Testing Data Pipelines

Himalayan peaks

Ksenia Tomak, Dodo Engineering
Pasha Finkelshteyn, JetBrains

Who we are

What is Big Data

Who are DEs?

What is pipeline?

Who needs pipelines

QA of pipeline

QA ?= QC

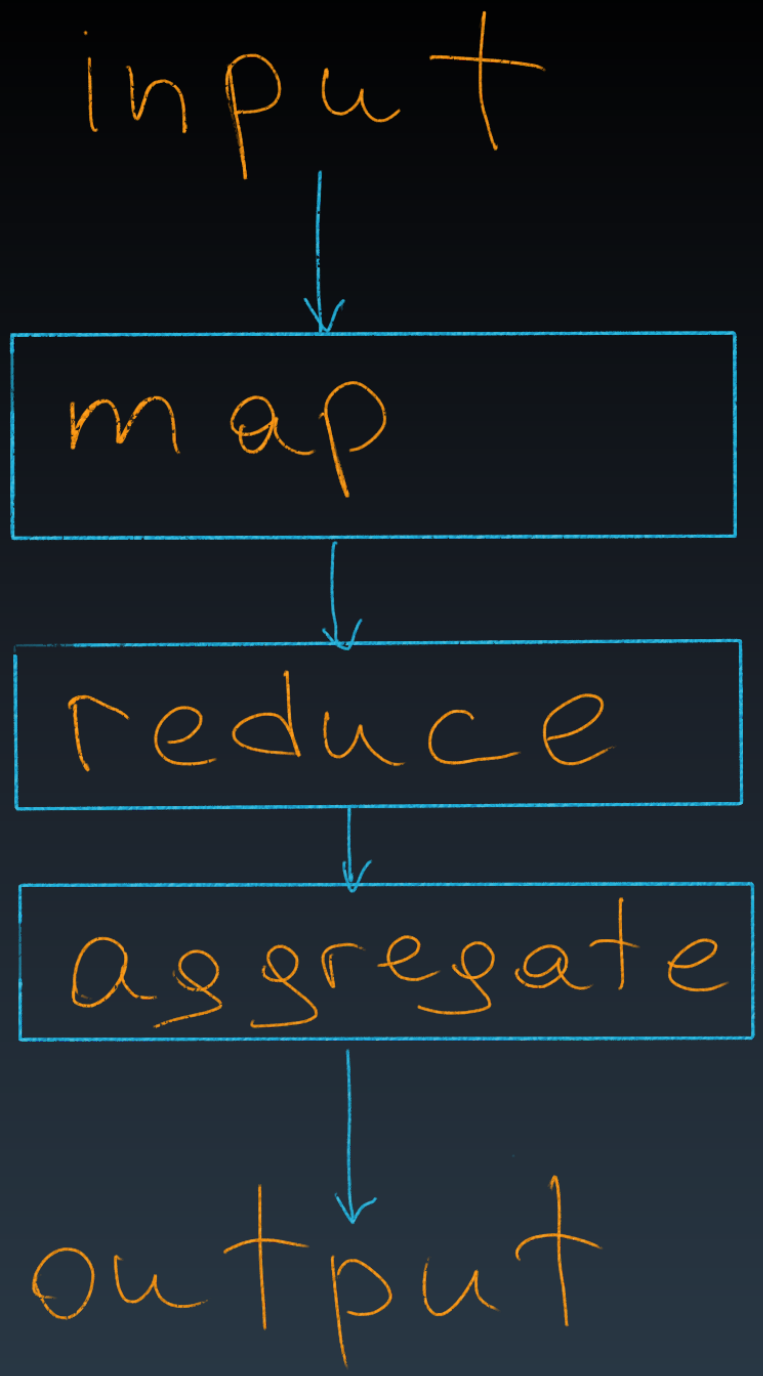
QA of pipeline

QA ≠ QC

QA is about processes, and not only about software quality.

Pyramid of testing. Unit

Typical pipeline



Unit testing of pipeline

What may we test here?

A pipeline should transform data correctly!

Correctness is a business term

Let's paste fakes!

Fake/mock input data

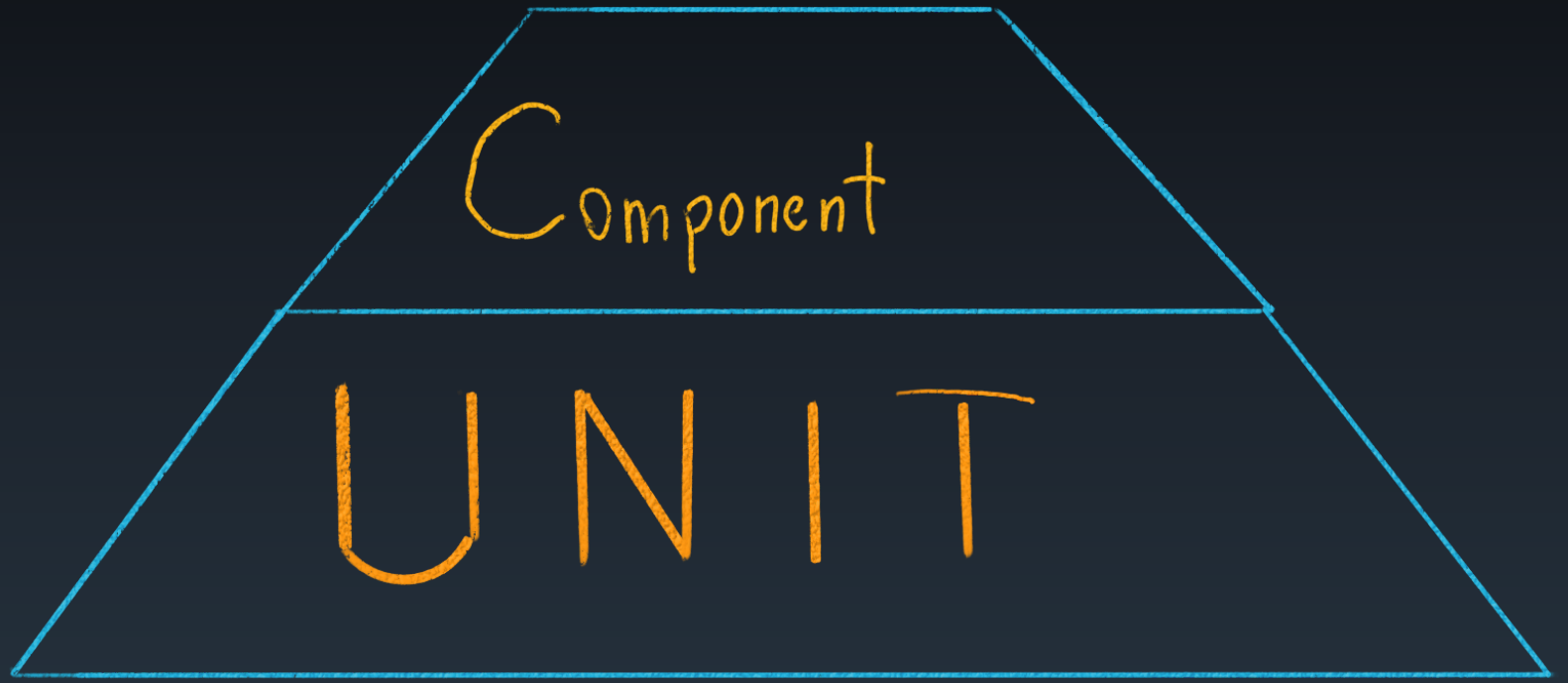
Reference data at the end of pipeline

Separate Function

Tools

holdenk/spark-testing-base ← Tools to run tests
MrPowers/spark-daria ← tools to easily create test data

Component testing



TESTCONTAINERS

Test Containers

```
import sqlalchemy
from testcontainers.mysql import MySqlContainer

with MySqlContainer('mysql:5.7.17') as mysql:
    engine = sqlalchemy.create_engine(mysql.get_connection_url())
    version, = engine.execute("select version()").fetchone()
    print(version)  # 5.7.17
```

TestContainers

Supported languages:

- Java (and compatibles: Scala, Kotlin, etc.)
- Python
- Go
- Node.js
- Rust
- .NET

TestContainers

Separate Function