JAVA
Meeting
Point.

SHARING IS CARING

Spark for Java Devs

Pasha Finkelshteyn, JetBrains

Moscow

asm0di0

BY DINS

Who am I

- ex system administrator
- ex developer
- ex team lead
- ex data egineer
- developer advocate for big data

Together >14 years in IT

Who are data engineers?

Responsibilities:

- Build your DWH
- Build your DMP
- Transfer and store your data

As effective and fast as possible

What is Big Data

- Doesn't fit the single node (or Excel)
- Maybe scaled when growing
- Enough data to make reliable business solutions

What is the most popular tool in DE?

APACHE
Spark
TM

What `J(ava|VM)` dev should know

- Lazy sequences (streams, sequences, scala lazy)
- Functional operations
- SQL (or SQL-like)

Differences in handling of Big and Small data

Big data processing is

- Distributed
- Requires sending large amount of data between nodes
- Is built on map-reduce primitives

What did we learn?

- Spark is like streams
- Spark supports joins
- Spark allows us to work with big data in the same manner as with small
- There are 3 APIs in Spark:
    - Datasets (typed)
    - Dataframes (untyped)
    - SQL