Data Integration in the World of Microservices

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 %

Lieblingsprodukt suchen...

ZUM SALE › ZU DEN LOOKS › ZUR AUSWAHL ›

DAS ZALANDO FASHION HOUSE

ERLEBE MIT UNS
DIE WELT DER MODE

zalando

One of Europe's largest online fashion retailers

15 countries

4 fulfillment centers

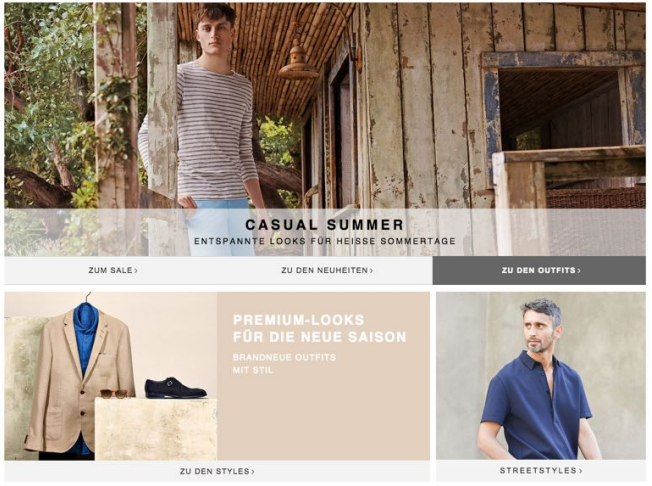
18+ million active customers

~3 billion € revenue

150,000+ products

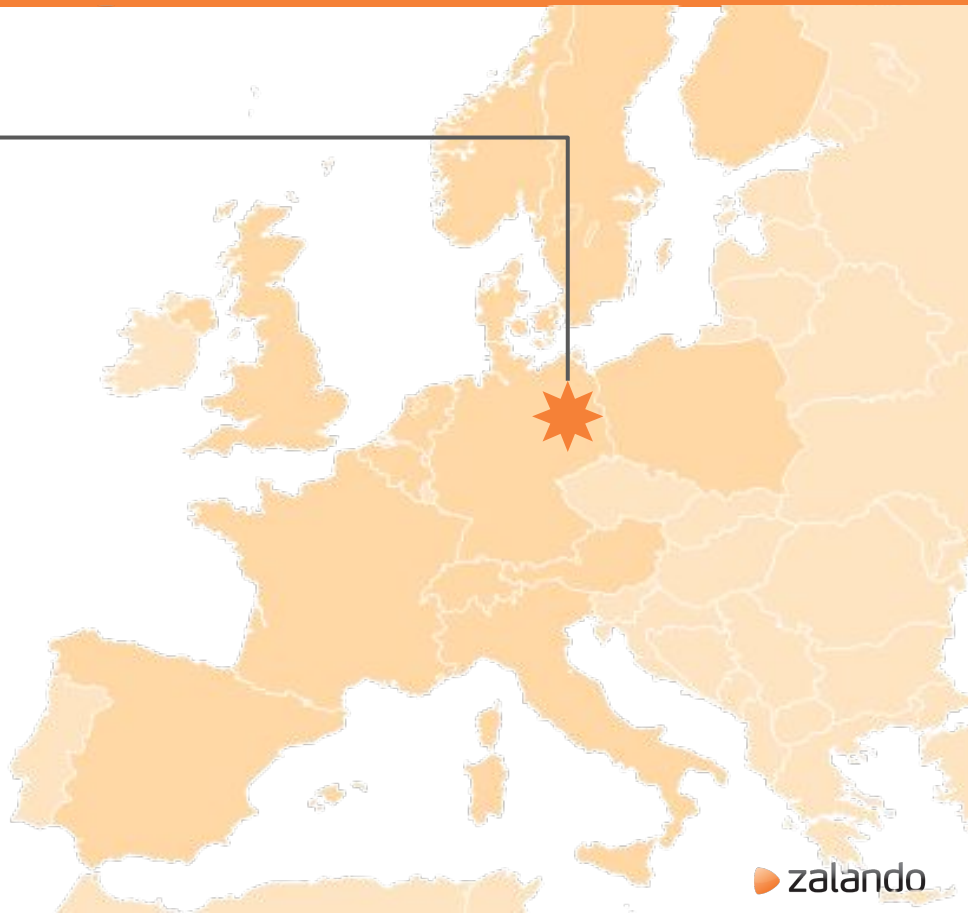
10,000+ employees

135 million visits per month



Zalando Technology

BERLIN



zalando

Zalando Technology

BERLIN

DORTMUND

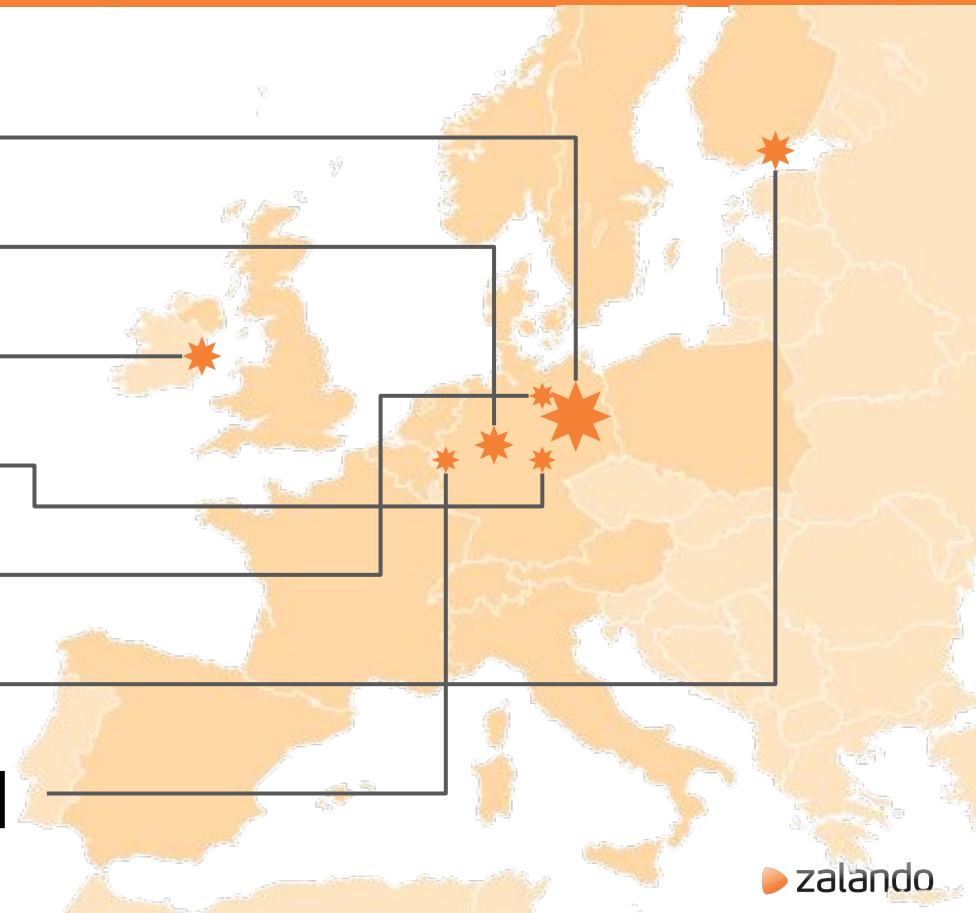
DUBLIN

ERFURT

HAMBURG

HELSINKI

MÖNCHENGLADBACH



zalando

Zalando Technology

1200+ TECHNOLOGISTS

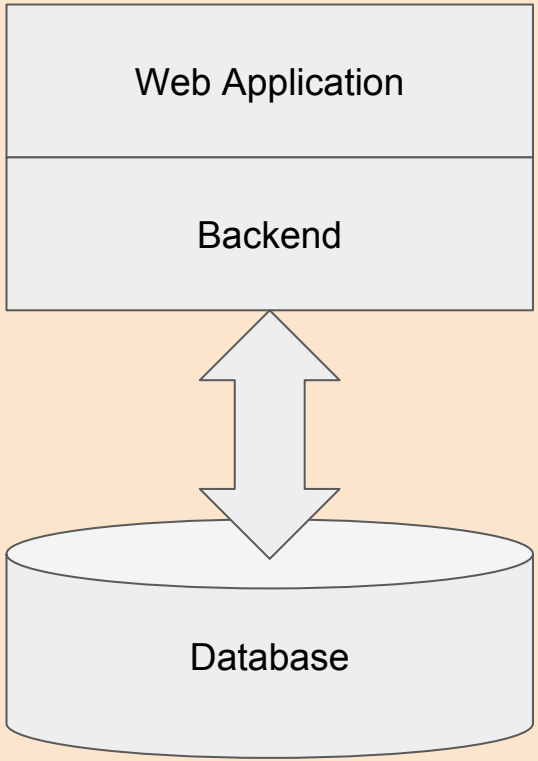
Rapidly growing
international team

http://**tech**.zalando.de

zalando

Good old small world

Once upon a time...



Started as a tiny online shop

Prototyped on Magento (PHP)

Used MySQL as a database

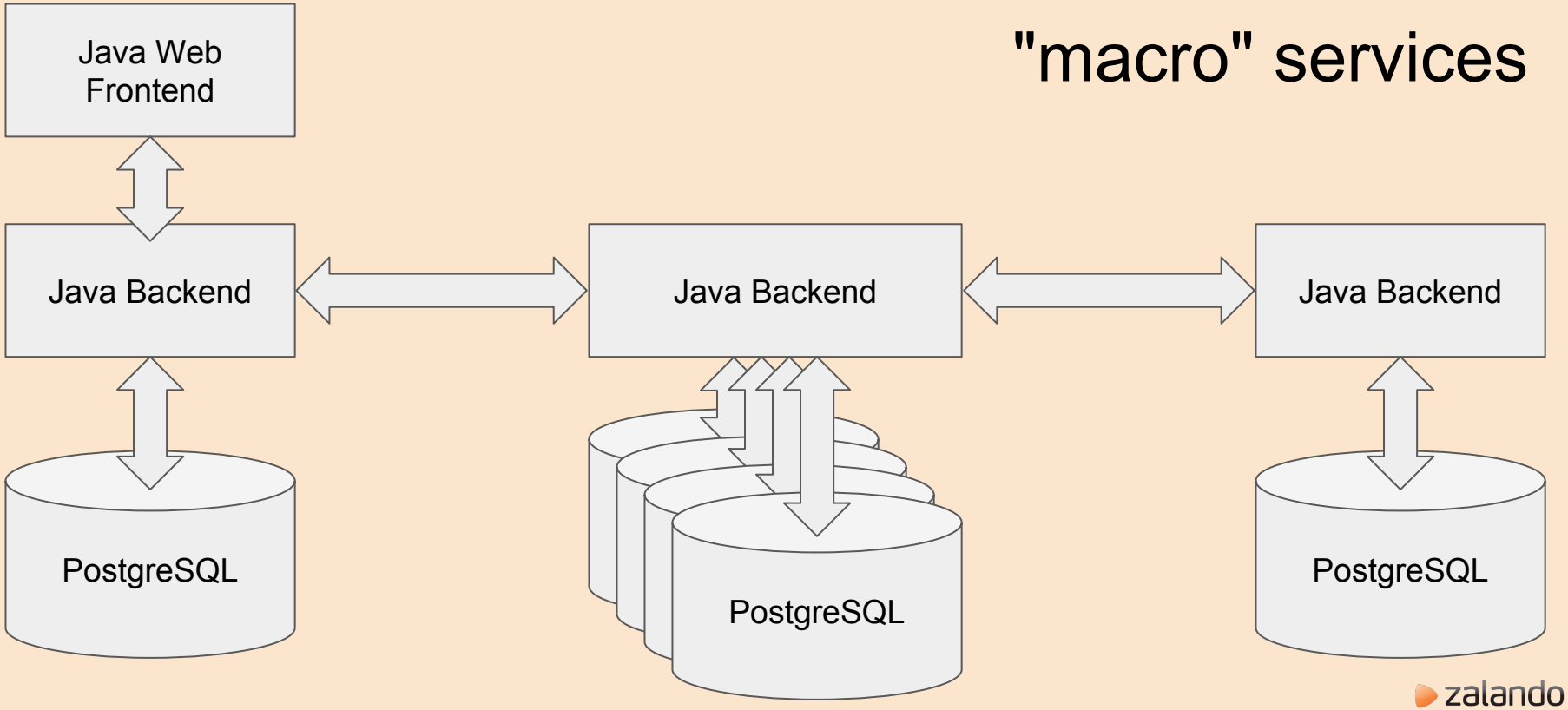
REBOOT

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



"macro" services

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

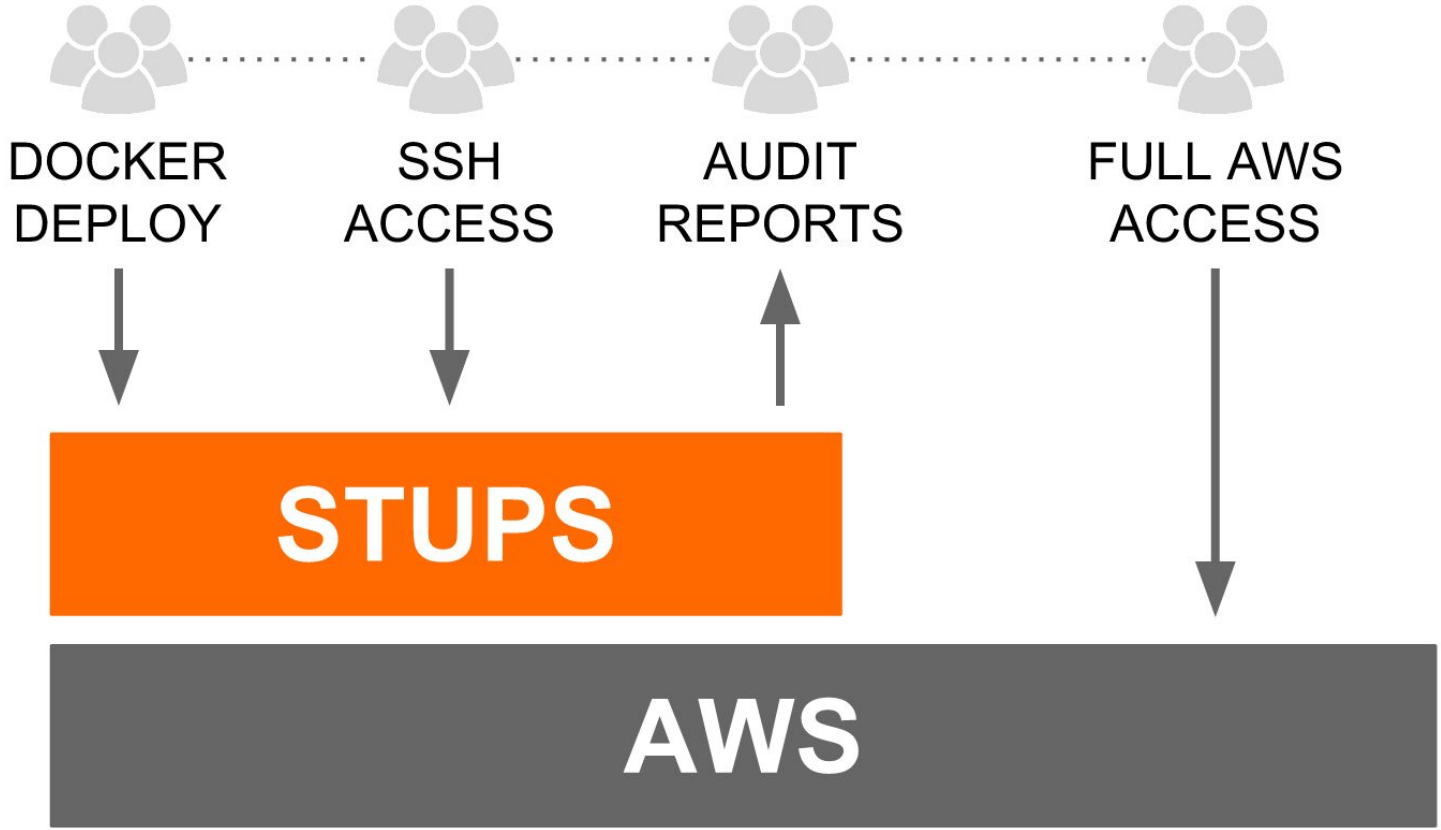
Autonomy

Autonomous teams

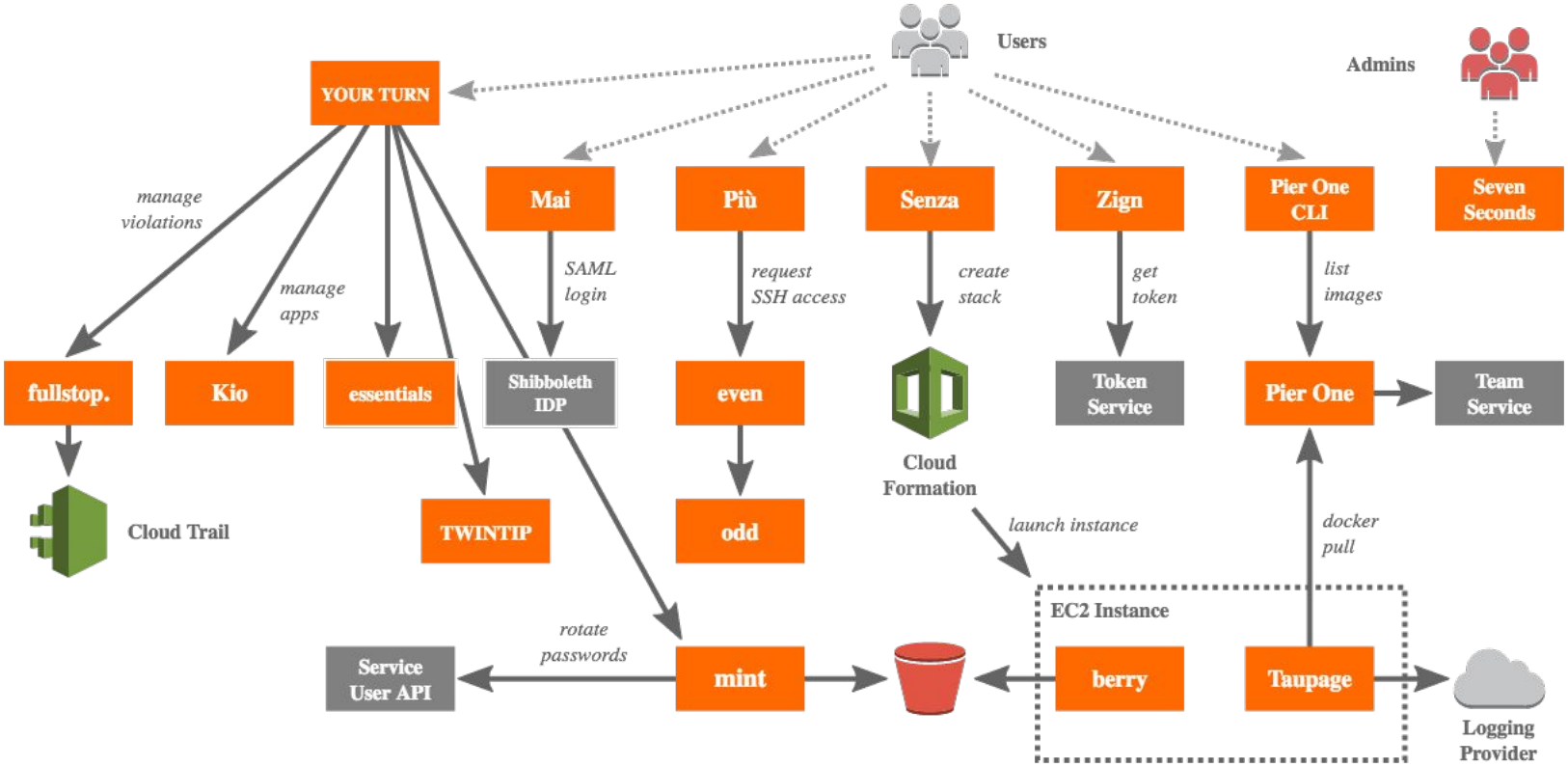
- can choose own technology stack
- including persistence layer
- are responsible for operations
- should use isolated AWS accounts

Autonomy is not an Anarchy

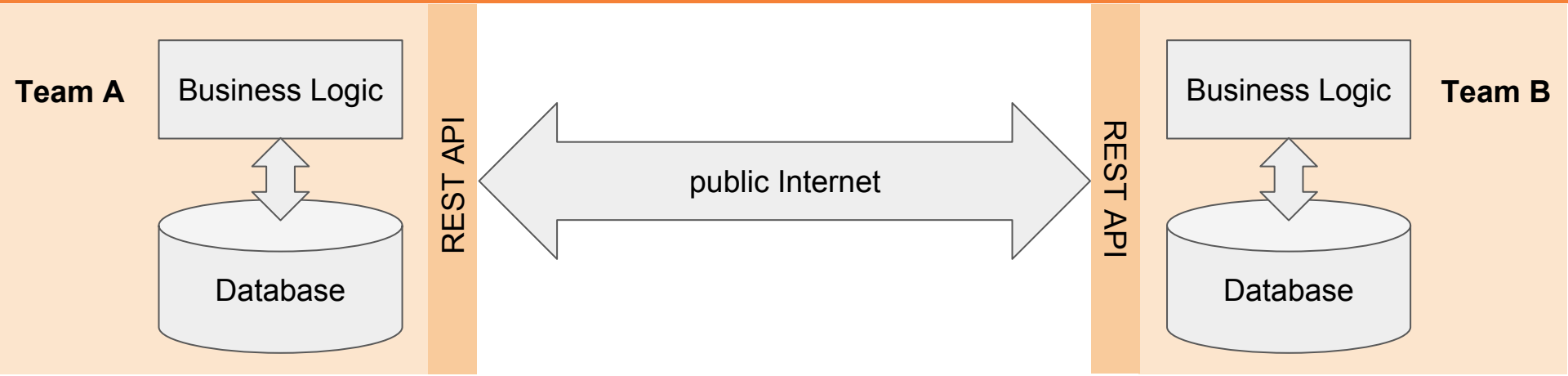
Supporting autonomy — STUPS



Supporting autonomy — STUPS

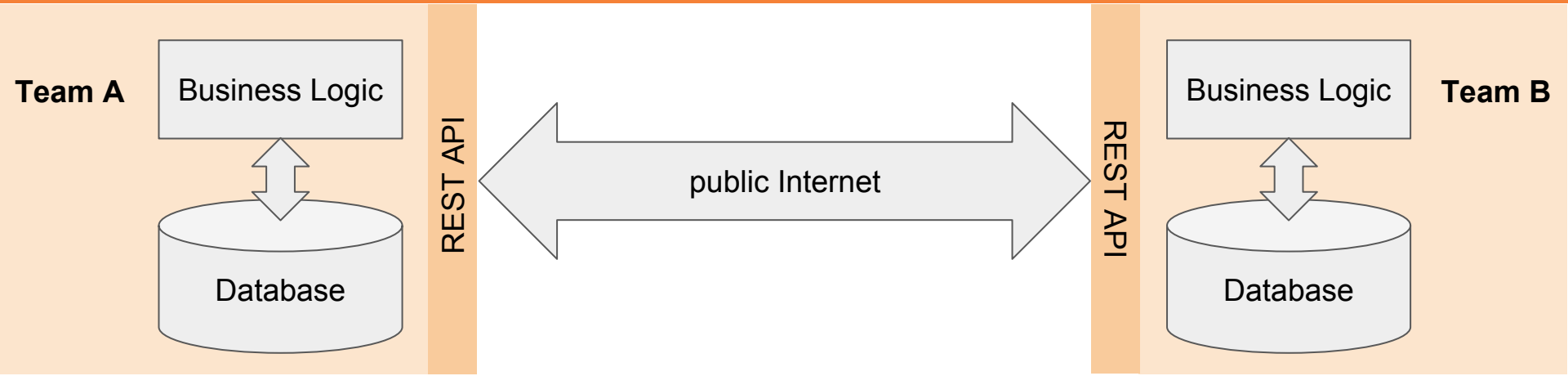


Supporting autonomy — Microservices



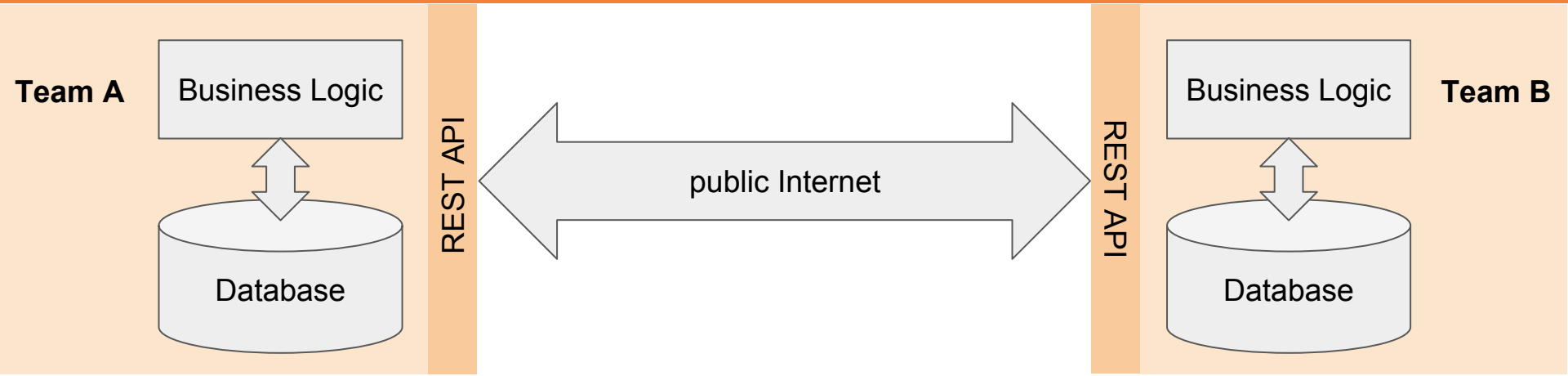
- Applications communicate using REST APIs
- Databases hidden behind the walls of AWS VPC

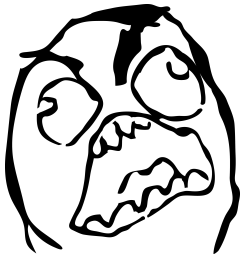
Supporting autonomy — Microservices



- Database team is consulting autonomous teams
- Spilo as STUPS service for PostgreSQL

Supporting autonomy — Microservices



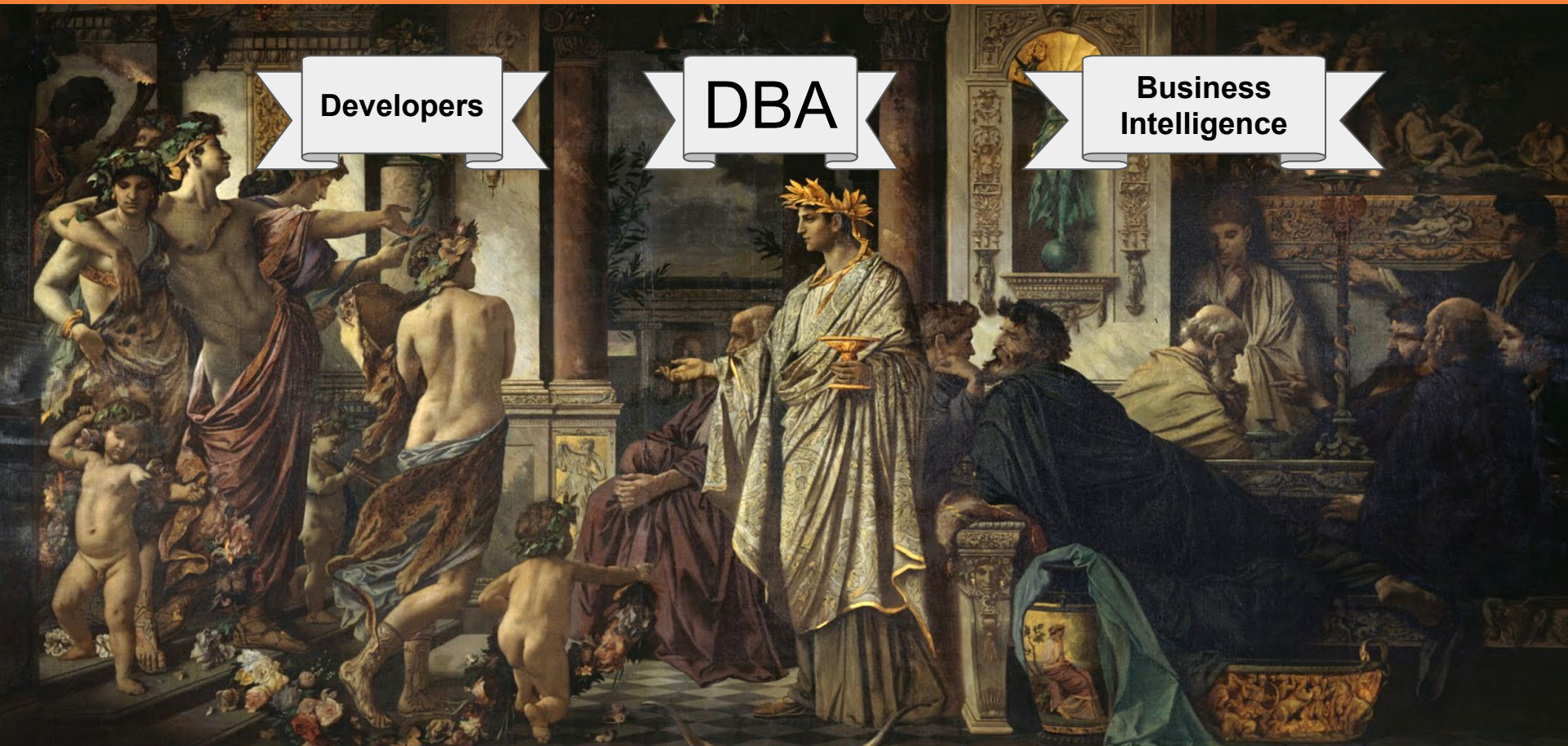


Classical ETL process is impossible!

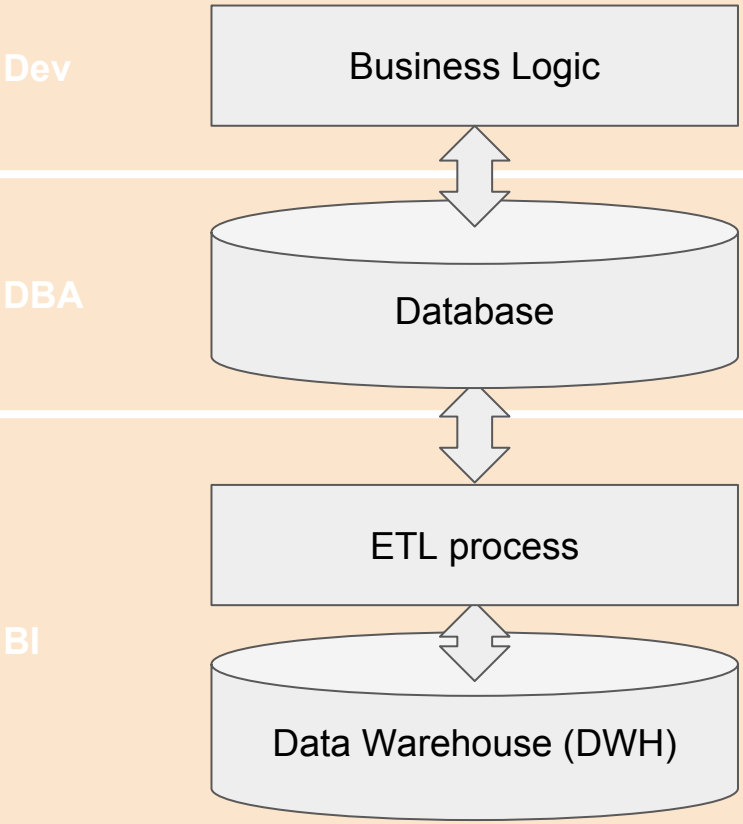
Data integration in the classical world

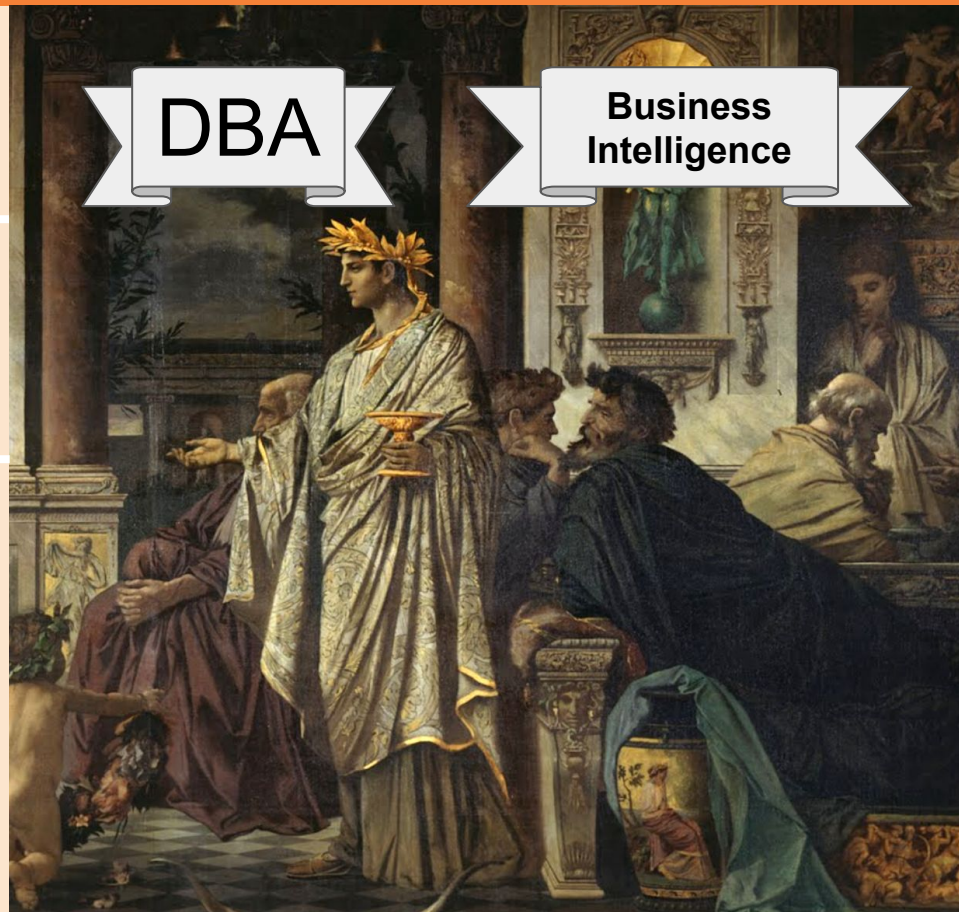
Data integration in the classical world

Data integration in the classical world

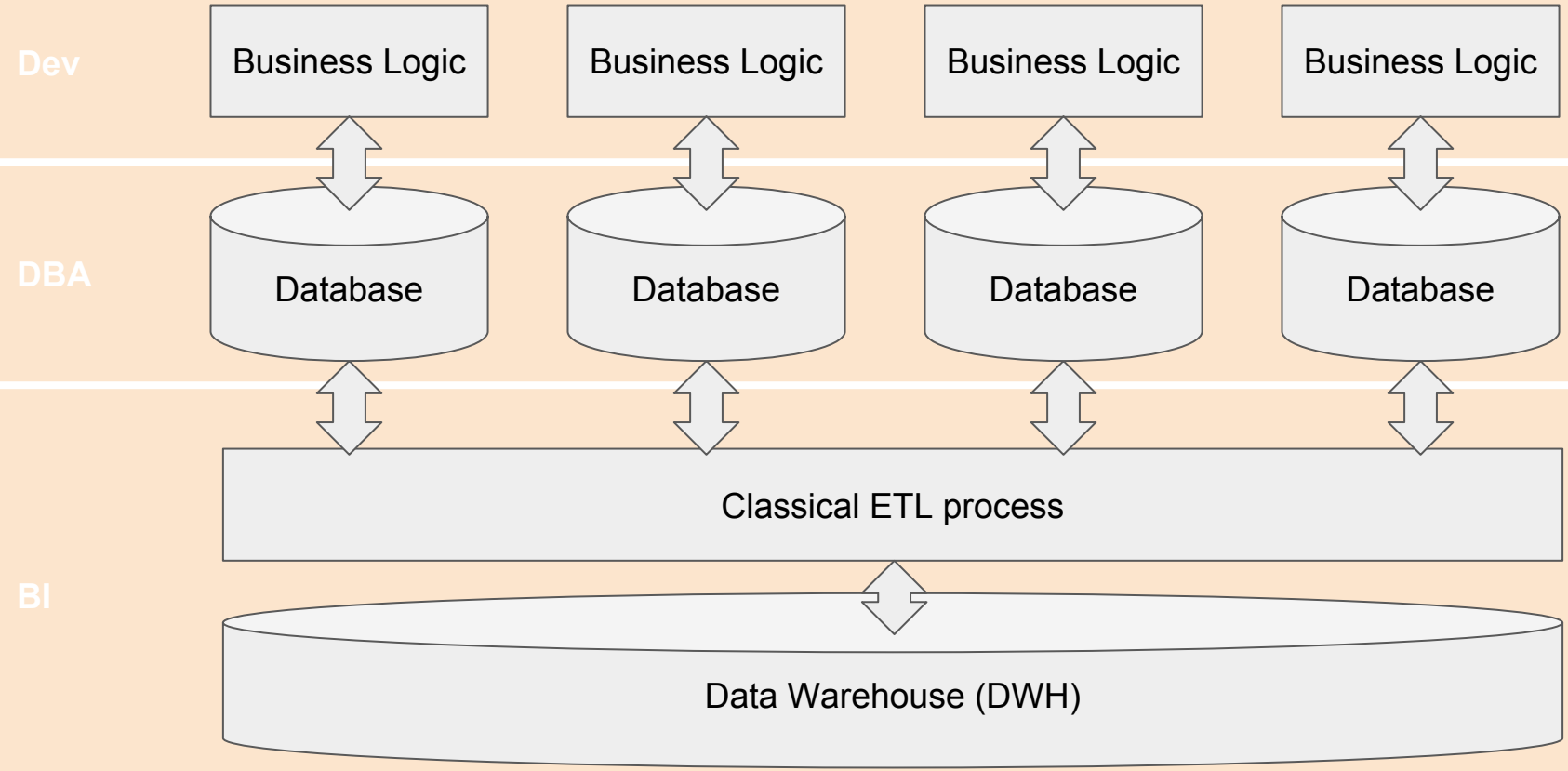


Data integration in the classical world





Data integration in the classical world



Data integration in the classical world

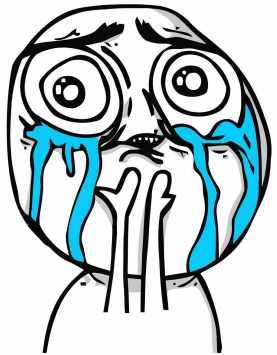
Classical ETL process

- Use-case specific — a lot of manual work

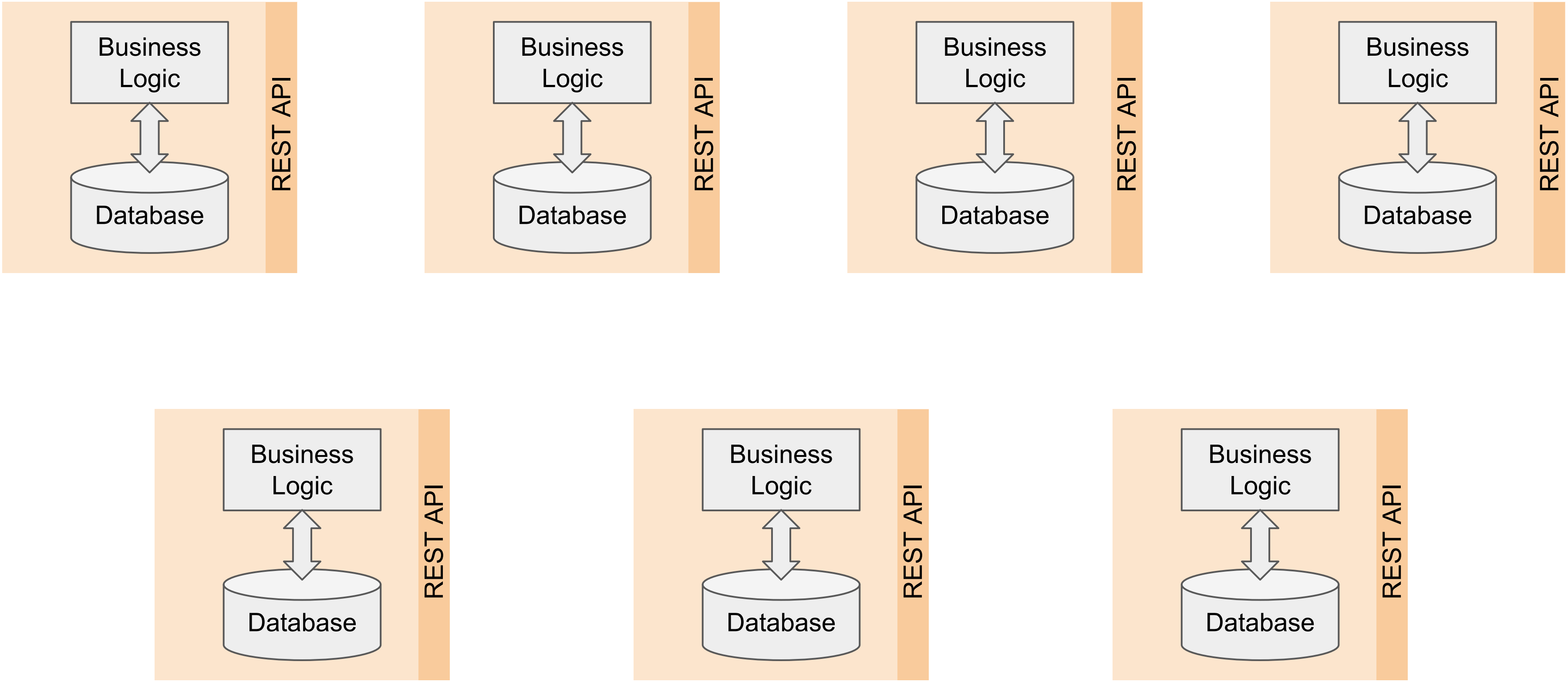
- Usually outputs data into a Data Warehouse
  - well structured
  - easy to use by the end user (SQL)

Data integration in the world of microservices

Data integration in the world of microservices

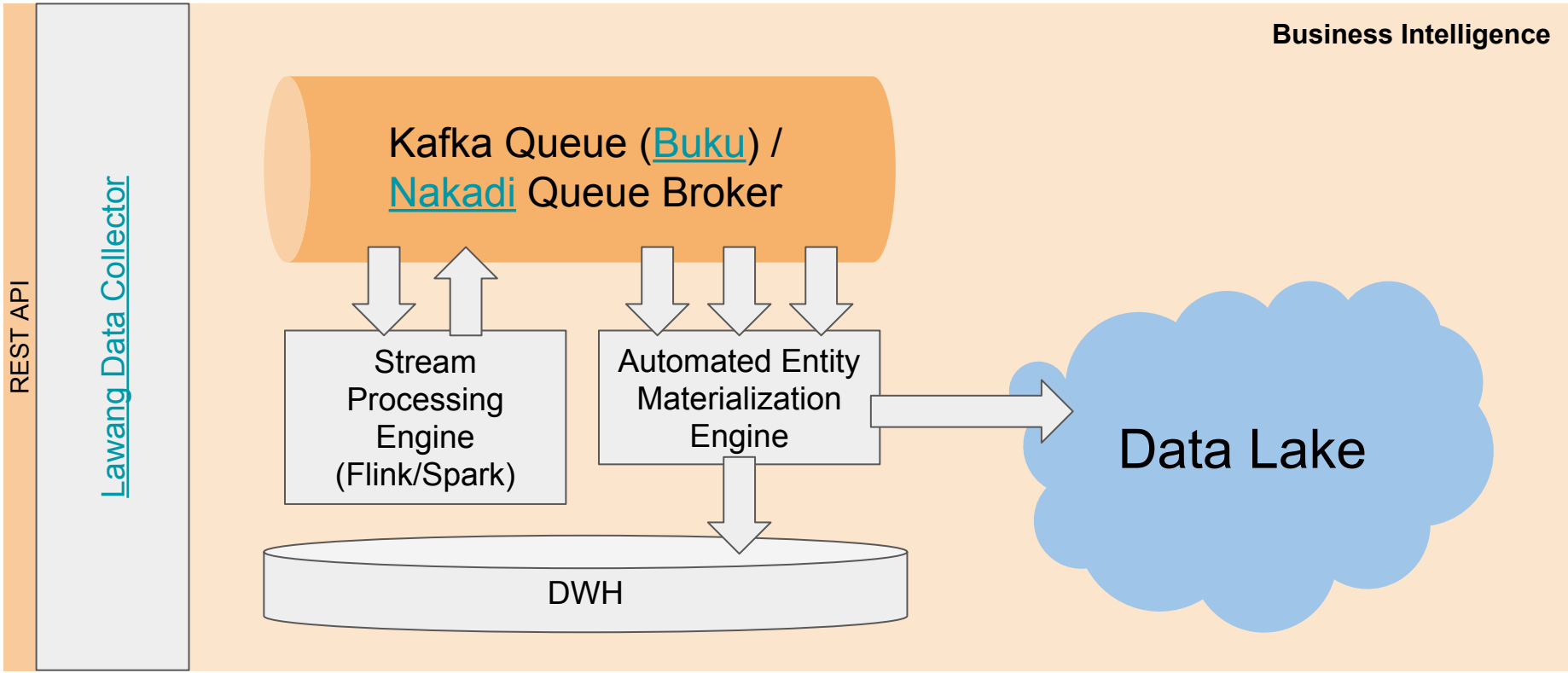


Data integration in the world of microservices

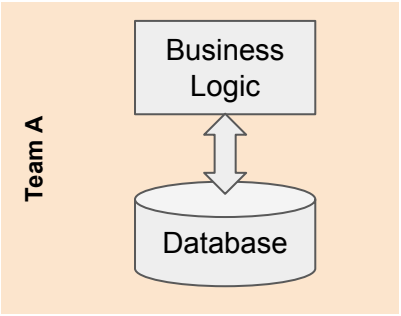


zalando

Data integration in the world of microservices



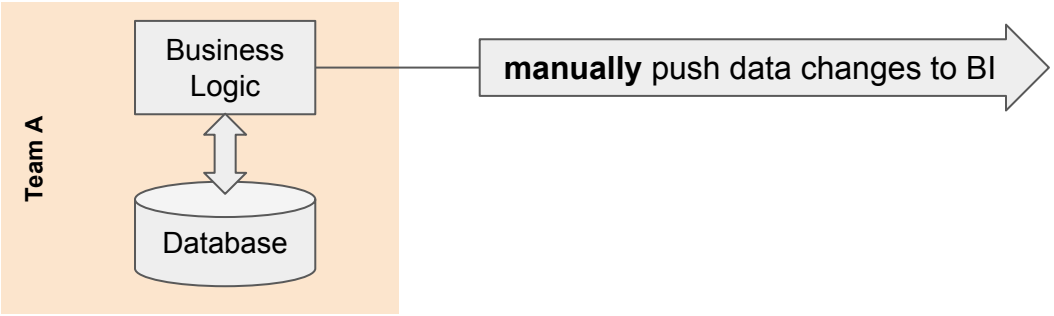
Data integration in the world of microservices



REST API

Business Intelligence

Data integration in the world of microservices

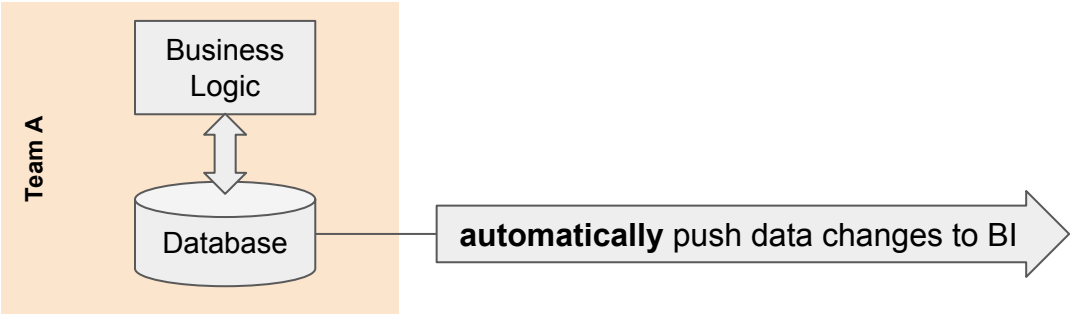


- Error prone
- 2PC is actually needed
- Very difficult to implement

REST API

Business Intelligence

Data integration in the world of microservices



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

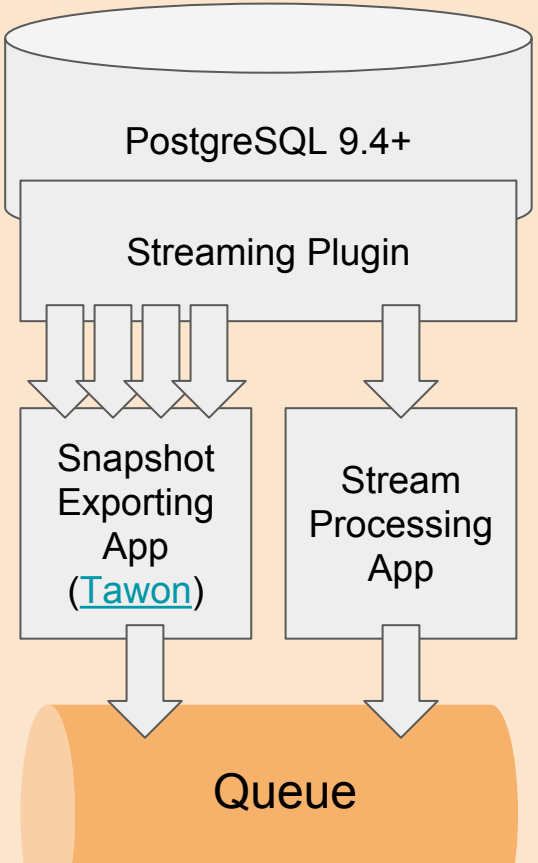
Data integration in the world of microservices

PostgreSQL Logical Replication

Enables automatic Data Change Event Extraction

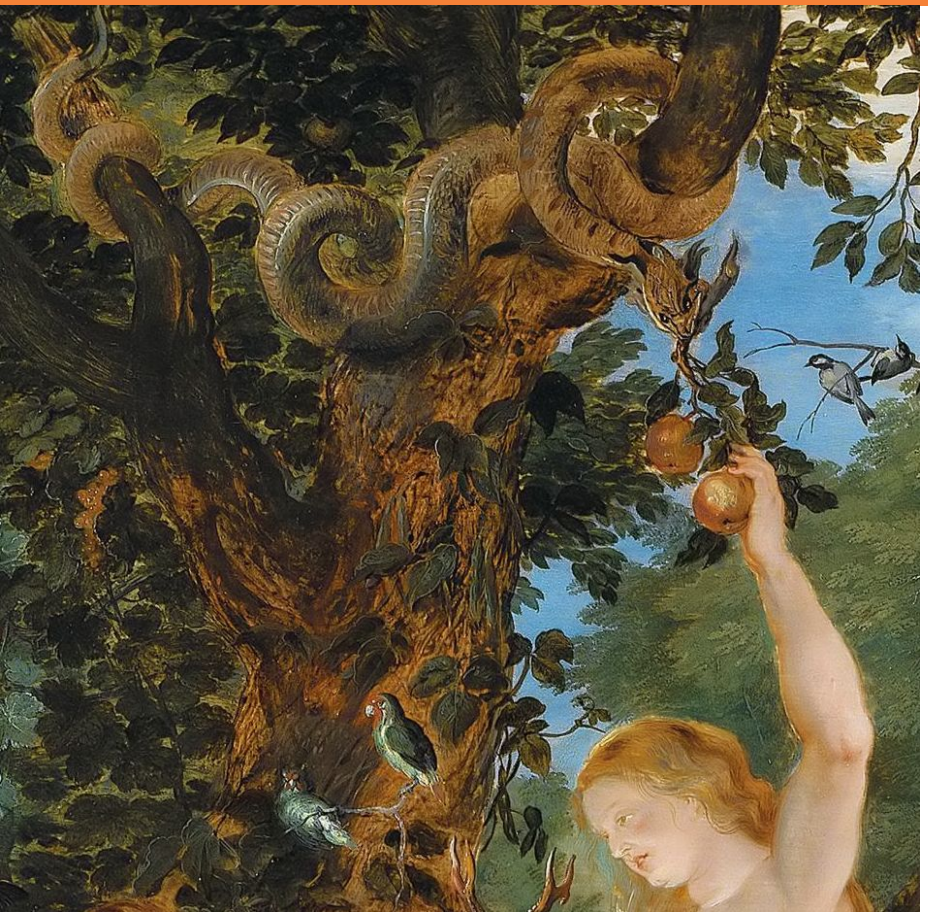
- pglogical by 2nd Quadrant (streaming plugin)
- BottledWater by Confluent (Avro to Kafka streaming)
- Tawon by Zalando (parallel snapshotting)

Data integration in the world of microservices



Try it yourself

Python support for replication protocol



python™

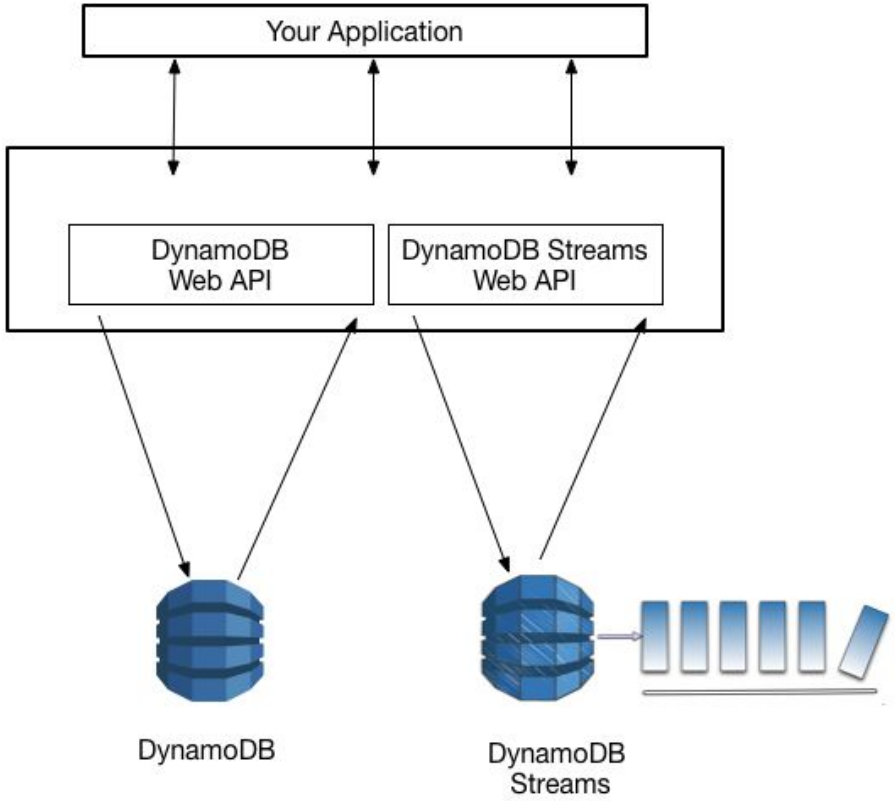
is the best to make you do experiments

- psycopg2 PR for replication protocol
- psycopg2 replication protocol usage
- Tawon initial snapshot export tool
- streaming initial database snapshot

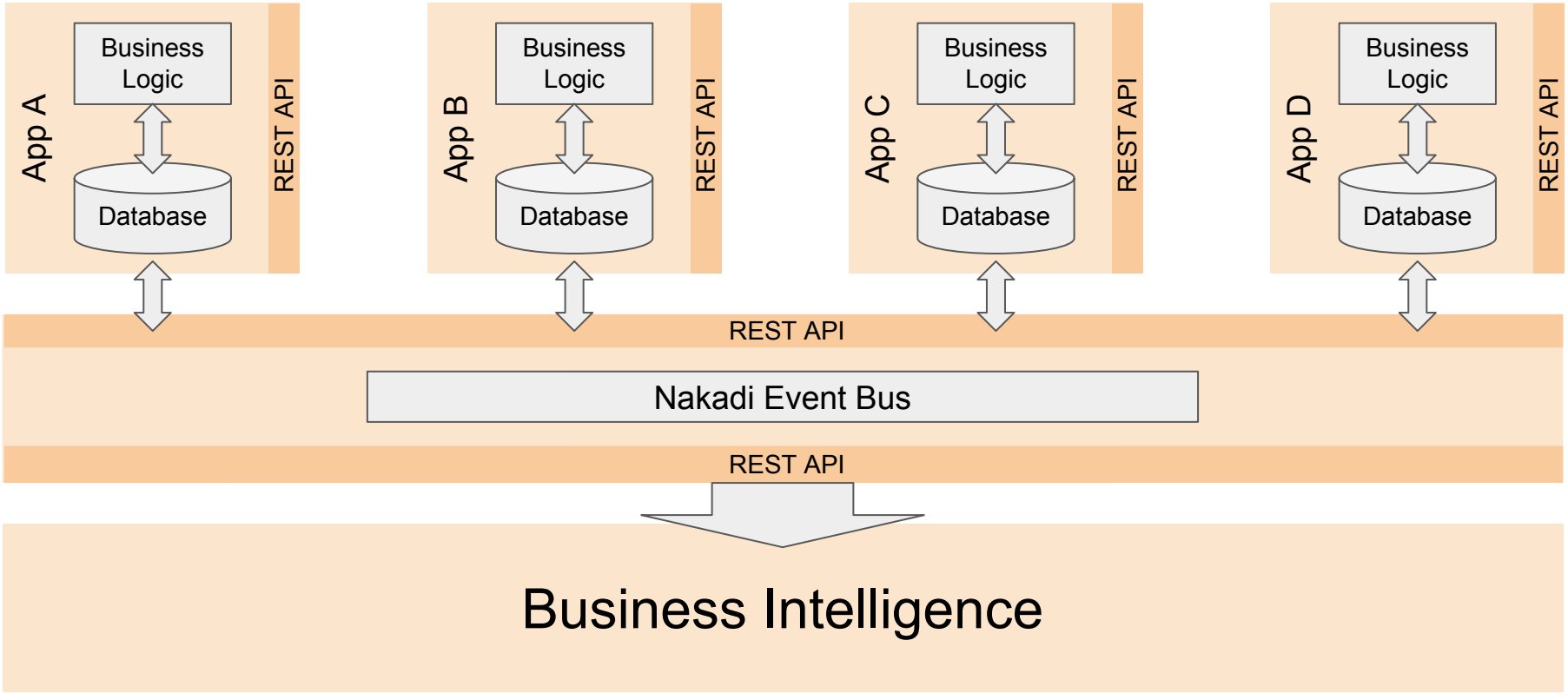
Data integration in the world of microservices

DynamoDB Streams

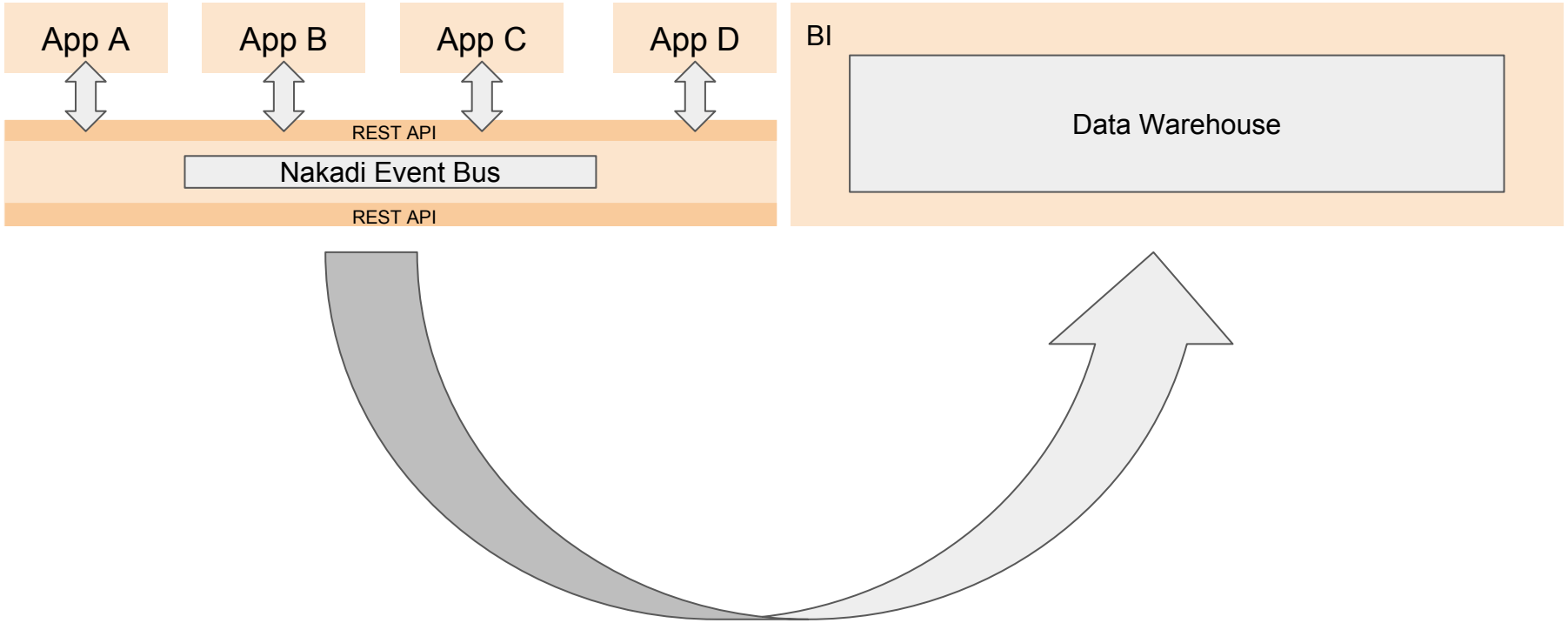
Enables automatic Data Change Event Extraction



Data integration in the world of microservices

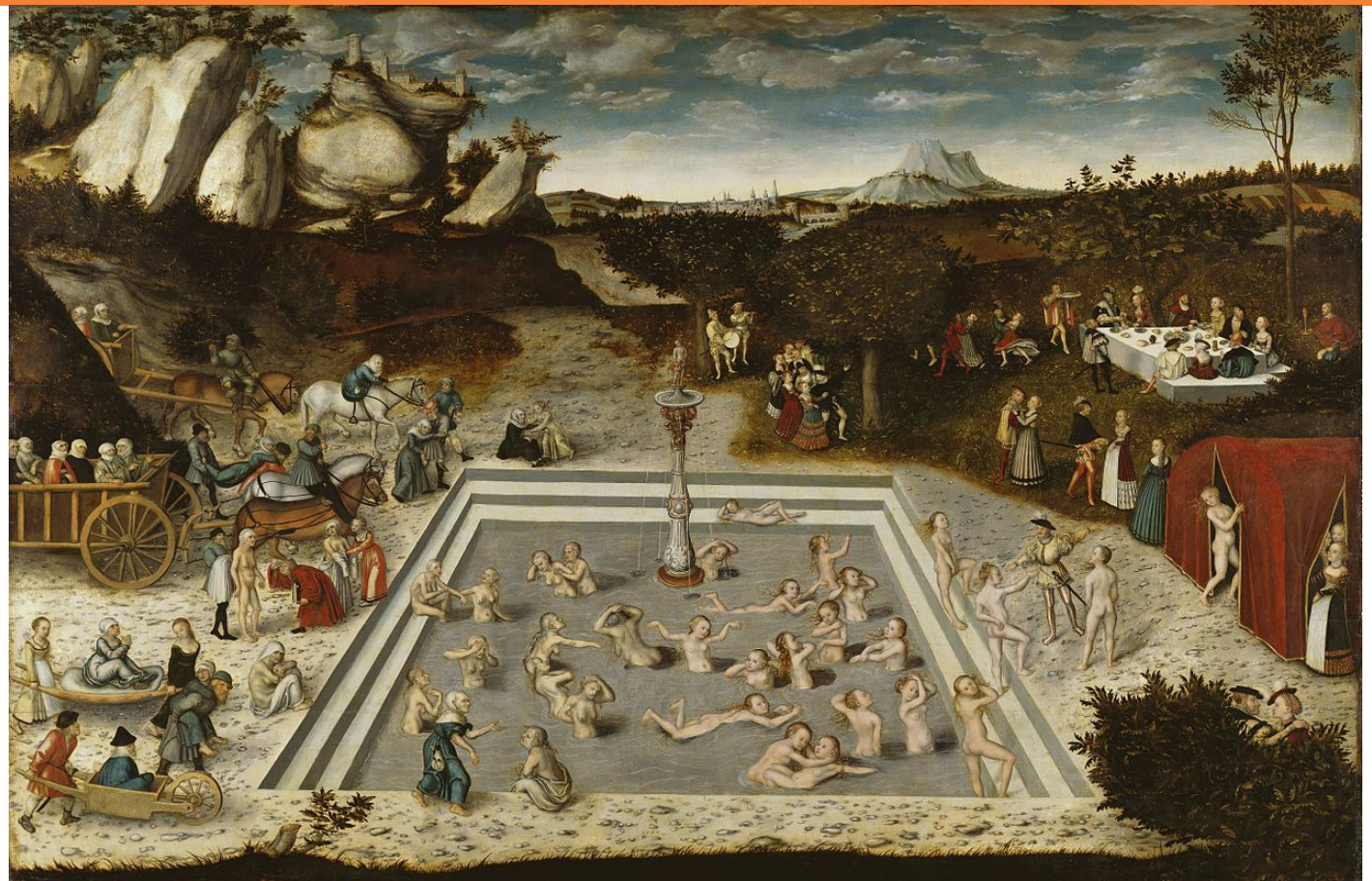


Data integration in the world of microservices



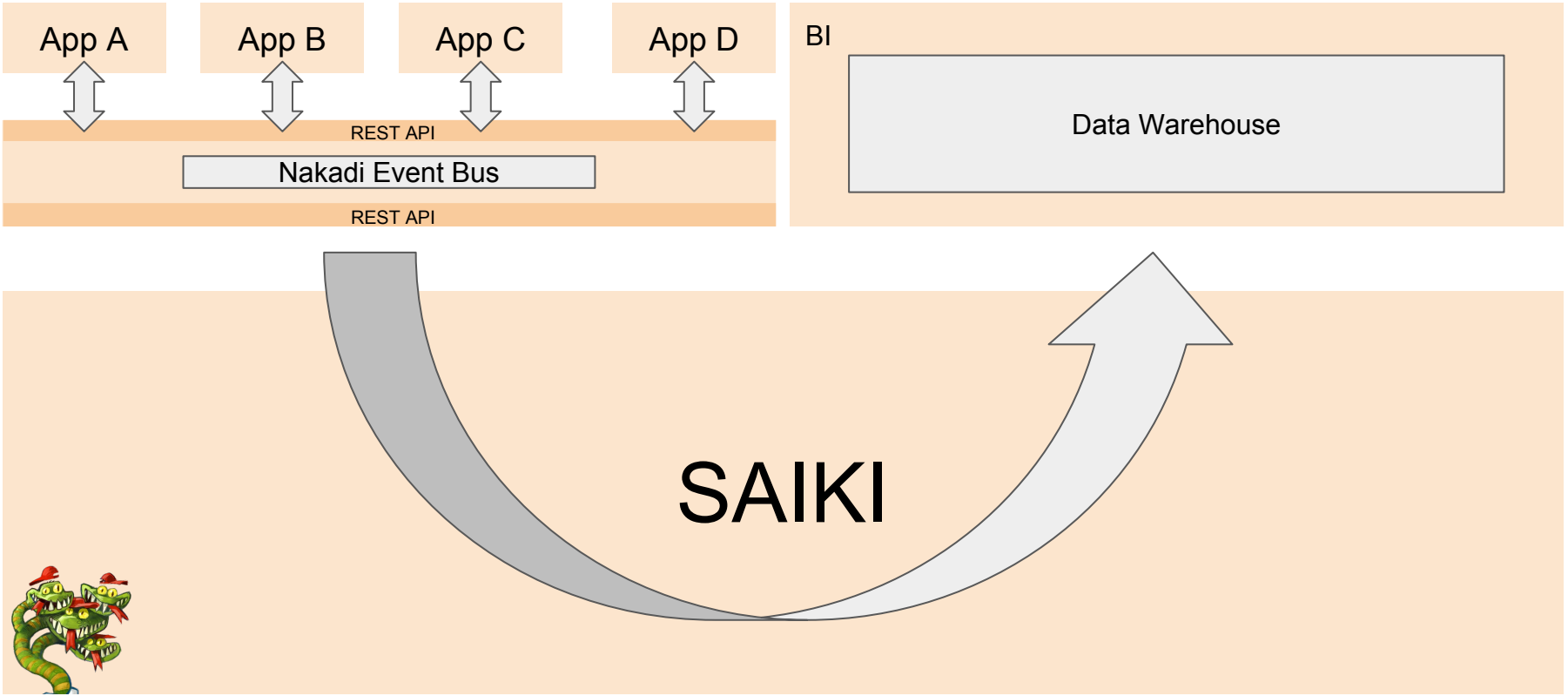
SAIKI
Data Platform

Saiki Data Jungbrunnen (1546)

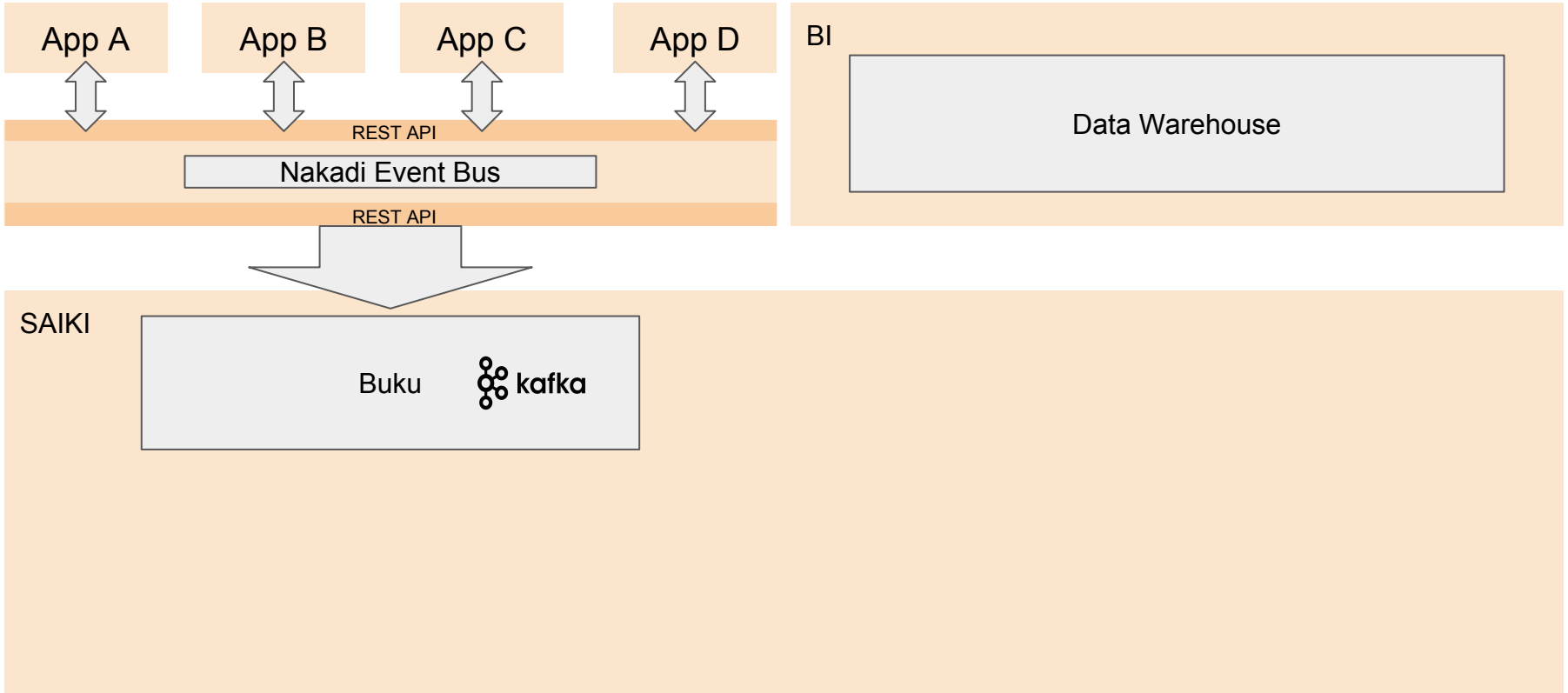


zalando

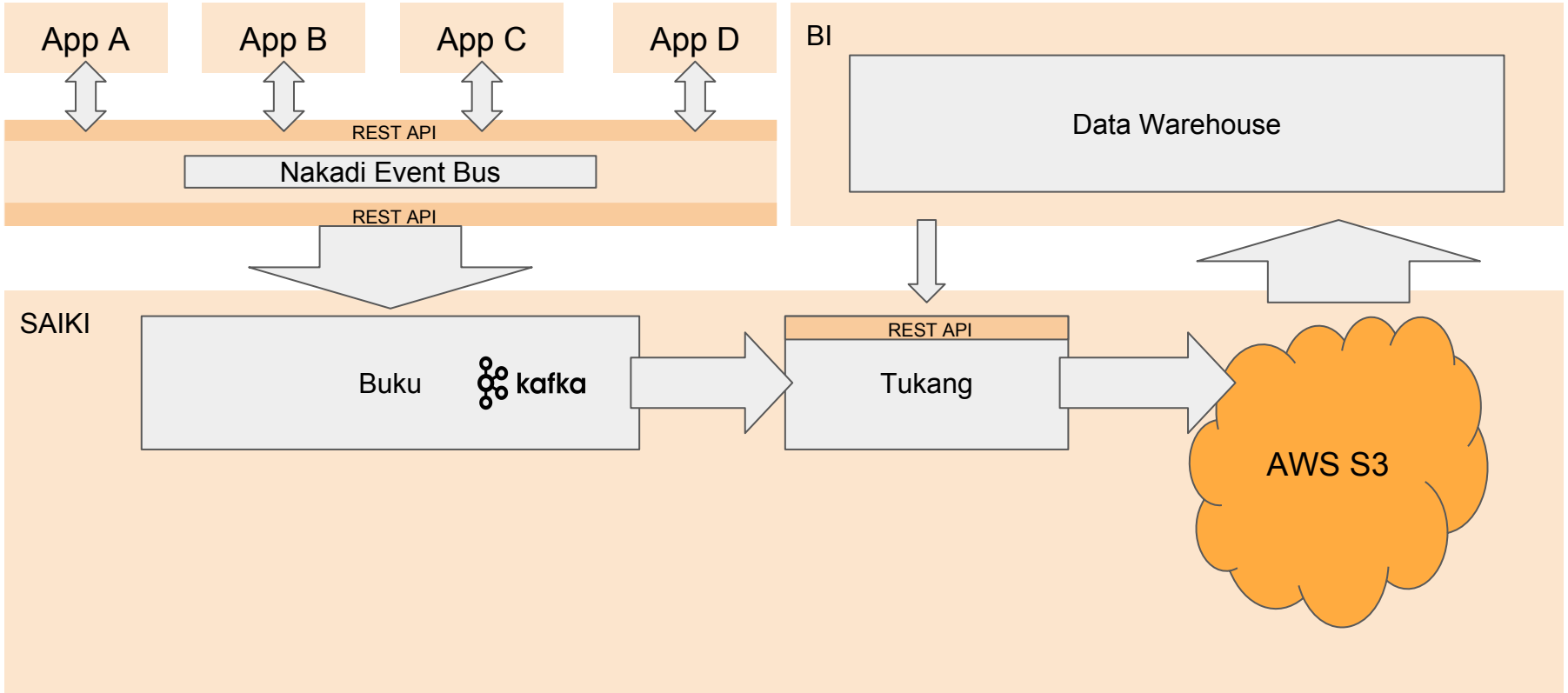
Saiki Data Platform



Saiki Data Platform



Saiki Data Platform

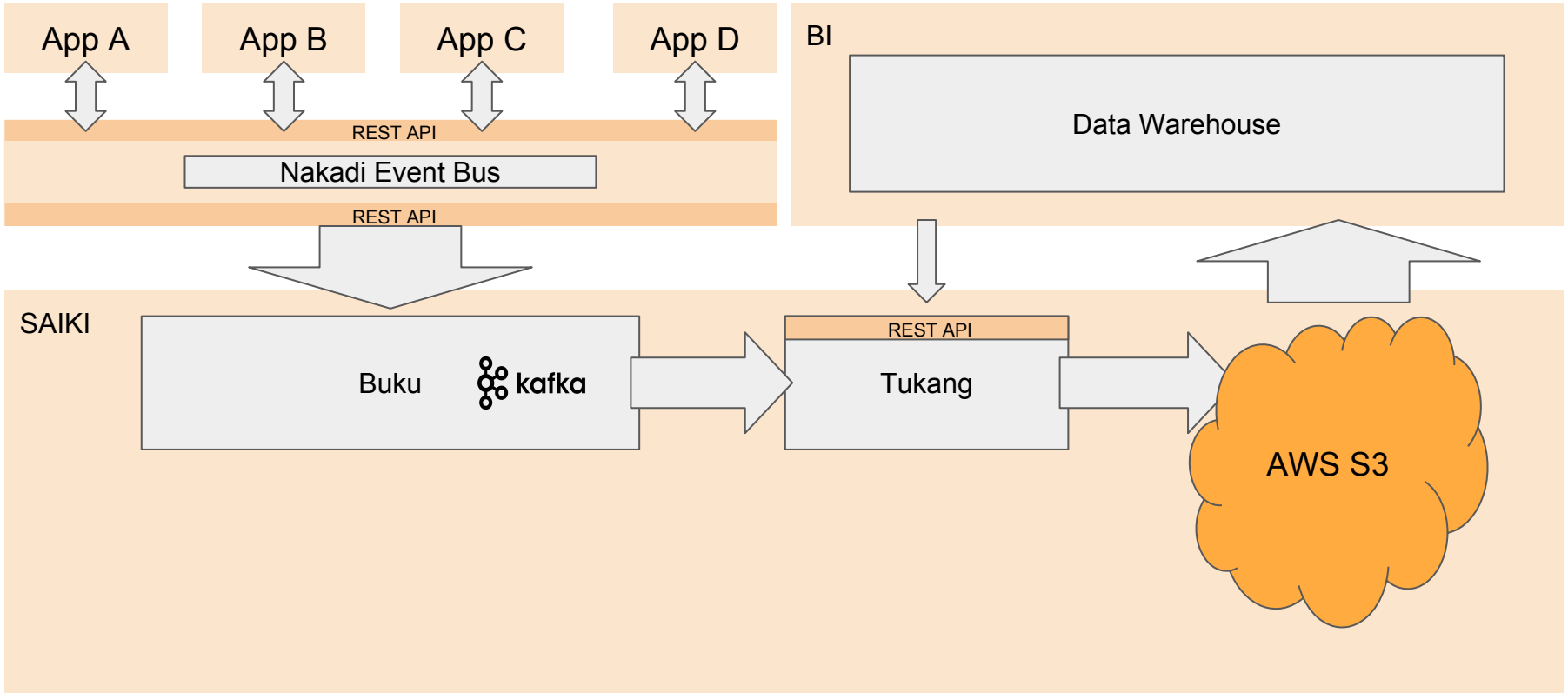


Saiki Data Platform

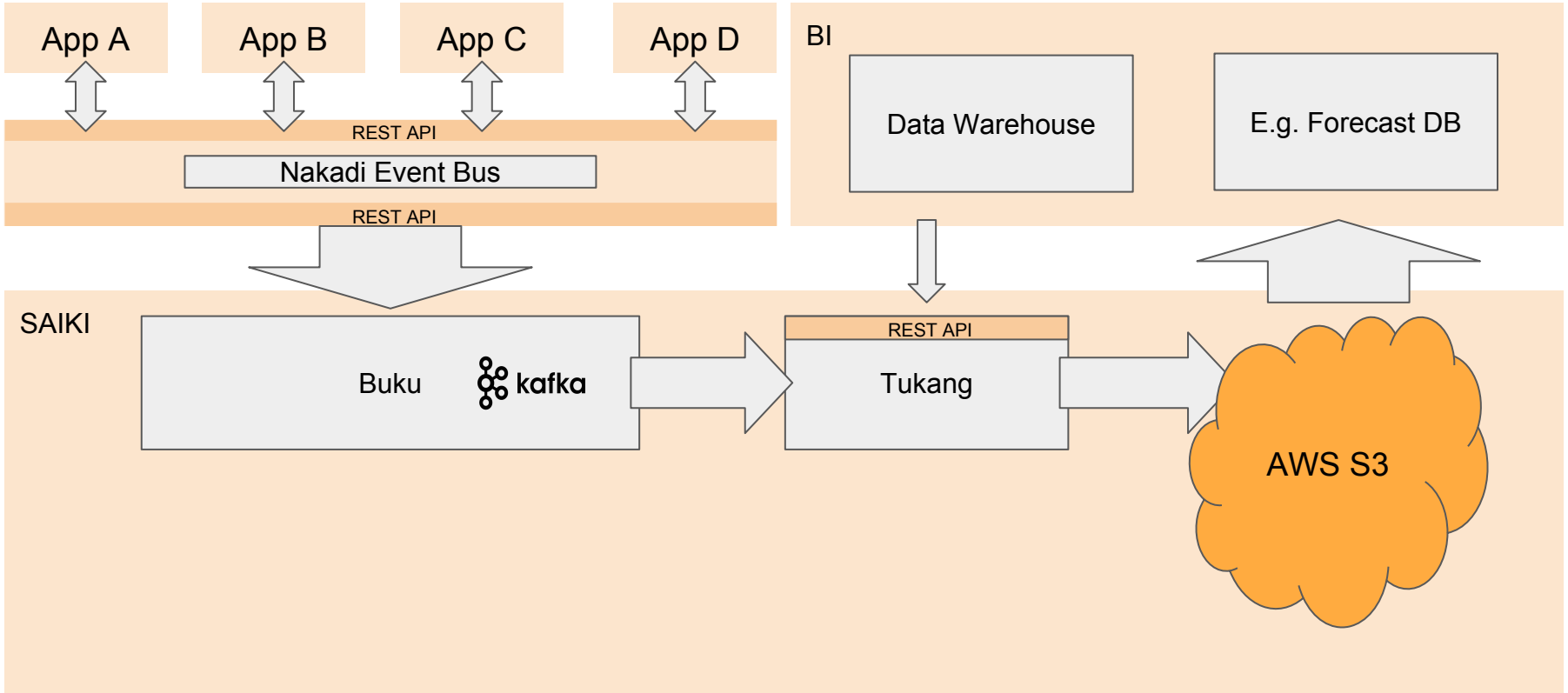
Saiki Tukang

- First cleansing of events (out of order, duplicates, etc.)
- Materialize data from Kafka in AWS S3
- Provide metadata via RESTful interface
- DWH downloads data directly from cloud storage

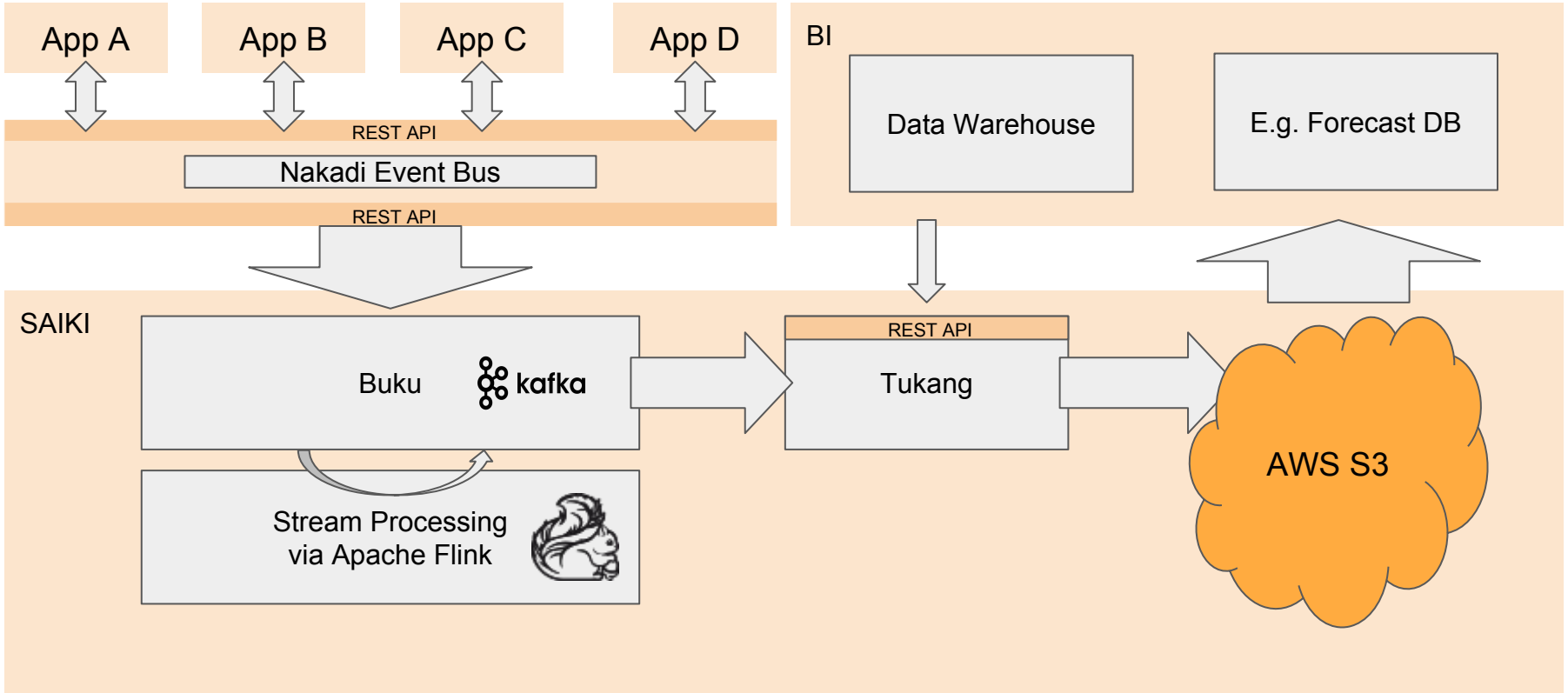
Saiki Data Platform



Saiki Data Platform



Saiki Data Platform



Saiki Data Platform

Apache Flink

- true stream processing framework
- process events at a consistently high rate with relatively low latency
- scalable
- support from Berlin/Europe

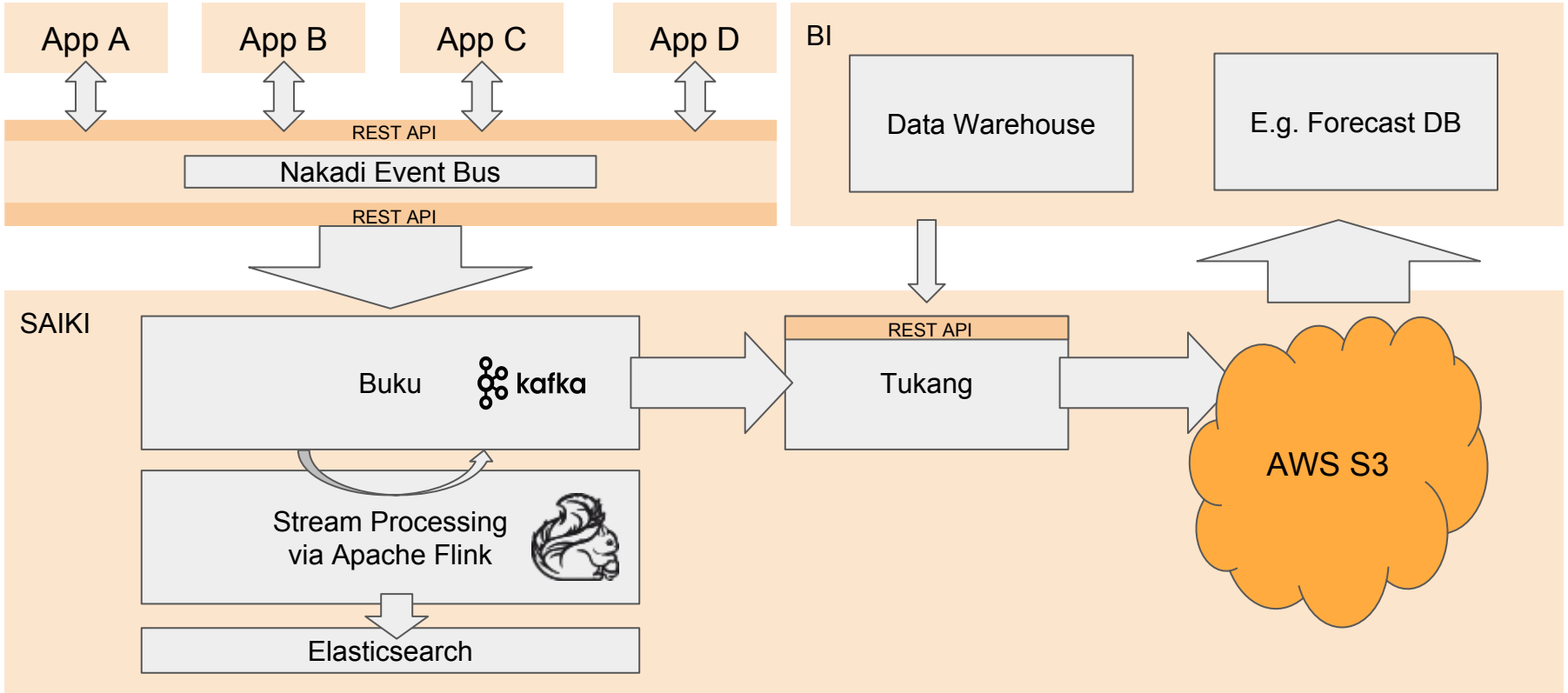
https://tech.zalando.com/blog/apache-showdown-flink-vs.-spark/

Saiki Data Platform

Apache Flink

- connectors
  - Kafka
  - Elasticsearch
  - etc.

Saiki Data Platform

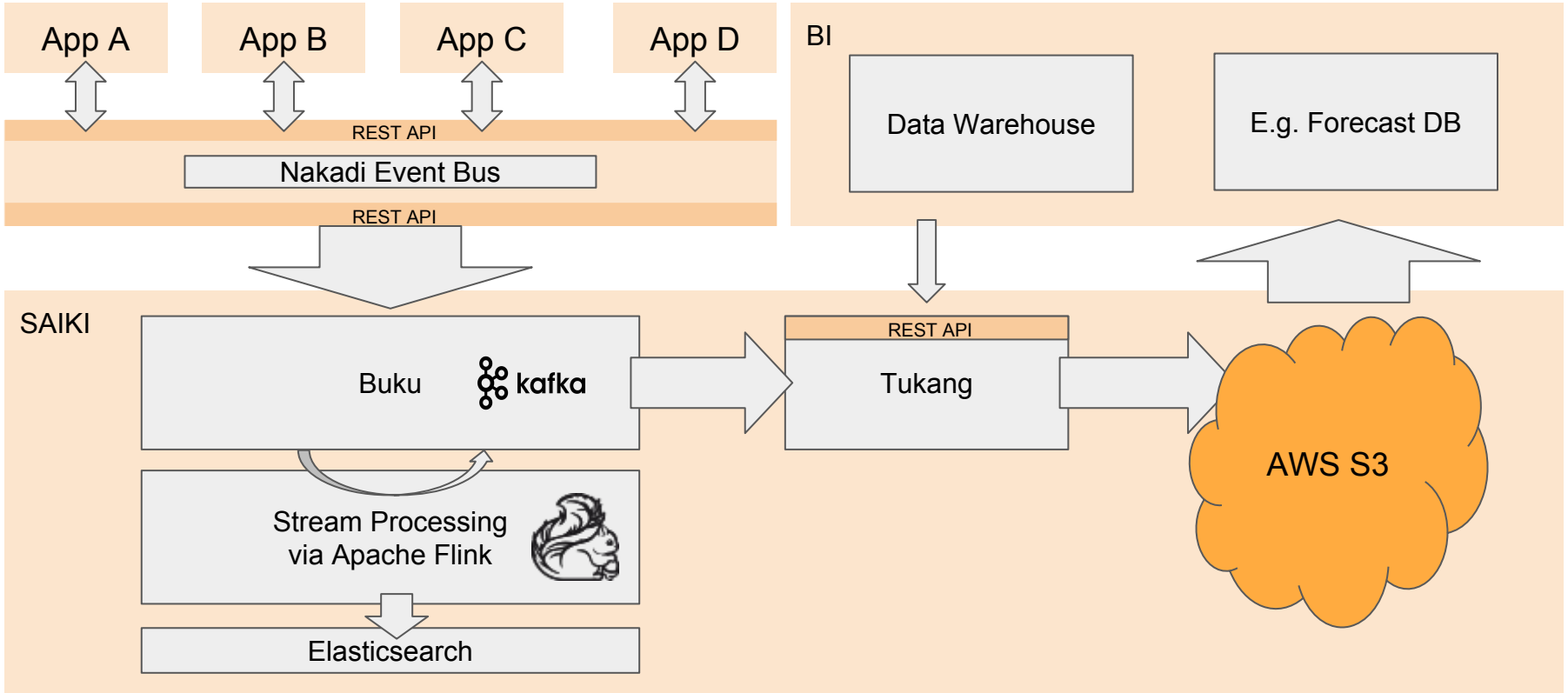


Saiki Data Platform

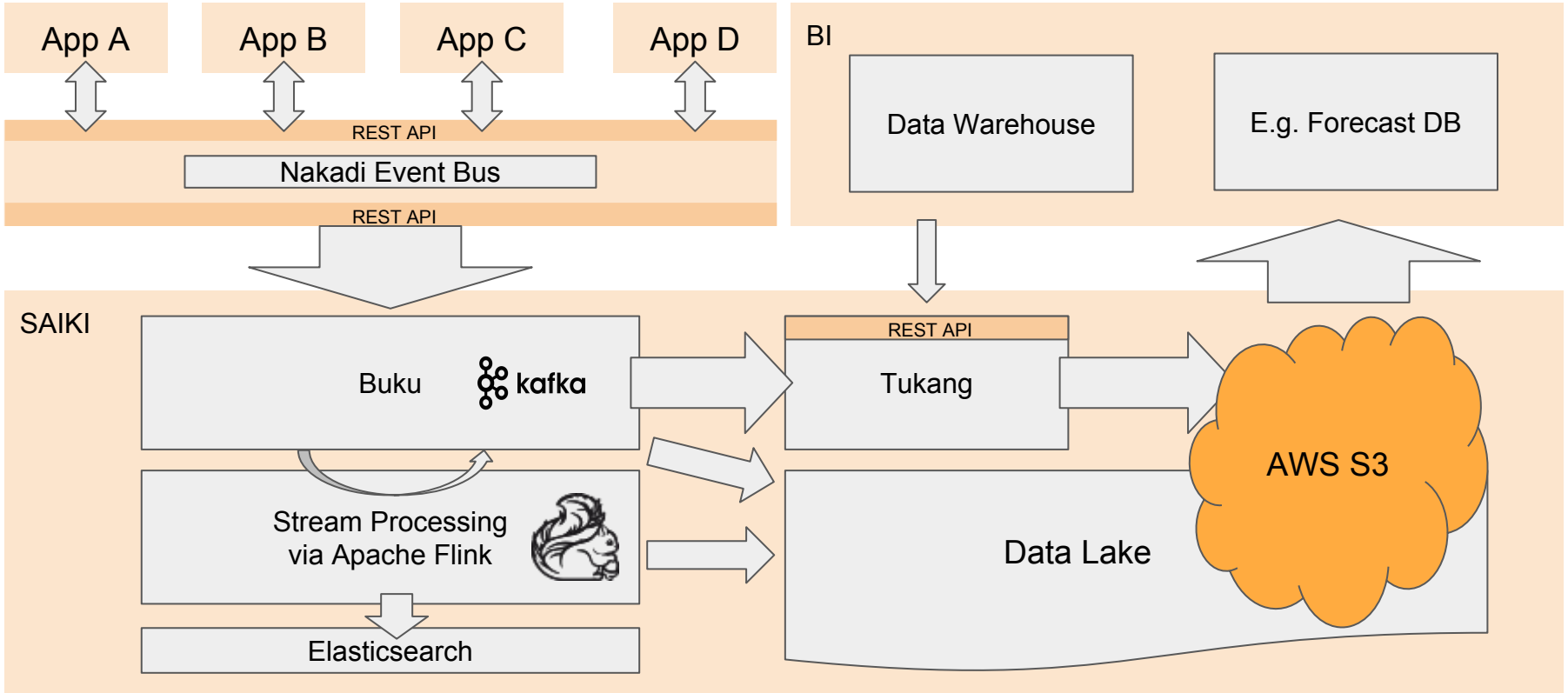
For example: Real-time Business Process Monitoring

- Check if business processes work as expected
- Analyze data on the fly
- Visualization with Python/Flask and Chart Frameworks

Saiki Data Platform



Saiki Data Platform



Open Source @ZalandoTech

Open Source @ZalandoTech

- https://tech.zalando.com - Technology Blog
- https://zalando.github.io - Open Source Projects
  - PostgreSQL Open Source Tools
  - STUPS.io for responsible organisation using AWS
  - **Saiki** Data Integration Platform projects
  - REST API on Swagger (OpenAPI)
    - https://github.com/zalando/restful-api-guidelines
    - https://github.com/zalando/connexion
      - To create REST servers on Python

Questions?