

DATA INTEGRATION

in the world of microservices

About me



Valentine Gogichashvili

Head of Data Engineering @ZalandoTech

twitter: @valgog

google+: +valgog

email: valentine.gogichashvili@zalando.de

DAMEN

HERREN

KINDER

 zalando

 Mein Konto ▾

 Wunschzettel

 Warenkorb

Neu

News&Style

Bekleidung

Schuhe

Sport

Accessoires

Wäsche

Premium

Marken

Sale %

Liebblingsprodukt suchen...



SOMMERSTRICK

DIE COOLE MASCHE FÜR HEISSE TAGE

ZUM SALE >

ZU DEN LOOKS >

ZUR AUSWAHL >



**DAS ZALANDO
FASHION HOUSE**

ERLEBE MIT UNS
DIE WELT DER MODE





One of Europe's largest online fashion retailers

15 countries

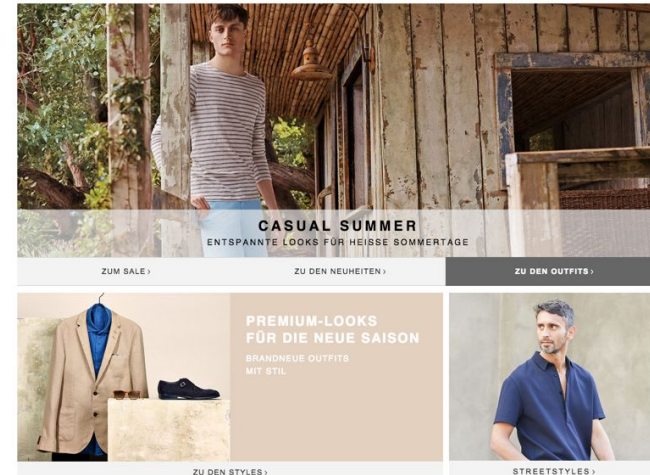
4 fulfillment centers

15+ million active customers

2.9 billion € revenue 2015

150,000+ products

9,000+ employees



Zalando Technology

BERLIN

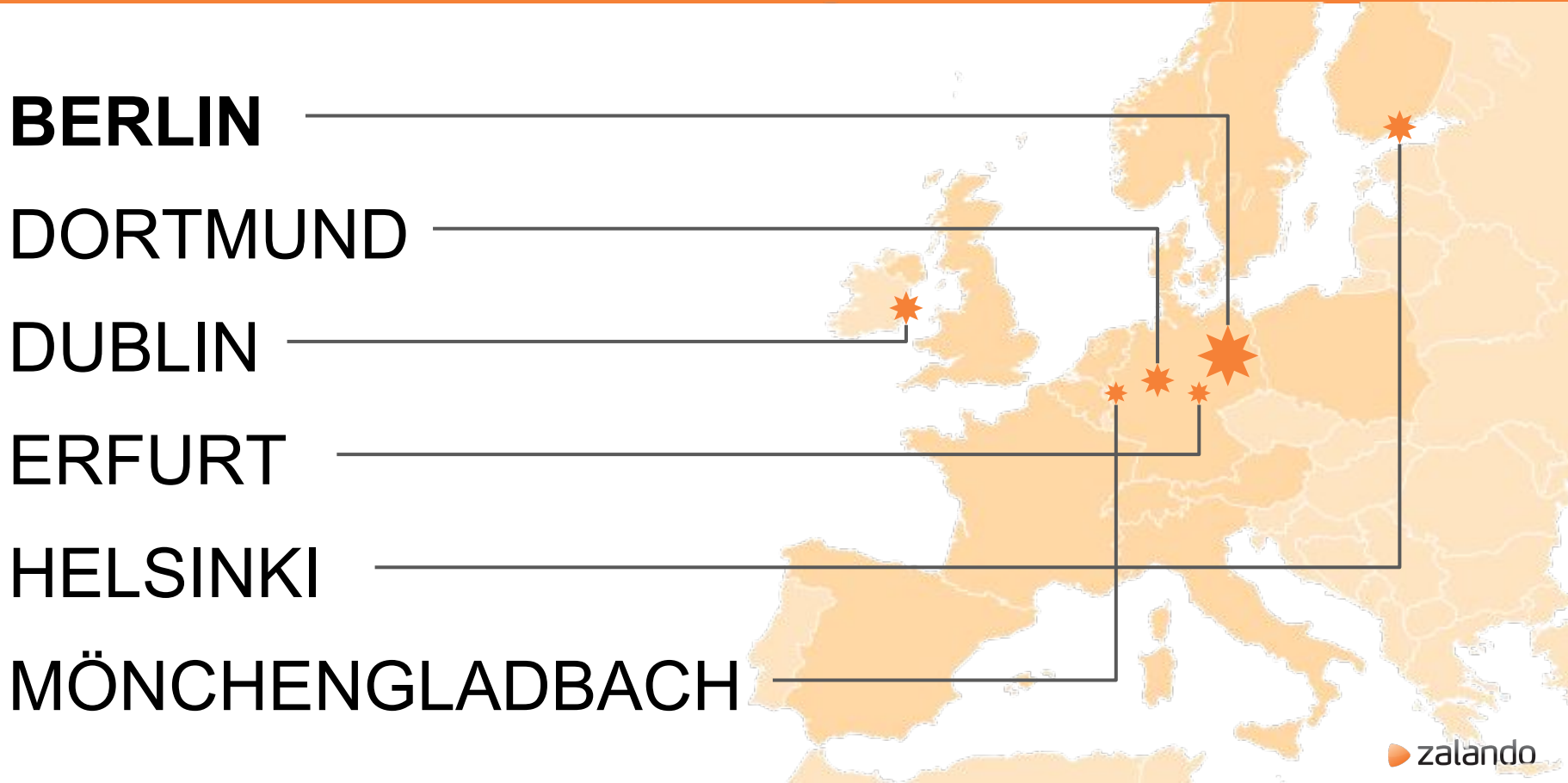
DORTMUND

DUBLIN

ERFURT

HELSINKI

MÖNCHENGLADBACH





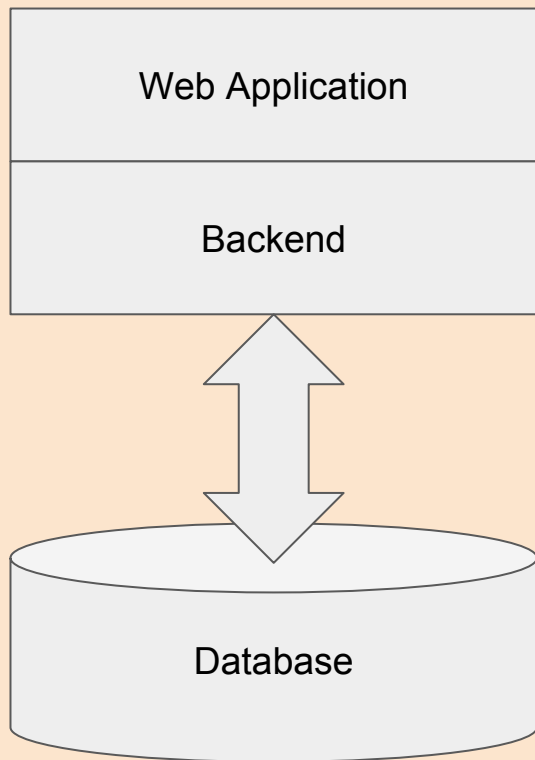
900+ TECHNOLOGISTS

Rapidly growing
international team

<http://tech.zalando.de>

Good old small world

Once upon a time...



Started as a tiny online shop

Prototyped on Magento (PHP)

Used MySQL as a database

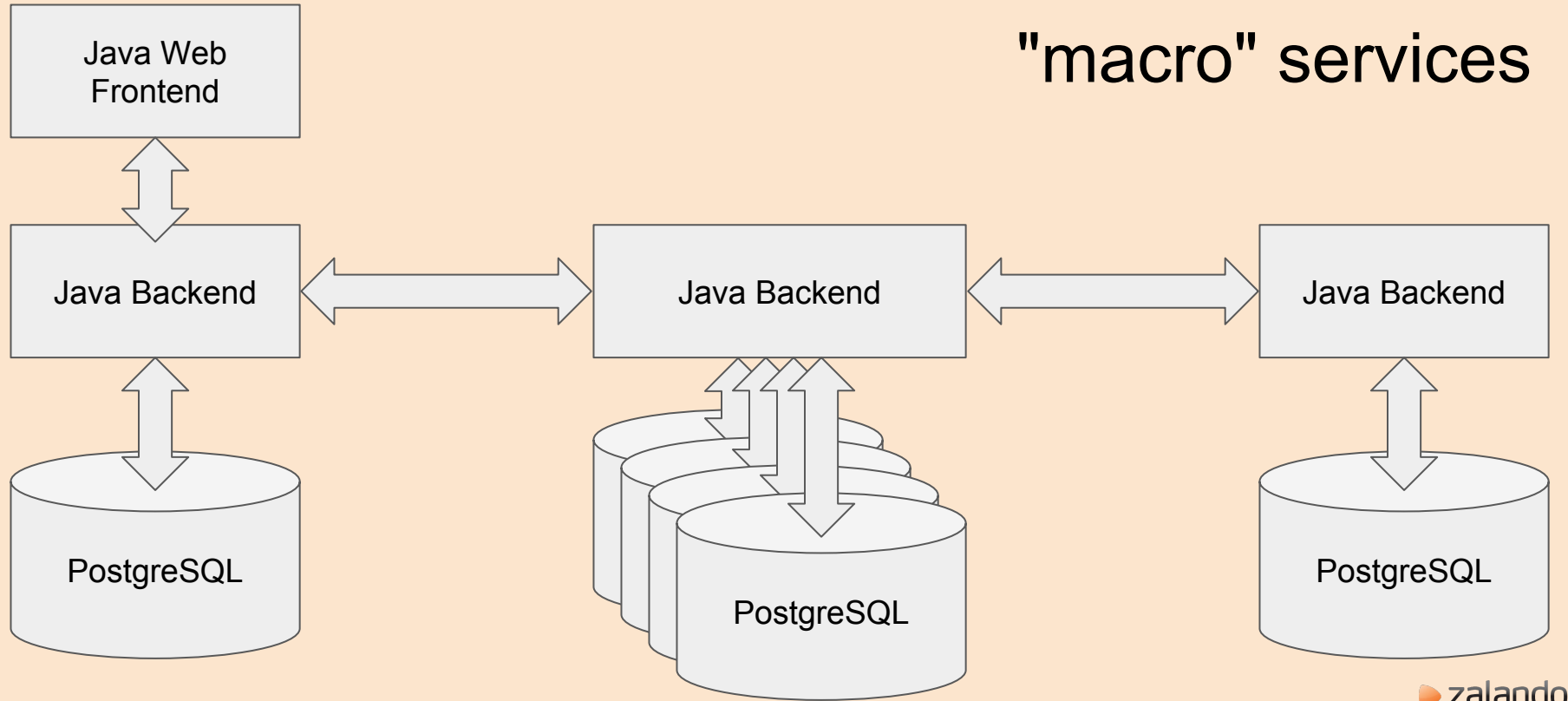
REBOOT

REBOOT

5½ years ago

- Java
 - macro service architecture with SOAP as RPC layer
- PostgreSQL
 - Heavy usage of Stored Procedures
 - 4 databases + 1 sharded database on 2 shards
- Python for tooling (i.e code deploy automation)

REBOOT



REBOOT

- PostgreSQL
 - Heavy usage of Stored Procedures
 - clean transaction scope
 - very clean data
 - processing close to data

REBOOT

- PostgreSQL
 - [Java Sproc Wrapper](#)
 - complex type mapping
 - transparent sharding

REBOOT

- PostgreSQL
 - introduced DBDIFF database schema management
 - schema based Stored Procedure versioning

Live long and prosper...

Very stable architecture that is still in use in the oldest (vintage) components

We implemented everything ourselves starting from warehouse and order management and finishing with Web Shop and Mobile Applications

Live long and prosper...



"I want to code in Scala/Clojure/Haskell because it is cool and compact"



"But nobody will be able to support your code if you leave the company, everybody should use Java, learn SQL and write Stored Procedures"



"SQL is cool but f*ck you, I am moving on to another company where I can use cool technologies!"

RADICAL AGILITY

Radical Agility



AUTONOMY

PURPOSE

MASTERY

Autonomy

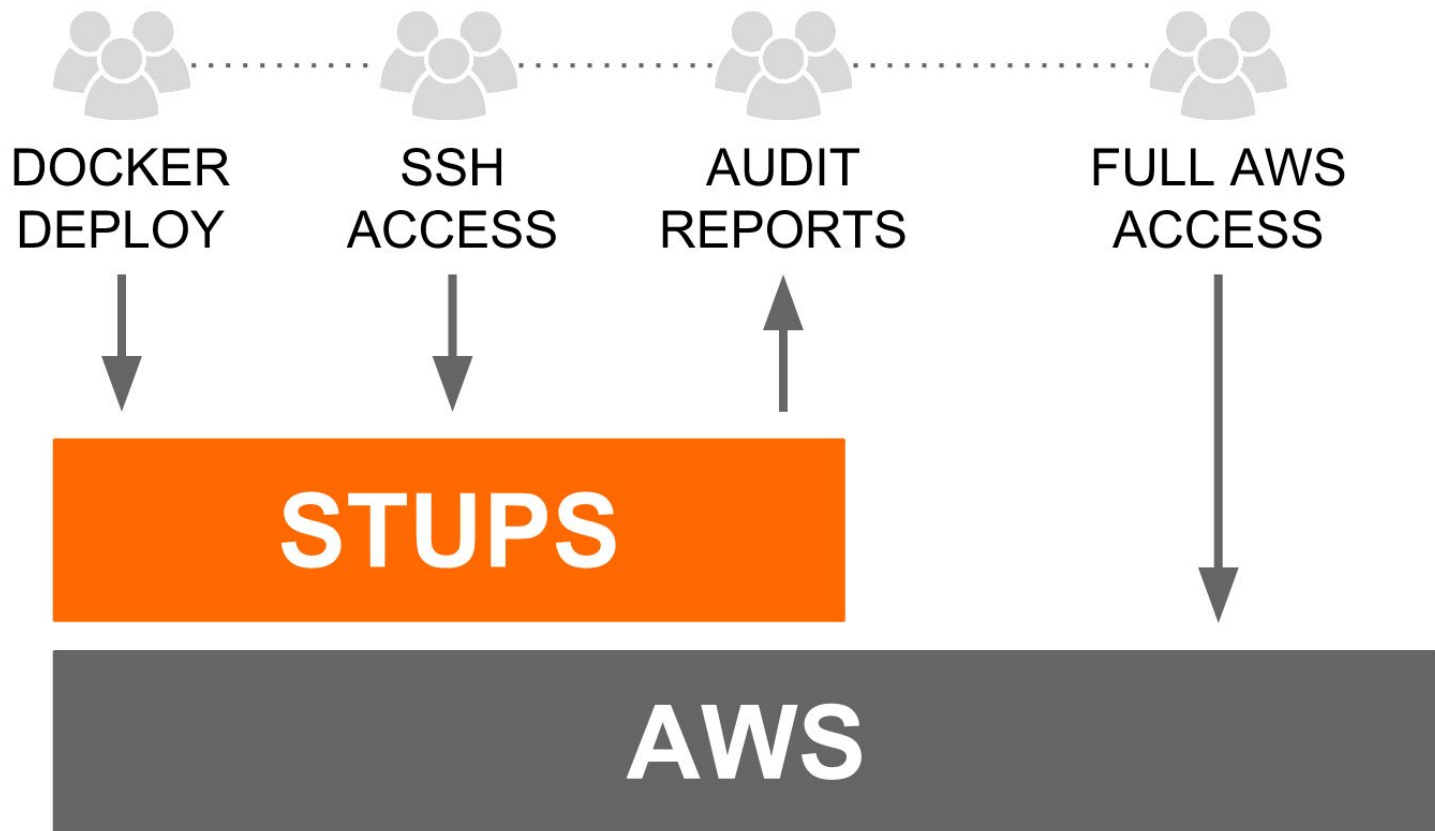
Autonomous teams

- can choose own technology stack
- including persistence layer
- are responsible for operations
- should use isolated AWS accounts

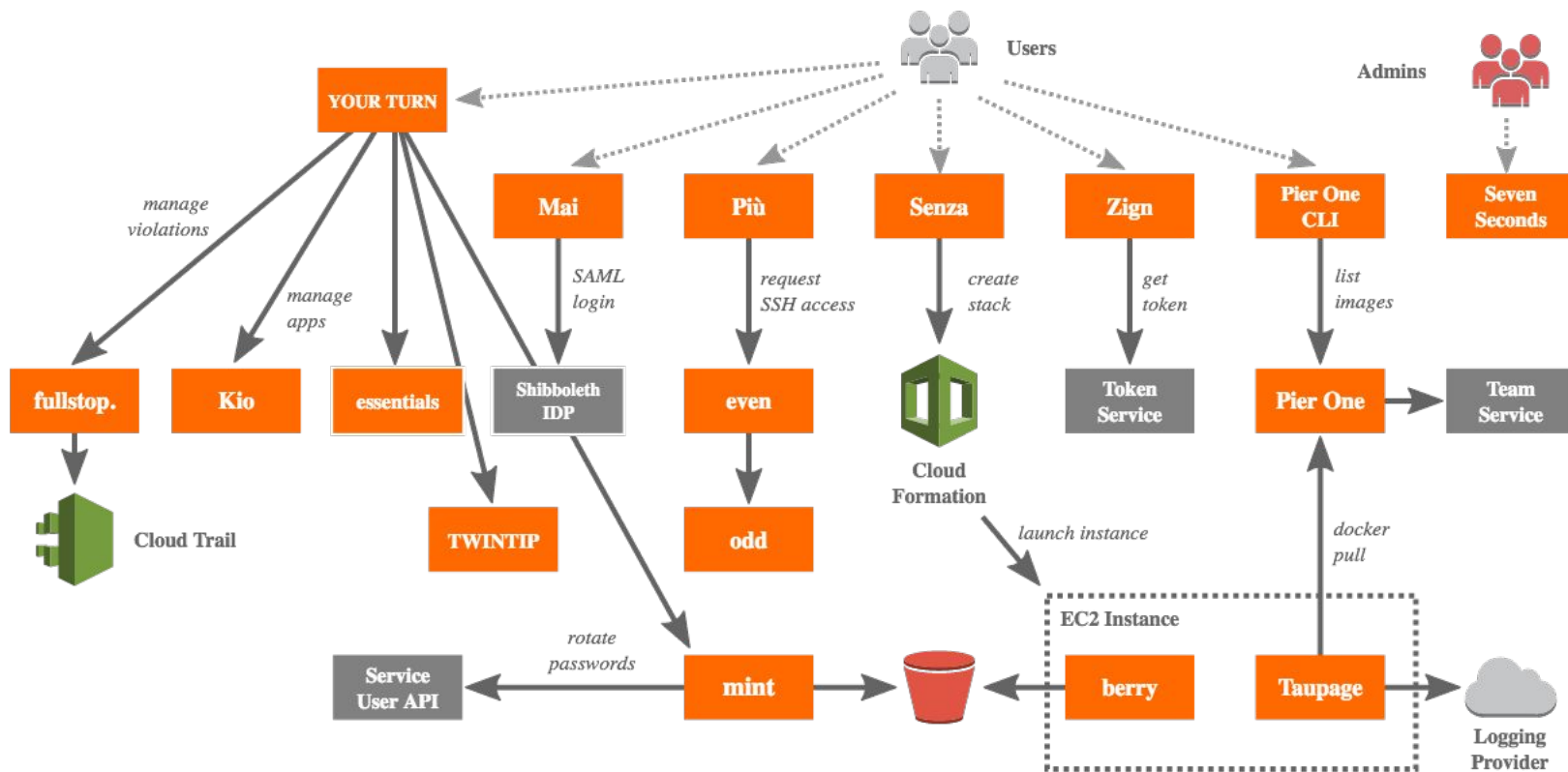
Autonomy is not an Anarchy



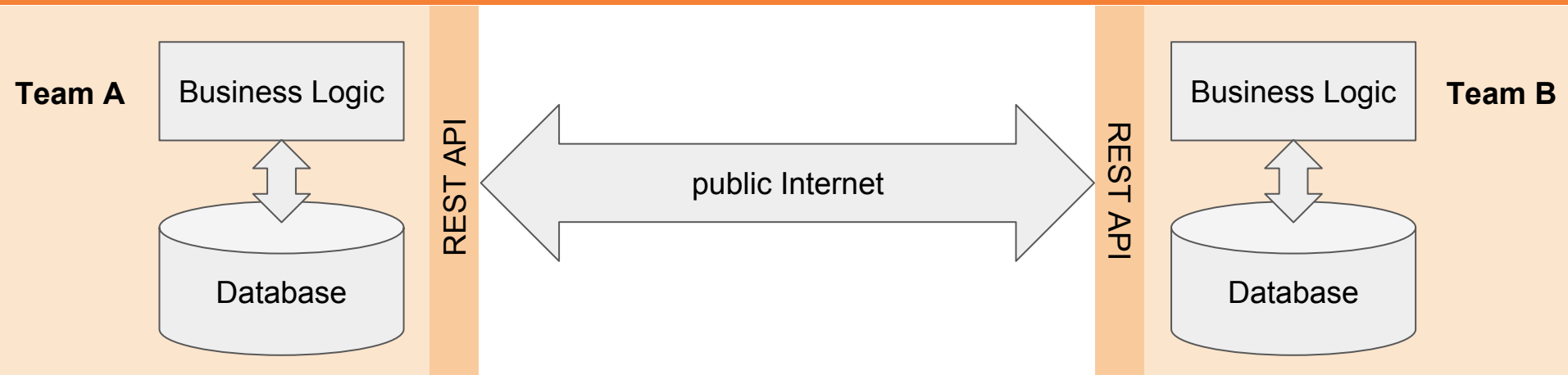
Supporting autonomy — STUPS



Supporting autonomy — STUPS

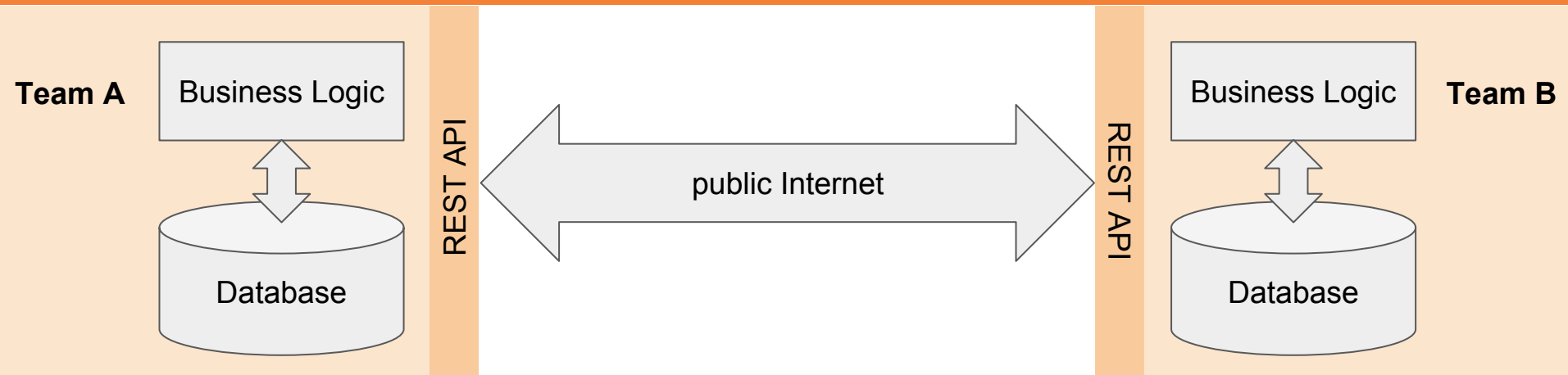


Supporting autonomy — Microservices



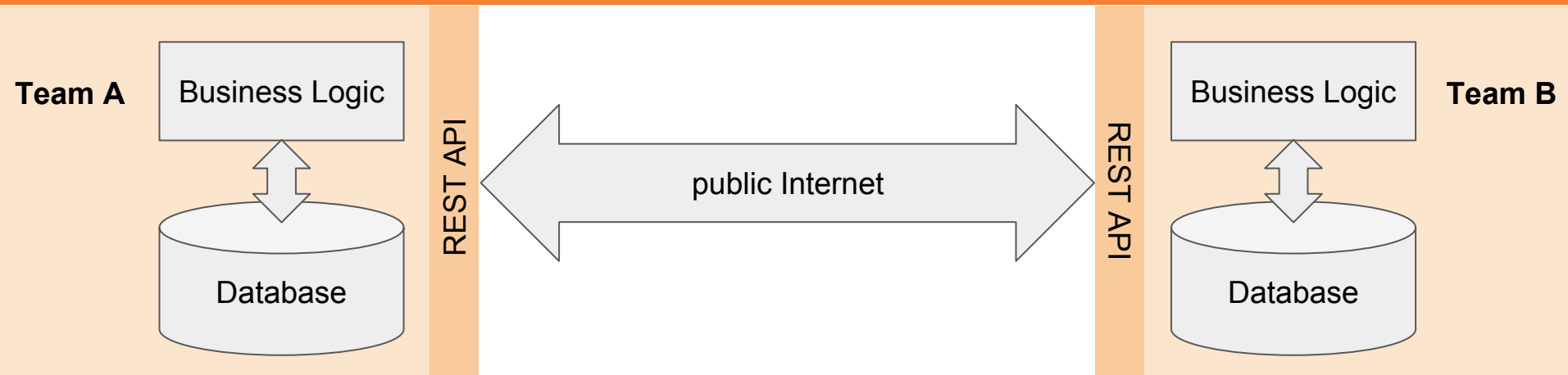
- Applications communicate using REST APIs
- Databases hidden behind the walls of AWS VPC

Supporting autonomy — Microservices



- Database team is consulting autonomous teams
- [Spilo](#) as STUPS service for PostgreSQL

Supporting autonomy — Microservices



Classical ETL process is impossible!

Data integration in the classical world

Data integration in the classical world



Data integration in the classical world



Developers

DBA

**Business
Intelligence**

Data integration in the classical world

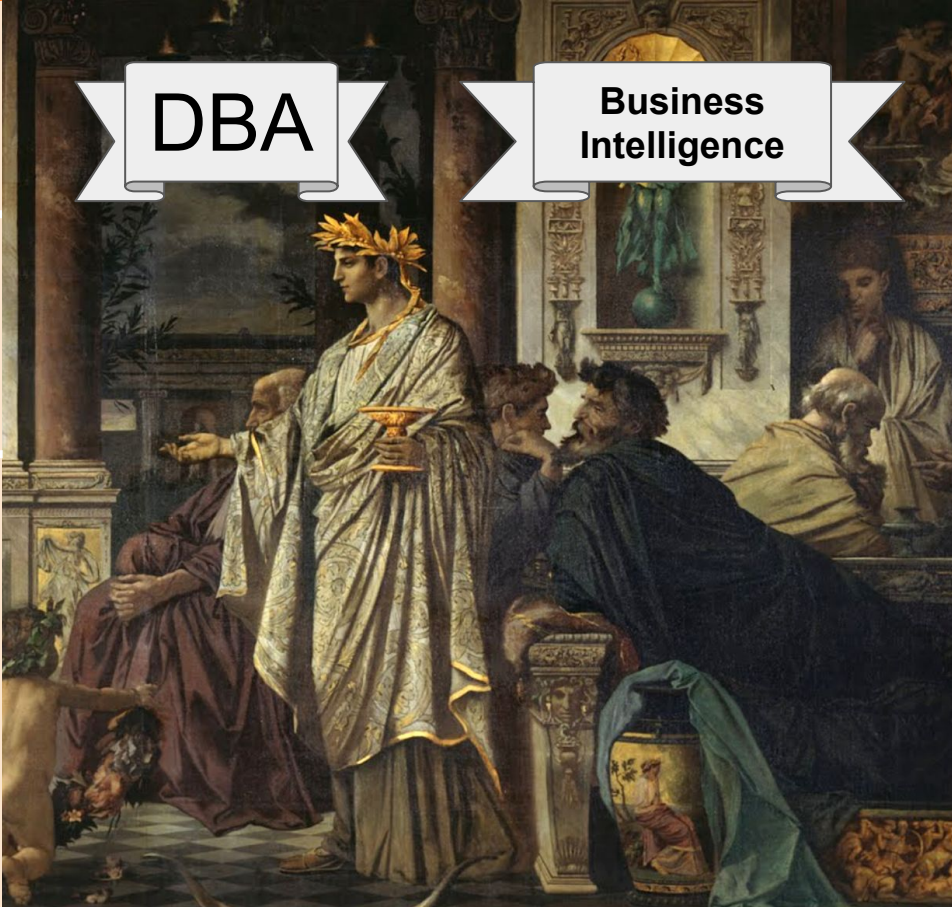
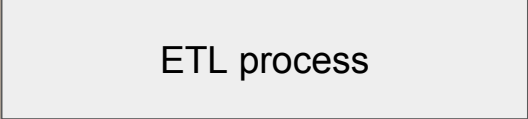
Dev



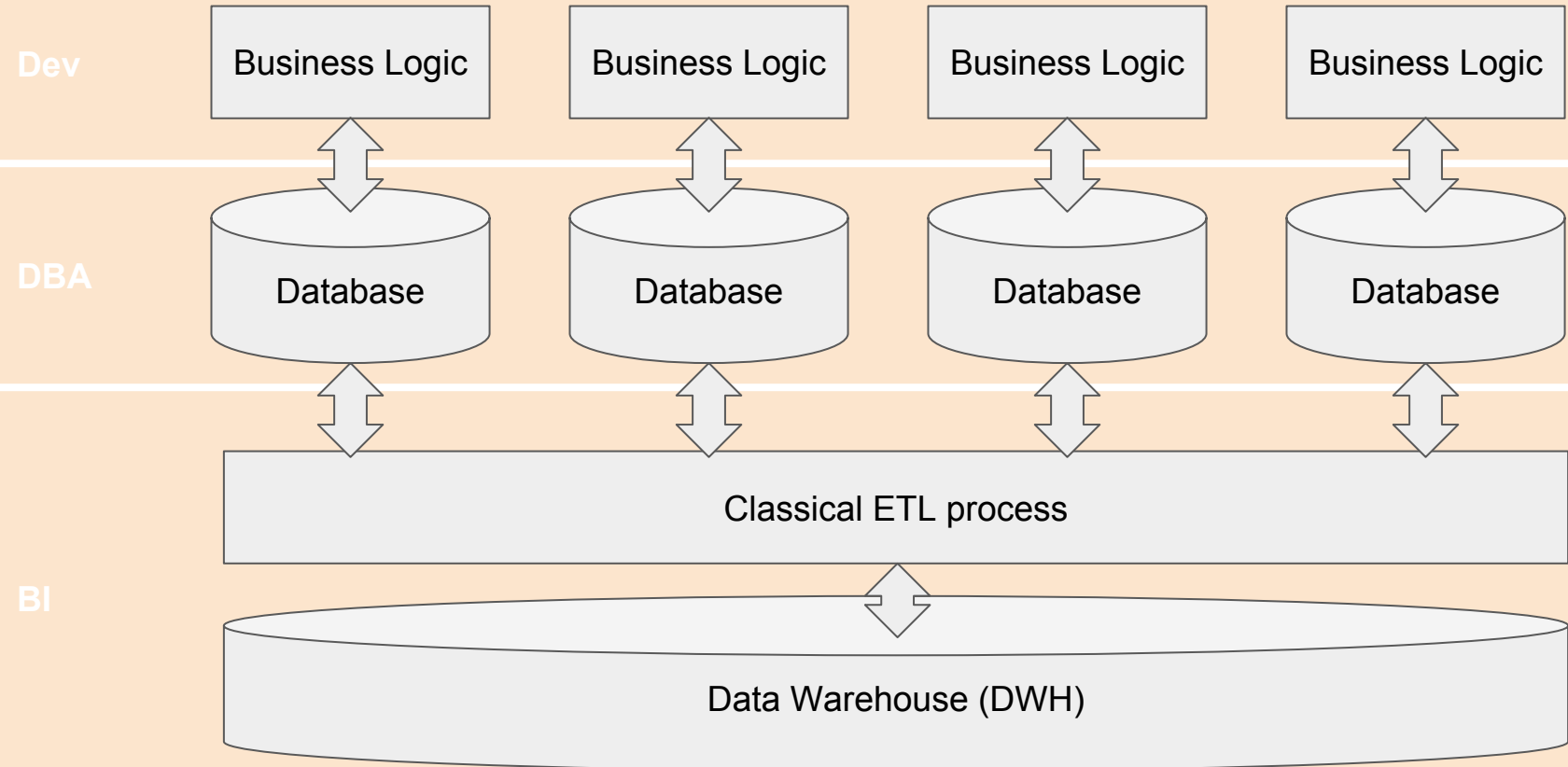
DBA



BI



Data integration in the classical world



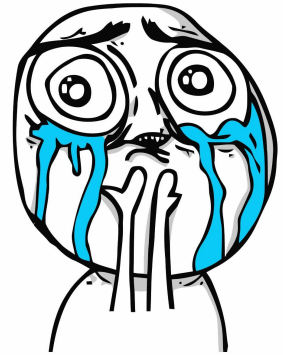
Data integration in the classical world

Classical ETL process

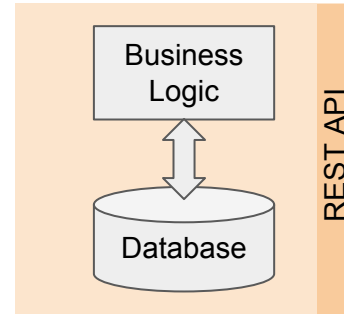
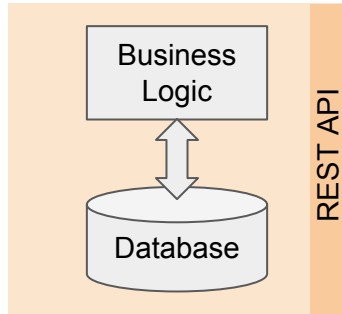
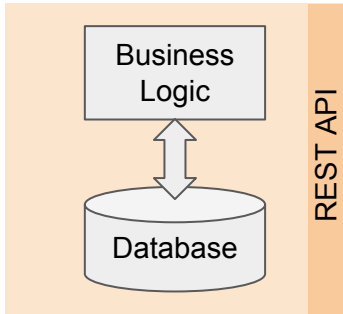
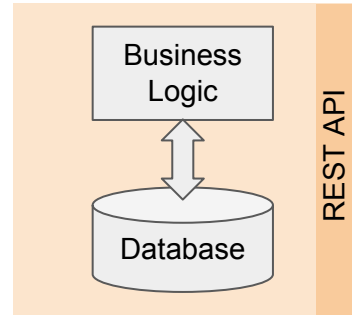
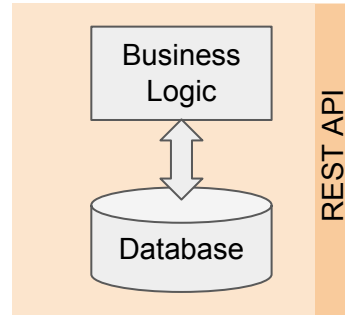
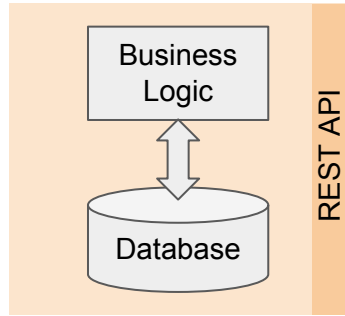
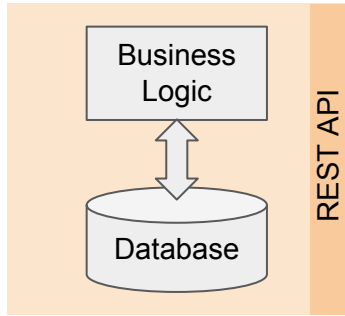
- Use-case specific — a lot of manual work
- Usually outputs data into a Data Warehouse
 - well structured
 - easy to use by the end user (SQL)

Data integration in the world of microservices

Data integration in the world of microservices

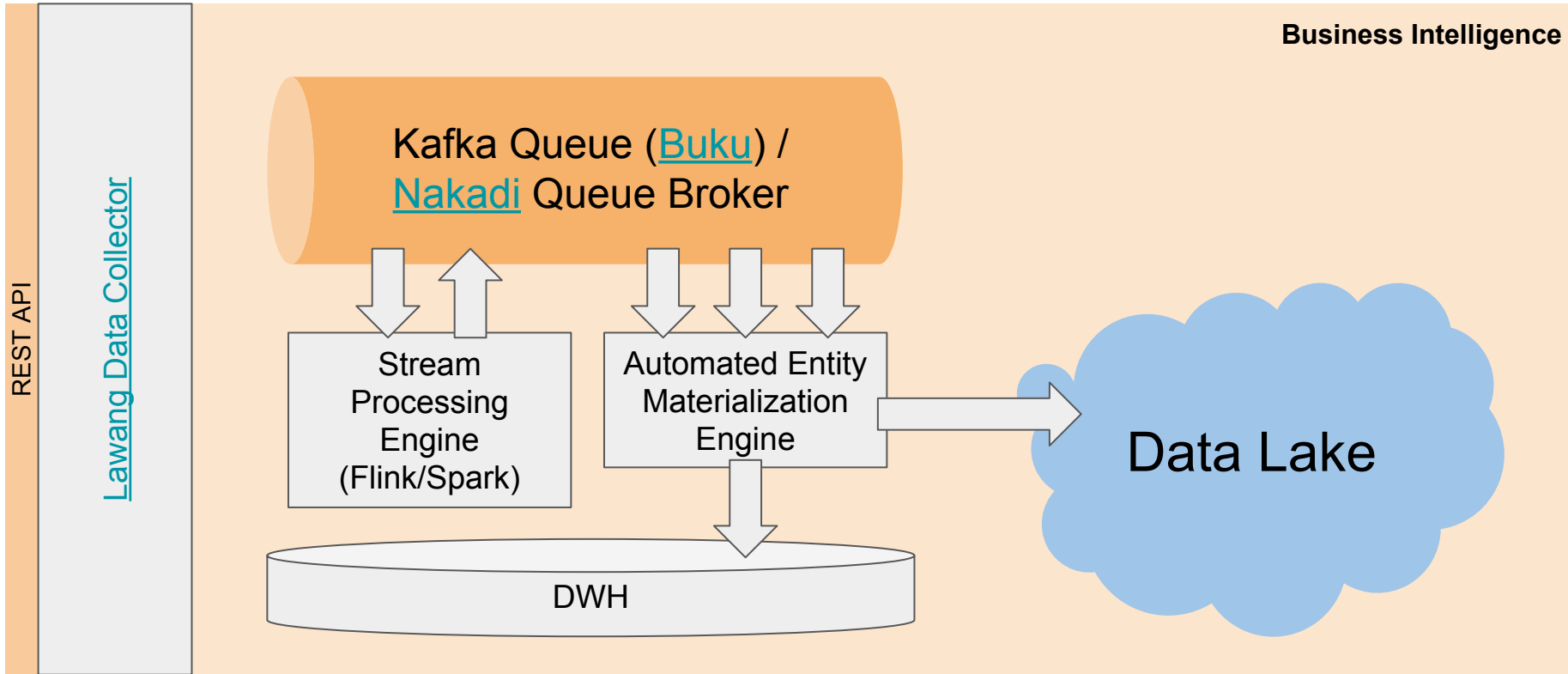


Data integration in the world of microservices

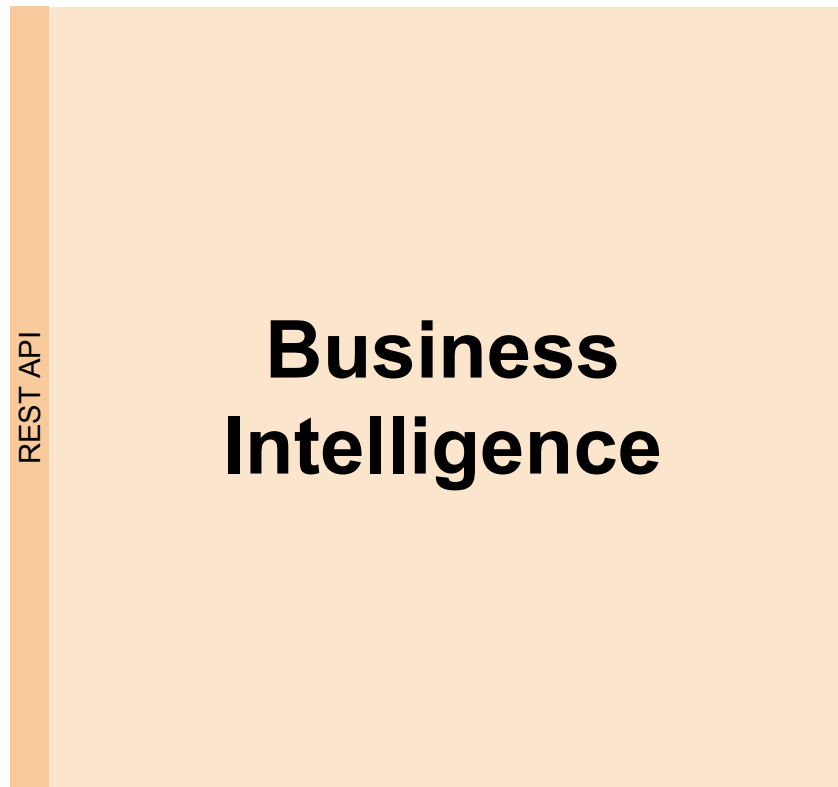
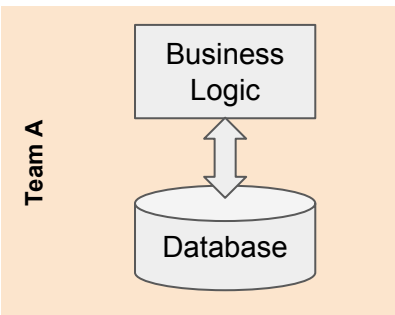


Data integration in the world of microservices

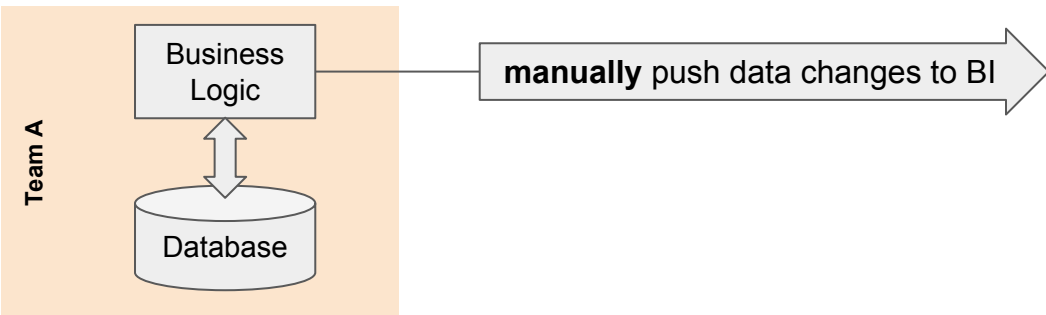
Business Intelligence



Data integration in the world of microservices



Data integration in the world of microservices

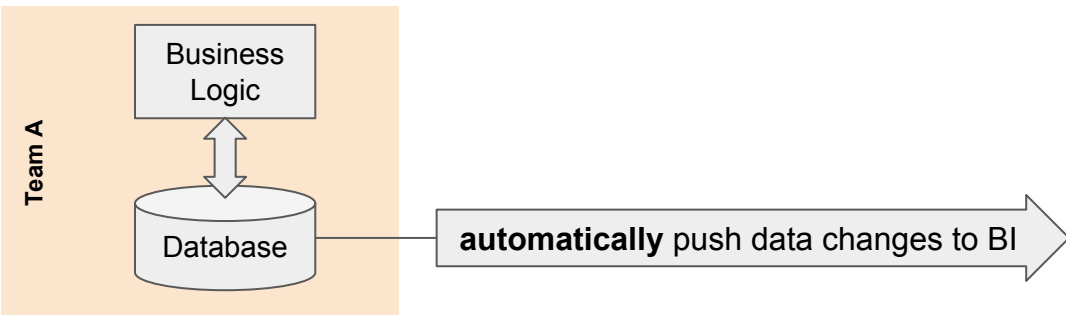


- Error prone
- 2PC is actually needed
- Very difficult to implement

REST API

Business Intelligence

Data integration in the world of microservices



- You cannot miss anything
- No additional work needed on the business logic side

REST API

Business Intelligence

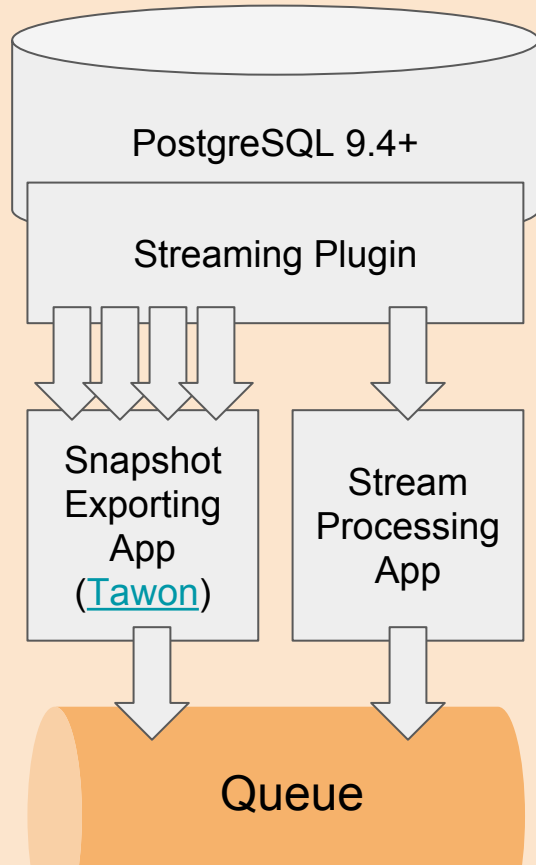
Data integration in the world of microservices

PostgreSQL Logical Replication

Enables automatic Data Change Event Extraction

- [pglogical](#) by 2nd Quadrant (streaming plugin)
- [BottledWater](#) by Confluent (Avro to Kafka streaming)
- [Tawon](#) by Zalando (parallel snapshotting)

Data integration in the world of microservices



Try it yourself

Python support for replication protocol



is the best to make you do experiments

- [psycopg2 PR for replication protocol](#)
- [psycopg2 replication protocol usage](#)
- [Tawon initial snapshot export tool](#)
- [streaming initial database snapshot](#)

Open Source at Zalando Technology

- [Tech Blog of Zalando Technology](#)
- <https://zalando.github.io/> - Open Source Projects
 - pg_view
 - PGObserver
 - Java Sproc Wrapper
 - Patroni
 - Spilo
 - ...

Questions?