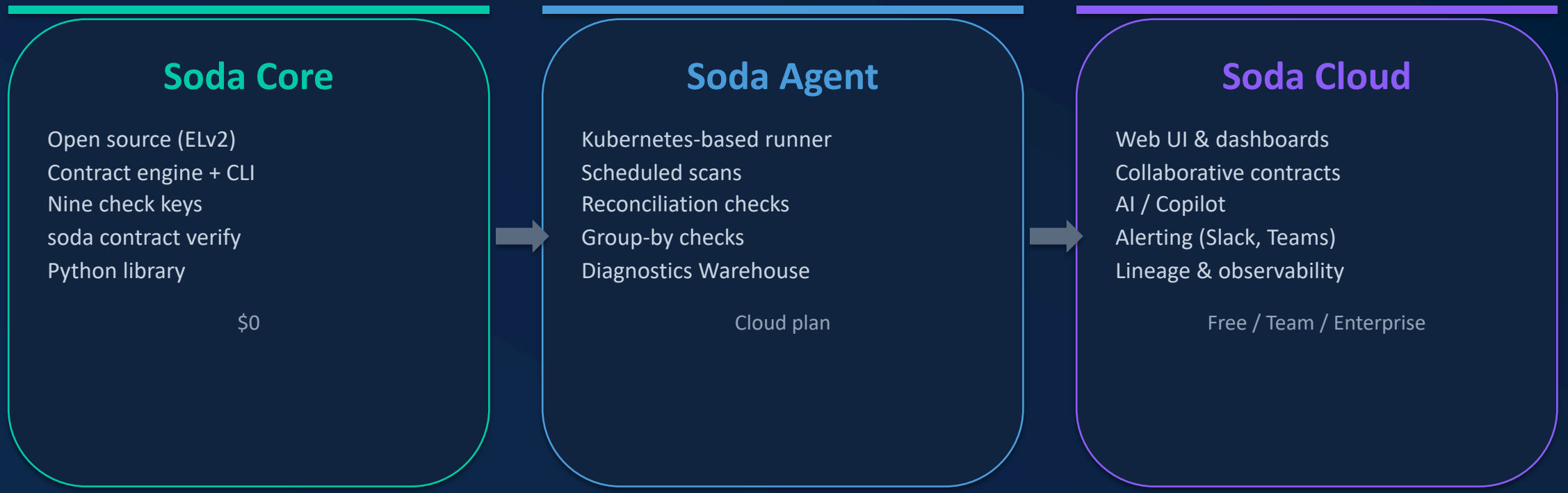
Per-column or dataset-level (bottom)

Built-in types

missing, duplicate, invalid, aggregate, schema, row_count, freshness, failed_rows, metric — 9 keys, lots of options

The Soda Ecosystem

Core (open source) + Agent + Cloud



Soda Library (v3 commercial extension) has been retired in v4.

Soda × Microsoft Fabric

Two paths to data quality monitoring

Fabric Data Warehouse

Package: soda-fabric
Protocol: TDS (T-SQL)
Auth: service principal or interactive

Soda generates SQL queries and sends them to the Fabric SQL Analytics endpoint. Results are fetched via pyodbc.

Fabric Lakehouse

Package: soda-spark-df
Protocol: Spark DataFrames
Auth: notebook session context

Soda runs inside a Spark session in a Fabric Notebook. Scans execute as Spark jobs on the Lakehouse engine.

Both connect to OneLake — choose based on your compute engine.

Getting Connected

Configuration for Fabric Data Warehouse

data_source.fabric.yml

```yaml
type: fabric
name: dwh_sales
connection:
  host: ${env.FABRIC_HOST}
  port: 1433
  database: ${env.FABRIC_DB}
  authentication: activedirectoryserviceprincipal
  client_id: ${env.FABRIC_SP_CLIENT_ID}
  client_secret: ${env.FABRIC_SP_CLIENT_SECRET}
  driver: ODBC Driver 18 for SQL Server
```

Run your first contract

```
pip install soda-fabric
soda contract verify \
  -ds config/data_source.fabric.yml \
  -c contracts/fact_orders.yml
```

Authentication options

`activedirectoryserviceprincipal`

Client ID + secret
For pipelines and scheduled scans

`activedirectoryinteractive`

Browser popup (MFA)
Great for local development

`activedirectorypassword / sql`

Username + password fallback.
Avoid in production.

Demo

Soda on Microsoft Fabric Data Warehouse

- Connect Soda to a Fabric Data Warehouse
- Define a data contract for a dataset
- Run soda contract verify
- Explore results and check outcomes
- Integrate in a Fabric Notebook

Going Further

Check types, anomaly detection, AI, and integrations

Built-in Check Types

Nine check keys, lots of options

Per column

missing
duplicate
invalid
aggregate

Per dataset

schema
row_count
freshness
failed_rows

Validity options

valid_values
valid_format (regex)
valid_min / valid_max
valid_length

Custom SQL

metric: expression
metric: query
failed_rows: query
Anomaly: in Soda Cloud

Custom SQL checks let you write any validation logic.

Anomaly Detection

Statistical monitoring without manual thresholds

How it works

Soda Cloud collects metric values over time and detects anomalies using MAD (Median Absolute Deviation) — robust against outliers.

No more guessing thresholds.

Instead of row_count > 1000, let Soda Cloud learn what 'normal' looks like and alert you when values deviate significantly.

Where it lives

```
# In your contract, push metrics to
# Soda Cloud. Configure anomaly
# detection on row_count / metrics
# from the Cloud UI — no YAML needed.
```

MAD-based

Robust statistical method.
70% fewer false positives.

Cloud-only in v4

Anomaly detection moved out of Core; it lives in Soda Cloud.

Integration Patterns

Embed Soda in your existing workflows

Fabric Notebook

Run Soda inside a Spark session.
Schedule via Data Factory pipelines.
Use managed identity for auth.

Best for Lakehouse + scheduled scans

CI/CD Pipeline

Run soda contract verify in GitHub Actions, Azure DevOps, or GitLab CI.
Fail the build on contract violations.

Best for deployment-time validation

dbt Integration

Soda can ingest dbt test results from manifest.json and run_results.json.
Combine dbt tests + Soda contracts.

Best for dbt-powered pipelines

Orchestrators

Programmatic scans from Airflow, Dagster, or Prefect DAGs.
Community guides available.

Best for complex pipelines

Soda Cloud

The collaborative data quality layer

Observability Dashboard

Track metrics over time. Spot trends and anomalies in a visual timeline.

Smart Alerting

Slack, Teams, Jira, PagerDuty. Route alerts to the right people automatically.

Collaborative Contracts

Engineers write in Git. Business users review in UI. Shared ownership of DQ.

No-code Check Builder

Non-technical users can add checks without writing YAML or SQL.

AI Copilot

Auto-generate contracts from production data. Ask AI to explain failures.

Diagnostics Warehouse

Store failed records in your own warehouse for root-cause analysis.

Free tier available · Team: $750/month · Enterprise: custom

Soda AI

AI-powered data quality — from CLI to Cloud

Soda AI CLI (beta)

An AI agent built into soda-core that helps you write and migrate data contracts.
Run: soda ai

Open beta since January 2026.

ML Monitoring (NannyML)

Soda acquired NannyML (Leuven, June 2025) for ML drift detection without ground-truth labels. Their OSS library stays open.

Extends Soda beyond tabular DQ into ML.

Contracts Autopilot

Auto-generate contracts from production data. Soda profiles your datasets and proposes appropriate checks based on patterns it finds.

Soda Cloud Enterprise only.

AI in Soda Cloud

AI Checks Assistant: describe what you want in natural language, get contract YAML.
Ask AI: ask questions about your DQ results.

Available on Team and Enterprise plans.

Licensing & Pricing

What changed and what it means for you

License: Elastic License 2.0

Since June 2025, Soda Core is no longer Apache 2.0. It's now Elastic License 2.0 (source-available, not OSI-approved).

What's restricted:

- Cannot offer Soda Core as a managed service
- Cannot circumvent license keys

Internal use at your organization: no change.

Soda Cloud Pricing

Free — $0/month

Pipeline testing, metrics observability, alerting, free SPUs. No credit card.

Team — $750/month

Catalog integrations, pay-as-you-go SPUs, unlimited users.

Enterprise — Custom

Collaborative contracts, no-code UI, advanced AI, RBAC/SSO, private deploy.

For strict OSS policies: evaluate ELv2 restrictions. Alternatives exist (Provero, Great Expectations).

Wrapping Up

Key Takeaways

Data quality won't fix itself

Proactive monitoring catches issues before your stakeholders do.

Soda makes DQ accessible

Declarative YAML contracts. No custom framework needed.

First-class Fabric support

soda-fabric for Data Warehouse, soda-spark-df for Lakehouse.

Start small, grow naturally

Free tier + CLI → CI/CD integration → Cloud for team-wide observability.

Contracts are the future

Schema + checks in one artifact. Version-controlled. Collaborative.

Get Started

From zero to first contract in 5 minutes

```
# Install
pip install soda-fabric

# Create config/data_source.fabric.yml
# Create contracts/my_table.yml

# Verify
soda contract verify \
  -ds config/data_source.fabric.yml \
  -c contracts/my_table.yml
```

Resources

docs.soda.io	Official documentation
github.com/sodadata/soda-core	Source code (ELv2)
cloud.soda.io	Soda Cloud (free tier)
docs.soda.io/soda/connect-fabric.html	Fabric-specific guide

Thank You!

Questions?

sam@debruyn.dev · debruyn.dev

Raffle Prizes

EUROPEAN MICROSOFT FABRIC Community Conference









Download Slides

debruyn.dev/dpns26

Session Feedback

Share your thoughts!

From Flat to Sparkling: Monitoring Data Quality with Soda in Microsoft Fabric