APACHE BIG DATA EUROPE 2016

The Role of Apache BigData Stack in Finance: A Real-World Experience on Providing Added Value to Online Customers

Luca Rosellini, KEEDIO

Twitter handles: @KEEDIO, @luca_rosellini

WHO WE ARE

www.keedio.com

@keedio

Keedio History

2006

Our mother company has been involved in distributed computing projects for 10 years (grid computing).

2014

A Joint initiative of Banco Santander and the private university "Universidad Alfonso X el Sabio".

Focused on Big Data technologies (and Cloud Computing)

2016

Team composition:

- 20% Data analysts
- 50% Developers
- 30% System engineers

KEEDIO

Platform

Keedio builds and maintains the "Keedio Data Stack" an OpenSource BigData platform built from scratch.

Services

Big Data Assessment (use cases discovery).

Project development and after sales.

Platform-agnostic, we do not push our own platform, we work with the customer chosen one or suggest which one is most appropiate for a specific project.

Technical and commercial training.

We approach project with partners, the success of a project comes from our partner's domain knowledge and our technical competences.

BANKS: THE EVOLUTION TO BIG DATA

THE PAST, THE PRESENT AND THE FUTURE

Banks: Evolution of their challenges and the relationship with customers

Banco di San Giorgio

(Genova)

1405

1406

1407

Banks: Evolution of their challenges and the relationship with customers

1406

1407

Customer Bank

- Very few people were using bank services.
- We still didn't know how to do banking properly

Banks: Evolution of their challenges and the relationship with customers

Customer Bank

- Banks fully knew their customers.
- Customers were loyal to their banks.
- Banks only had to provide a few financial services.

Banks: Evolution of their challenges and the relationship with customers

'90s

TODAY

- Shift to a virtual world.
- Banks face a plethora of new challanges to overcome.
- Legal voids
- New financial “actors”, native to the virtual world, come into play.
- Customers are not loyal anymore. Milton Friedman, “the freedom to choose”.

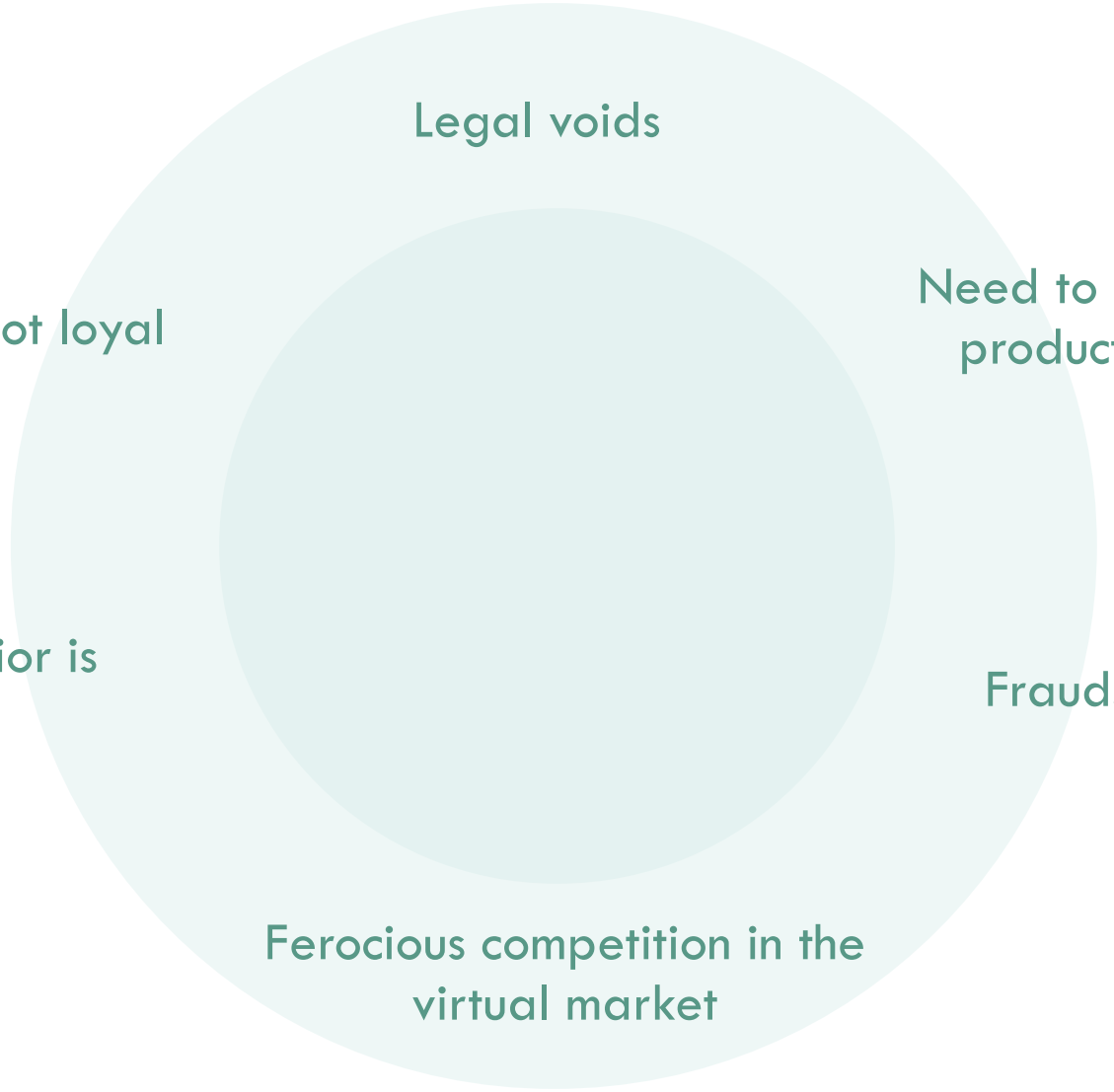
Banks: Evolution of their challenges and the relationship with customers

> Banks are unaware of who their customers are.

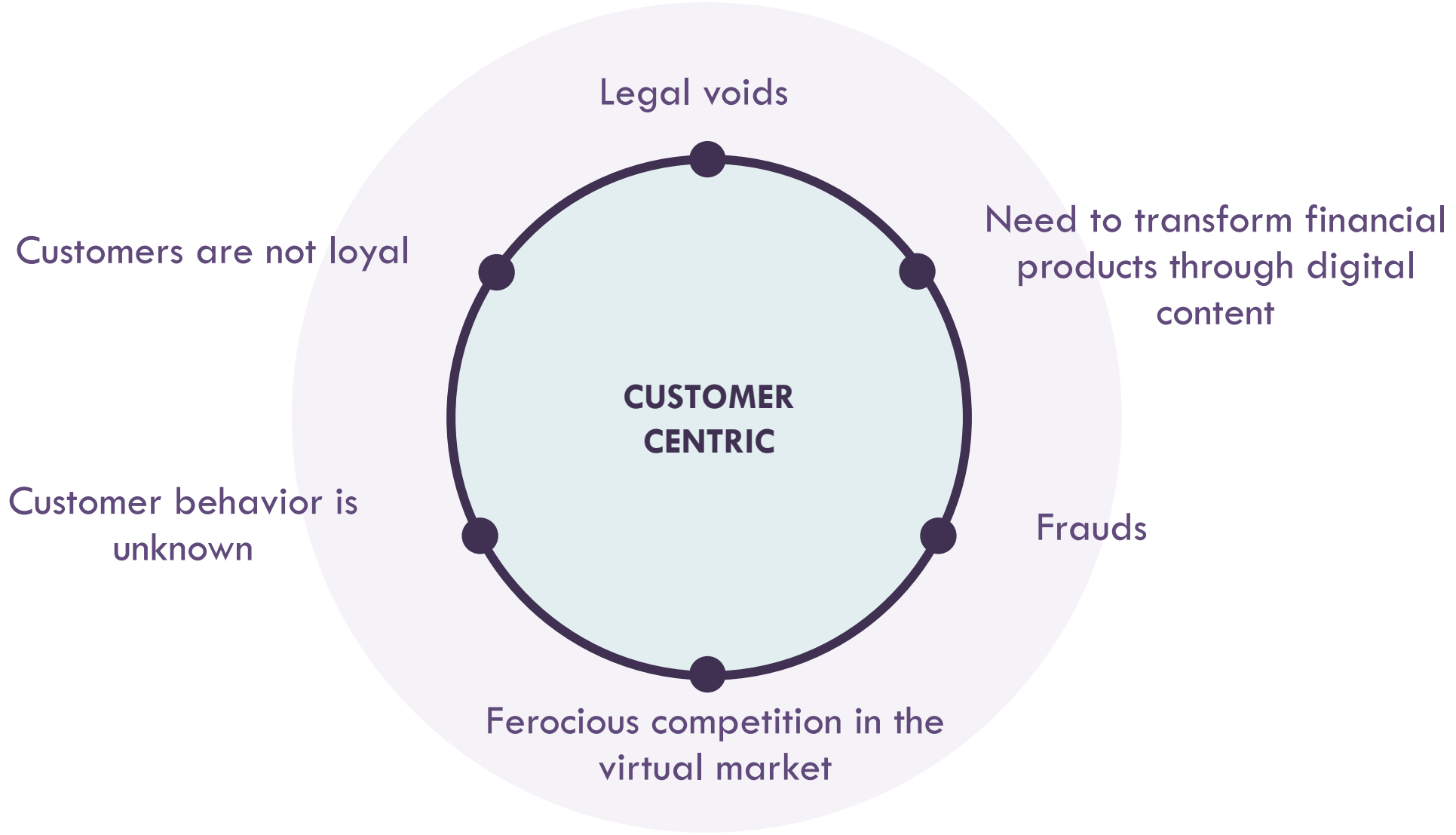
Banks: Evolution of their challenges and the relationship with customers

> They need to take back this knowledge.

Banks: Evolution of their challenges and the relationship with customers



Banks: Evolution of their challenges and the relationship with customers




“War is ninety percent information”


“War is ninety percent information”
Napoleon Bonaparte

Banks: Evolution of their challenges and the relationship with customers

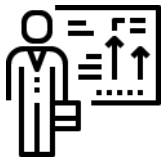
THE CHALLENGE

KNOWING YOUR CUSTOMER

The Challenge (business)

Our customer wanted to embrace the “customer centric” strategy.

Enrich their traditional core banking services with advanced analytics.



Get more knowledge of customer behavior.

Provide custom-fitted, tailored-banking.

Offer the right product at the right moment, on a 24x7 basis.

The Challenge (infrastructure)

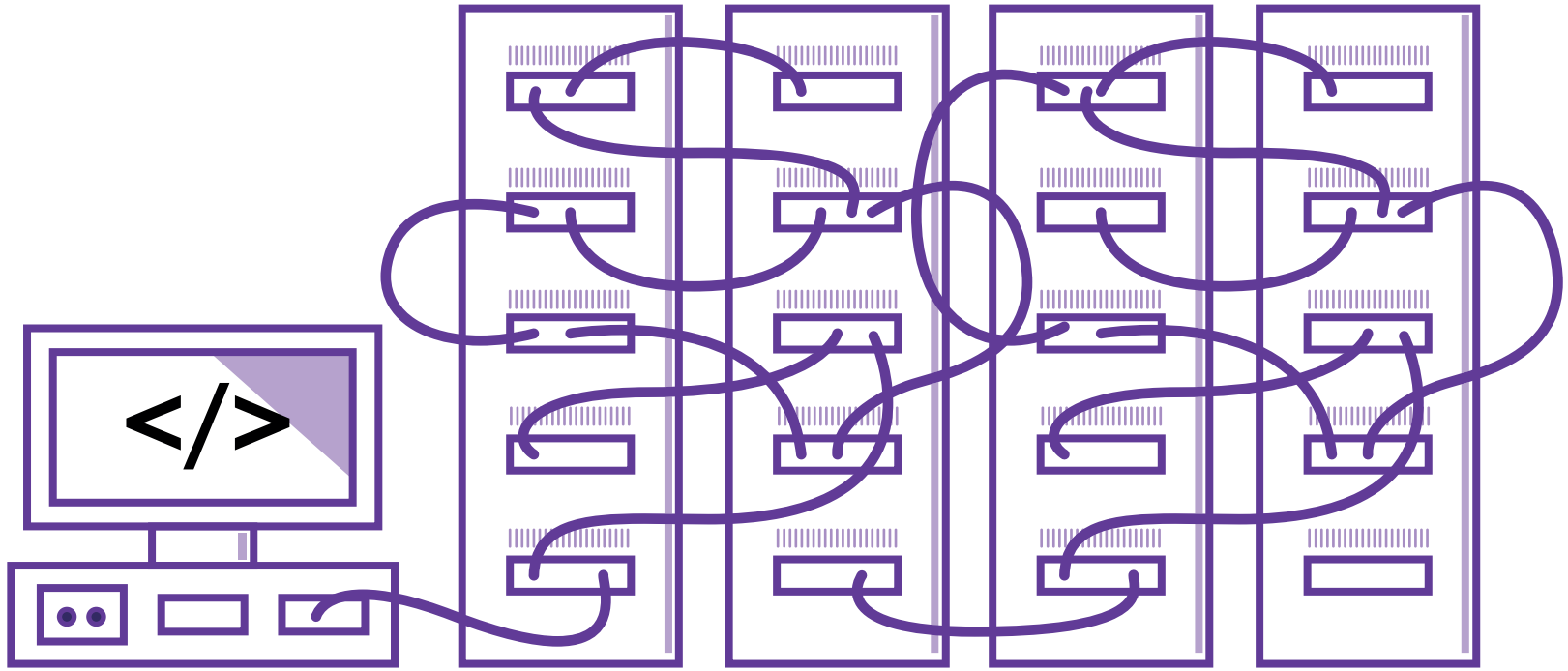
Monolithic, expensive, propietary legacy platforms

→

Platform composed of open source, easily replaceable components, whose development is backed by a whole community.

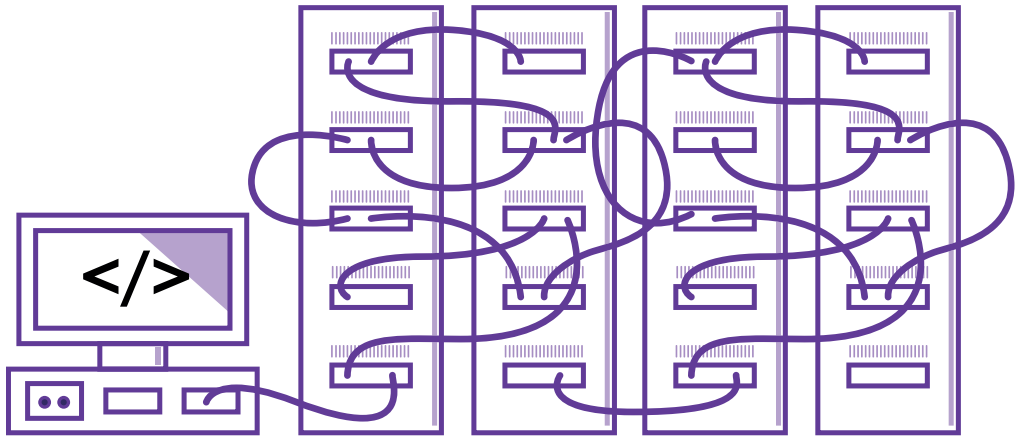
THE SOLUTION

ARCHITECTURE



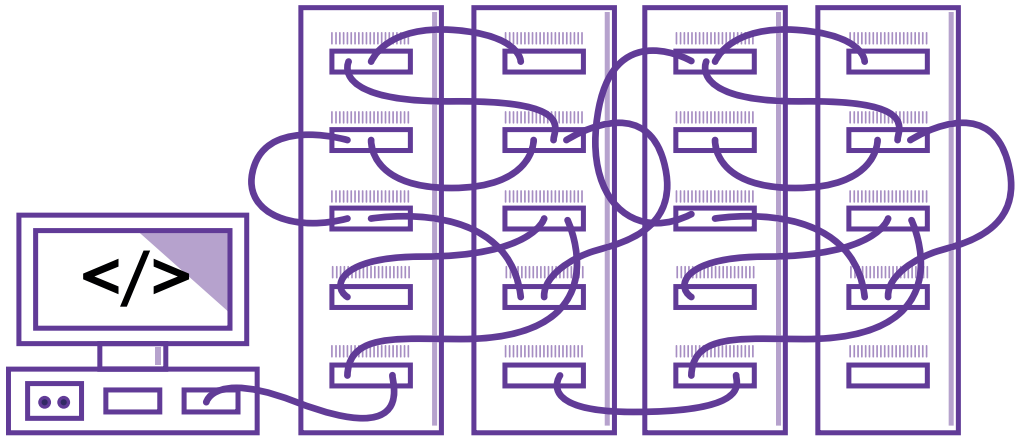
On-premises

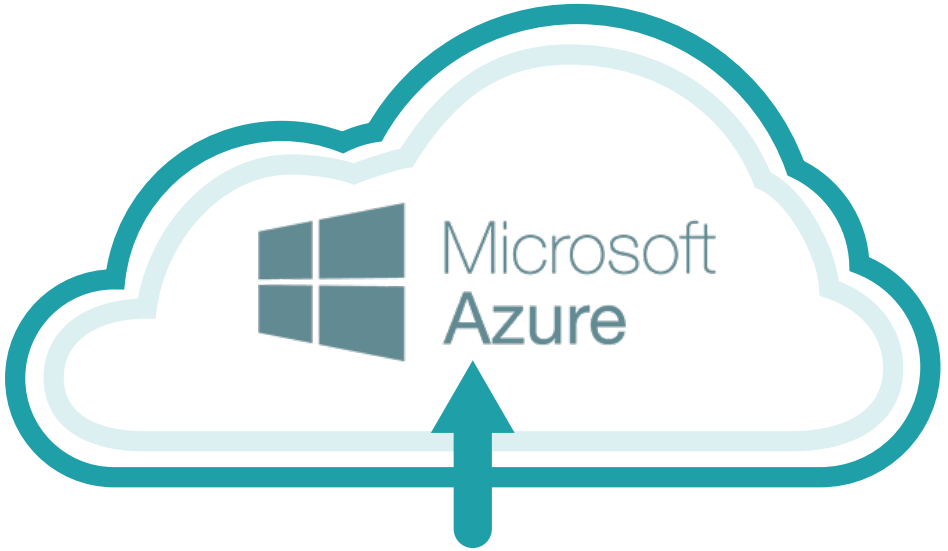
ARCHITECTURE



On-premises → Cloud

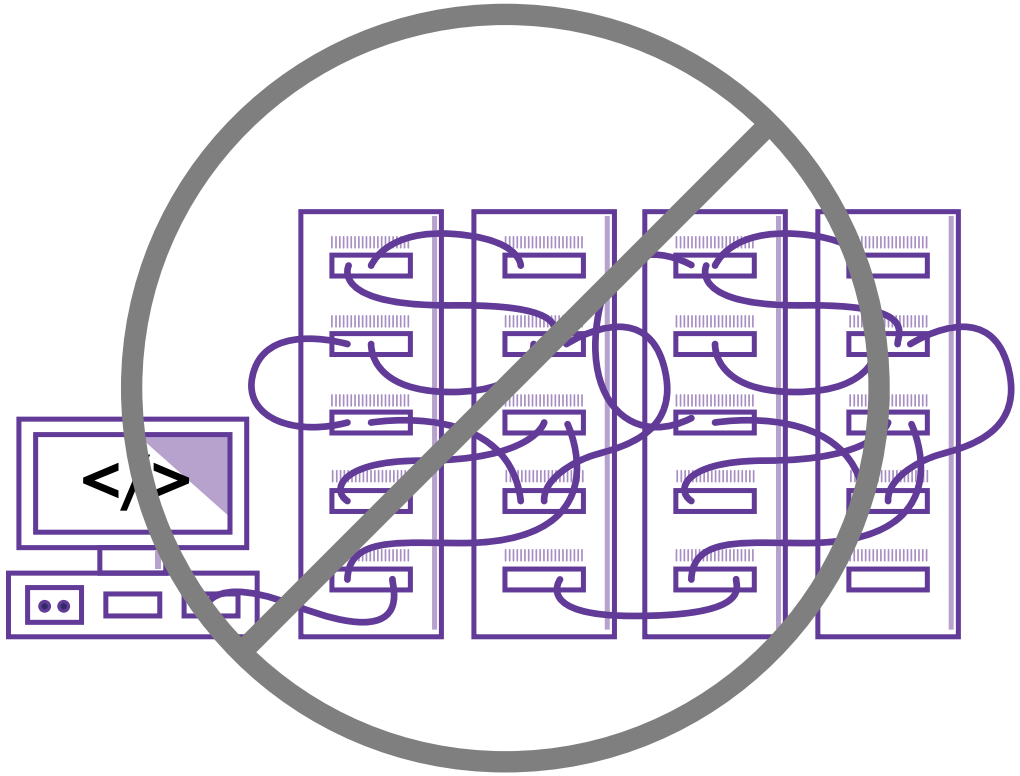
ARCHITECTURE



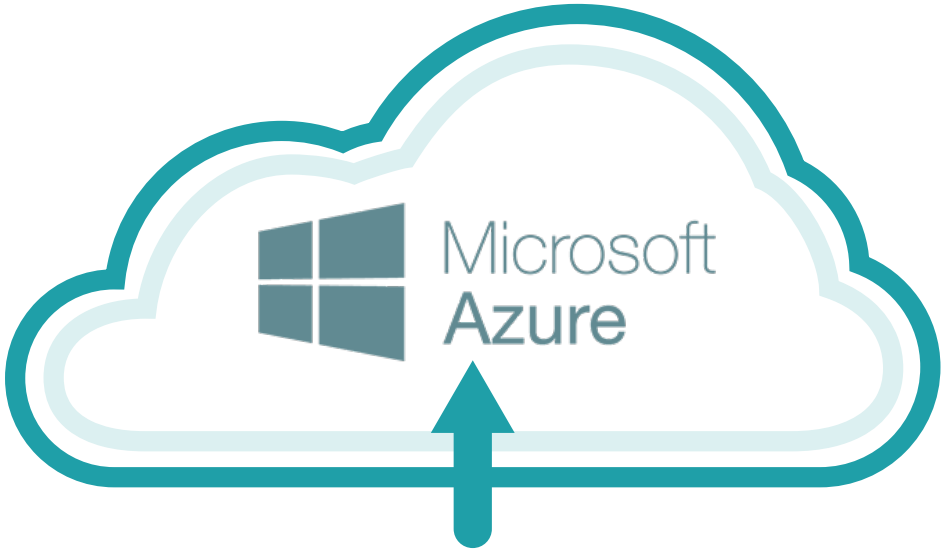


On-premises → Cloud

ARCHITECTURE

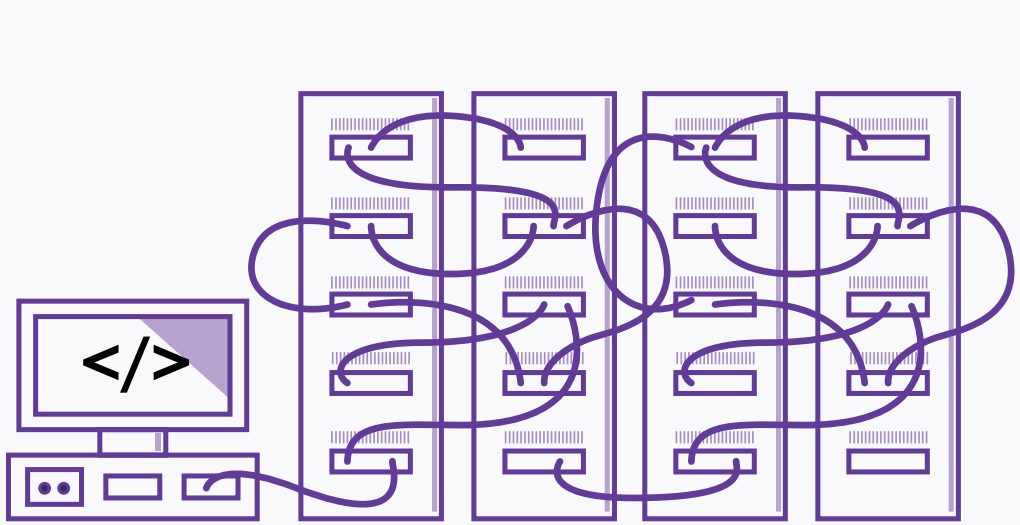


On-premises



Cloud

ARCHITECTURE

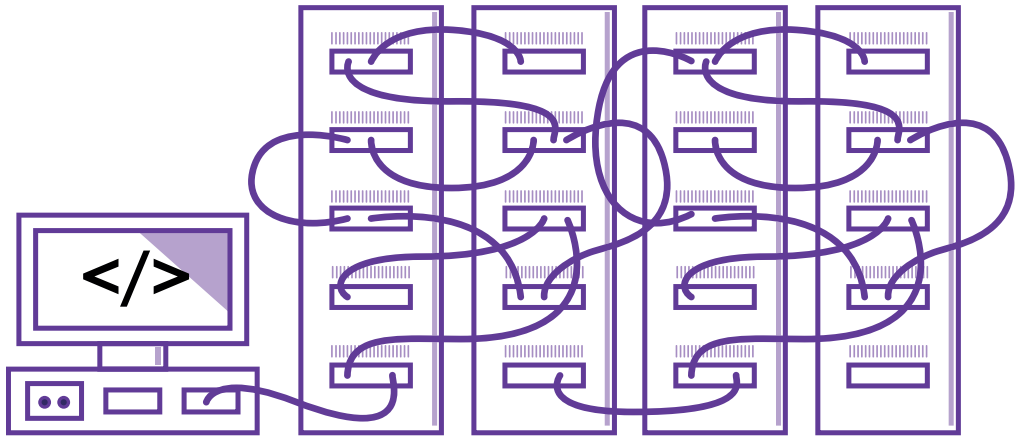


On-premises

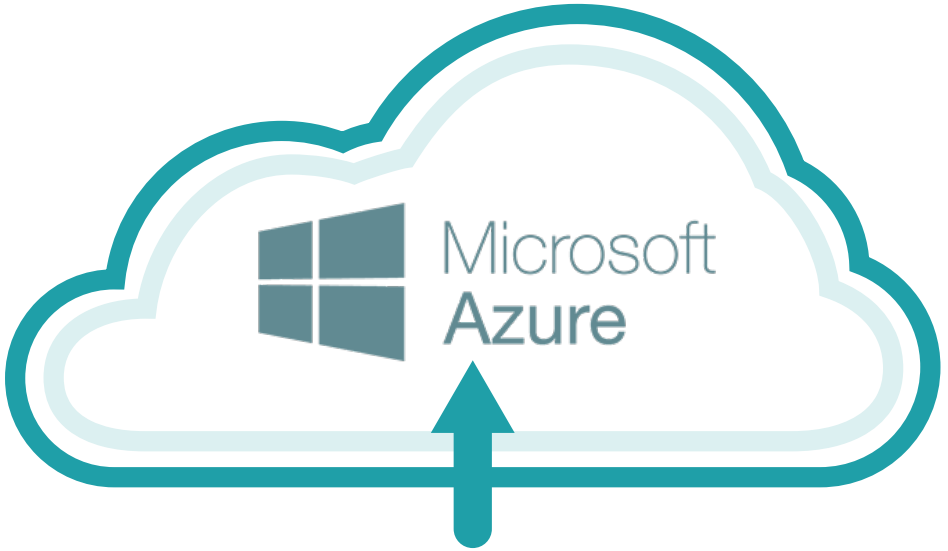
Cloud

Hybrid on-premises/cloud infrastructure

ARCHITECTURE



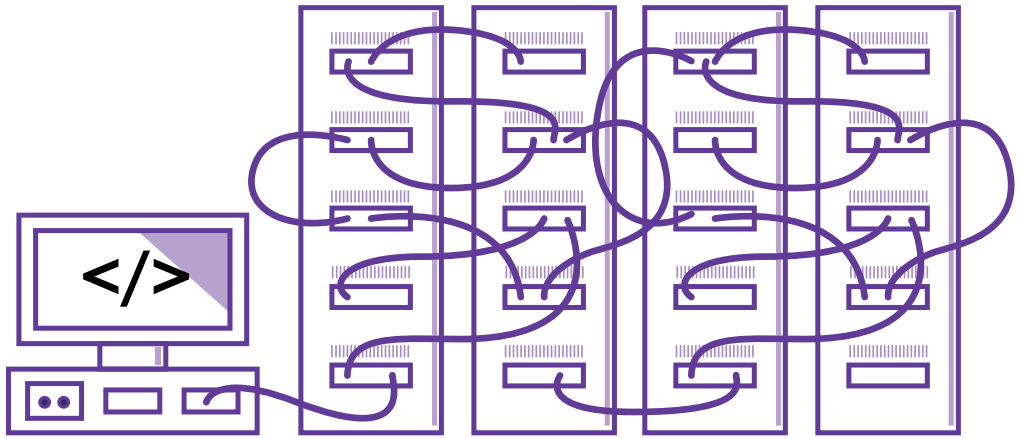
On-premises



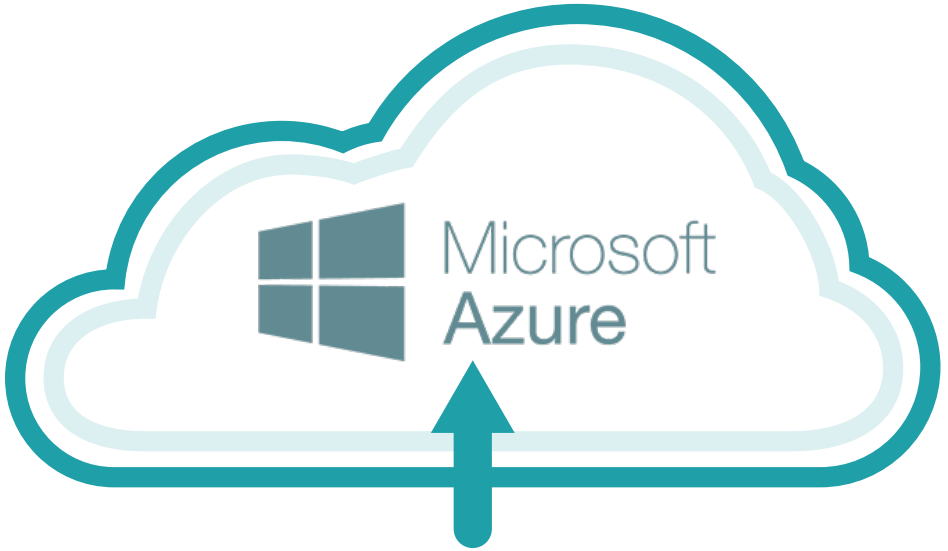
Cloud

End users

ARCHITECTURE



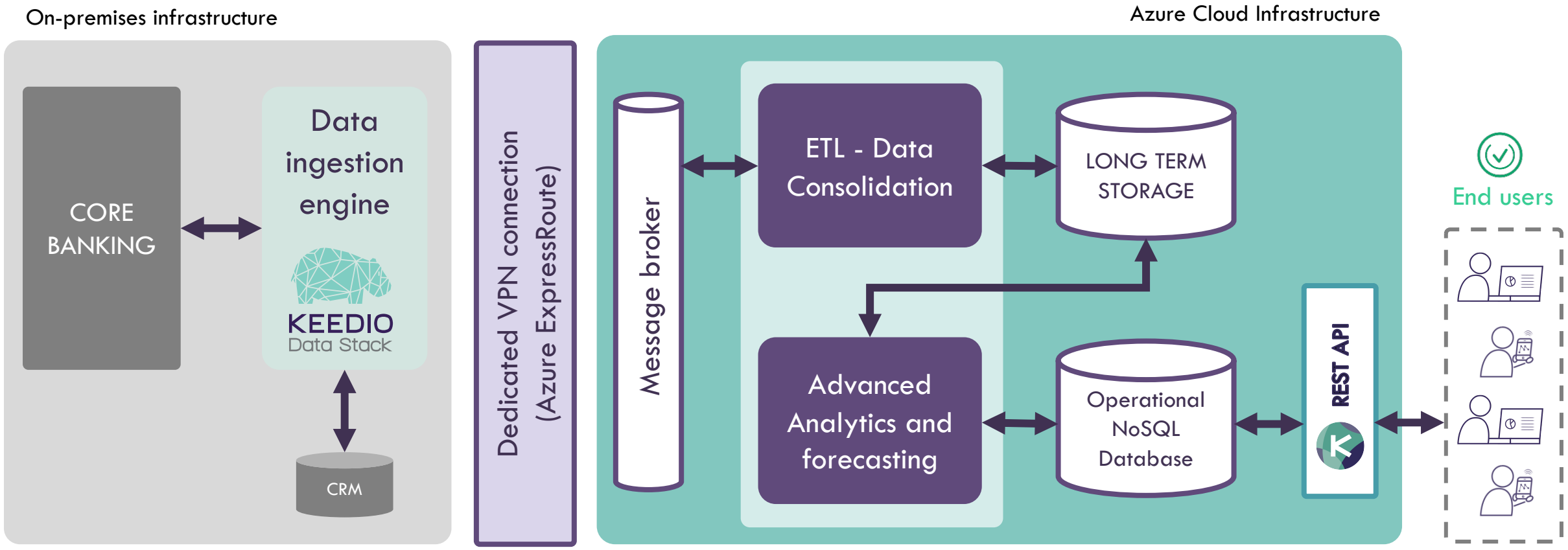
On-premises



Cloud

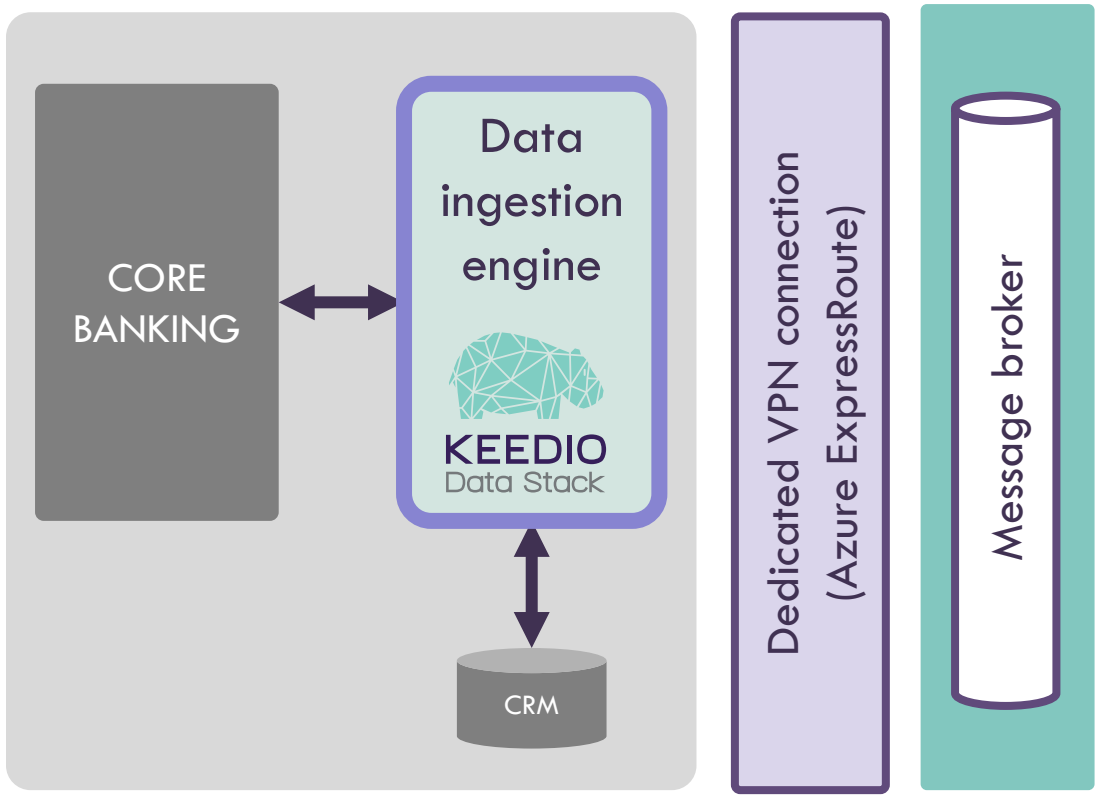
End users

ARCHITECTURE



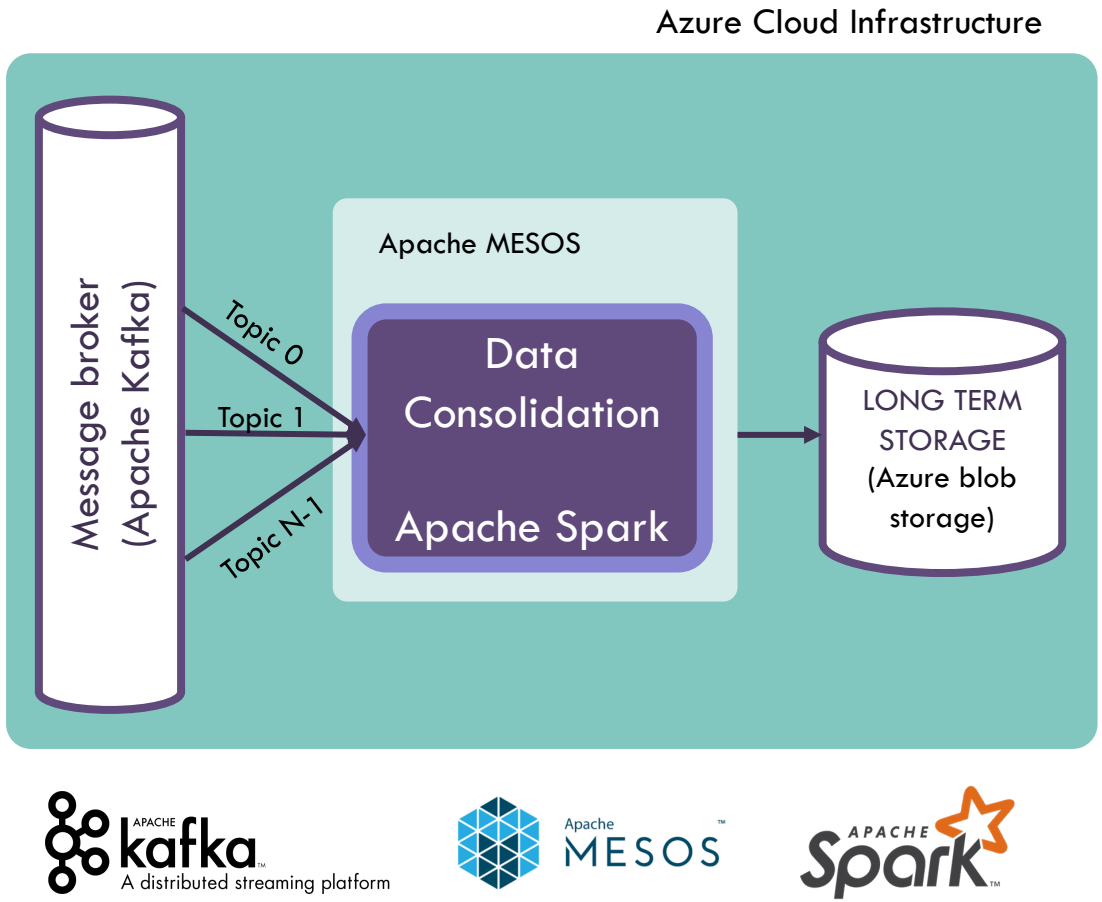
Data ingestion Engine – Legacy systems integration

On-premise infrastructure



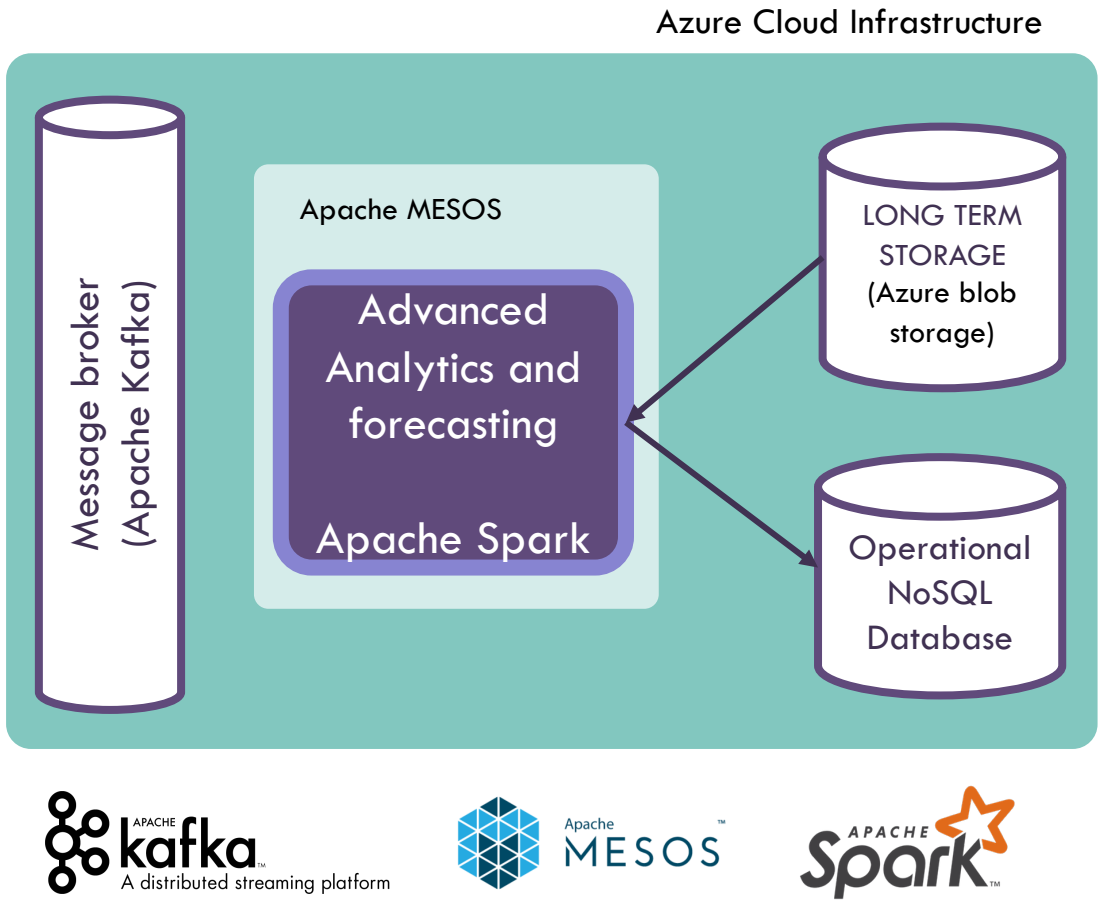
- Based on Apache NiFi
- Efficient incremental data retrieval from relational data bases.
- Data orchestration
- Data anonymization
- Efficient serialization via Apache AVRO
- Snappy compression
- Streaming-centric data ingestion layer

Data ingestion Engine – Data Consolidation



- ETL developed with Apache Spark
- Running in an Apache MESOS cluster
- Data validation and normalization.
- Complex incremental consolidation of “batch” core banking data.
- Incremental consolidation of historical data with online data.

Processing Engine



- Running in an Apache MESOS cluster
- Auto scale-out
- Auto scale-down
- Bank movement categorization through supervised learning.
- Next bill prediction
- Per-category expense forecast

Processing Engine – How do scale out and down?

- Azure VM Scale Sets trigger the scale-out, new virtual machines are provisioned

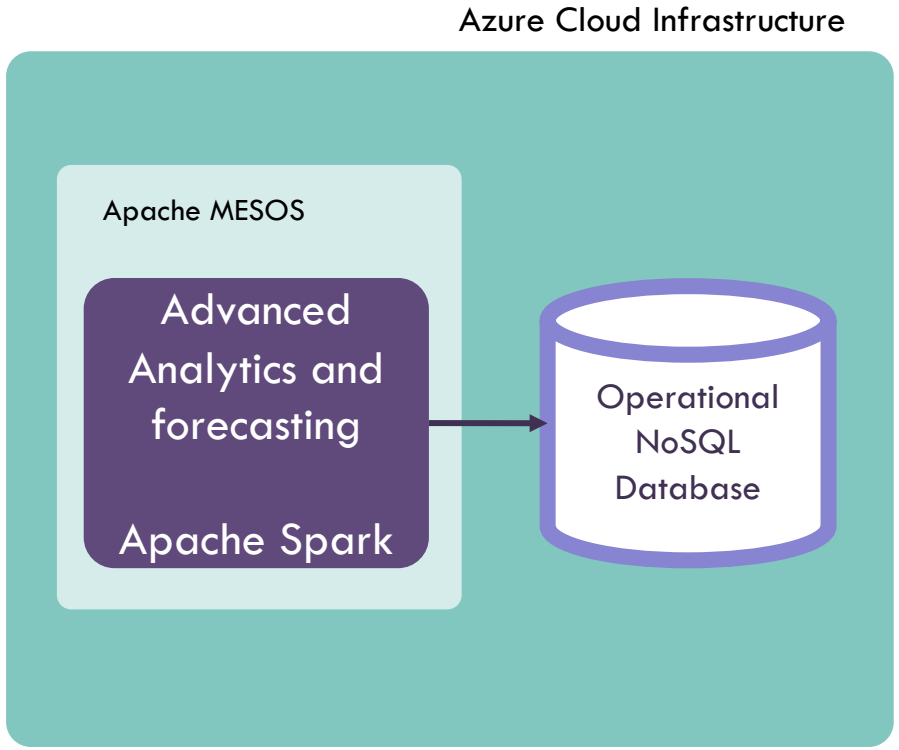
- Foreman discovers the newly provisioned VMs and applies a Host Group.

- Puppet configures and updates the newly spawned VMs.

- New hardware resources are available to Apache MESOS and Apache Spark jobs.

< 15 minutes

Operational Database

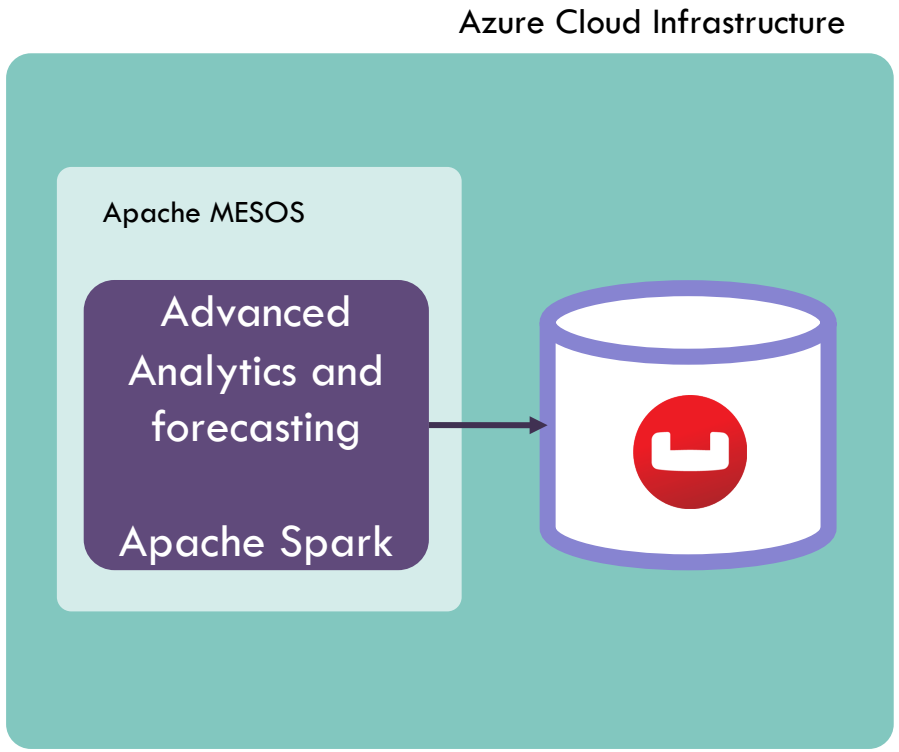


Requirements:

- Open-Source
- Document-oriented
- Data de-normalization was not an option.
- Need for a NoSQL database with advanced SQL-like interface.
- Need for a NoSQL database with advanced secondary index feature.
- Seamless data replication between different environments.
- Low latency data access.
- Production-ready.

No Apache Top level project solution matches completely these requirements.

Operational Database

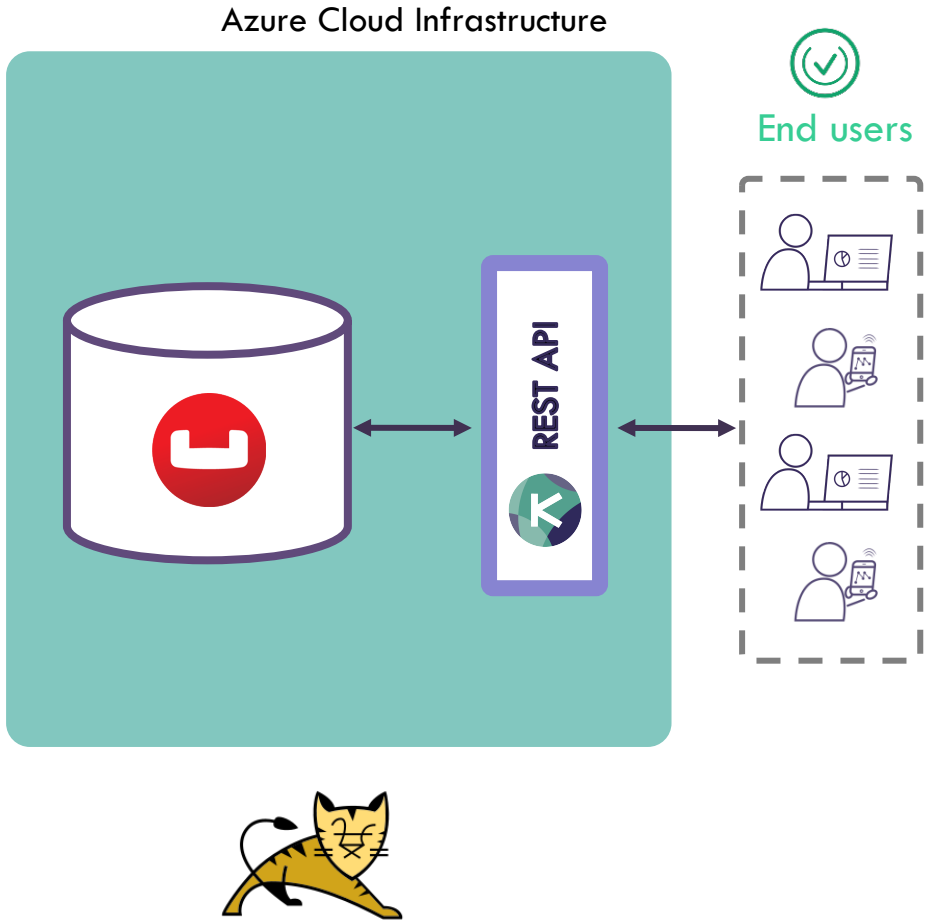


Requirements:

- Open-Source
- Document-oriented
- Data de-normalization was not an option.
- Need for a NoSQL database with advanced SQL-like interface.
- Need for a NoSQL database with advanced secondary index feature.
- Seamless data replication between different environments.
- Low latency data access.
- Production-ready.

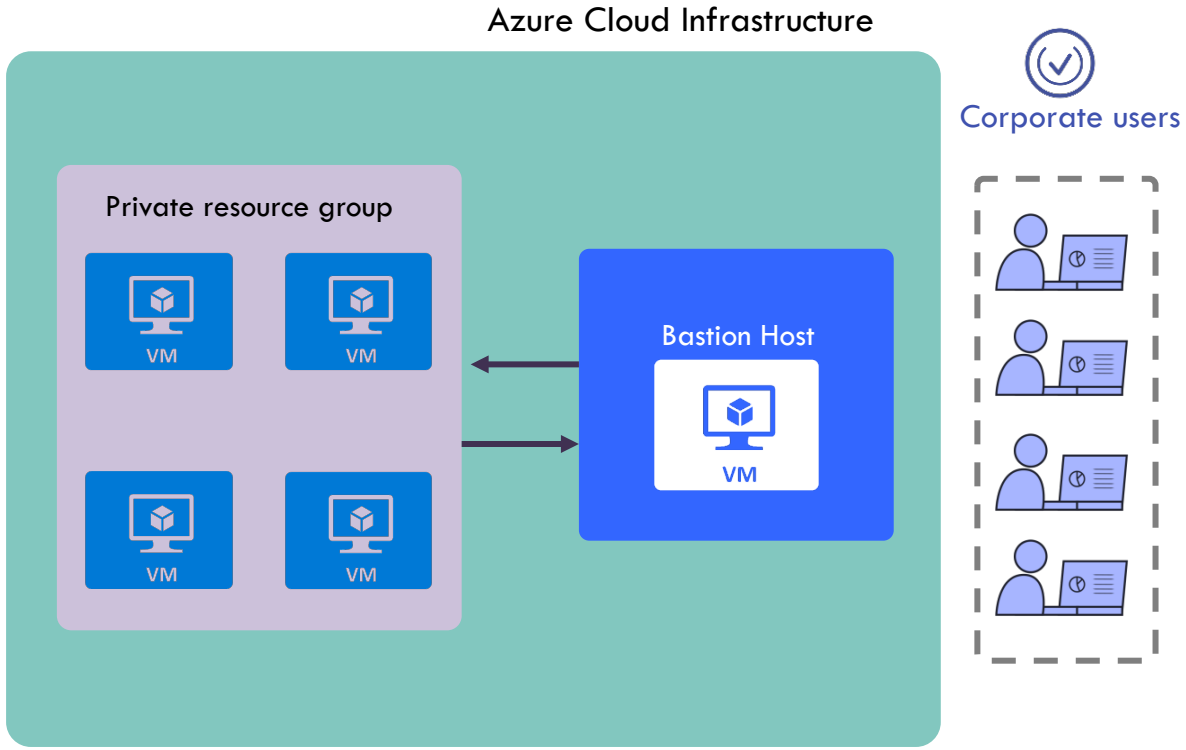
No Apache Top level project solution matches completely these requirements.

SERVING LAYER



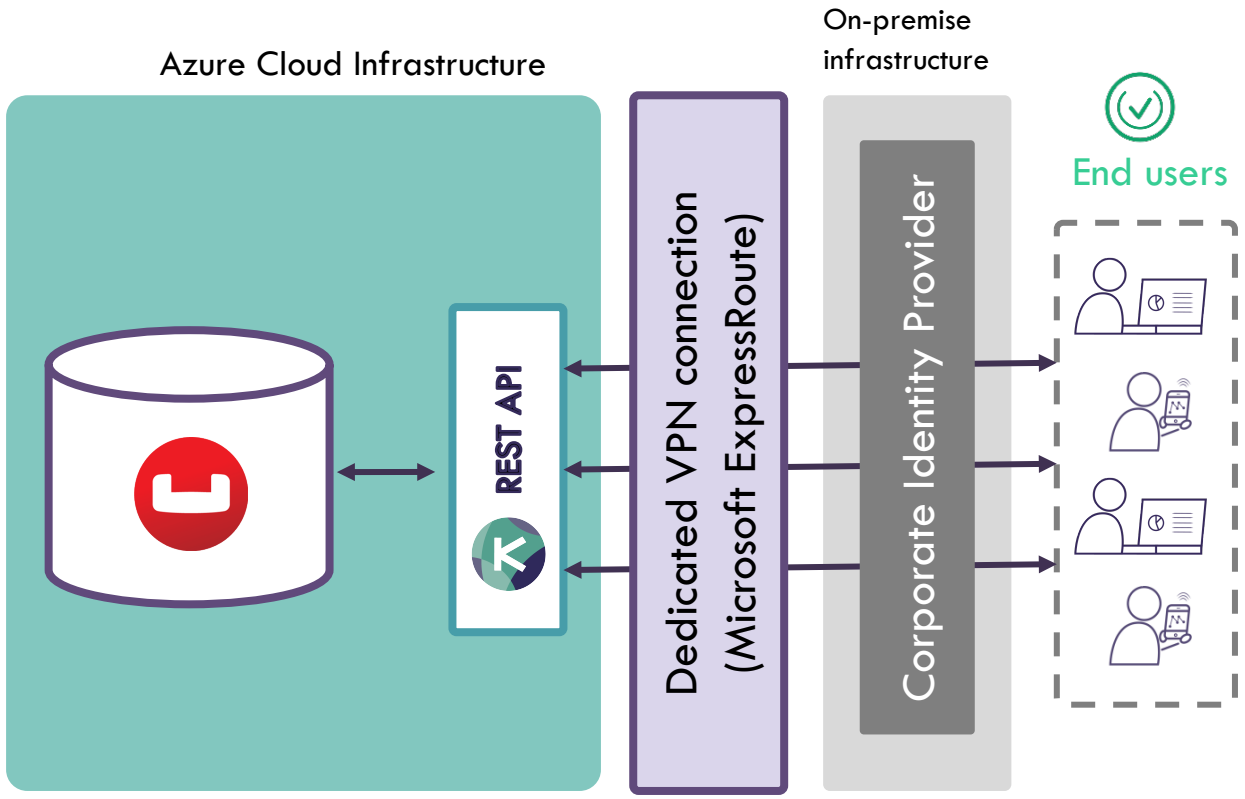
- Ad-hoc REST API
- Deployed on a scalable cluster of Apache Tomcat.

SECURITY (I)



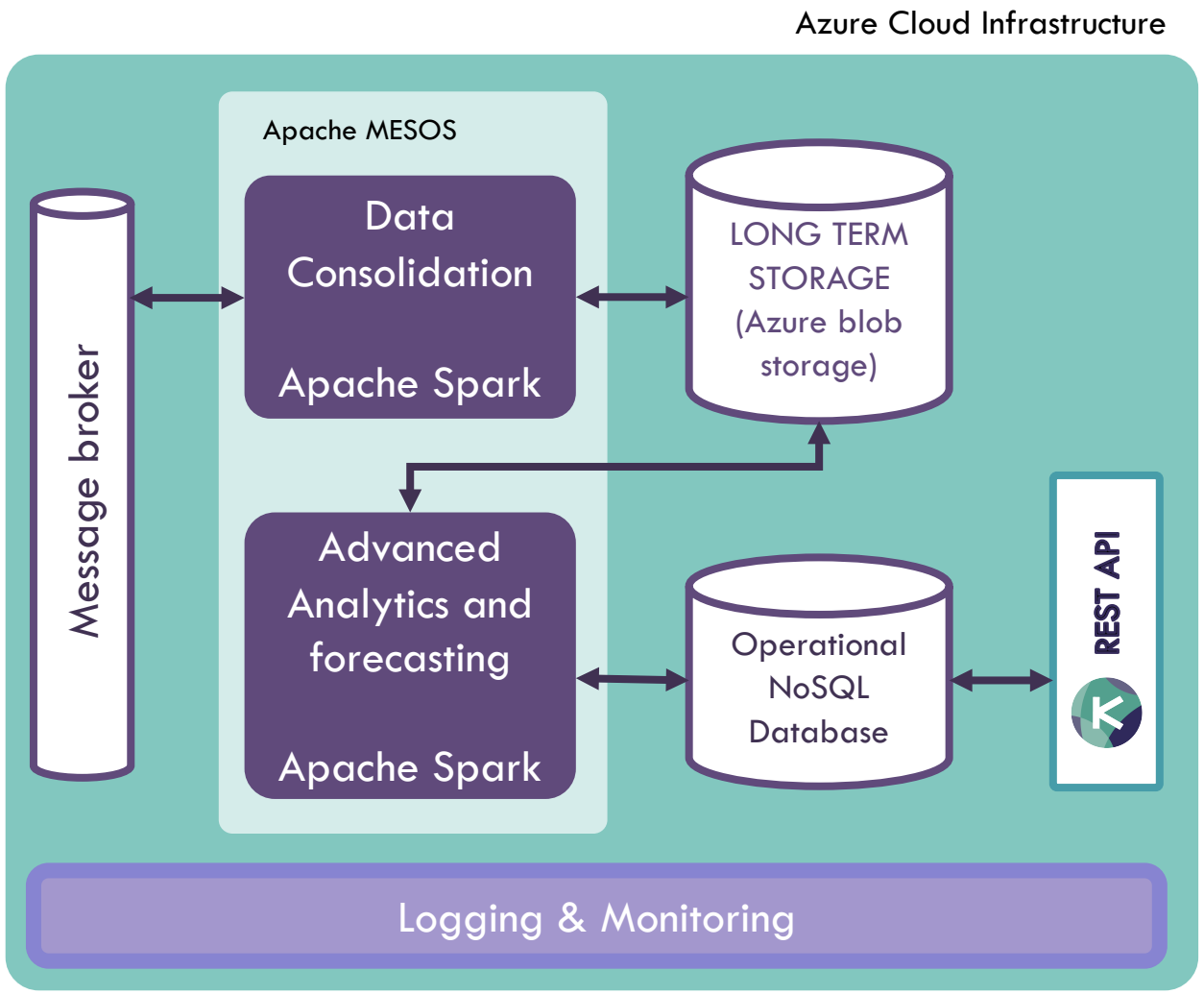
- Virtual Machines are not public
- Corporate users can only access IaaS machines through an hardened gateway.

SECURITY (II)



- Traffic passes through the corporate identity provider.
- Scalability bottleneck: all traffic must pass through the on-premise/cloud VPN:
  - Federation between the REST API and the IdP.

Logging & Monitoring



- On-premises metrics
- IaaS metrics
- PaaS metrics
- Business metrics:
  - Ingestion engine
  - Consolidation engine
  - Analytics engine
  - REST API traffic
- Email alerting on unexpected or failure scenarios.

THE OUTCOME

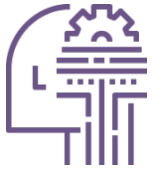
OPERATIONAL PLATFORM

The outcome (I)

Builds the foundation for the customer-centric strategy of our customer.

Platform able to enrich core banking services with advanced services to end users, 24x7.

Core banking systems are seamlessly integrated with cloud infrastructure.

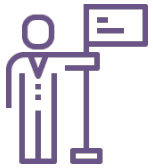


Better knowledge of customer behavior.

The outcome (II)

Infrastructure changes are quick and cheap.

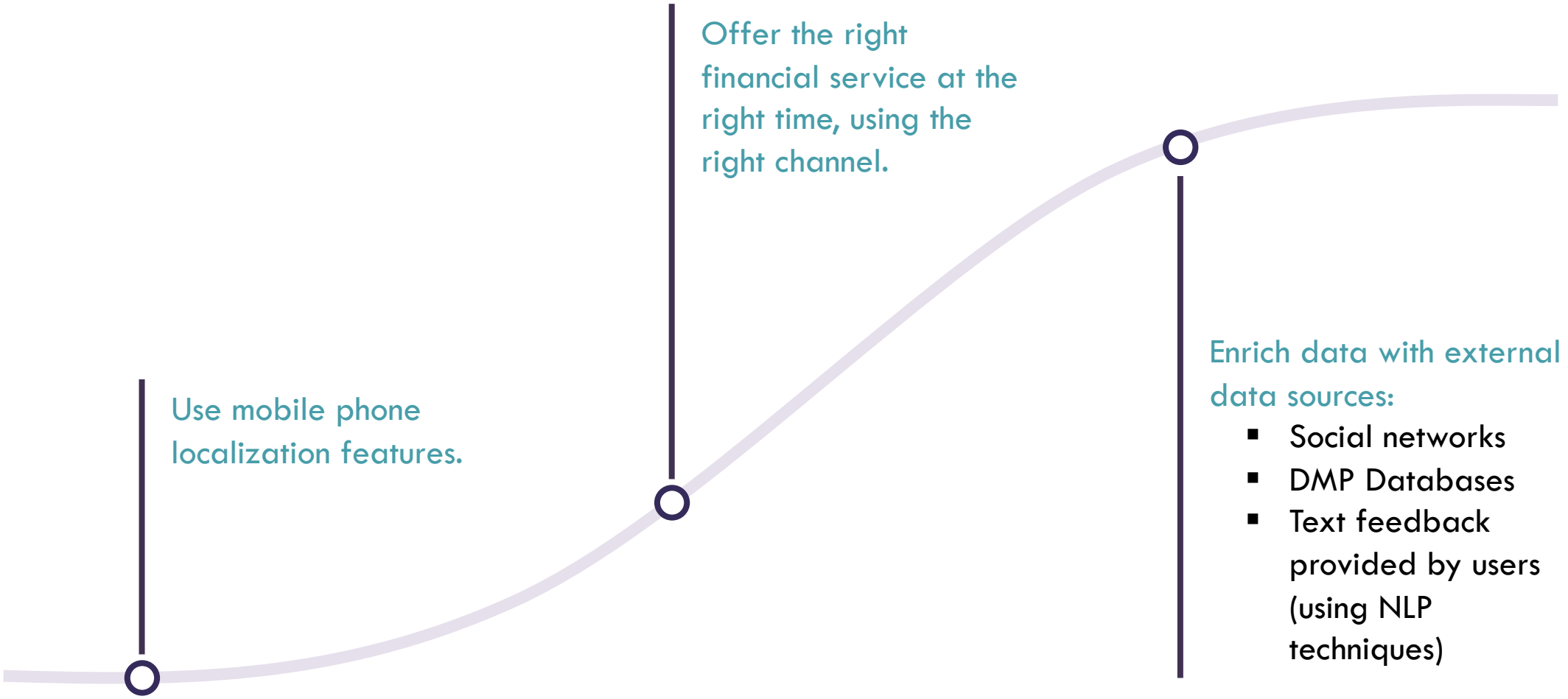
Processing power scales up and down dependening on the actual processing needs.



Data in the public cloud is anonymized as requested by national regulation.

Development and pre-production environments can be switched off on demand.

Roadmap


Use mobile phone localization features.
Offer the right financial service at the right time, using the right channel.
Enrich data with external data sources:
Social networks
DMP Databases
Text feedback provided by users (using NLP techniques)

We’re hiring!

www.keedio.com info@keedio.com @keedio keedio

Calle Virgilio 25
Edificio Ayessa I, Bajo D
Pozuelo de Alarcón
28223 Madrid

Thanks!

You have
Questions

We have
Answers