SODA

Meet Soda and the presenters

Tom

CTO

Vijay

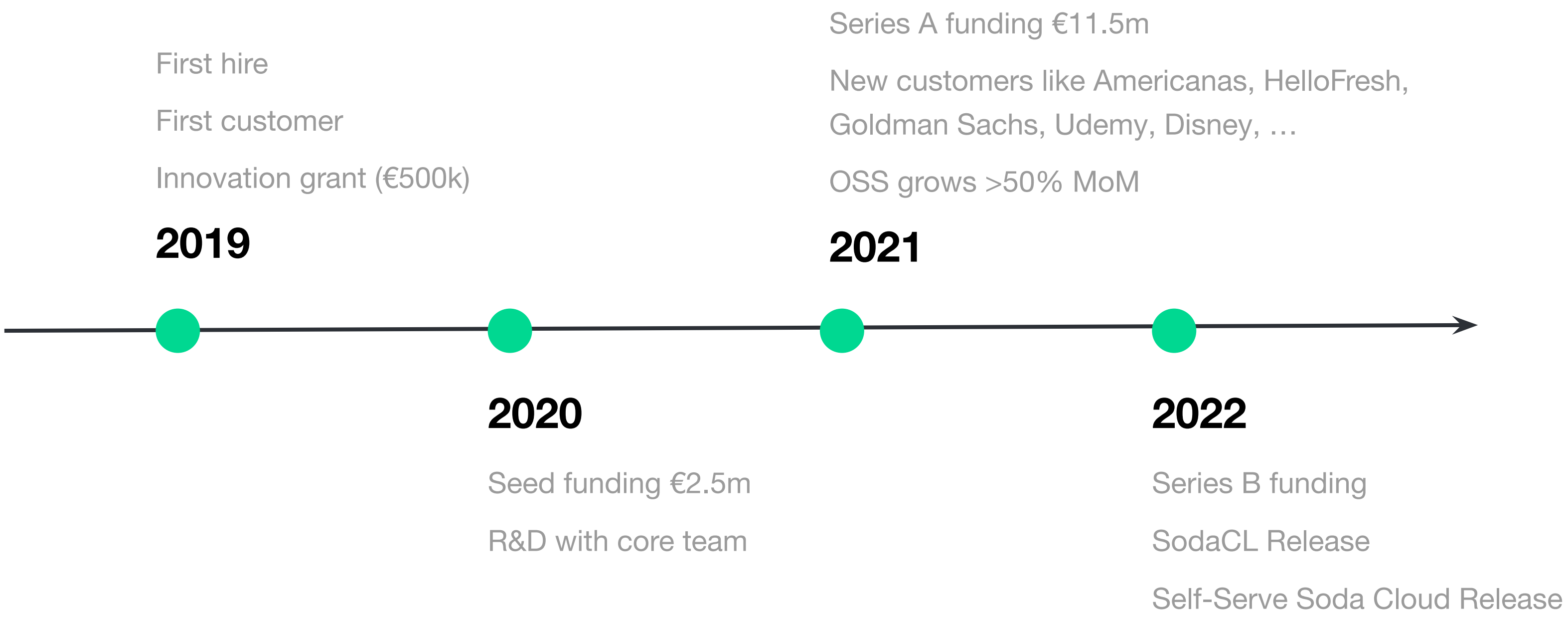
Engineering Lead

Maarten

CEO

Soda helps teams build on top of reliable, high-quality data by providing a workflow to find, analyze, and resolve data issues.

Some history



2019

First hire

First customer

Innovation grant (€500k)

2020

Seed funding €2.5m

R&D with core team

2021

Series A funding €11.5m

New customers like Americanas, HelloFresh, Goldman Sachs, Udemy, Disney, …

OSS grows >50% MoM

2022

Series B funding

SodaCL Release

Self-Serve Soda Cloud Release

How data teams ensure data is reliable, and of high quality.

Find problems automatically

Automatically monitor key patterns about your data that could lead to issues.

Align on data expectations

Configure data quality agreements to ensure that your data is fit for purpose.

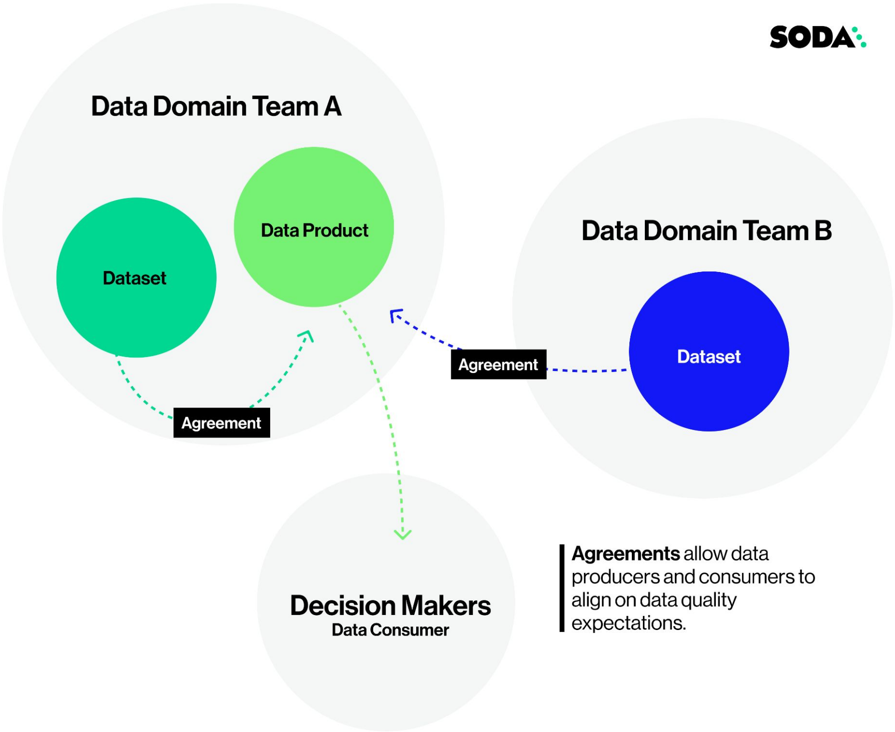
Manage & keep track of issues

Break down data silos by triaging, analysing, and resolving data incidents in your favorite tools.

Prevent issues w circuit breakers

Test data as soon as it flows into your data pipelines so you can stop bad data in its tracks.

Aligning on data expectations



Agreements allow data producers and consumers to align on data quality expectations.

SODA

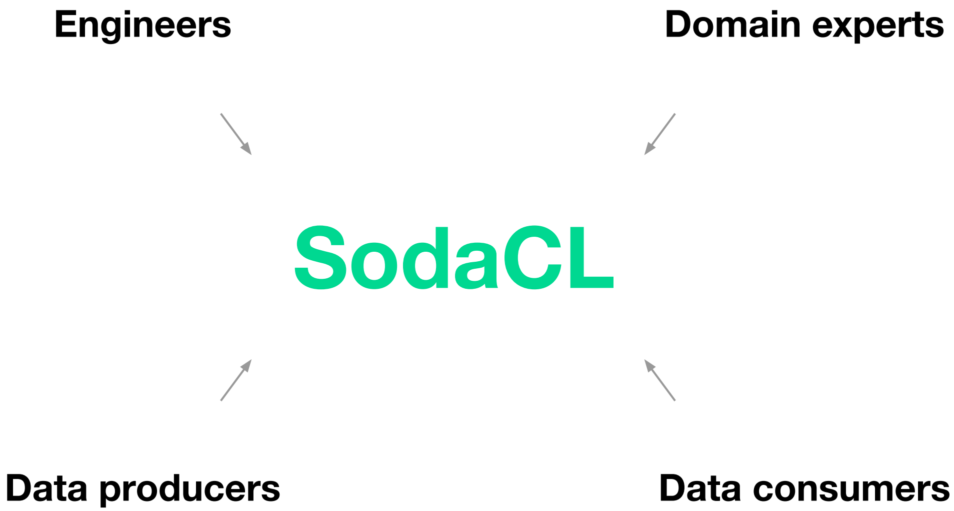
Introducing…

SodaCL

What is SodaCL?

A simple, yet powerful, declarative language for data reliability-as-code.

It is used to find problems automatically, and align data producers & consumers on data expectations.



Engineers

Domain experts

SodaCL

Data producers

Data consumers

Underserved

A language that includes data analysts

- Domain knowledge
- Manage data as a product
- Copy-n-paste from examples
- Engineering help if needed
- Towards full self-serve

Bottleneck

Advanced and flexible for data engineers

- Embedded
  - Airflow, Prefect, …
  - CLI & library
  - CI/CD
- Whitebox & control
- Many OOTB check types
- Extensible with SQL

Self-serve

All the benefits of data reliability-as-code

- Data reliability is part of the data infrastructure
- Source controlled
- Recreate setup
- Compliance
- Also includes automated monitoring and observability

Types of data issues

Data issue type	Description
Schema changes	Unexpected structural changes to data, including data types
Inconsistencies over time	Unexplainable changes relative to a previous point in time
Invalid categorical values	Out-of-domain values or unstandardized values
Missing rows or values	Partially or completely missing rows or values, including null values
Inconsistent data	Data doesn’t match an authoritative or alternative source
Duplicate keys	Data was duplicated as part of the transformation process
Untimely data	Data that didn’t refresh or was not delivered on time
Concept drift	Statistical properties of a variable change too much over time
Insufficient coverage	The number of rows containing a value is too small

Objective of SodaCL:

Make data reliability a ubiquitous part of the data stack:

- Run anywhere: from ingestion to consumption
- Treat data reliability as infrastructure, hence as code
- Simple language support all types of checks OOTB
- Thriving user community

SodaCL Feature Highlights

Many check types

Use declarative test types to check for common types of data issues.

Historical checks

Check for changes over time. Leverage Soda Cloud time-series storage with a simple configuration.

Anomaly detection

No need to set hard boundaries. Detect anomalies in time-series data.

Cross-warehouse checks

Run scans cross multiple data sources to compare row counts, for example.

Partitions

Create time windows for consistency-over-time checks, or to filter for only newest records.

For each checks

Apply checks to hundreds of thousands of tables or columns in one go.

SODA

SodaCL Showcase

SodaCL

Showcase

1. Set up the prepared environment, including a demo database.
2. Use SodaCL to write a few checks for data quality in a dataset.
3. Run a scan of the data to execute the checks and find out which data is good and which is bad.
4. Tell us what you think! Honest feedback is a gift.

Next Steps

Next steps

- Access GitHub Repo: https://github.com/sodadata/sodacl-workshop
- Join the private #sodacl-preview-program channel on Slack.
- Follow the included instructions to experiment with SodaCL as you write checks that test data in the demo database for quality.
- Record your feedback on your experience with SodaCL.
- February 28 to March 4: Hands-on testing and feedback and listening sessions
- March 7: Wrap-up party!

Thank you.

SODA

Vijay Kiran
Head of Data Engineering
@vijaykiran
vijay@soda.io