zalando

A Journey from Scikit-learn to Spark

Stanimir Dragiev
Patrick Baier
Zalando SE

SPARK SUMMIT
EUROPE 2016

Speakers

- Stanimir Dragiev
  - PhD in Robotics + Machine Learning
  - Machine Learning engineer at Zalando

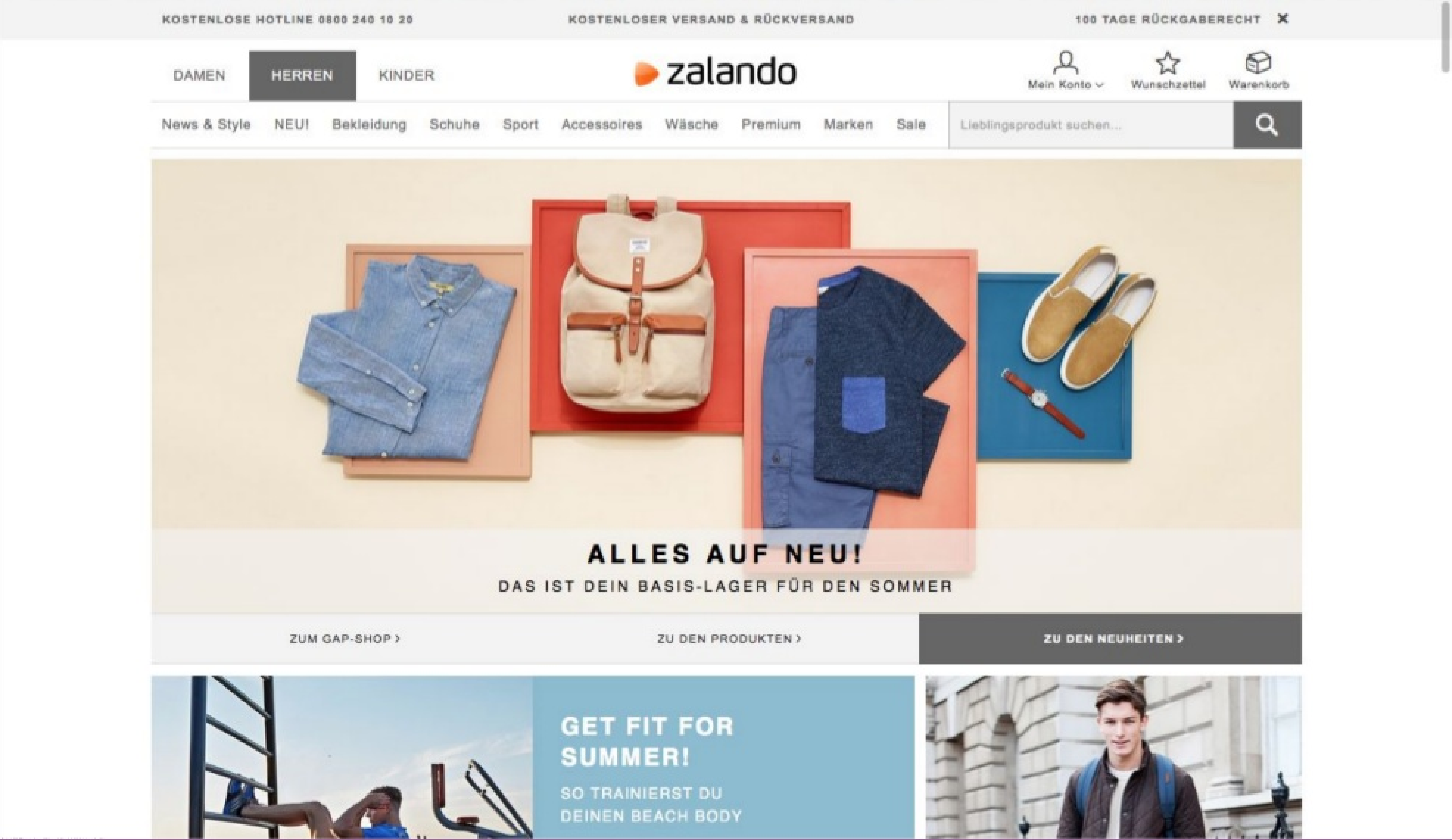
- Patrick Baier
  - PhD in Computer Science
  - Machine Learning engineer at Zalando

Europe's Leading Fashion Platform

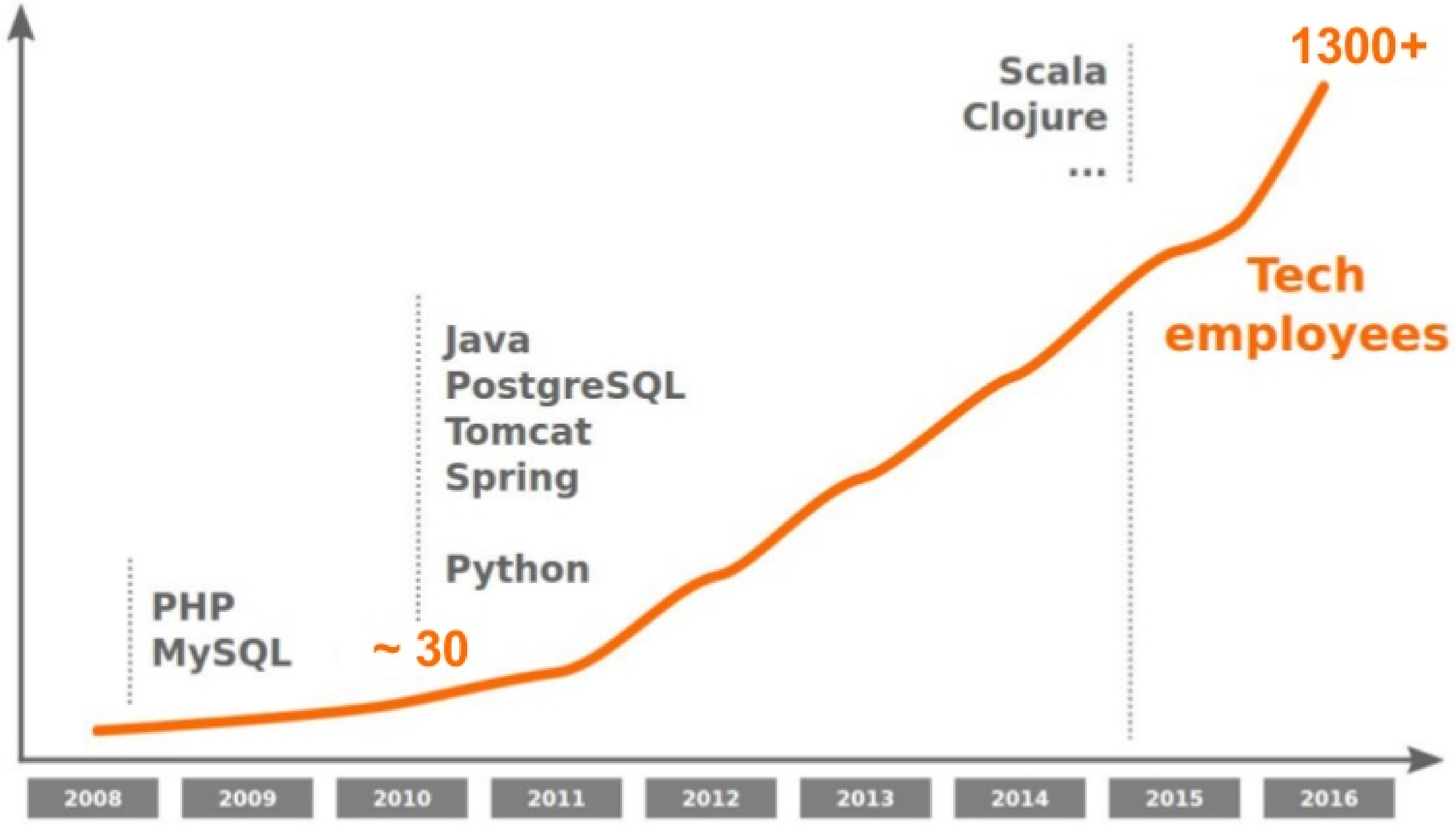
15 countries
3 fulfillment centers
19+ million active customers
3.0+ billion € revenue
160+ million visits per month
1.300+ employees in tech

Visit us: tech.zalando.com
radicalagility.org

SPARK SUMMIT EUROPE 2016


KOSTENLOSE HOTLINE 0800 240 10 20
KOSTENLOSER VERSAND & RÜCKVERSAND
100 TAGE RÜCKGABERECHT
DAMEN
HERREN
KINDER
zalando
Mein Konto
Wunschzettel
Warenkorb
News & Style
NEU!
Bekleidung
Schuhe
Sport
Accessoires
Wäsche
Premium
Marken
Sale
Lieblingsprodukt suchen...
ALLES AUF NEU!
DAS IST DEIN BASIS-LAGER FÜR DEN SOMMER
ZUM GAP-SHOP ›
ZU DEN PRODUKTEN ›
ZU DEN NEUHEITEN ›
GET FIT FOR SUMMER!
SO TRAINIERST DU DEINEN BEACH BODY

Zalando's Technology History



What we do

Team “Payment Analytics”

Goal: Estimate fraud risk for incoming orders

This requires us to master:

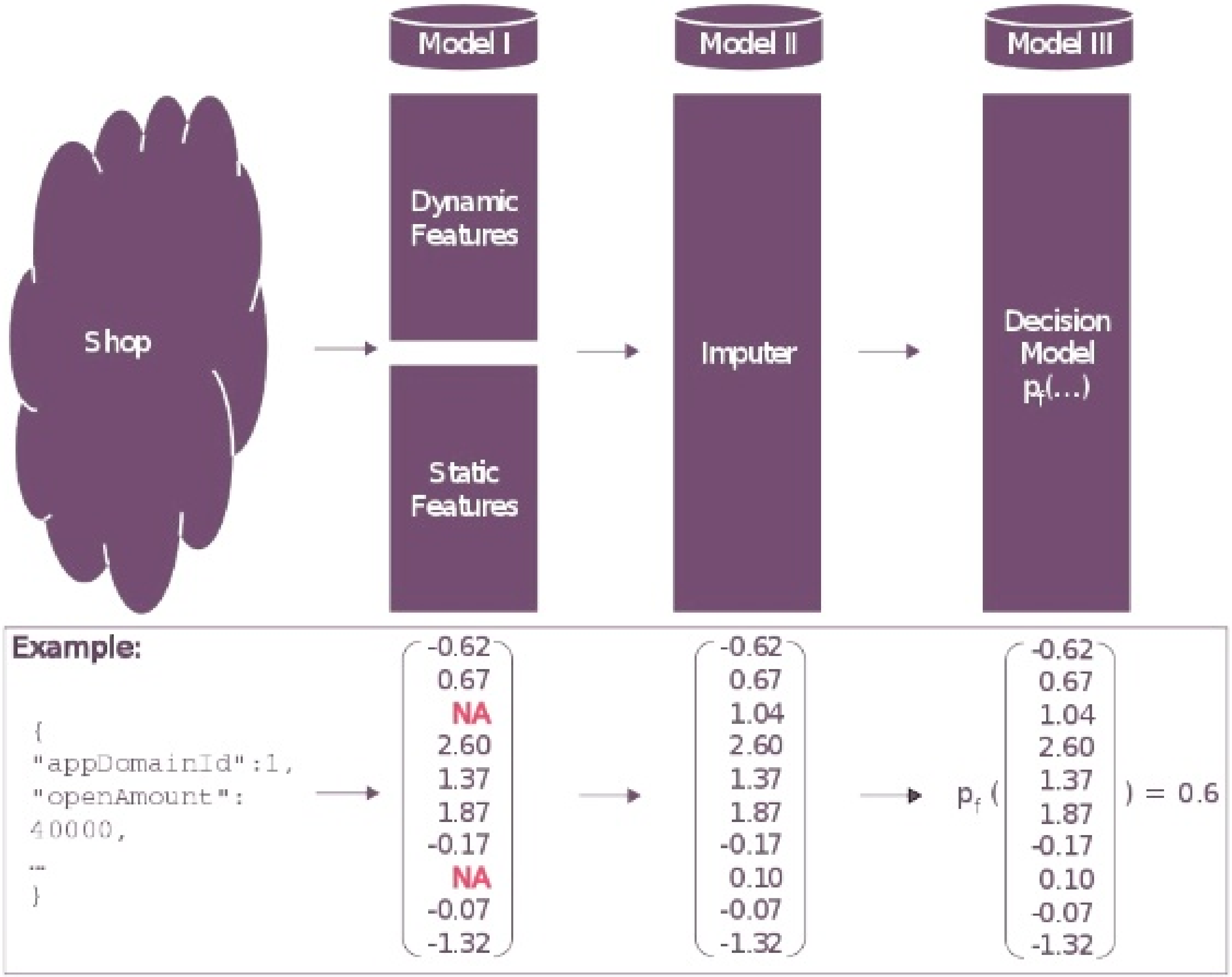
- Machine Learning
- Production code
- A diverse tech stack (Spark, Scala, R, SQL, AWS, Jenkins, Docker, Python, sh, ...)

Modeling Fraud Probabilities

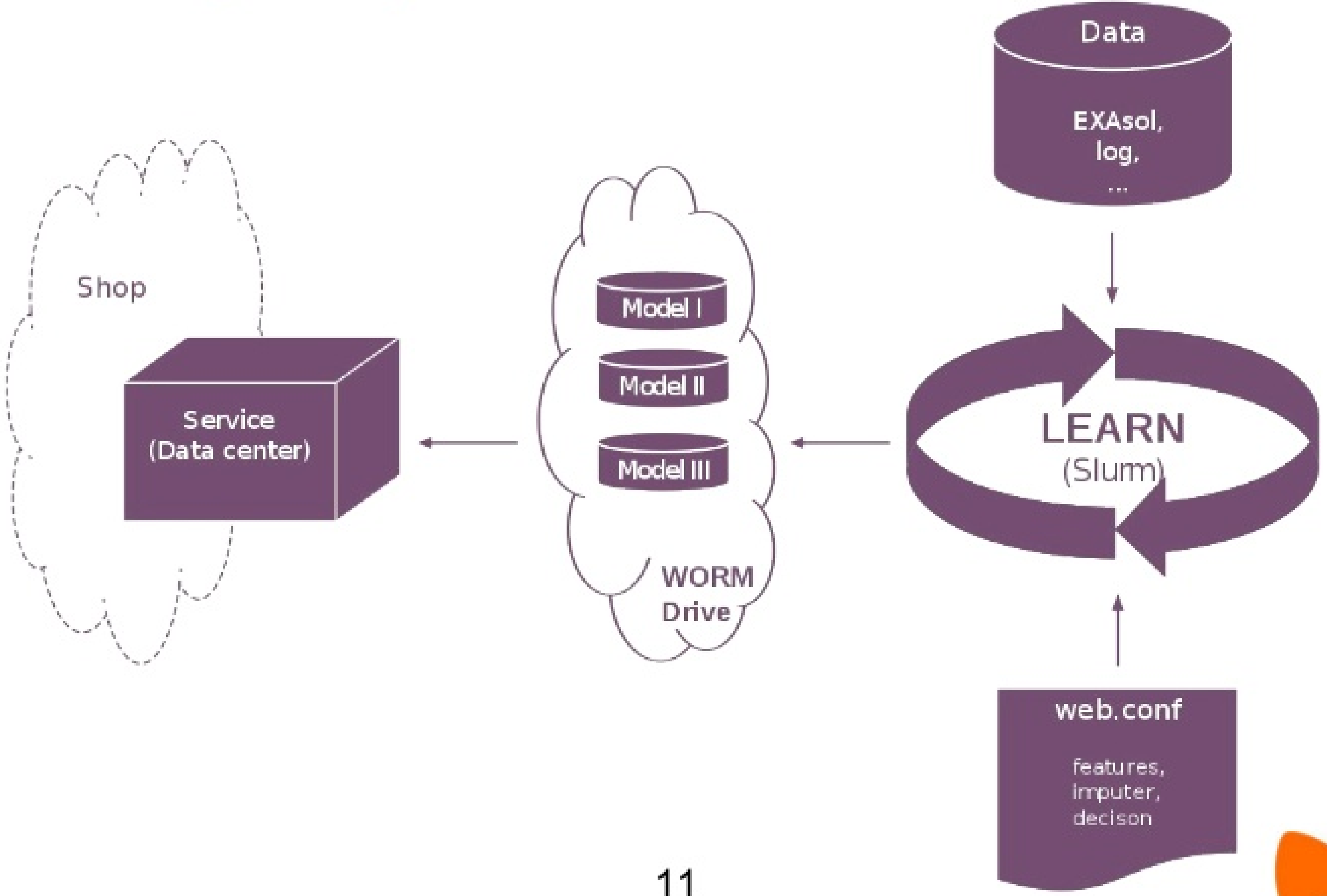
- Fraud probability p_{fraud} beyond business rules ...
- … but can be modeled via machine learning
  – data-driven
  – unbiased
  – reproducible
- Different application domains require large variety of models with frequent updates

Current setup

ML Architecture



Deployment of Models



Scalability

- Zalando evolves from online shop to fashion platform[1]
  - Connect all stakeholders in the fashion world: online and offline retailers, advertising agencies, logistic services, ...

- Number of orders continue to increase[2]
  - 2014: 41.4m orders
  - 2015: 55.3m orders

→ Scalability of our services is now major concern

[1] https://blog.zalando.com/en/blog/how-zalando-becoming-online-fashion-platform-europe
[2] https://corporate.zalando.com/en/zalando-continues-high-growth-path

Pain Points of Current Setup

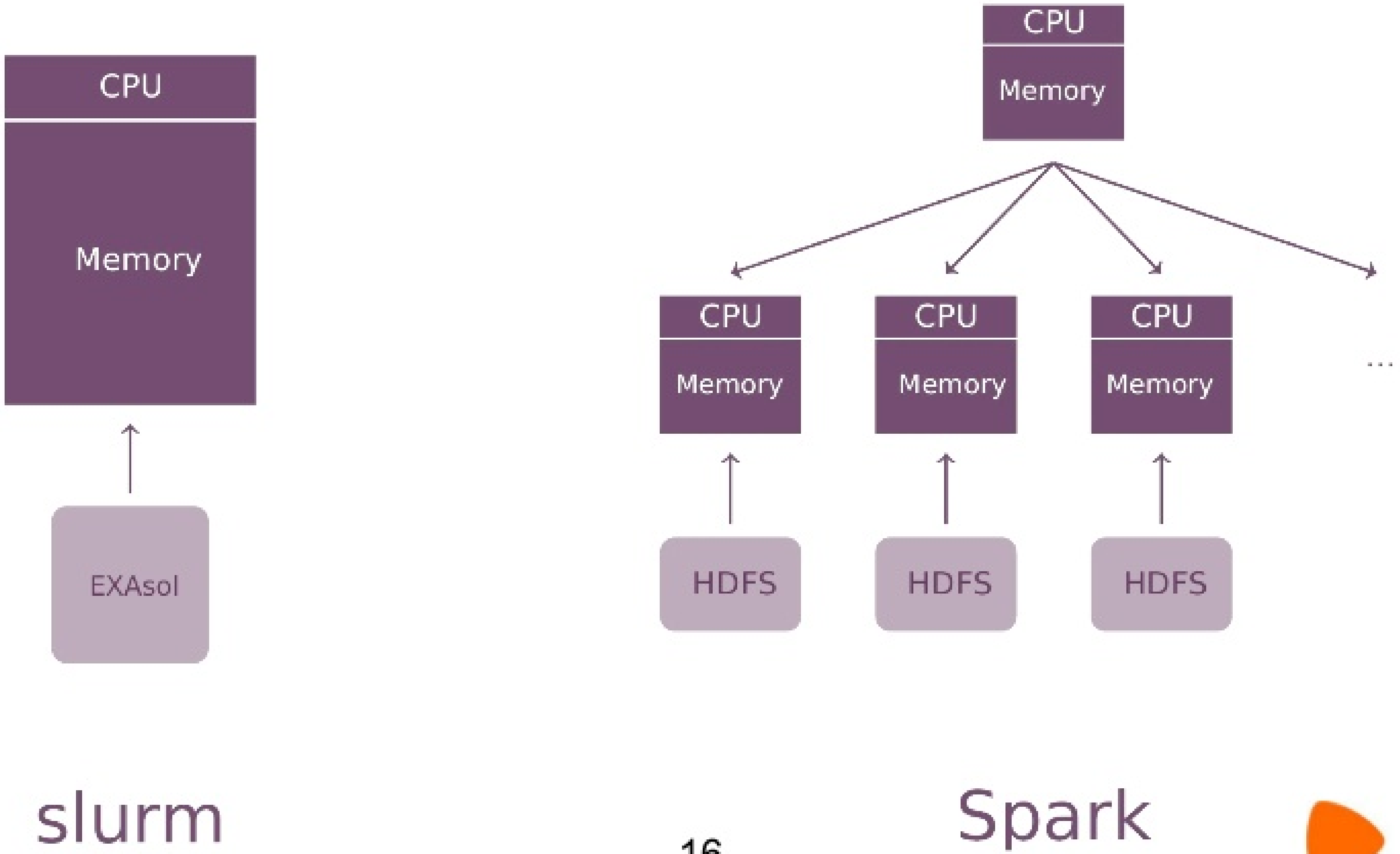
- No horizontal scaling of learning (memory bottleneck)
- Python: fast initial development, <u>but:</u> bad maintenance
- Coupled learning and data fetching (bottleneck: EXAsol)
- Learning on in-house cluster ⇝ bottleneck

Redesign

Spark in Scala on AWS

- Spark promises:
  - Seamless horizontal scaling
  - High level APIs for machine learning (MLlib)
- Scala promises:
  - Type safety
  - Real multithreading
- AWS promises:
  - (Nearly) unlimited computing power
  - Cheap data storage for unifying input data

Horizontal Scaling



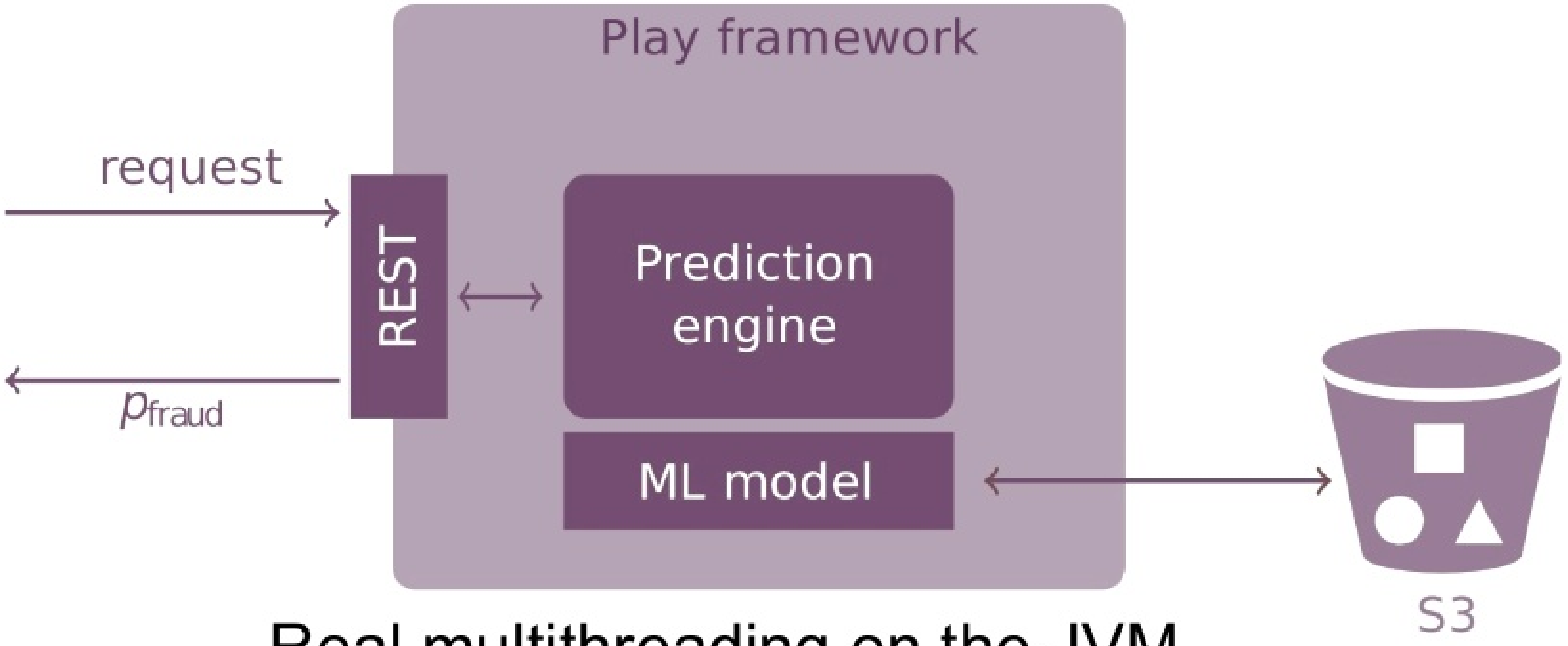
Type Safety for Configuration

old config

```scala
package de.zalando.payana.lf.model.appDe

case class DynamicFeat1() extends BasicScorer(
  "SecretScore1", SecretScorer1(), Seq("1970-01"))

case class DynamicFeat2() extends BasicScorer(
  "SecretScore2", SecretScorer2(), Seq("1970-02", "1970-03"))

case class DynamicFeat3() extends BasicScorer(
  "SecretScore3", SecretScorer3(), Seq("1970-04"))

case class DynamicFeat4() extends BasicScorer(
  "SecretScore4", SecretScorer4(), Seq("1970-05", "1970-06"))

case class Forest_1970_09() extends Model[OrderDao](
  randomForest("model_1970_09"),
  BasicFeatures ++ CustomerHistoryFeatures ++
    Seq(DynamicFeat1(), DynamicFeat2(), DynamicFeat3(), DynamicFeat4()),
  Seq("1970-09", "1970-10", "1970-11")
)

class Job_2016_02_02 extends Job {
  override def execute(implicit ctx: Context): Unit = {
    ...
    val allModels = Seq(Forest_1970_09(), Ridge_1970_09())
    for (model <- allModels) {
      Learn(model, ...) andThen Predict(model, ...)
    }
  }
}
```

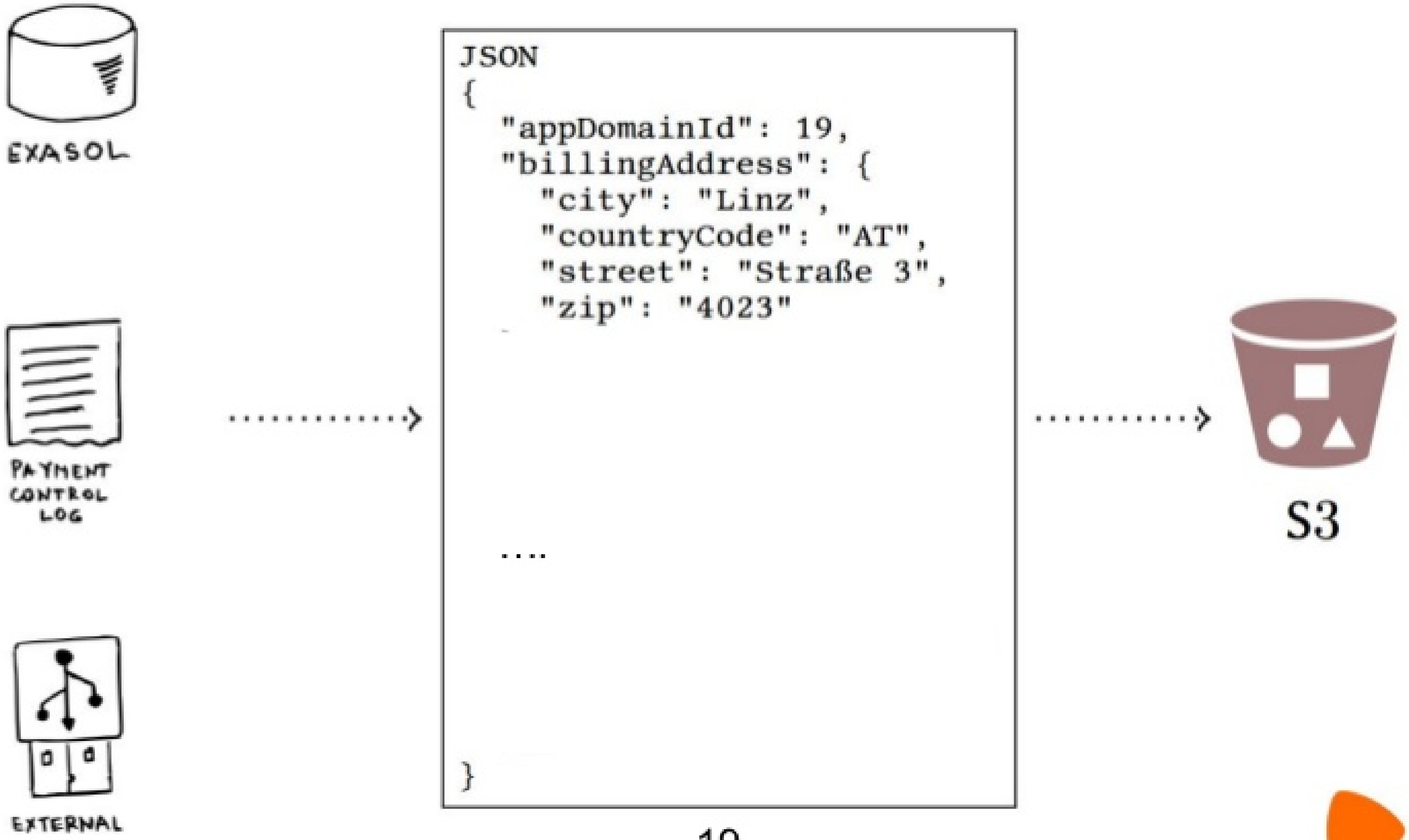
new config as code

Scala Runtime



Real multithreading on the JVM

Unify Data Sources on AWS



Comparison

Lines of Code

Language	Comment	Code
Scala	1192	3322
Python	411	6314

http://cloc.sourceforge.net

Learning Time

- Scenario 1: Old solution
  - Python-based learning framework
  - In-house cluster on a single machine with 10 cores

- Scenario 2: New solution
  - Spark-based learning framework
  - run on AWS with 1 master and 5 workers

→ Overall learning time drops by **factor two**

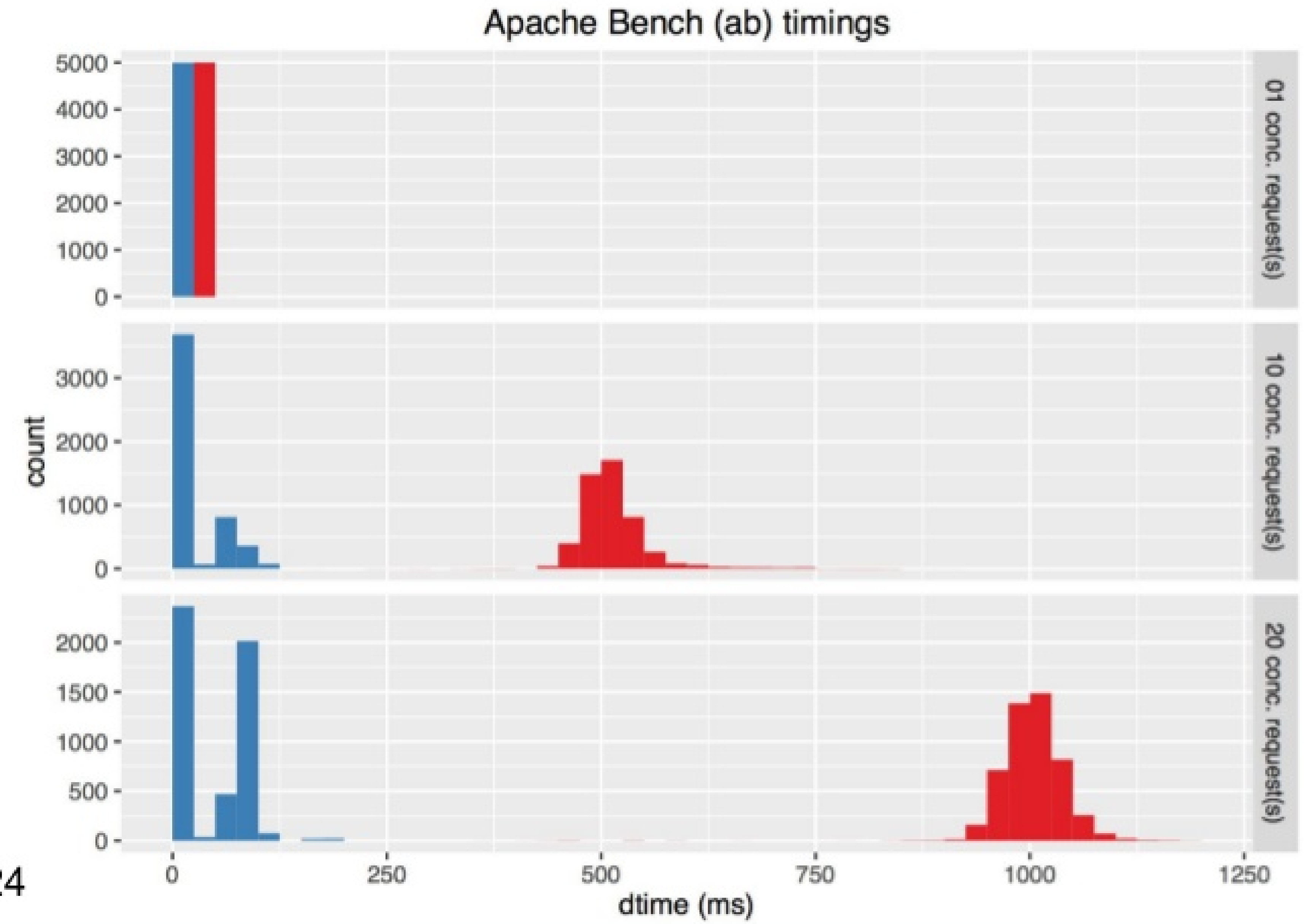
Data size

- Py/slurm
  - limited by single node's memory

- Spark/AWS
  - horizontal scaling
  - 10x more data points in same time

→ Able to train with more data; higher accuracy

Prediction Time



Lessons Learned

\+

Spark & AWS eliminated all pain points
Scaling works as expected
Speedup in learning and execution

\-

Steep learning curve for Scala
Hard to debug distributed execution
Maturity level of MLlib (as of version 1.6)

zalando

THANK YOU.

Stanimir Dragiev, Patrick Baier
firstname.lastname@zalando.de

https://tech.zalando.com/blog/scalable-fraud-detection-fashion-platform

SPARK SUMMIT EUROPE 2016