Routing billions of events a day: How we do routing in Schibsted

Carlos Manuel Duclos-Vergara, Staff Engineer

About me

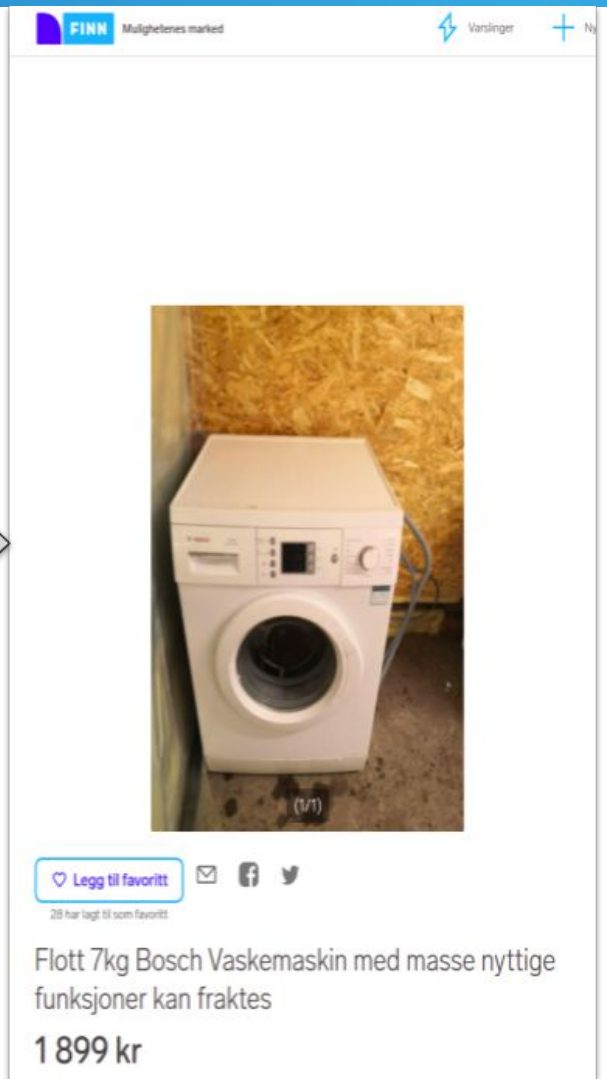
Schibsted

Agenda

- Schibsted
- A short story
- GDPR
- Pulse (our tracking solution)
  - Overview
  - Internals

Schibsted

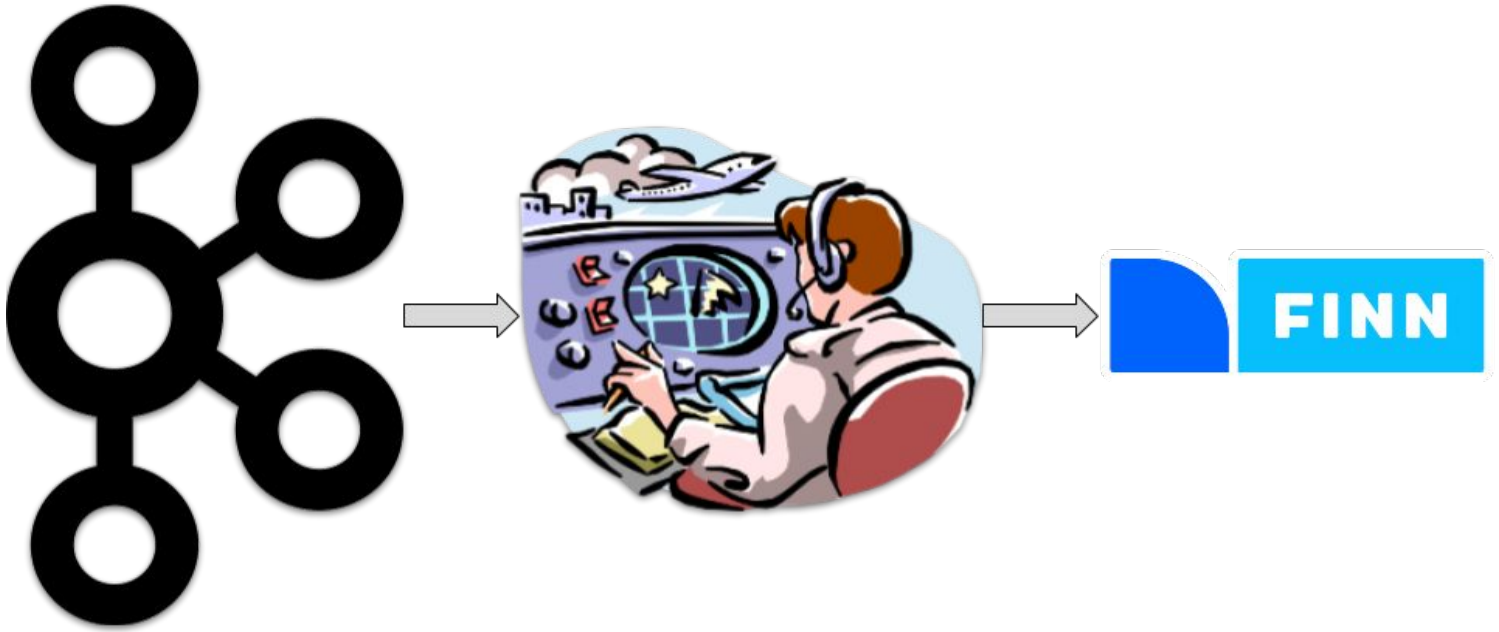
Event generation



Event routing

Collector EventLog EventProcessorRouter

Event dispatching



Kafka Dispatcher Consumer

Event consumption

GDPR and data collection

User rights

1. Data portability
2. Right to be forgotten

Legal basis for data collection

1. Consent
2. Processing obligation
3. Legal obligation
4. Vital interest
5. Public interest
6. Legitimate interest

End to end event processing solution

Pulse

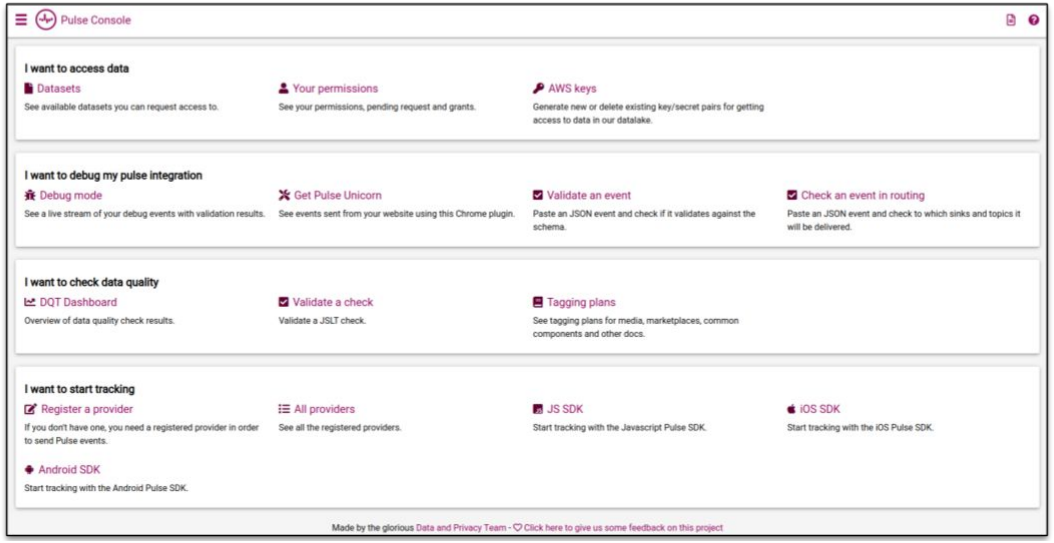
Pulse ecosystem

Tracking SDK

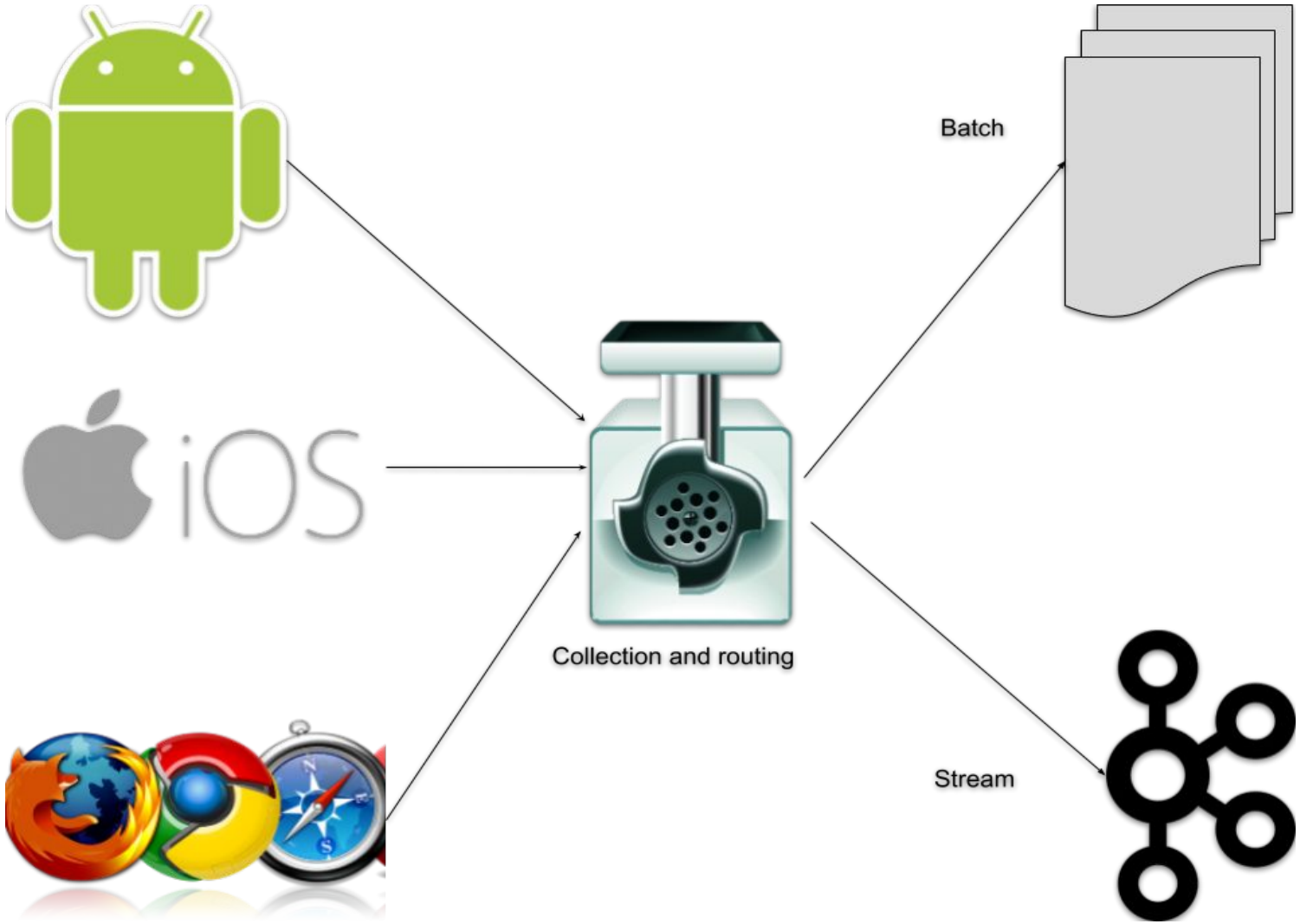
Data Quality

Data and Statistics

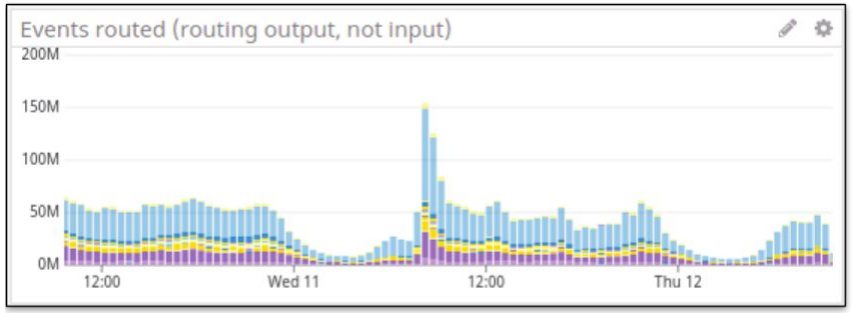
Administration portal



Lifetime of an event



Side track: How much is 1 billion events

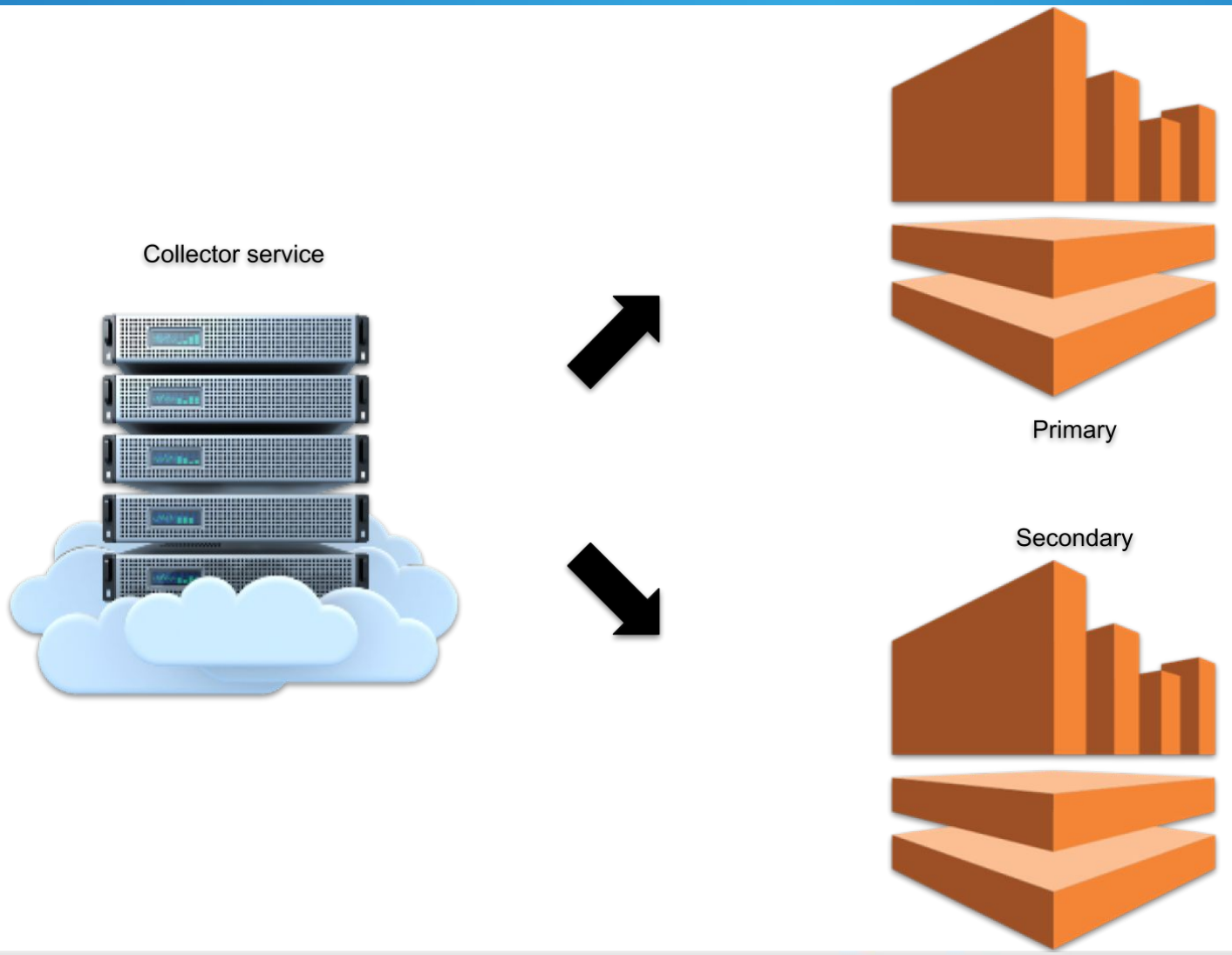


86400 seconds per day

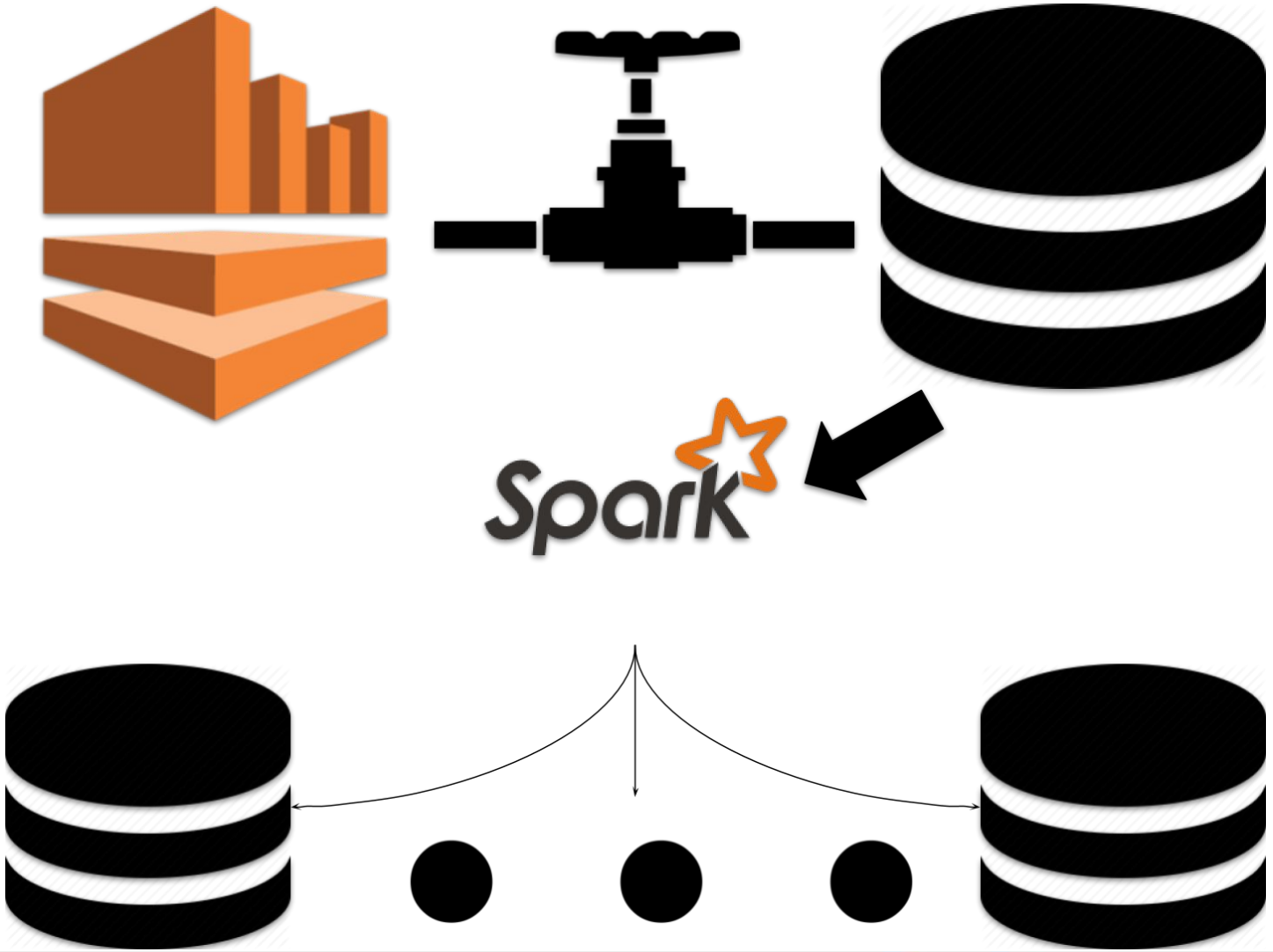
864000 miliseconds per day

Low traffic hours	10 M / 30 minutes ~ 5000 / s
High traffic hours	50 M / 30 minutes ~ 25000 / s
Peaks	> 100 M / 30 minutes > 50000 / s

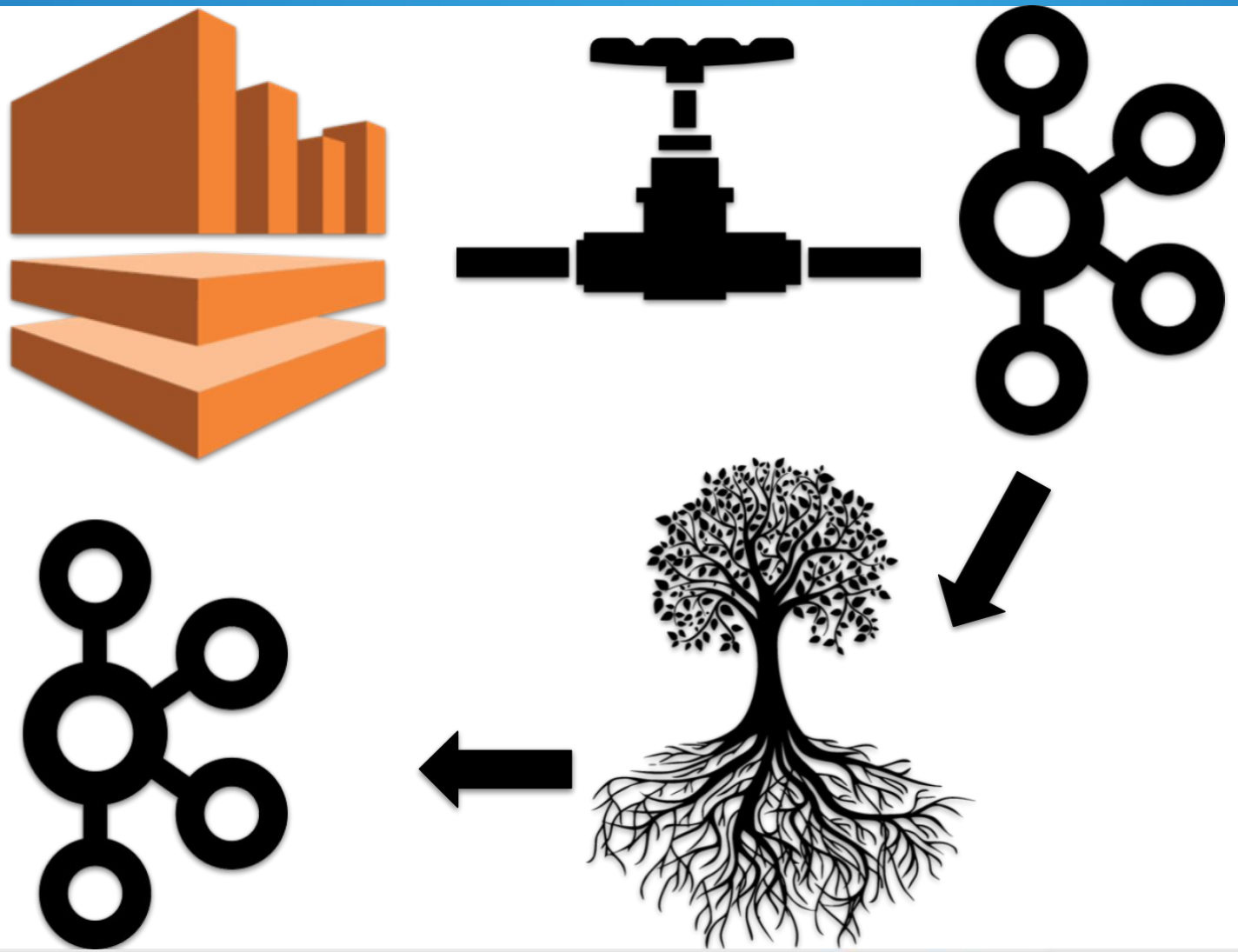
Common pipeline



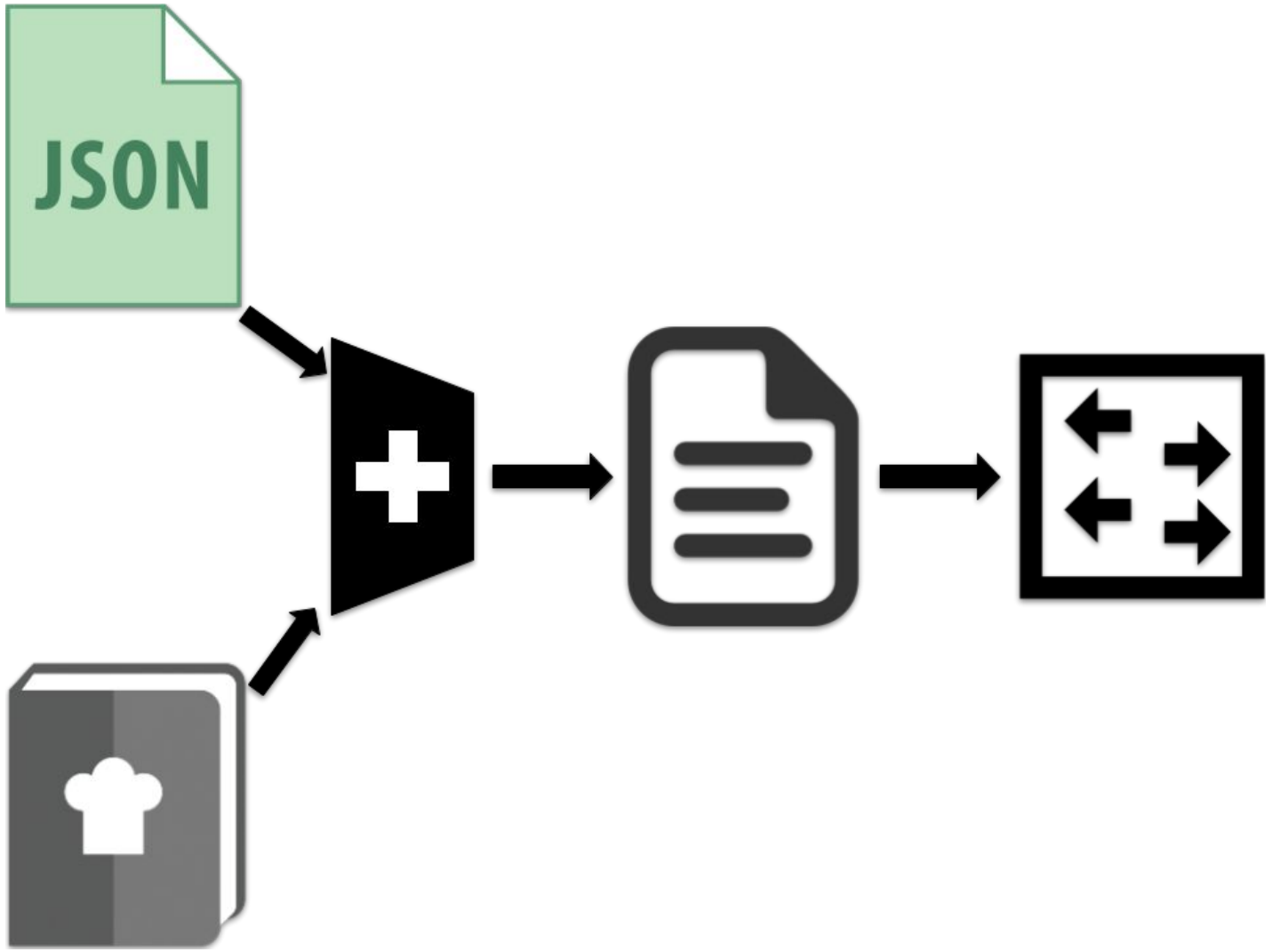
Batch pipeline



Streaming pipeline

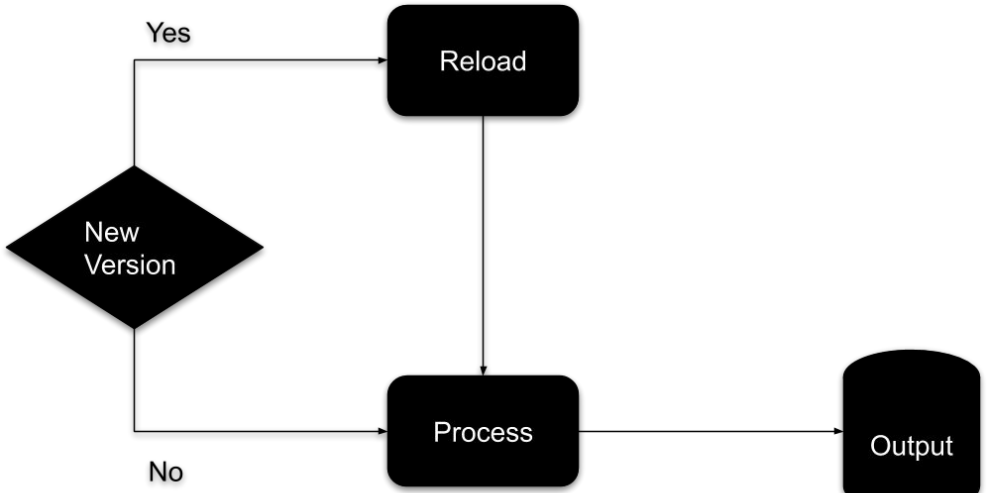


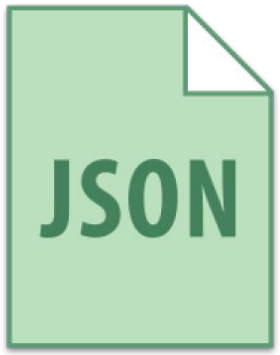
Processing and routing internals



Routing lib

VersionWatcher





Processing: routing language

```
SinkName:
  eventType: event schema
  filter: inline || stored || null
  transform: stored || null
  SinkType:
    SinkDetails:
```

```
ProbeEvent-1:
  eventType: ProbeEvent
  kafka:
    topic: probe-topic
```

Event formats: probe event

```
{
   "$schema": "http://json-schema.org/draft-04/schema#",
   "allOf": [
       {
           "$ref": "base-routable-event.json#"
       }
   ],
   "description": "Events sent by Data Platform Probe to measure latencies and missing events in the pipeline",
   "id": "http://schema.schibsted.com/events/backend-probe-event.json#",
   "properties": {
       "senderId": {
           "description": "Sender ID, in case several instances of Probe is running",
           "type": "integer"
       },
       "sequenceNumber": {
           "description": "Probe sequence number",
           "type": "integer"
       },
       "timeSent": {
           "$ref": "../common-definitions.json#/definitions/timestamp",
           "description": "UTC timestamp of when the event is generated by Probe"
       }
   },
   "title": "BackendProbeEvevnt",
   "type": "object"
}
```

JSLT: The magic sauce of processing

JSON query and transformation language

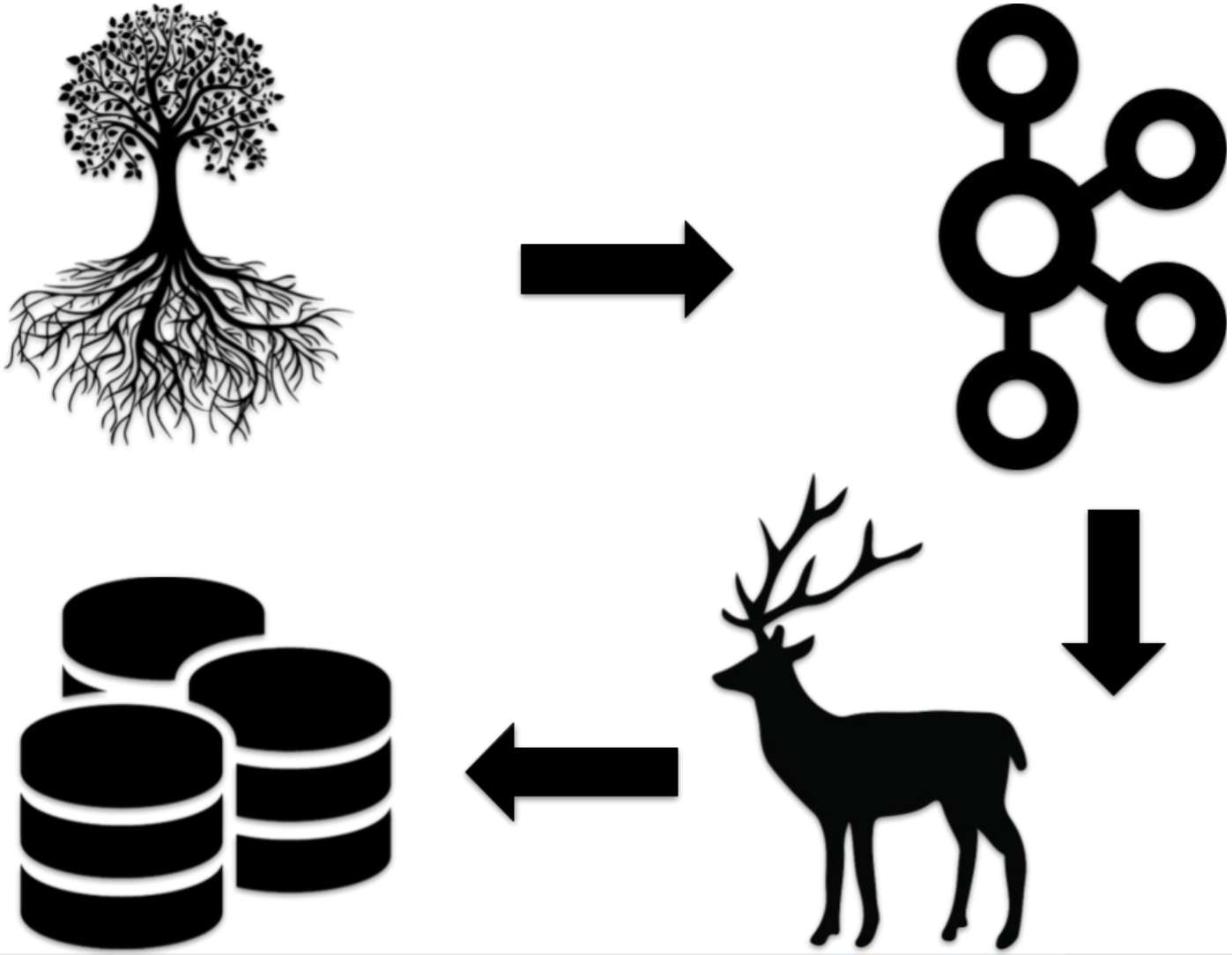
Github repo: https://github.com/schibsted/jslt

License: Apache 2.0

```
{
    "time": round(parse-time(.published, "yyyy-MM-dd'T'HH:mm:ssX") * 1000),
    "device_manufacturer": .device.manufacturer,
    "device_model": .device.model,
    "language": .device.acceptLanguage,
    "os_name": .device.osType,
    "os_version": .device.osVersion,
    "platform": .device.platformType,
    "user_properties": {
        "is_logged_in" : boolean(.actor."spt:userId")
    }
}
```

Routing: batch

Routing: streaming



Lessons learned (so far…)

- Schemas and versions
- Backfilling and recovery
- Logging and metrics
- Auditing

And finally

Extra

About Schibsted

1839 - Schibsted Forlag established

1860 - Aftenposten is founded

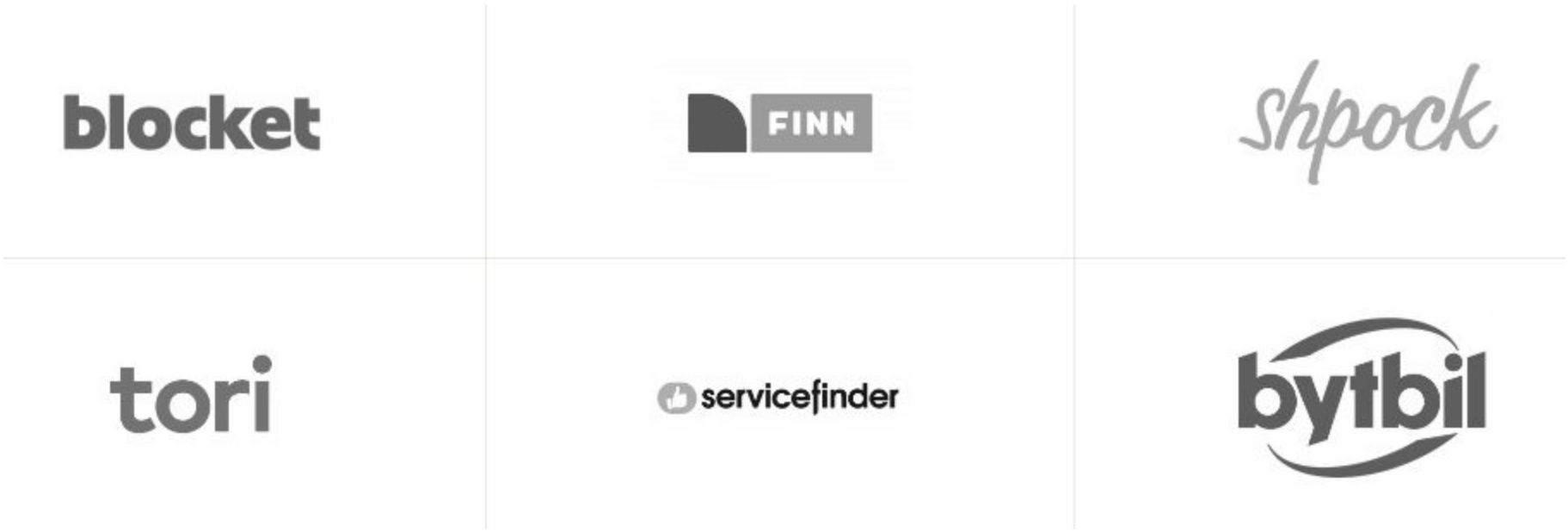
1966 - VG acquired

1992 - Listed in Oslo exchange

2000 - Finn.no launched

Today - Leader in Media and marketplaces in the Nordics

Marketplaces



News Media

VG

Aftenposten

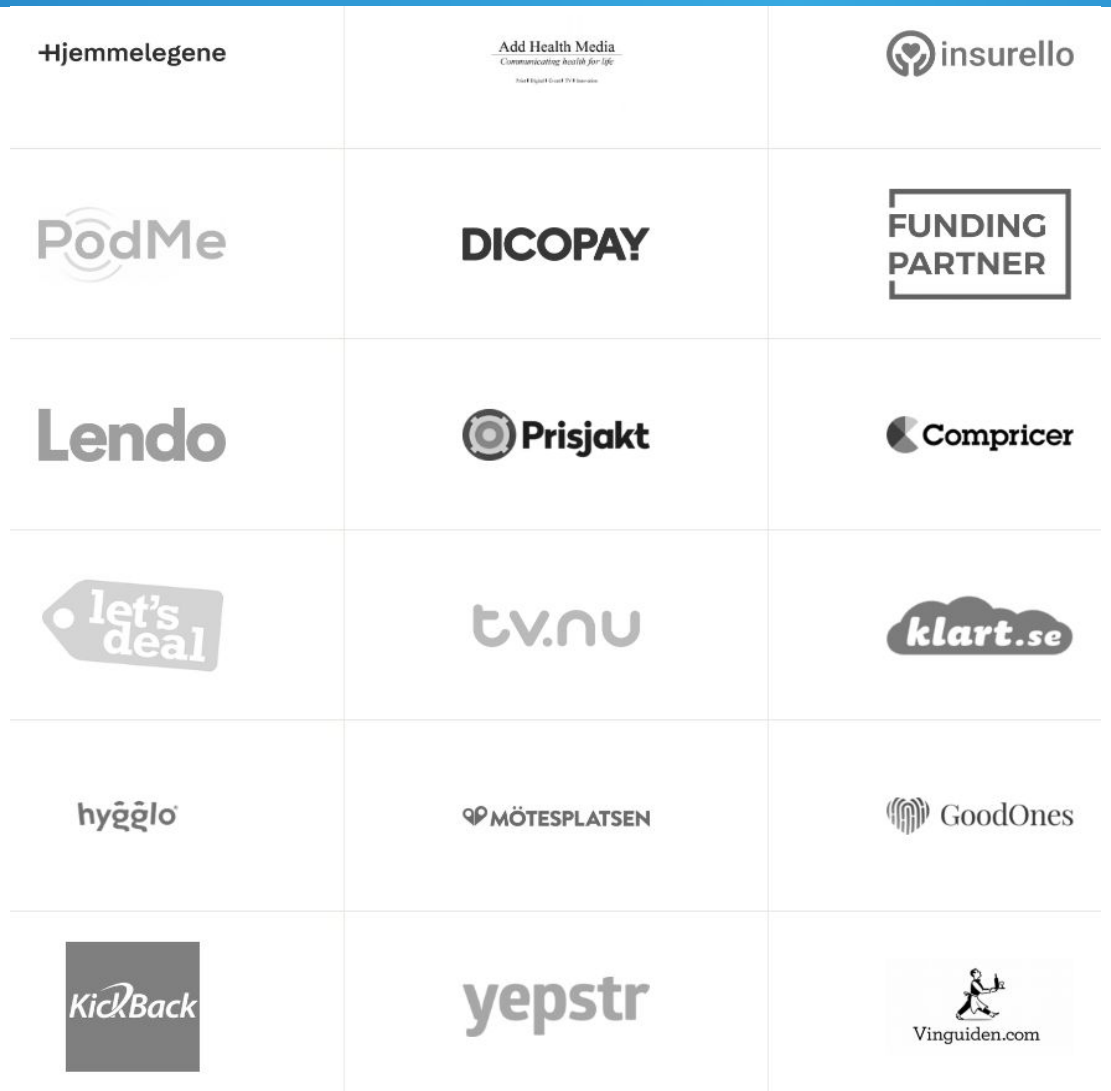
AFTONBLADET

SvD

Bergens Tidende

Omni

Fædrelandsvennen

Stavanger Aftenblad

Some of our Next companies