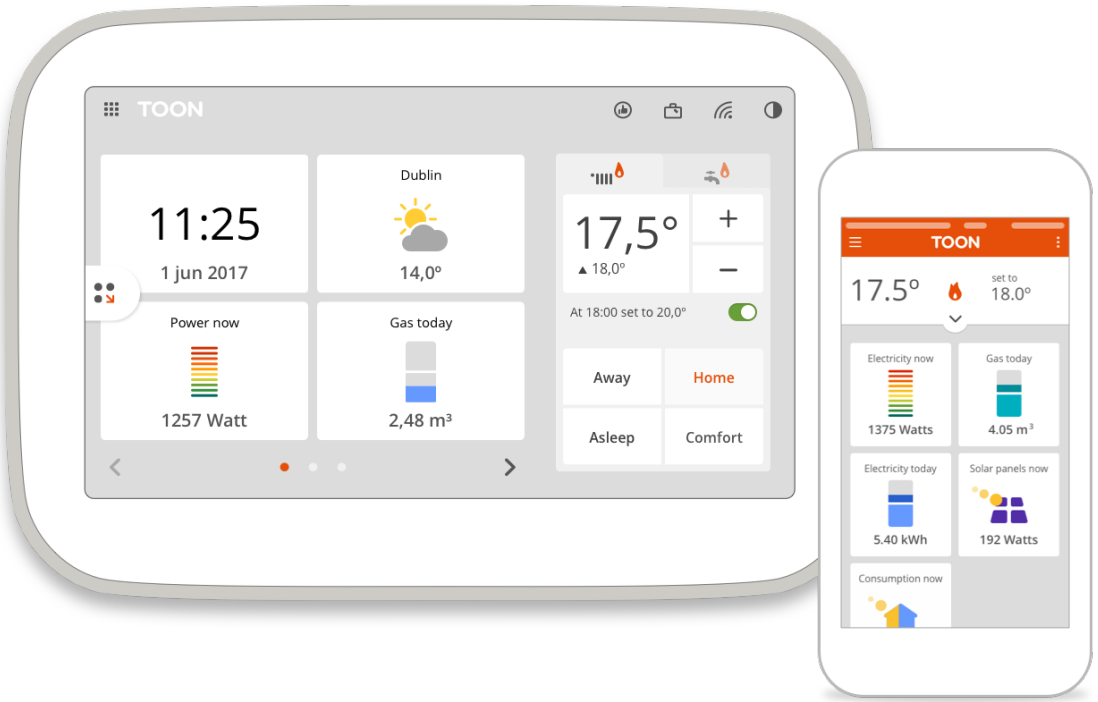
Using Spark in the Cloud: A Devops perspective

Telmo Oliveira, Toon


TOON
11:25
1 jun 2017
Dublin
14,0°
Power now
1257 Watt
Gas today
2,48 m³
17,5°
▲ 18,0°
At 18:00 set to 20,0°
Away
Home
Asleep
Comfort
TOON
17.5°
set to
18.0°
Electricity now
1375 Watts
Gas today
4.05 m³
Electricity today
5.40 kWh
Solar panels now
192 Watts
Consumption now

TOON®

2004

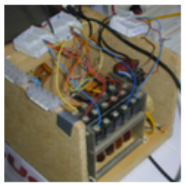
Home Automation Europe is founded

2005

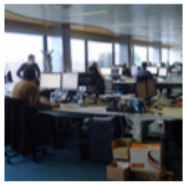
The Home Control Box is launched

2006

Home control alternatives are prototyped



2007



Home Automation Europe moves into a new office

2008

The shift to smart thermostats

2009

2010

2011

2012

Home Automation Europe partners with Eneco to create Toon

2013

Home Automation Europe becomes Quby

2014

2015

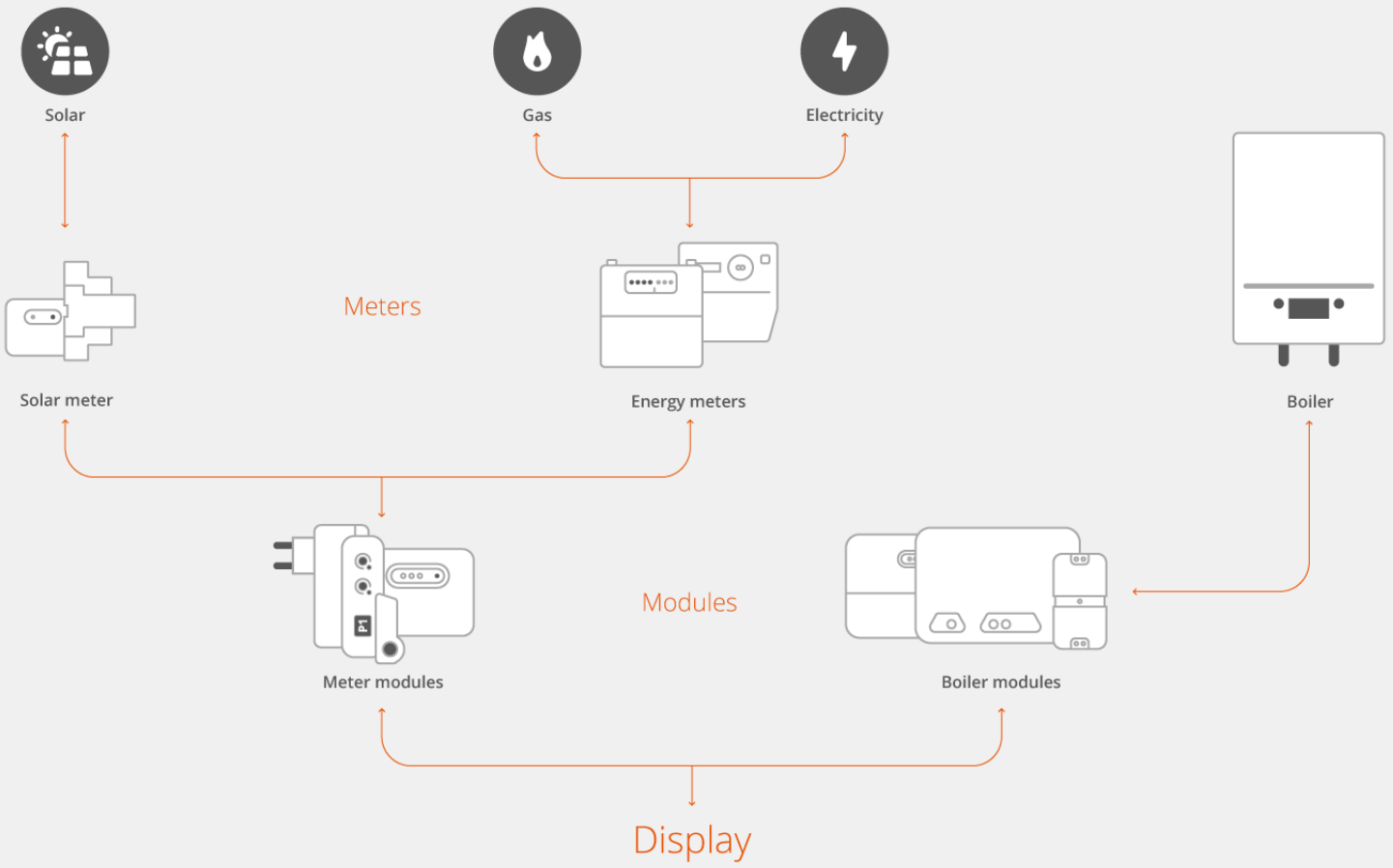
Eneco acquires 100% of Quby's shares

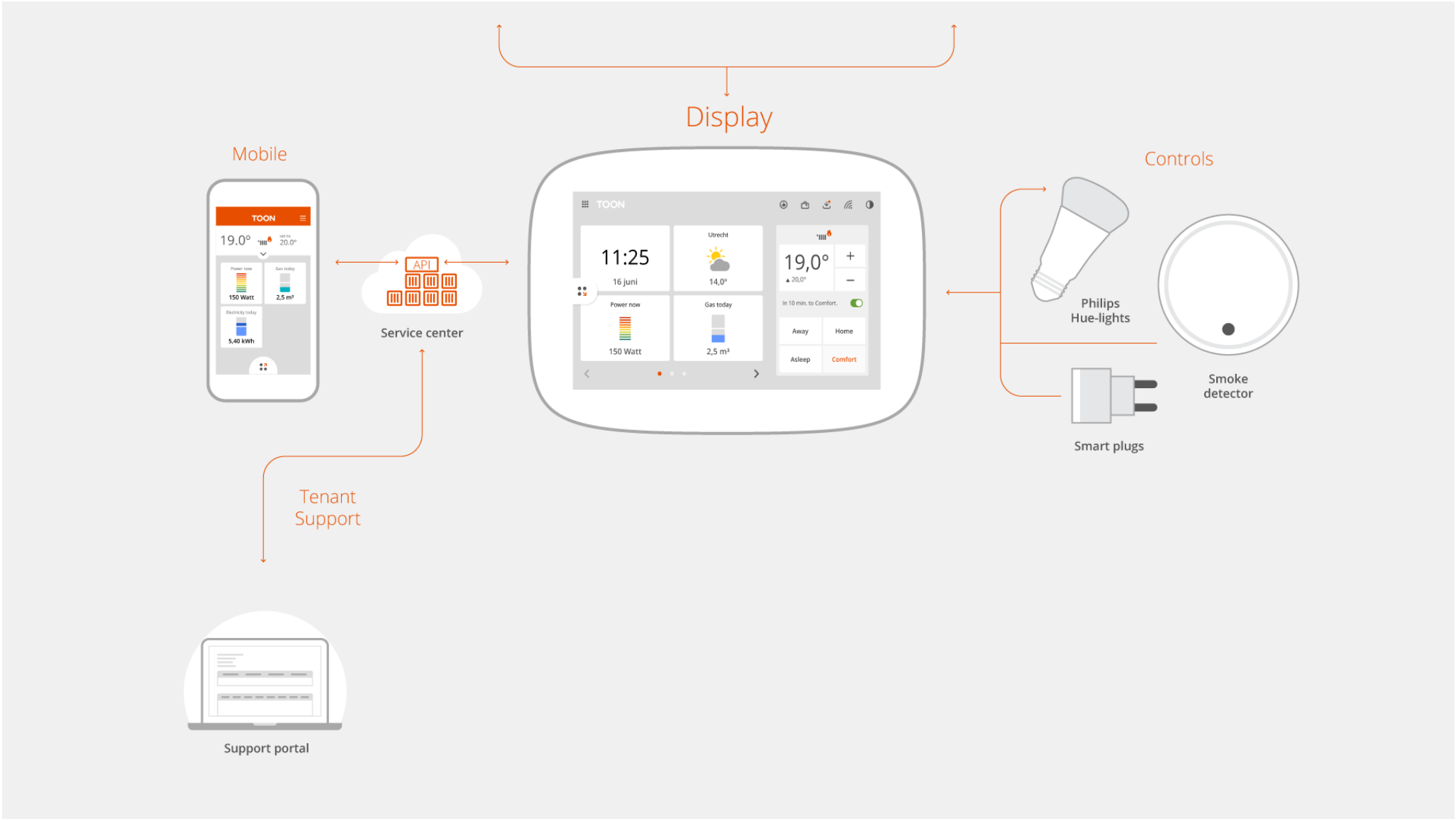
2016

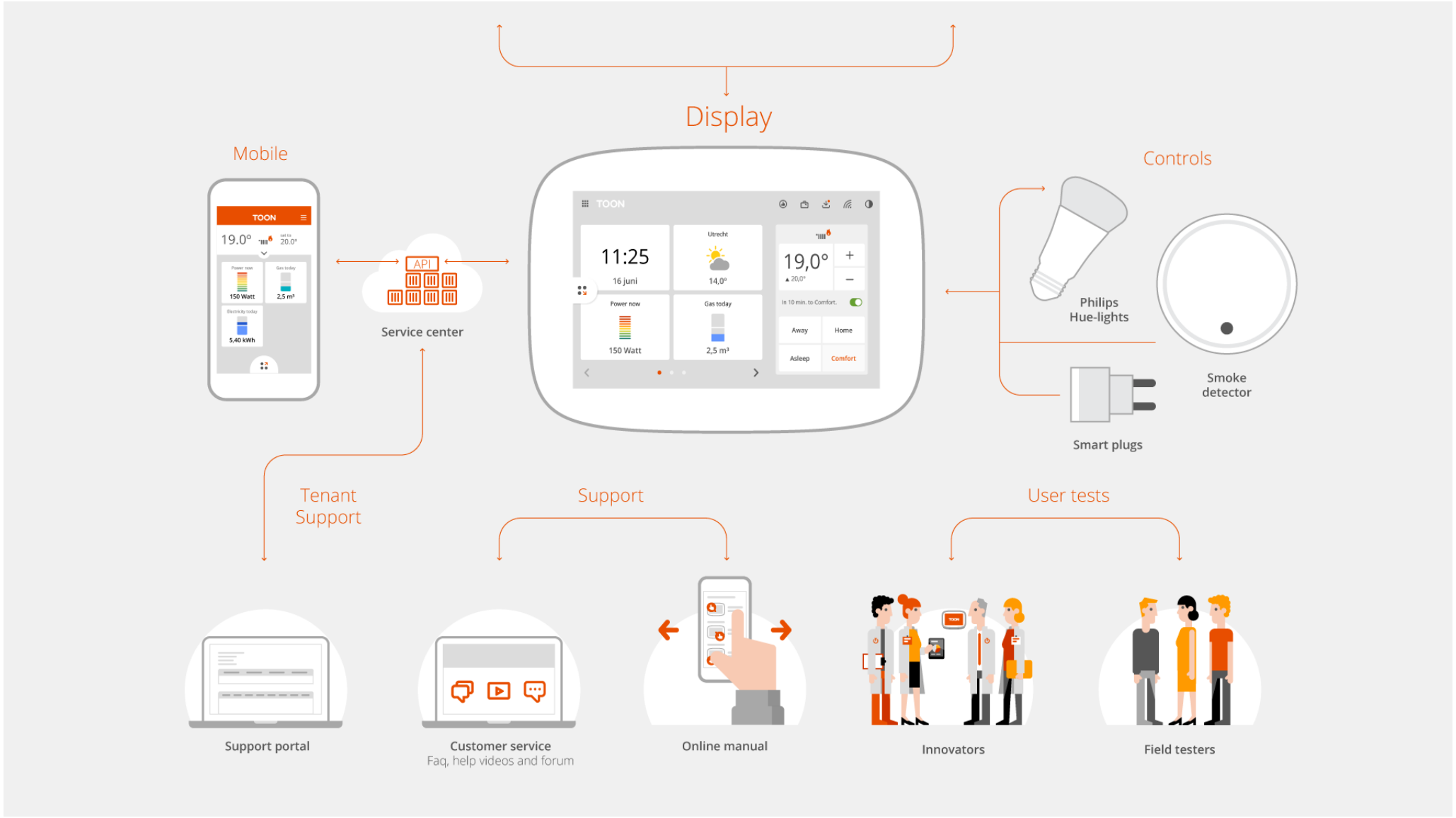
Quby partners with Engie to create Boxx

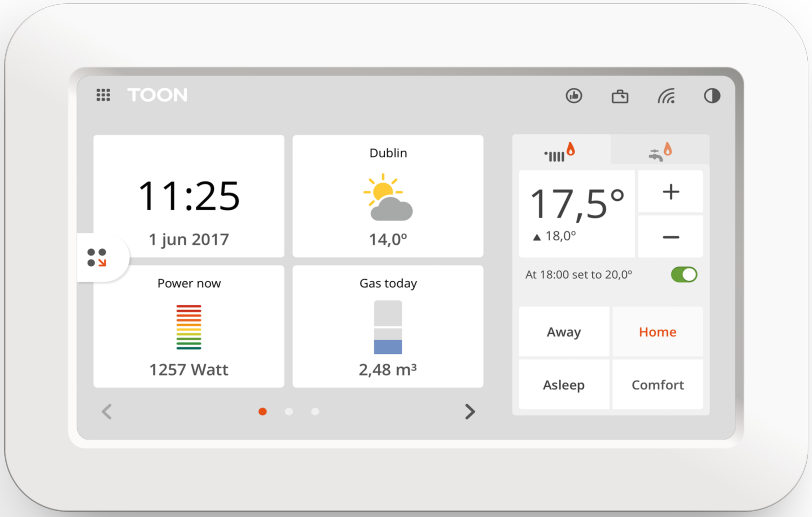
2017

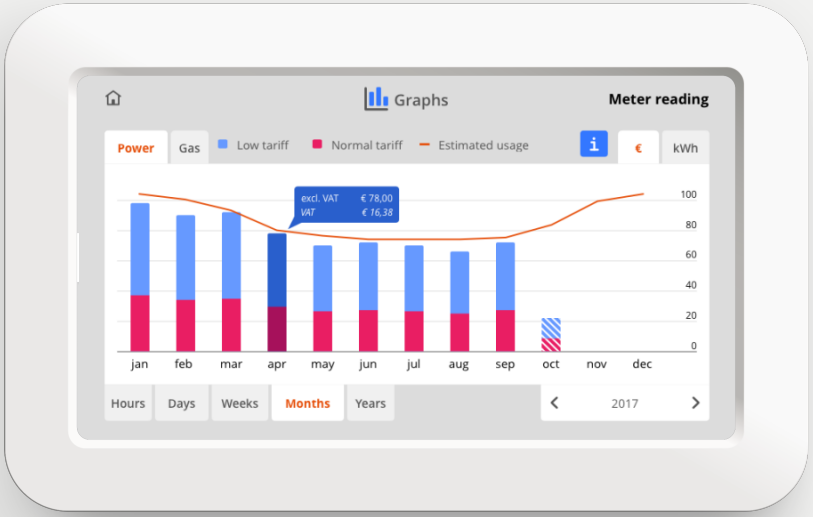
Quby's white label product is rebranded internationally as Toon


Solar
Gas
Electricity
Meters
Solar meter
Energy meters
Boiler
Modules
Meter modules
Boiler modules
Display

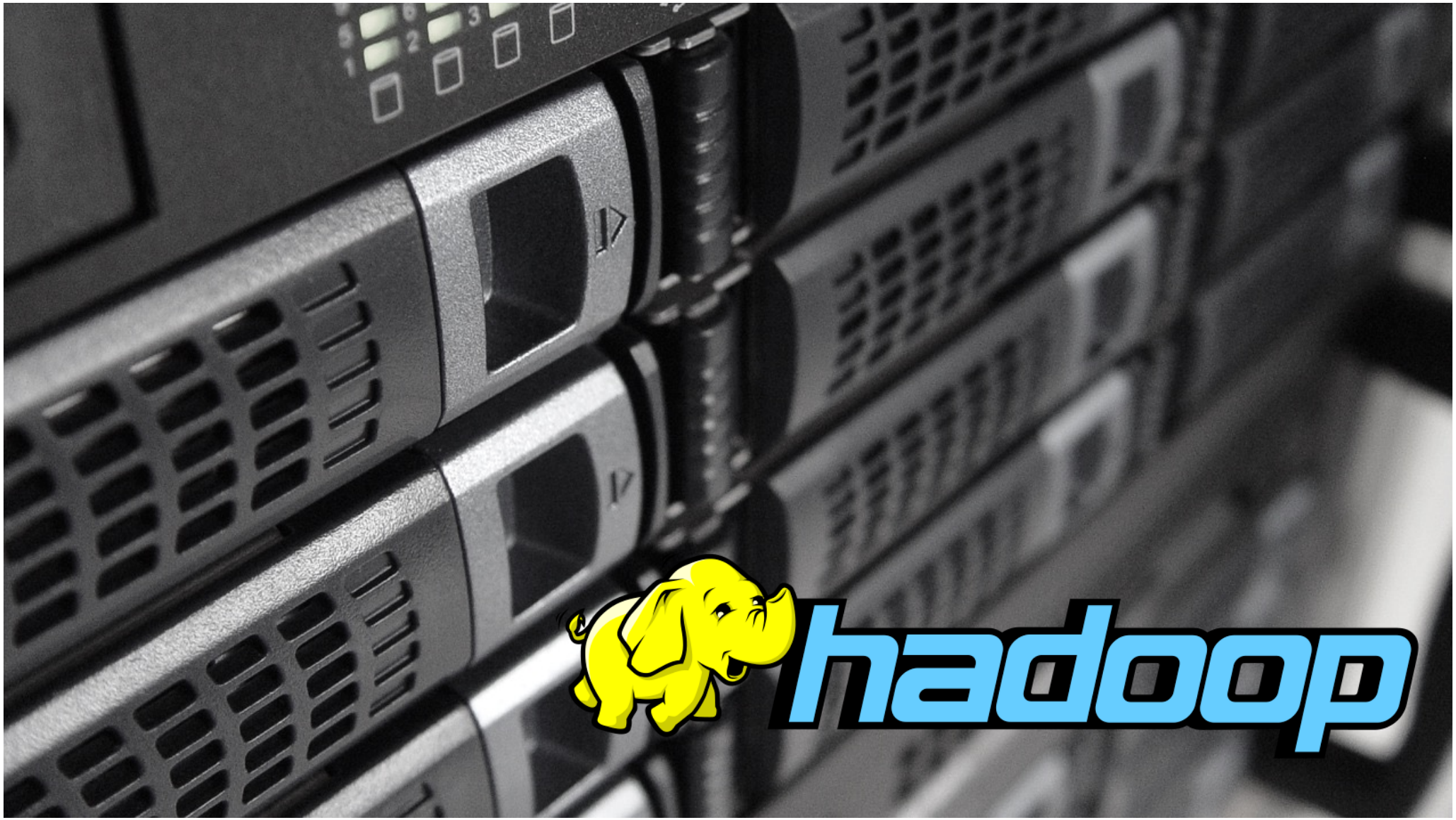

Display
Mobile
TOON
19.0°
150 Watt
2,5 m³
5,40 kWh
API
Service center
TOON
11:25
16 juni
Utrecht
14,0°
19,0°
Power now
150 Watt
Gas today
2,5 m³
Away
Home
Asleep
Comfort
Controls
Philips
Hue-lights
Smoke
detector
Smart plugs
Tenant
Support
Support portal

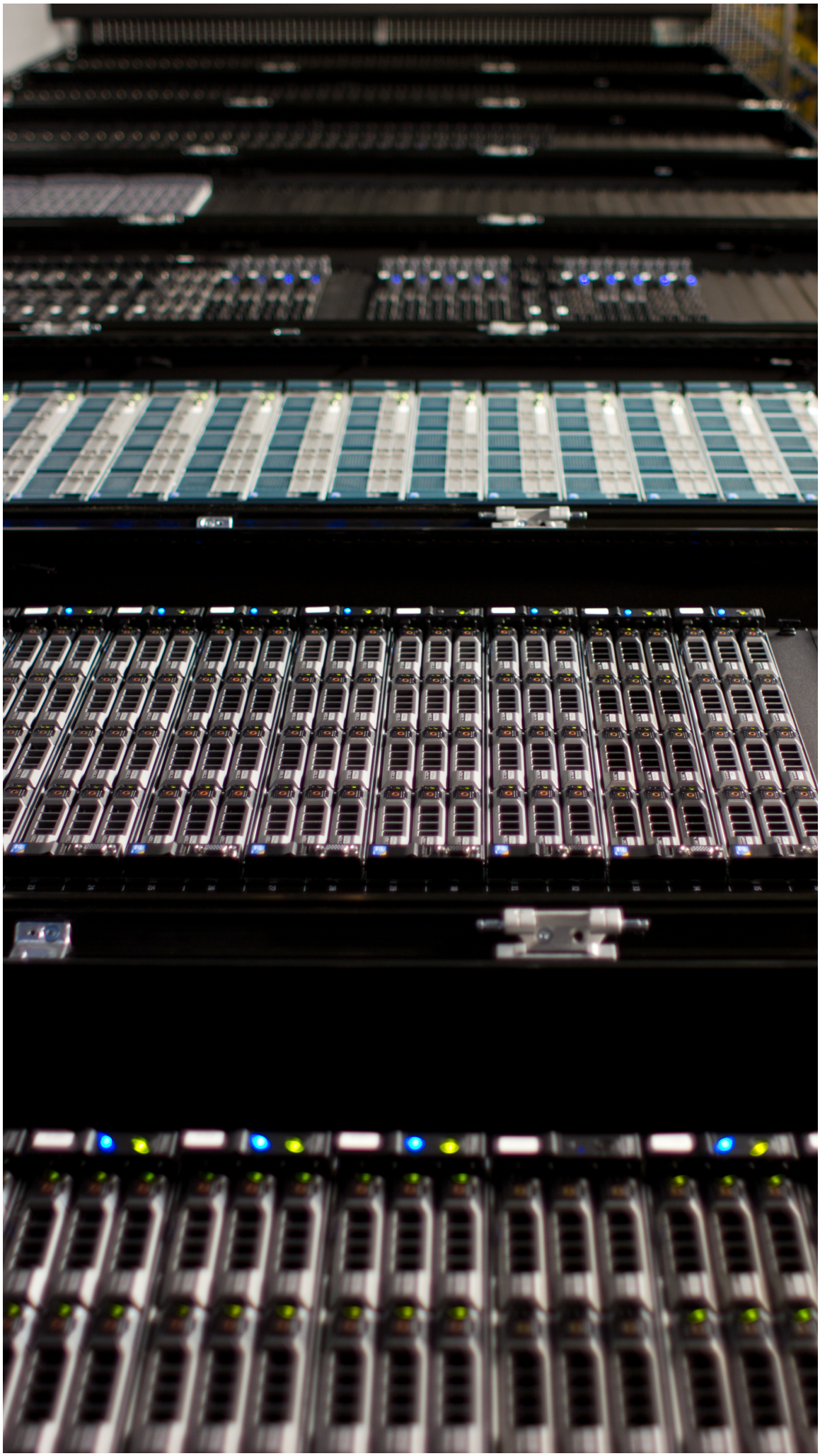

Display
Mobile
TOON
19.0°
Service center
API
TOON
11:25
16 juni
Utrecht
14,0°
Power now
150 Watt
Gas today
2,5 m³
19,0°
Away
Home
Asleep
Comfort
Controls
Philips
Hue-lights
Smoke
detector
Smart plugs
Tenant
Support
Support
User tests
Support portal
Customer service
Faq, help videos and forum
Online manual
Innovators
Field testers


TOON
11:25
1 jun 2017
Dublin
14,0°
Power now
1257 Watt
Gas today
2,48 m³
17,5°
▲ 18,0°
At 18:00 set to 20,0°
Away
Home
Asleep
Comfort


Graphs
Meter reading
Power
Gas
Low tariff
Normal tariff
Estimated usage
kWh
excl. VAT € 78,00
VAT € 16,38
100
80
60
40
20
0
jan
feb
mar
apr
may
jun
jul
aug
sep
oct
nov
dec
Hours
Days
Weeks
Months
Years
2017

TOON®


hadoop



Requirements

- Seamless transition
- Ensure data anonymity
- Move fast, optimise later
- Ensure multi-tenancy
- As little disturbance as possible to the DS team

TOON®

TOON®

TOON®

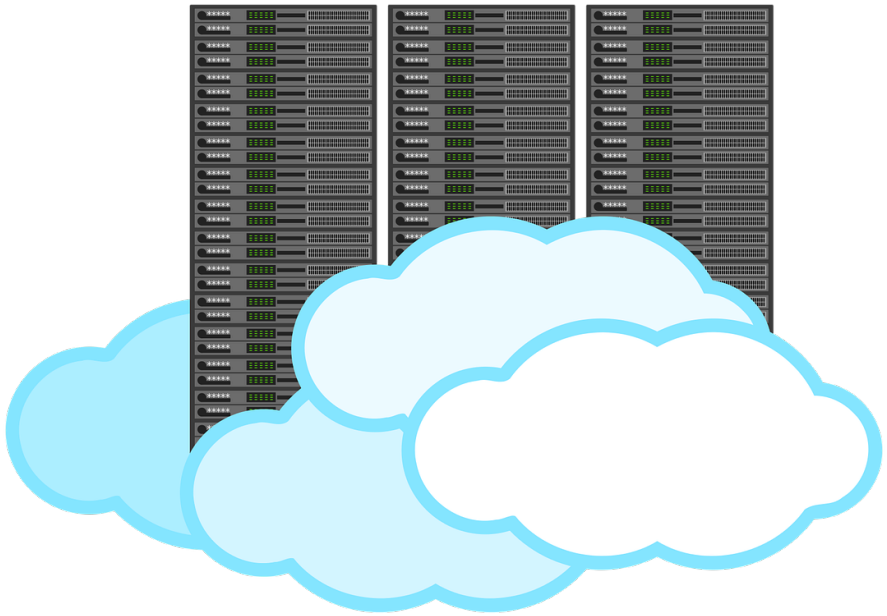
TOON®

- Cluster timeouts
- Autoscaling
- Spot instances
- Well documented API

TOON®

Infrastructure as code

- Repeatability
- Fast deployment
- Resilience
- Documentation



TOON®

TOON®

Terraform

- S3 Buckets
- EC2 instances
- Network topology
- Log management
- RDS instances
- IAM roles/policies

TOON®

Terraform

- S3 Buckets
- EC2 instances
- Network topology
- Log management
- RDS instances
- IAM roles/policies

```
./bin/provision relational-db <target> apply
```

```
./bin/provision buckets-configuration <target> apply
```

TOON®

Ansible

- User management
- Databases and ACLs
- Custom app deployment

TOON®

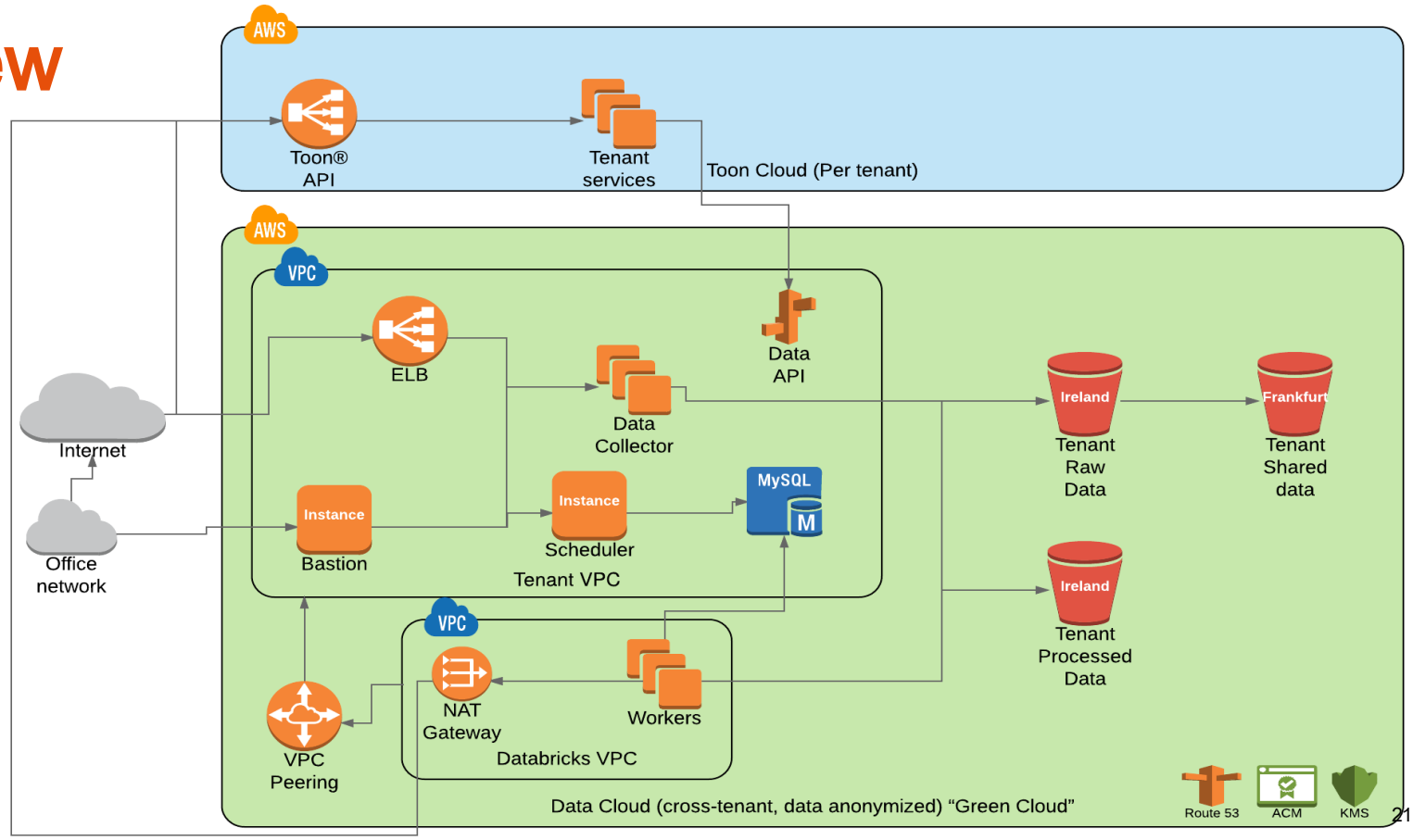
Ansible

- User management
- Databases and ACLs
- Custom app deployment

```
./bin/configure hive-metastore.yml <target>
./bin/configure create-cluster.yml <target>
```

TOON®

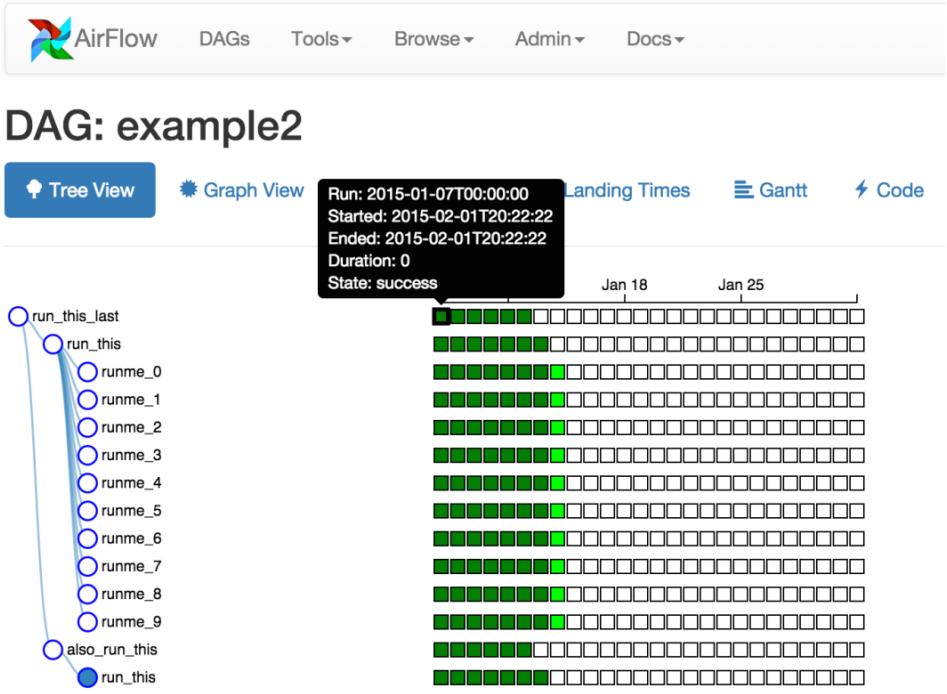
Architecture Overview

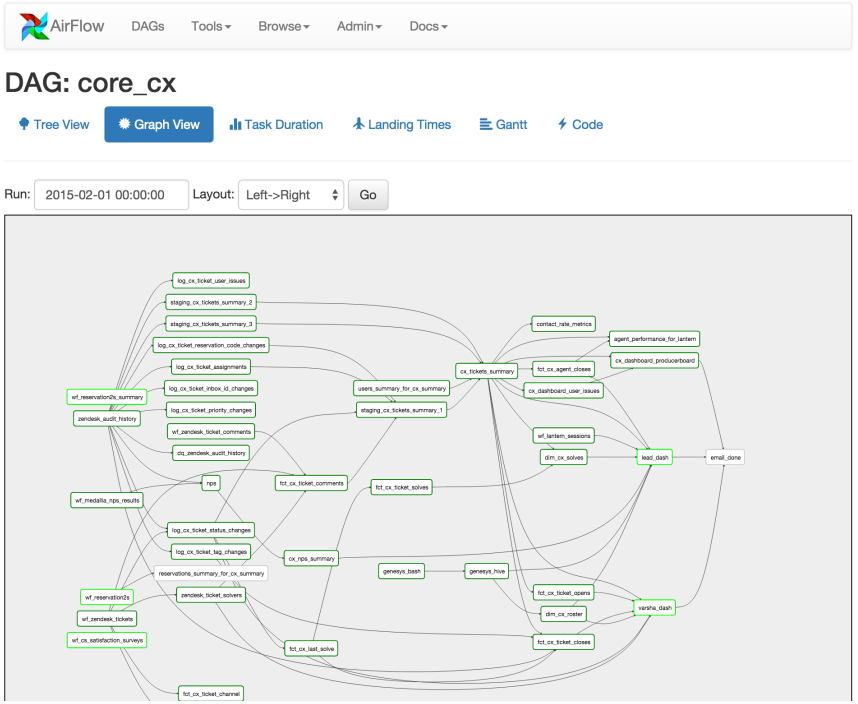


TOON®


8 Mi
9 Do
10 Fr
11 Sa
12 So
13 Mo

Airflow





TOON®

```
dag_config:
  dag_id: <tenant>_preprocessed_dag
  schedule_interval: '30 5 * * *'
  start_date: 2017-08-28
  max_active_runs: 1
```

```
task_defaults:
  cluster_name: '<tenant>_preprocessed_dag'
  databricks_conn_id: 'databricks_<tenant>'
  retries: '3'
  jar: '<location_of_jar>'
```

TOON®

```
tasks:
  download_<user_group>_acc:
    class: eu.toon.etl.preprocessed.LoadPreProcessedData
    parameters:
      endpoint:                   '<data_endpoint>'
      source_db:                  '<tenant_db>'
      source_table:               '<table_name>'
      destination_db:             '<tenant_db>'
      destination_table:          '<target_table_name>'
      authenticator:              '<authenticaton_type>'
      aws_region:                 '<aws_region>'
      source:                     'acc'
      service_name:               '<service_name>'
      aws_wso2_initial_token_key: '<token_key_on_ssm>'
      aws_wso2_auth_endpoint:     '<auth_endpoint>'
      date:                       '{{ ds }}'
    depends_on:
    - <task_name>
```

TOON®

```
new_cluster:
    spark_version: '3.2.x-scala2.11'
    node_type_id: 'i3.xlarge'
    autotermination_minutes: '20'
    spark_conf:
        spark.hadoop.javax.jdo.option.ConnectionDriverName: '{{ <hive_driver> }}'
        spark.hadoop.javax.jdo.option.ConnectionUserName: '{{ <hive_username> }}'
        spark.hadoop.javax.jdo.option.ConnectionPassword: '{{ <hive_passowrd> }}'
        spark.hadoop.javax.jdo.option.ConnectionURL: '{{ <hive_url> }}'
        spark.serializer: 'org.apache.spark.serializer.KryoSerializer'
        spark.hadoop.fs.s3.impl: 'com.databricks.s3a.S3AFileSystem'
        spark.sql.hive.metastore.jars: 'builtin'
        spark.io.compression.codec: 'snappy'
        spark.sql.hive.metastore.version: '1.2.1'
    aws_attributes:
        first_on_demand: '2'
        availability: 'SPOT'
        zone_id: '<aws_region>'
        instance_profile_arn: '<arn_of_profile>'
        spot_bid_price_percent: '75'
    num_workers: '5'
    custom_tags:
        Environment : '<tenant>_prod'
        GroupName : 'BigData'
        Tenant: '<tenant_name>'
        Provisioner : 'Databricks'
        CostCenter : 'international-operations'
        Stage : 'prod'
```

TOON®

TOON®

- External Hive metastore
- Send logs to S3
- Authorisation
- i3.2xlarge nodes

TOON®

Future plans

- Streaming
- Real time services
- Improve CI/CD

TOON®

What's all this for?

TOON®

What's all this for?

TOON®

Thanks to the team

Aemro Amare

Barend Garvelink

Bert Jan Katsman

Kliment Markovski

Miquel Monreal

Stanislava Potupchik

TOON®

Questions?

TOON®