SODA

Data is the center of every digital experience.

In order to excel in digital, companies need to manage these 4 components through people, process & systems:

Infrastructure Monitoring

IT / Infrastructure: “DevOps”

Platform: Data Dog

+

Data Monitoring

Data Owners & Engineers: “DataOps”

Platform: Soda

+

Model Monitoring

ML & Data Engineers: “MLOps”

Platforms: Dataiku

+

Application Monitoring

Software Development: “DevOps”

Platforms: New Relic

What could possibly go wrong?

Definitely. Many issues are actually silent (go unnoticed), until someone shouts from the rooftops.

- **Data ownership**

  Use of poorly architected systems

  Data contracts / SLAs not met

  Lack of process enforcement

- **Data engineering**

  Software/firmware changes

  (Partially) missing data

  Schema evolution

- **Data analytics**

  Errors in interpretation

  Changing data requirements

  Annotation distribution

⬇ Customer experience

⬆ Reputation risk

⬆ Operational inefficiencies

© Soda Data. Proprietary and Confidential.

Solution: Data Monitoring-as-a-Service

1 **Automatically instrument your data to create visibility & transparency at scale.**

ML-driven analysis of data at rest and in-motion, whether it’s in your data warehouse, lake or elsewhere. Analytics teams can layer-in their data requirements and track data usage SLAs.

2 **Analyse and collaborate on meaningful alerts using rich context.**

Get alerts based on your role and inspect those using rich diagnostics and contextual info. Involve SMEs, find the appropriate resolution, and escalate problems where needed.

3 **Proactively share the status of datasets via the Data Quality Catalog.**

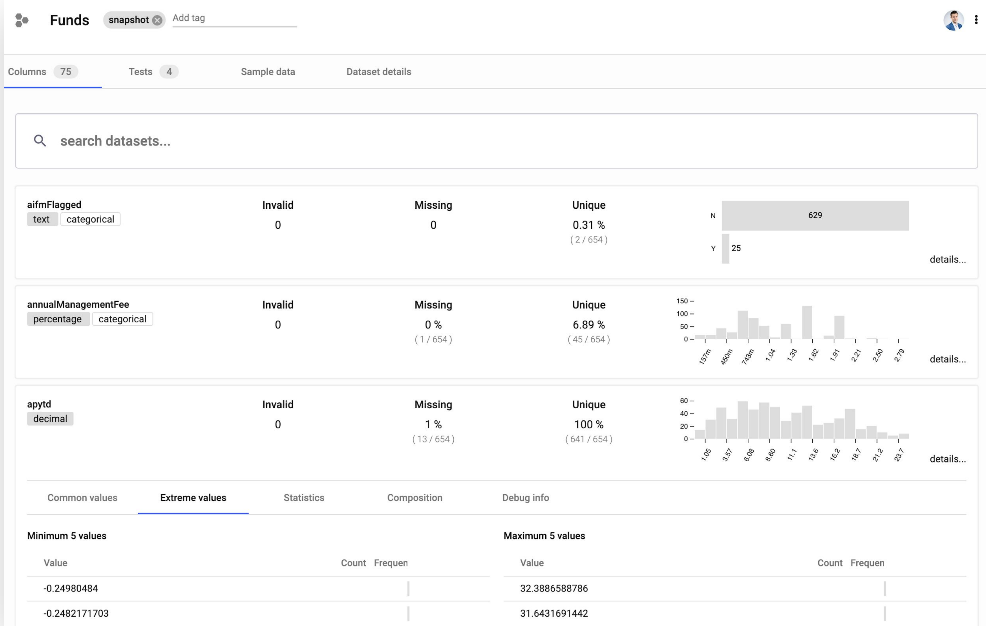
Give consumers of data products real-time insight into the quality of the data. Publish data quality indicators where data is searched and consumed to help increase trust in data.

Feature Highlights

SODA

Soda automates data profiling for Data Quality

When new data comes in, Soda automatically re-runs profiling and analyzes the results historically. The approx. 50 out-of-the-box profiling metrics can be extended by adding custom metrics or test definitions.



	Invalid	Missing	Unique
aifmFlagged text categorical	0	0	0.31 % (2 / 654)
annualManagementFee percentage categorical	0	0 % (1 / 654)	6.89 % (45 / 654)
apytd decimal	0	1 % (13 / 654)	100 % (641 / 654)

Minimum 5 values

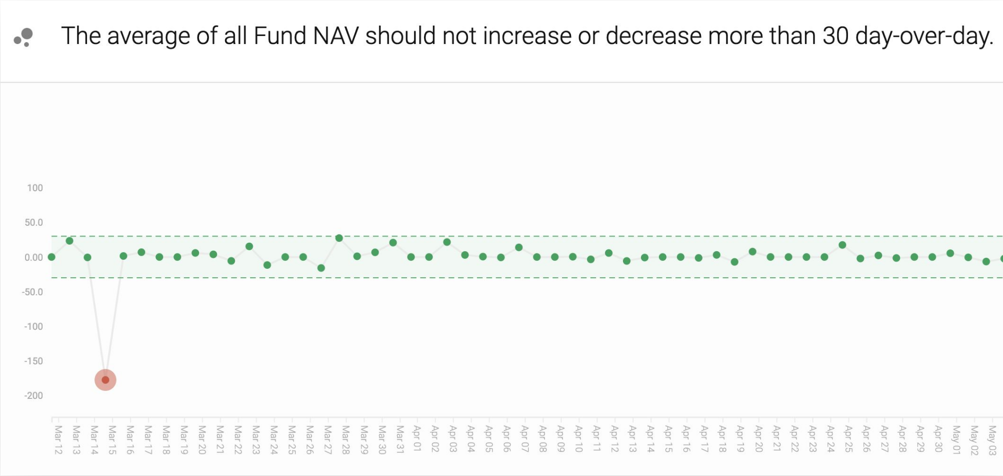
Value	Count	Frequen
-0.24980484		
-0.2482171703		

Maximum 5 values

Value	Count	Frequen
32.3886588786		
31.6431691442		

Soda finds and analyzes issues with your data.

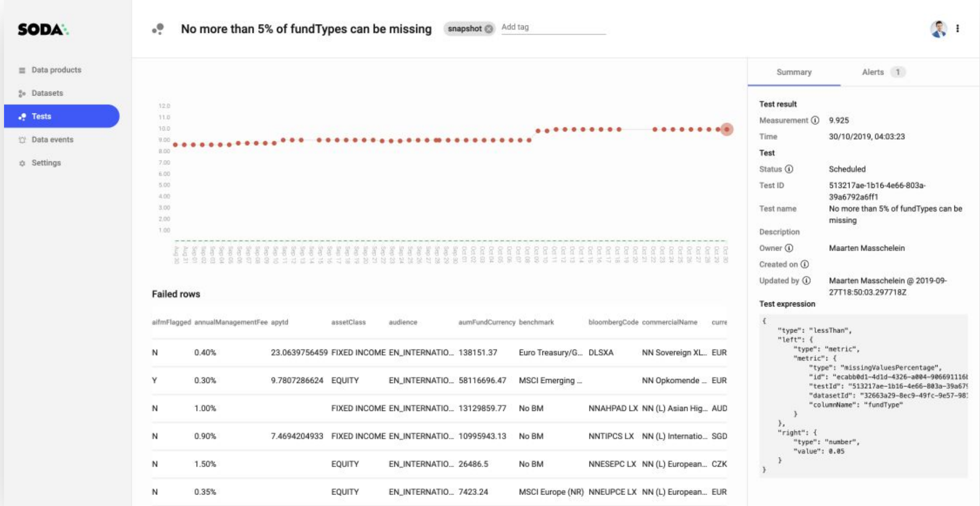
Soda automatically measures key characteristics of your data over time in search of anomalies. Layer in business logic (via self-service) to get a robust operating model for data quality in place.



The average of all Fund NAV should not increase or decrease more than 30 day-over-day.

Route non-compliant records to the right people

Every metric Soda monitors contains details on what was calculated. If you're validating record-by-record, these details include all non-compliant records.



No more than 5% of fundTypes can be missing

Failed rows

aifmFlagged	annualManagementFee	apytd	assetClass	audience	aumFundCurrency	benchmark	bloombergCode	commercialName	curre
N	0.40%	23.0639756459	FIXED INCOME	EN_INTERNATIO...	138151.37	Euro Treasury/G...	DLSXA	NN Sovereign XL...	EUR
Y	0.30%	9.7807286624	EQUITY	EN_INTERNATIO...	58116696.47	MSCI Emerging ...		NN Opkomende ...	EUR
N	1.00%		FIXED INCOME	EN_INTERNATIO...	13129859.77	No BM	NNAHPAD LX	NN (L) Asian Hig...	AUD
N	0.90%	7.4694204933	FIXED INCOME	EN_INTERNATIO...	10995943.13	No BM	NNTIPCS LX	NN (L) Internatio...	SGD
N	1.50%		EQUITY	EN_INTERNATIO...	26486.5	No BM	NNESEPC LX	NN (L) European...	CZK
N	0.35%		EQUITY	EN_INTERNATIO...	7423.24	MSCI Europe (NR)	NNEUPCE LX	NN (L) European...	EUR

Test result

Measurement ⓘ	9.925
Time	30/10/2019, 04:03:23

Test

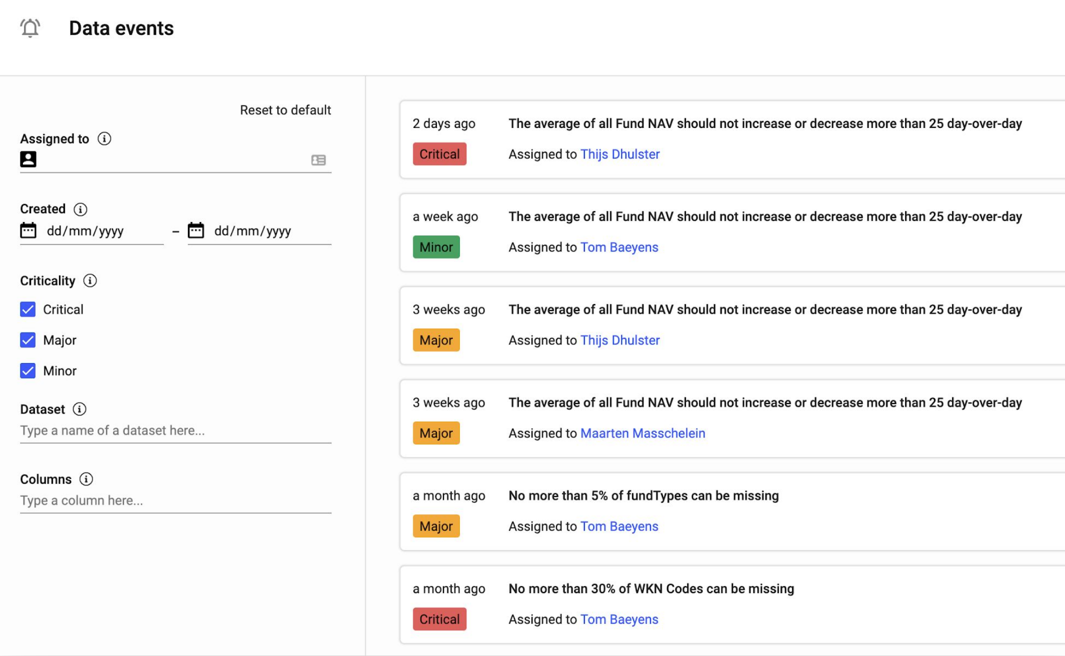
Status ⓘ	Scheduled
Test ID	513217ae-1b16-4e66-803a-39a6792a6ff1
Test name	No more than 5% of fundTypes can be missing
Description	
Owner ⓘ	Maarten Masschelein
Created on ⓘ	
Updated by ⓘ	Maarten Masschelein @ 2019-09-27T18:50:03.297718Z

Test expression

```
{
    "type": "lessThan",
    "left": {
        "type": "metric",
        "metric": {
            "type": "missingValuesPercentage",
            "id": "ecabb0d1-4d1d-4326-a004-906691116
            "testId": "513217ae-1b16-4e66-803a-39a67
            "datasetId": "32663a29-8ec9-49fc-9e57-98
            "columnName": "fundType"
        }
    },
    "right": {
        "type": "number",
        "value": 0.05
    }
}
```

Soda helps you prioritize & analyze alerts

Soda helps you understand which alerts are important to solve first by analyzing how many data products will be impacted and what the criticality of those are.



Data events

Reset to default

Assigned to

Created
dd/mm/yyyy – dd/mm/yyyy

Criticality
☑ Critical
☑ Major
☑ Minor

Dataset
Type a name of a dataset here...

Columns
Type a column here...

2 days ago | Critical | The average of all Fund NAV should not increase or decrease more than 25 day-over-day | Assigned to Thijs Dhulster

a week ago | Minor | The average of all Fund NAV should not increase or decrease more than 25 day-over-day | Assigned to Tom Baeyens

3 weeks ago | Major | The average of all Fund NAV should not increase or decrease more than 25 day-over-day | Assigned to Thijs Dhulster

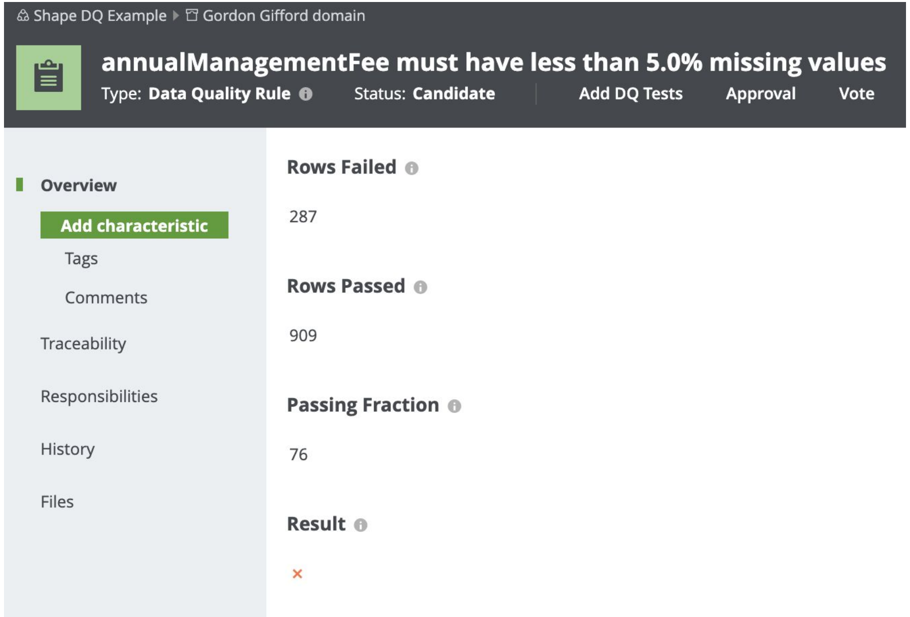
3 weeks ago | Major | The average of all Fund NAV should not increase or decrease more than 25 day-over-day | Assigned to Maarten Masschelein

a month ago | Major | No more than 5% of fundTypes can be missing | Assigned to Tom Baeyens

a month ago | Critical | No more than 30% of WKN Codes can be missing | Assigned to Tom Baeyens

Soda integrates natively with data catalogs

Whether you use Alation, Colliba, Hive Metastore, or Amundsen, we natively integrate into your catalog to help provide operational context (real-time intelligence) to your datasets. This way, users can truly trust the data they use or produce.



annualManagementFee must have less than 5.0% missing values

Type: Data Quality Rule

Status: Candidate

Rows Failed

287

Rows Passed

909

Passing Fraction

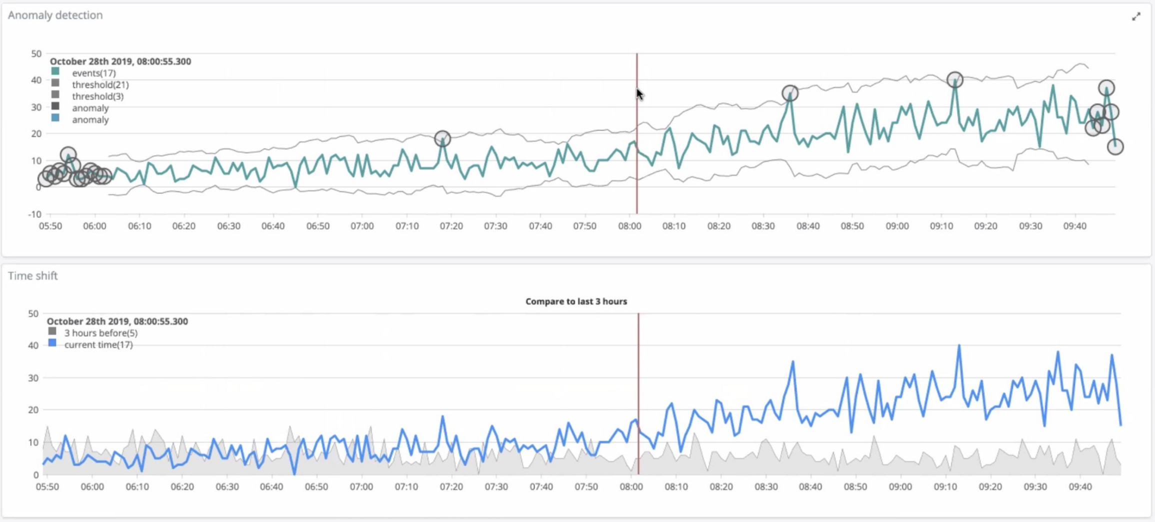
76

Result

×

Soda uses ML based anomaly detection for DQ

Managing rules on data is unwieldy at scale. This is where Machine Learning comes in. Soda look at outliers in your data delivery process (when does data arrive, how does it look, ...).



Architecture

01

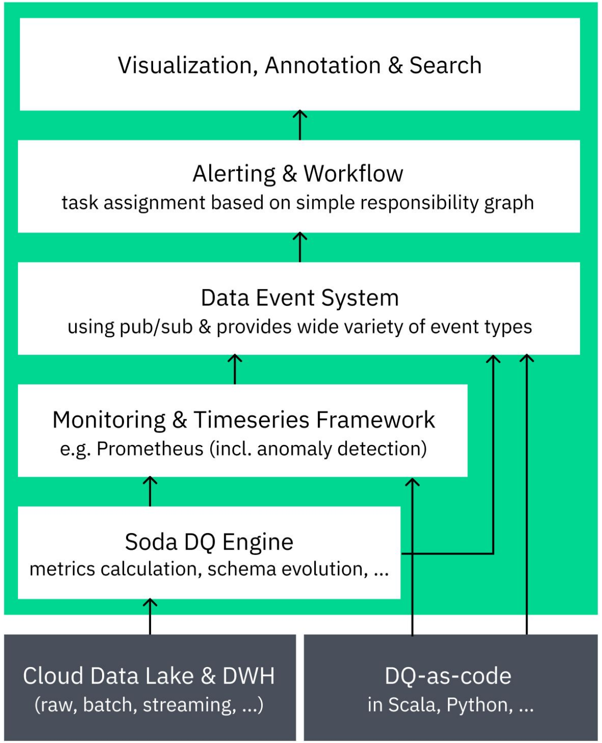
Soda Compute Engine

Big data engine calculates a wide variety of base metrics and allows business users to define & measure custom metrics that will be evaluated as new data comes in (and where possible also historically).

E.g. “Percentage of valid values”
= (countNulls(ds1.c1) + customValueCount(“N/A”) + outOfDomainValueCount(ds1.c1, ds2.c1)) / rowCount(ds1)

Soda supports the creation of metrics on filtered datasets as well.

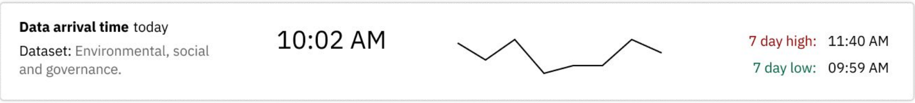
Soda Solution Architecture


Visualization, Annotation & Search
Alerting & Workflow
task assignment based on simple responsibility graph
Data Event System
using pub/sub & provides wide variety of event types
Monitoring & Timeseries Framework
e.g. Prometheus (incl. anomaly detection)
Soda DQ Engine
metrics calculation, schema evolution, ...
Cloud Data Lake & DWH
(raw, batch, streaming, ...)
DQ-as-code
in Scala, Python, ...

02

Monitoring & Timeseries Framework

Leverages open-source metrics framework like Prometheus & provides anomaly detection tools. Externally managed metrics can also be pushed in this framework (data quality as code). Add DQ specific functionality to show e.g. backfills, patched data, ...



Data arrival time today

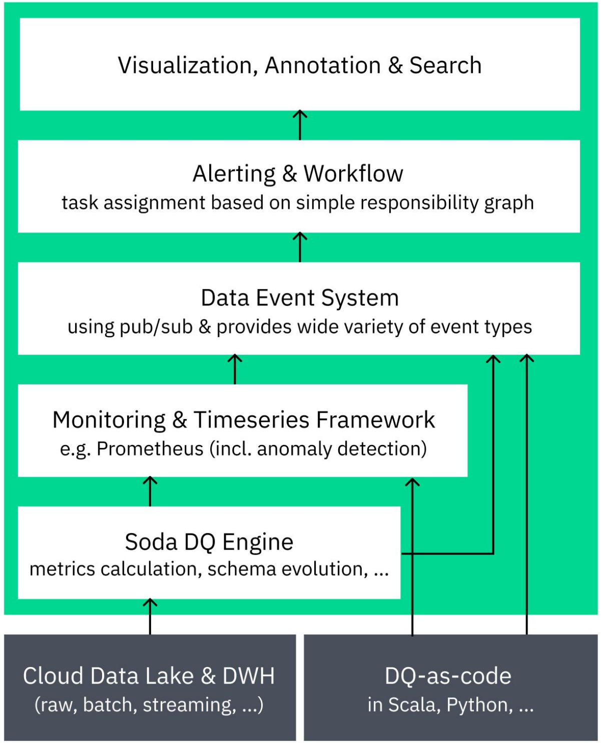
Dataset: Environmental, social and governance.

10:02 AM

7 day high: 11:40 AM

7 day low: 09:59 AM

Soda Solution Architecture



Visualization, Annotation & Search

Alerting & Workflow
task assignment based on simple responsibility graph

Data Event System
using pub/sub & provides wide variety of event types

Monitoring & Timeseries Framework
e.g. Prometheus (incl. anomaly detection)

Soda DQ Engine
metrics calculation, schema evolution, ...

Cloud Data Lake & DWH
(raw, batch, streaming, ...)

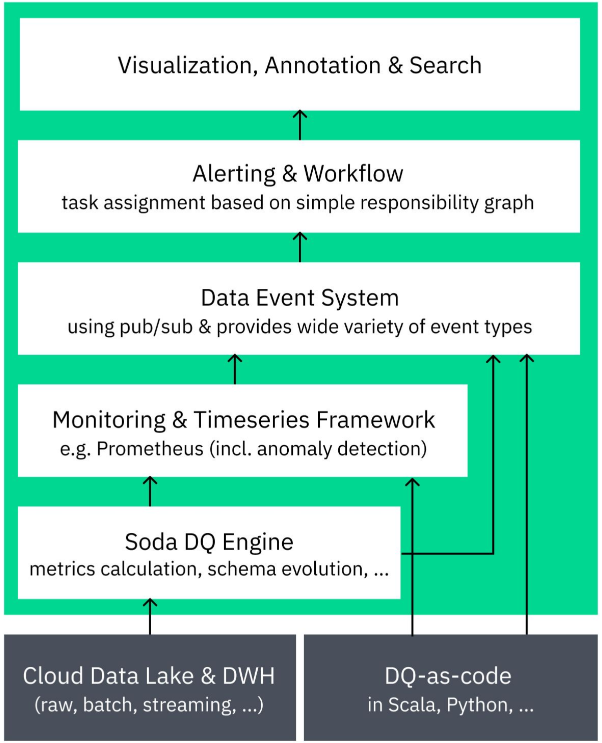
DQ-as-code
in Scala, Python, ...

03

Data Event System

Events are created from (i) metrics that have a measurement that exceeds a thresholds, (ii) schema service or (iii) external code. Data consumers can subscribe to datasets to receive alerts. Events are displayed in an event stream that can be filtered and aggregated at different levels (e.g. dataset, user, ...).

Soda Solution Architecture

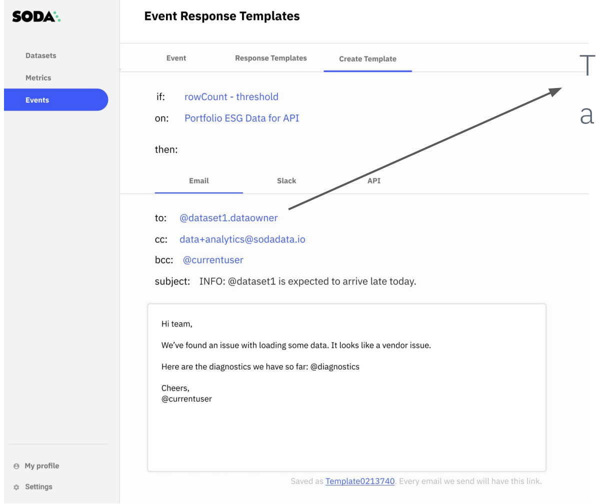

Visualization, Annotation & Search
Alerting & Workflow
task assignment based on simple responsibility graph
Data Event System
using pub/sub & provides wide variety of event types
Monitoring & Timeseries Framework
e.g. Prometheus (incl. anomaly detection)
Soda DQ Engine
metrics calculation, schema evolution, ...
Cloud Data Lake & DWH
(raw, batch, streaming, ...)
DQ-as-code
in Scala, Python, ...

© Soda Data. Proprietary and Confidential.

04

Alerting & Workflow

Provide, easy yet flexible notification & task mgmt.



if: rowCount - threshold

on: Portfolio ESG Data for API

then:

to: @dataset1.dataowner

cc: data+analytics@sodadata.io

bcc: @currentuser

subject: INFO: @dataset1 is expected to arrive late today.

Hi team,

We've found an issue with loading some data. It looks like a vendor issue.

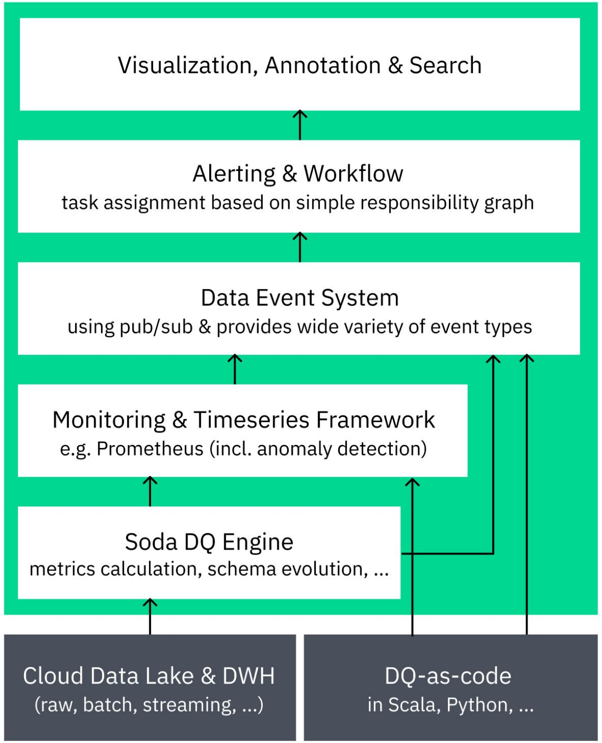
Here are the diagnostics we have so far: @diagnostics

Cheers,
@currentuser

Saved as Template0213740. Every email we send will have this link.

Traverse the graph to add the users you want

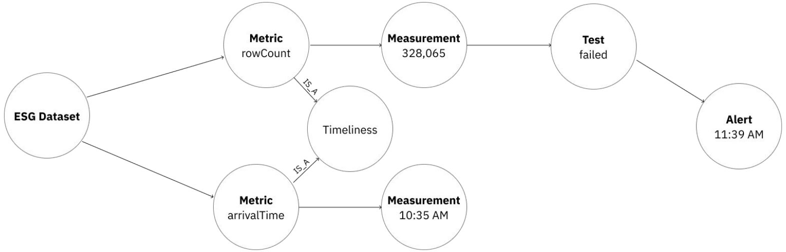
Soda Solution Architecture


Visualization, Annotation & Search
Alerting & Workflow
task assignment based on simple responsibility graph
Data Event System
using pub/sub & provides wide variety of event types
Monitoring & Timeseries Framework
e.g. Prometheus (incl. anomaly detection)
Soda DQ Engine
metrics calculation, schema evolution, ...
Cloud Data Lake & DWH
(raw, batch, streaming, ...)
DQ-as-code
in Scala, Python, ...

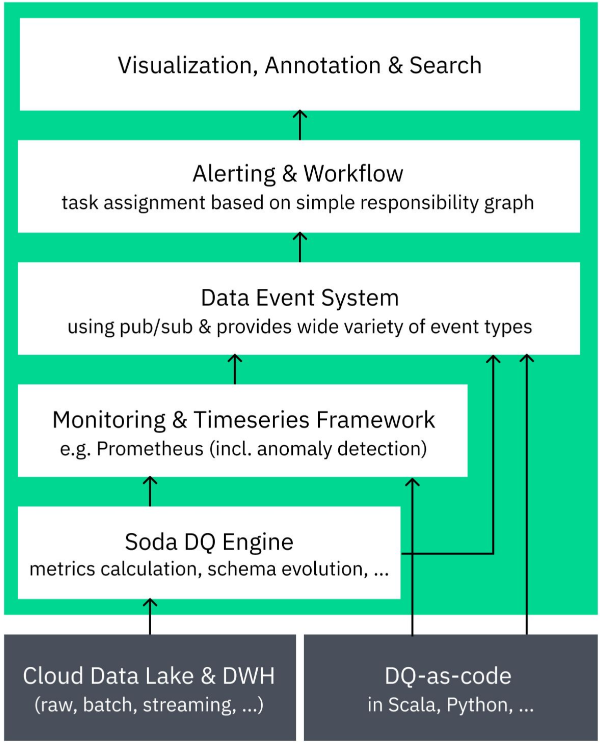
05

Visualization, Annotation & Search

Search & correlate events to find the root cause. Provide out-of-the-box aggregation dashboards per persona or connect to third party tools (Tableau, ...).



Soda Solution Architecture



Visualization, Annotation & Search

Alerting & Workflow
task assignment based on simple responsibility graph

Data Event System
using pub/sub & provides wide variety of event types

Monitoring & Timeseries Framework
e.g. Prometheus (incl. anomaly detection)

Soda DQ Engine
metrics calculation, schema evolution, ...

Cloud Data Lake & DWH
(raw, batch, streaming, ...)

DQ-as-code
in Scala, Python, ...

Thank you.

SODA

Maarten Masschelein
Founder & CEO
(929) 920-1414
maarten@sodadata.io