SPARK+AI SUMMIT EUROPE

Lessons Learned Developing and Managing High Volume Apache Spark Pipelines in Production

Erni Durdevic, Quby B.V.

#SAISML4

Over 350,000 connected homes across Europe

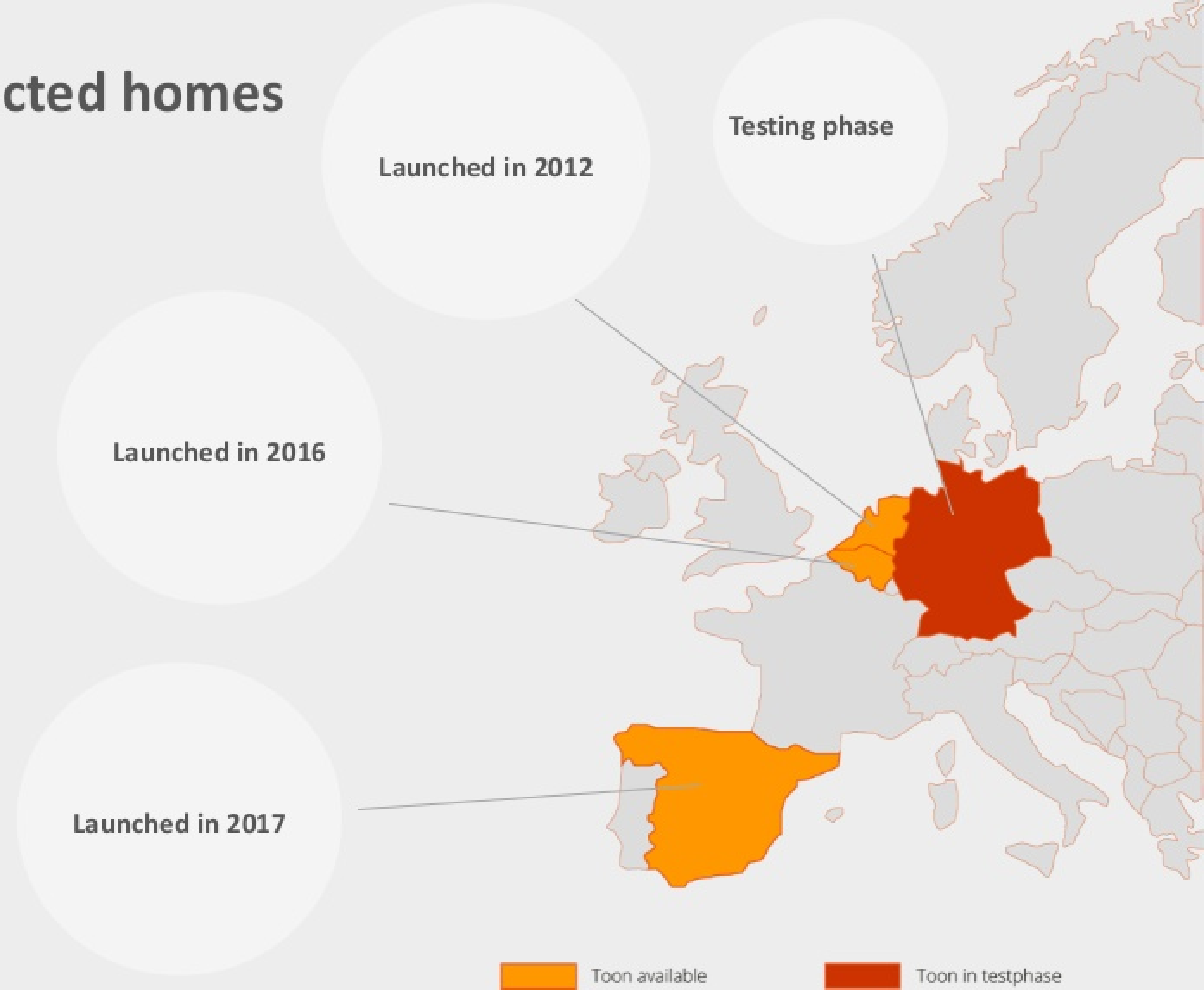
Our partners:

Eneco

Interpolis.

engie

viesgo



TOON

ENERGY INSIGHTS APP

SMART METER DONGLE & APP

SMART THERMOSTAT & APP

SECURITY PACKAGE & APP

DATA SERVICES

WASTE CHECKER

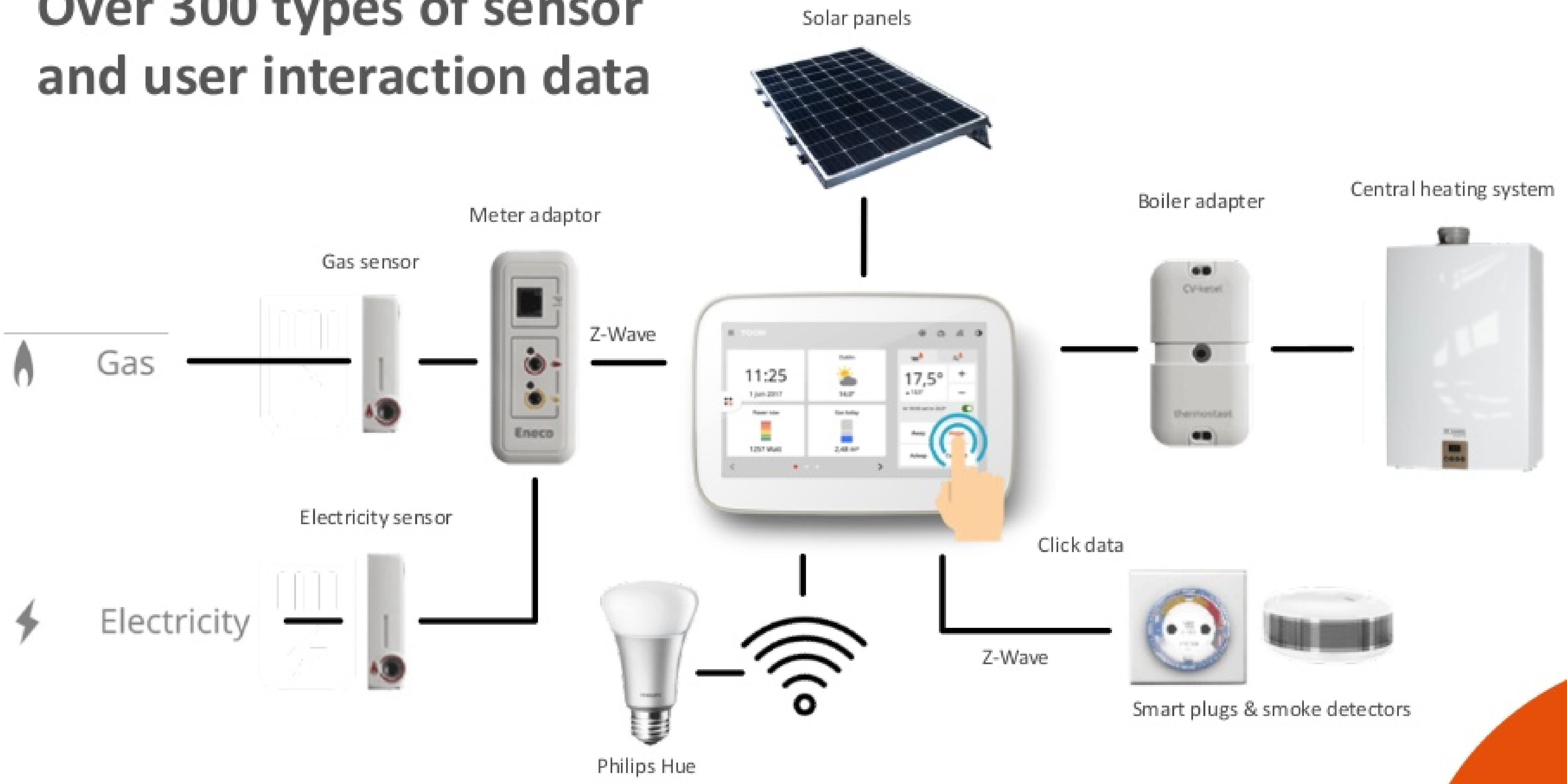
WATER INSIGHT

MONTHLY ENERGY INSIGHT

TOON SOLAR

BOILER MONITORING

Over 300 types of sensor and user interaction data

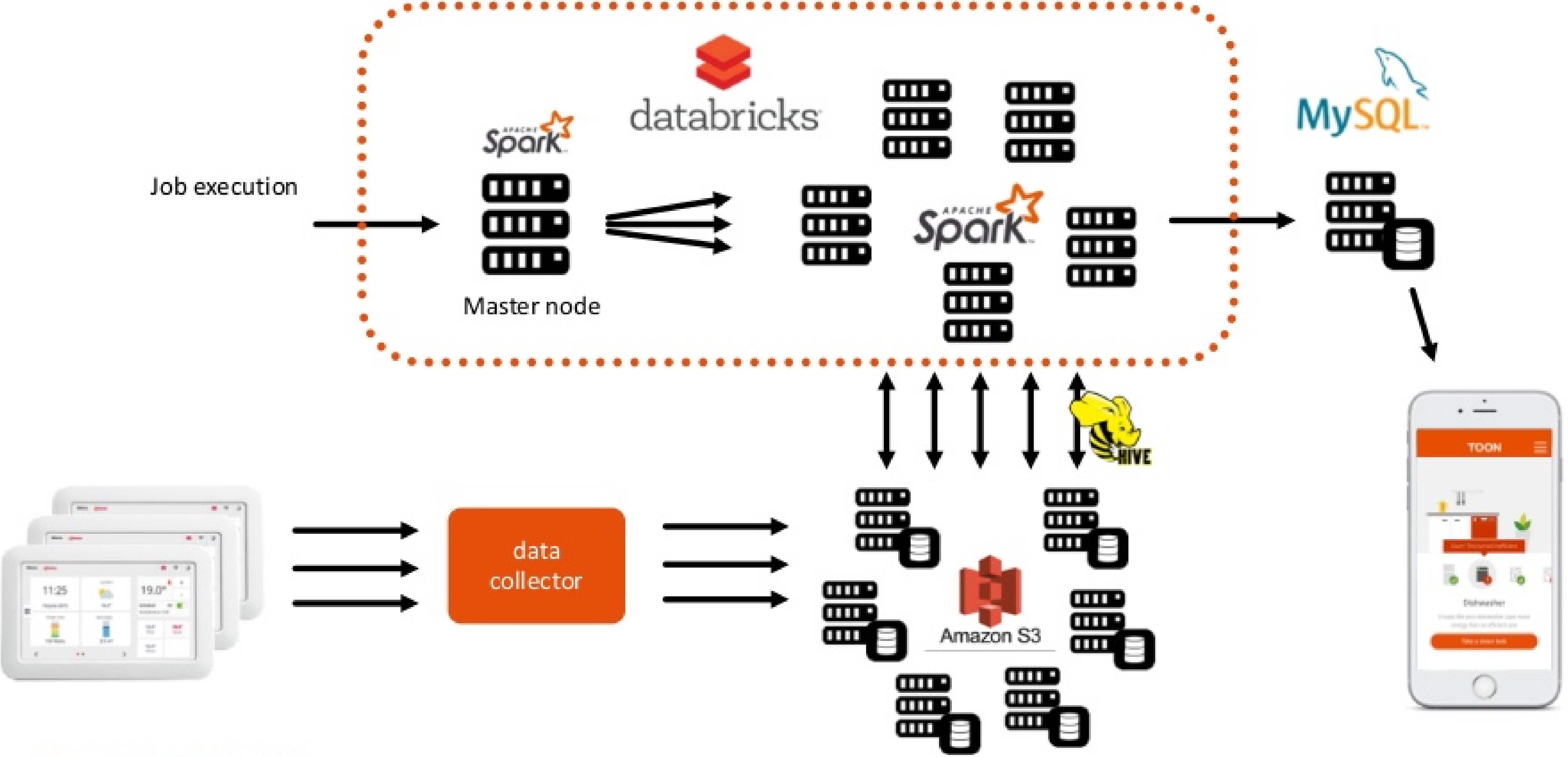


>1200 TB in total

TOON®

#SAISML4

Batch pipeline



TOON®

#SAISML4

USE CASE #1

Waste Checker

TOON®

Energy Waste Checker

"We don't always notice how much energy we're wasting. Toon can now expose the energy guzzlers in your home."

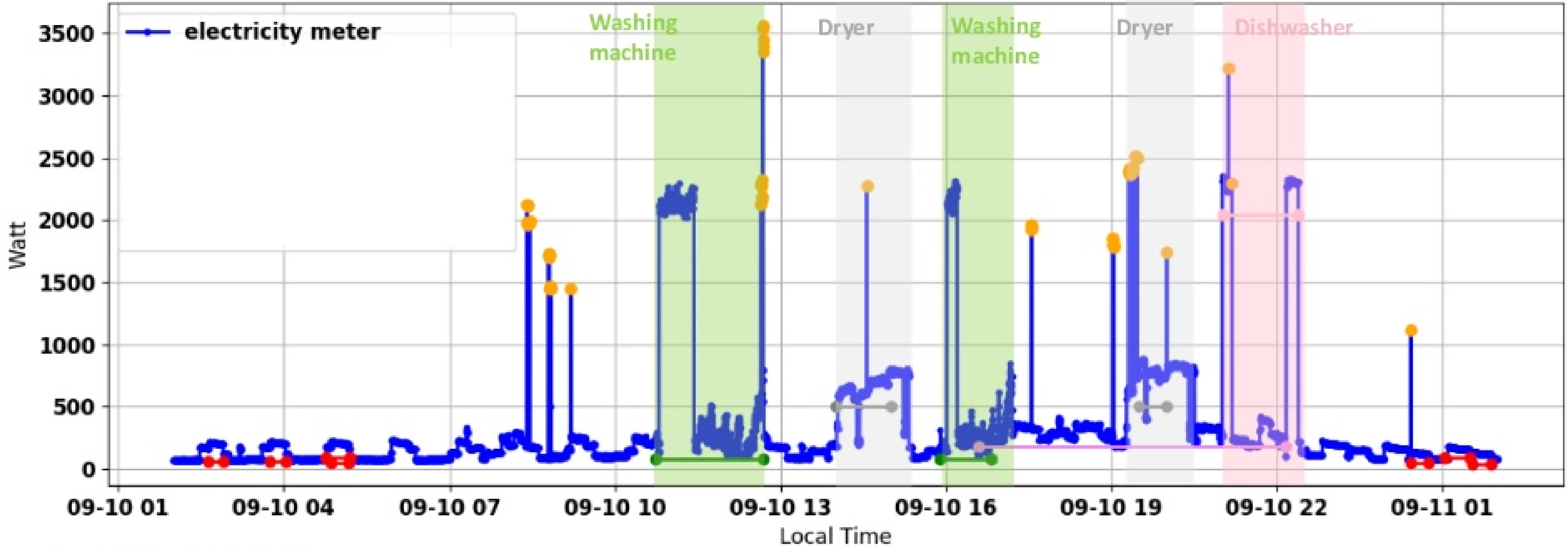
Launched in December 2017 to all Eneco Toon users

TOON®

#SAISML4

Quby's disaggregation algorithms

Patent pending algorithms can detect appliances from 10 second resolution electricity meter data

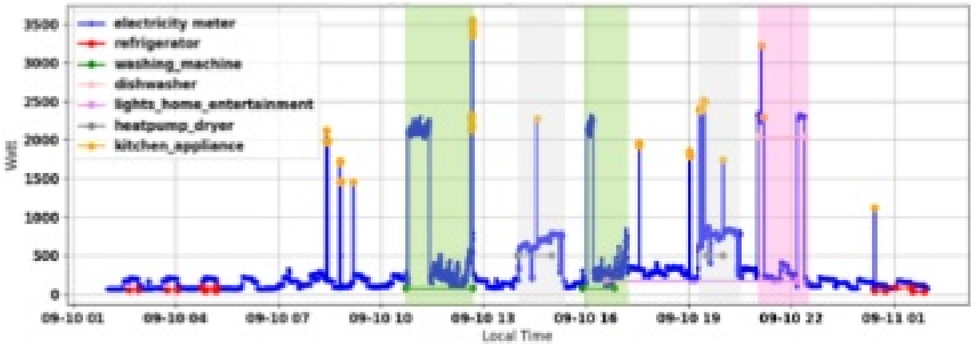


TOON®

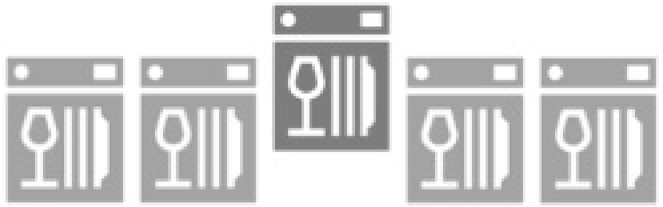
#SAISML4

Use case example: Inefficient dishwasher diagnosis

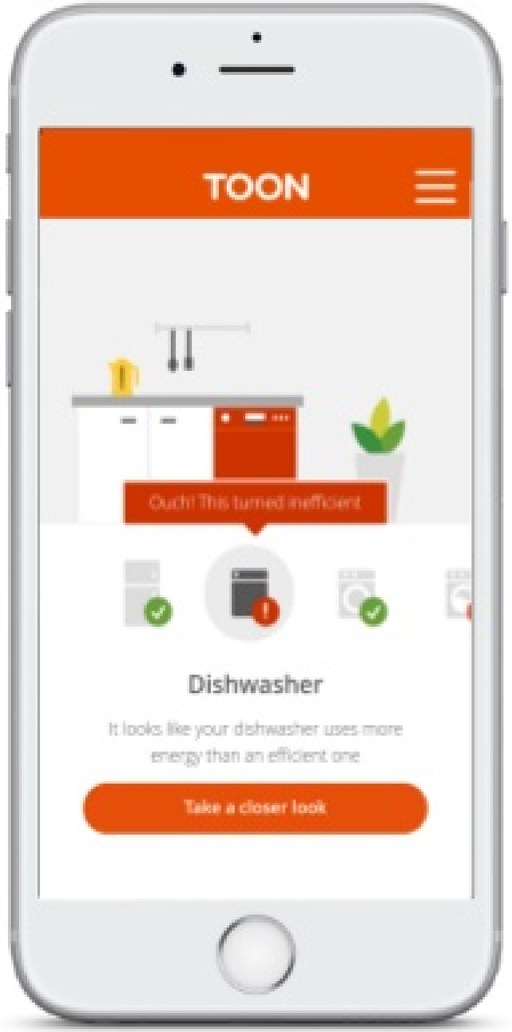
Disaggregation algorithms run on the 10s electricity meter data



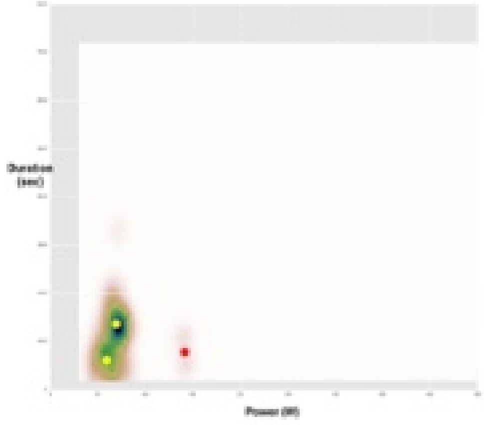
Compared with industry standards and peers



Translated to personalised advice for the end user



Toon determines the “fingerprint” of the appliance through features



TOON®

#SAISML4

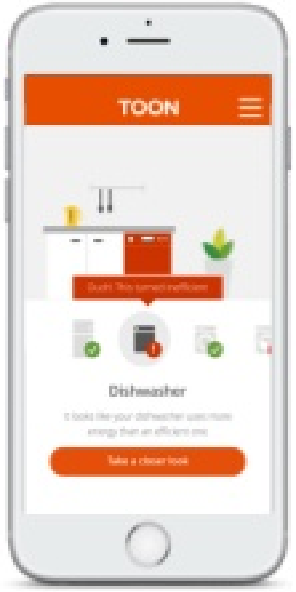
Scale of the Waste checker

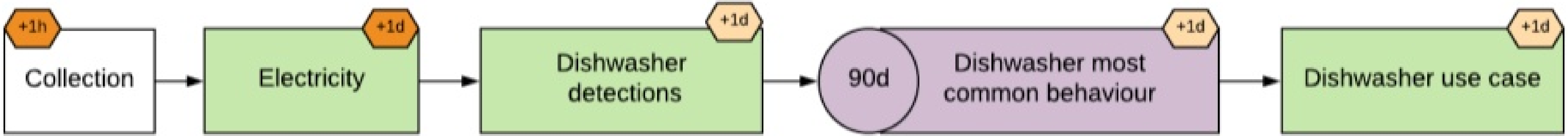
Each day we detect over
75,000 dishwasher
cycles

That's
13 years
of dishwashers running
continuously

and over
25% are
used
inefficiently

Example data flow

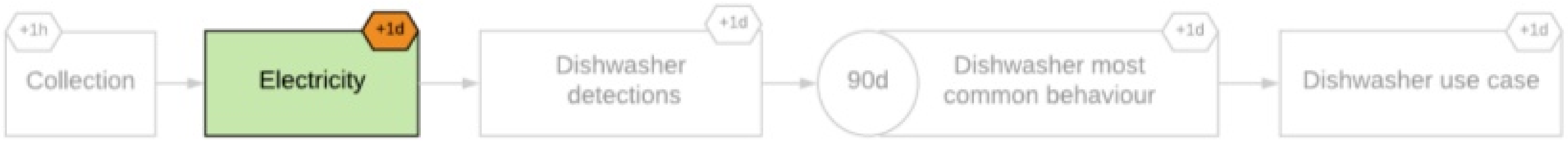




TOON®

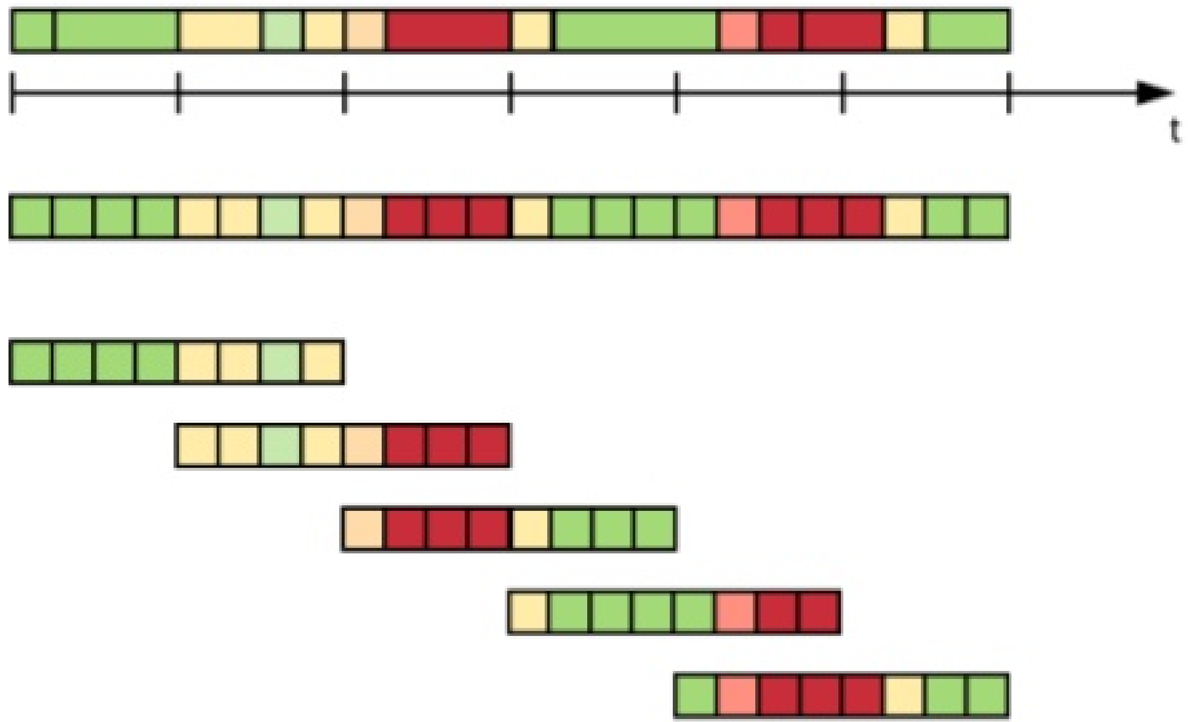
#SAISML4

Example data flow

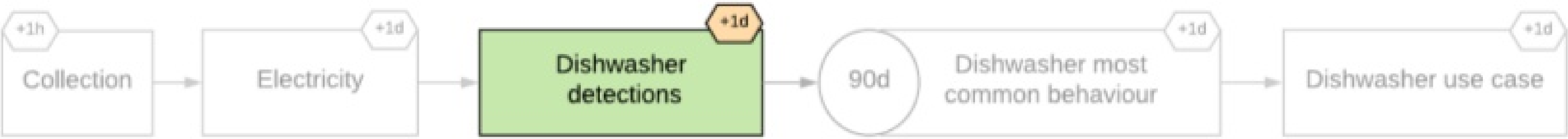


Getting the data ready

- Extraction & Cleaning
- Resampling
- Vectorization



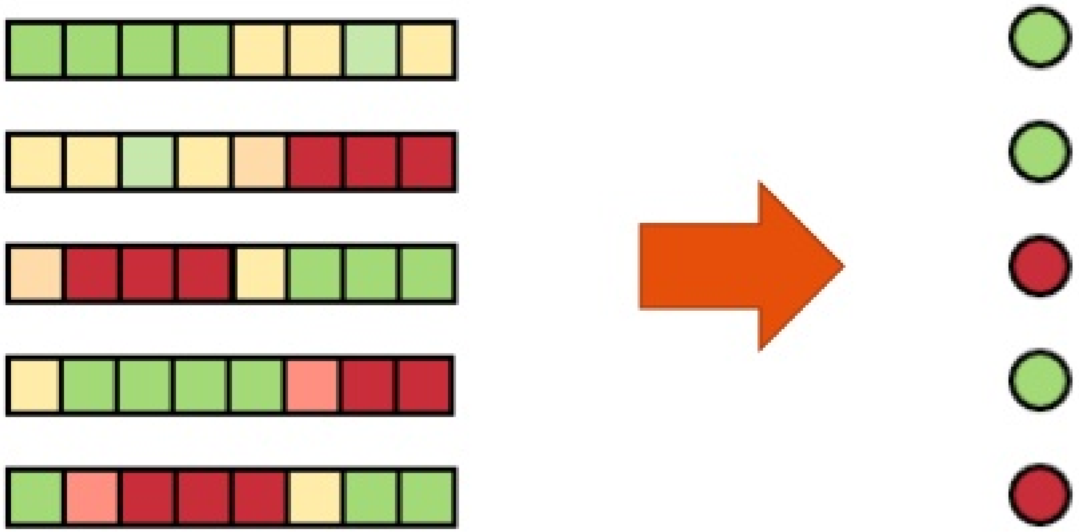
Example data flow



Detecting appliances

- **Signal processing**
- Machine learning (One Big model)

Keras TensorFlow APACHE Spark



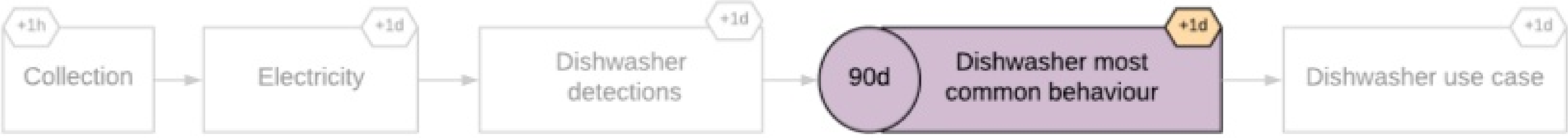
Vectorized electricity

Predictions

TOON®

#SAISML4

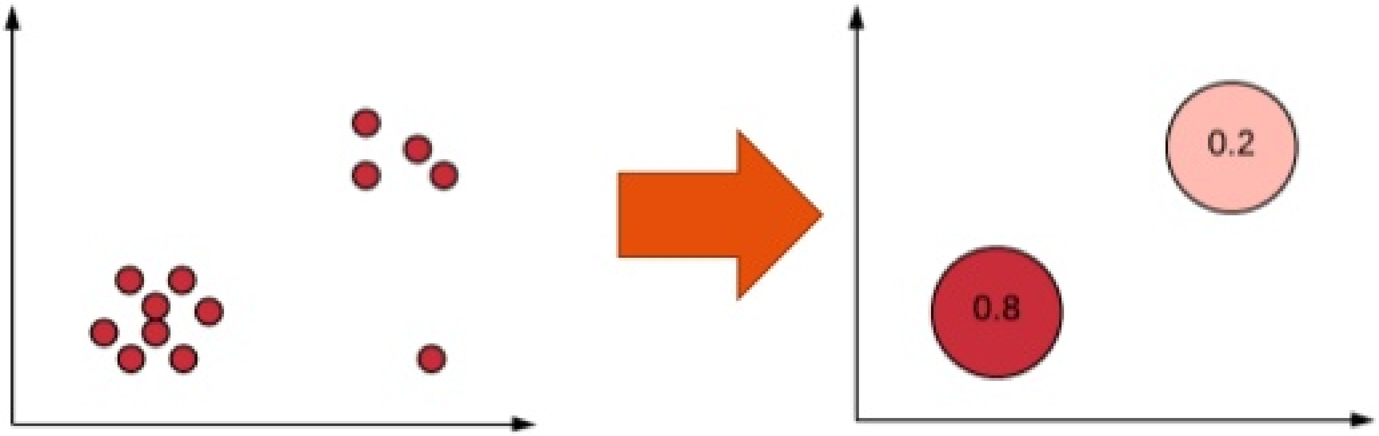
Example data flow



Finding user behavior

- Clustering per user
- Many, many, **many** (small) models

scikit learn + APACHE Spark

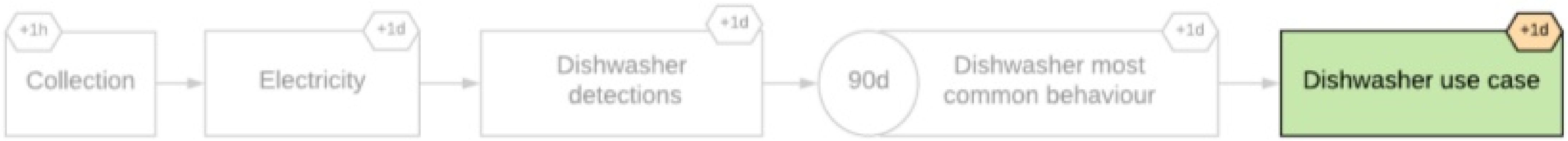


Detections Behaviour

TOON®

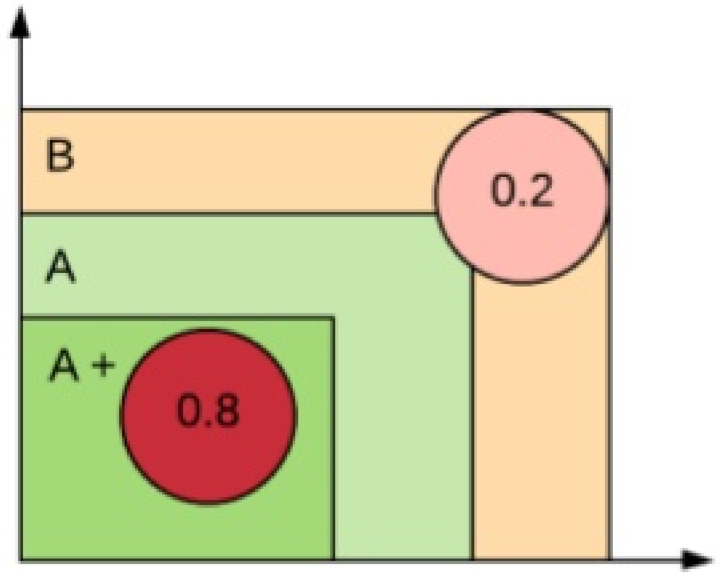
#SAISML4

Example data flow



Drawing conclusions

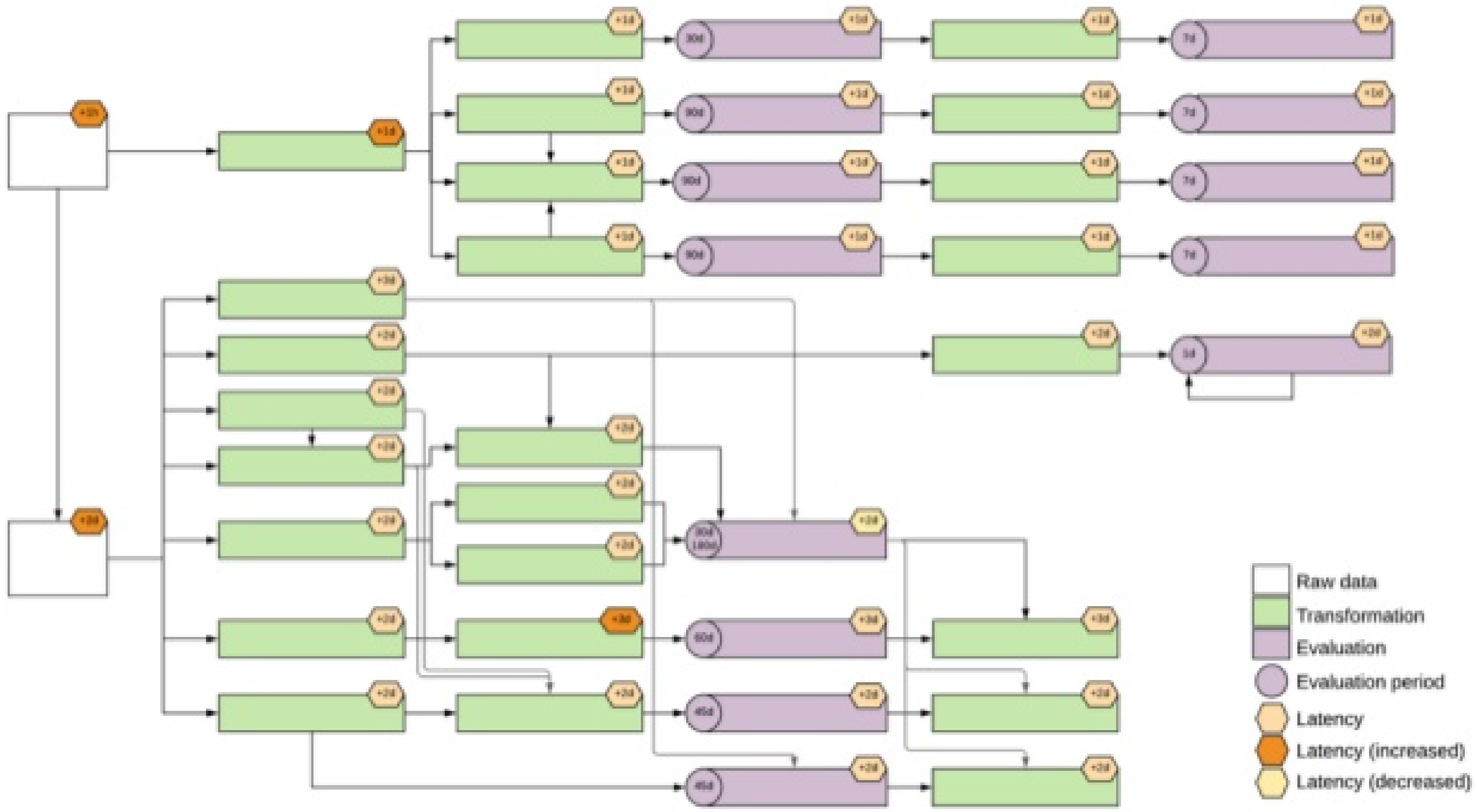
- Comparing to other users
- Comparing to industry standards



TOON®

#SAISML4

The Data Pipeline



#SAISML4

Managing jobs (Option 1)

- Databricks Jobs & Notebook Workflows
  https://databricks.com/blog/2016/08/30/notebook-workflows-the-easiest-way-to-implement-apache-spark-pipelines.html

Active runs

Run	Run ID	Start Time	Launched	Duration	Spark	Status
Run Now / Run Now With Different Parameters						

Completed in past 60 days

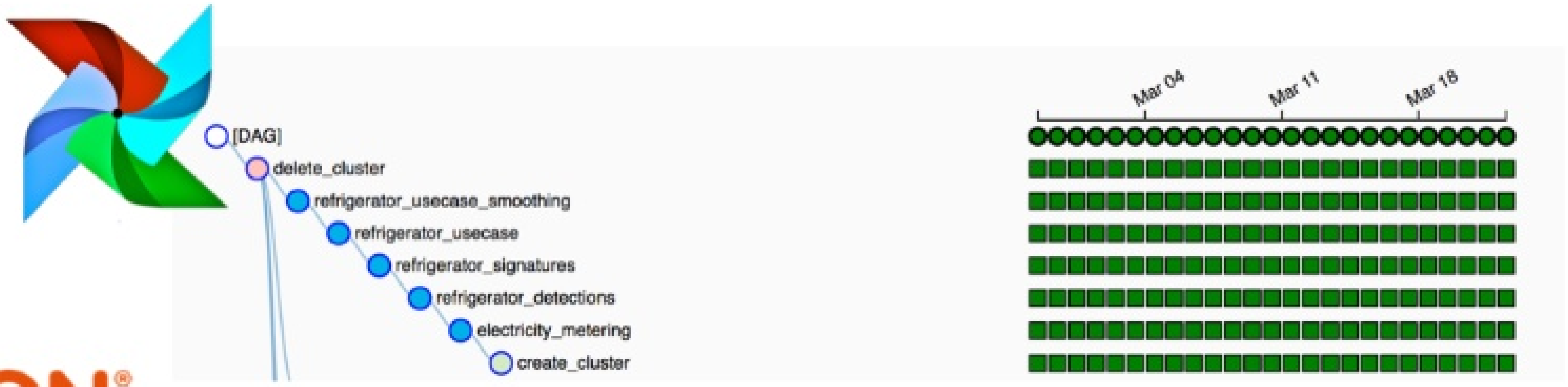
Latest successful run (refreshes automatically)

‹ Previous 20 Next 20 ›

Run	Run ID	Start Time	Launched	Duration	Spark	Status	
Run 206	114438	2018-09-28 11:21:46 CEST	Manually	47m 57s	Spark UI / Logs / Metrics	Succeeded	×
Run 205	114350	2018-09-28 04:00:00 CEST	By scheduler	3m 26s	Spark UI / Logs / Metrics	Failed	×
Run 204	114109	2018-09-27 04:00:00 CEST	By scheduler	44m 54s	Spark UI / Logs / Metrics	Succeeded	×
Run 203	113874	2018-09-26 04:00:01 CEST	By scheduler	46m 27s	Spark UI / Logs / Metrics	Succeeded	×
Run 202	113635	2018-09-25 04:00:00 CEST	By scheduler	46m 35s	Spark UI / Logs / Metrics	Succeeded	×
Run 201	113370	2018-09-24 04:00:00 CEST	By scheduler	46m 31s	Spark UI / Logs / Metrics	Succeeded	×
Run 200	113132	2018-09-23 04:00:00 CEST	By scheduler	45m 22s	Spark UI / Logs / Metrics	Succeeded	×
Run 199	112893	2018-09-22 04:00:00 CEST	By scheduler	45m 45s	Spark UI / Logs / Metrics	Succeeded	×

Managing jobs (Option 2)

- Airflow DAGs
  https://airflow.apache.org/
- Must read:
  - ETL principles: https://gtoonstra.github.io/etl-with-airflow/principles.html
  - Gotcha's: https://gtoonstra.github.io/etl-with-airflow/gotchas.html



TOON®
#SAISML4

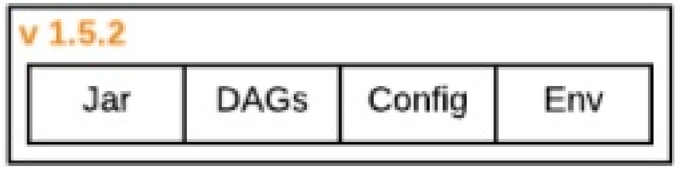
When designing data pipelines

- Enforce idempotent constraints

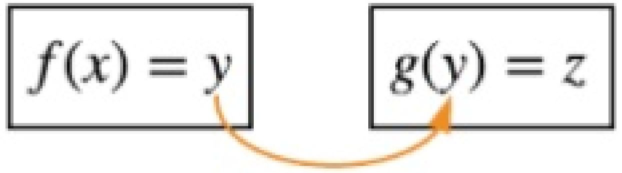
$$f(x) = y$$

Always

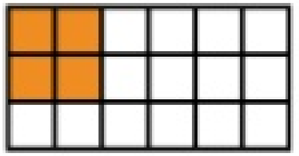
- Enforce reproducibility



- Let data transformations be chainable

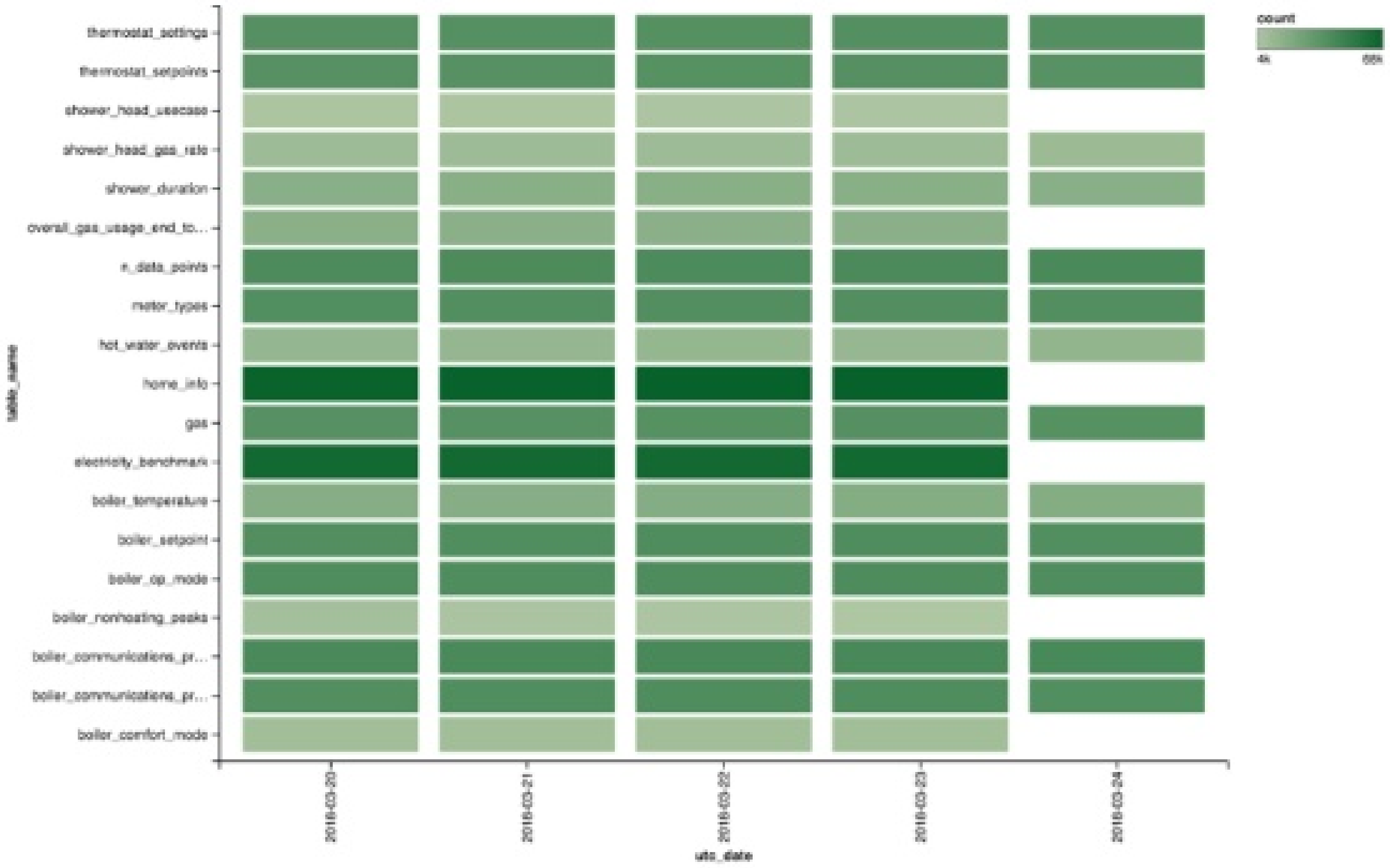


- Leverage partitioning and data locality



Monitoring

- Live dashboards with aggregated data

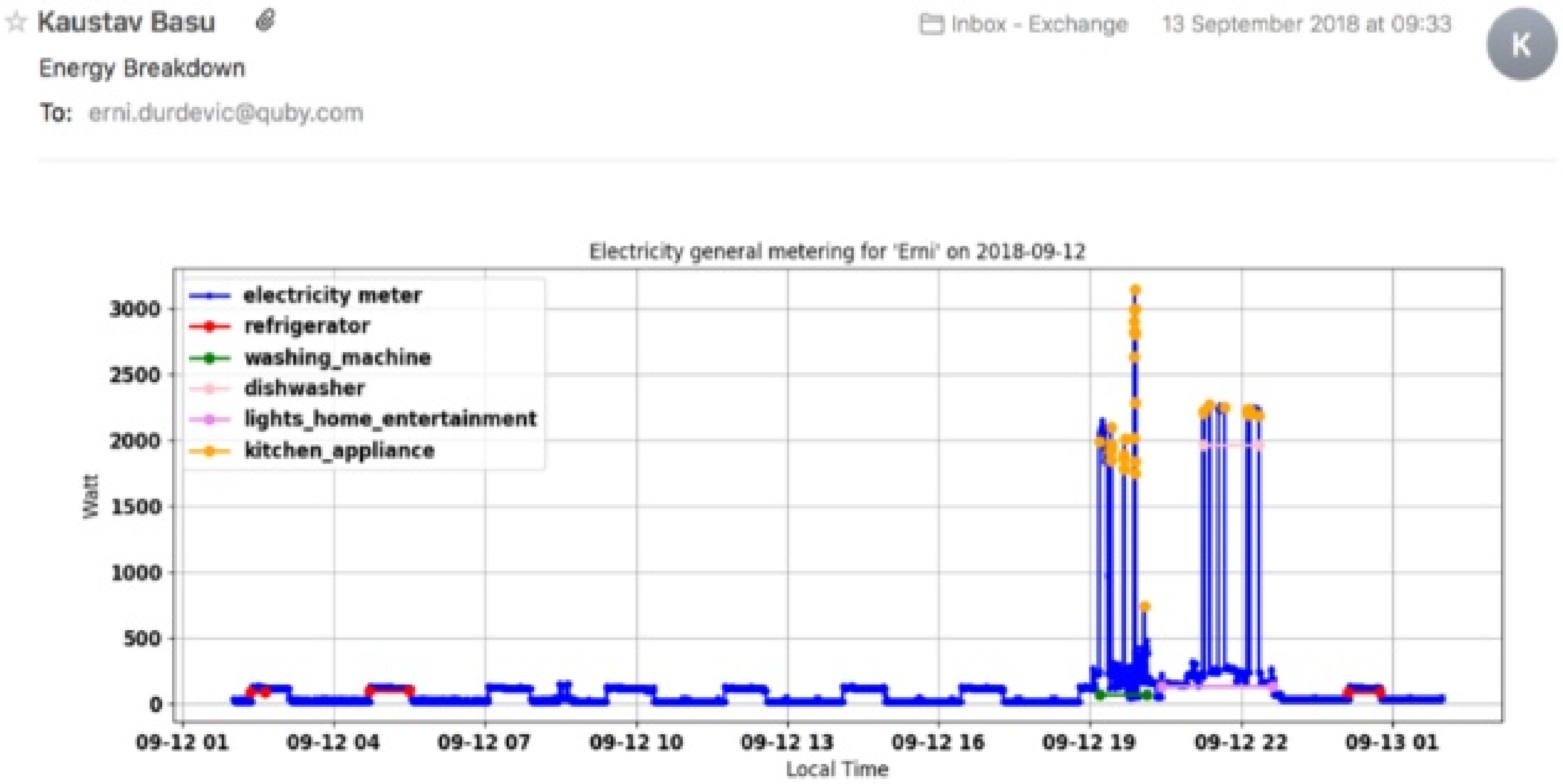


TOON®

#SAISML4

Monitoring and Validation

- Daily email to Quby's VIP employees



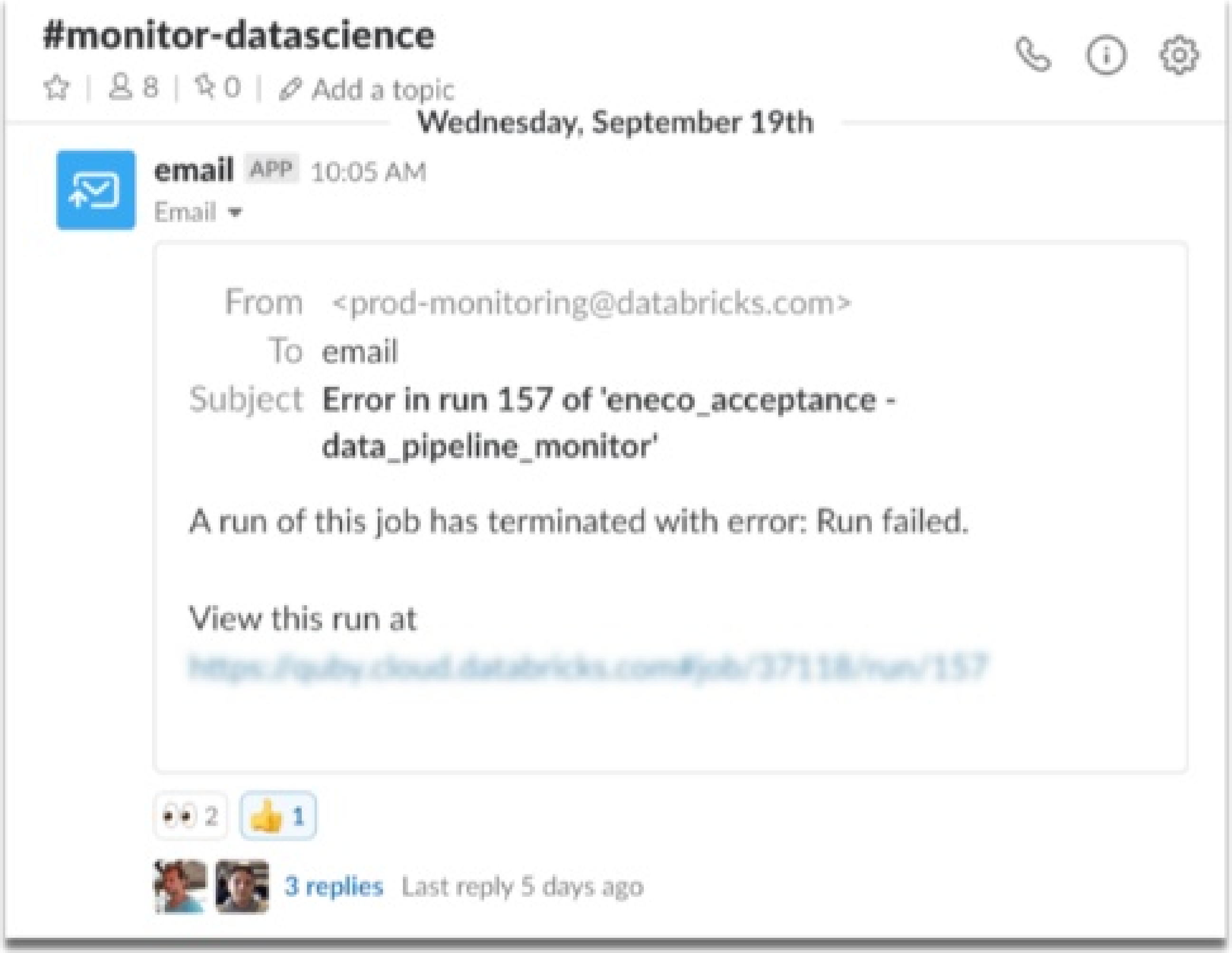
TOON®

#SAISML4

Alerting

Alerting (via ~~Email~~ / Slack)

- If anything goes wrong
- If an independent monitoring job detects missing data



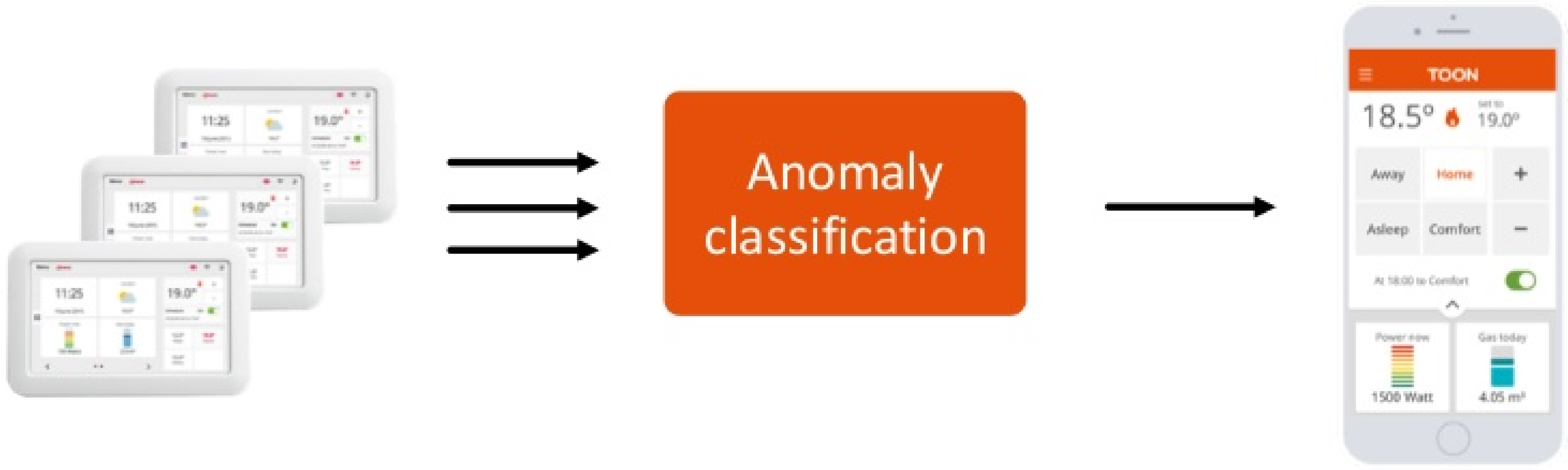
USE CASE #2

Detecting anomalies in heating systems

TOON®

#SAISML4

Structured Streaming



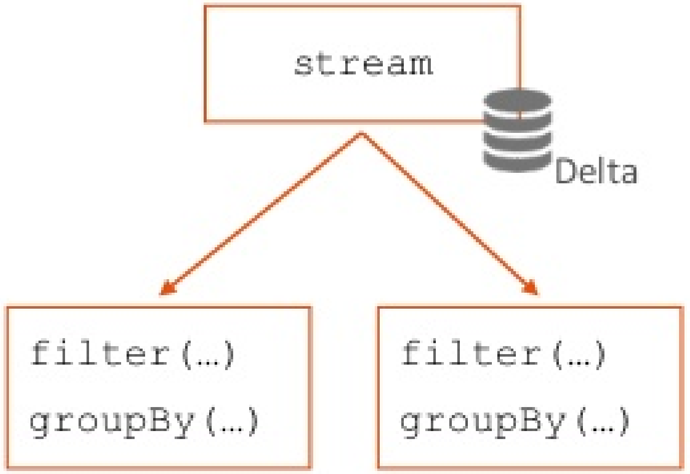
TOON®

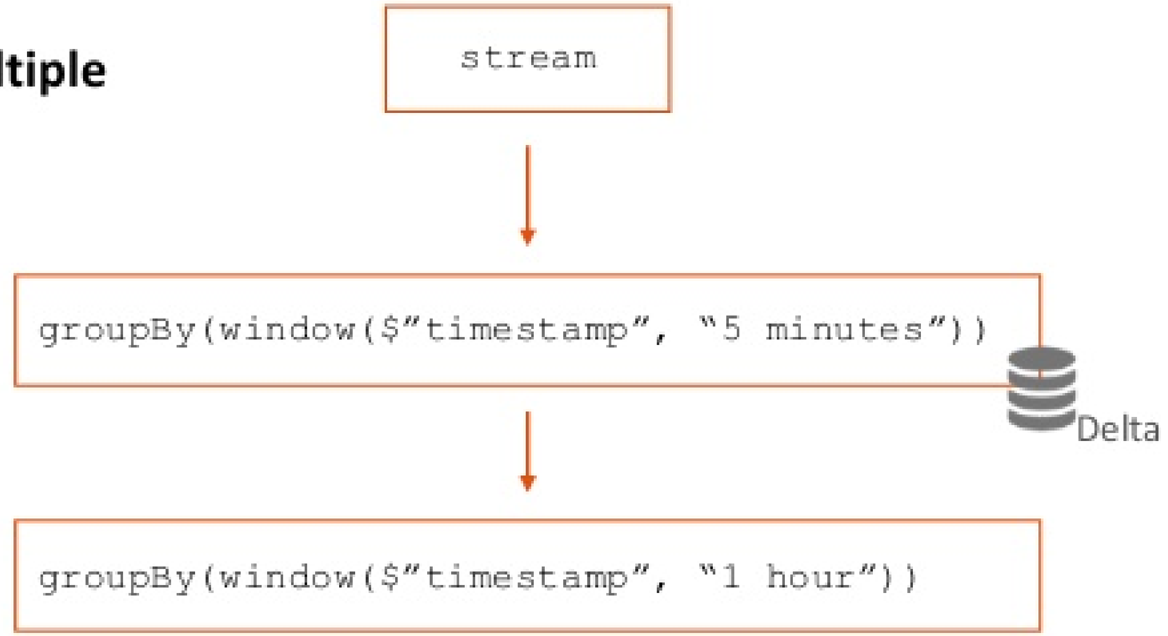
#SAISML4

Multiple Streaming aggregations

When working with streams in spark it is not possible to do **multiple aggregations** on the same stream

- E.g. Forking a stream in multiple streams
- E.g. Do consecutive aggregations





Work around

Output the aggregations on a sink and read it back in Spark
(E.g. Kafka, Kinesis, Delta tables)

Non time-based windows on streams

- When working with streams in spark it is not possible to compute **non time-based window** operations
  - E.g. Compute the derivative of a signal

timestamp	lag(timestamp)
2018-10-04 14:00:00	
2018-10-04 15:00:00	2018-10-04 14:00:00
2018-10-04 16:00:00	2018-10-04 15:00:00

Our solution

Compute **non time-based window** operations

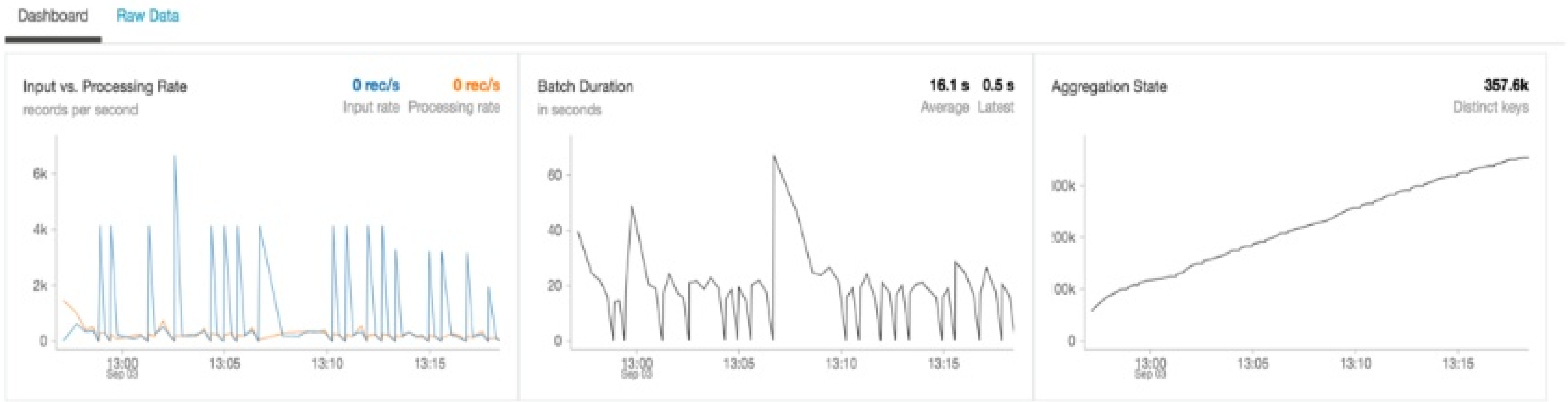
- Use (Flat) mapGroupWithState (Beware: no ordering guarantee)
- Inside a time-based window by collecting a list of Struct(Timestamp, Value)

TOON®

#SAISML4

Stream to stream joins

- When doing **stream to stream joins**, keep an eye on the distinct key count on aggregation state



TOON®

#SAISML4

Streaming in production

- Structured Streaming in Production checklist
  - Setup recovery of queries from failure
    - Configure Checkpointing
    - Query restart
  - Configure Spark scheduler pool for efficiency
  - Optimize performance of stateful streaming queries
  - Configure multiple watermark policy

Reference: https://docs.databricks.com/spark/latest/structured-streaming/production.html

TOON®

#SAISML4

Erni Durdevic
Machine Learning Engineer

+31 (0) 638 11 94 12
erni.durdevic@quby.com

TOON®
#SAISML4