



FROM FLAT TO SPARKLING

MONITORING DATA QUALITY WITH
SODA IN MICROSOFT FABRIC

Sam Debruyne

Future Data Driven Summit
September 2024





Who am I?

Sam Debruyn

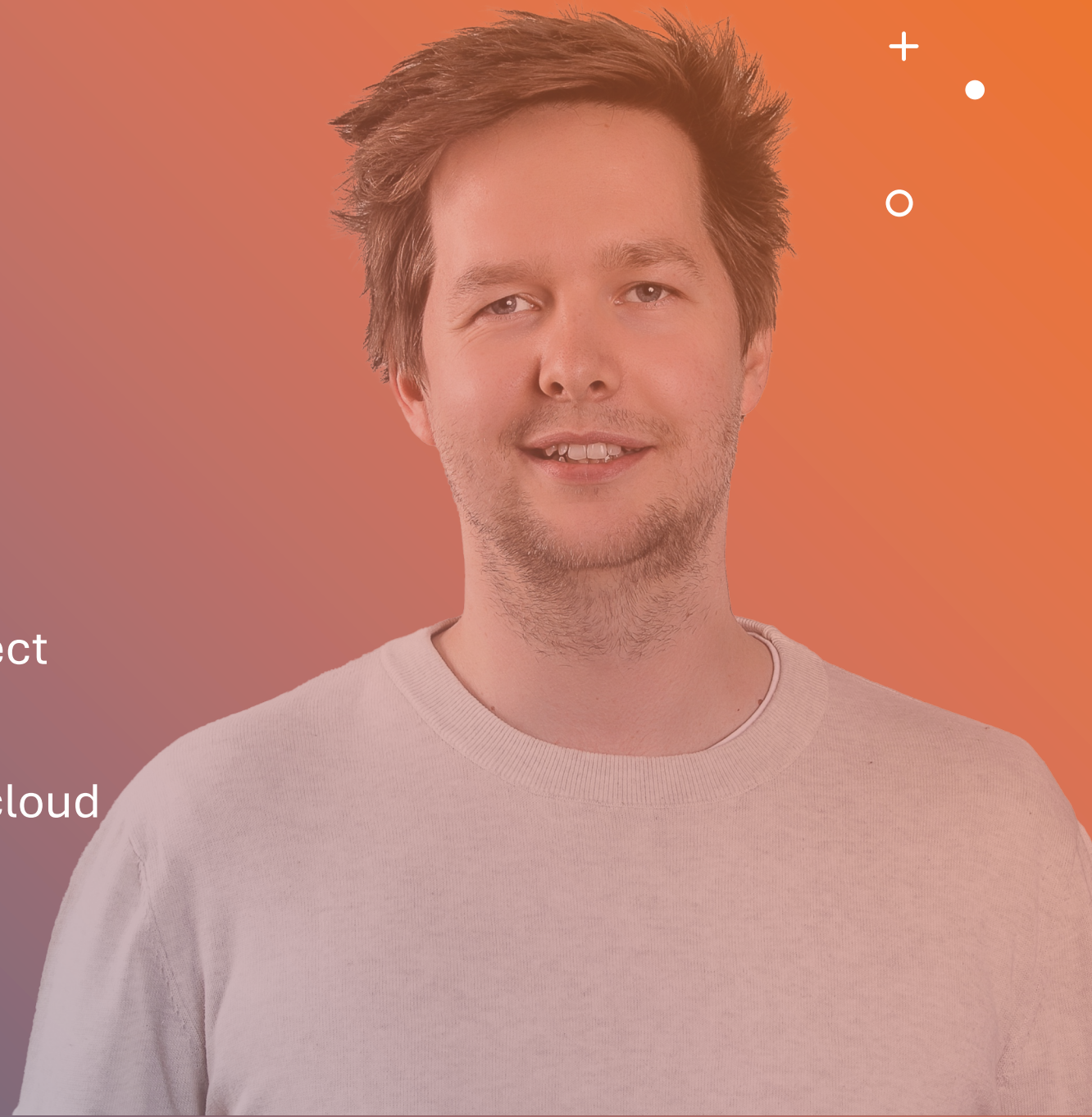
 Heist-op-den-Berg, BE

 Consultant / Data & Cloud Architect

 5 years in data

 10 years in software / architecture / cloud

 Fabric, Azure, modern data stack



What we'll talk about



Data quality

Data quality management

What is Soda?

Soda in Fabric



Why
proactive
data quality
monitoring
is important

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

[source 1](#) | [source 2](#)





NASA (1999)

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



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.

The Samsung logo, consisting of the word "SAMSUNG" in a bold, sans-serif font, is centered within a blue oval shape. The oval is tilted slightly to the right.

SAMSUNG

SECURITIES

Samsung Securities (2018)

A data entry error led to the distribution of 30 times more shares than what was available to employees.

Market value loss: \$300M



Spanish Navy (2013)

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

British Post Office / Fujitsu (1999-2015)

Accounting software Horizon,
developed by Fujitsu, reported
incorrect data.

Hundreds of postmasters were
wrongfully accused and convicted of
fraud, resulting in destroyed
livelihoods, imprisonments, suicides,
...

The logo for the British Post Office, featuring the words "POST OFFICE" in a bold, white, sans-serif font. The text is centered within a large, red, rounded rectangular shape. To the right of the top of this shape is a small red circle containing a white registered trademark symbol (®).

POST
OFFICE



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

JR's Demolition (2020)

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

Data quality vs. data quality management



+

•

○

Data quality dimensions

Accuracy: Equal to the “real-life” value

Completeness: All mandatory values present, all optional, but implied values present

Consistency: A value in one data set aligns to another data set

Currency: How “fresh” is the data compared to the real world

Precision: The level of detail in the data element

Privacy: Does the data need restricted access or monitoring?

Reasonableness: Is the current behaviour in line with previous behaviour (or average)?

Referential Integrity: All child records must have a parent

Timeliness: The difference between when the data is available and needed

Uniqueness: Are there duplicates within the data (that do not reflect reality!)?

Validity: Is the data within the allowable bounds of the “domain”?

+




•

○

What is data quality management?

“... 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-DMBOK2)

+
○ Different ways to perform data quality monitoring in Fabric

- 1) Frequent manual reviews of tables and reports
- 2) Making data pipelines fail in certain cases
- 3) Building your own framework to validate a set of tables/files
- 4)  Purview
- 5)  great expectations
- 6)  **SODA**

+

•

○

Soda concepts

Data source: a data store containing the data you want to monitor

Dataset: a table/view/relation/file in a data source

Metric: a property of the dataset

Check: measuring the value of a metric and verifying if it is within a certain threshold for the dataset

on the dataset level;

or on the level of a column;

or for a subset of rows;

or ...

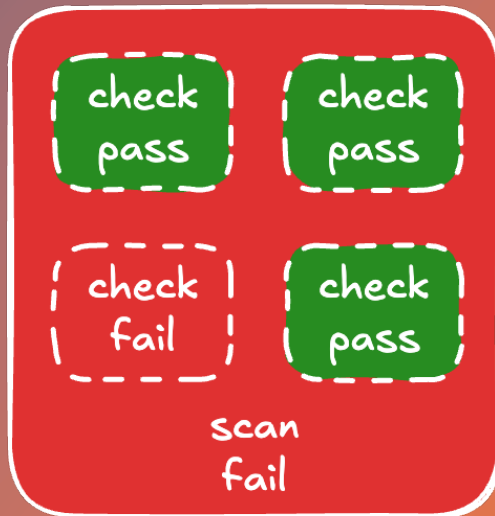
Scan: validating a data source with a selection of checks

+

•

○

Soda concepts



Running a Soda scan results in an outcome for the entire scan, based on the outcomes of the individual checks included in the scan.

The outcome can be:

- **Fail:** the metric was not within the treshold defined in the check
- **Pass:** the metric was within the treshold defined in the check
- **Warn:** an extra user-defined treshold – not as severe as a fail – was not met
- **Error:** Soda failed to execute the check

Different flavours of Soda



Soda Core

Open-source

Free

Python package and CLI



Soda Library

Closed source

Paid

Python package and CLI

Complex checks



Soda Agent

Containerized version of Soda Library meant to run checks and submit results to Soda Cloud

Pull architecture (no open ports needed)



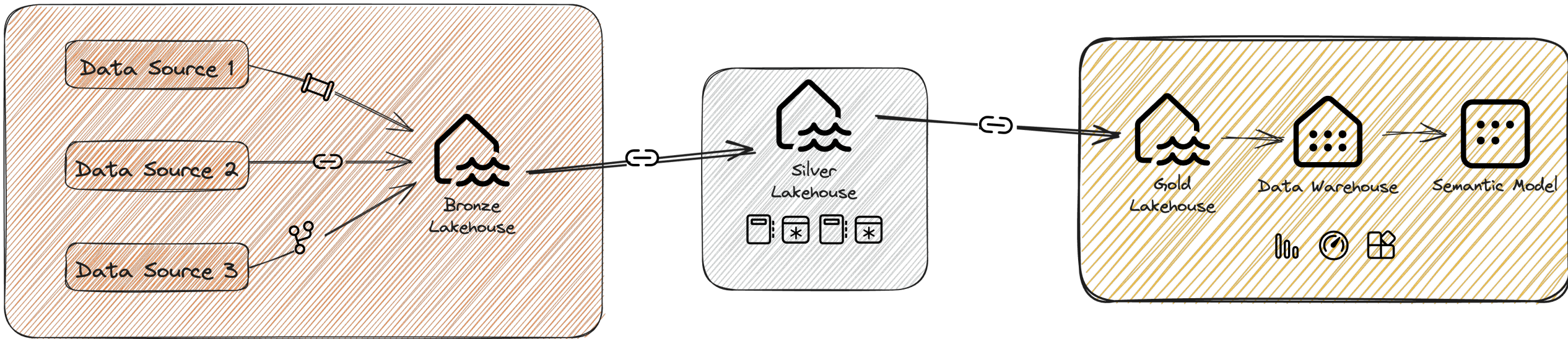
Soda Cloud

SaaS / cloud product

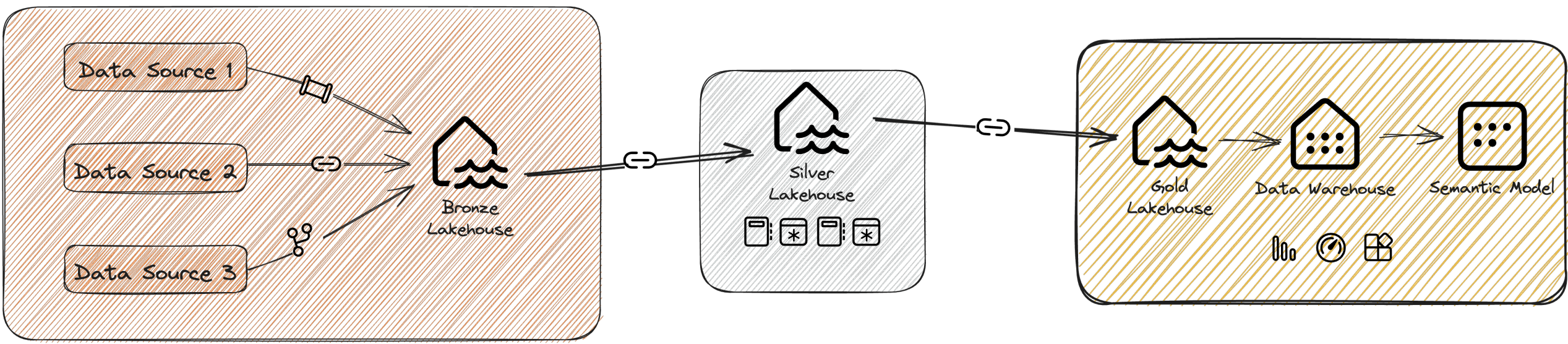
Extensive dashboard

Actionable alerting

How Soda integrates with Fabric

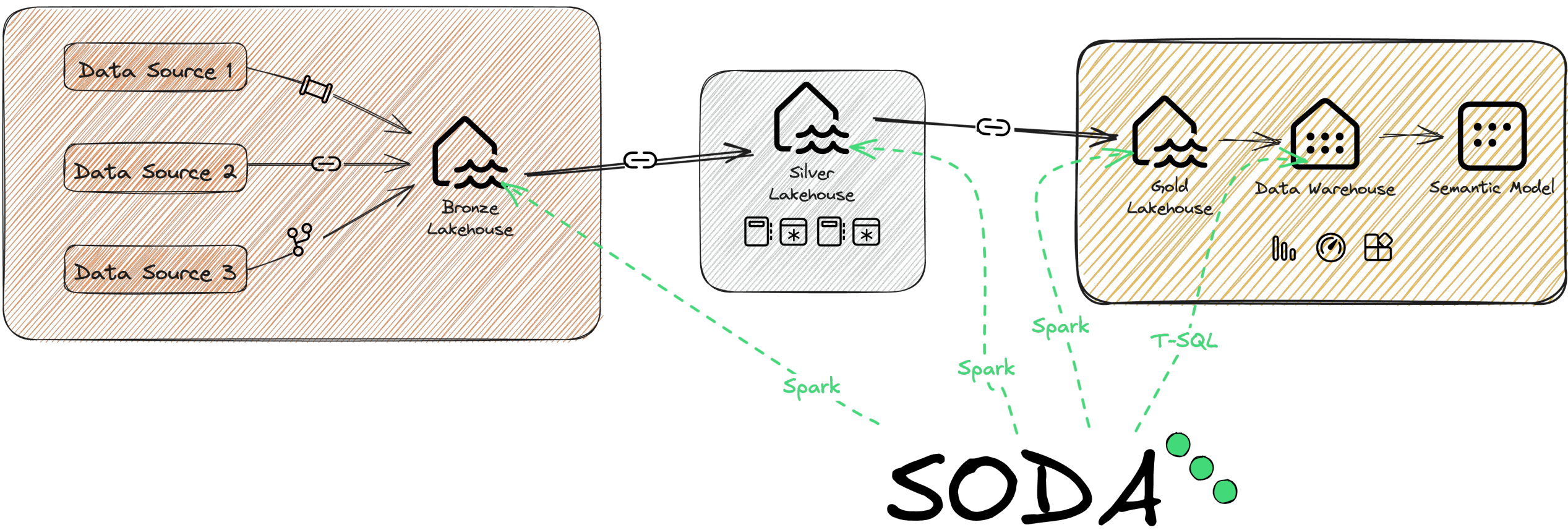


How Soda integrates with Fabric



SODA^{•••}

How Soda integrates with Fabric





Getting started

Decide how you want to monitor your data:

using Spark DataFrames (Lakehouse)

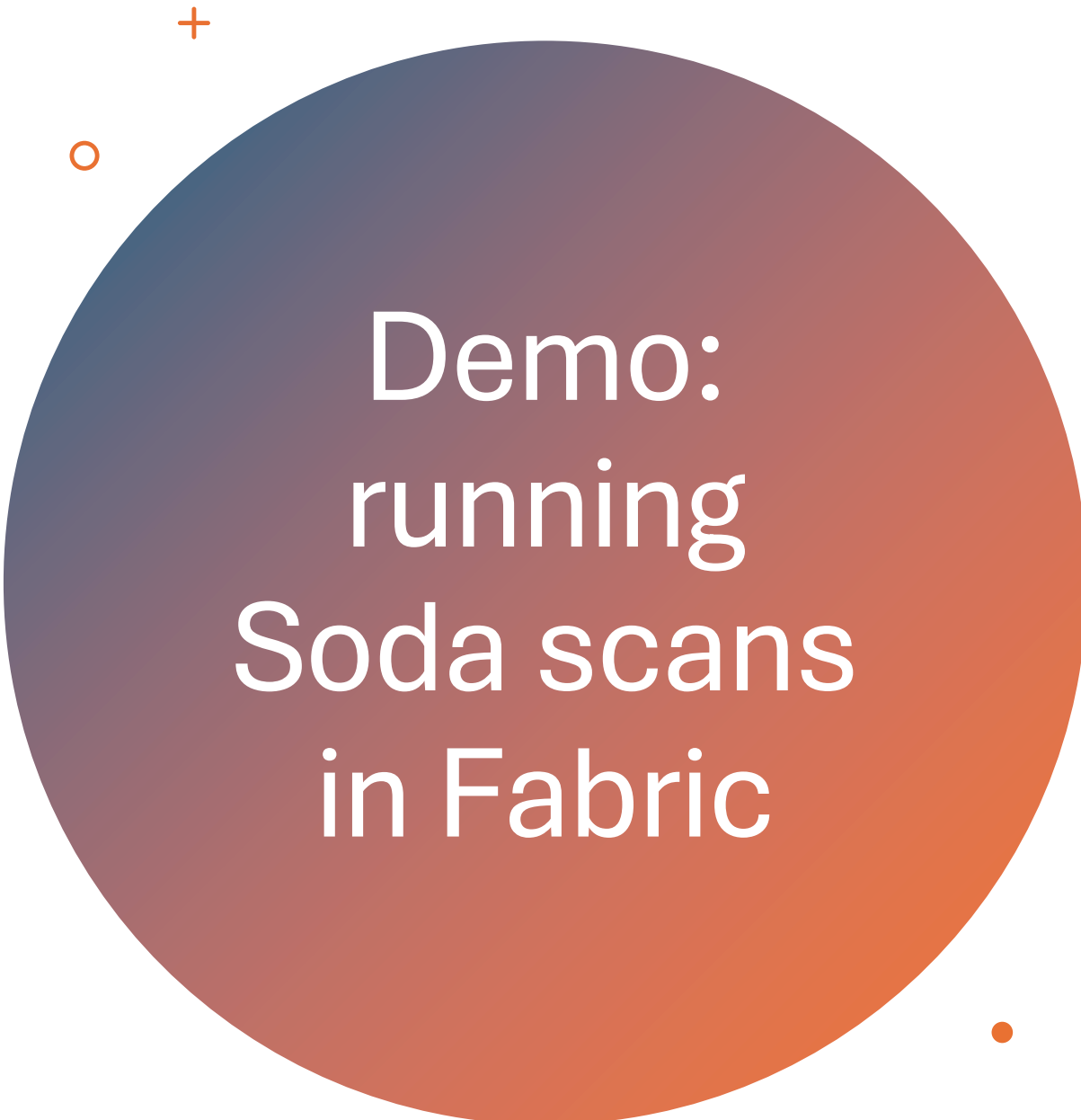
```
pip install soda-core-spark-df
```

or SQL (Data Warehouse)

```
pip install soda-core-fabric-samdebruyn
```

Write a configuration.yaml file to connect to Fabric

```
data_source adventureworks:  
  type: fabric  
  host: ...datawarehouse.fabric.microsoft.com  
  authentication: auto  
  database: adventureworks  
  schema: dbo
```



Demo:
running
Soda scans
in Fabric

Connecting Soda to a Lakehouse
or a Data Warehouse

Writing checks and storing them
in git

Integration in notebooks/Spark
Jobs / schedule Soda scans

Loading Soda results in
Lakehouse/Data Warehouse

Soda Cloud

There is more

Monitor the characteristics of a column with *distribution monitoring*

Write and verify *data contracts*

Integrate with Soda Agent (coming soon)

Data discussions

Integrate with Slack, Teams, ... for alerting

Integrate with data catalogs like Purview, Atlan, ...

Use SodaGPT/AskAI to add new checks

...





Accomplish great things

Version controlled

- Collaboration within the team

Connect to the most common data sources

- Monitor your data where it is today

Simple syntax to add new checks

- Make DQ accessible to less tech-savvy users

Store results & metrics

- Gain insight in long-term data quality

Questions?



sam@debruyne.dev

<https://debruyne.dev>

