DATA INTEGRATION

in the world of microservices

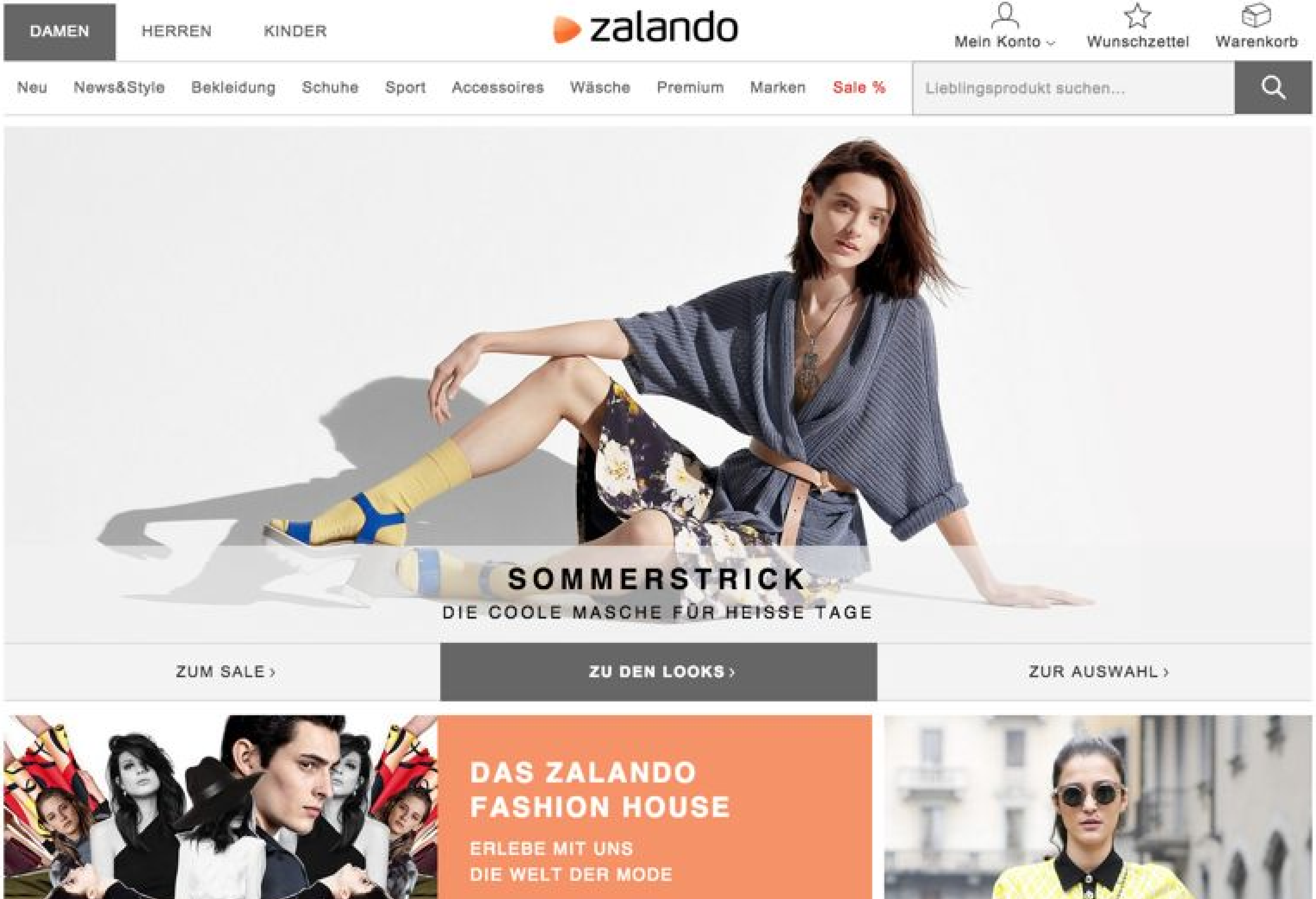
About us

Valentine Gogichashvili

Head of Data Engineering @ZalandoTech
twitter: @valgog
google+: +valgog
email: valentine.gogichashvili@zalando.de

Fabian Wollert

Data Engineer Business Intelligence
github: @drummerwolli
email: fabian.wollert@zalando.de



SOMMERSTRICK

DIE COOLE MASCHE FÜR HEISSE TAGE

DAS ZALANDO FASHION HOUSE

ERLEBE MIT UNS DIE WELT DER MODE

Zalando Technology

BERLIN

DORTMUND

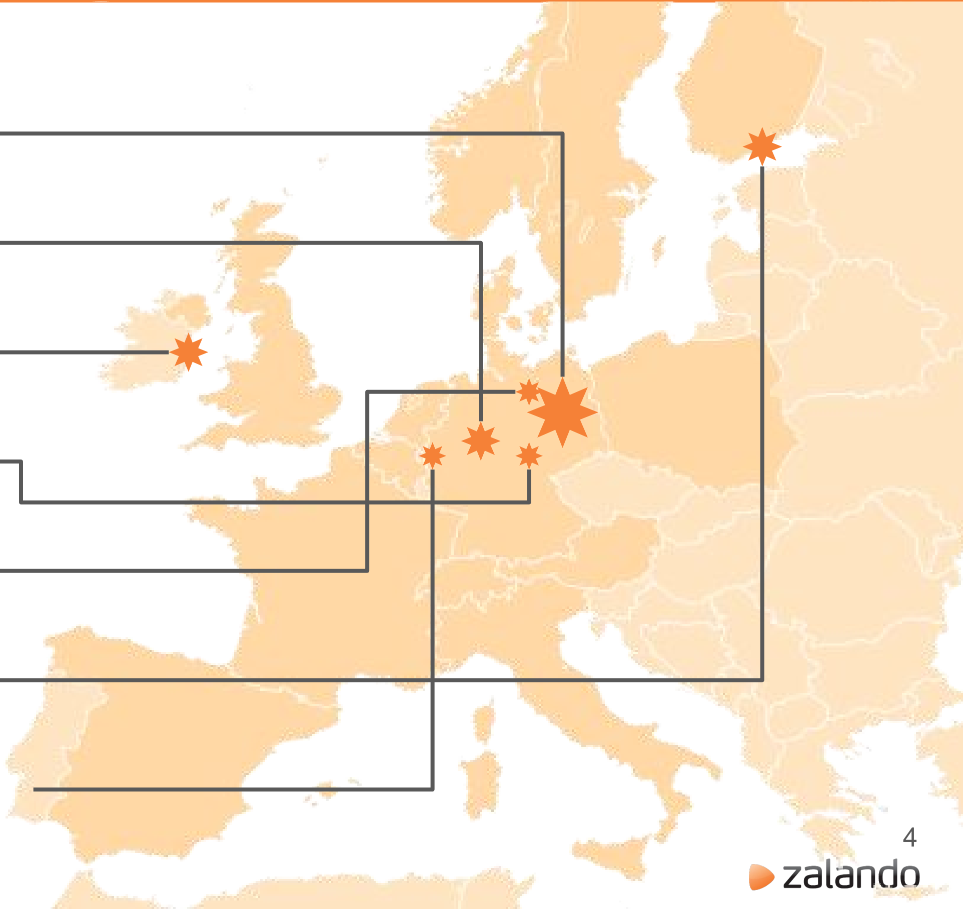
DUBLIN

ERFURT

HAMBURG

HELSINKI

MÖNCHENGLADBACH



Zalando Technology

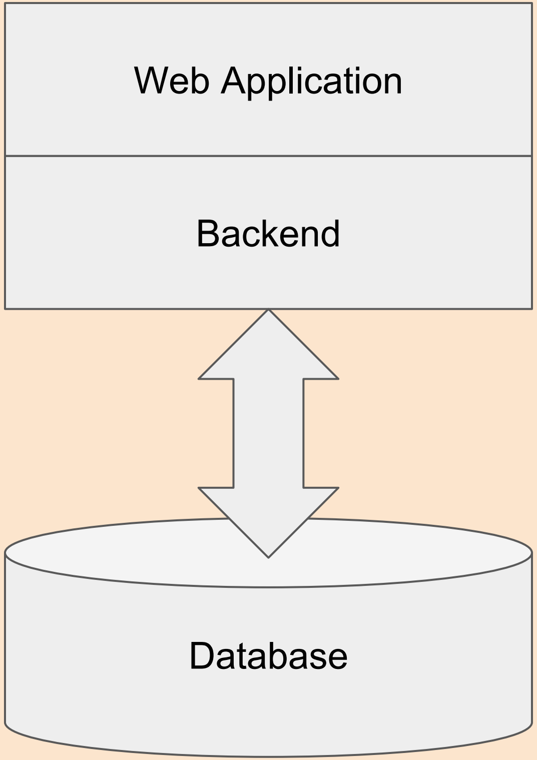
1100+ TECHNOLOGISTS

Rapidly growing international team

http://**tech**.zalando.com

Good old small world

Once upon a time...



Started as a tiny online shop

Prototyped on Magento (PHP)

Used MySQL as a database

REBOOT

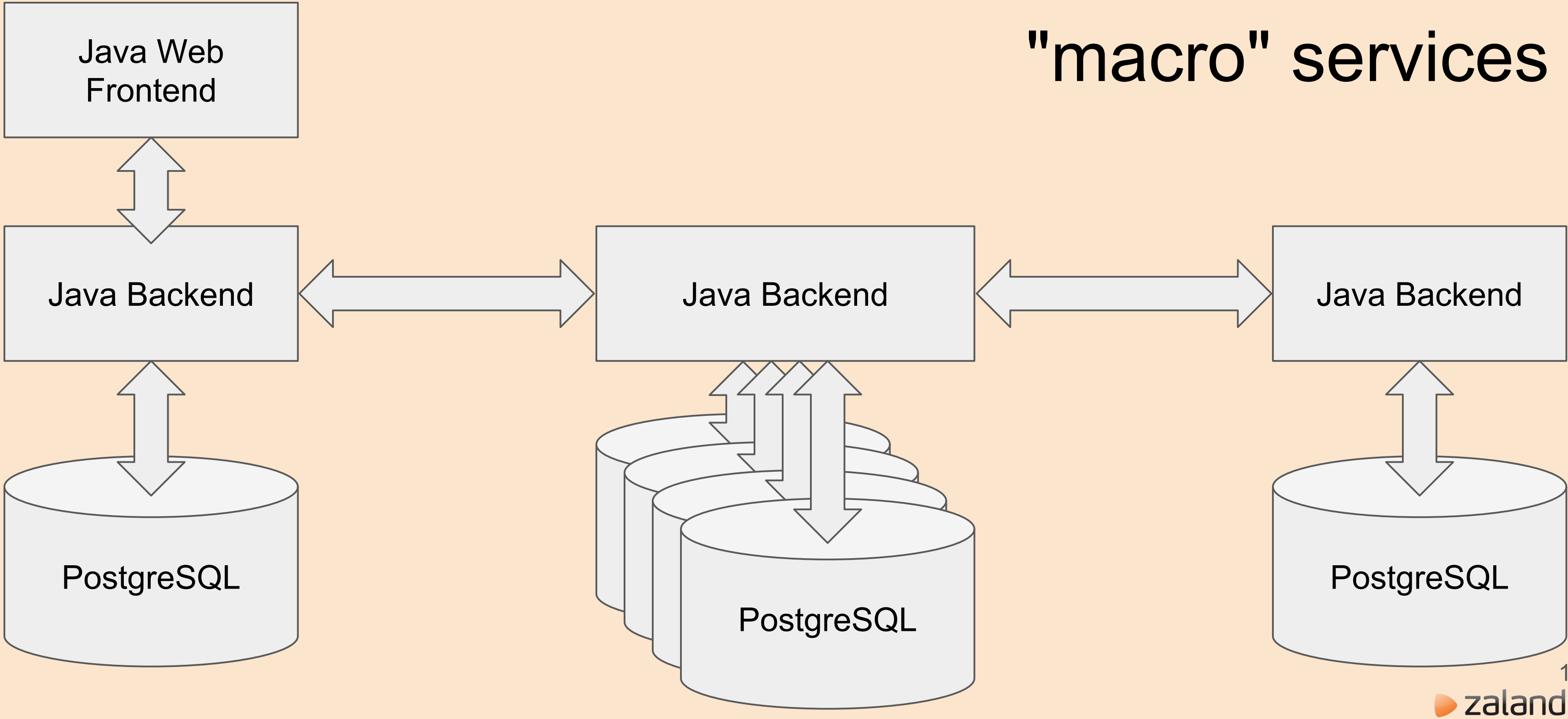
REBOOT

5½ years ago

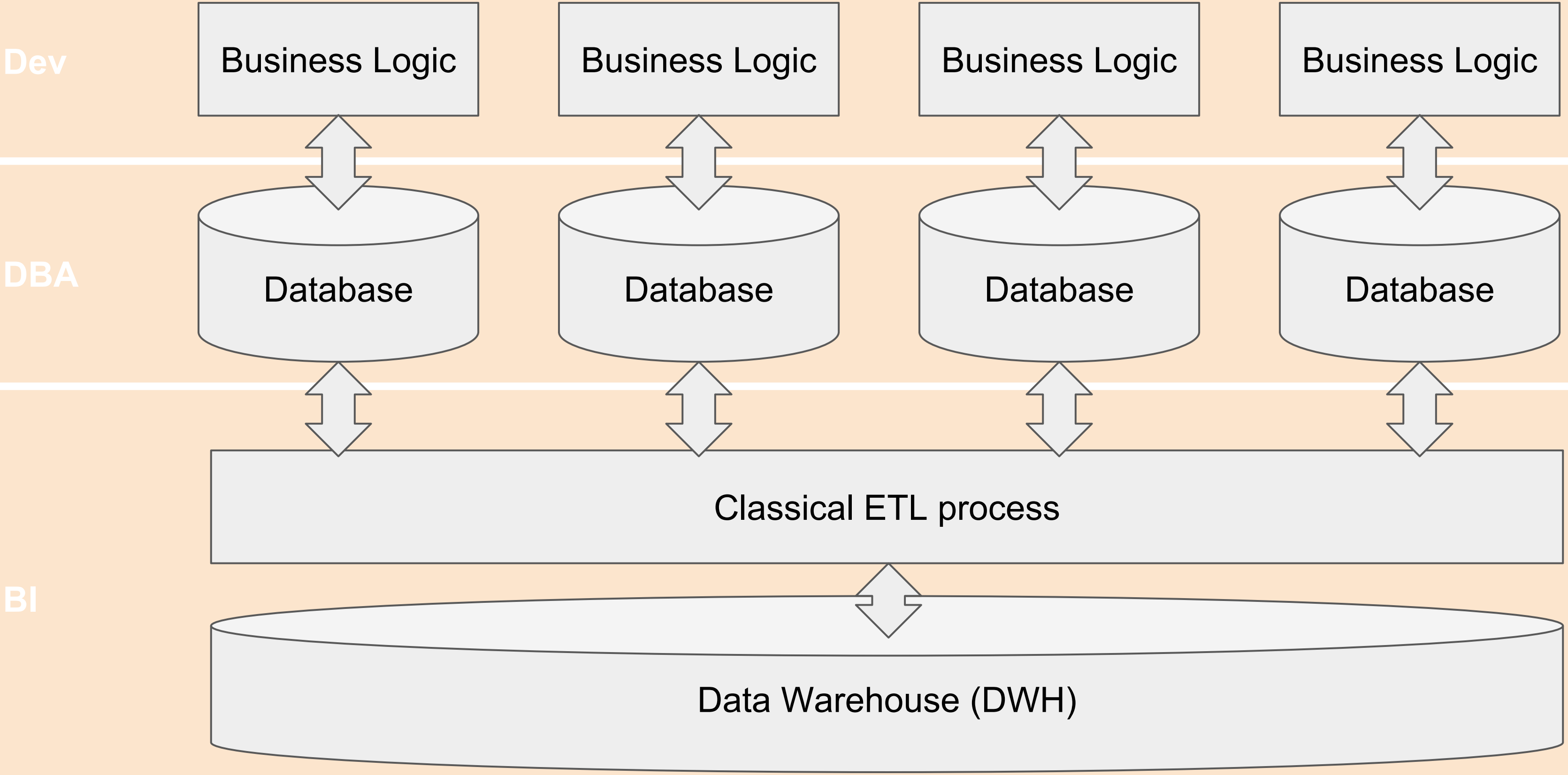
- Java
  - macro service architecture with SOAP as RPC layer
- PostgreSQL
  - Heavy usage of Stored Procedures
  - 4 databases + 1 sharded database on 2 shards
- Python for tooling (i.e code deploy automation)

REBOOT

"macro" services



REBOOT


Dev
DBA
BI
Business Logic
Business Logic
Business Logic
Business Logic
Database
Database
Database
Database
Classical ETL process
Data Warehouse (DWH)

REBOOT

Classical ETL process

- Use-case specific

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

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"

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"

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"

"Zalando 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

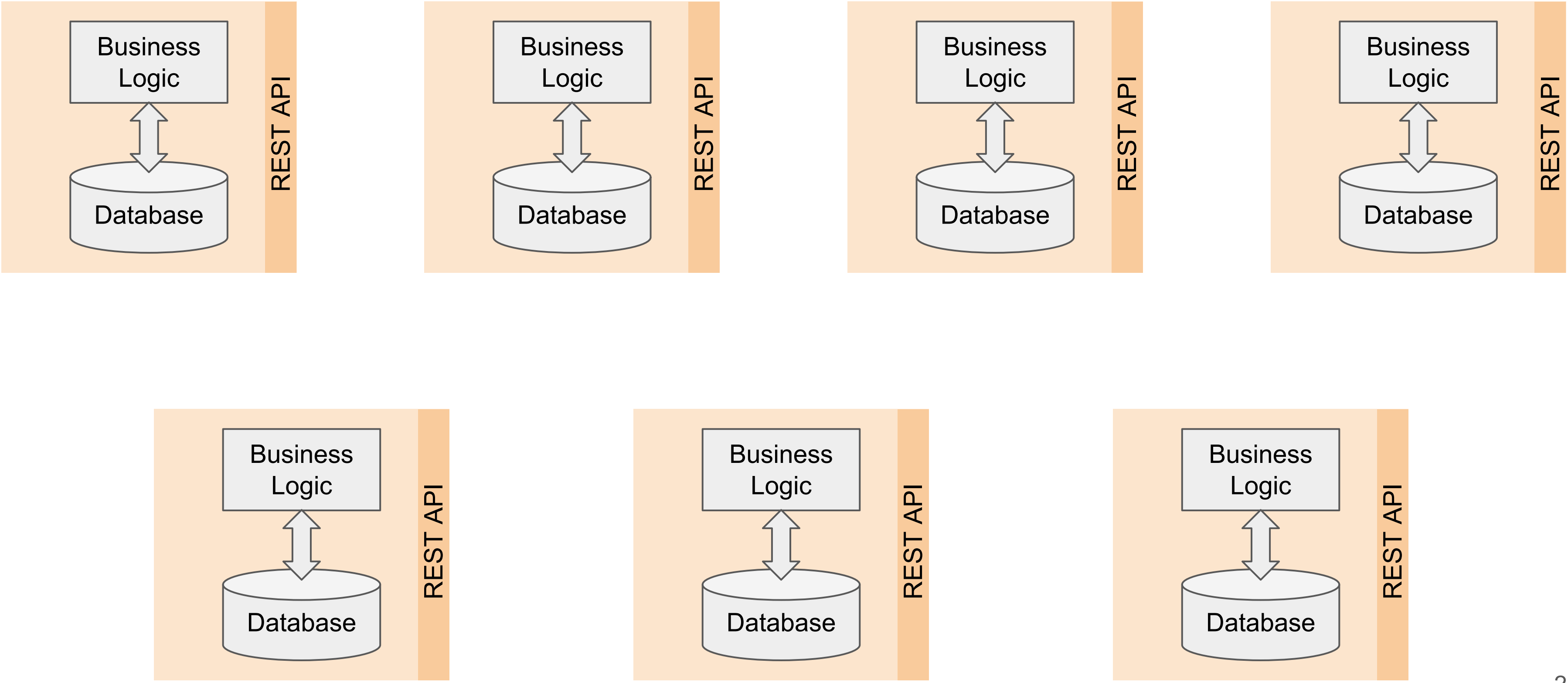
MASTERY

Autonomy

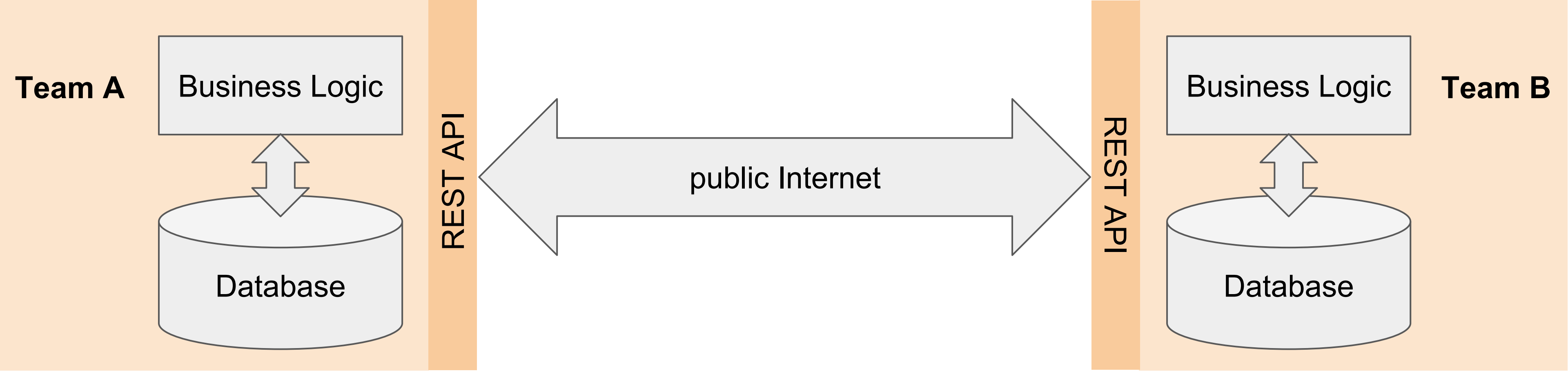
Autonomous teams

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

Supporting autonomy — Microservices

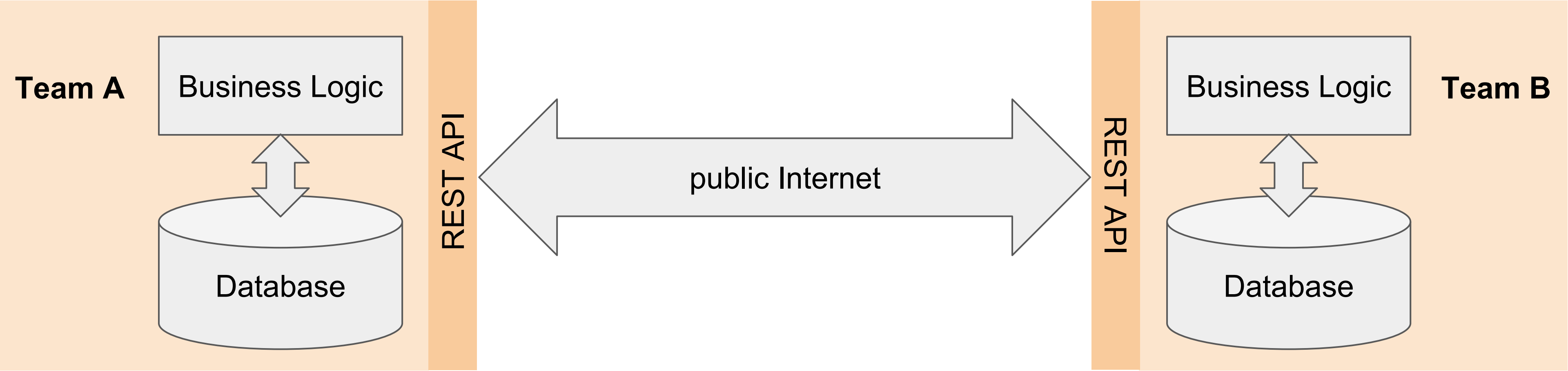


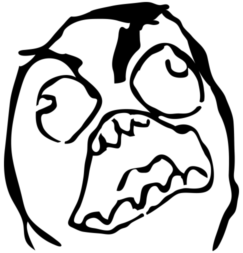
Supporting autonomy — Microservices



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

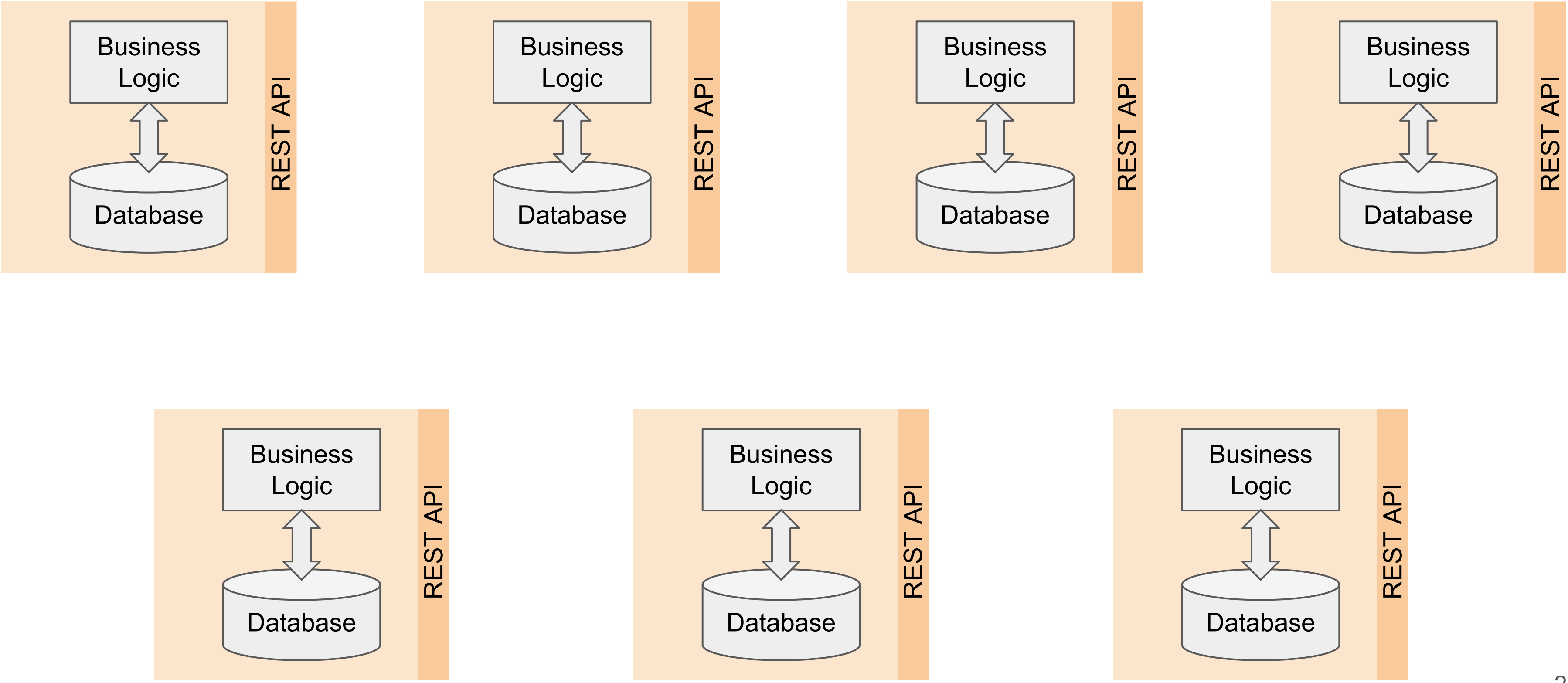
Supporting autonomy — Microservices



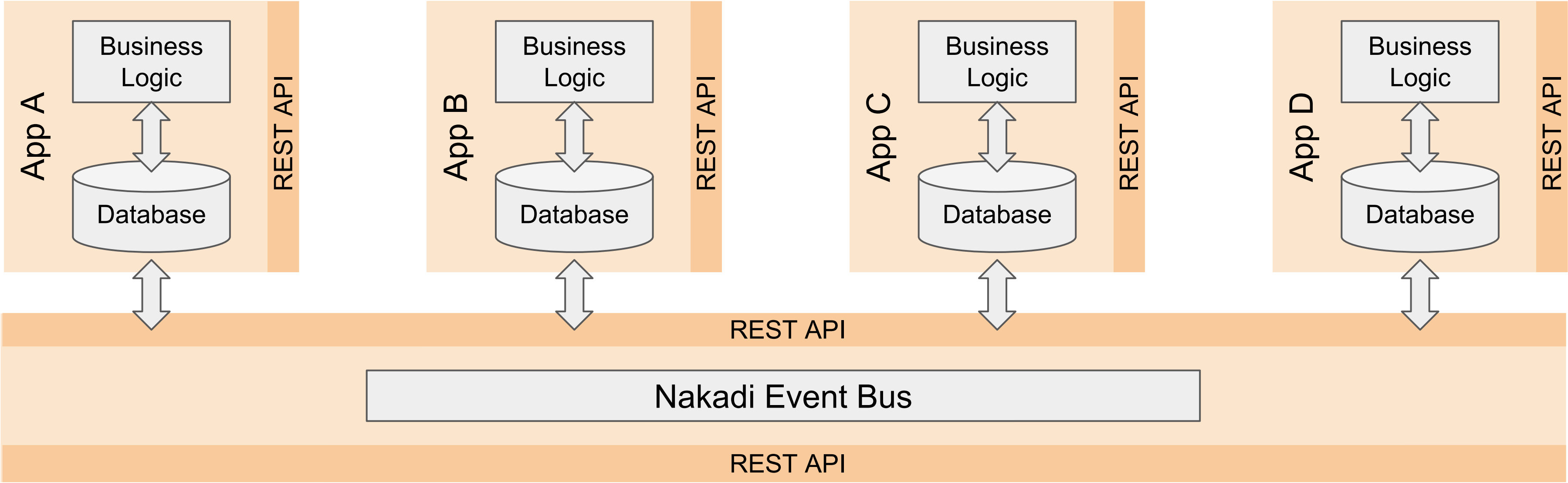


Classical ETL process is impossible!

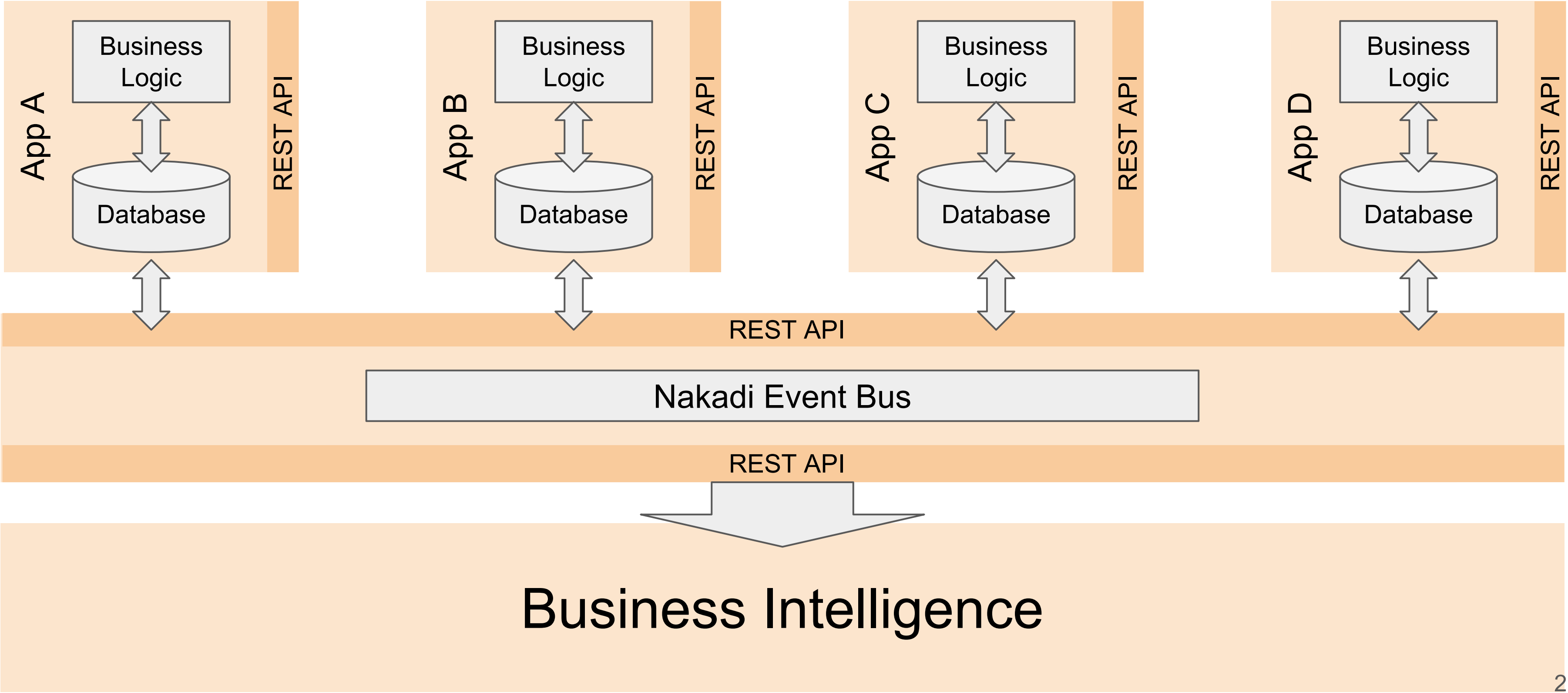
Supporting autonomy — Microservices



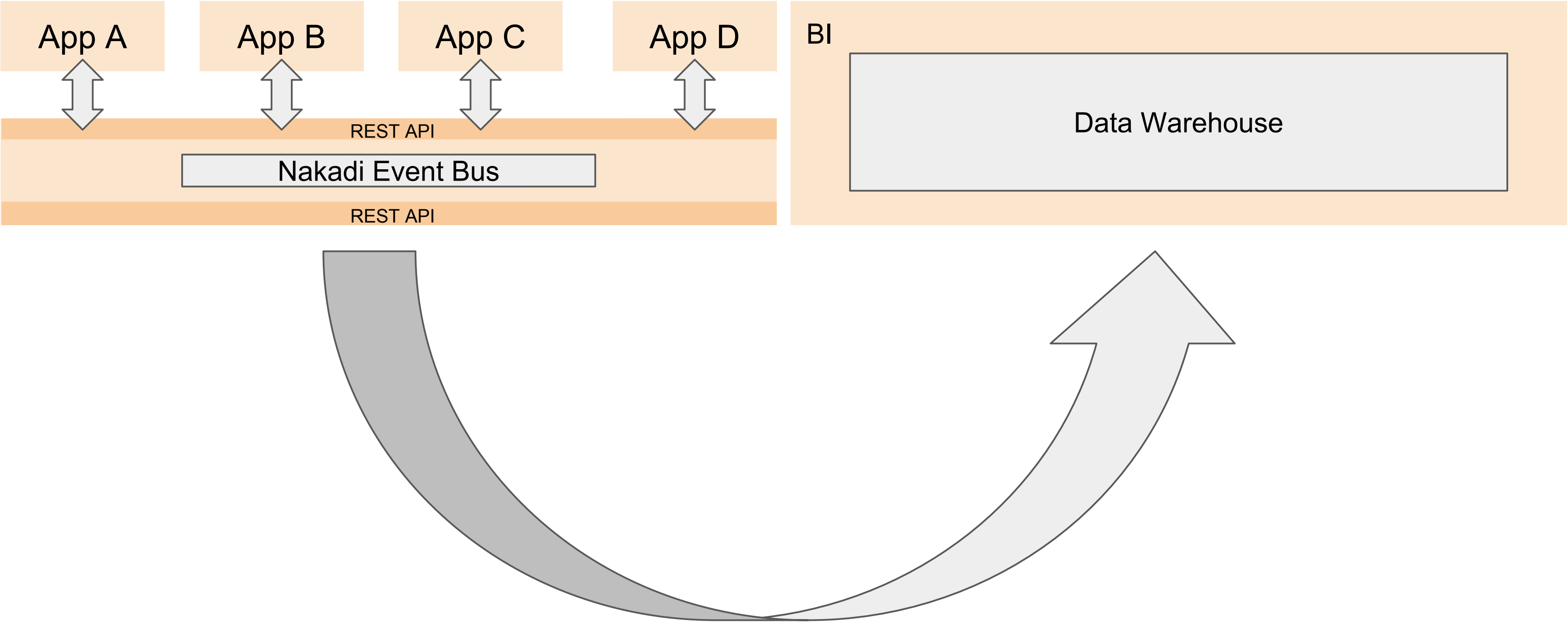
Supporting autonomy — Microservices



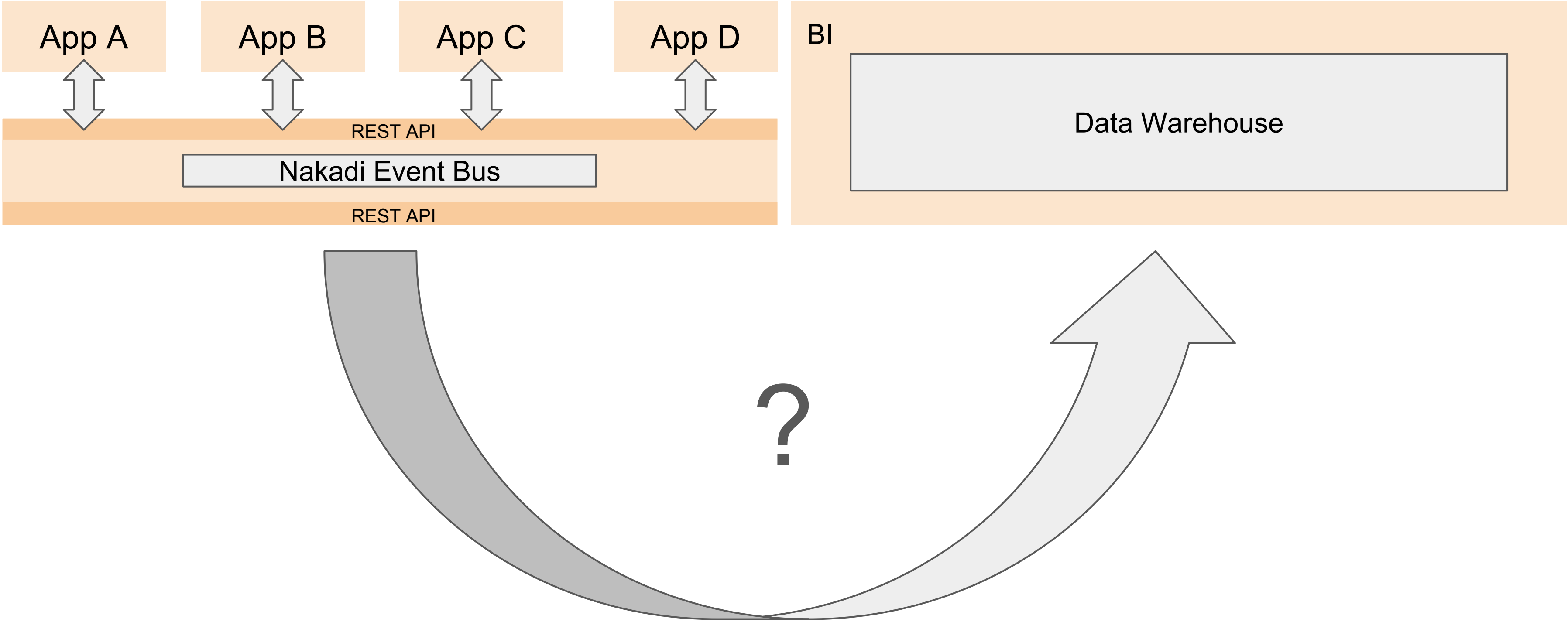
Supporting autonomy — Microservices



Supporting autonomy — Microservices



Supporting autonomy — Microservices



App A

App B

App C

App D

REST API

Nakadi Event Bus

REST API

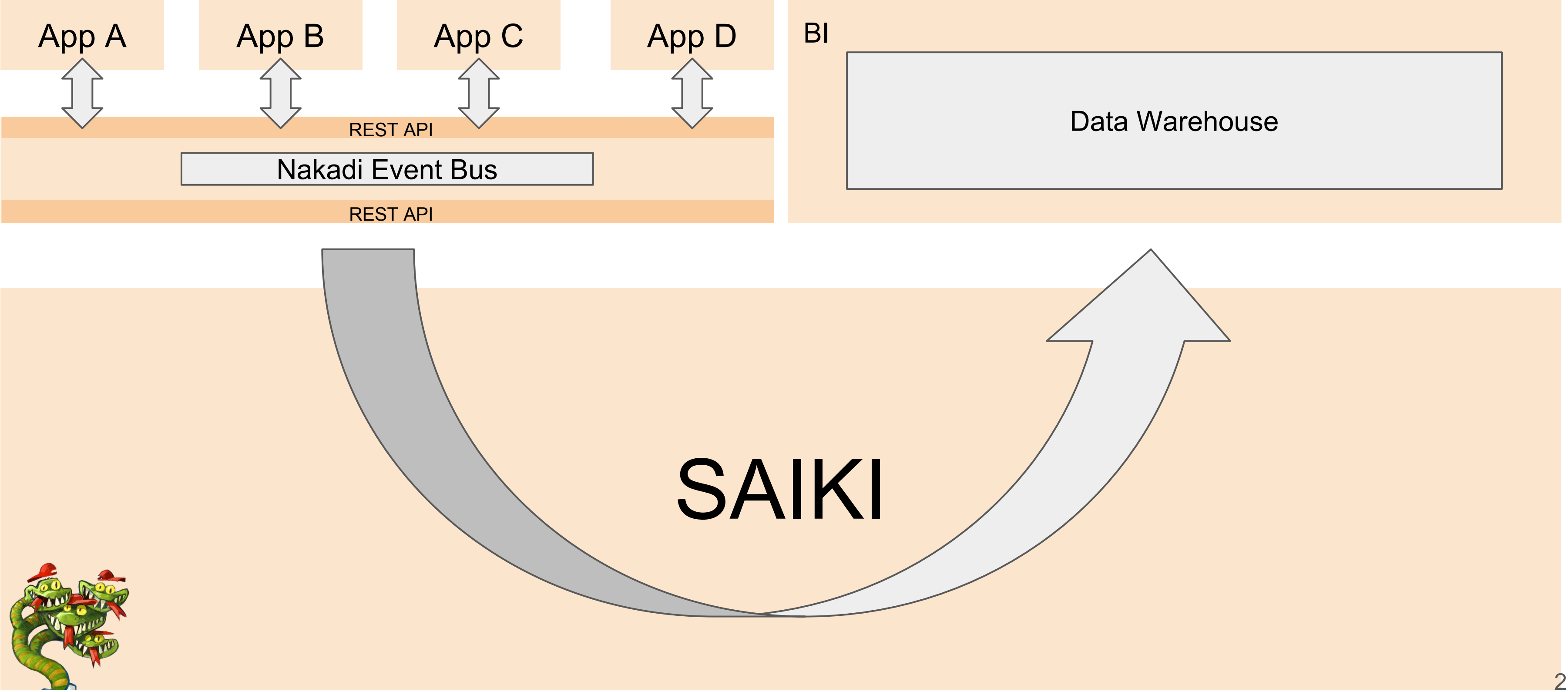
BI

Data Warehouse

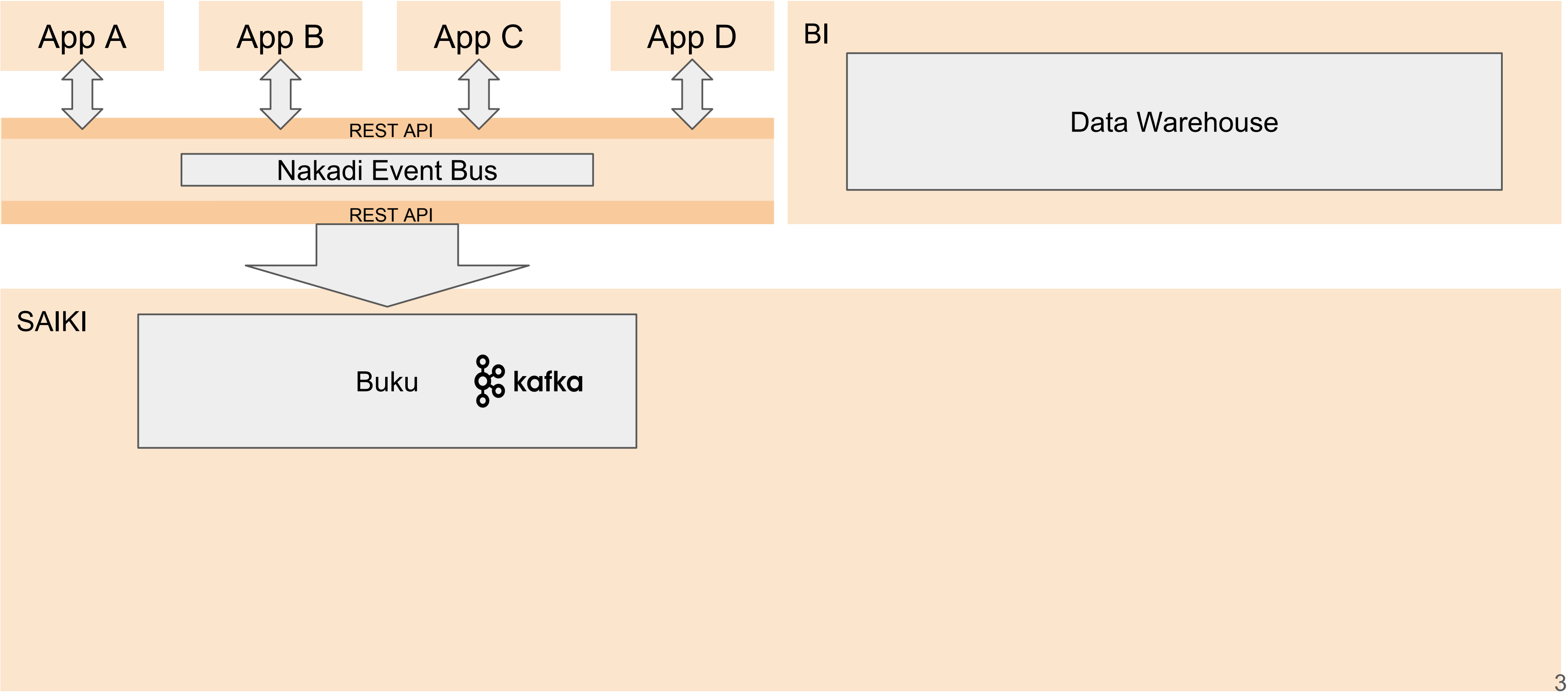
?

SAIKI

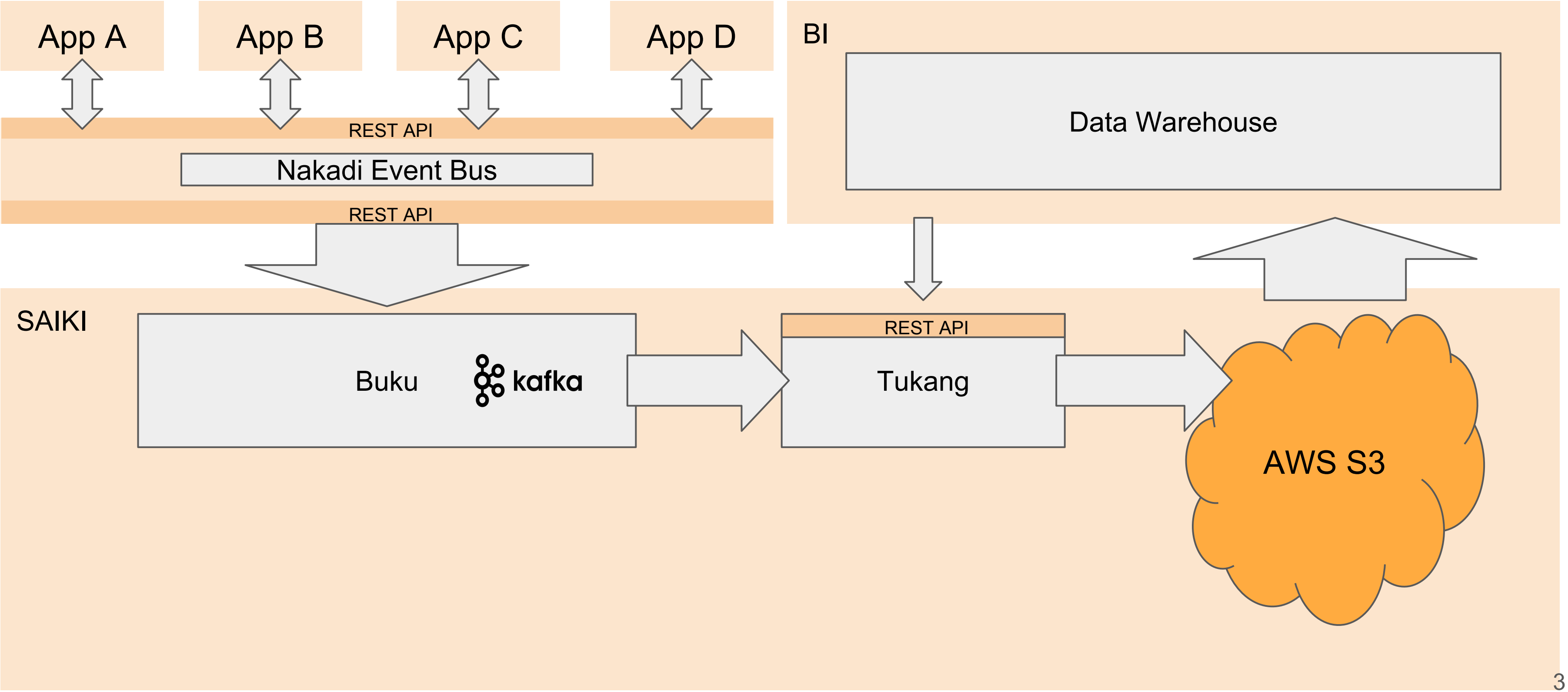
Saiki Data Platform



Saiki Data Platform



Saiki Data Platform

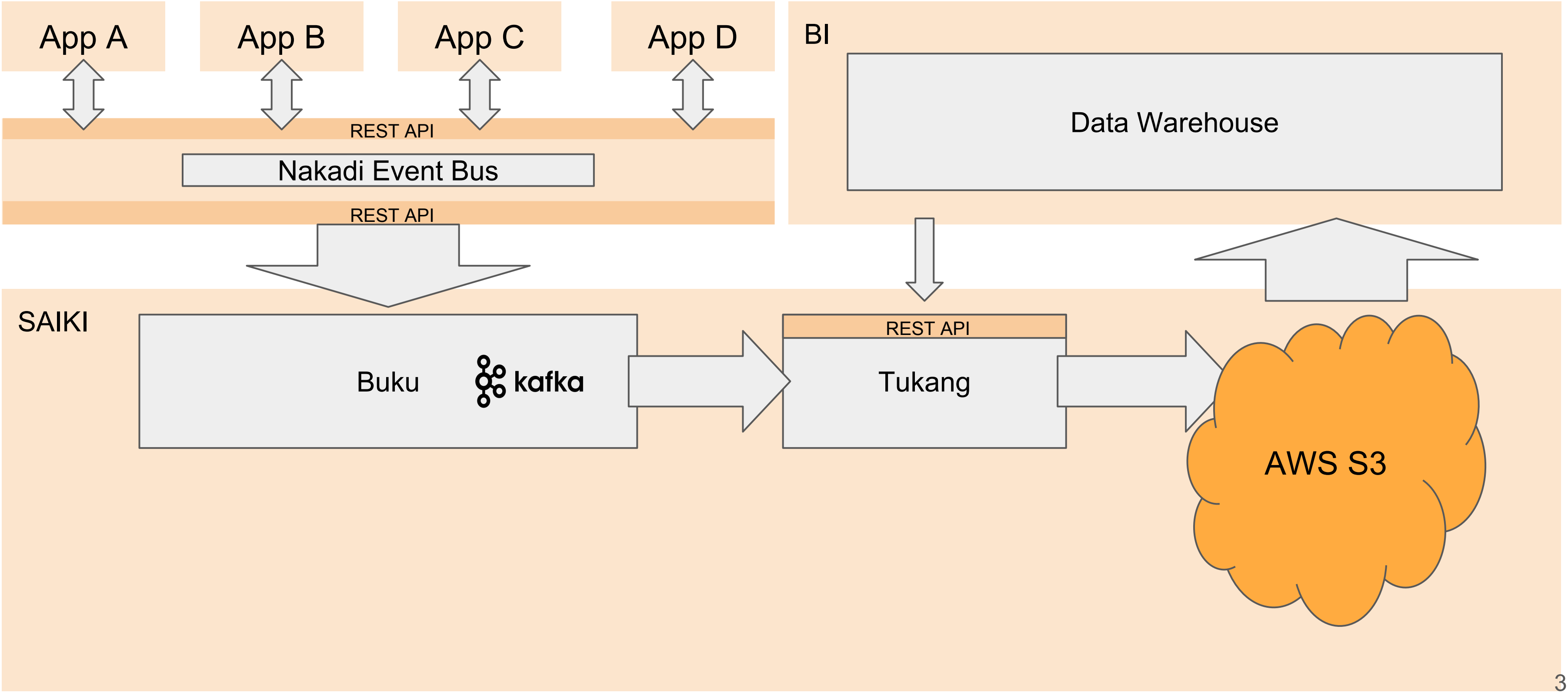


Saiki Data Platform

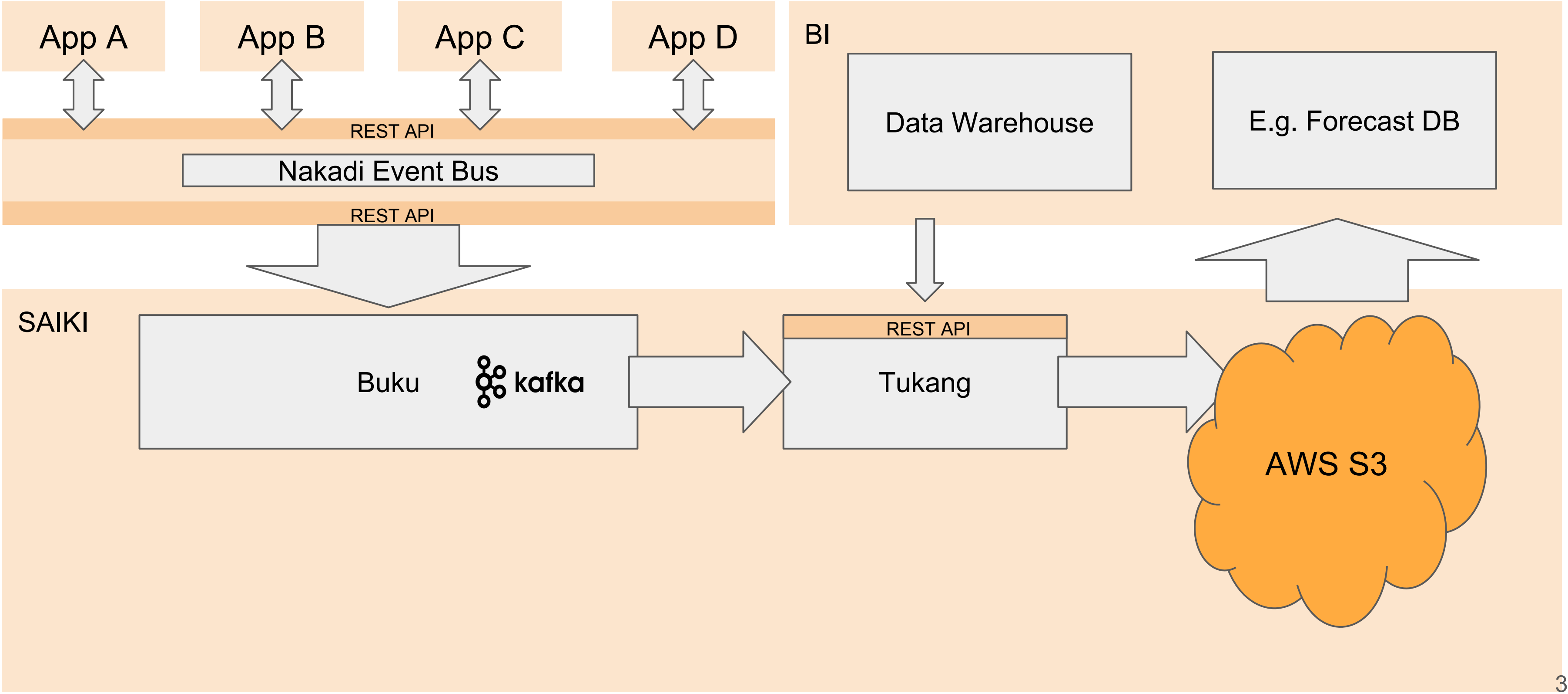
Saiki Tukang

- 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

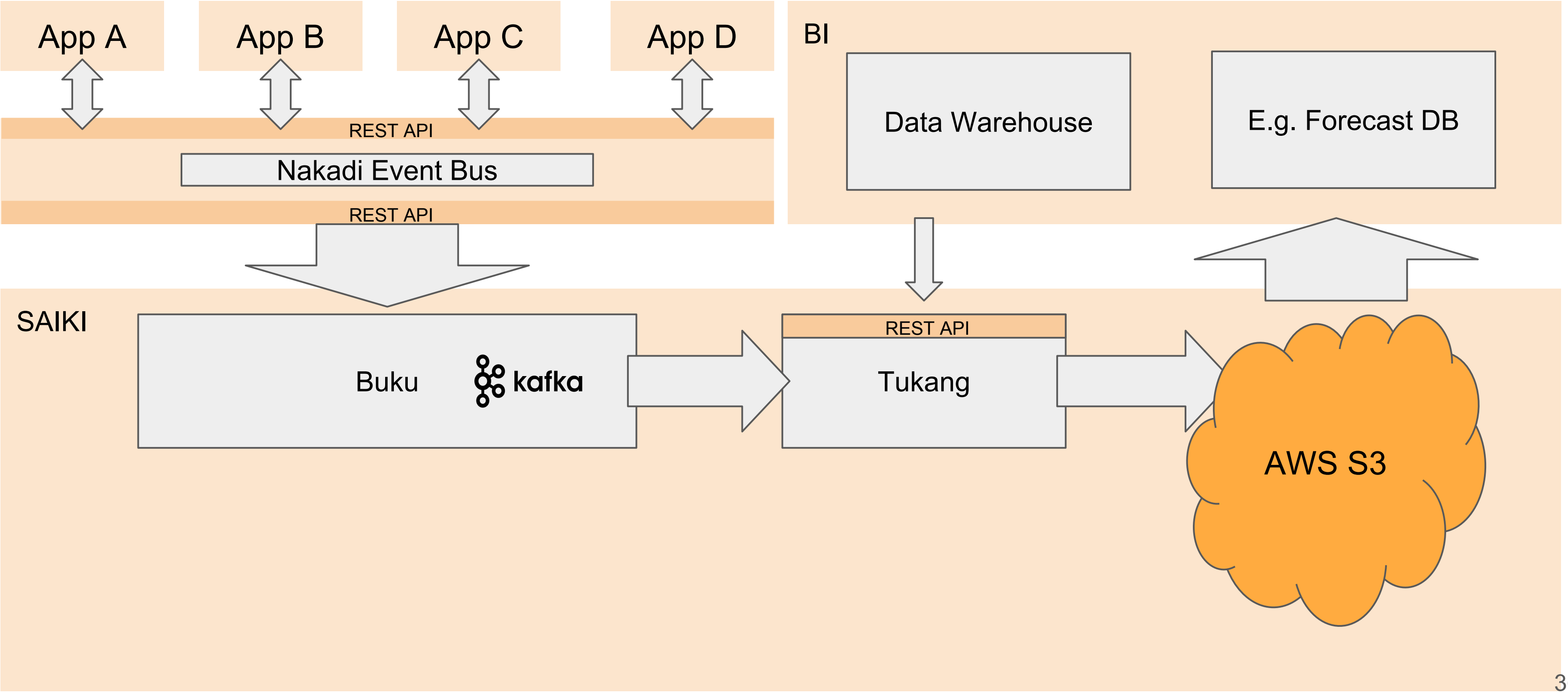


zalando

Saiki Data Platform

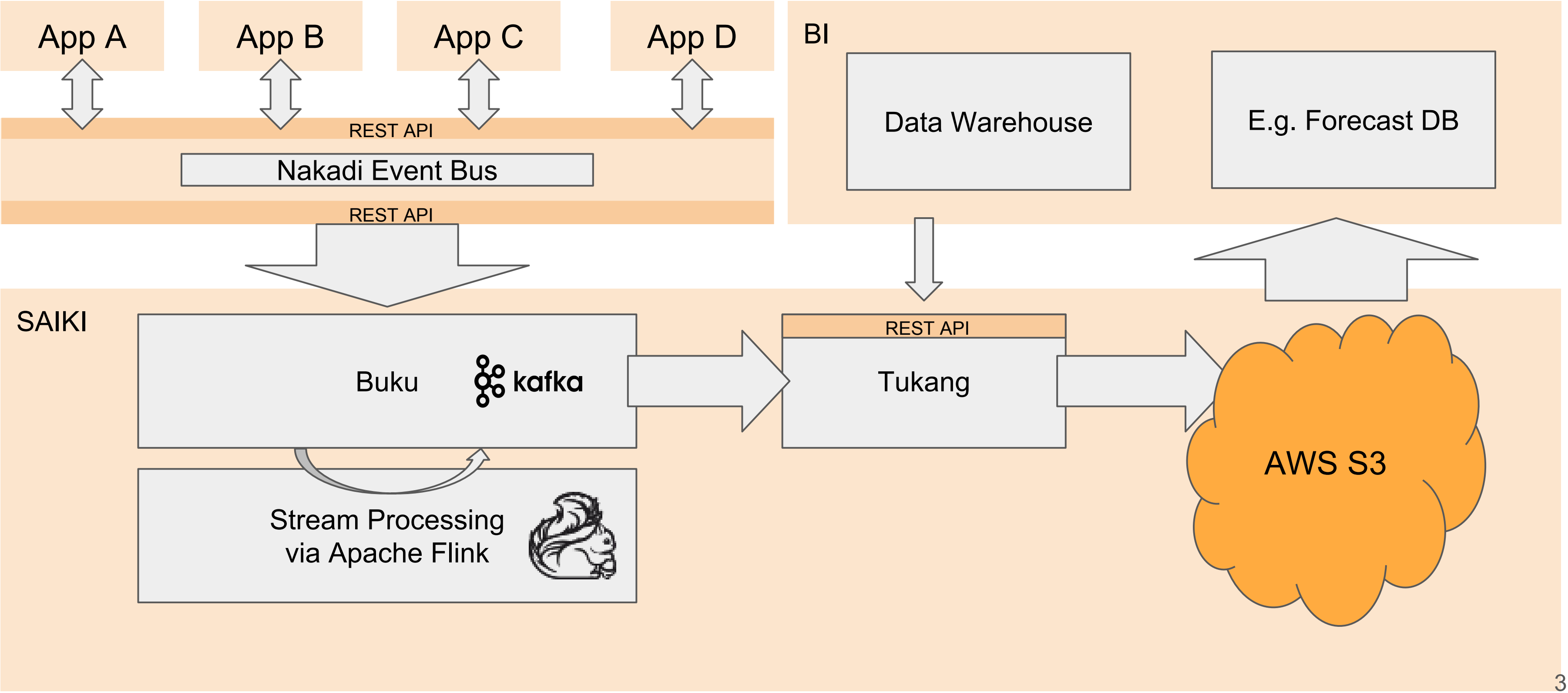
Old Load Process	New Load Process
relied on Delta Loads	relies on Event Stream
JDBC Connection	RESTful HTTPS Connections
data quality could be controlled by BI independently	Trust for correctness of data in the delivery teams
PostgreSQL dependent	Independent of the source technology stack
N to 1 data stream	N to N stream, no single data sink

Saiki Data Platform



zalando

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

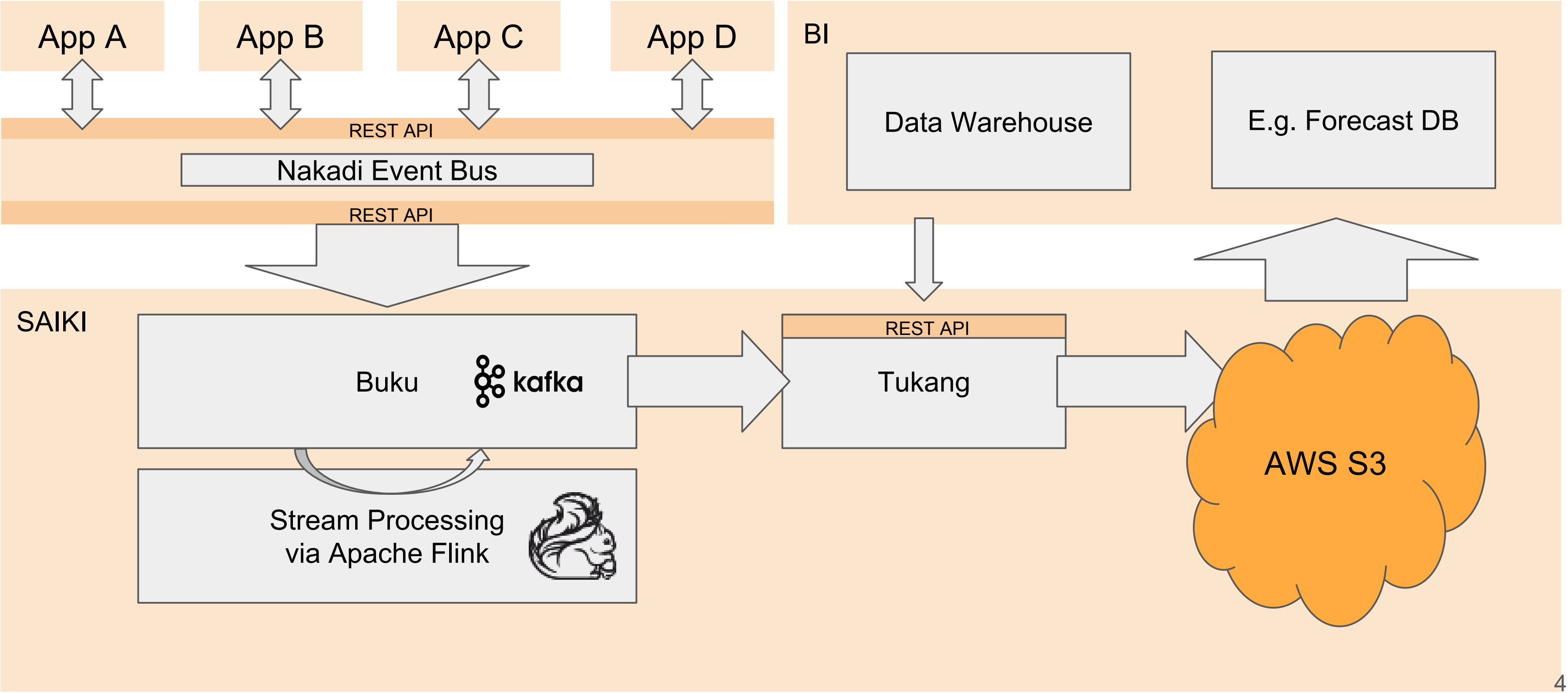
https://tech.zalando.com/blog/apache-showdown-flink-vs.-spark/

Saiki Data Platform

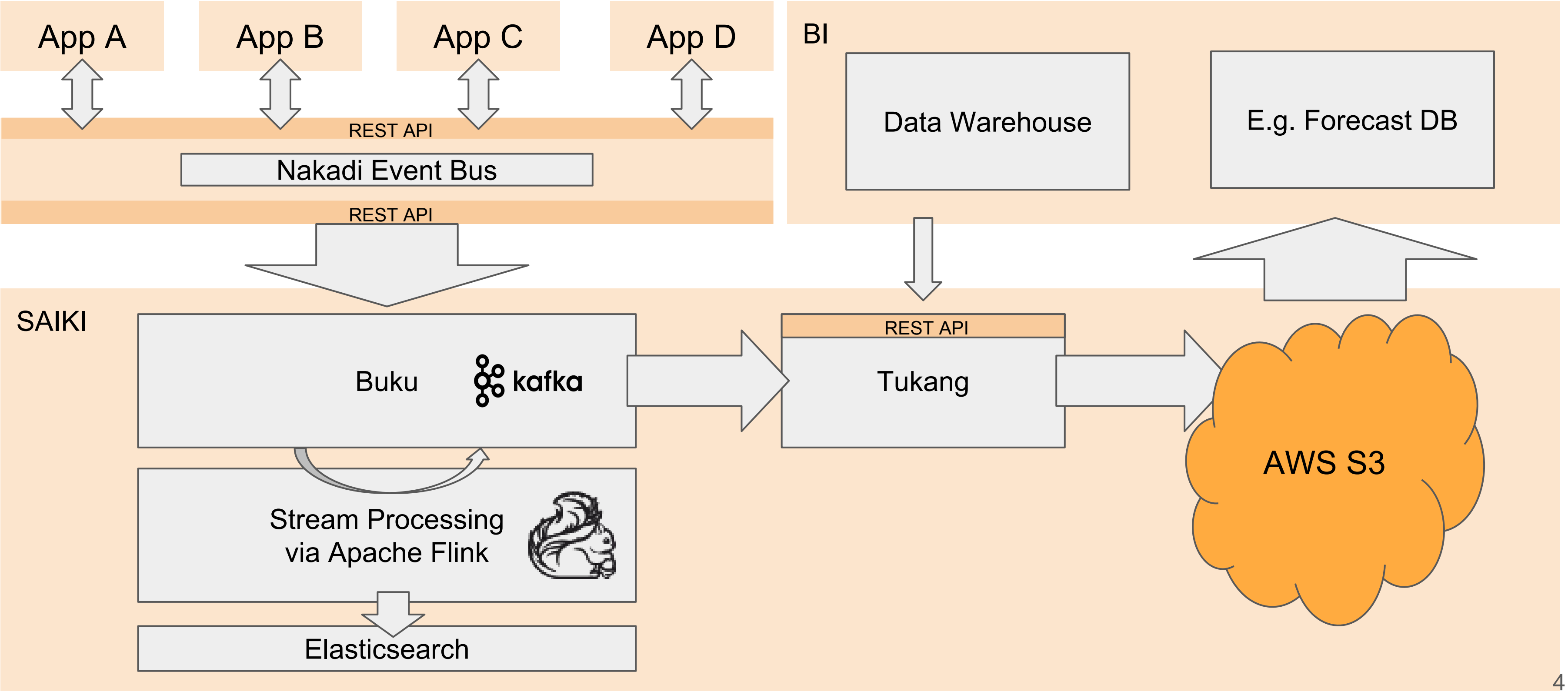
Apache Flink

- connectors
    - Kafka
    - Elasticsearch
    - etc.

Saiki Data Platform



Saiki Data Platform

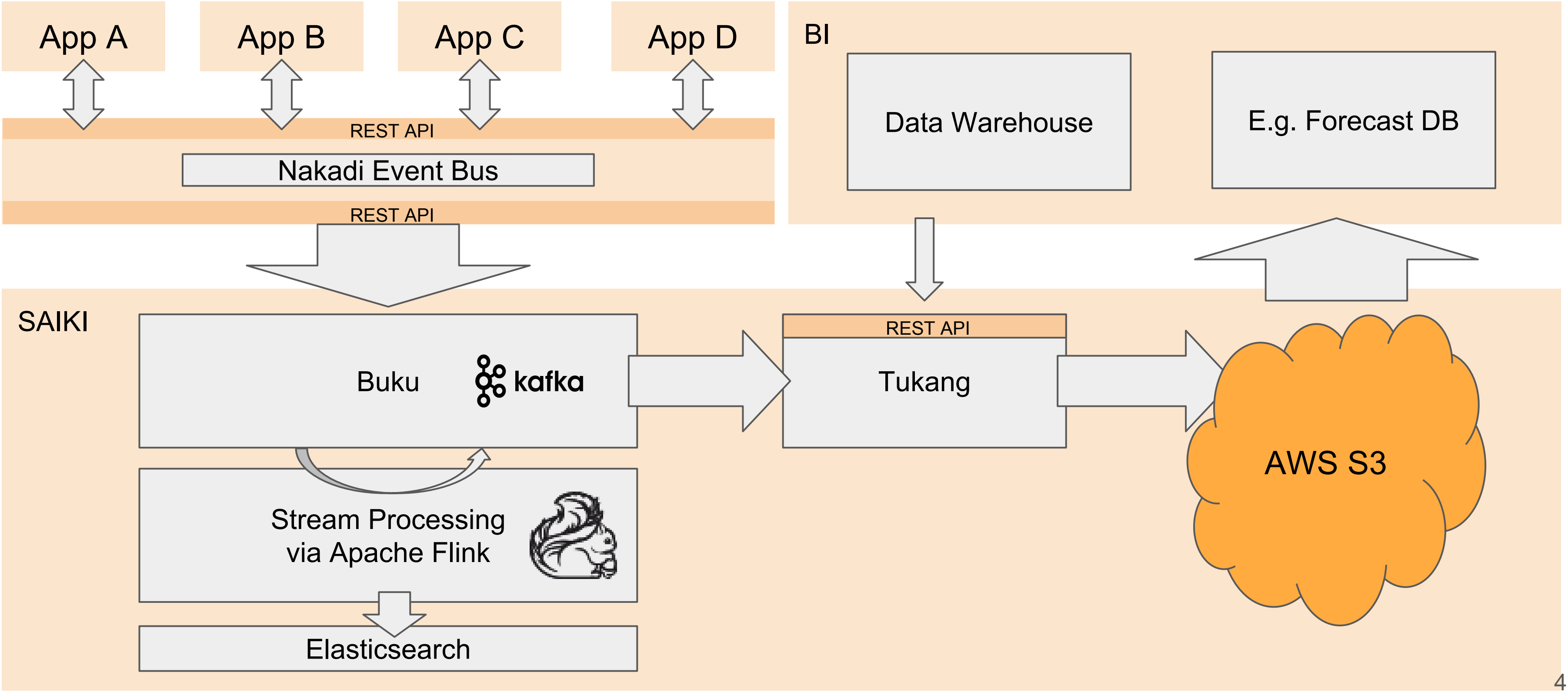


Saiki Data Platform

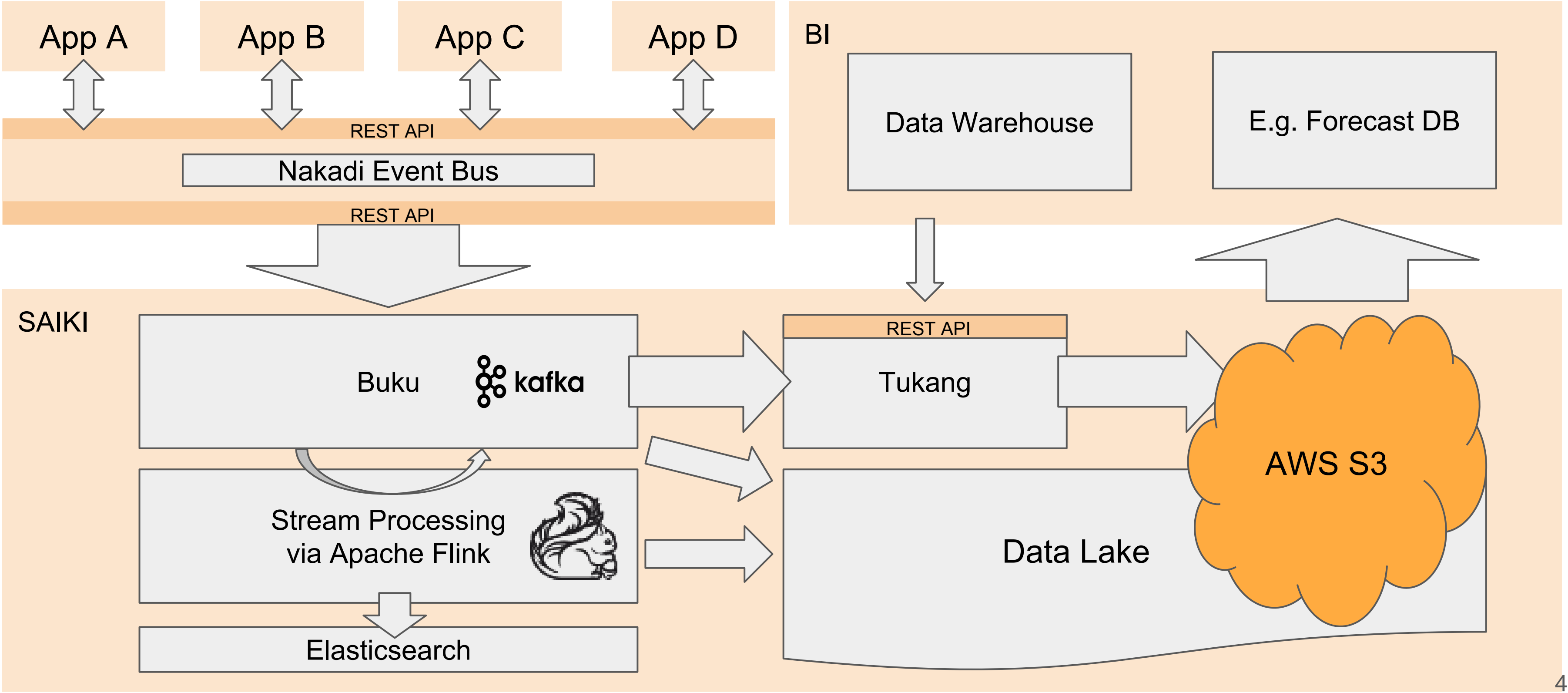
For example: Real-time Business Process Monitoring

- Check if technically the platform works
- Analyze data on the fly
- Visualization with Python/Flask and Chart Frameworks

Saiki Data Platform



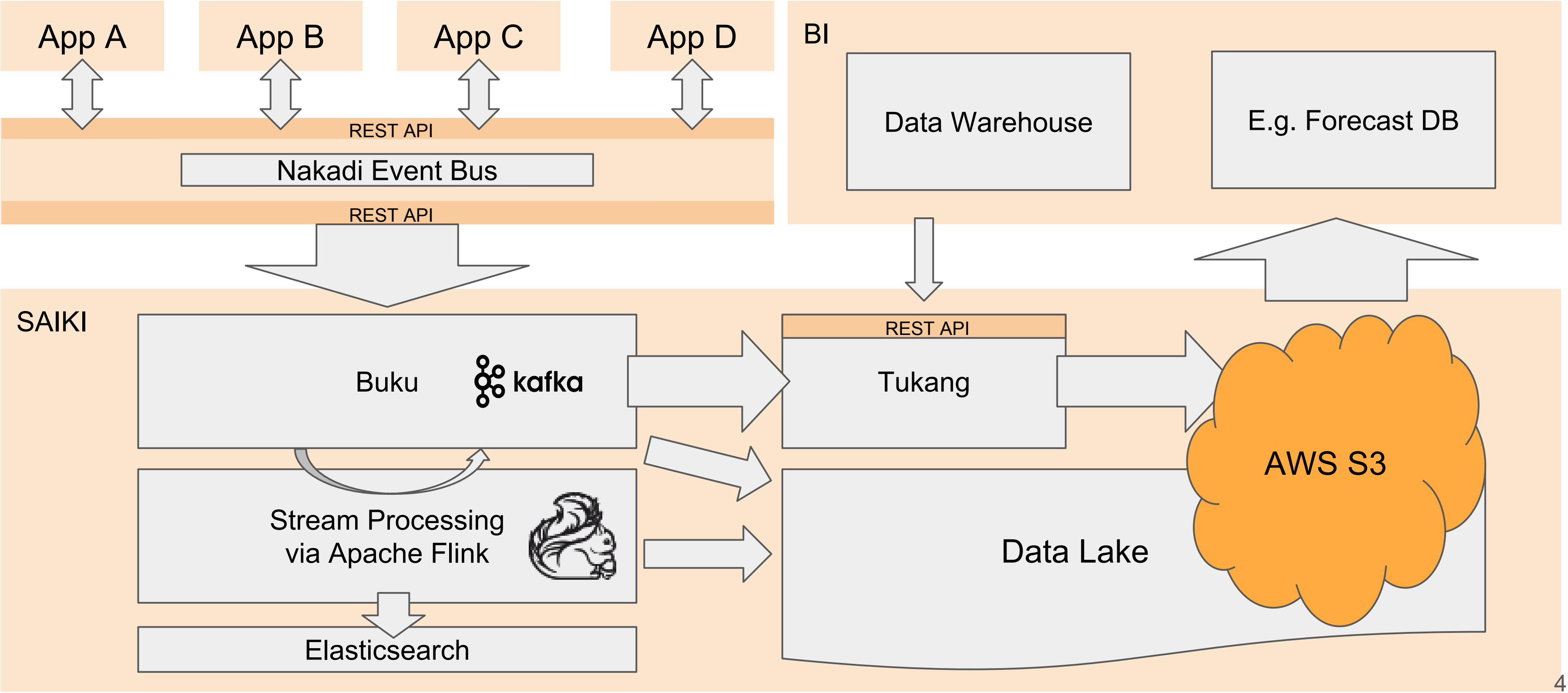
Saiki Data Platform



zalando

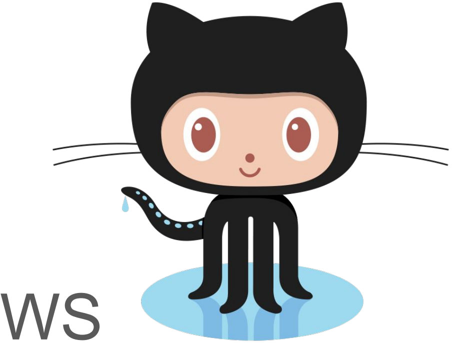
Free the data from the silos!

Saiki Data Platform



Open source @ZalandoTech

- https://zalando.github.io/
- https://tech.zalando.de/blog
- https://github.com/zalando/**saiki**/wiki
- STUPS.io for responsible organizations in AWS



- REST API on Swagger (OpenAPI)
  - https://github.com/zalando/restful-api-guidelines
  - https://github.com/zalando/connexion
  - https://github.com/zalando/play-swagger