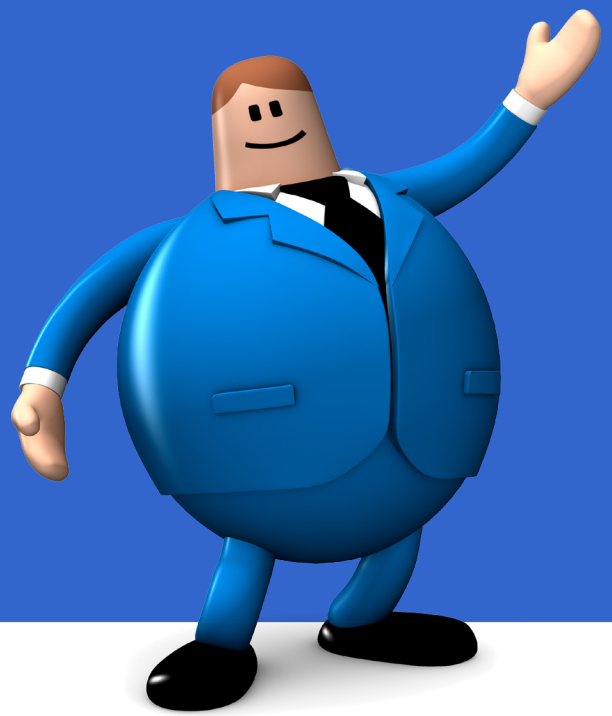
Half a year of Measuring 2.0 in production at bol.com



bol.com
de winkel van ons allemaal

Agenda

- Motivation
- Design
- Production experiences
  - Measurements
  - Kafka
  - Trending products
  - Avro/Parquet
- Future plans

bol.com

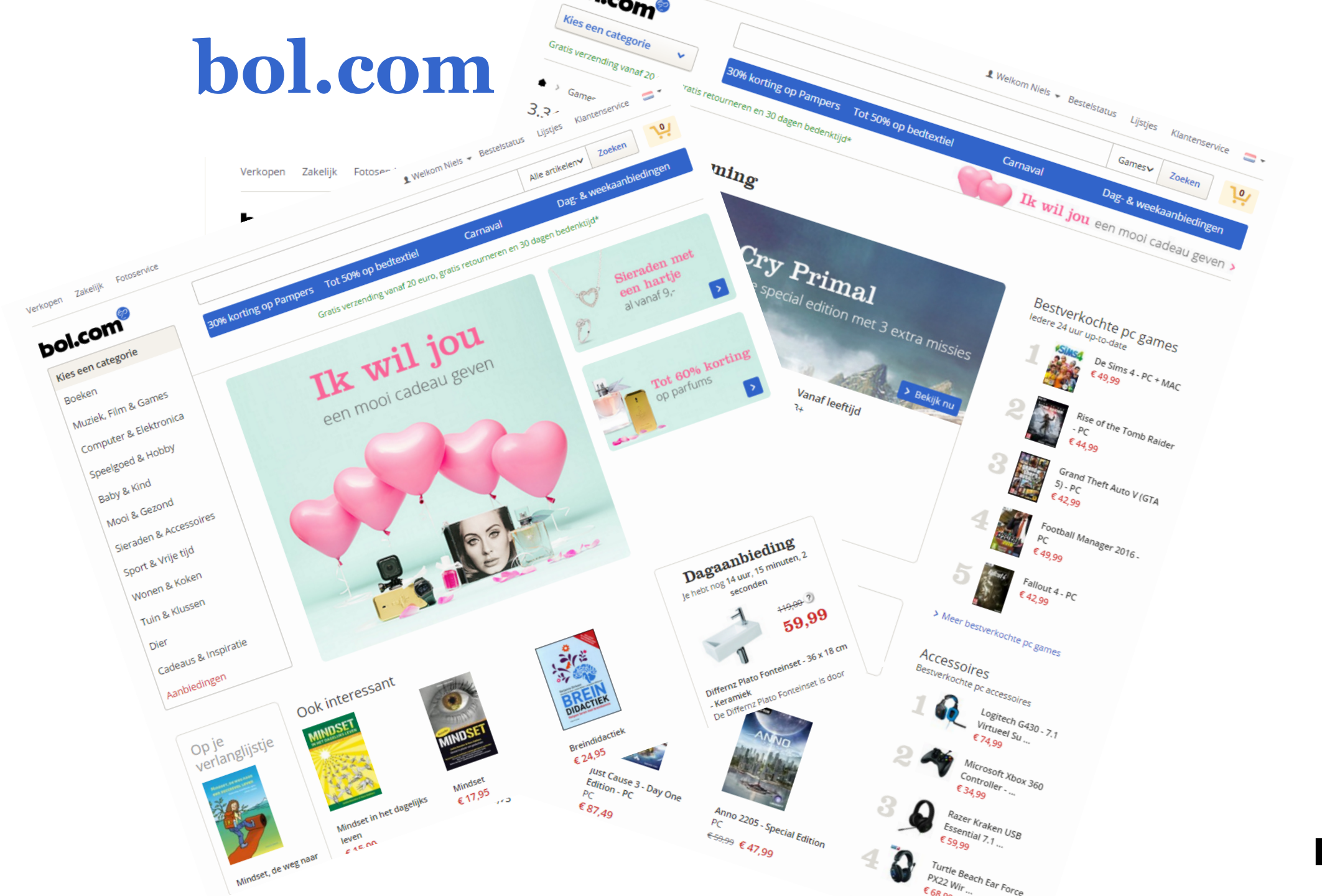
Ivan Budinčević

- ibudincevic@bol.com
- PhD in nuclear physics KU Leuven 2015
- Data analysis with python
- Java developer at bol.com since 2016
- Interested in big data technologies, streaming data processing

bol.com

Motivation

bol.com



bol.com

bol.com

bol.com

bol.com

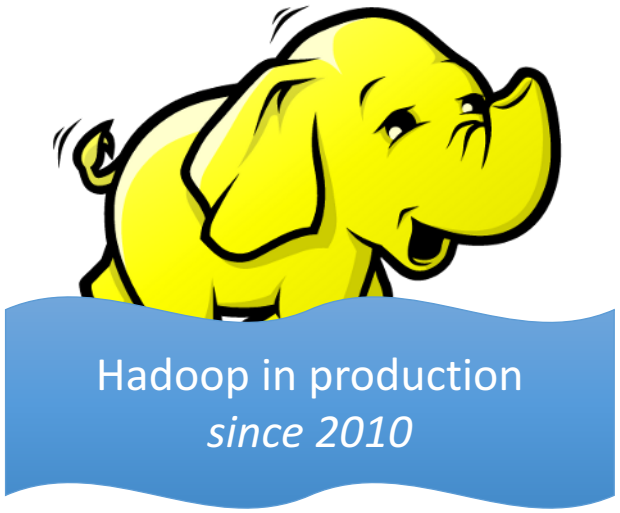
bol.com

> 15 million products for sale
> 50 million in catalog

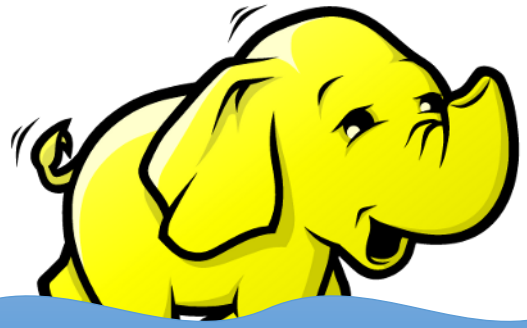
> 6 million active customers
> 27 million visits per month

> 4 billion
pageviews/year

bol.com



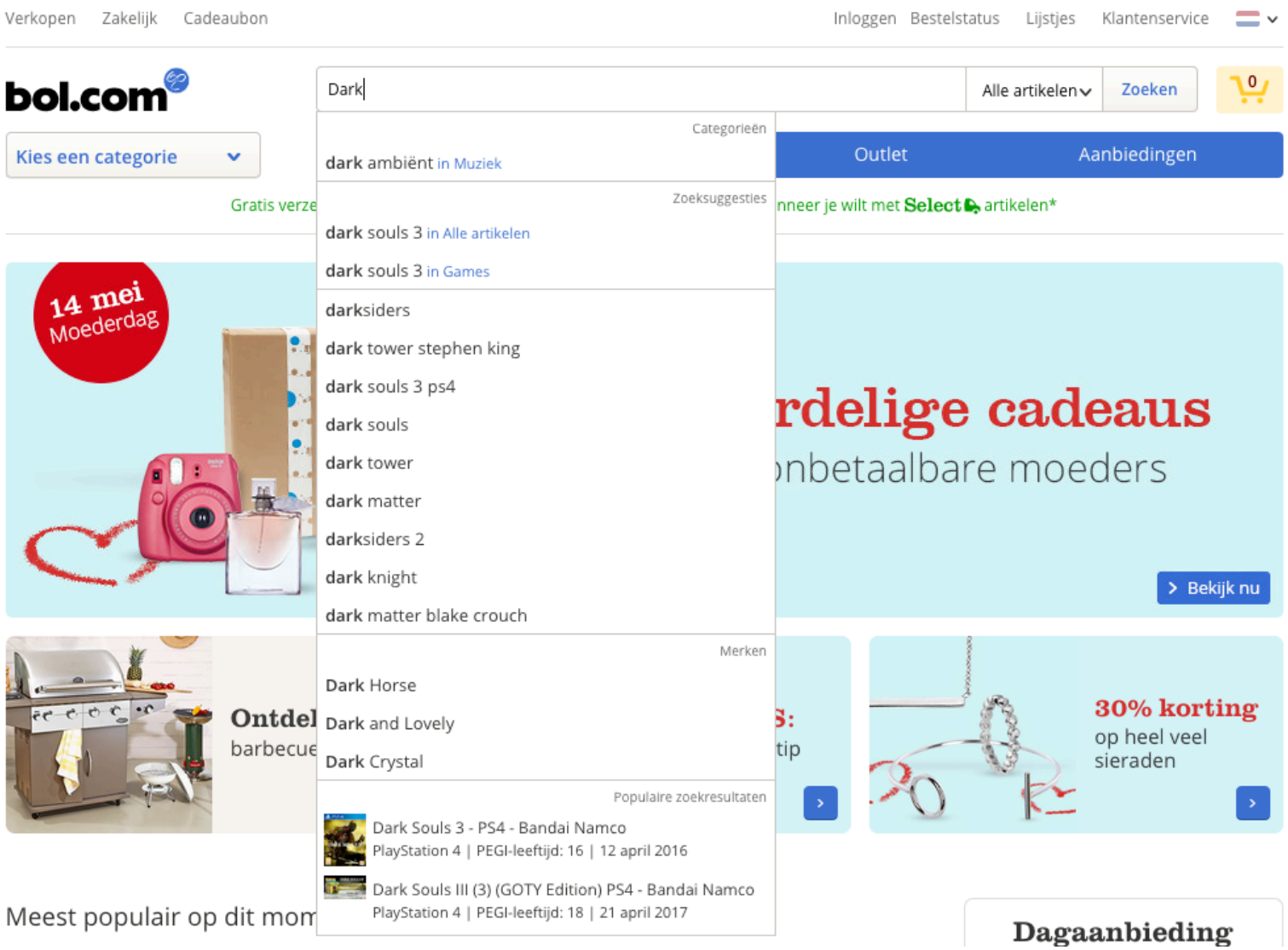
bol.com



Hadoop in production
since 2010

bol.com

Search suggestions



bol.com

Personalization

Je bekeek Ook interessant:

In The Mood Wicker - Sierkussen - Paa...

In The Mood Pailletten Klein -...

€ 18,99

In The Mood Turijn - Sierkussen - Paa...

€ 13,99

In The Mood Koffieboon -...

€ 18,99

› Bekijk al je persoonlijke aanbevelingen

Speciaal voor jou geselecteerd

Microsoft Office 365 Personal - Neder...

€ 69,00

12 - 99 jaar

United Entertainment - Magic Wine Dec...

€ 17,95

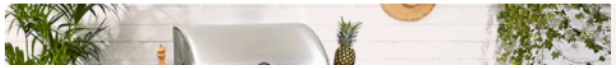
Microsoft Windows 7 Professional - En...

€ 69,99

Microsoft Office 365 Home - Nederland...

€ 99,00

› Bekijk al je persoonlijke aanbevelingen



Dagaanbieding

Je hebt nog **10 uur, 40 minuten, 35 seconden**

Populaire cd's en films voor 10 euro!

Alleen vandaag: populaire cd's en films waar moeders dol op zijn voor maar 10 euro! Zoals de nieuwe John Mayer, O'G3NE en Soof 2!

› Naar de actie

› Ontvang elke dag de dagaanbieding

Tip voor jou

Misschien heb je interesse in dit artikel?

Dark Souls 3 - Windows

Windows

€ 29,99

bol.com

Personalization

Je bekeek Ook interessant:

In The Mood Wicker - Sierkussen - Paa...

In The Mood Pailletten Klein -...

€ 18,99

In The Mood Turijn - Sierkussen - Paa...

€ 13,99

In The Mood Koffieboon -...

€ 18,99

› Bekijk al je persoonlijke aanbevelingen

Speciaal voor jou geselecteerd

Microsoft Office 365 Personal - Neder...

€ 69,00

12 - 99 jaar

United Entertainment - Magic Wine Dec...

€ 17,95

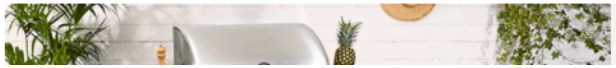
Microsoft Windows 7 Professional - En...

€ 69,99

Microsoft Office 365 Home - Nederland...

€ 99,00

› Bekijk al je persoonlijke aanbevelingen



Dagaanbieding

Je hebt nog **10 uur, 40 minuten, 35 seconden**

Populaire cd's en films voor 10 euro!

Alleen vandaag: populaire cd's en films waar moeders dol op zijn voor maar 10 euro! Zoals de nieuwe John Mayer, O'G3NE en Soof 2!

› Naar de actie

› Ontvang elke dag de dagaanbieding

Tip voor jou

Misschien heb je interesse in dit artikel?

Dark Souls 3 - Windows

Windows

€ 29,99

bol.com

How do we react to customer needs in real-time?

bol.com

How do we react to customer needs in real-time?

- Measure customer behavior and take action in real-time

bol.com

How do we react to customer needs in real-time?

- Measure customer behavior and take action in real-time

- Measure EVERYTHING

bol.com

Goals of “Measuring 2.0”

- Measure everything on our website
  - All interactions (also AJAX)
  - All channels (also mobile, email, …)
  - All countries
  - All details
  - All visitors
- More reliable data
- Lowest possible load on the client
- Lowest possible latency (< 1 second)

bol.com

Goals of “Measuring 2.0”

- Developer
  - Easy to build
  - Easy to validate
  - Test automation
- Business
  - Always measure everything
  - Data is “independent”
    - New questions are allowed

bol.com

Goals of “Measuring 2.0”

- Privacy
  - Personal Data Protection Act
    - *Wet bescherming persoonsgegevens*
  - No profiling beyond 2 years
- Security
  - Avoid storing “login” info.
- Business
  - Do “profiling” on everything
  - For many years (>2)

bol.com

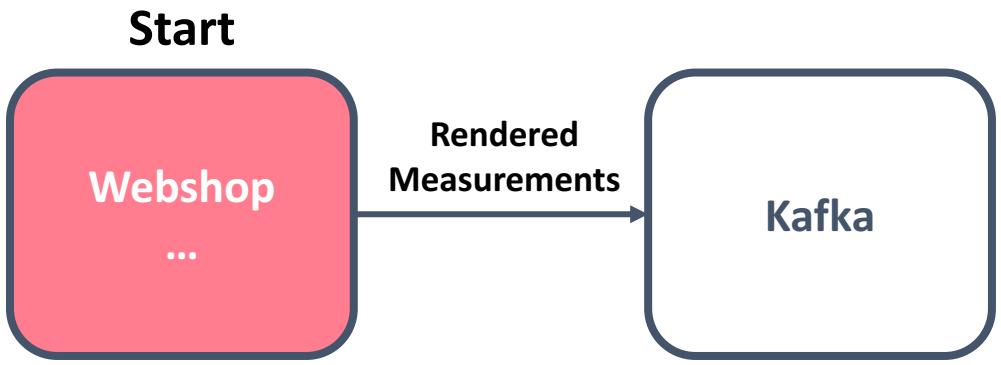
Design

Design

Start

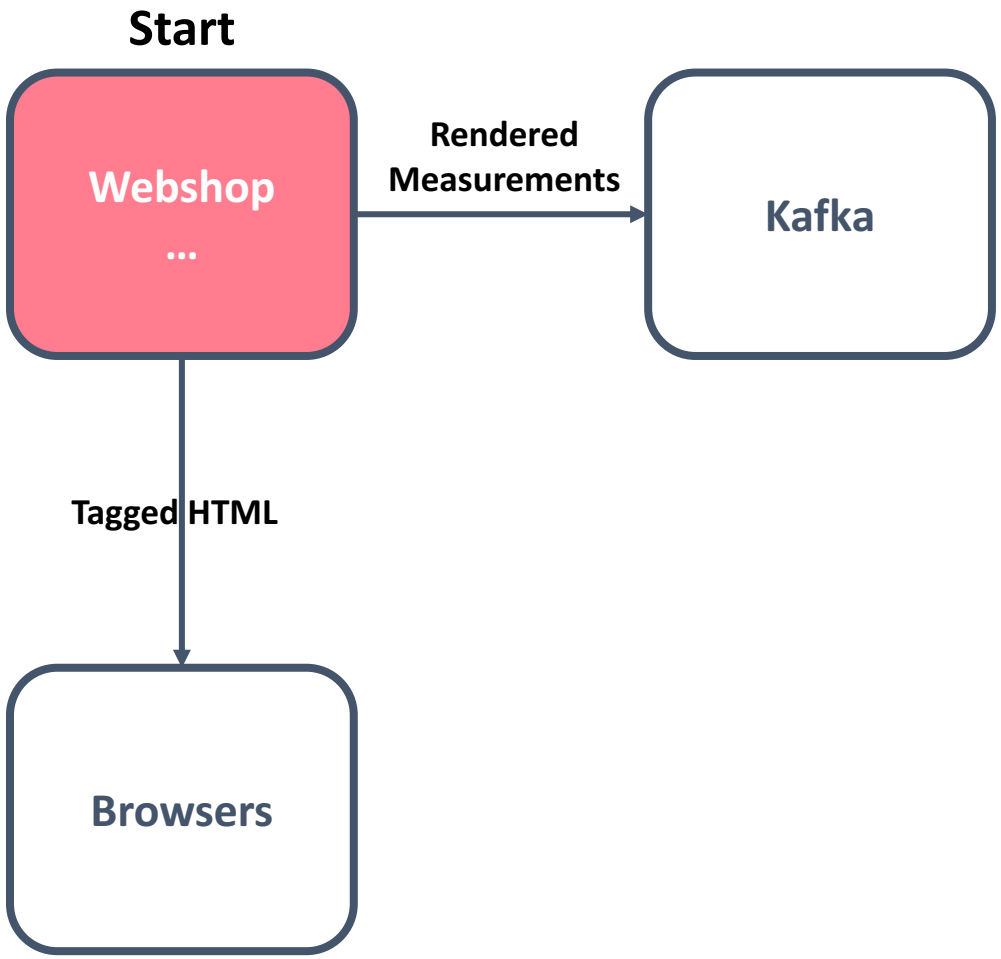
bol.com

Design



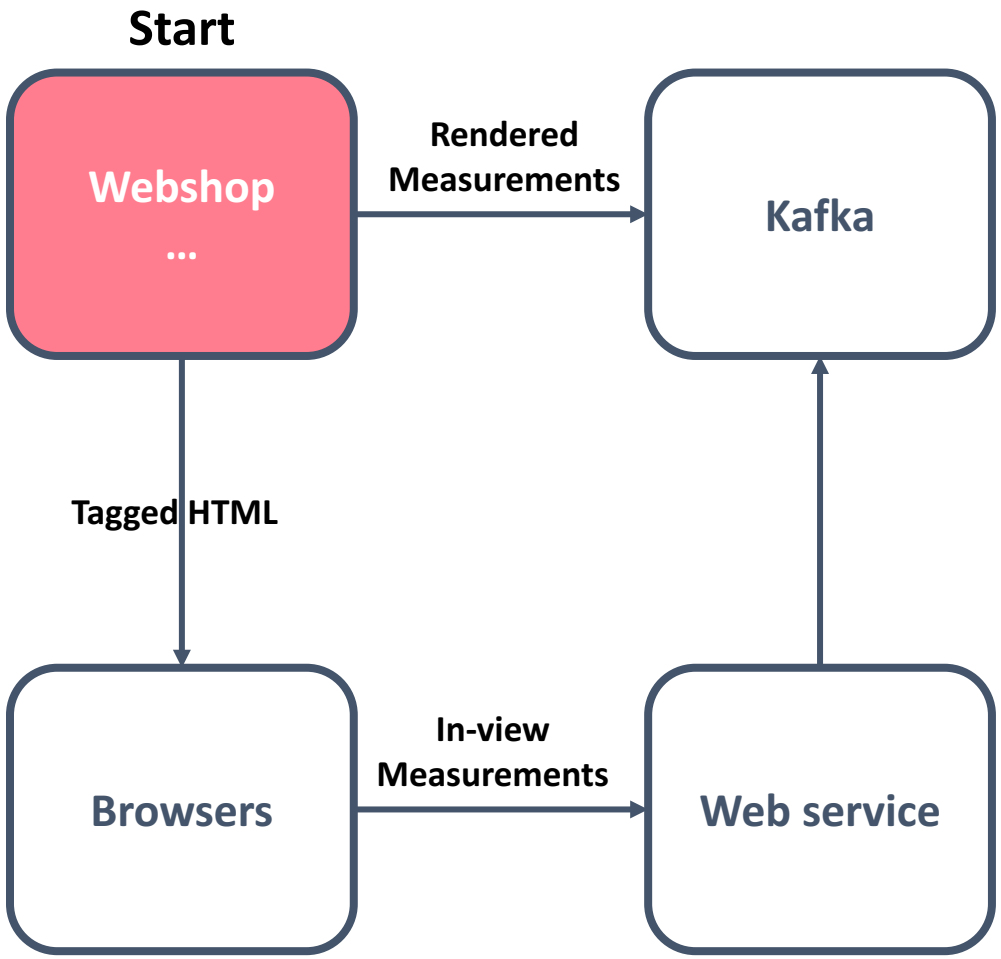
bol.com

Design



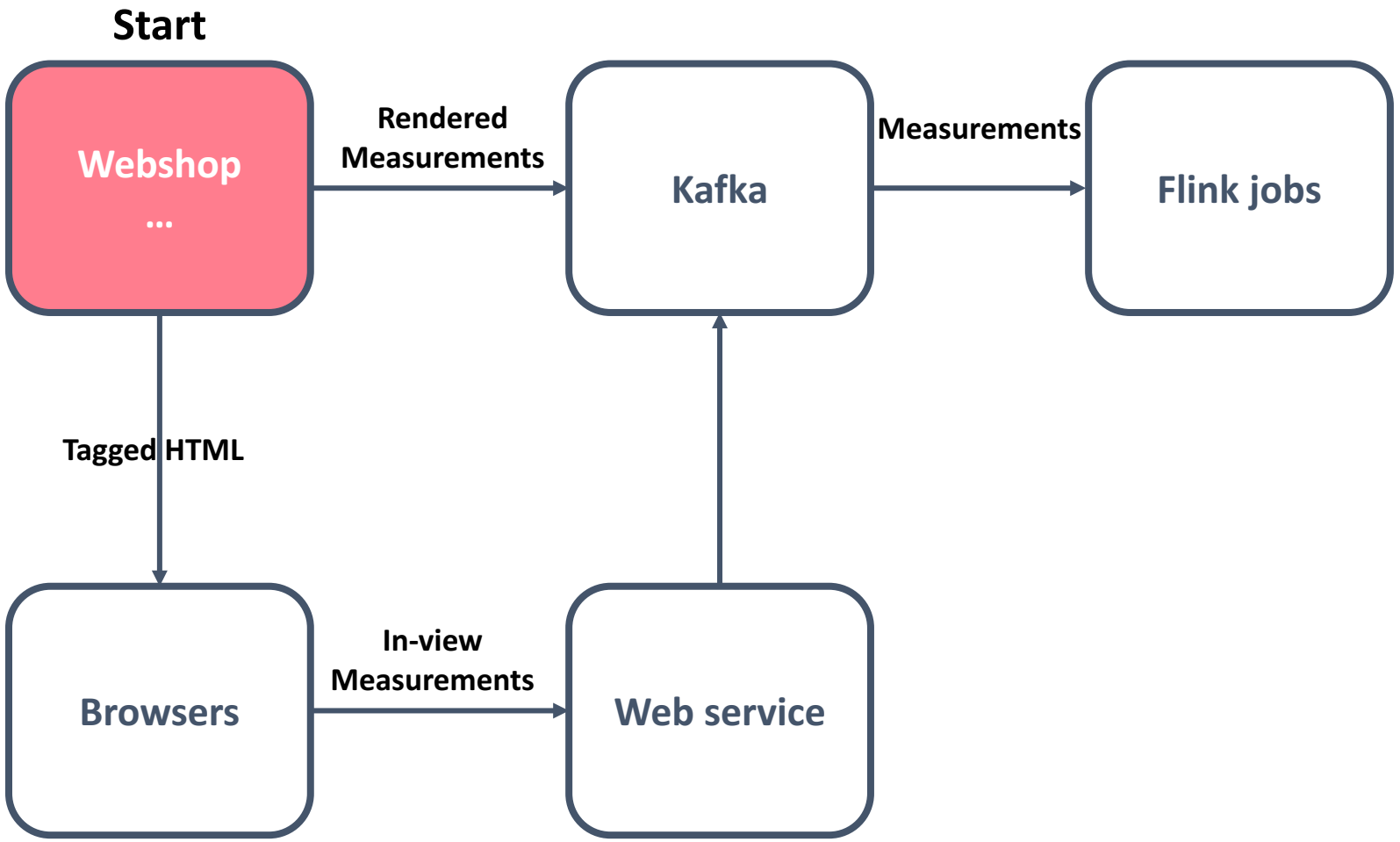
bol.com

Design



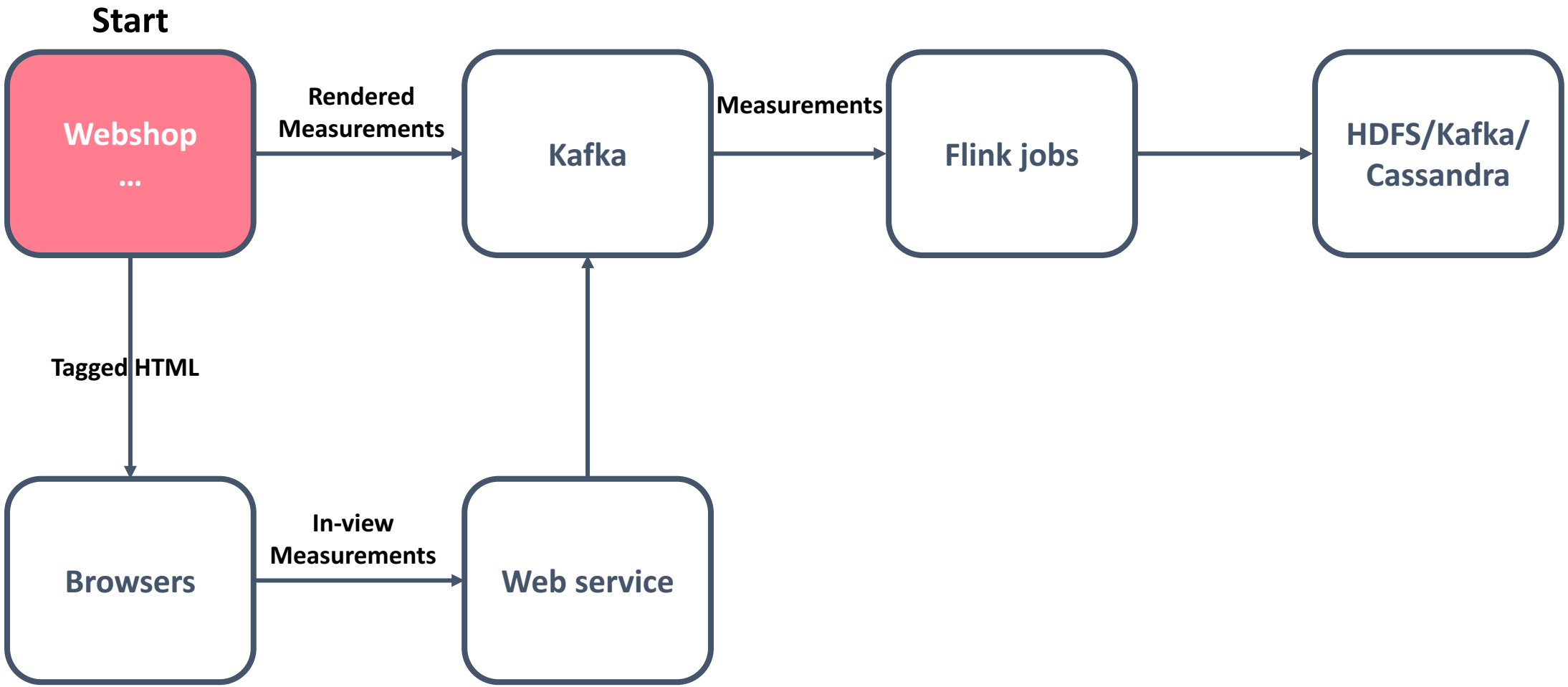
bol.com

Design



bol.com

Design



bol.com

Data structure

Verkopen Zakelijk Cadeaubon Inloggen Bestelstatus Lijstjes Klantenservice

bol.com

Waar ben je naar op zoek? Alle artikelen Zoeken 2 Bestellen

Kies een categorie 1+1 gratis op Nivea Sun Korting op tuinsets & parasols Outlet Aanbiedingen

Gratis verzending vanaf 20 euro, gratis retourneren, bezorging waar en wanneer je wilt met Select artikelen*

BulkAlert

Winkelwagentje

Bestel nu en je bestelling wordt gratis verzonden!

The Elder Scrolls Online: Morrowind - Collector's Edition - Windows
Bethesda
Windows

Nog niet verschenen - reserveer een exemplaar Beschikbaar op 06-06-2017

Aantal 1

€ 99,99

Verplaats naar verlanglijstje
Verwijder

Hand Fidget Spinner Glow in the Dark - met hoogwaardige keramische lagers

Op voorraad. Voor 16:00 uur besteld, morgen in huis

Verkoper: **BestGadgets4U**

Aantal 1

€ 11,99

Verplaats naar verlanglijstje
Verwijder

Cadeaukaart- of kortingscode

Code invoeren

Totaal artikelen (2)	€ 111,98
Verzendkosten	Gratis
Totaal	€ 111,98

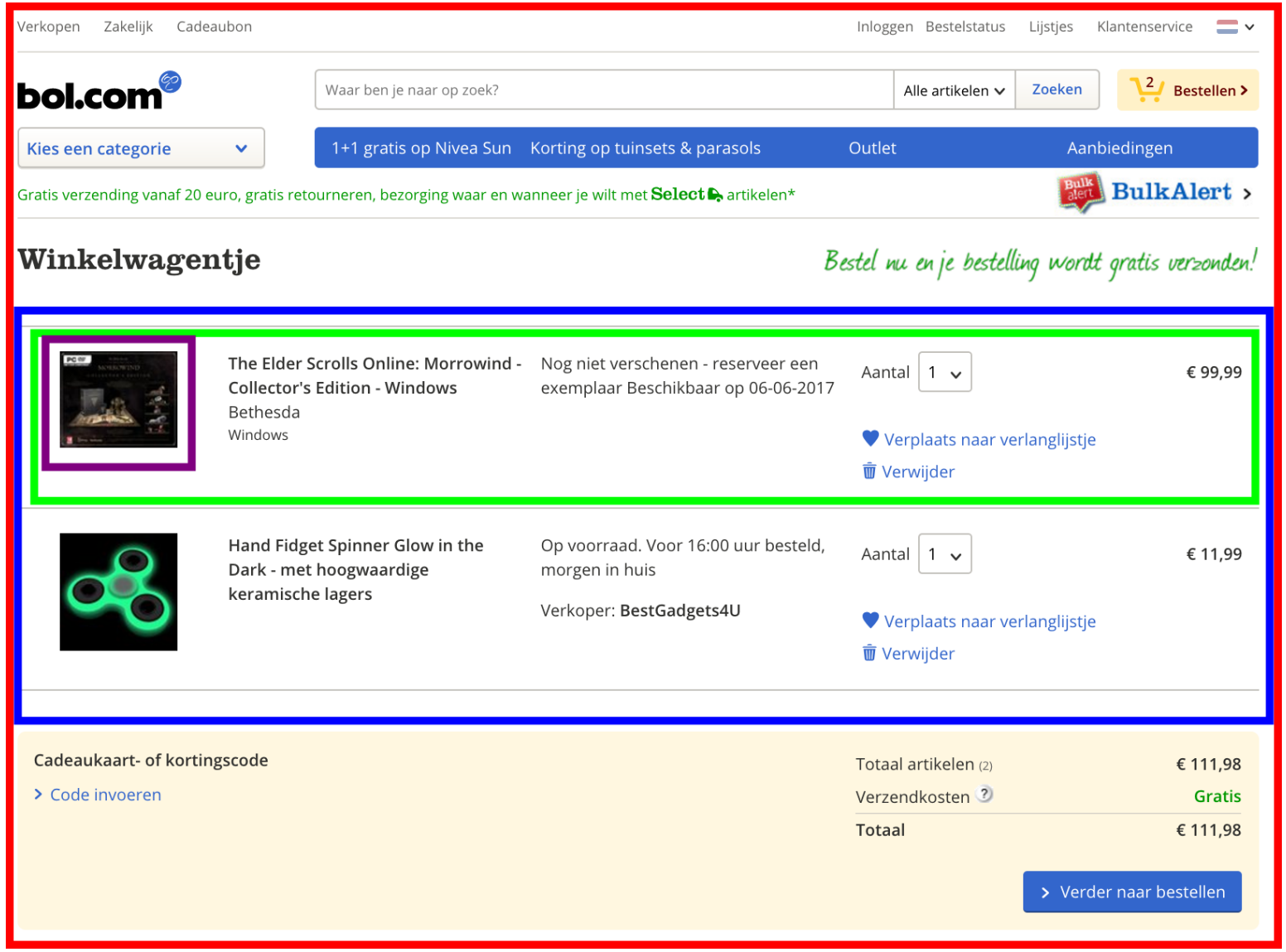
Verder naar bestellen

bol.com

Data structure

Measurements

- Container — ID = a
- Group — ID = a.b
- Item — ID = a.b.c
- Hook — ID = a.b.c.ProductImage

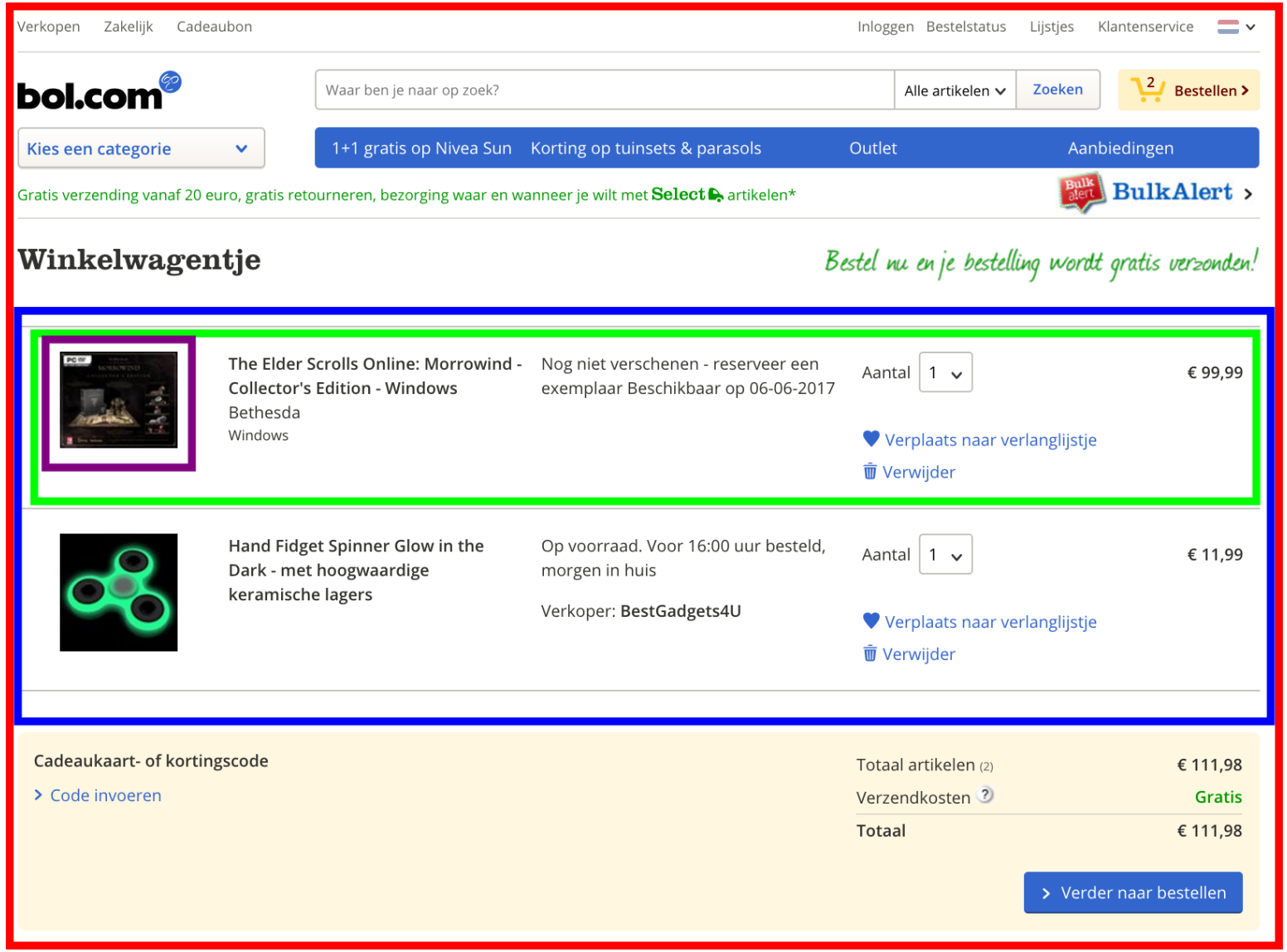


Data structure

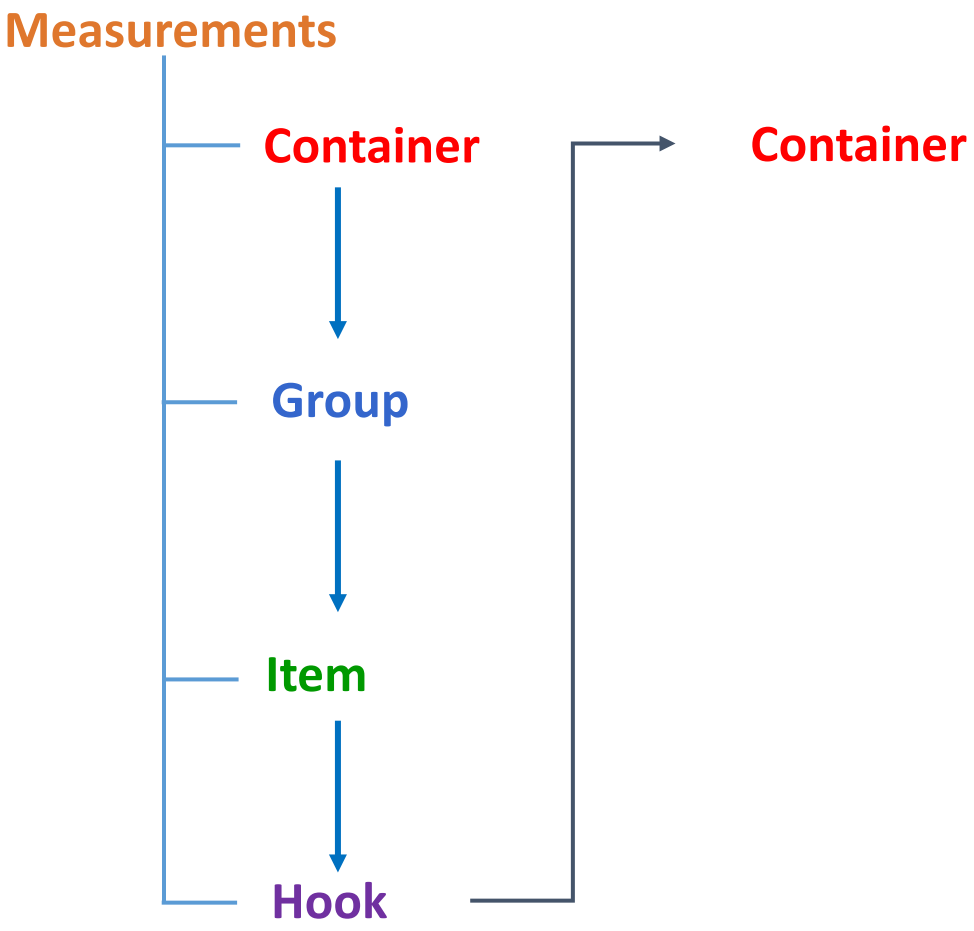
Measurements

- Container
- Group
- Item
- Hook

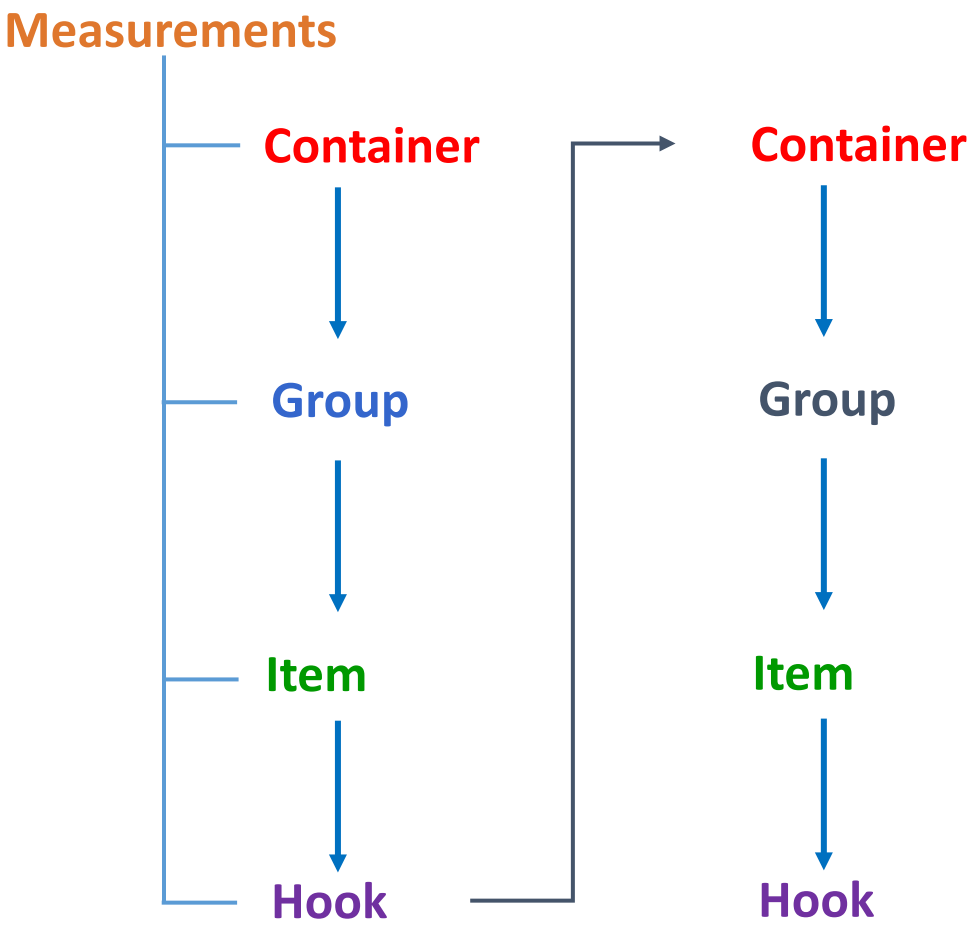
https://www.bol.com/nl/p/daten-is-meten/1001004010720589/?bltgh=ItemID.ProductImage

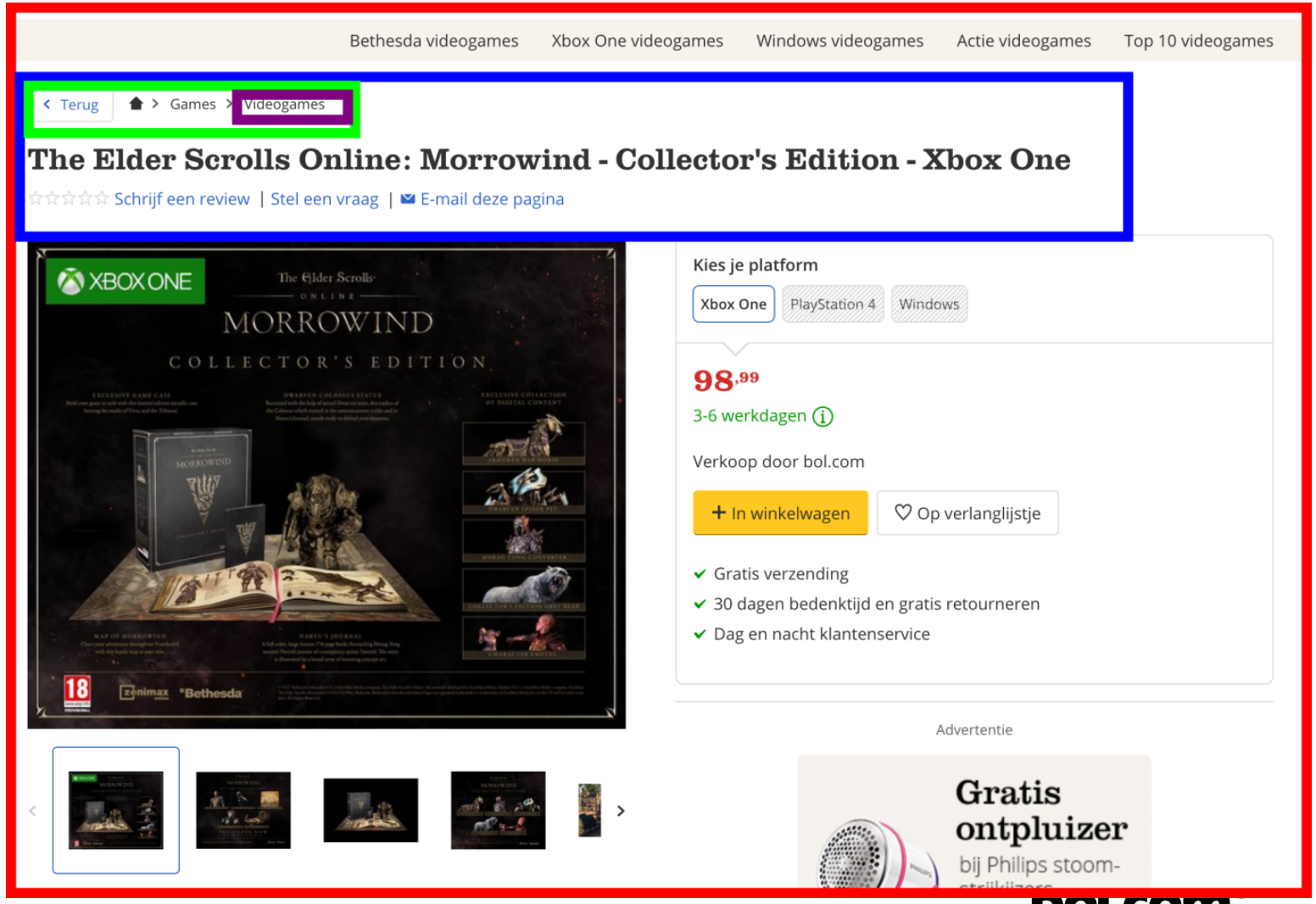


Data structure

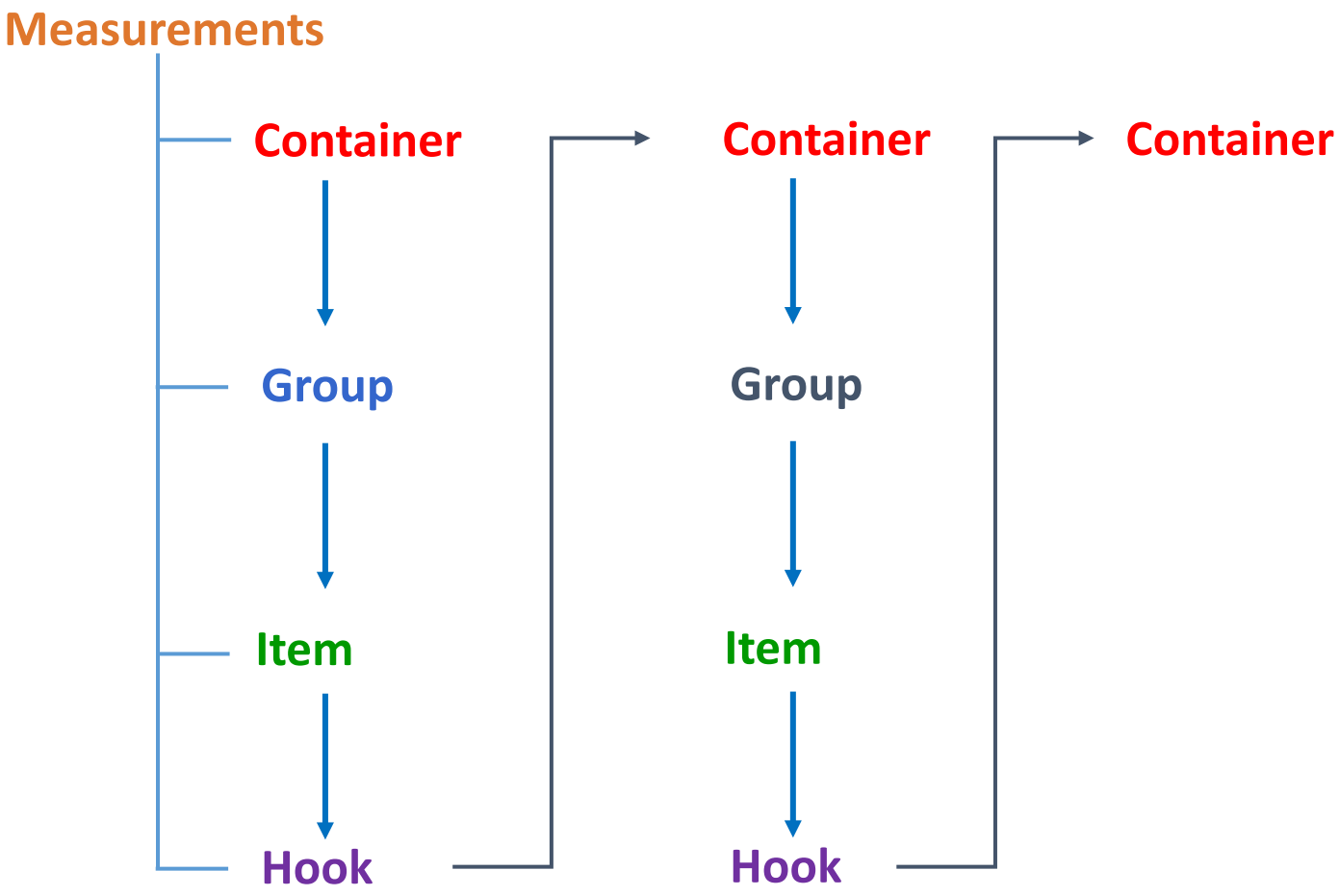


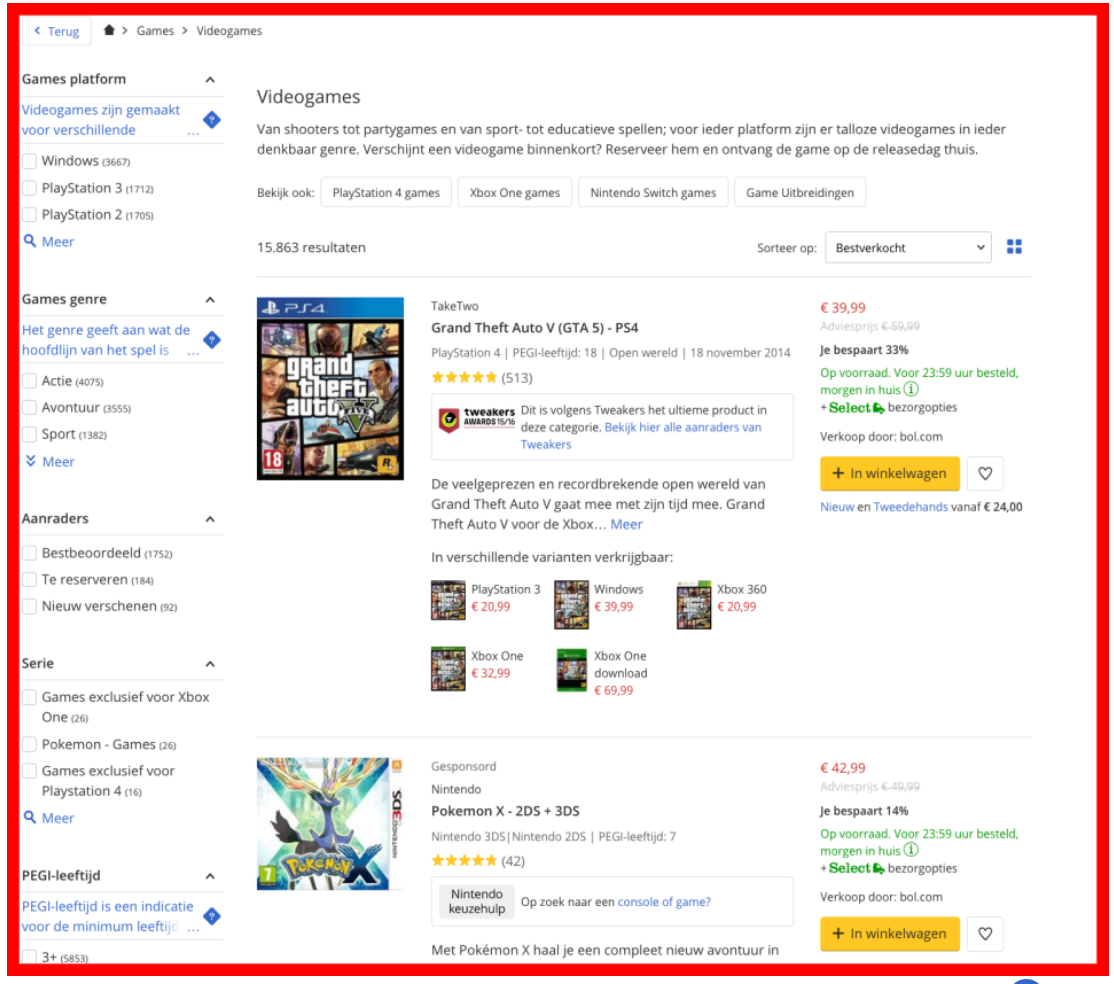
Data structure





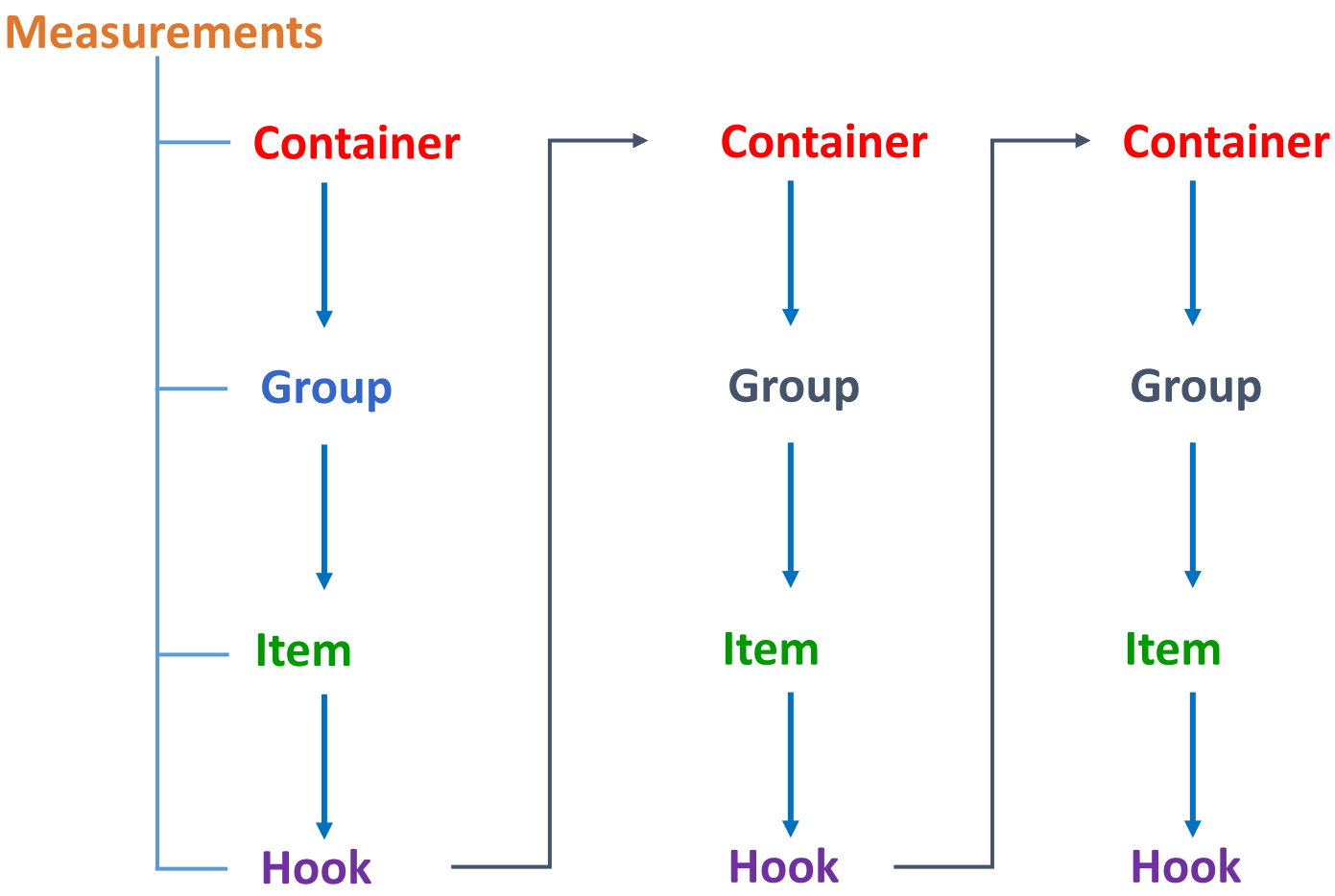
Data structure

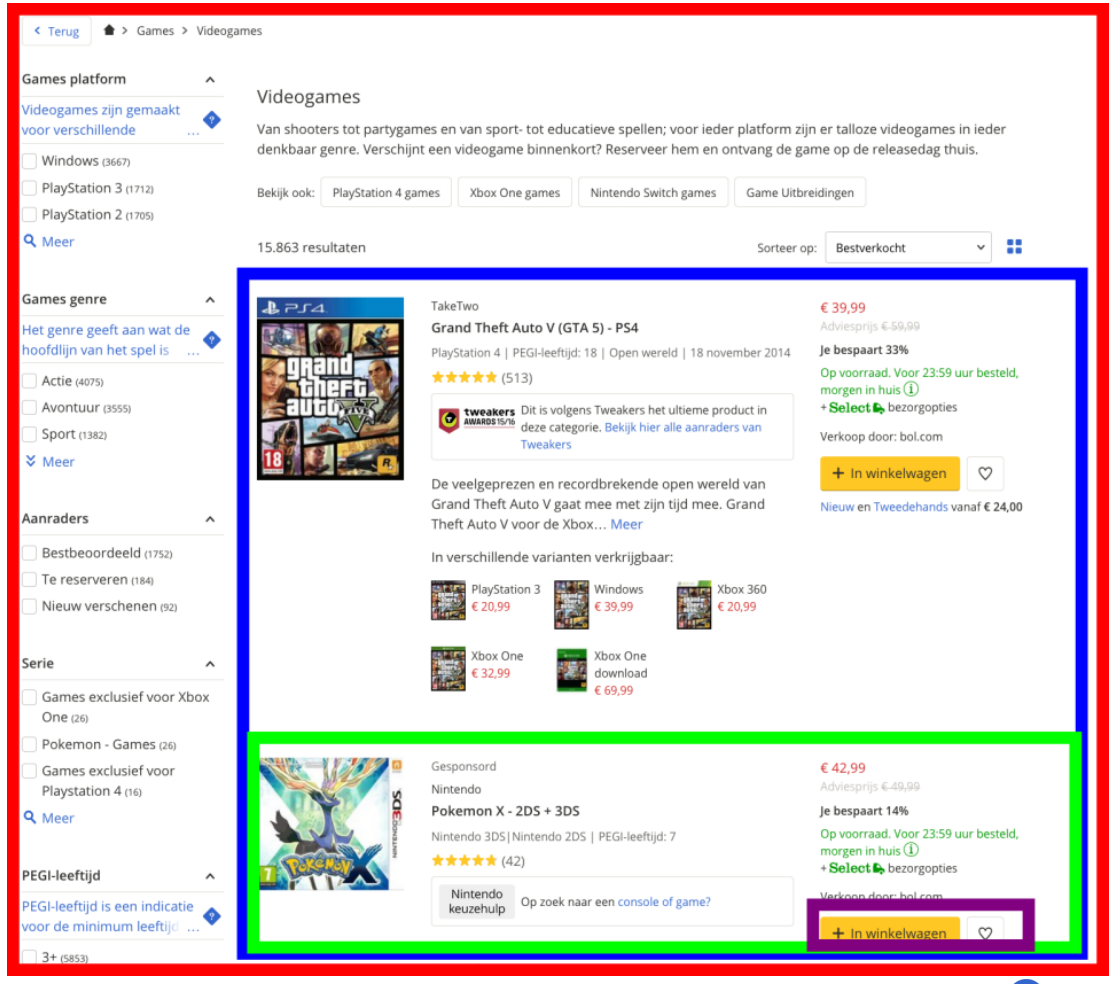




bol.com

Data structure

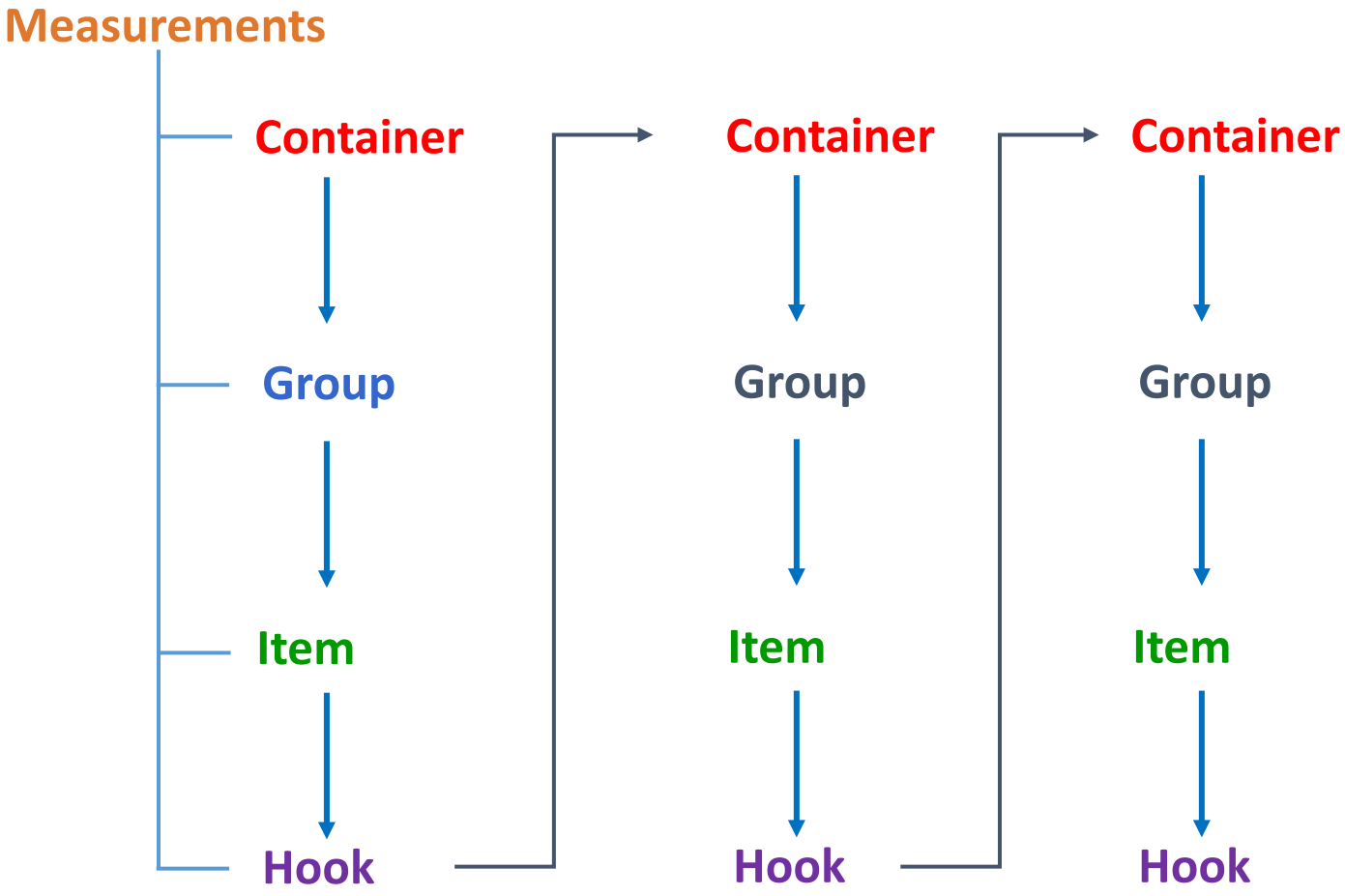


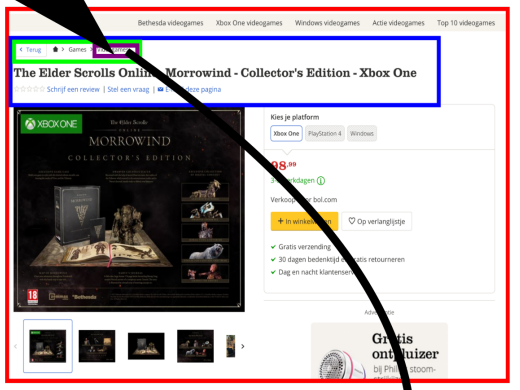


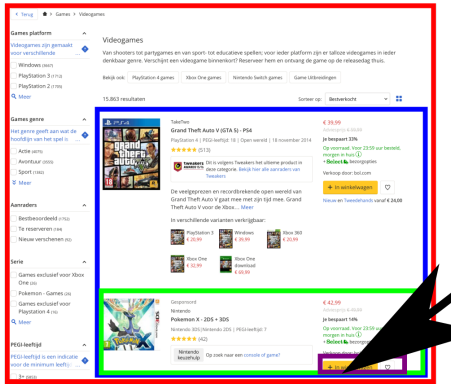
bol.com

Data structure

Measurements





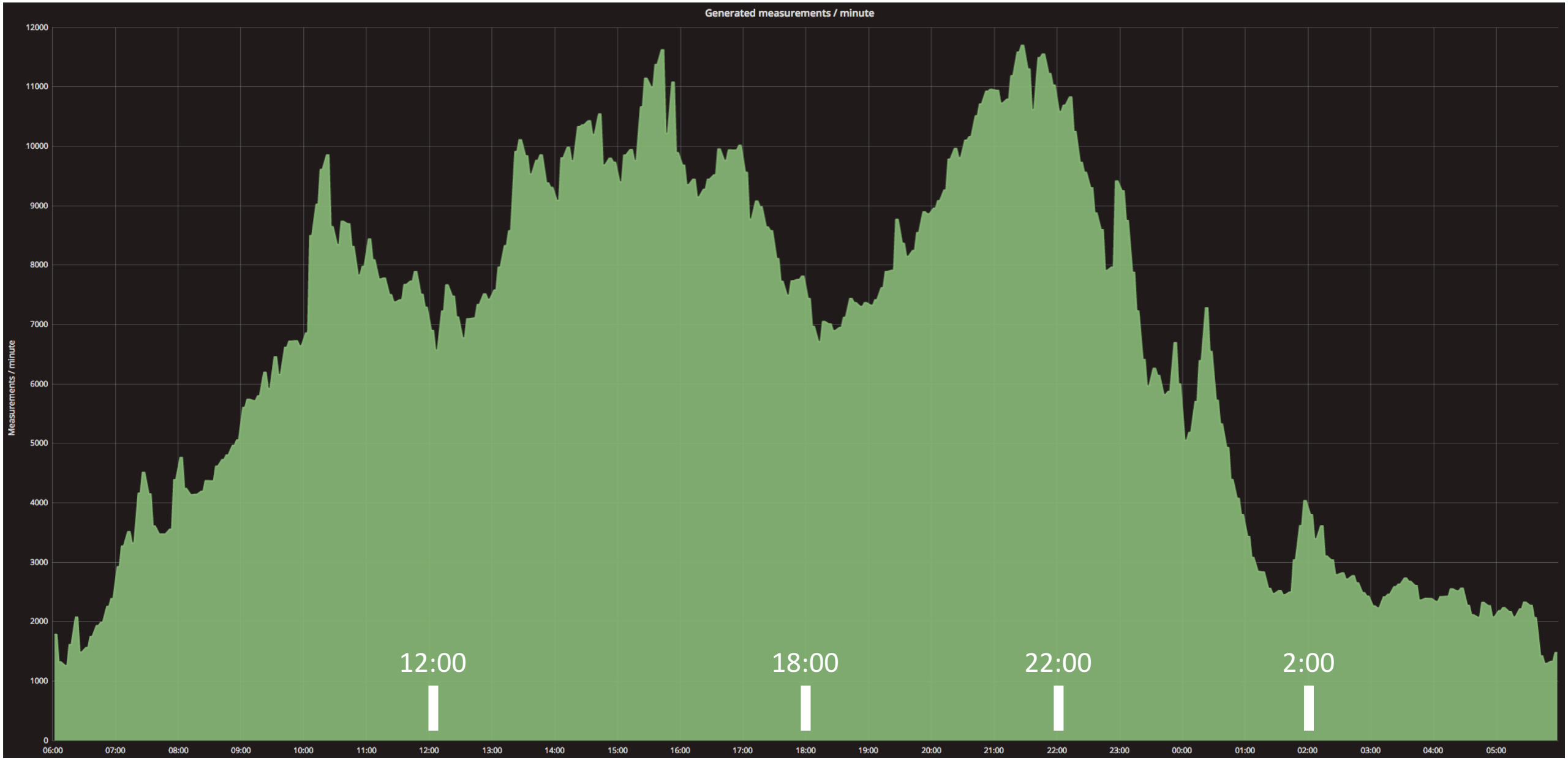


Customer's journey with cause and effect

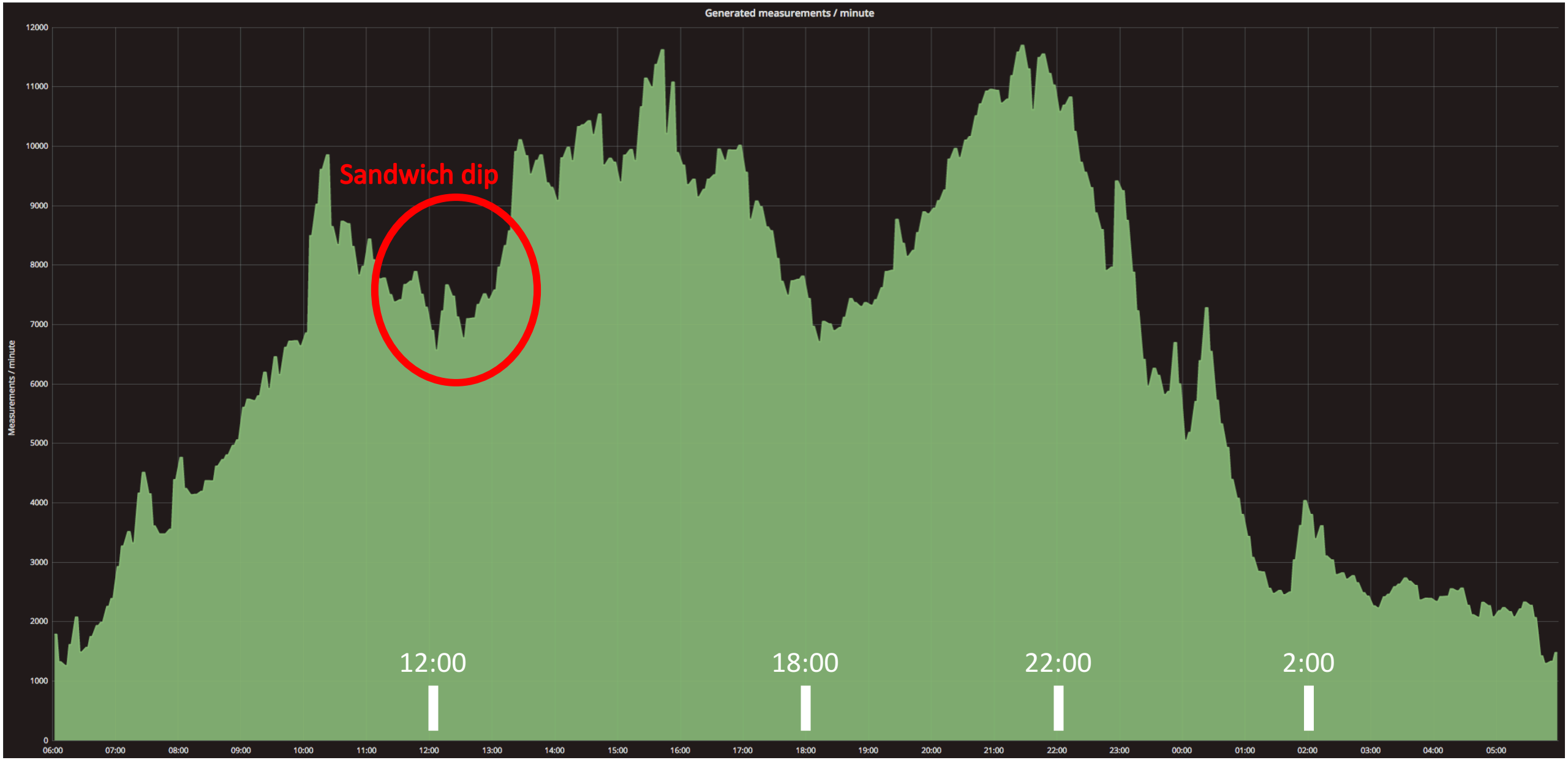
bol.com

Production experiences

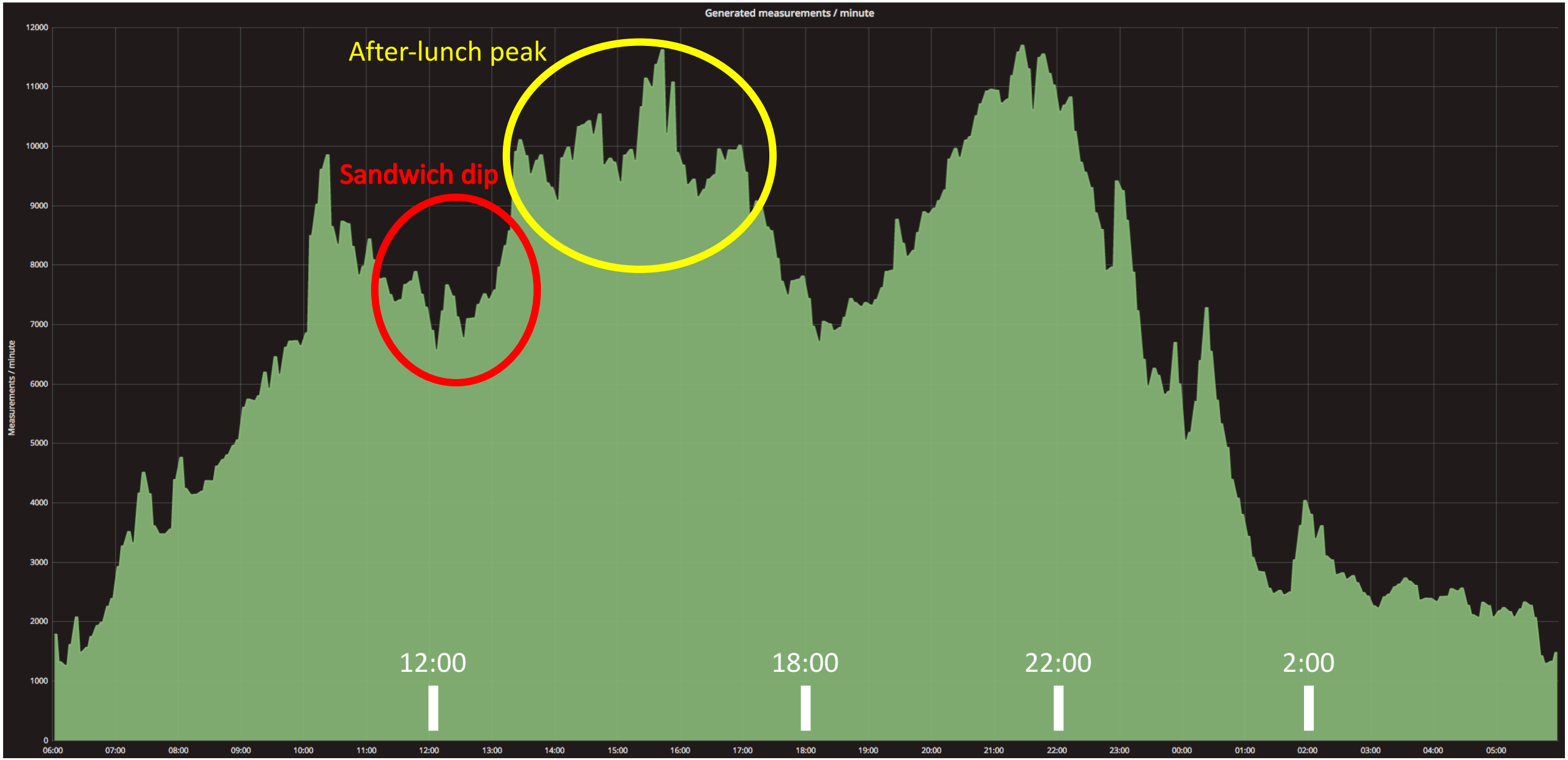
Production measurements



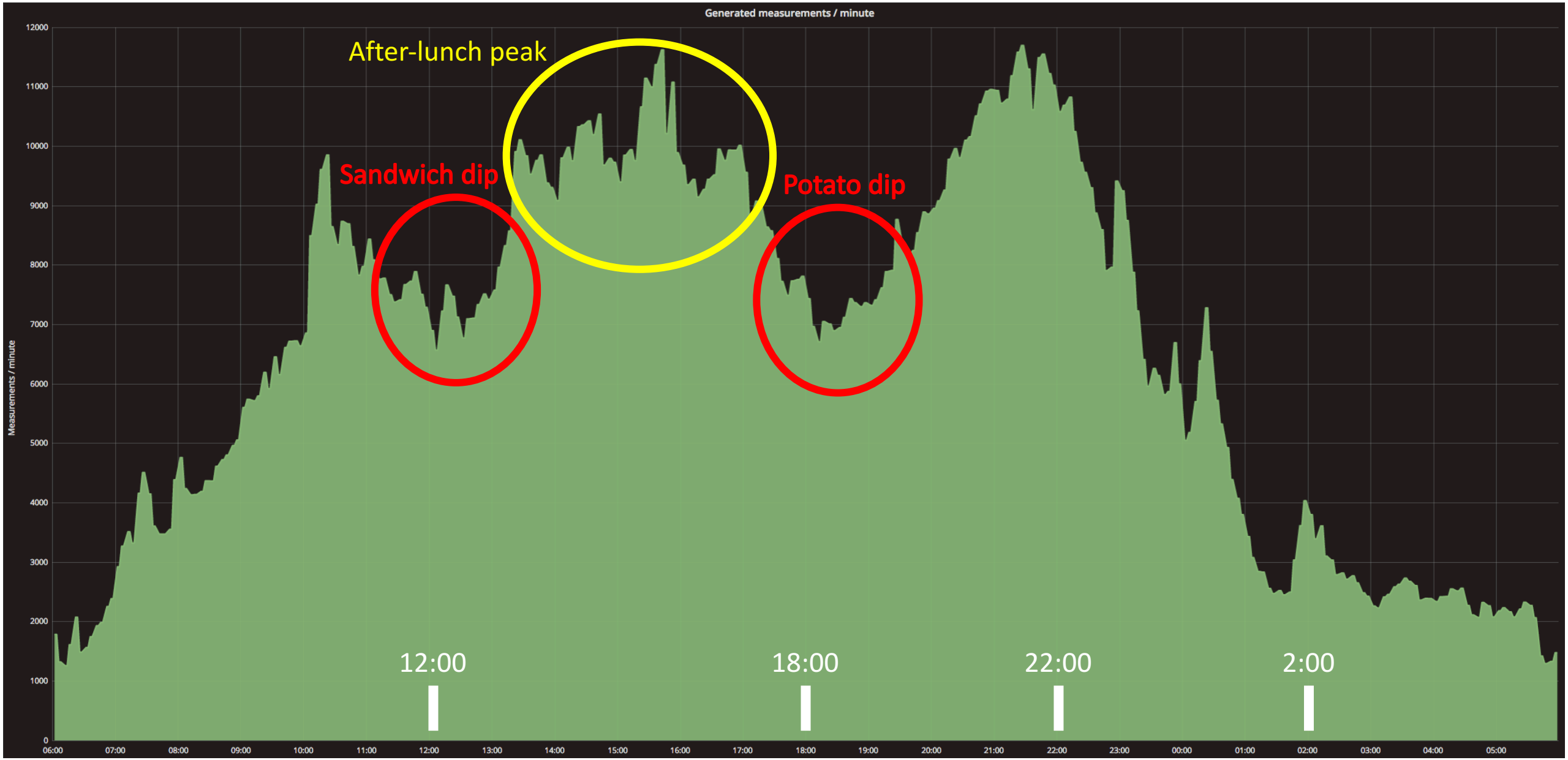
Production measurements



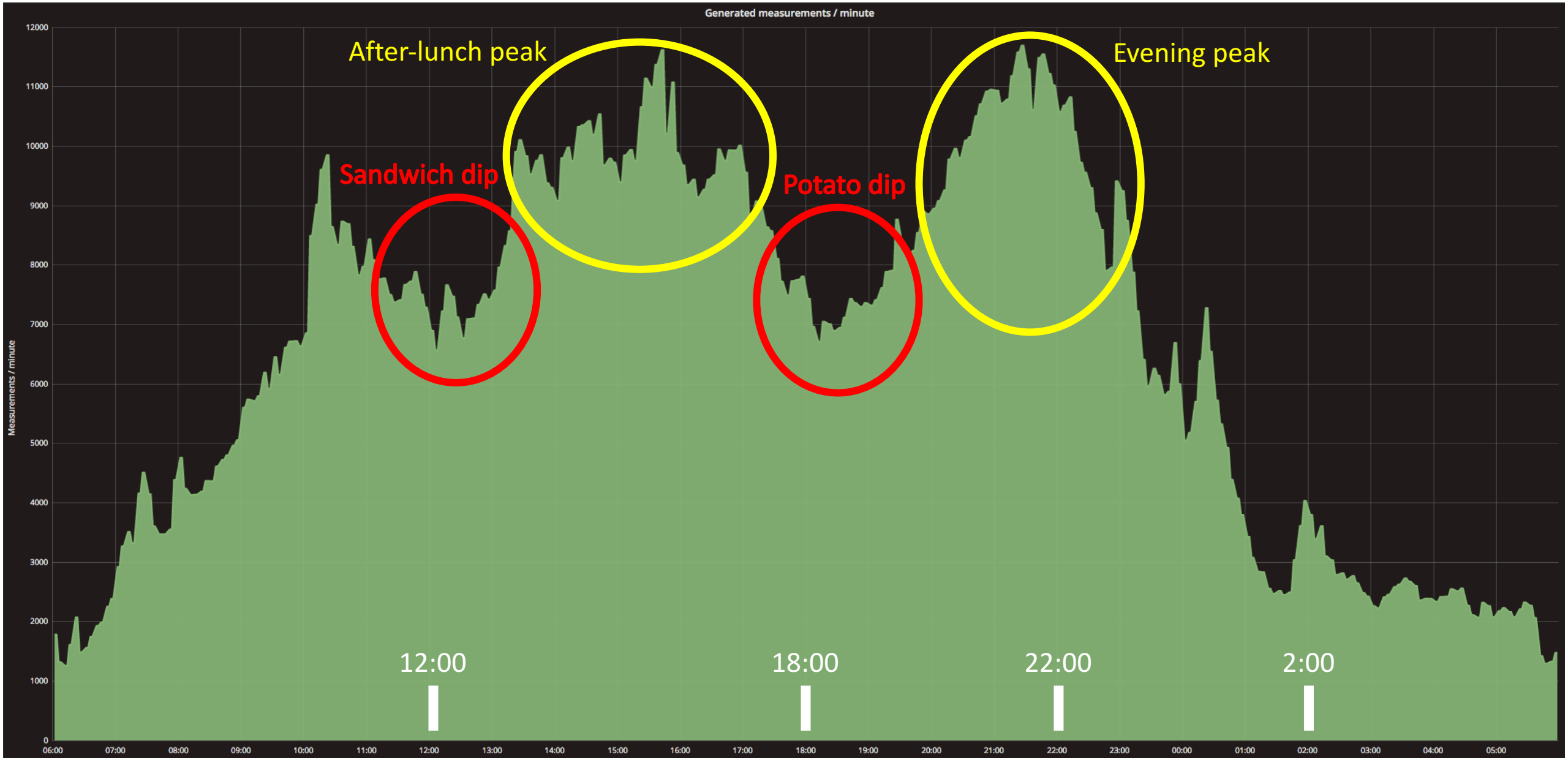
Production measurements



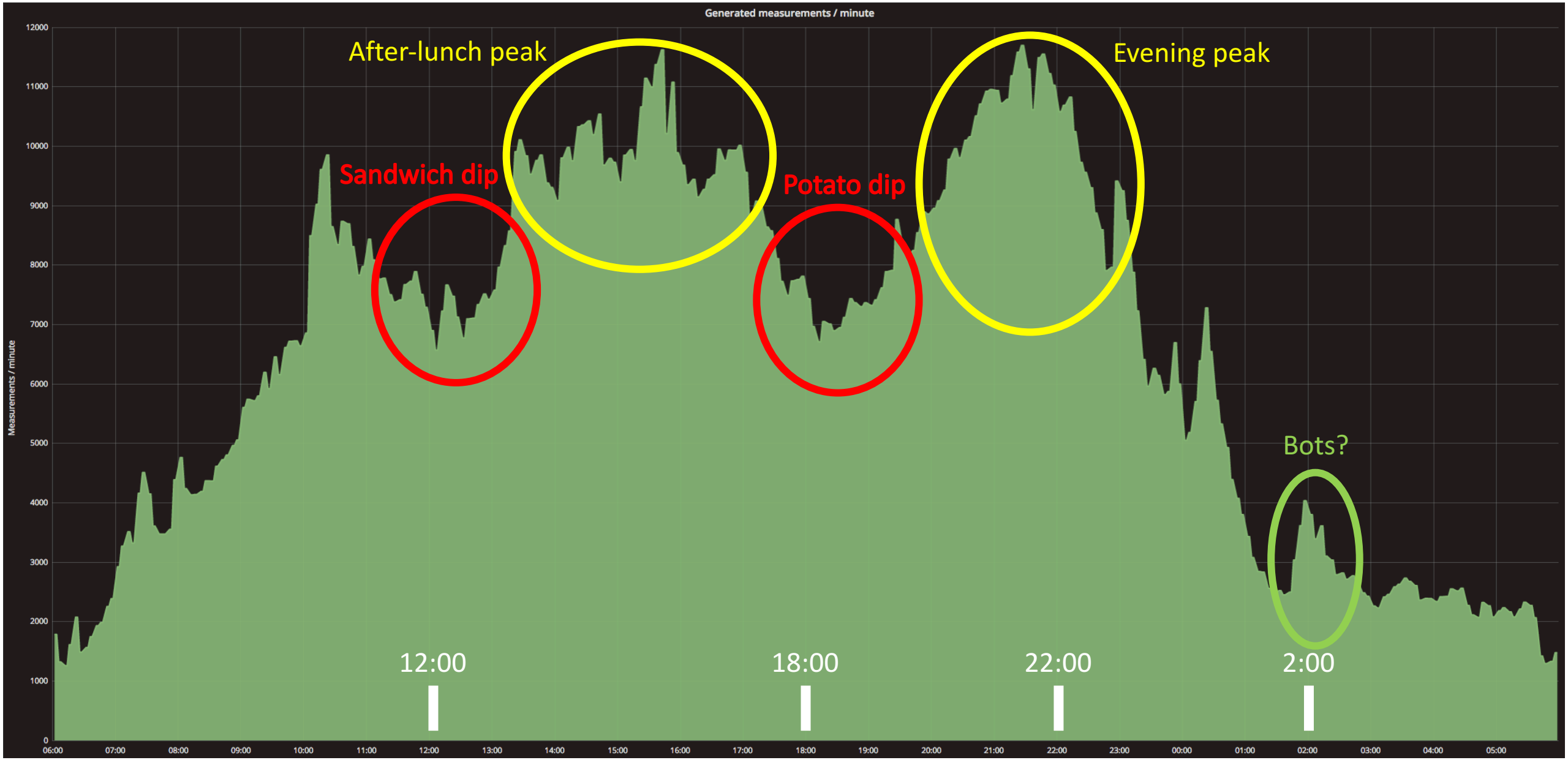
Production measurements



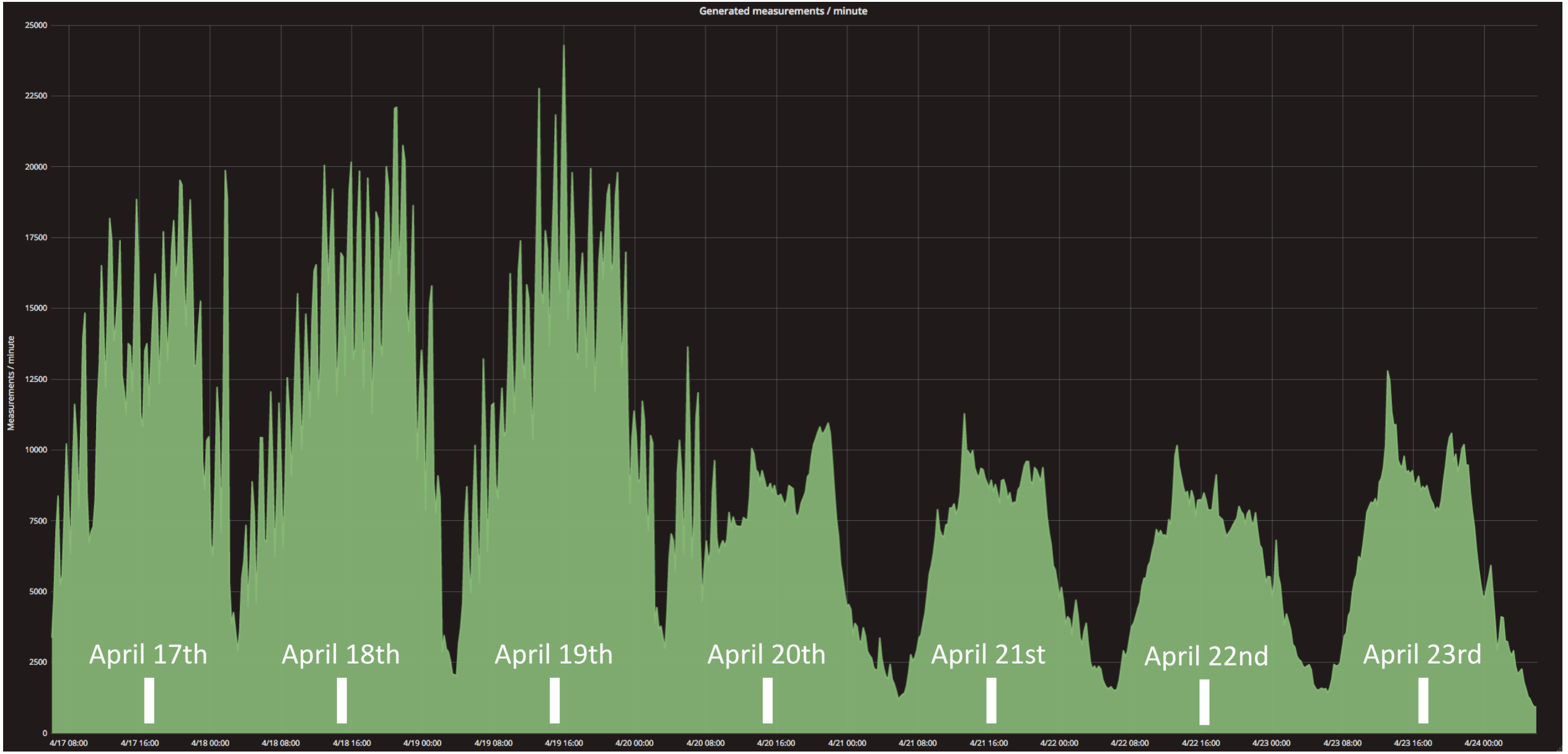
Production measurements



Production measurements

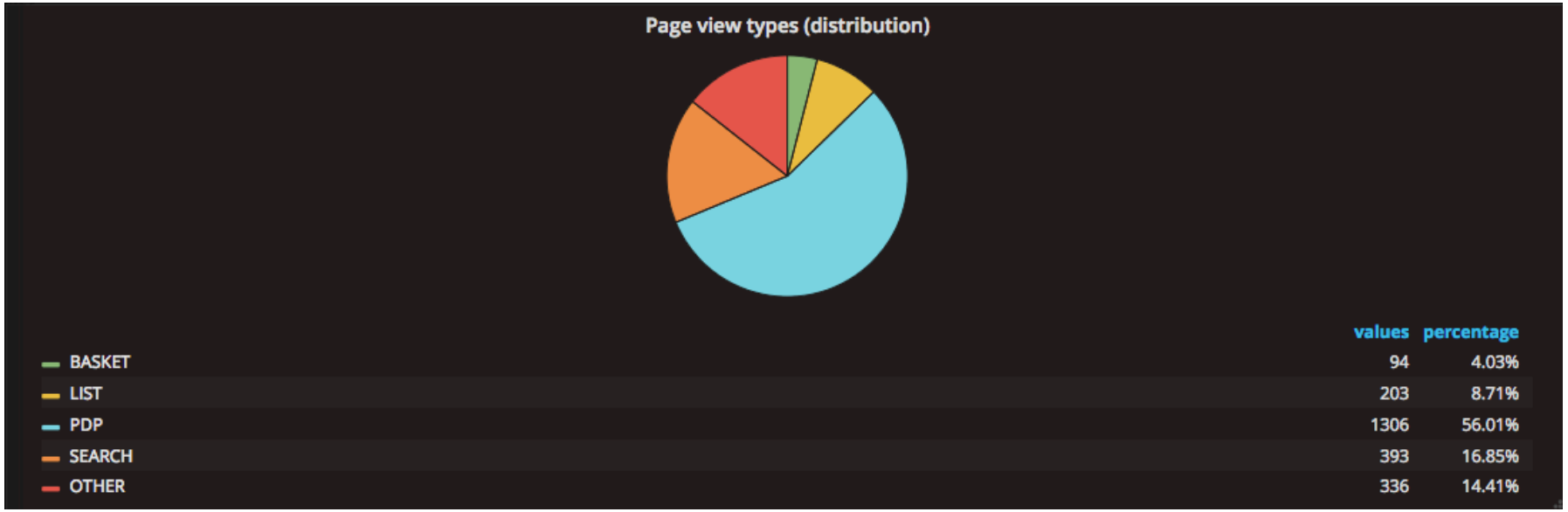


Filtering bot traffic



Some numbers

- Currently generated measurements/minute: ~15 500
- Expected peak load with website fully tagged: 6 000 000 measurements/minute
- Average message size: 500 bytes
- Most visited page: product page with 75% of views



bol.com

Kafka – cluster specs

- 3 nodes
  - 15x 126 GB disks
  - 3.6 TB RAM
  - 40 cores (2.3 GHz)

- First Kafka users at bol.com

Kafka – lessons learned

- Problem: disks spent a very long time on I/O operations

bol.com

Kafka – lessons learned

- Problem: disks spent a very long time on I/O operations

- Cause: property “flush.messages” set to 0

Kafka – lessons learned

- Problem: disks spent a very long time on I/O operations

- Cause: property “flush.messages” set to 0

flush.mes-sages	This setting allows specifying an interval at which we will force an fsync of data written to the log. For example if this was set to 1 we would fsync after every message; if it were 5 we would fsync after every five messages. In general we recommend you not set this and use replication for durability and allow the operating system's background flush capabilities as it is more efficient. This setting can be overridden on a per-topic basis (see the per-topic configuration section).

bol.com

Kafka – lessons learned

- Problem: disks spent a very long time on I/O operations

- Cause: property “flush.messages” set to 0

flush.mes-sages	This setting allows specifying an interval at which we will force an fsync of data written to the log. For example if this was set to 1 we would fsync after every message; if it were 5 we would fsync after every five messages. In general we recommend you not set this and use replication for durabili-ty and allow the operating sys-tem's background flush capa-bilities as it is more efficient. This setting can be overridden on a per-topic basis (see the per-topic configuration section).

bol.com

Kafka – lessons learned

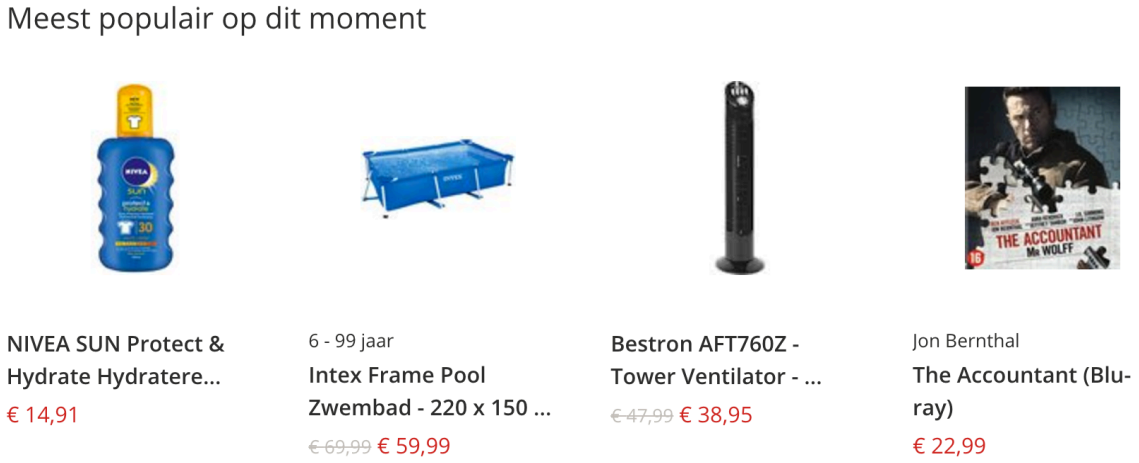
- **Problem**: disks spent a very long time on I/O operations
- **Cause**: property “flush.messages” set to 0
- **Lesson**: do not change a default property without a concrete justification

flush.mes-sages	This setting allows specifying an interval at which we will force an fsync of data written to the log. For example if this was set to 1 we would fsync after every message; if it were 5 we would fsync after every five messages. In general we recommend you not set this and use replication for durabili-ty and allow the operating sys-tem's background flush capa-bilities as it is more efficient. This setting can be overridden on a per-topic basis (see the per-topic configuration section).

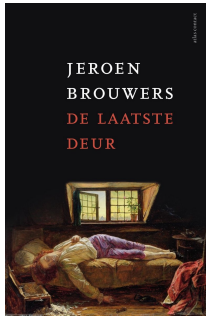
bol.com

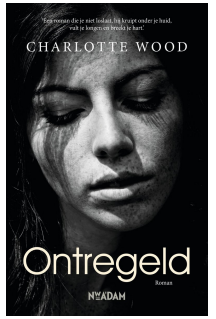
Applications – Real-time trending products

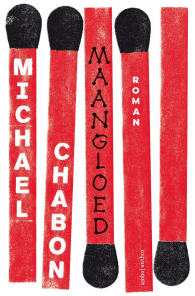
- Calculating products which are trending per 10 minute intervals compared to a reference period of 4 hours

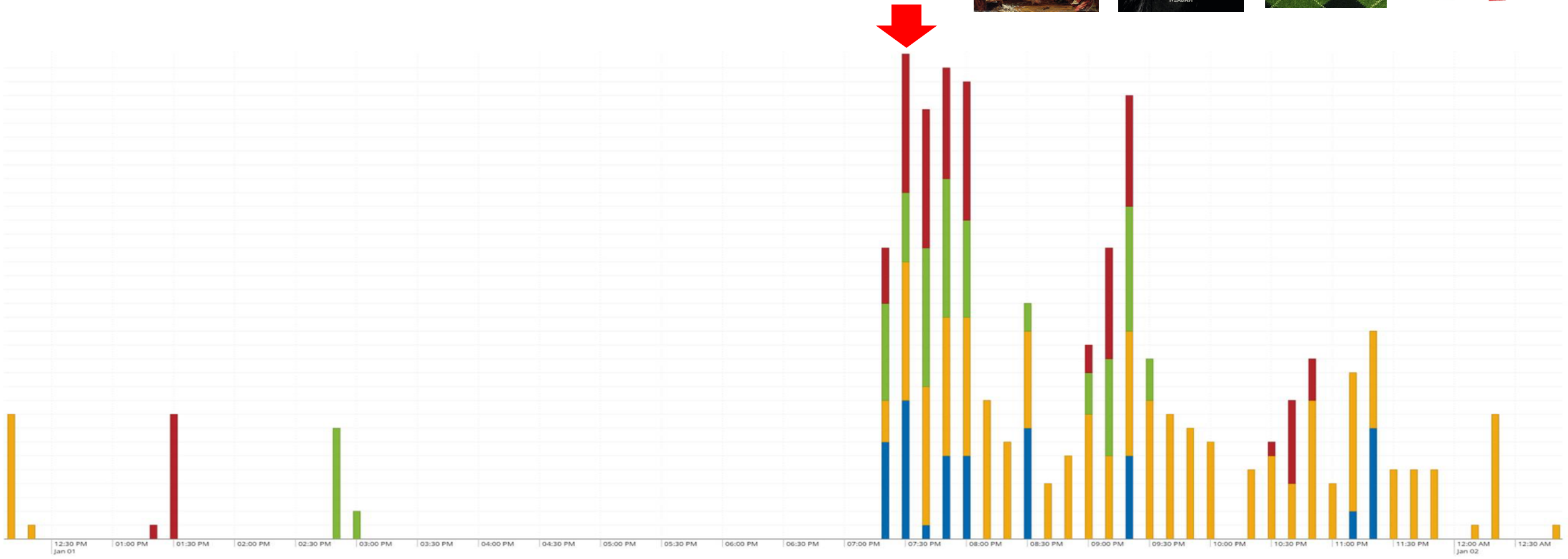


Real time trending





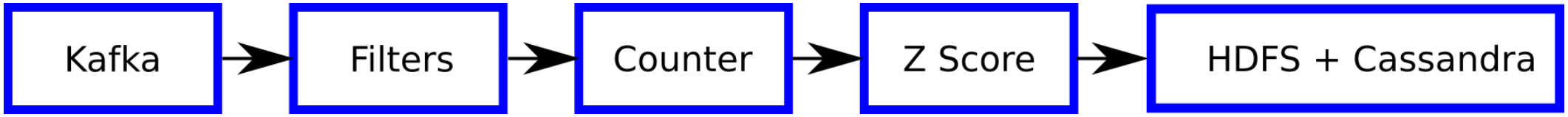




19:00 19:30

bol.com

How do we do it?



bol.com

Persisting data

- Define Schema in: Apache AVRO
  - Nice schema language (IDL)
  - Generates Java classes
  - Supports schema evolution

- Persist in files: Apache Parquet
  - Compresses well
  - Great on read use cases
  - Write straight from Avro classes.
    - See o.a.parquet.avro.AvroParquetWriter

Persisting data

- Apache AVRO
  - Standard byte[] for the record.
  - Needs the original Schema on read
- What if the Schema changes?
  - Kafka keeps the events for weeks
- AVRO-1704
  - Serialize record to a “Message” byte[]
    - schema fingerprint (hash)
    - pluggable schema ‘database’

Future plans

Future plans

- Tag the entire webshop
- Tag the mobile app/ the seller dashboard
- Fully enable in-view measurements (currently at 1%)
- Provide tooling for working with the data
- Generate live reporting in real-time
- ...

bol.com

Thanks for listening

Questions?

Ivan Budinčević

ibudincevic@bol.com