Himalayan Peaks of Testing Data Pipelines

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

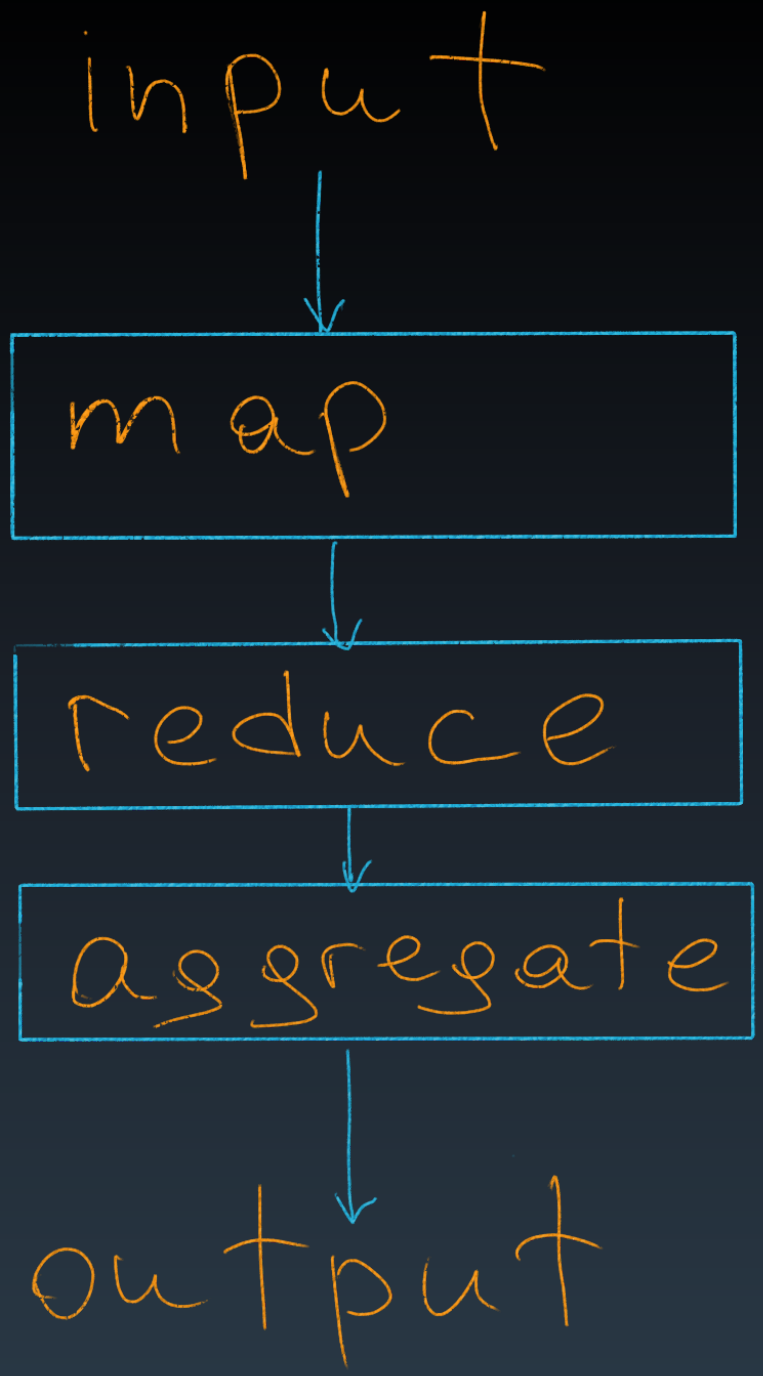
QA of pipeline

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

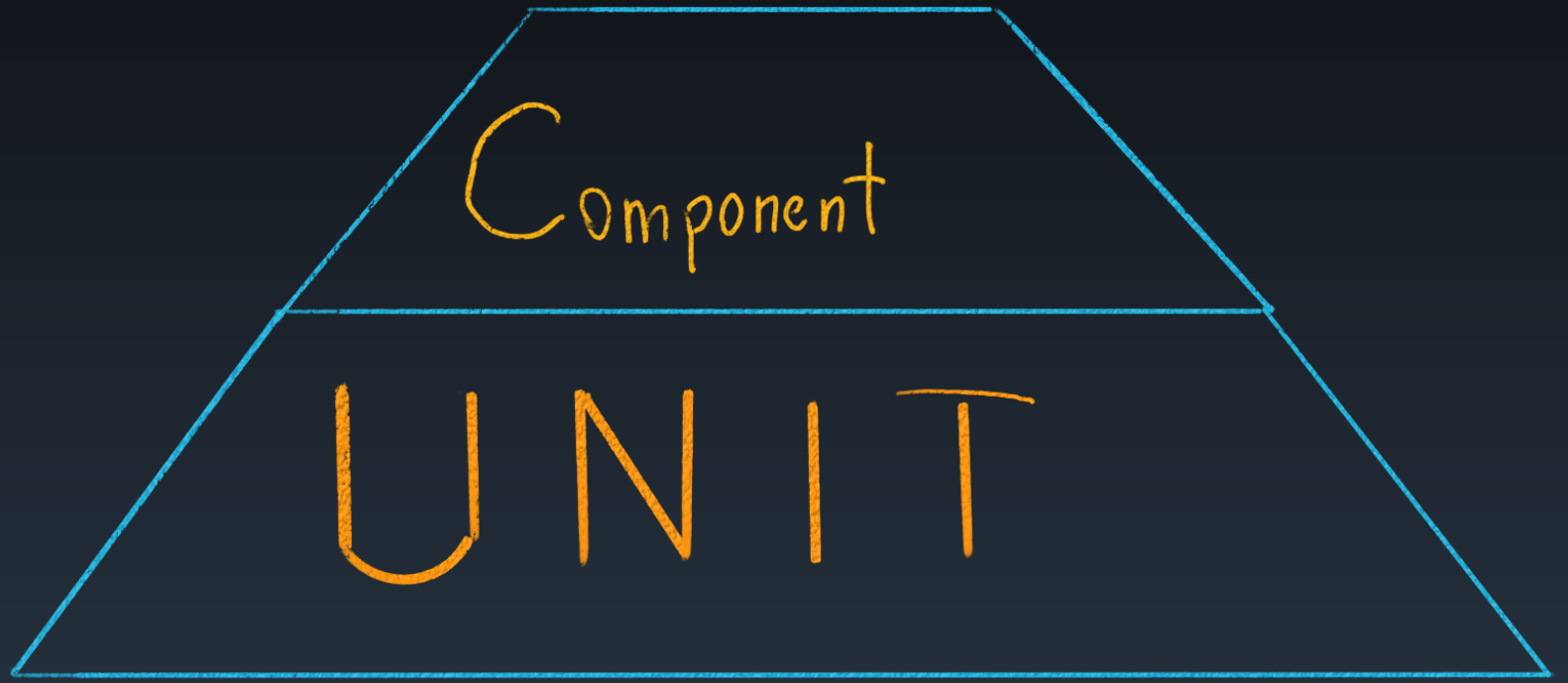
Reference data at the end of pipeline

Separate Function

Tools

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

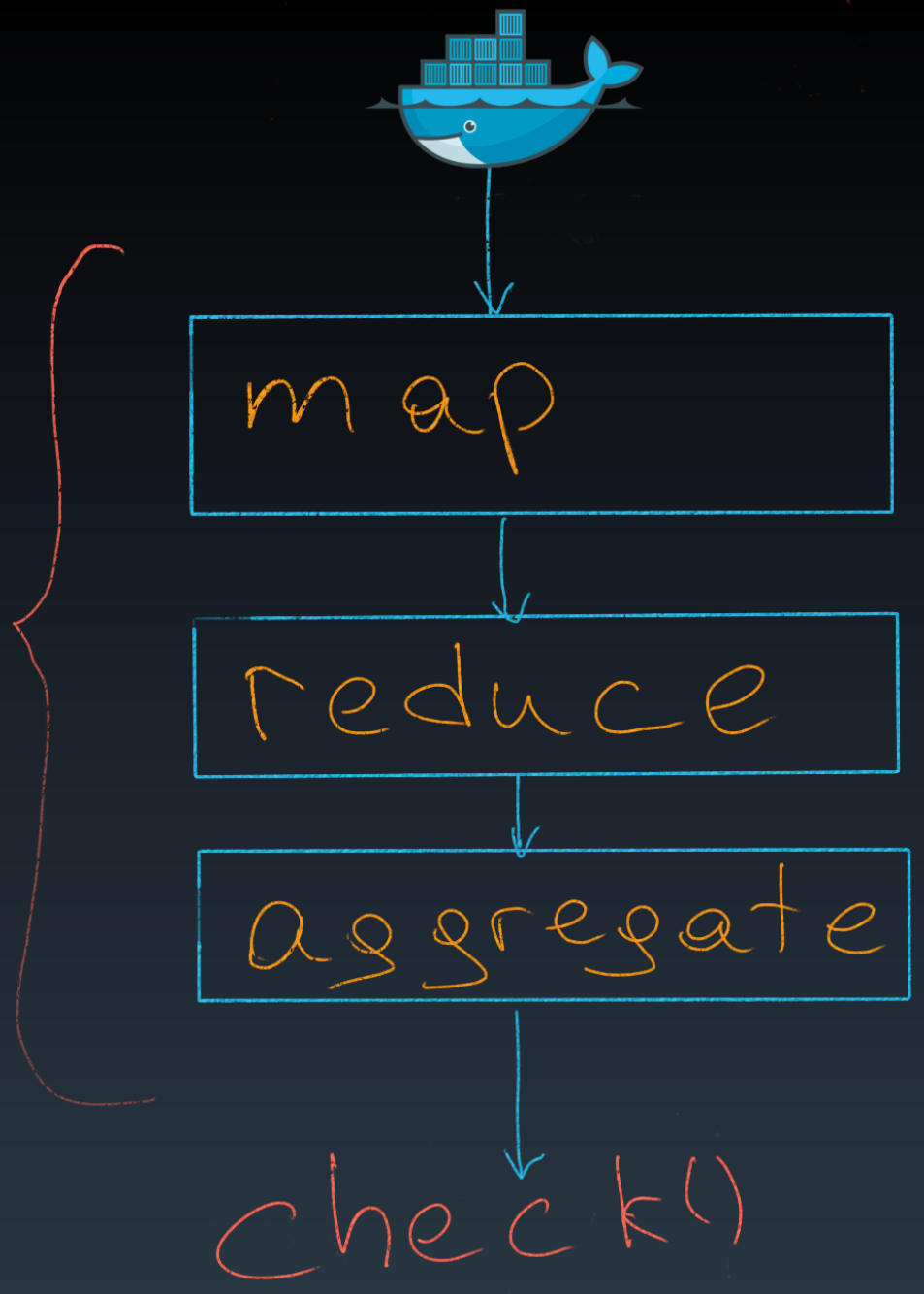
Component testing



TESTCONTAINERS

TestContainers

Separate Function



TestContainers

Supported languages:

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

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
```

Integration Tests

Why test containers are not enough?

- vendor lock tools (DB, processing, etc.)
- real data
- external error handling

Integration Tests: How to

- get data samples from prod, anonymize it
- deploy full data backup on stage, depersonalize it ($$$)
- run parallel job with different sink

Using production data for testing in a post GDPR world

Data expectations

Test:

- ✔ no data
- ✔ valid data
- ✔ empty partitions
- ? invalid data
- ? illegal data format

Data expectations. Tools:

- great expectations,
- Deequ

Use Dead letter queue pattern for broken data to prevent:

- data loss
- data traffic jam

Monitoring

Why?

How to collect:

- StreamingQueryListener, QueryExecutionListener
- foreachBatch aggregates, sink as logs

Monitoring visualization

[Visualizing Data Timeliness]

End-to-End tests

Compare with reports, old DWH

Multiple dimentions:

- data
- data latency
- performance, scalability

Performance Tests

Best performance test - initial data load

(image with initial data load + next microbatches loading)