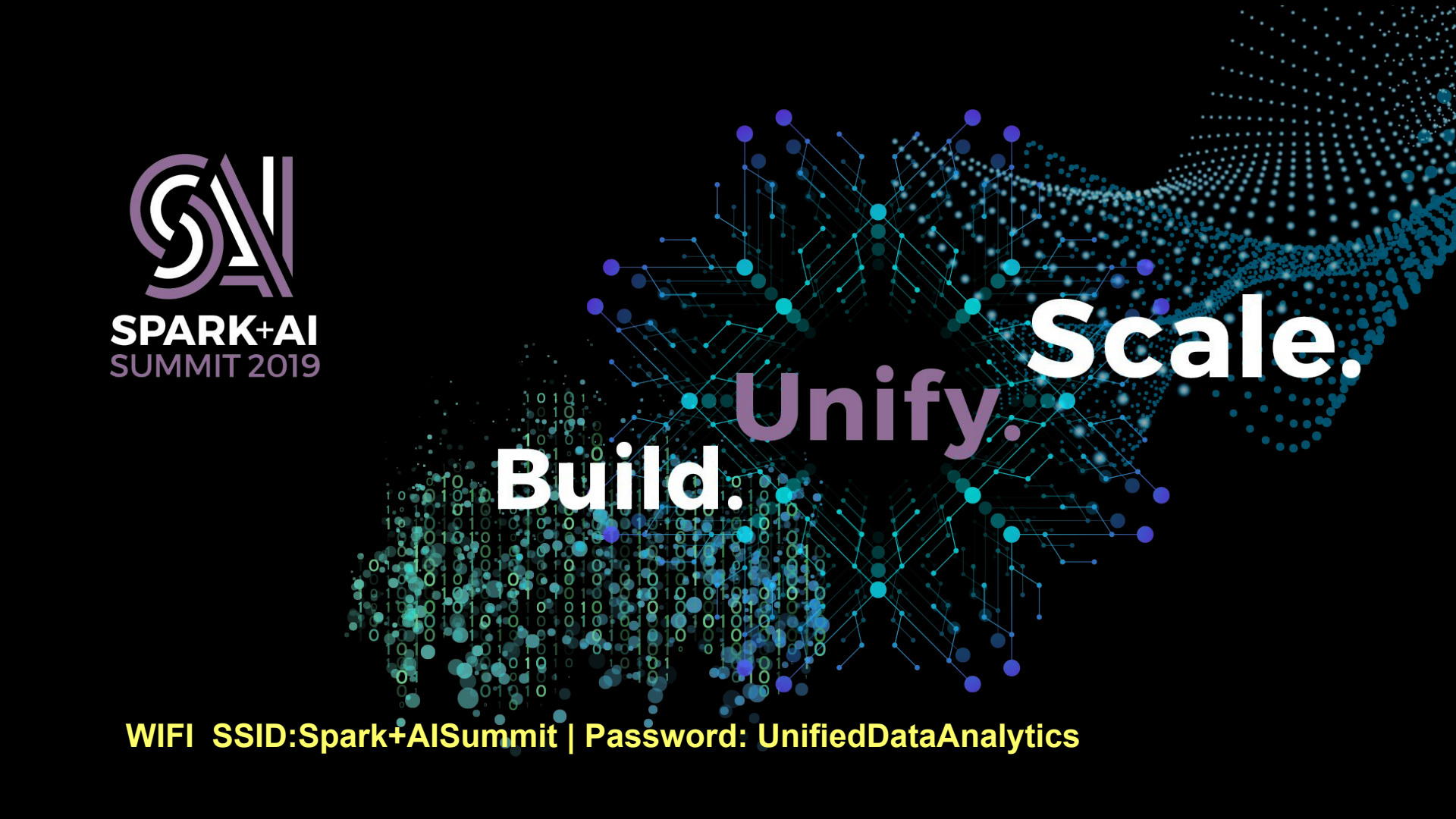


WIFI SSID:Spark+AISummit | Password: UnifiedDataAnalytics

Data Warehousing with Spark Streaming @ Zalando

Sebastian Herold, Zalando SE

Sebastian Herold

 # Principal Data Engineer / Architect

 # 7y @ Immo-/Scout24

 # DataDevOps Manifesto

 # Data Platform

 # 2y @ Zalando

 # ML Productivity

 # Streaming DWH

@heroldamus

WE BRING FASHION TO PEOPLE

17 markets

9 fulfillment centers

>28M active customers

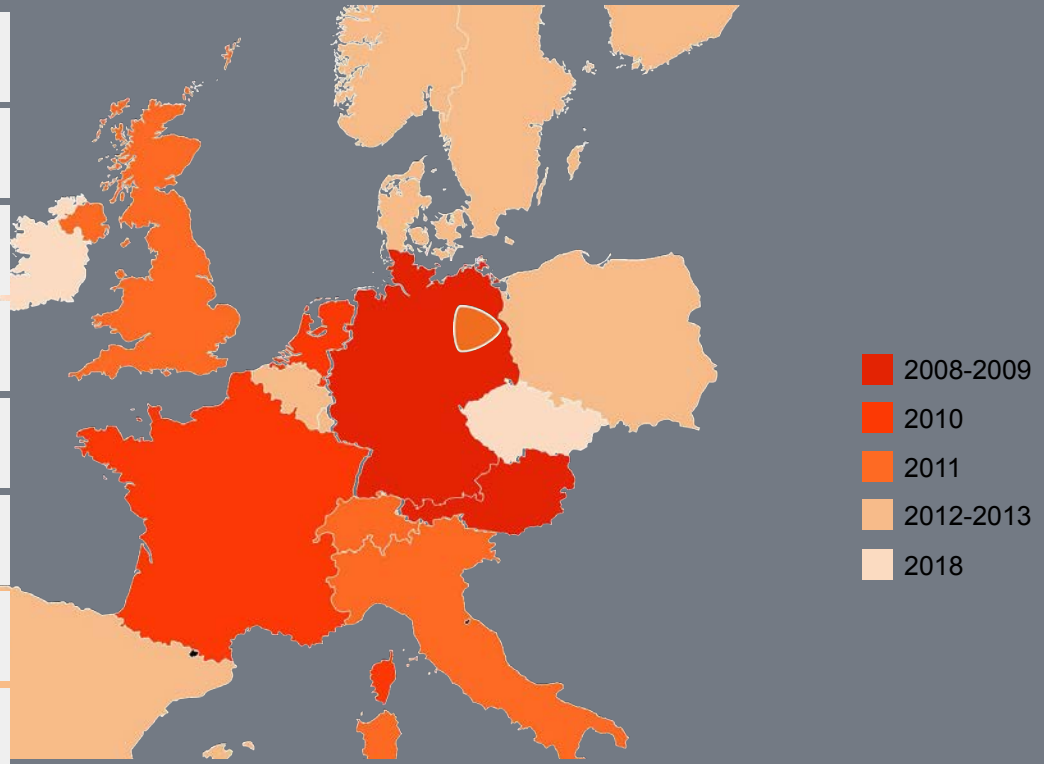
5.4B revenue 2018

>300M visits/month

>14k employees

>80% visits from mobile

>400k product choices



zalando

TECH@SCALE

aws

>350 accounts

>5 data lakes

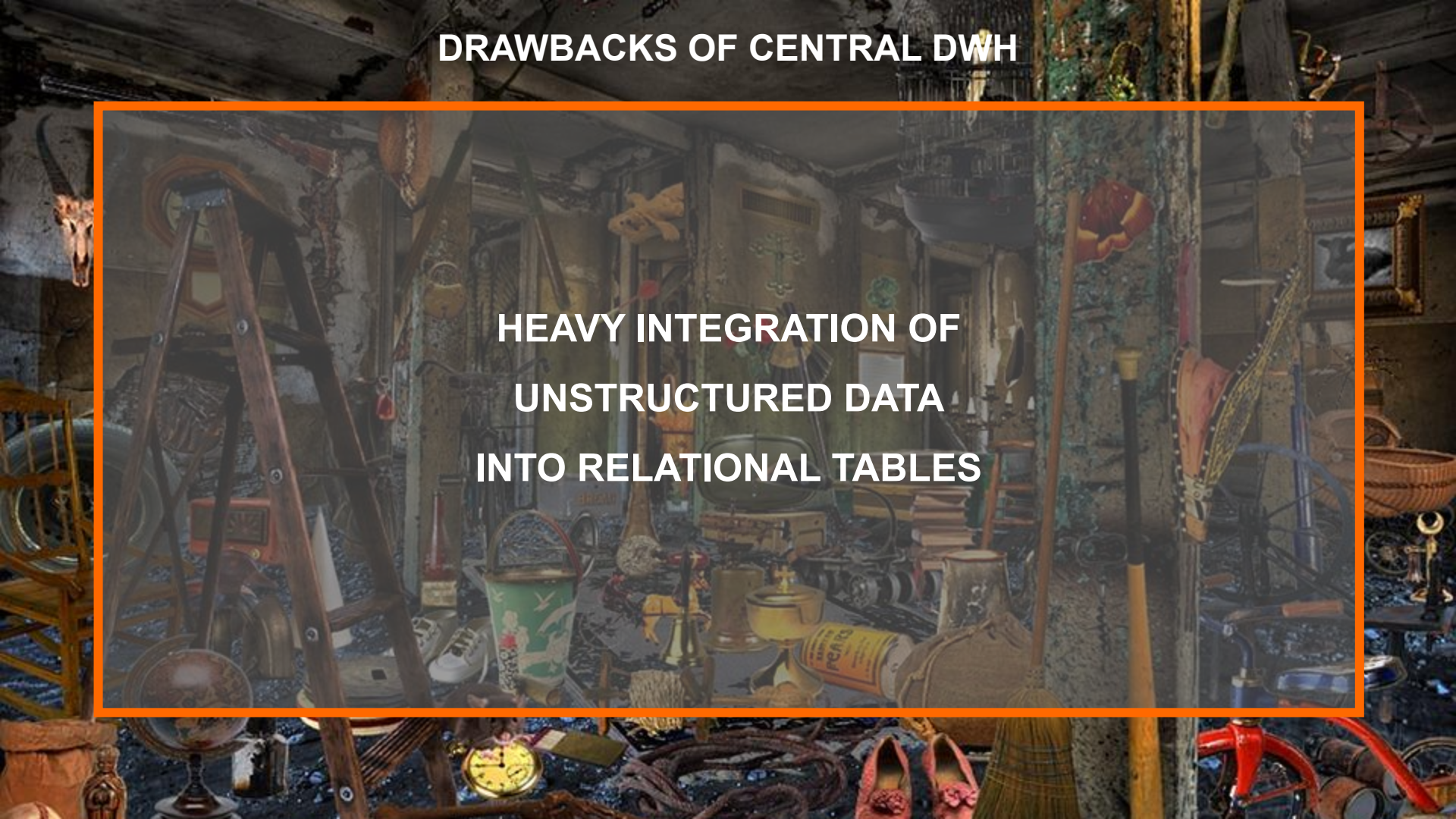
>250 teams

>800 micro services

>100 clusters

WHY OUR CENTRAL DWH DOES NOT SUCCEED ANYMORE?

DRAWBACKS OF CENTRAL DWH



HEAVY INTEGRATION OF

UNSTRUCTURED DATA

INTO RELATIONAL TABLES

DRAWBACKS OF CENTRAL DWH

DATASETS ARE NEEDED

DISTRIBUTED

DRAWBACKS OF CENTRAL DWH

LOWER LATENCY REQUIRED BY

AI USE-CASES,

OTHER DATA WAREHOUSES,

NEAR-REALTIME USE-CASES

DRAWBACKS OF CENTRAL DWH

MULTIPLE TEAMS DO SAME

LOW-LATENCY EVENT INTEGRATION

HEAVY INTEGRATION OF

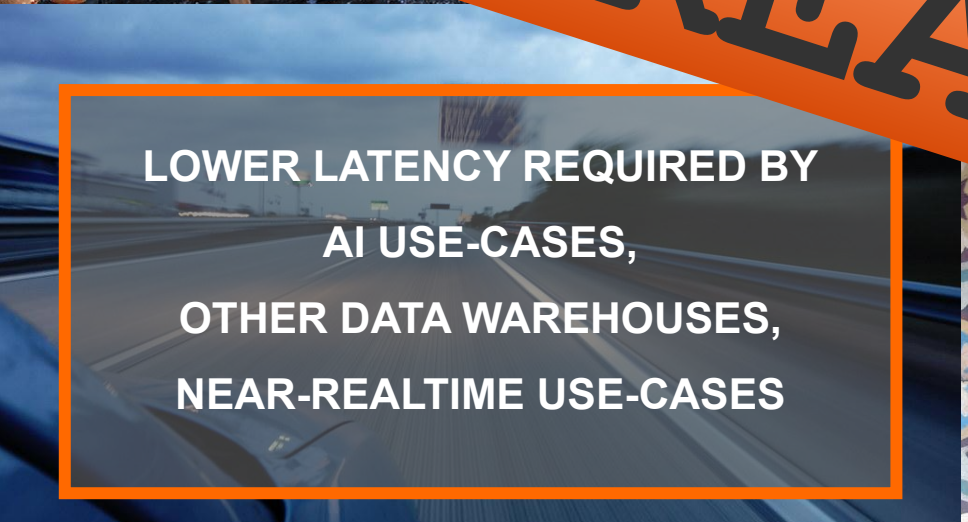
UN CTURED DATA

INT LES

DATASETS ARE NEEDED

DISTRIBUTED

STREAMING



LOWER LATENCY REQUIRED BY

AI USE-CASES,

OTHER DATA WAREHOUSES,

NEAR-REALTIME USE-CASES

AME

LOW-LA NT

INTEGRATION

SALES ORDER EXAMPLE

order.created

```
order_id,
order_date,
items,
...
```

payment.done

```
payment_id,
payment_date,
order_id,
...
```

shipment.created

```
order_id,
shipping_date,
shipped_items,
...
```

item.returned

```
order_id,
return_date,
returned_item,
...
```

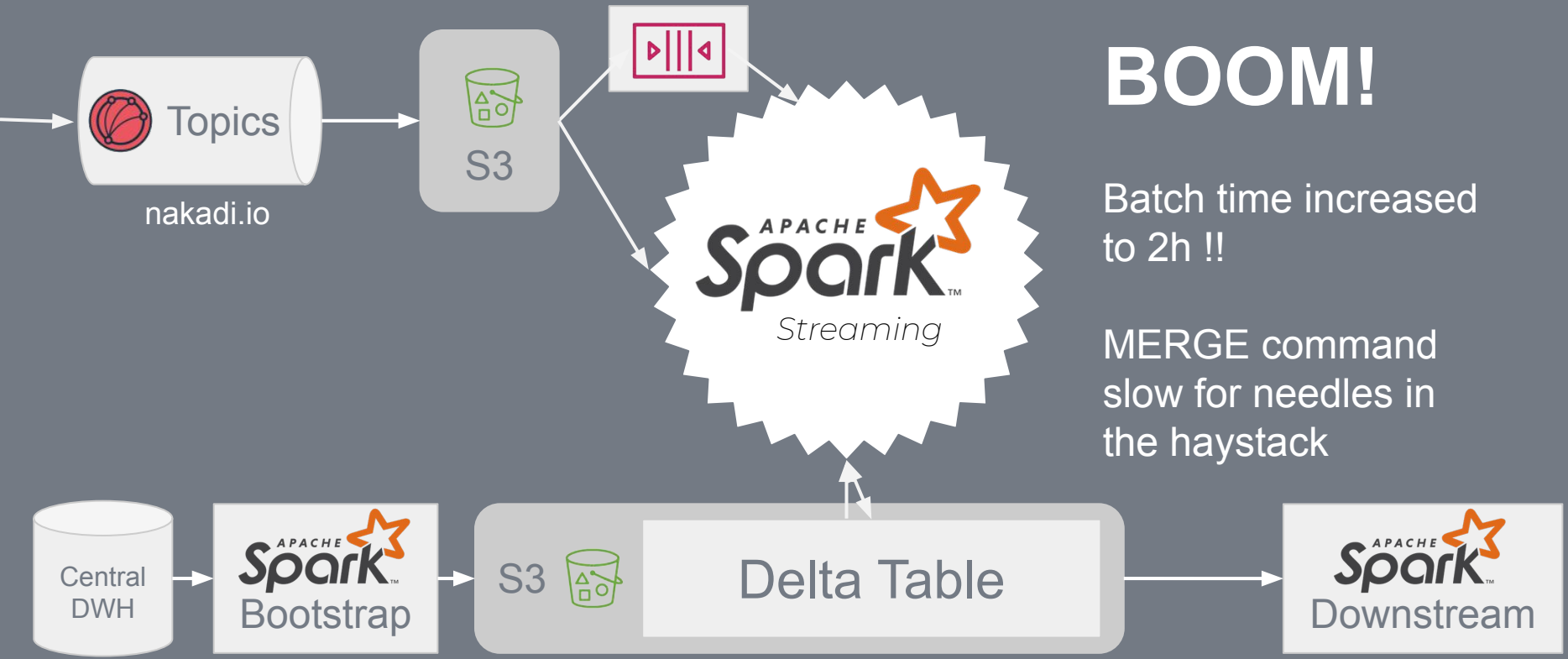
sales-order

```
order_id,
order_date,
payment_id,
payment_date,
items:
  shipped_at,
  returned_at,
  ...
calculated_1,
calculated_2
...
```

HOW WE STARTED?

zalando

INTEGRATION OF HISTORIC DATA



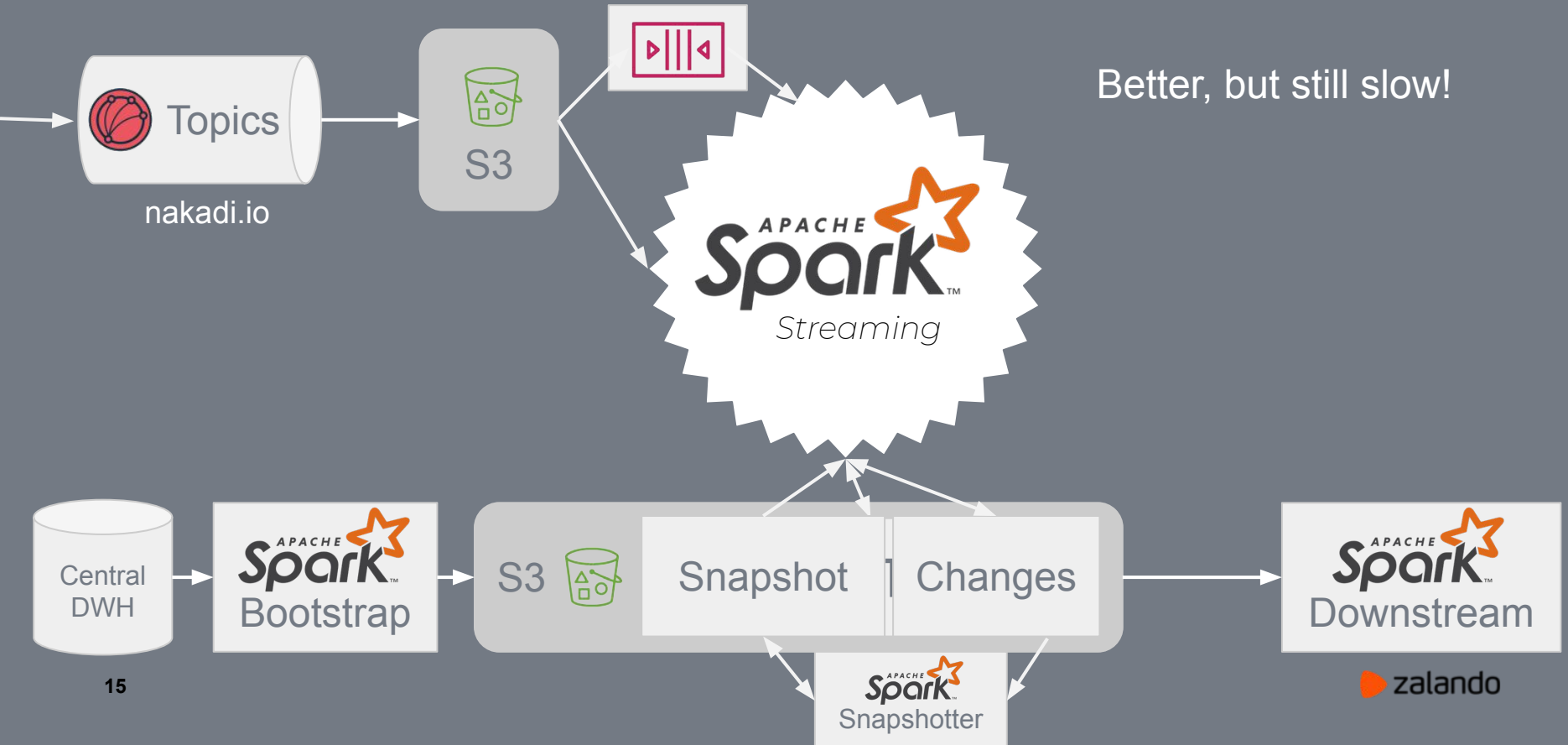
BOOM!

Batch time increased to 2h !!

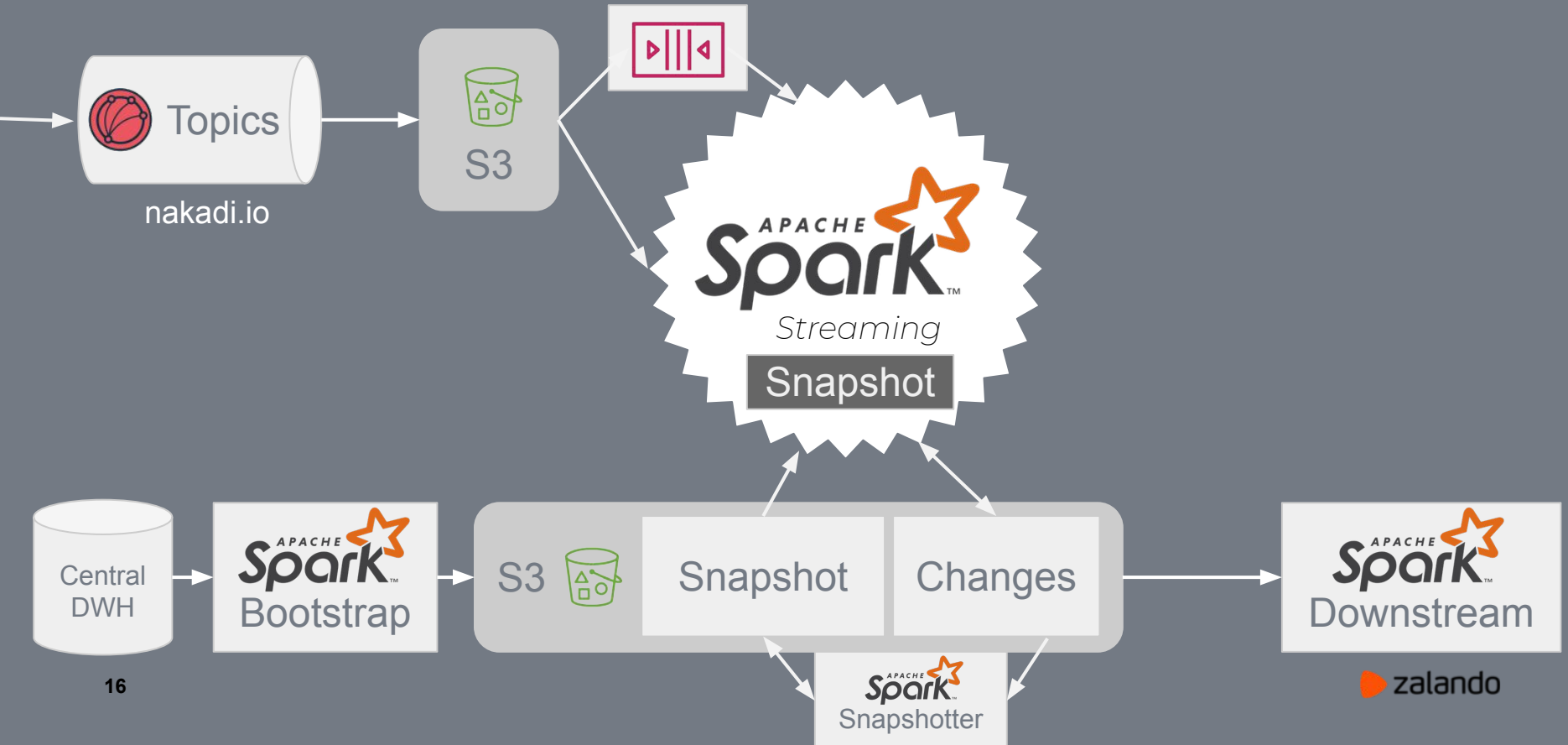
MERGE command slow for needles in the haystack

zalando

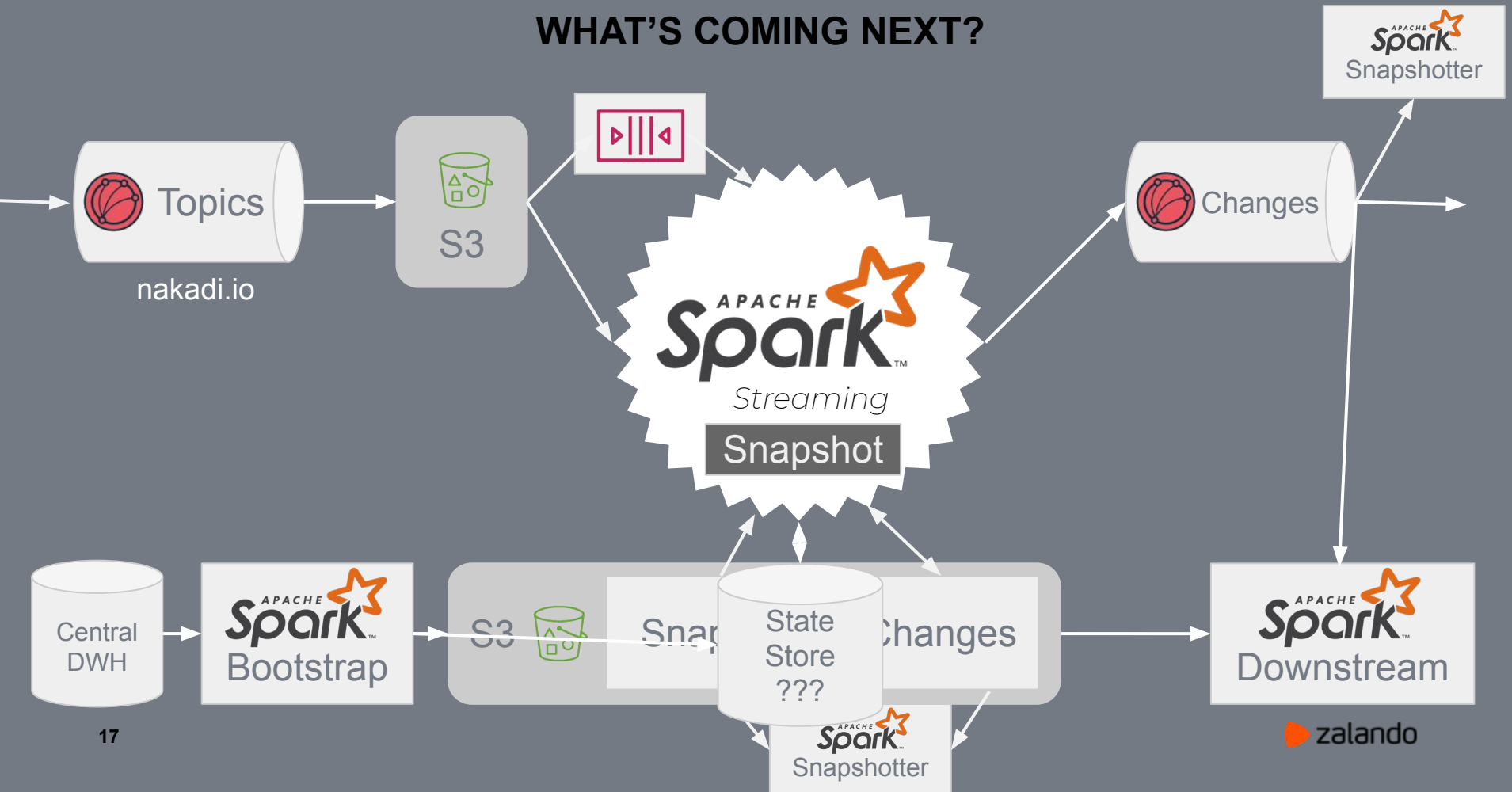
INTRODUCE SNAPSHOTS AND CHANGES TABLE



LOAD SNAPSHOT INTO CLUSTER



WHAT'S COMING NEXT?



SQL vs SCALA

SCALA

Started with 200 [illegible]es of SQL

Grew [illegible]

[illegible]nciple

[illegible] unit-test

Hard to refactor

Bad support for nested structures

LESSONS LEARNED

Streaming needs different thinking

DWH ~ Backend Programming

Don’t start with SQL because it’s easy

Databricks Delta succeeds Parquet

Make sure all data is available in S3

SPARK+AI SUMMIT 2019

THANKS A LOT!

QUESTIONS?

WE ARE HIRING!