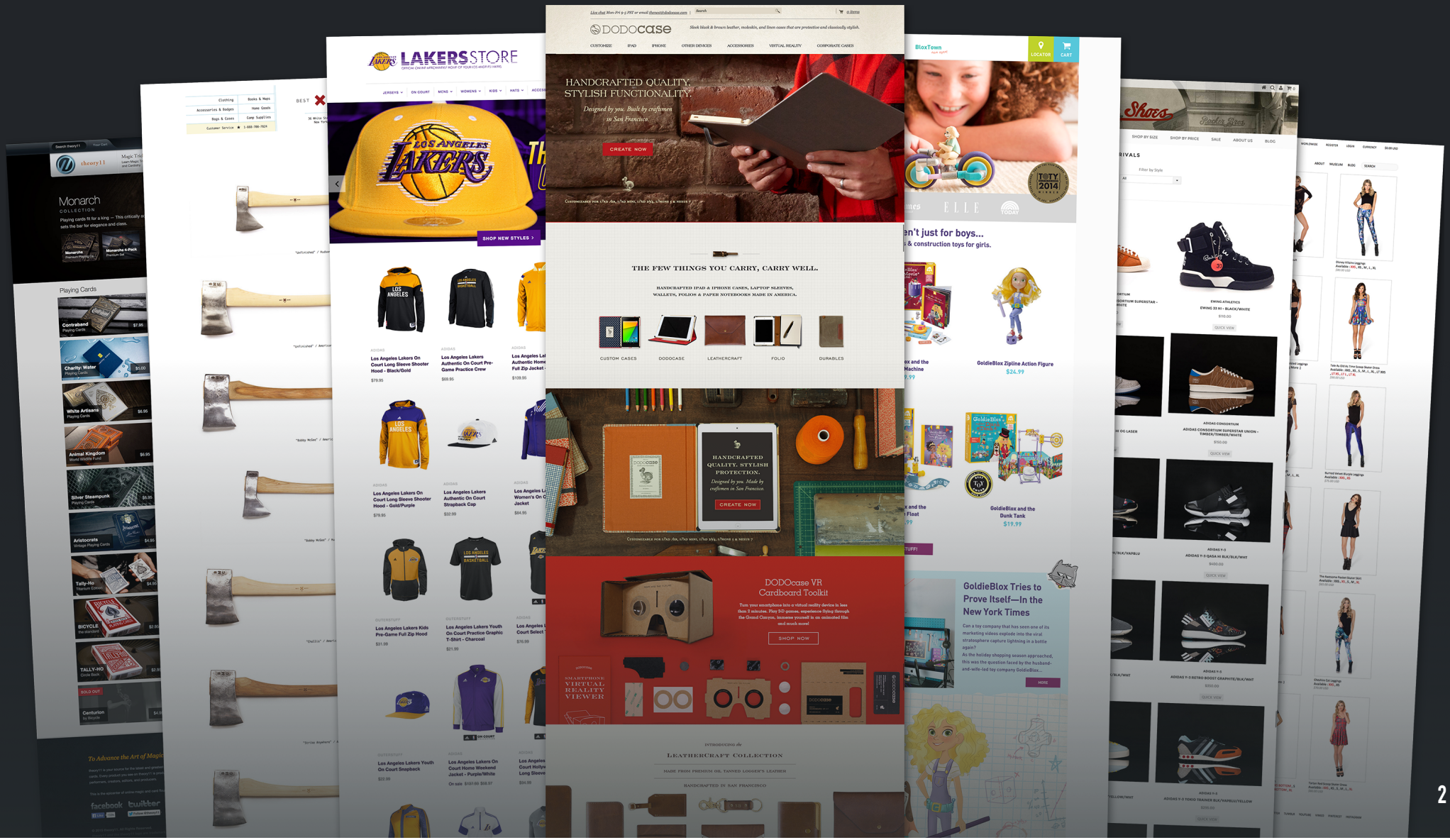
The Little Warehouse That Couldn't Or: How We Learned to Stop Worrying and Move to Spark

Yandu Oppacher (@yandu)
Data Infrastructure

shopify

Shopify Stores



August 2013

ETL

Ruby

Warehouse

Vertica

Reporting

Tillller

Why we had to move

- Data volume
- Data/Query complexity
- Performance issues

Couple of false starts

Pig + Oozie

Pig + Luigi

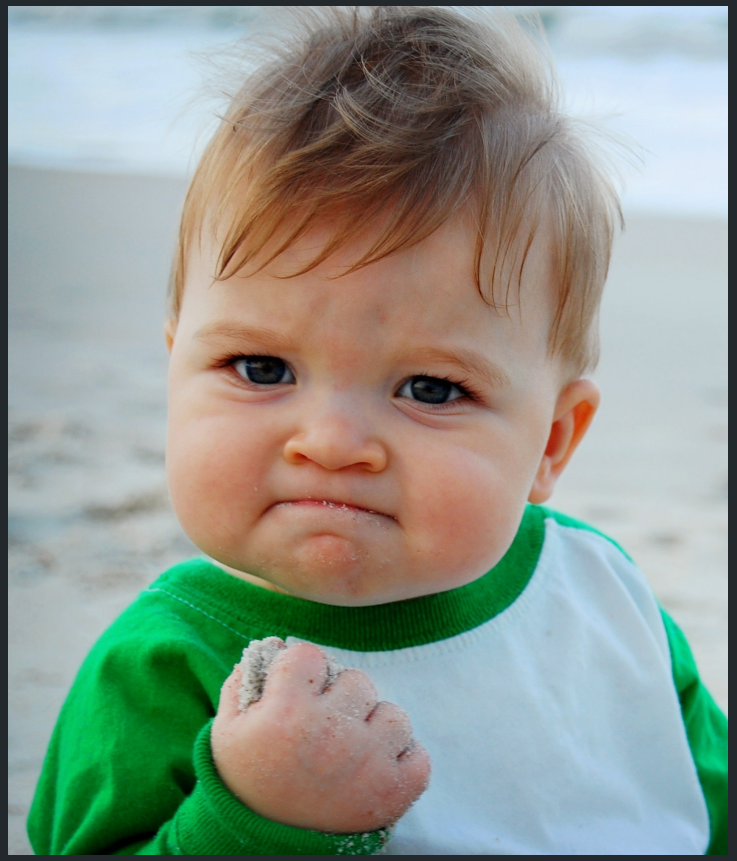
Platfora

“ a

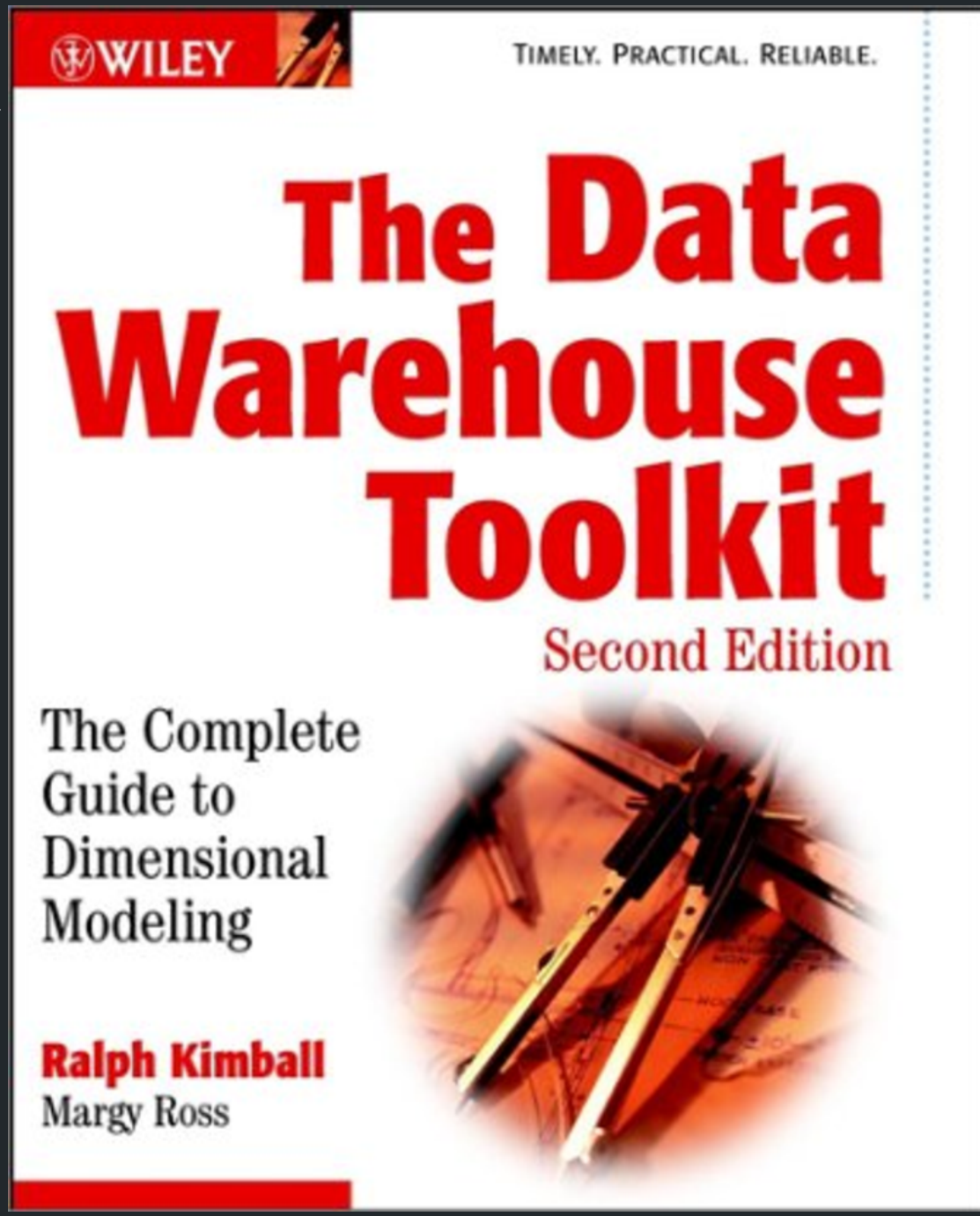
ware ng a

Enter Spark

- Fast
- Nice development model
- Python



The Good Book



GMV

A Case Study

165,000+

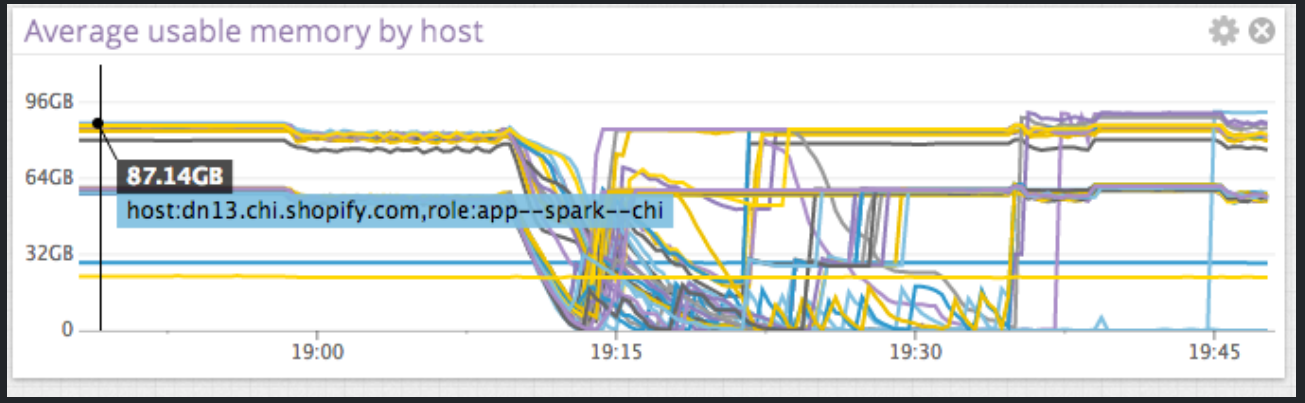
ACTIVE SHOPIFY MERCHANTS

$8 BILLION+

CUMULATIVE GMV

Growing pains

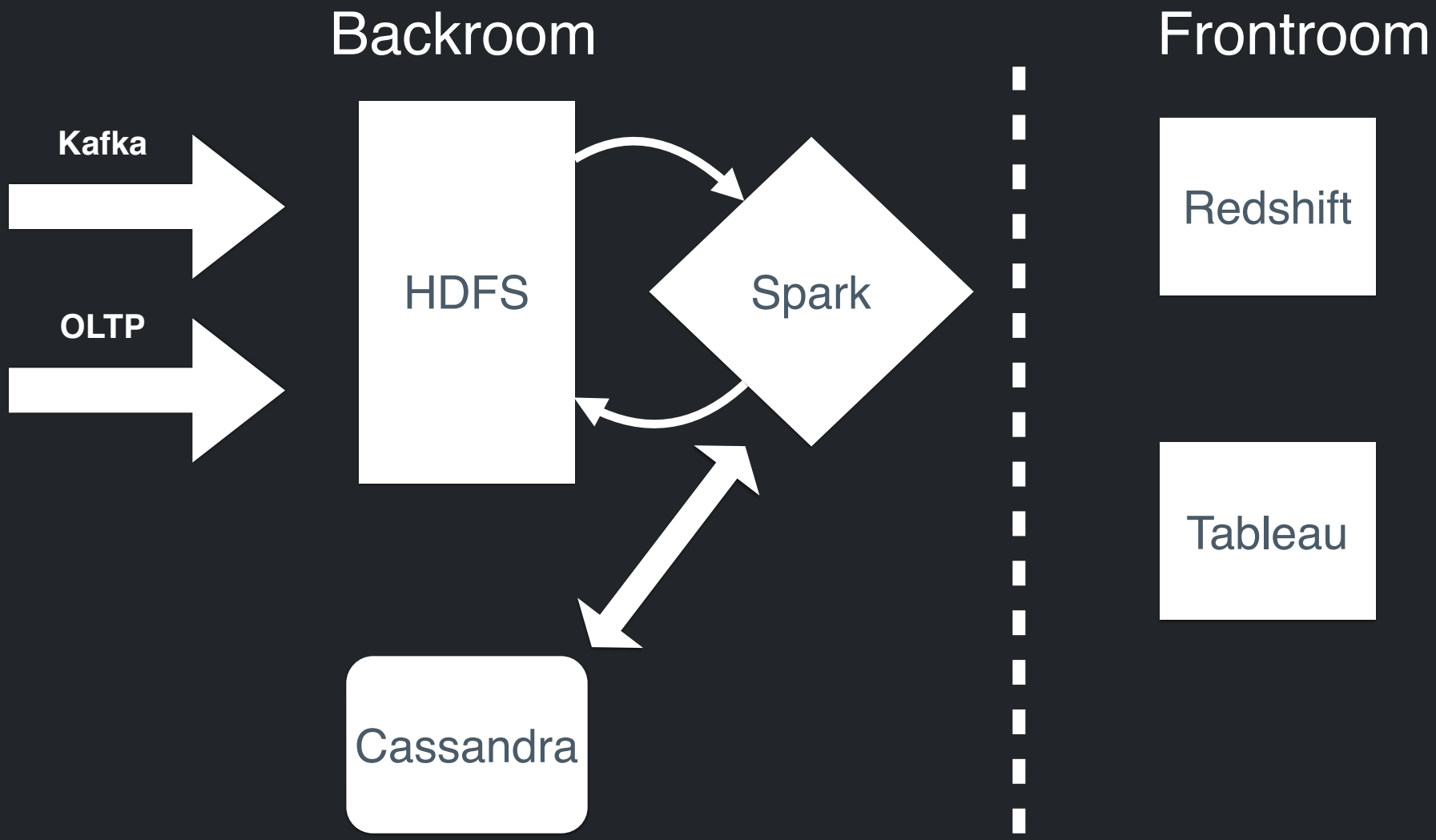
- Joins
- Groupings
- General data skew
- Getting to know python's performance quirks



Starscream

- specialized joins
- resolvers
  - range
  - cassandra
- overby
- contracts
- incrementalized fact builds

Our current stack



Thank you

Yandu Oppacher (@yandu)
Data Infrastructure

shopify