KPMG

KAVE: core platform for the Data Lake

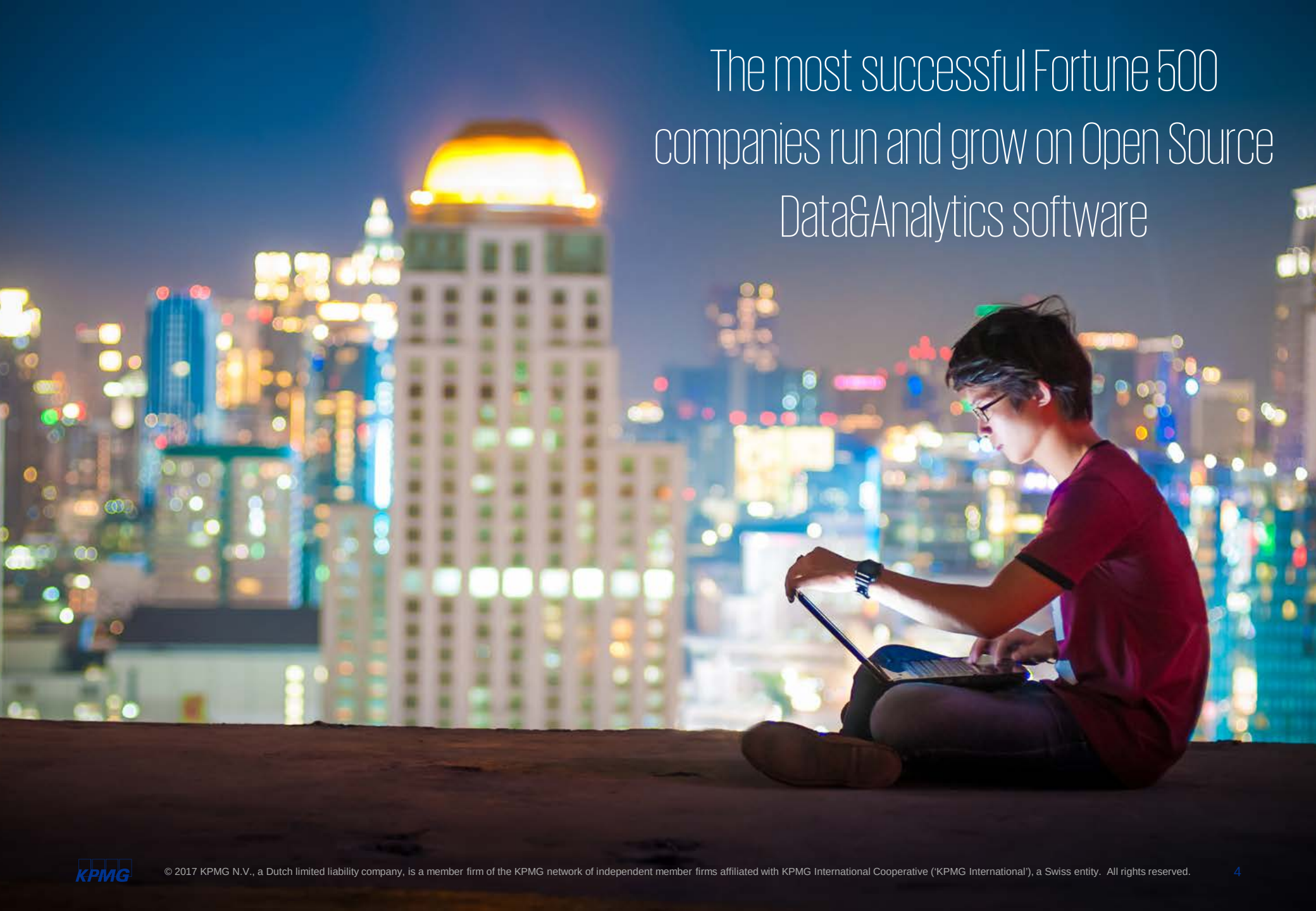
KAVE

How the KAVE offers the relevant technology to cover the use cases of a Data Lake solution

Table of contents

KPMG

Open Source for Analytics: an established yet evolving trend



OpenSource technology is empowering:

Processing 510,000 comment postings, 293,000 status updates, 136,000 photo uploads at FB, per second

An estimate 40K+ nodes cluster storing 500+ PB of data at Yahoo!

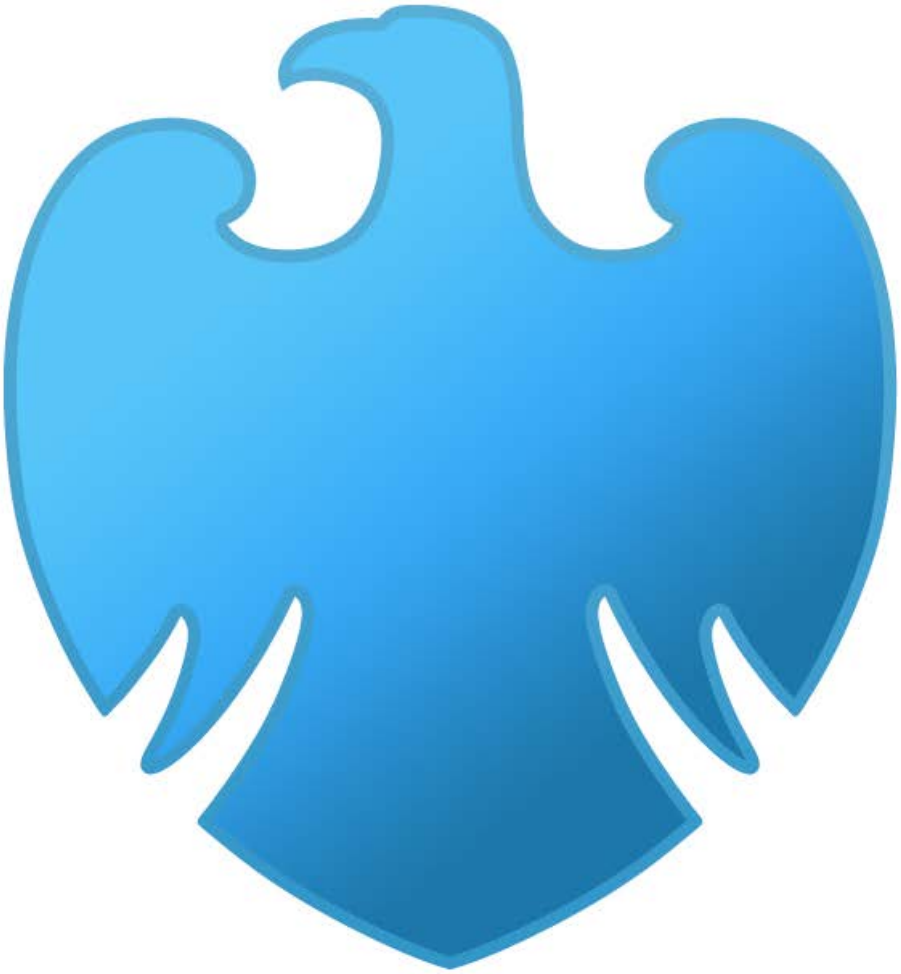
A repository of almost 1B citizens biometric data at Aadhaar India

Walmart's investment in open source is as big as they look: 6M+ $ (2016, estimate)

Google has released over 900 open source projects, totaling over 20M code lines.

Developer time spent on open source amounts to about 1B$ worth of salaries per year

Barclays claimed to have cut costs up to 90% in the last five years by adopting opensource for its cloud strategy



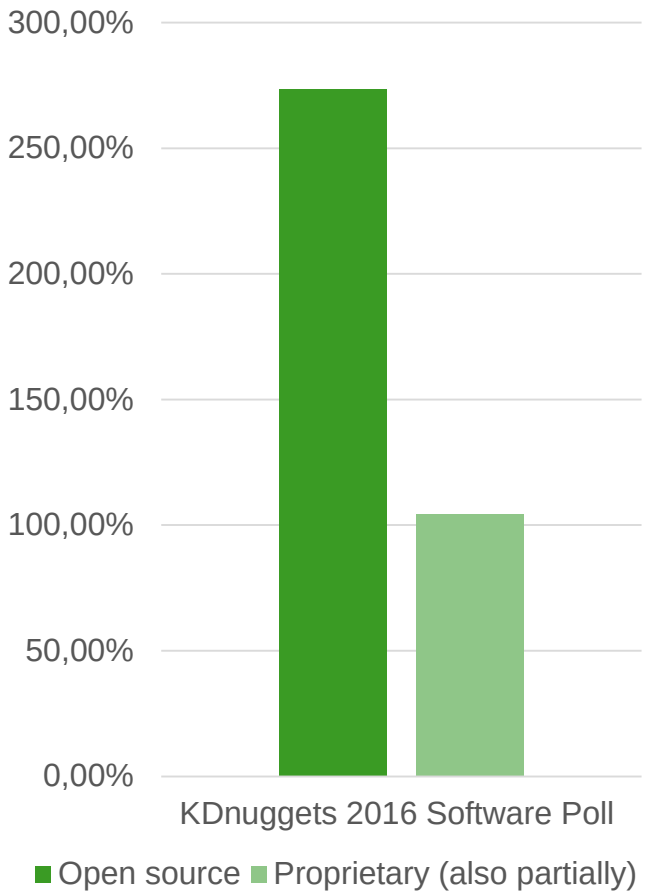
BARCLAYS

KPMG

Closed-source D&A: drawbacks & risks

Less efficiency & flexibility for data exploration: analysts' tools & techniques are too different

Data analytics software usage & relevance growth: Open vs proprietary



Lock-in solution: cannot easily integrate, customize and migrate

Federal Source Code Policy: 20% minimum of newly developed software released as open source: encourage usability, prevent lock-in

Fall-behind: cannot introduce state-of-art techniques or redesign

No agility: rigid licensing and scaling modeling, difficult market reaction & evolution

"In prior eras industry players lacking technical competence outsourced the job [...], game changes detemined innovation was not coming from there, and even if it did, licensing would be non-starter in scale-out environments" S. O'Grady

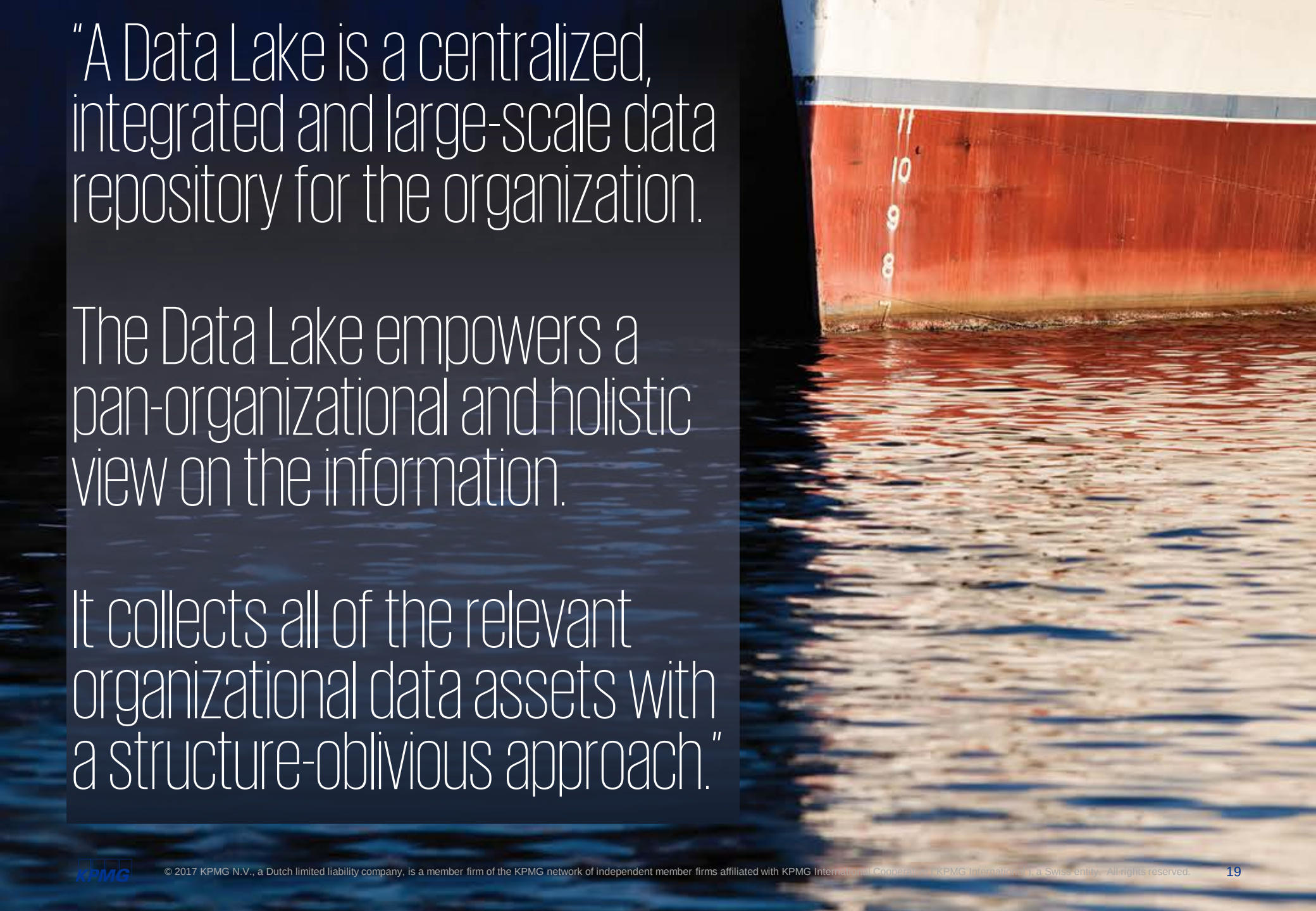
KPMG

Data Lakes

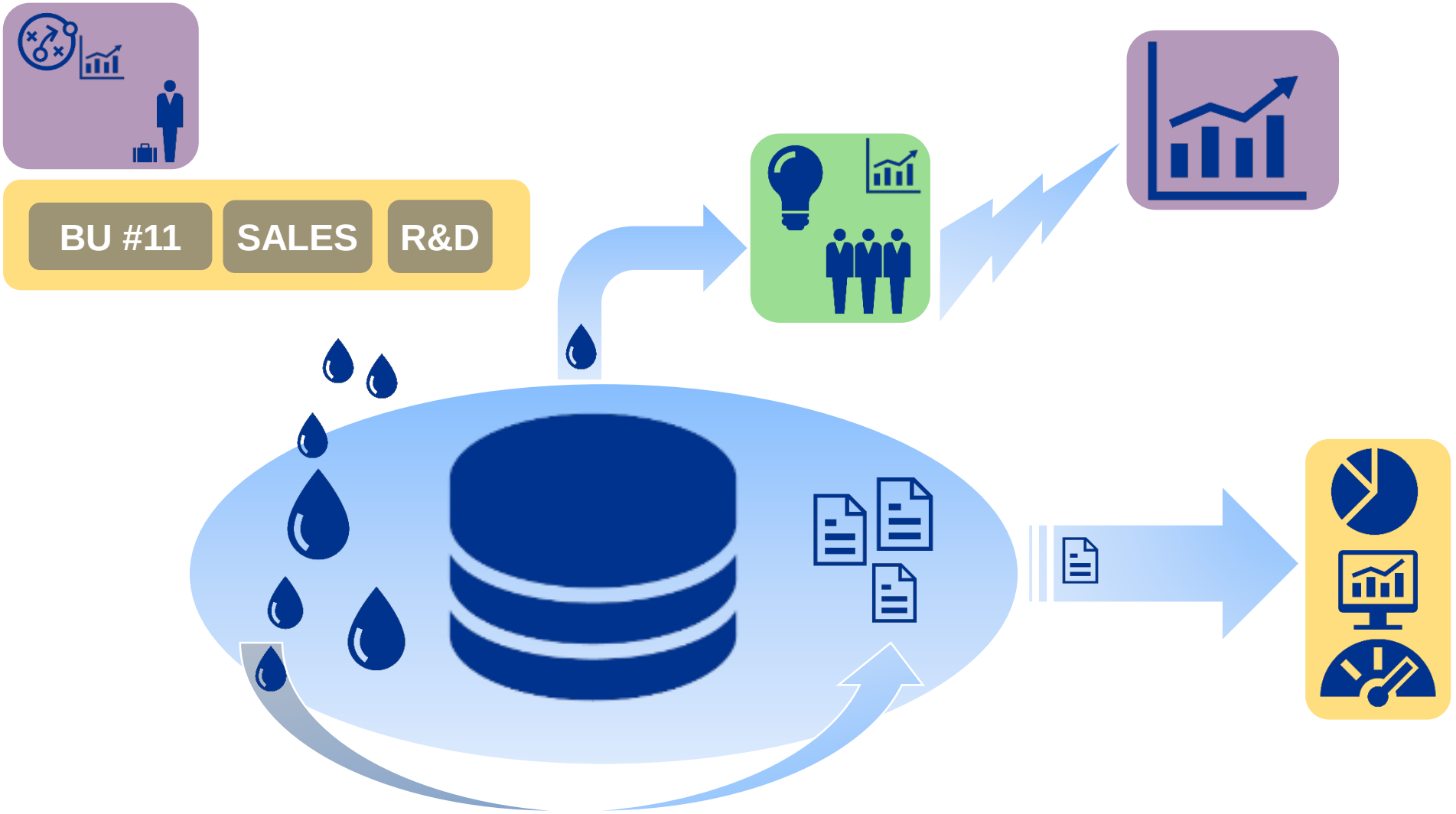
"A Data Lake is a centralized, integrated and large-scale data repository for the organization.

The Data Lake empowers a pan-organizational and holistic view on the information.

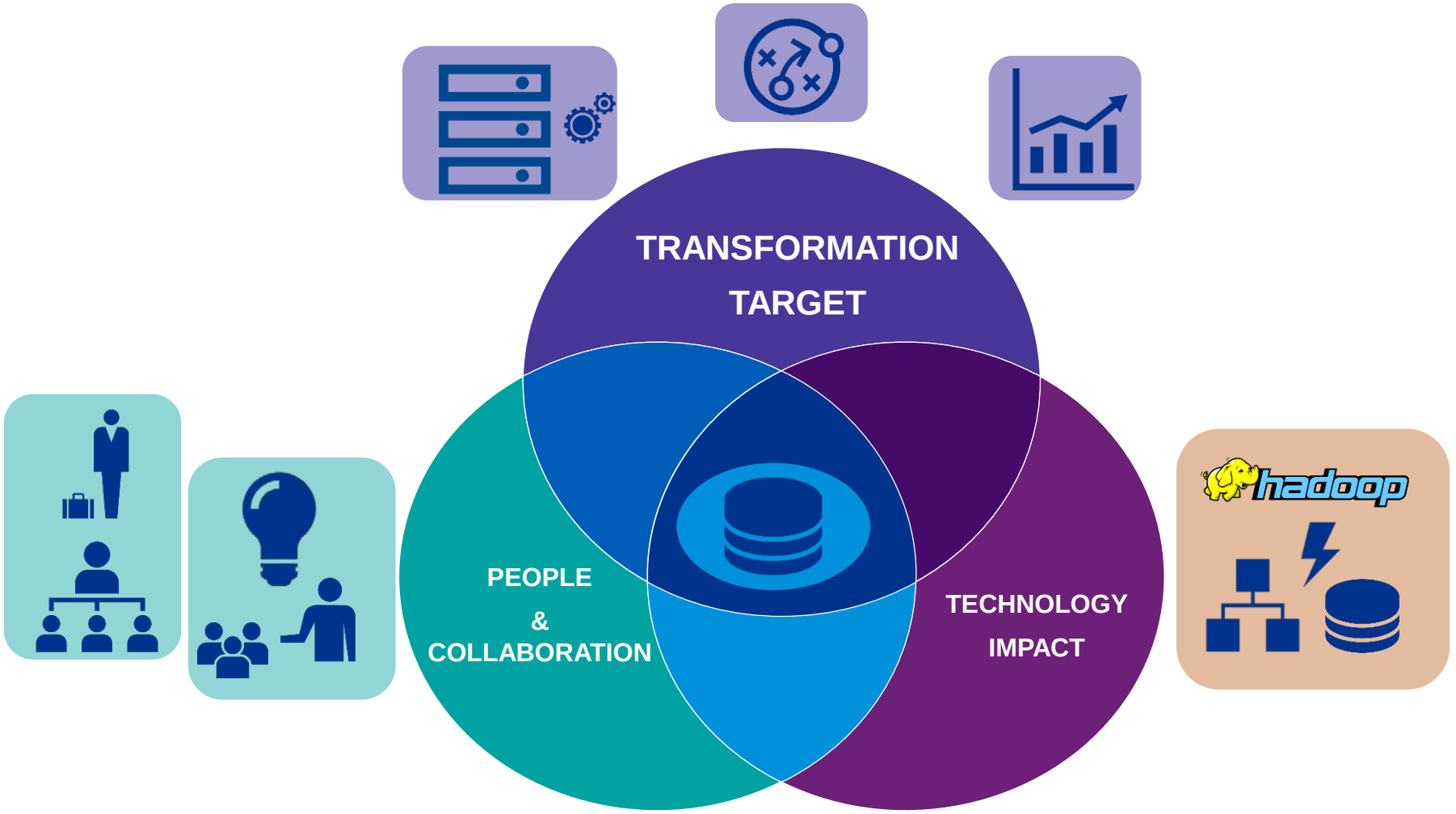
It collects all of the relevant organizational data assets with a structure-oblivious approach."



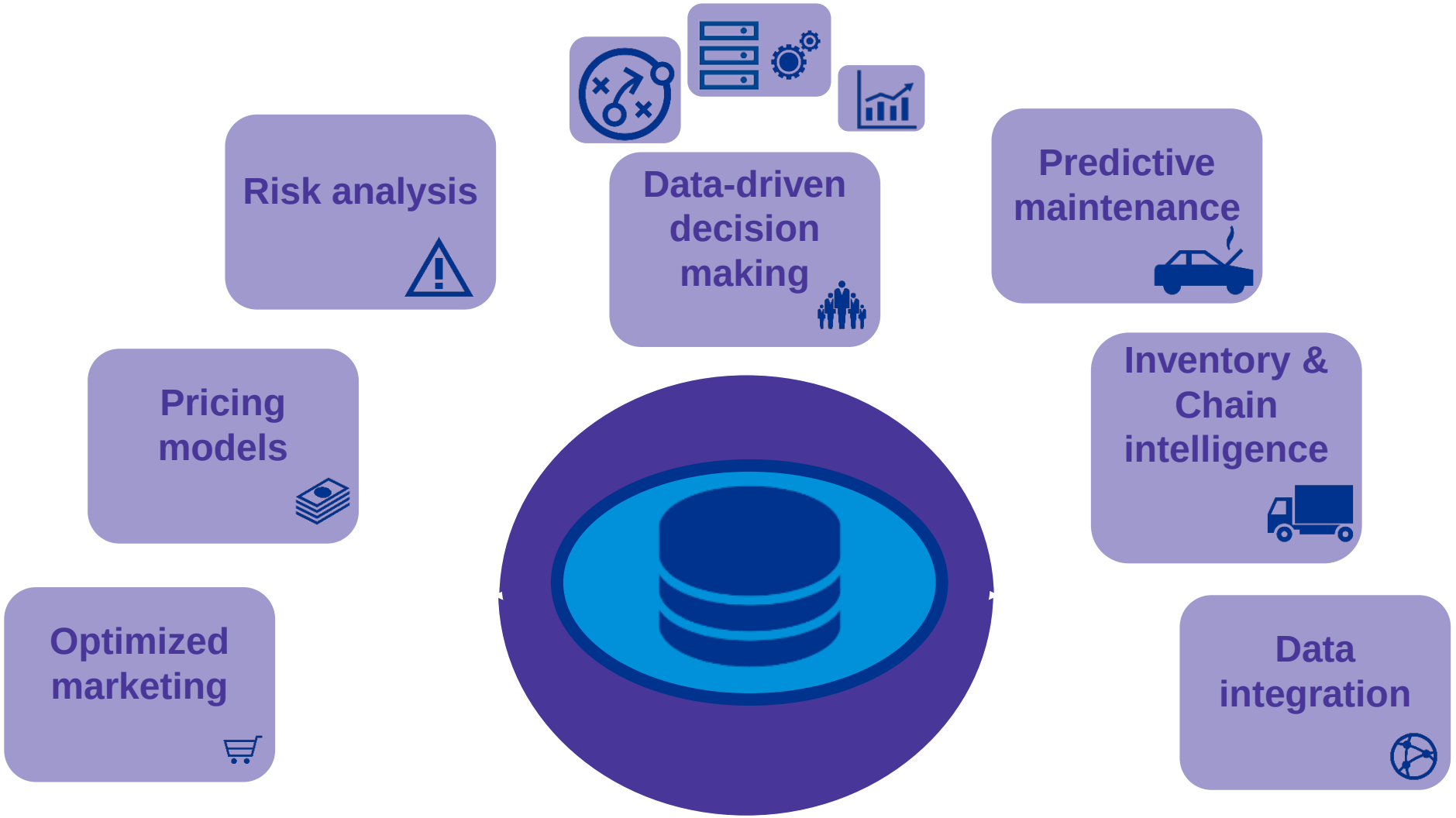
The Data Cycle



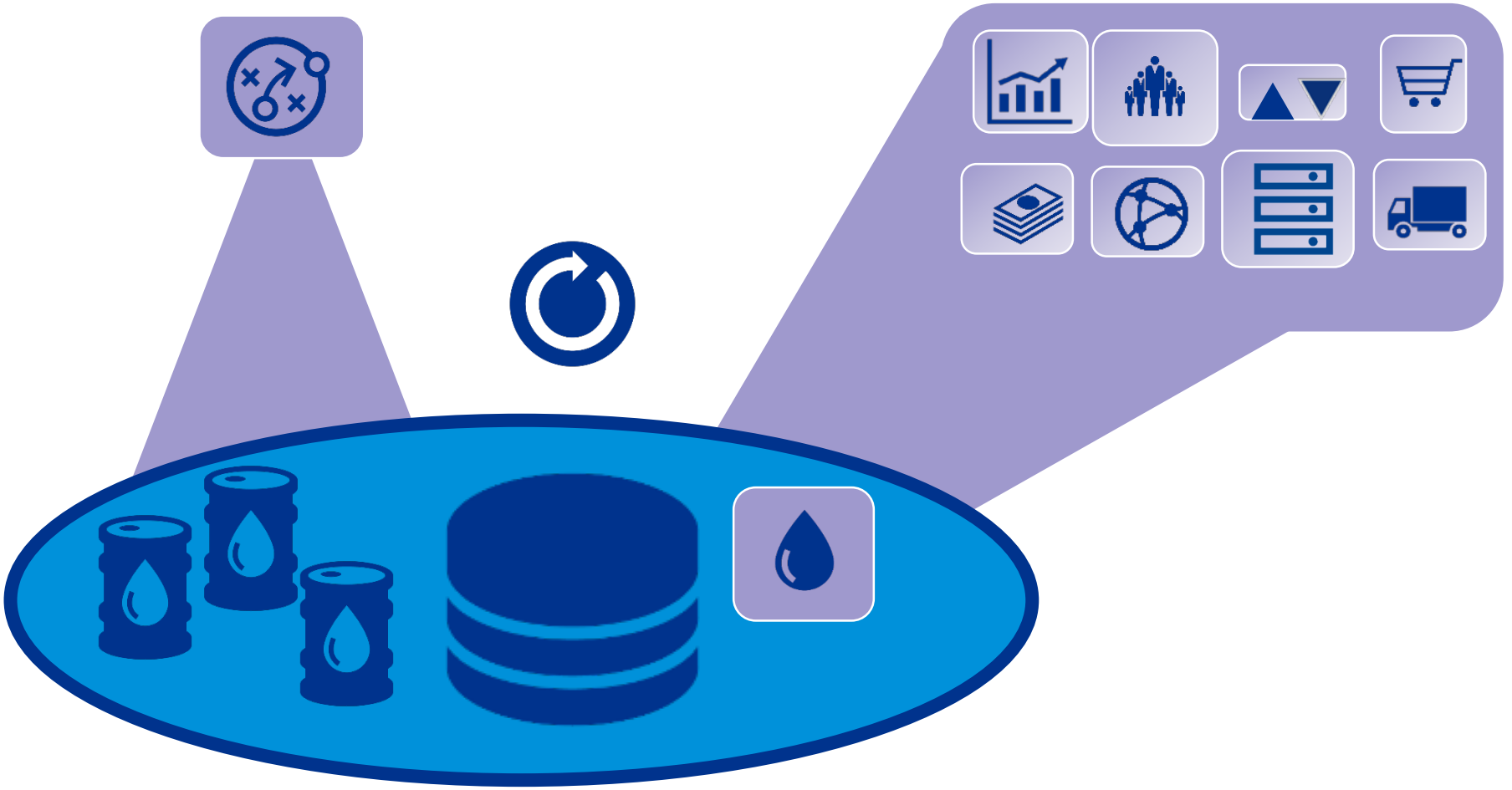
The Data Lake: driving the analytics evolution



The Data Lake: analytics processes & new strategies

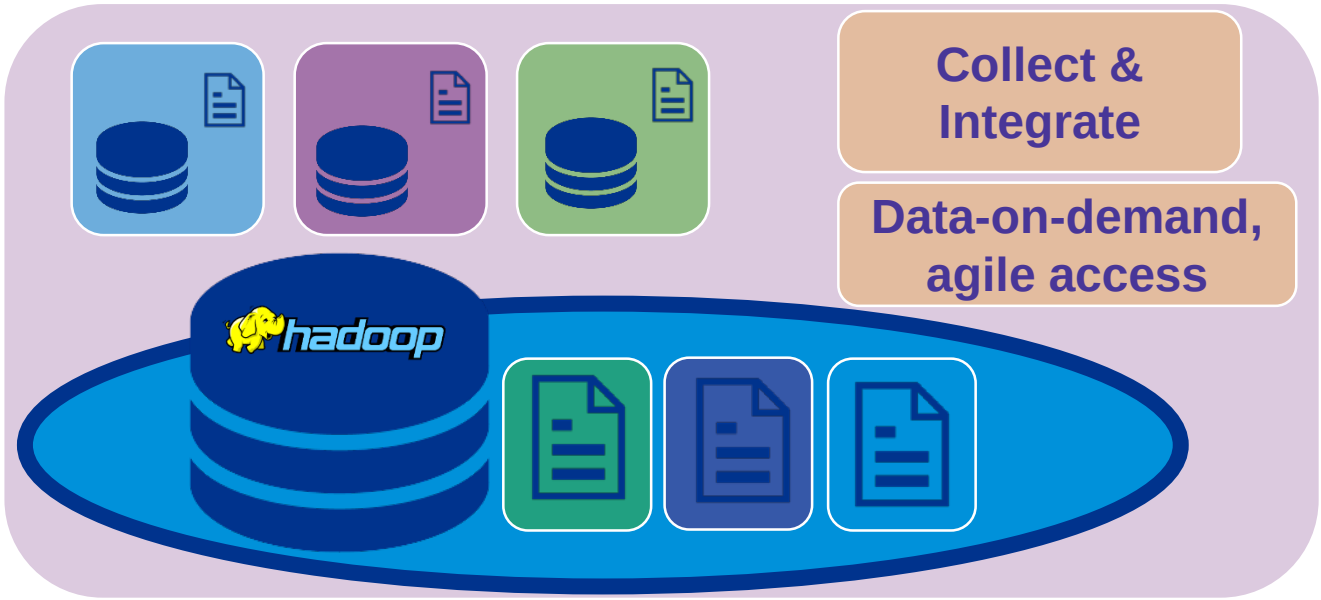


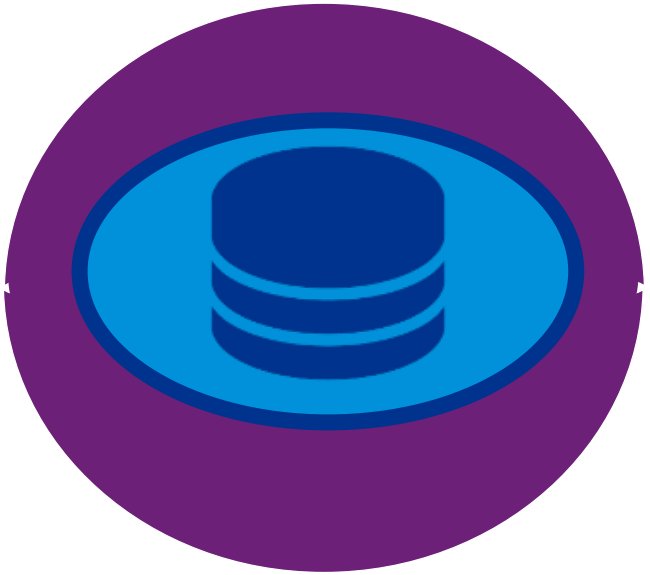
Data: not a by-product but a source of value

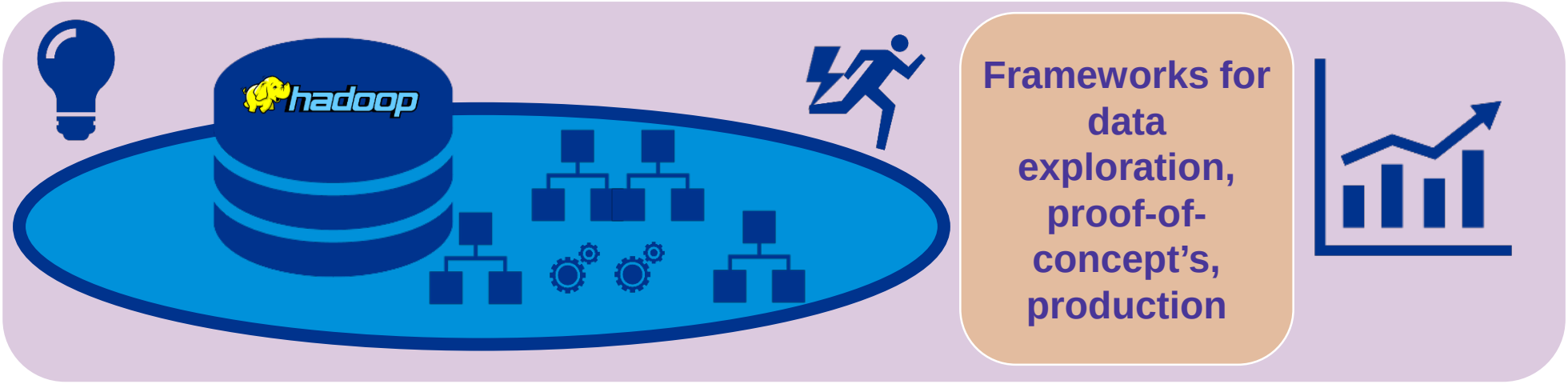


Enterprise Data Lake: analytics-driven organization

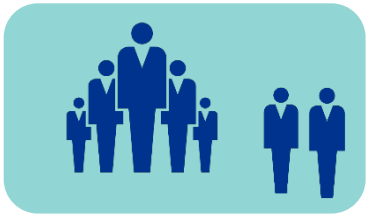
Data analytics: make value out of data





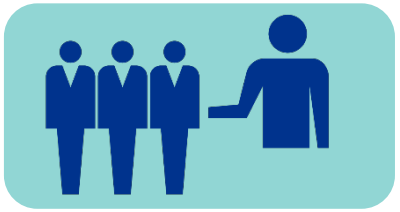


Focus: people



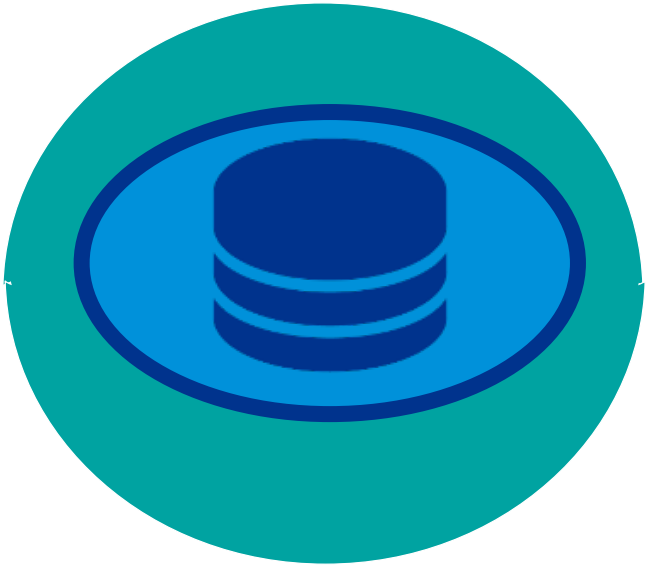
Not just tech-trend, real value for CIOs

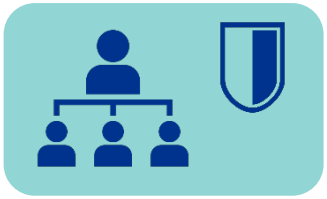
Main reference for CDOs



Enhanced customer experience, ad-hoc scenarios

Valorize your analysts team, attract new talent





Comply to the organization structure with respect to data

KPMG

What is the KAVE?

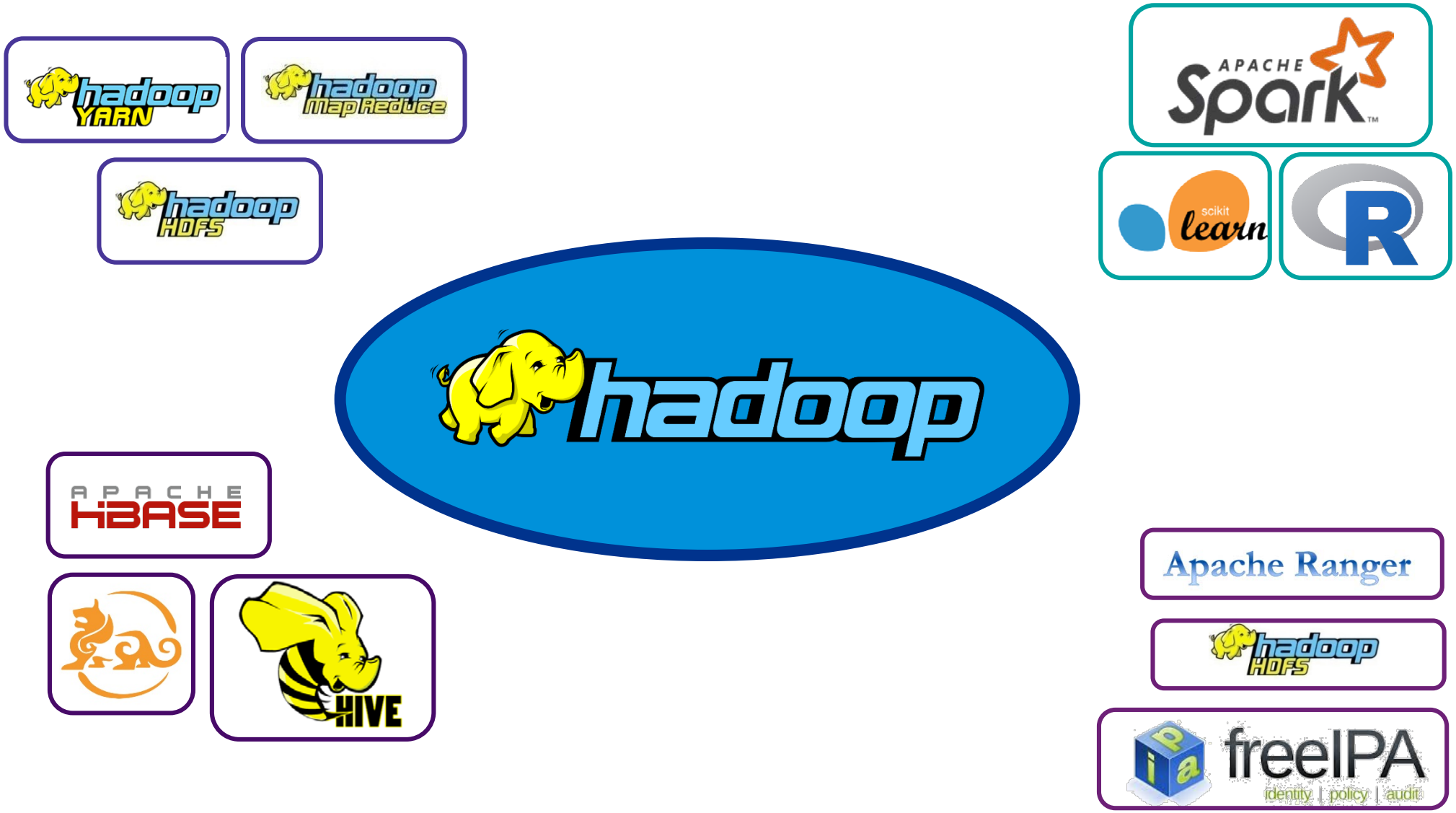
KAVE: extension of the HortonWorks Hadoop distribution

KAVE extension

HortonWorks Data Platform distribution

Hadoop core software

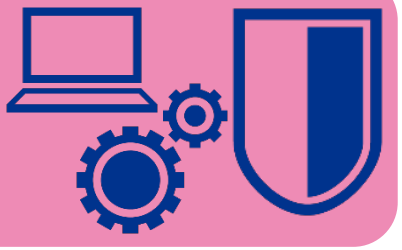
Data Lakes established technology ecosystem: Hadoop



Data Lakes established technology ecosystem: Hadoop

De-facto industry standard for Big Data

Modern architecture & service



hadoop

intel NTT DATA Y! IBM Microsoft

Opensource:

- **Free, no license cost**
- **OK commercial products**
- **Customizable - no lock-in**
- **Professional support**

KAVE: extension of the HortonWorks Hadoop distribution

Hortonworks Data Platform distribution:

- **Standard installation, partially automated**
- **Additional software (management, monitoring,…)**
- **Vendor solution: global tech support**

KAVE: extension of the HortonWorks Hadoop distribution

Data exploration & analysis

Collaboration & Development

BI/visualization integration

Web interfaces

Integrated security layer

Automated installation on Microsoft Azure

KAVE: extension of the HortonWorks Hadoop distribution

scikit learn
APACHE Spark
ROOT
GitLab
collaborate with TWiki
Zeppelin
BI
WildFly
LCM
DS
freeIPA
identity | policy | audit
Microsoft Azure

KAVE: extension of the HortonWorks Hadoop distribution

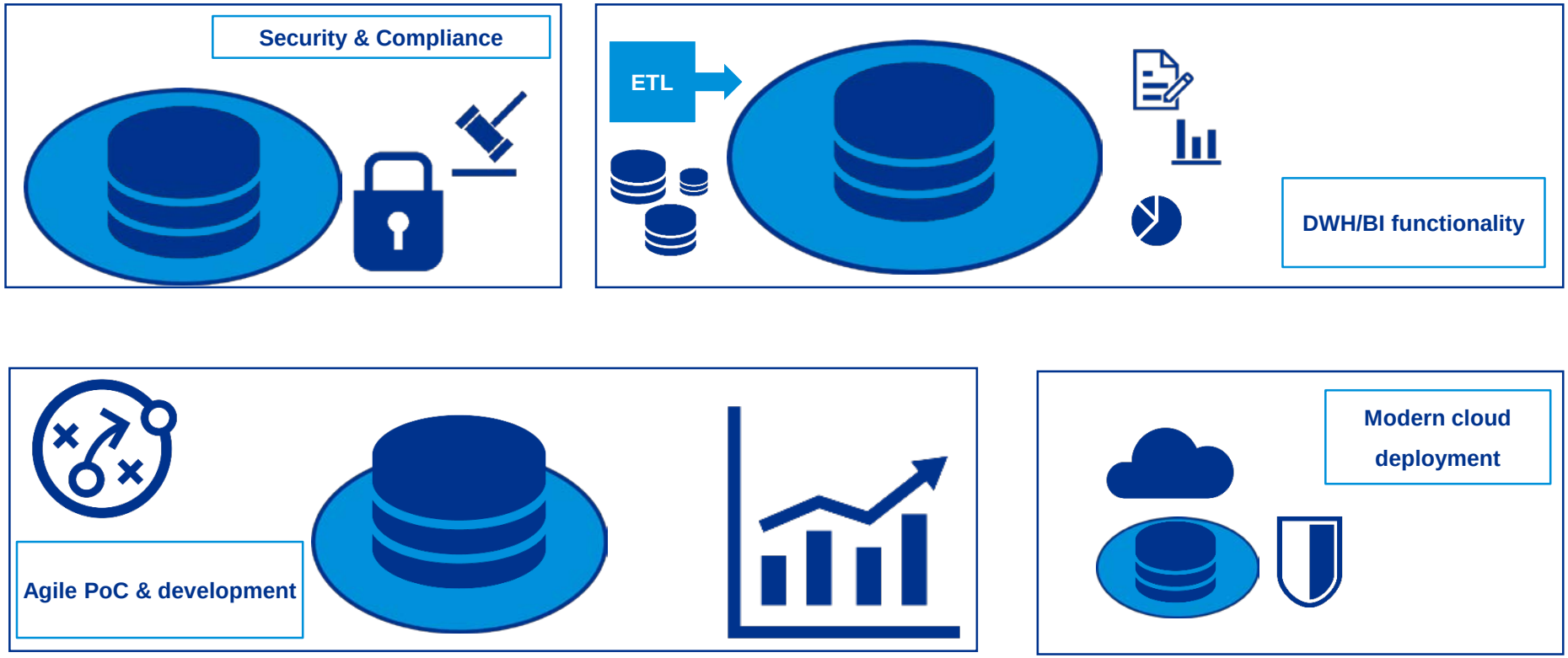
scikit learn
GitLab
Zeppelin
WildFly
APACHE Spark
collaborate with TWiki
JBoss
ROOT

Continuous improvement, up-to-date with Data Lake & Analytics technology

KPMG

KAVE & the fulfillment of the Data Lake evolution

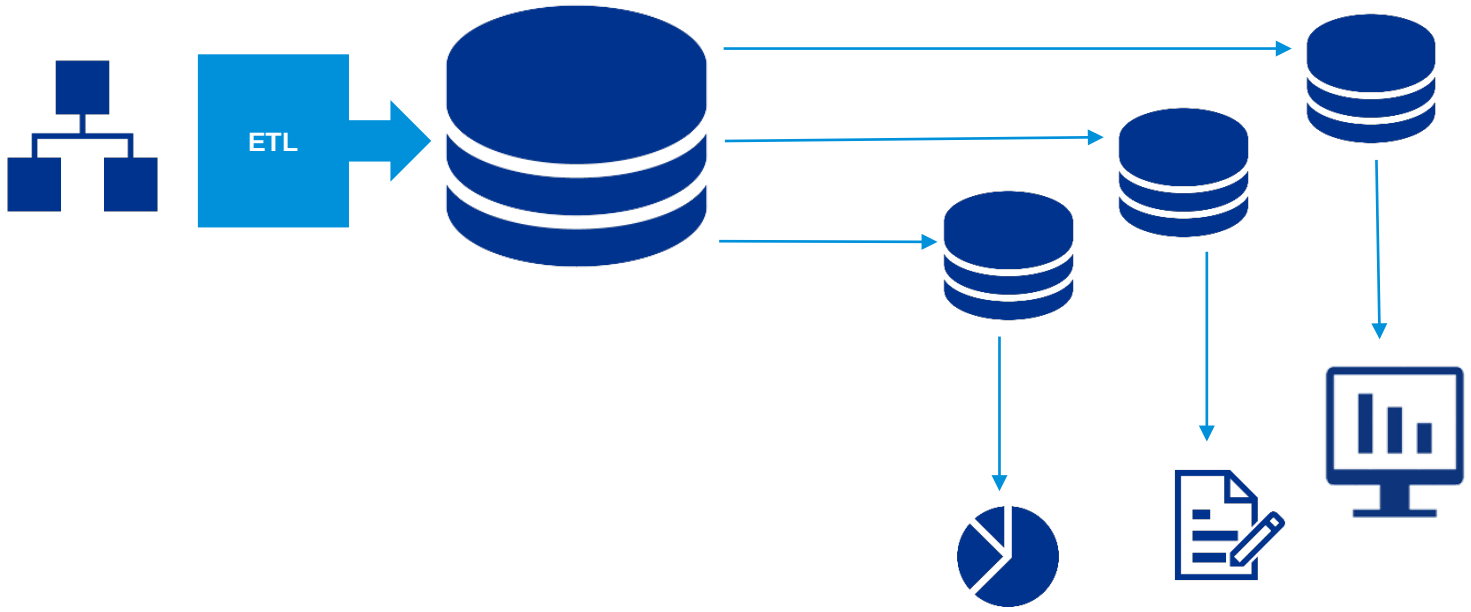
Enterprise Data Lake: topics & directions



KPMG

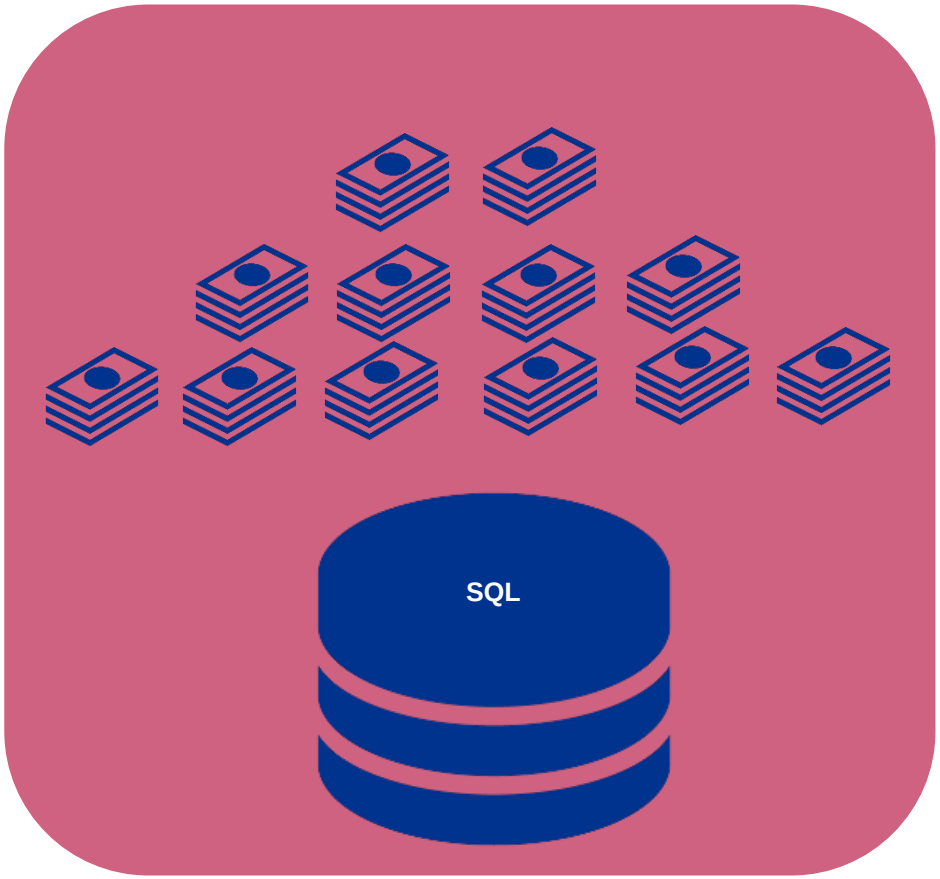
Data Warehouse & Business Intelligence functionalities

The traditional DWH/BI stack

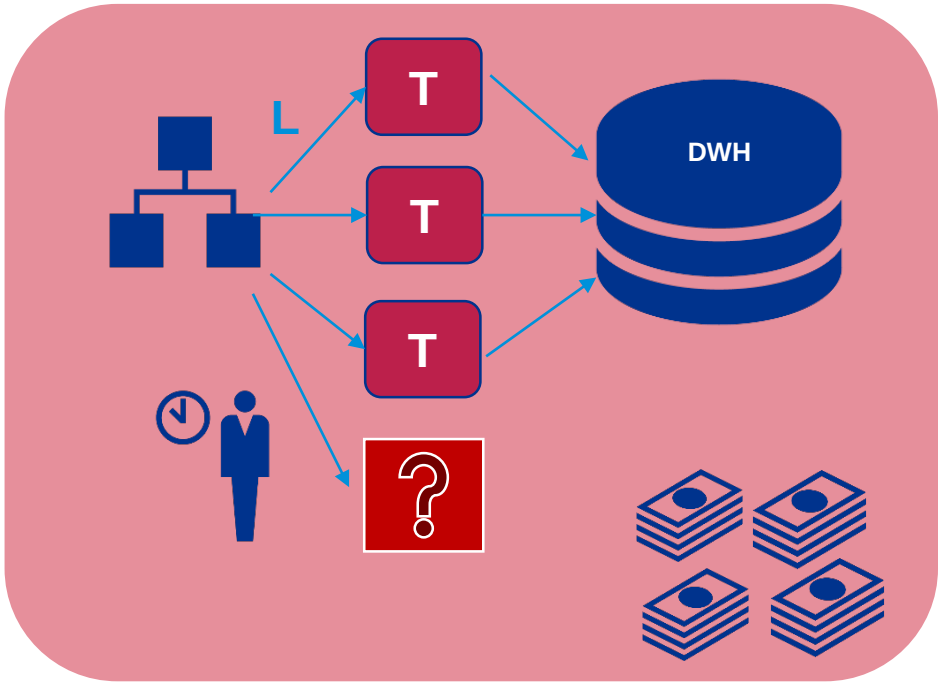


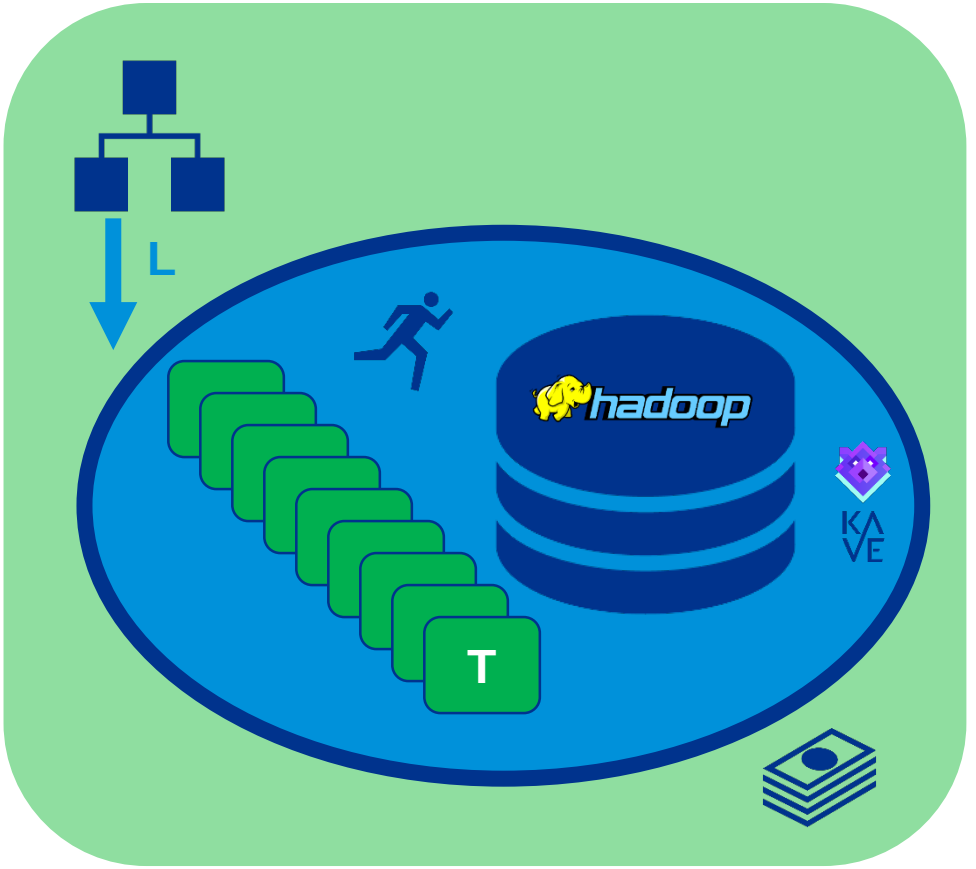
Traditional DWH/BI stack: capacity scale

- **Costs ?**
- **Performance ?**
- **SQL-only ?**

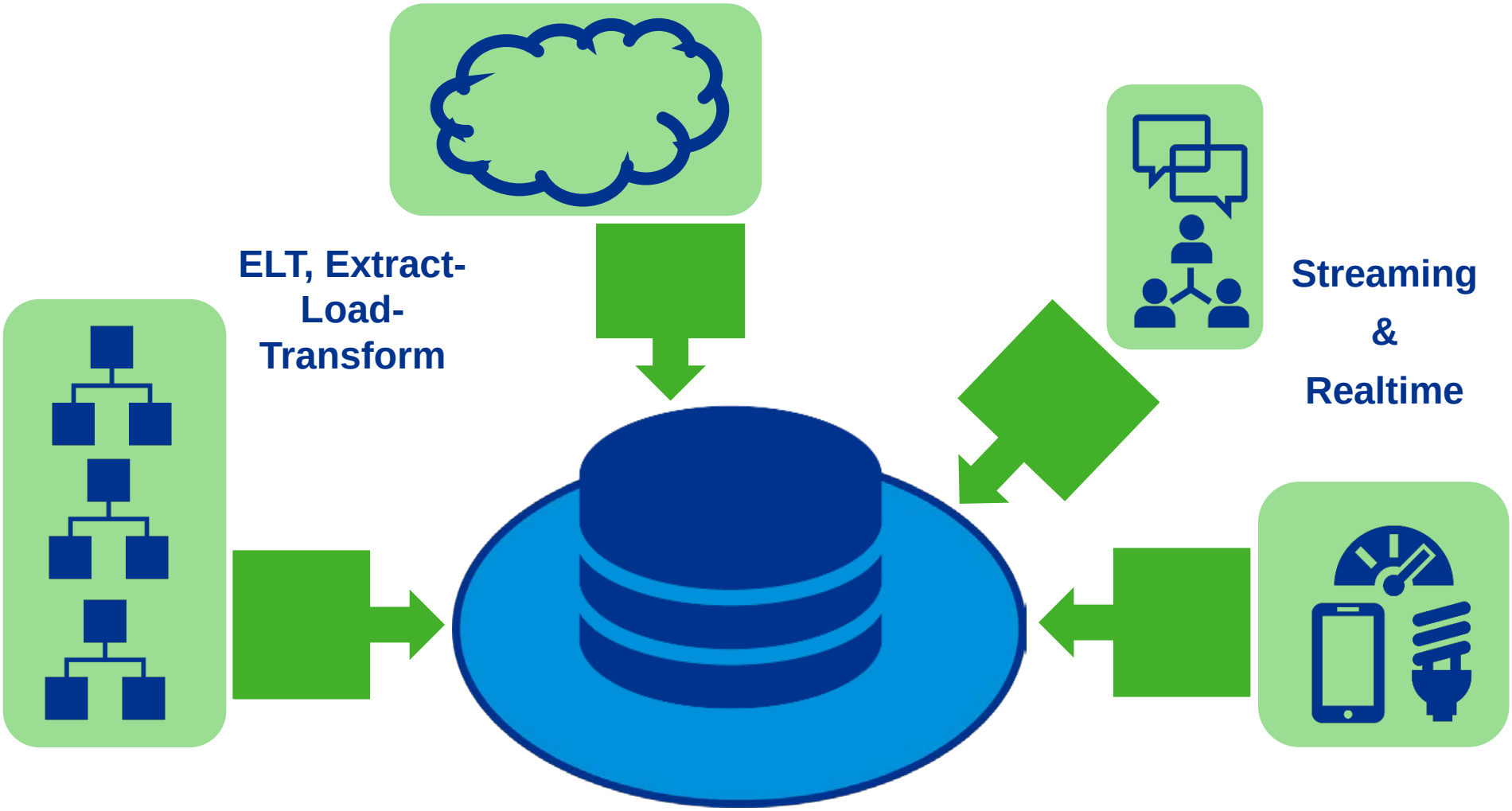


Enterprise Data Lake: ELT scaling in KAVE

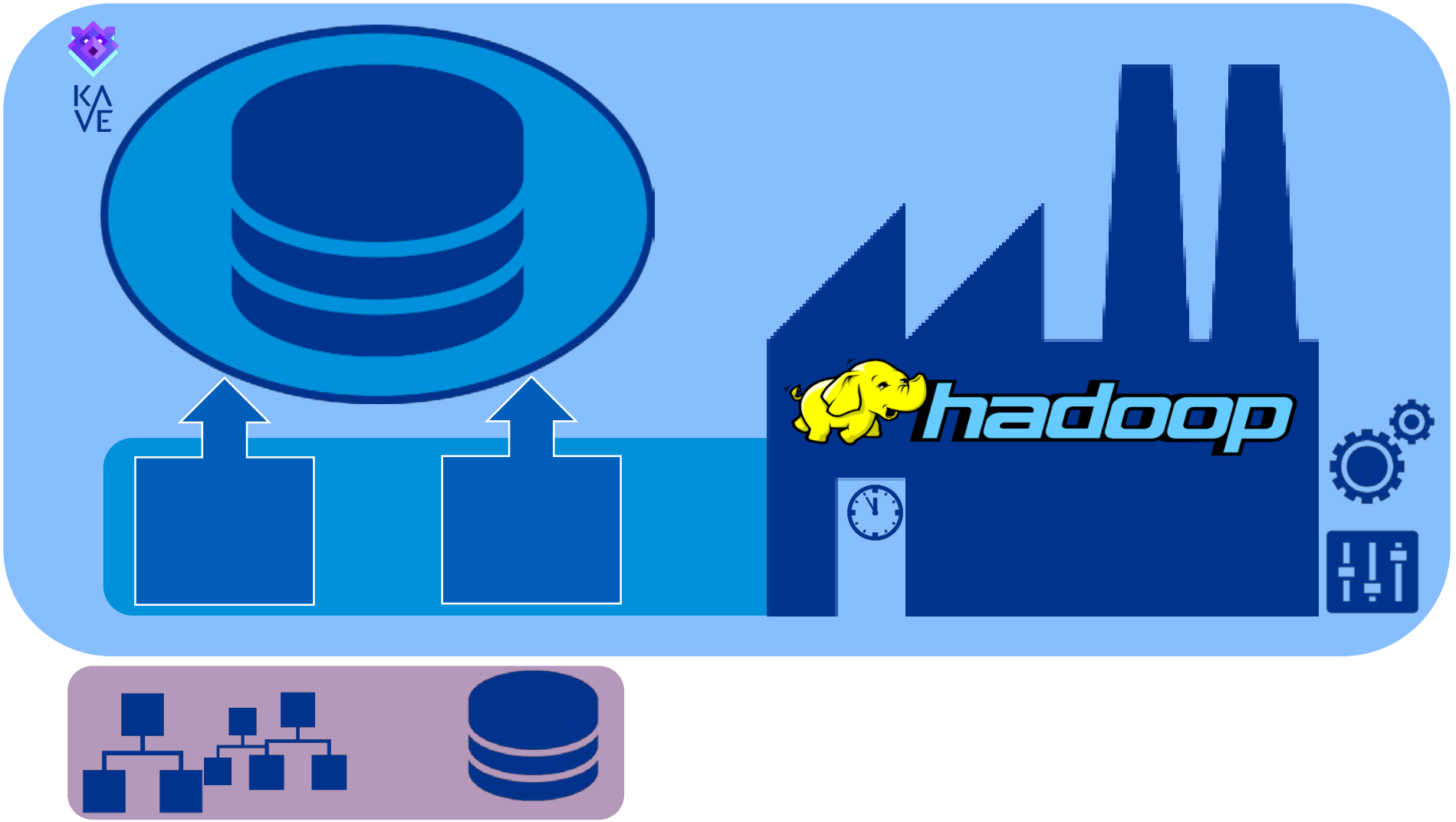




Enterprise Data Lake: evolution of the DWH/BI stack



KAVE: fully-automated ETL facilities



KAVE: fully-automated ETL facilities

APACHE FALCON

Define, schedule and manage ETL pipelines in a graphical way

Ad-hoc RDBMS import (Oracle, Postgres, MySQL…)

Build pipelines of any complexity for the best transformation strategy

Seamless import of heteogeneous data sources (logs, queues, files, webpages…)

KAVE

Apache Atlas

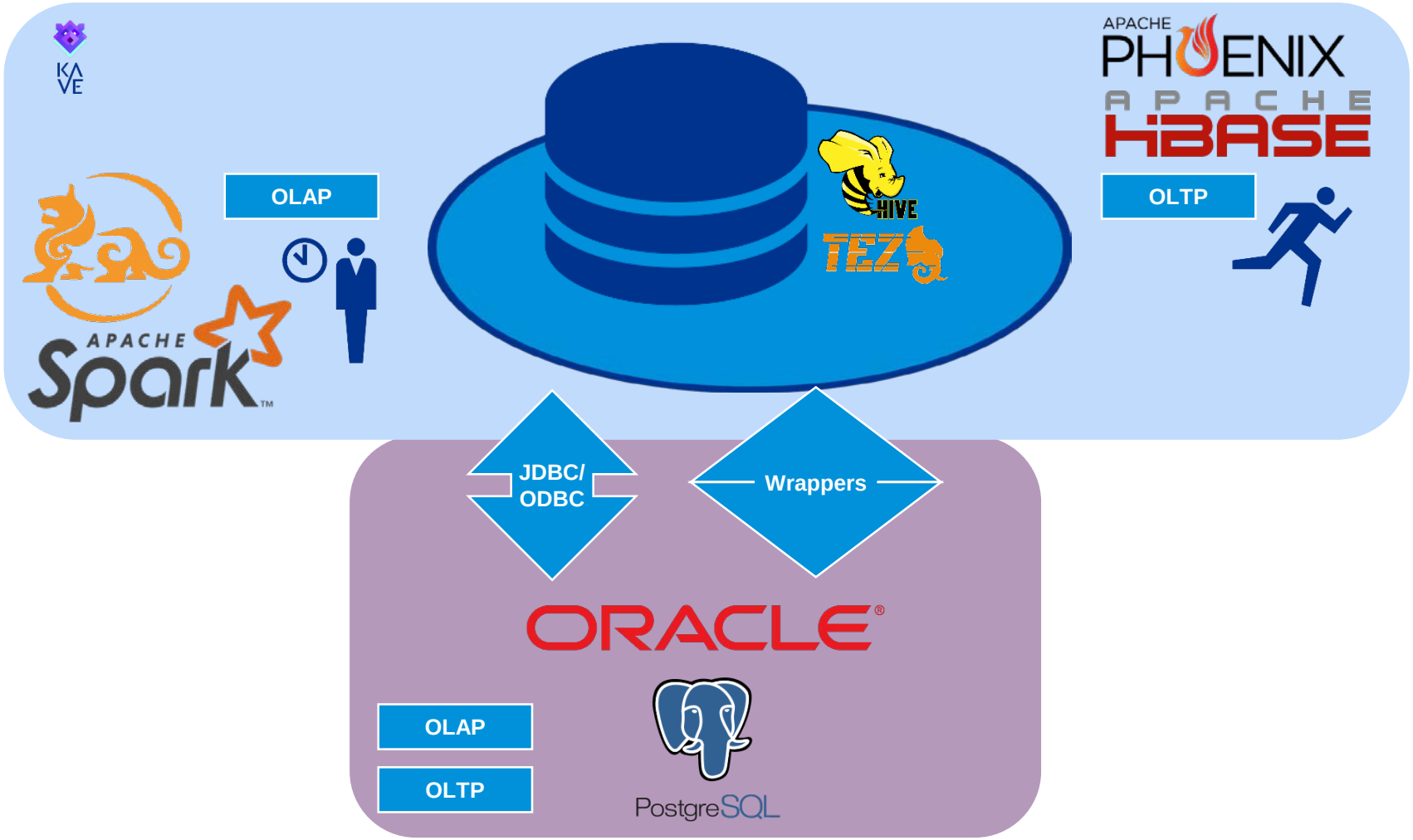
Integrated and automatic metadata creation and management

Apache orc™

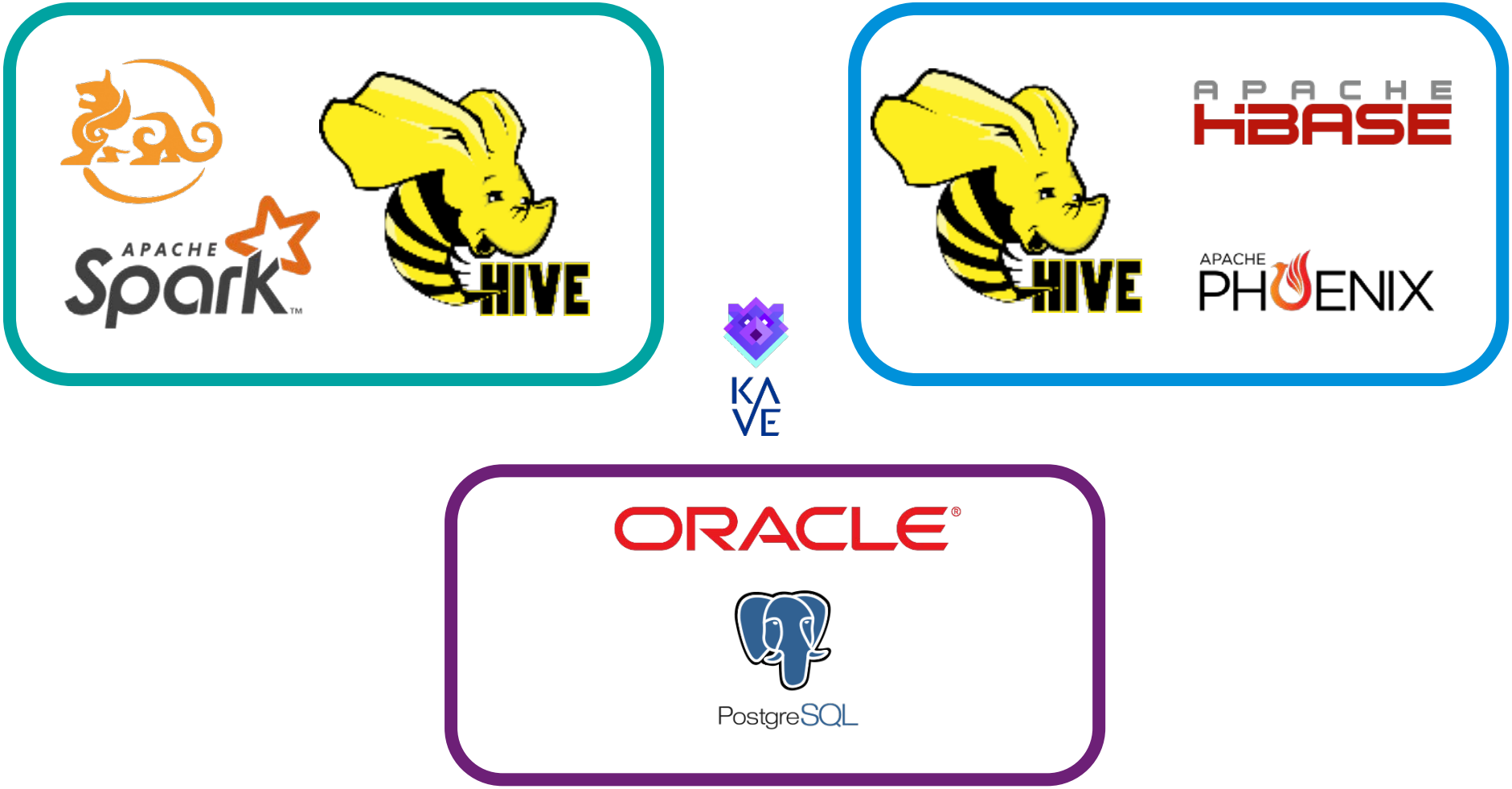
Parquet

Advanced and optimized Hadoop storage formats

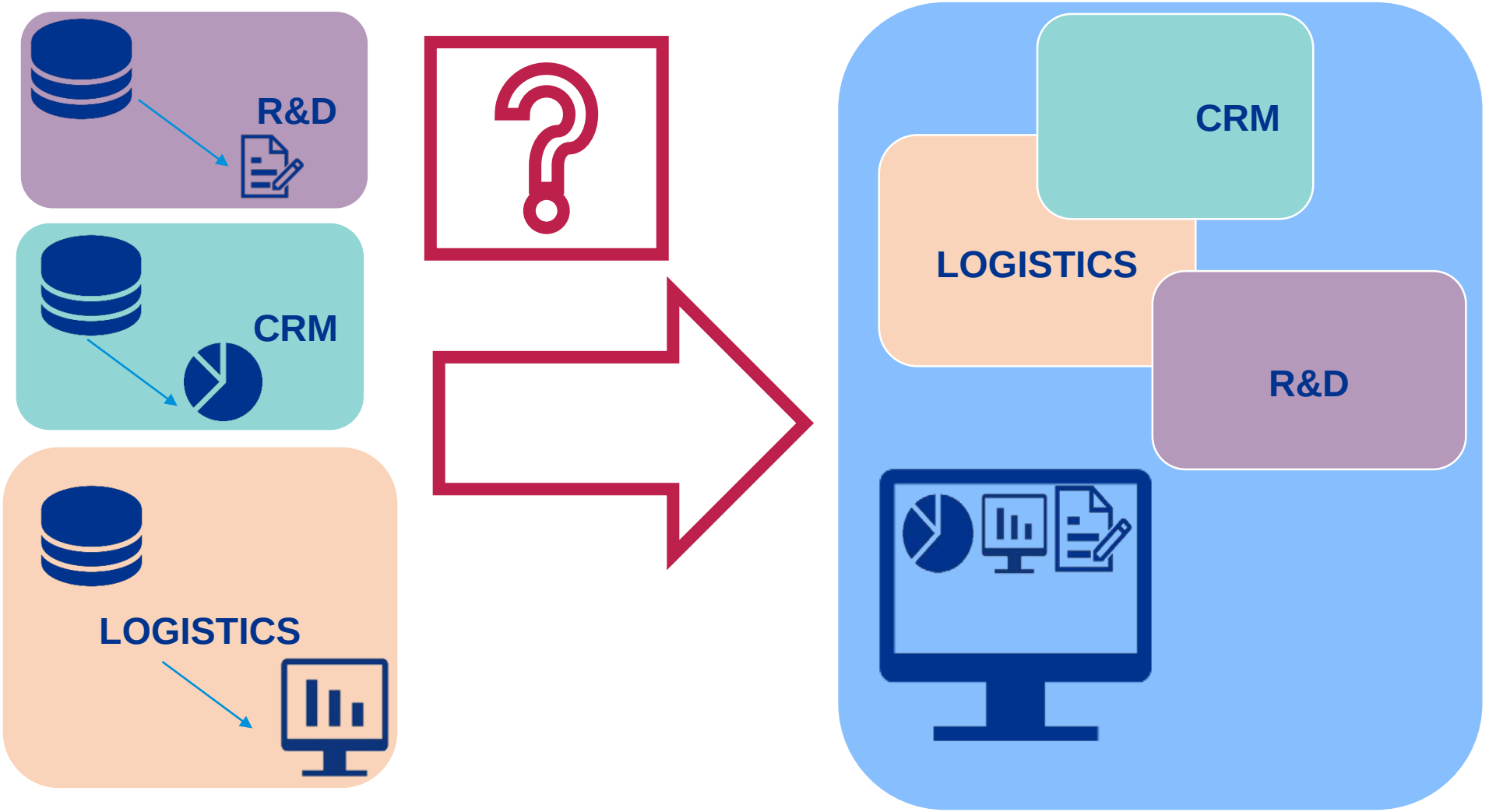
Enterprise Data Lake: OLAP & OLTP workloads



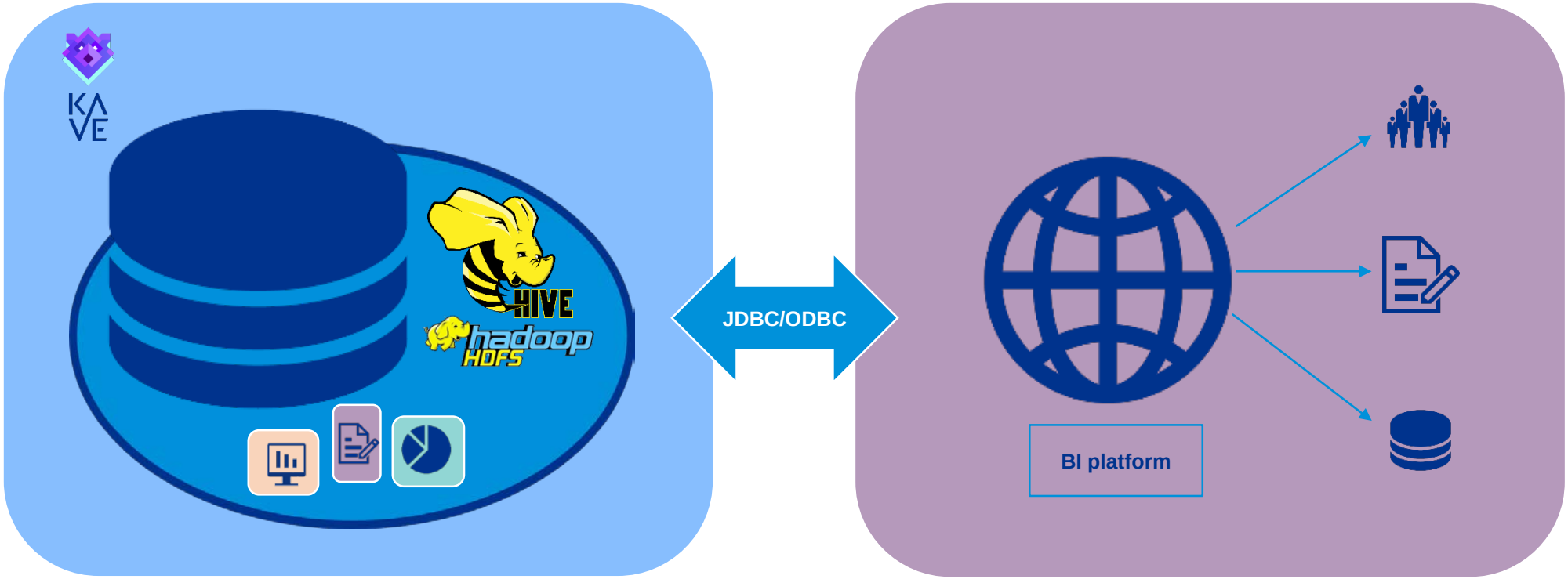
KAVE: OLAP & OLTP workloads



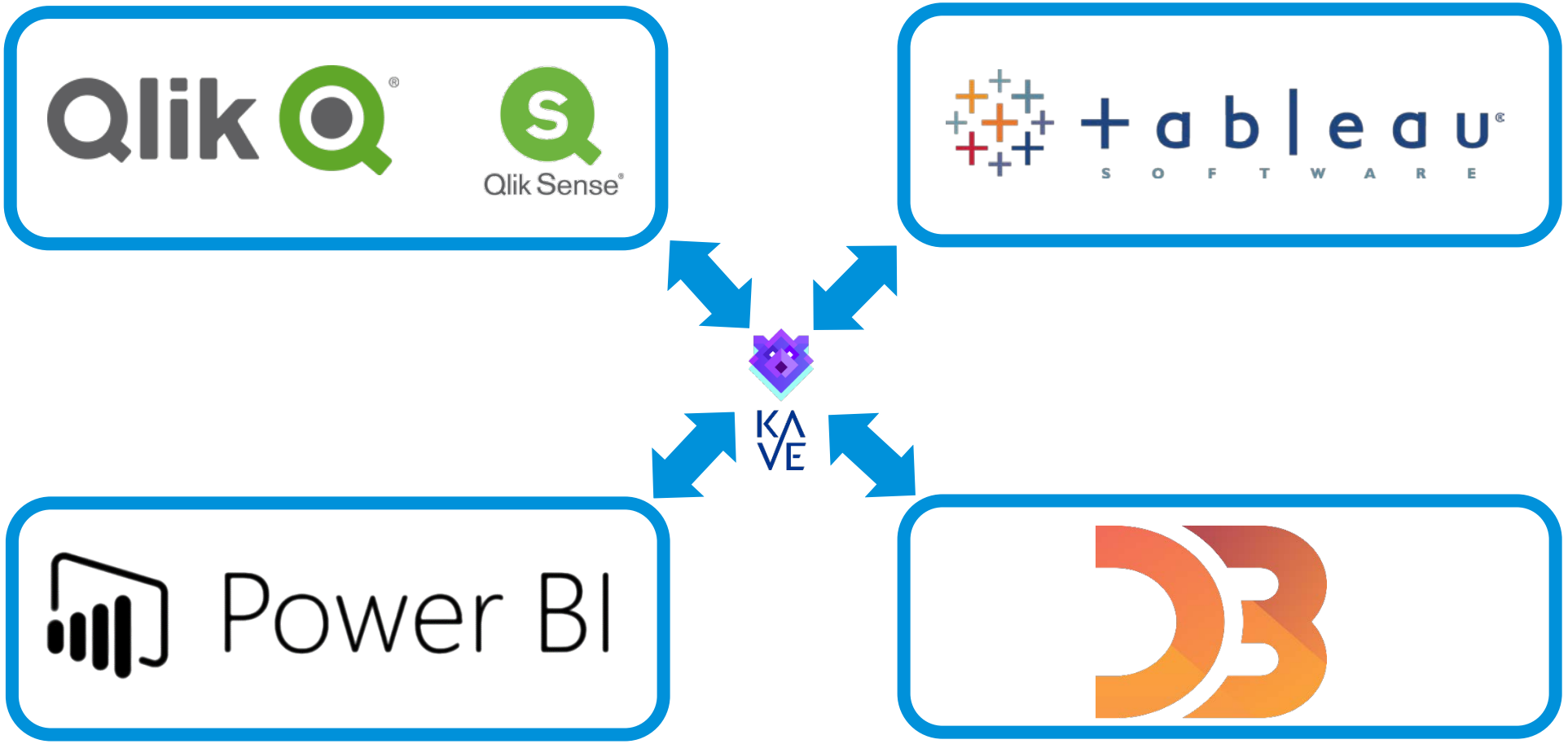
BI & BigData: are we there yet?



KAVE: reports & visualization



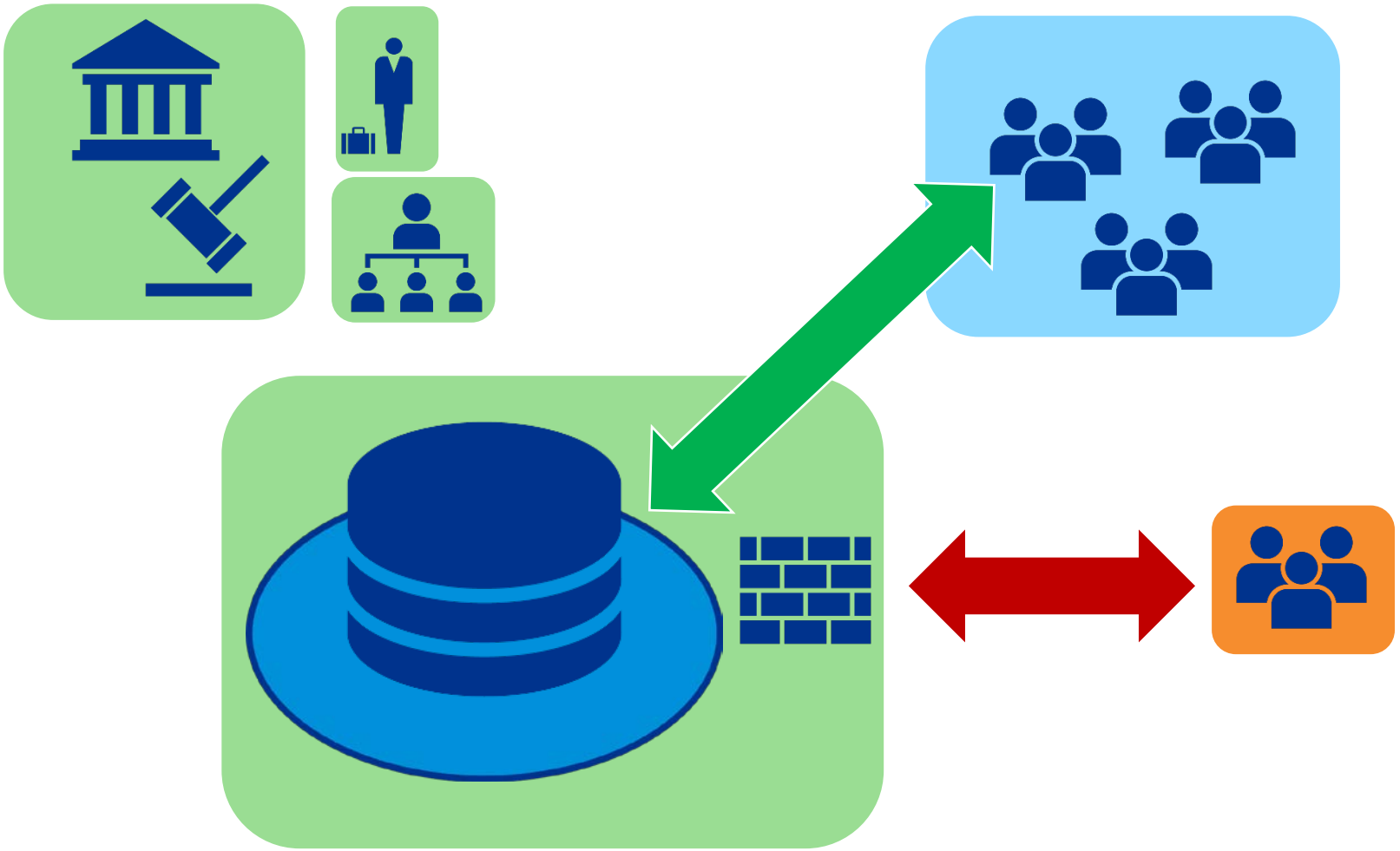
KAVE: reports & visualizations



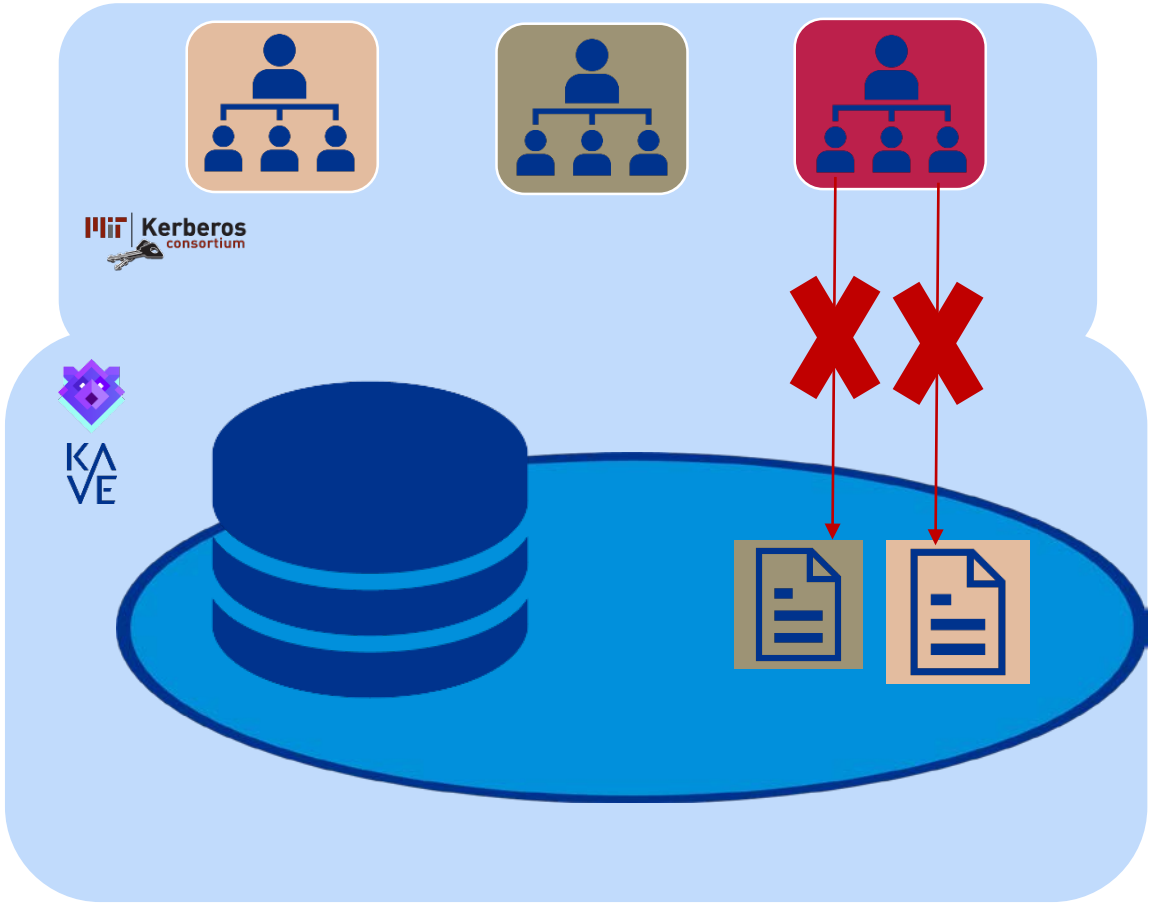
KPMG

Controlling the access and usage of the data

Enterprise Data Lake: security & governance



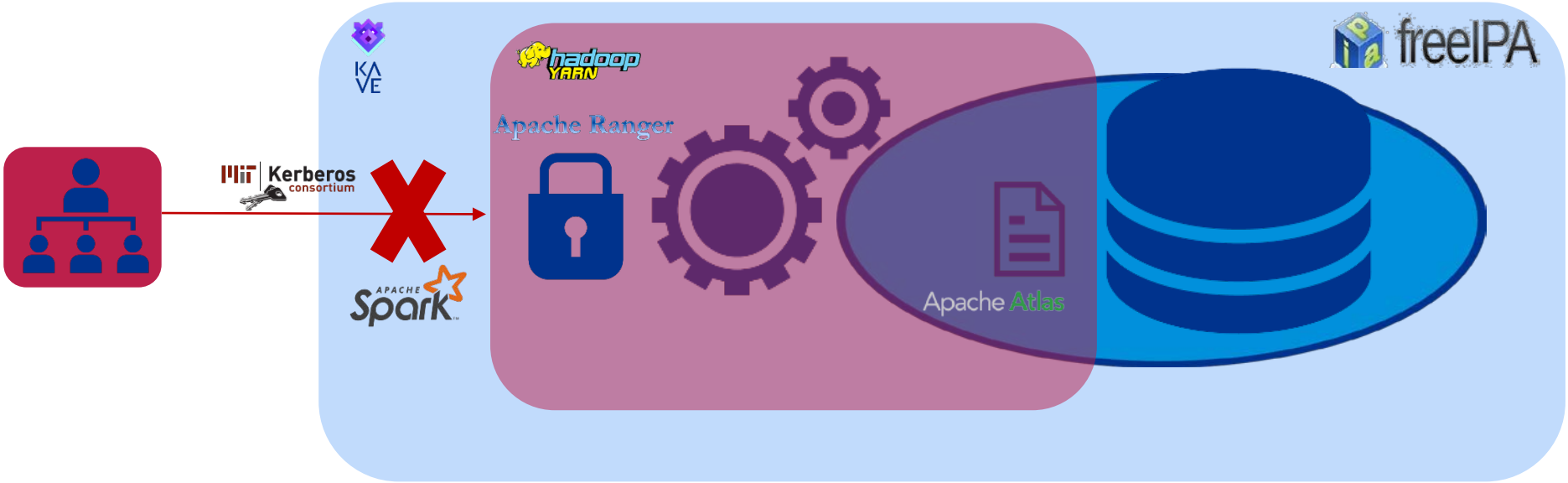
Enterprise Data Lake: security & governance



Sales and R&D departments data

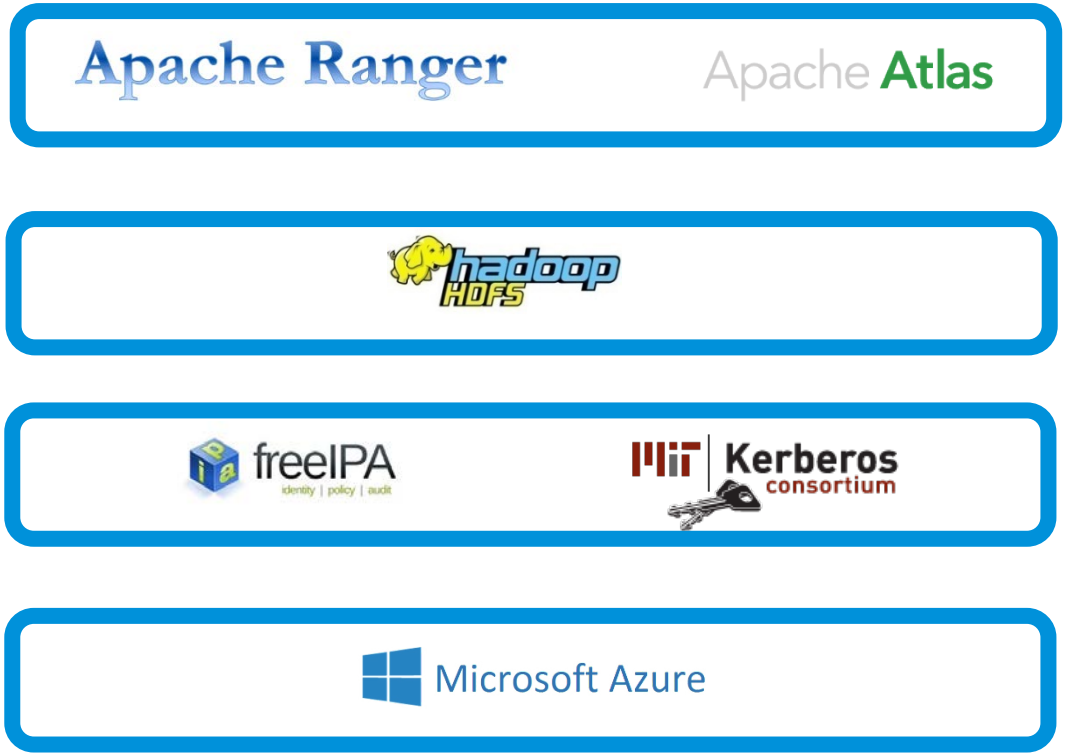
Finance department: no access to their data

Enterprise Data Lake: security & governance



Finance department cannot run Spark on test cluster

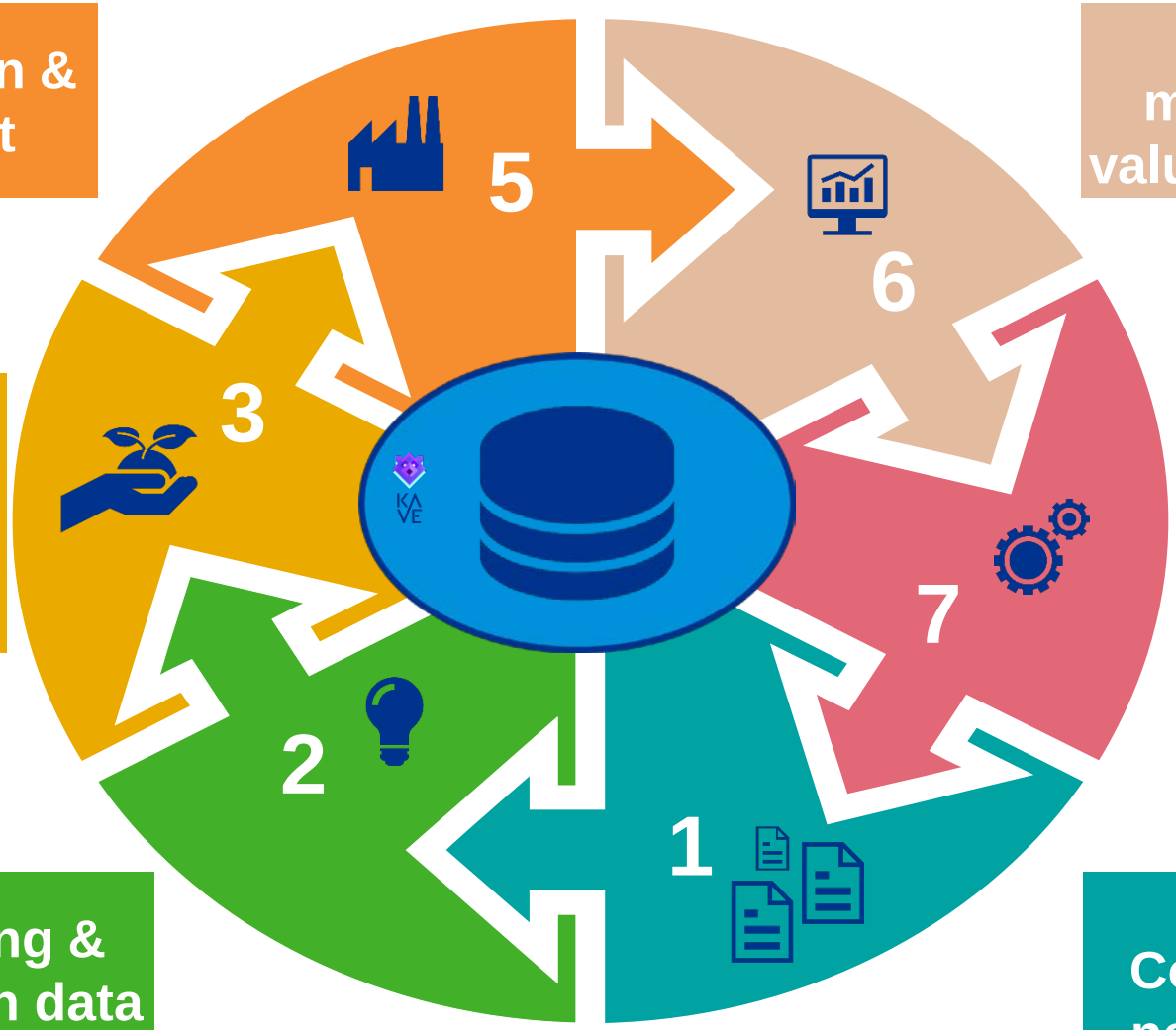
KAVE: full data management on secured infrastructure



KPMG

From experiments to production

Data-centric software products with cycle optimization



Productization & deployment

Solution monitoring and value measurement

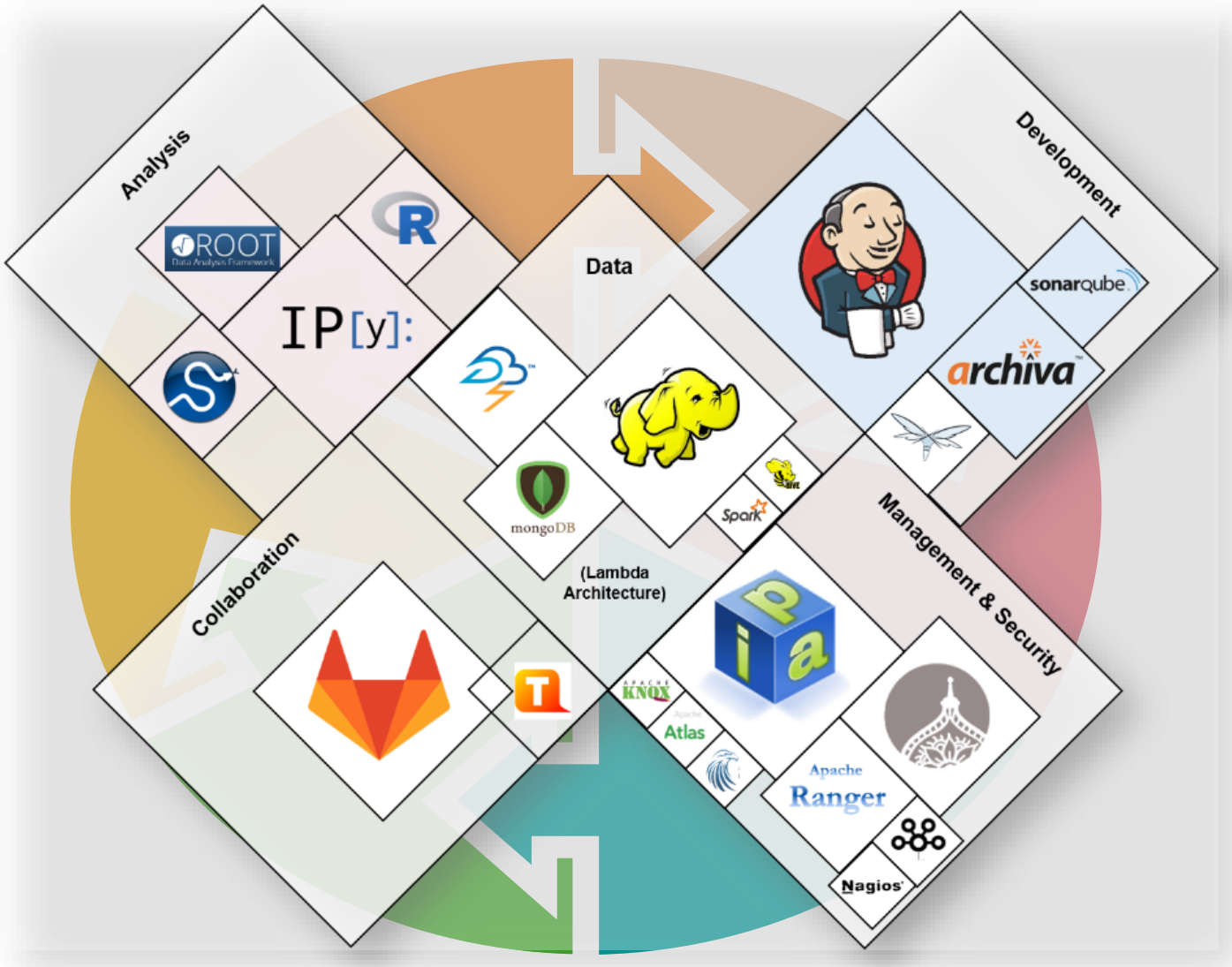
Business-significant proof-of-concept

Corrections & improvements

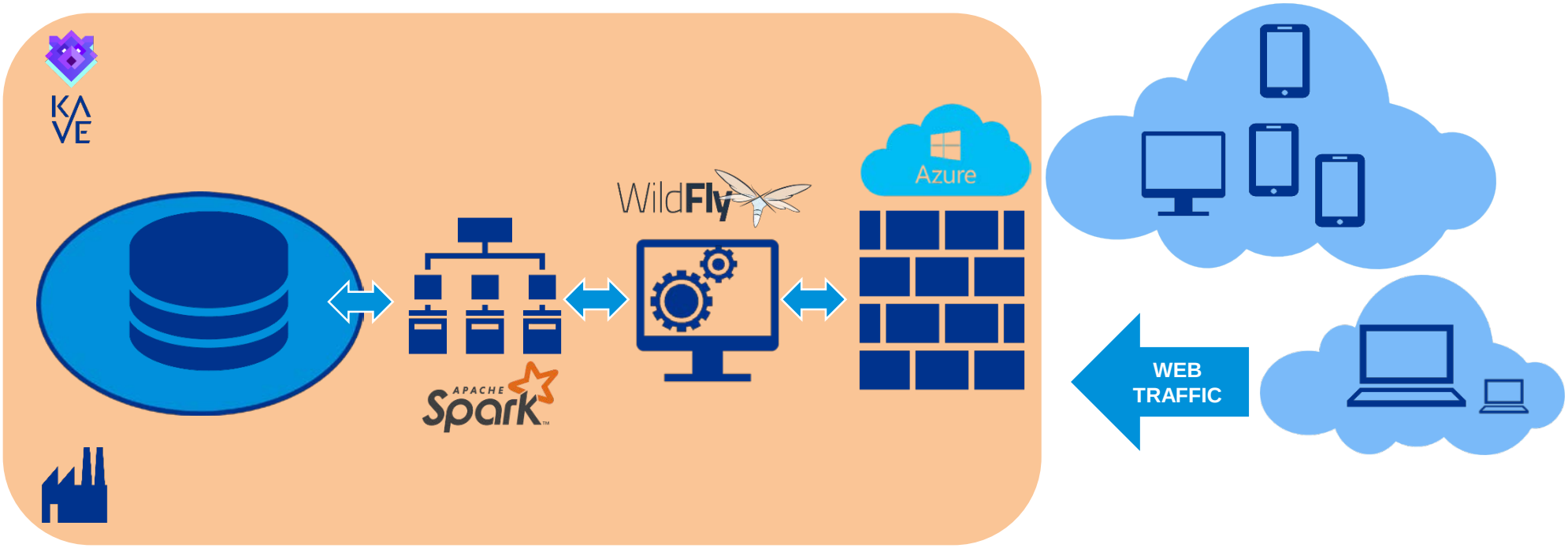
Brainstorming & exploration on data

Definition & Consolidation of needed datasets

KAVE & data-centric development model: a glance



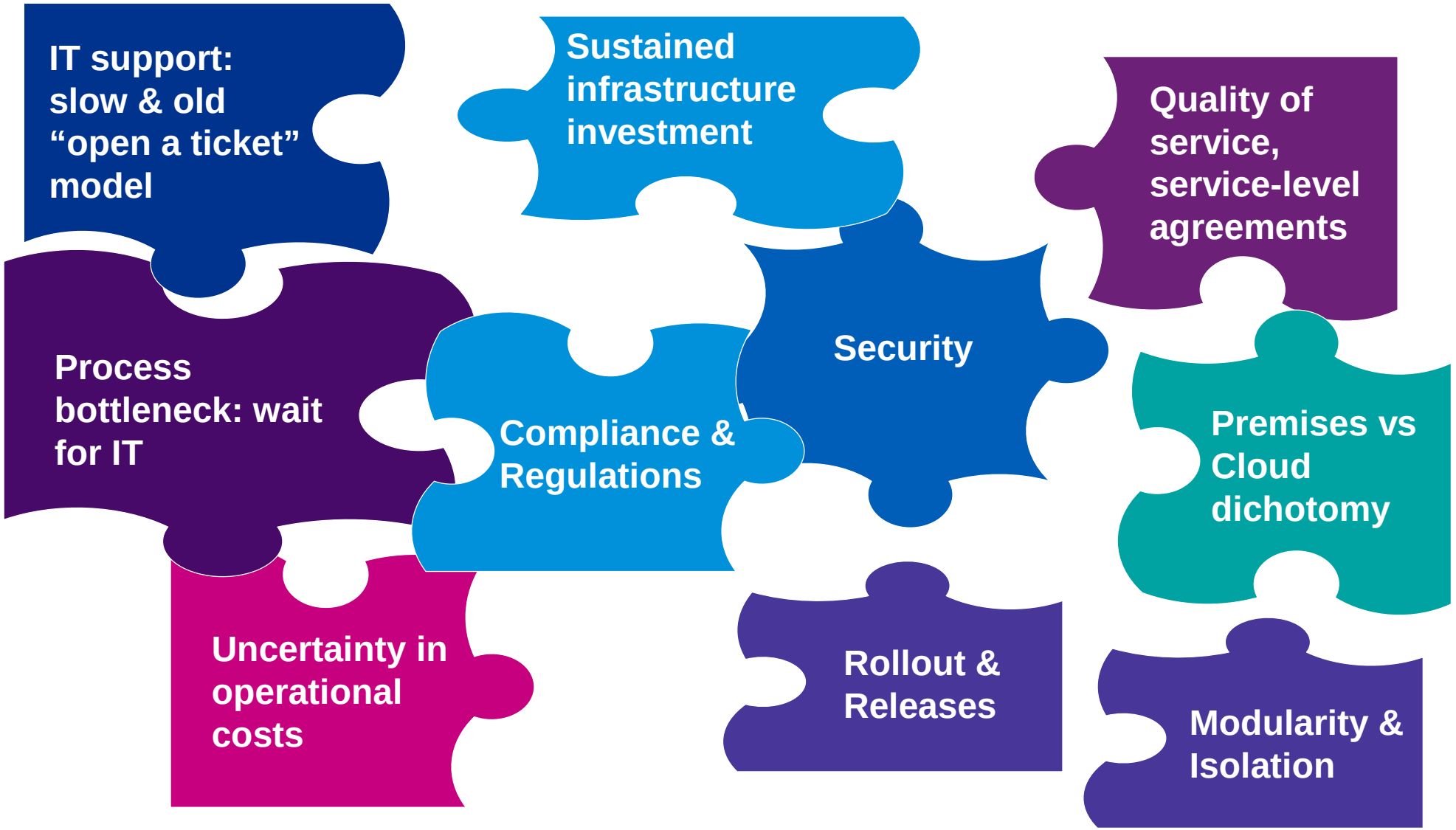
Prototype data product deployment for the web



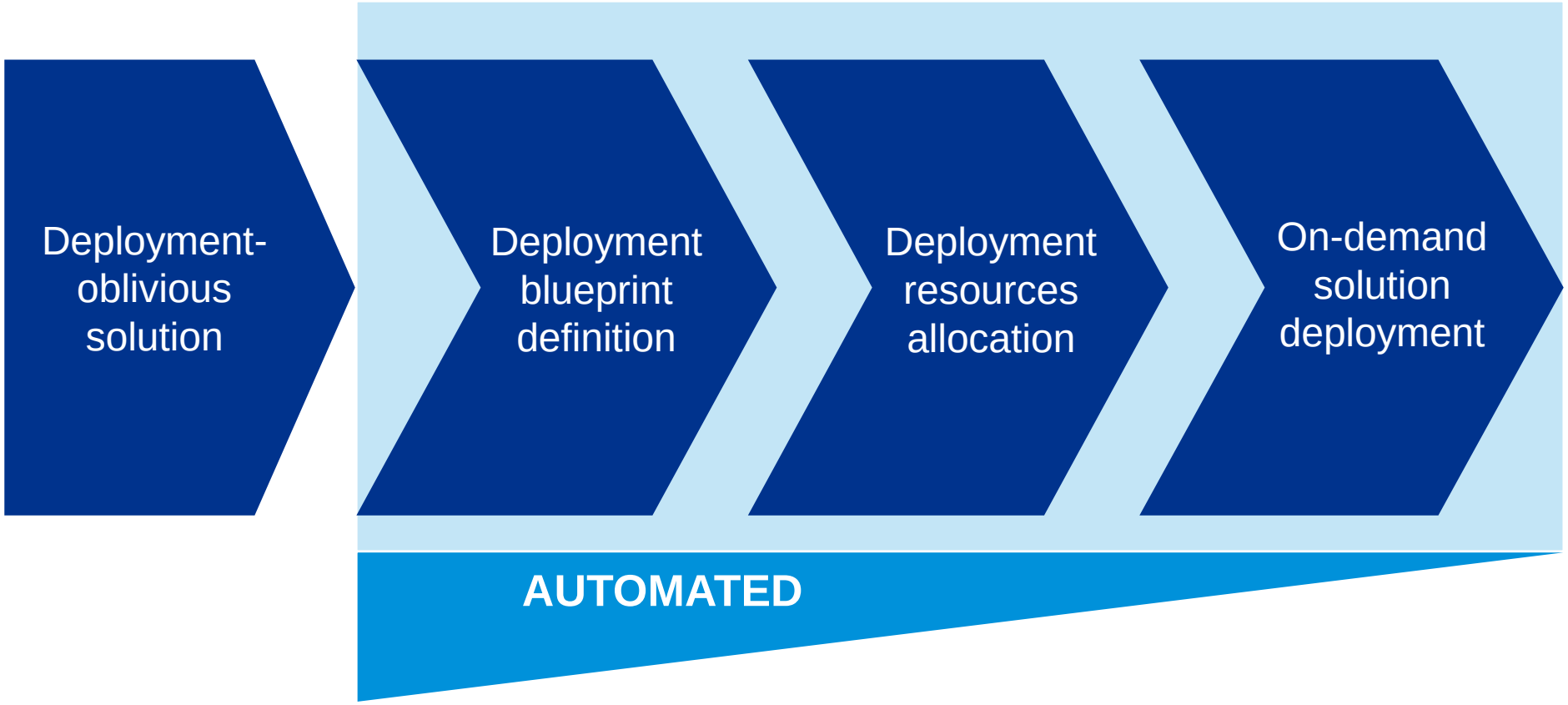
KPMG

The modern Cloud experience

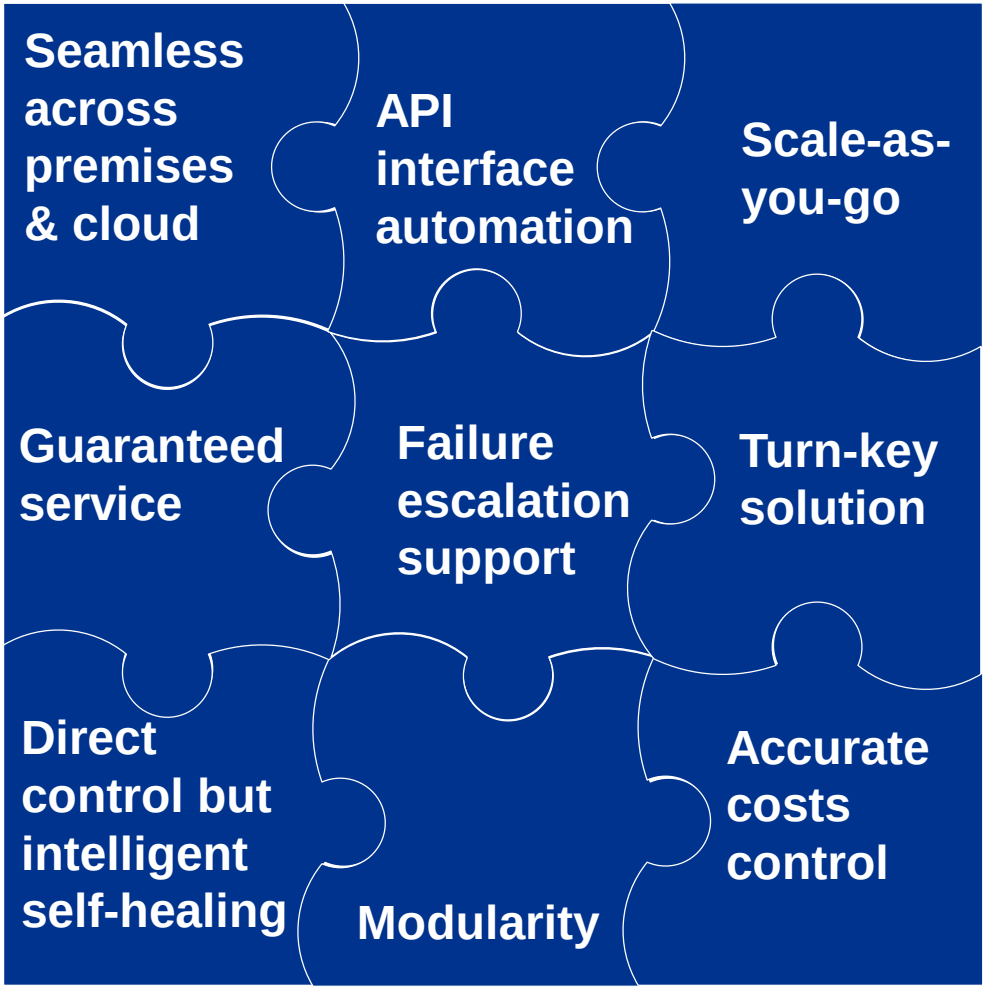
The many unknowns of low-automation infrastructure



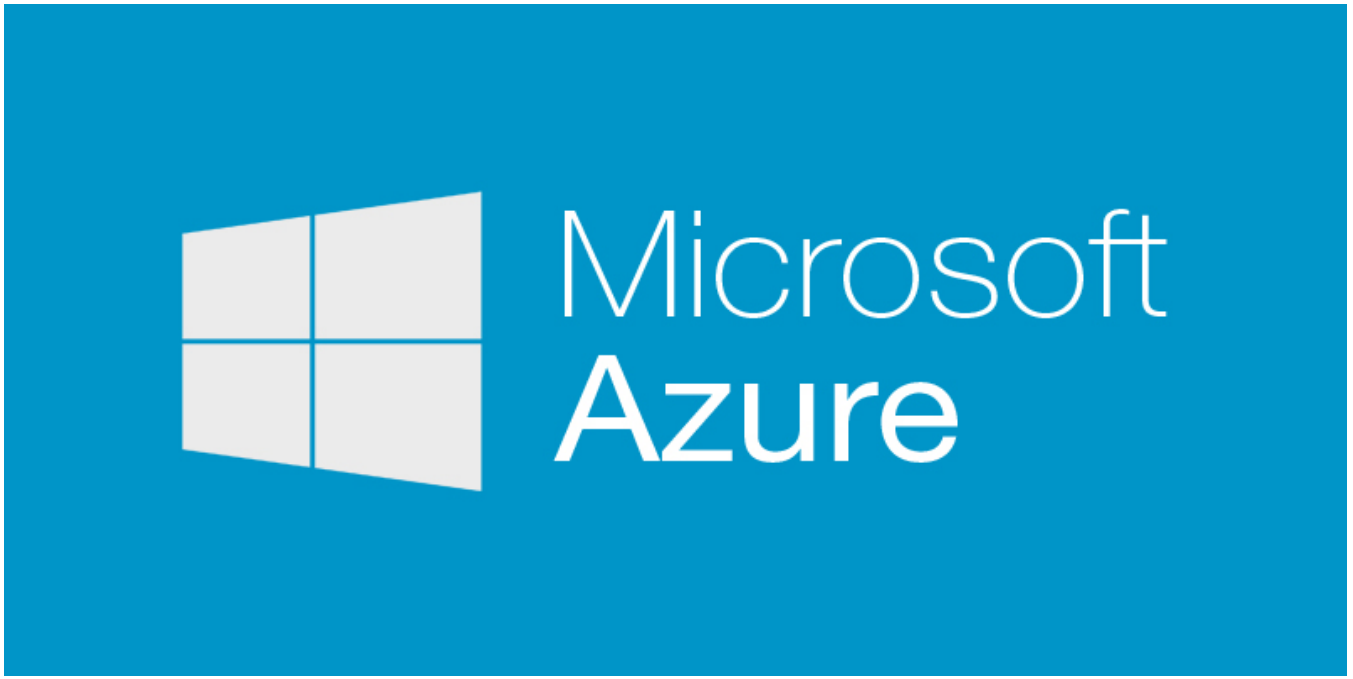
Satisfying the data product cycle: continuous delivery



The need for an integrated, dynamic and automated infrastructure



The preferred infrastructure model for KAVE: Azure



Azure

Datacenters

Tens of locations worldwide: compliance & localization

PaaS (platform as-a-service)

Extensive support for web applications

Security

Enterprise customizable security levels

IaaS (infrastructure as-a-service)

Fully automated virtual infrastructure management

Service & Availability

Per-service guarantees, virtually 100%

Costs & Billing

Ad-hoc minute-precision billing schemes; suspendable services

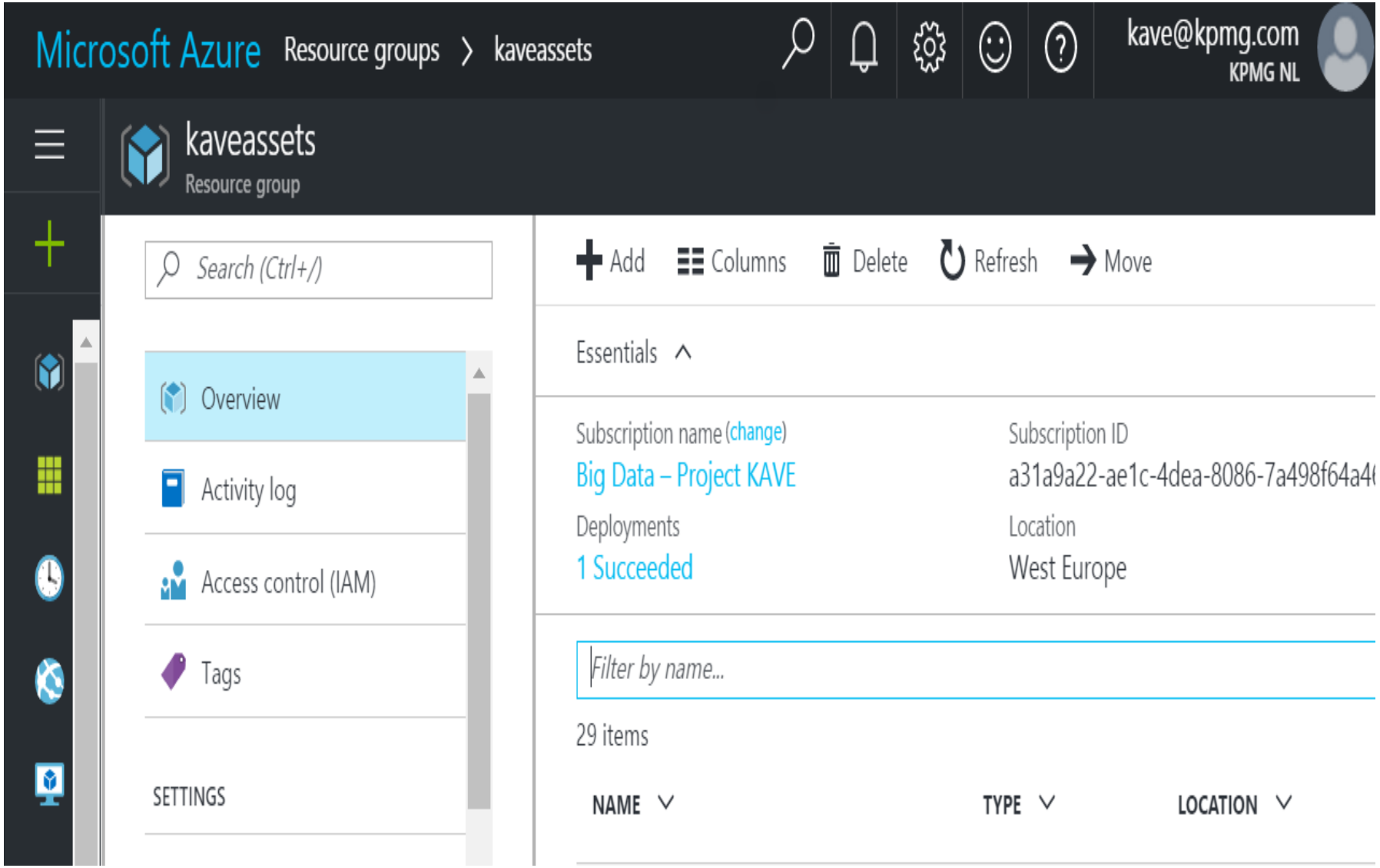
Marketplace

Basic Microsoft services and vendor offerings: vast offer, direct vendor contact

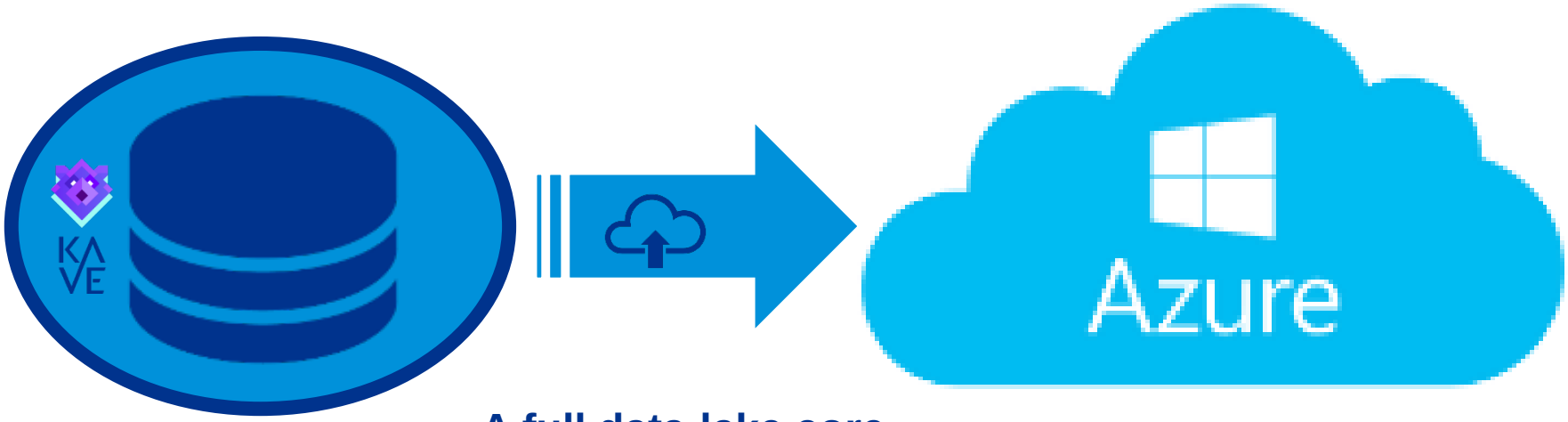
Modularity & Coverage

Dozens of independent and integrated services

Azure: modern web user experience

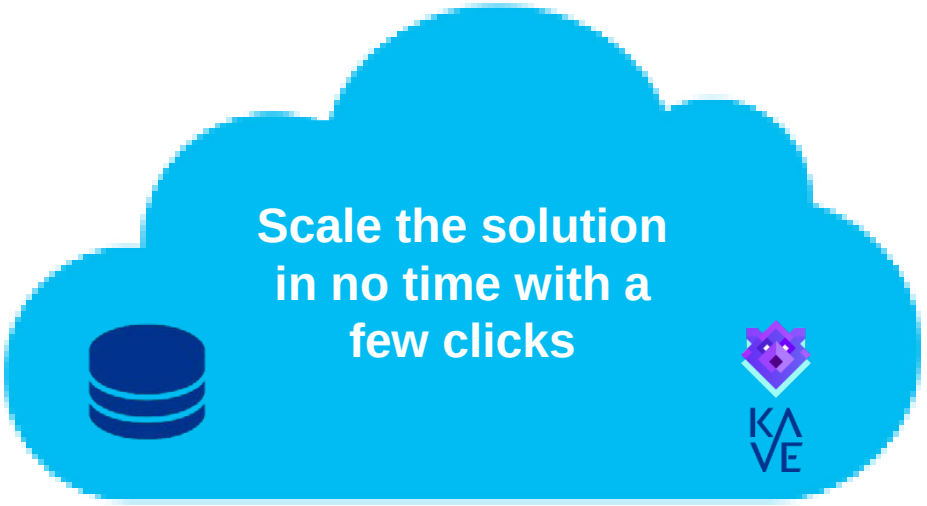


Azure: preferred infrastructure for KAVE

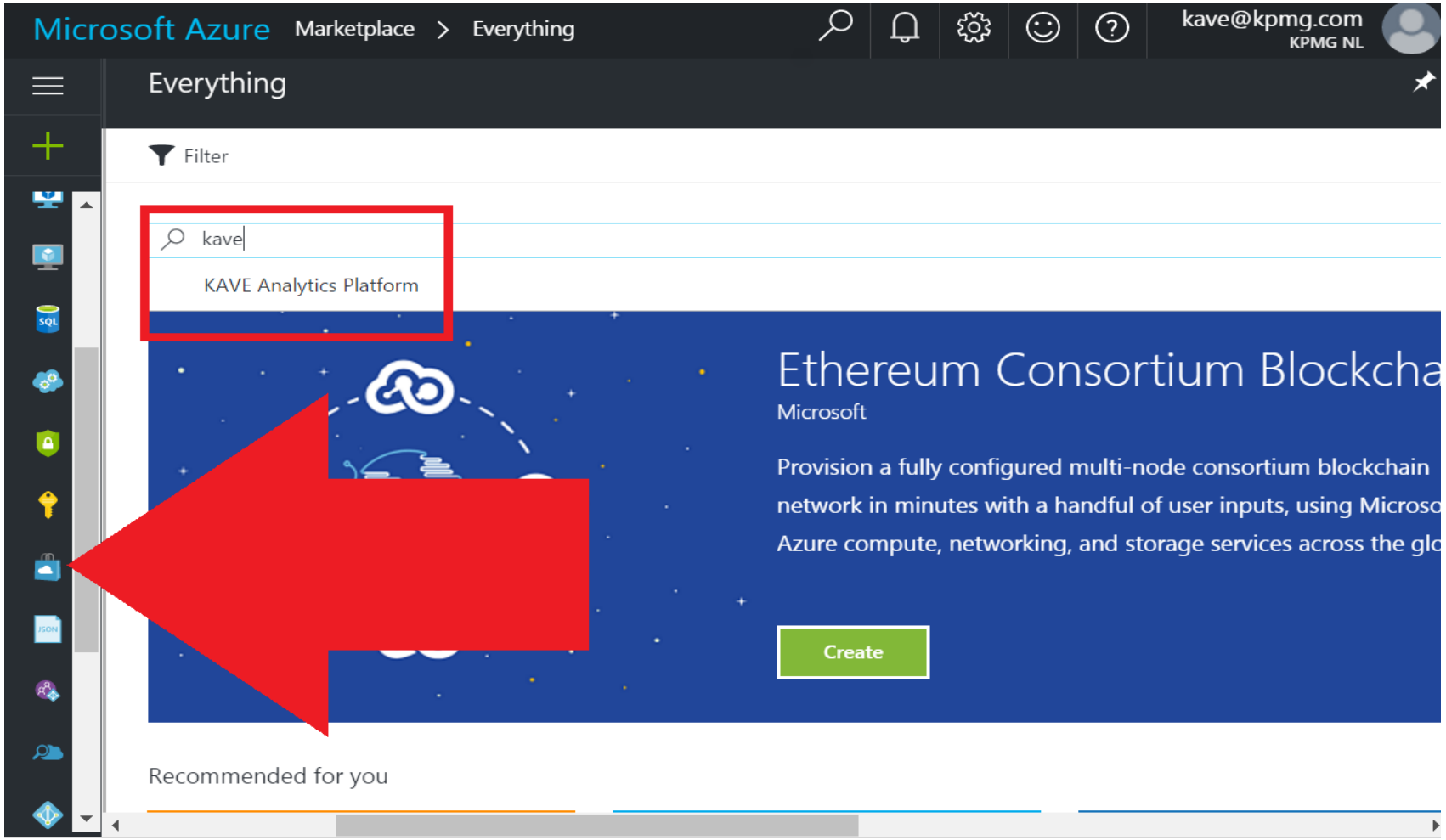


A full data-lake core deployment, in just a wizard!

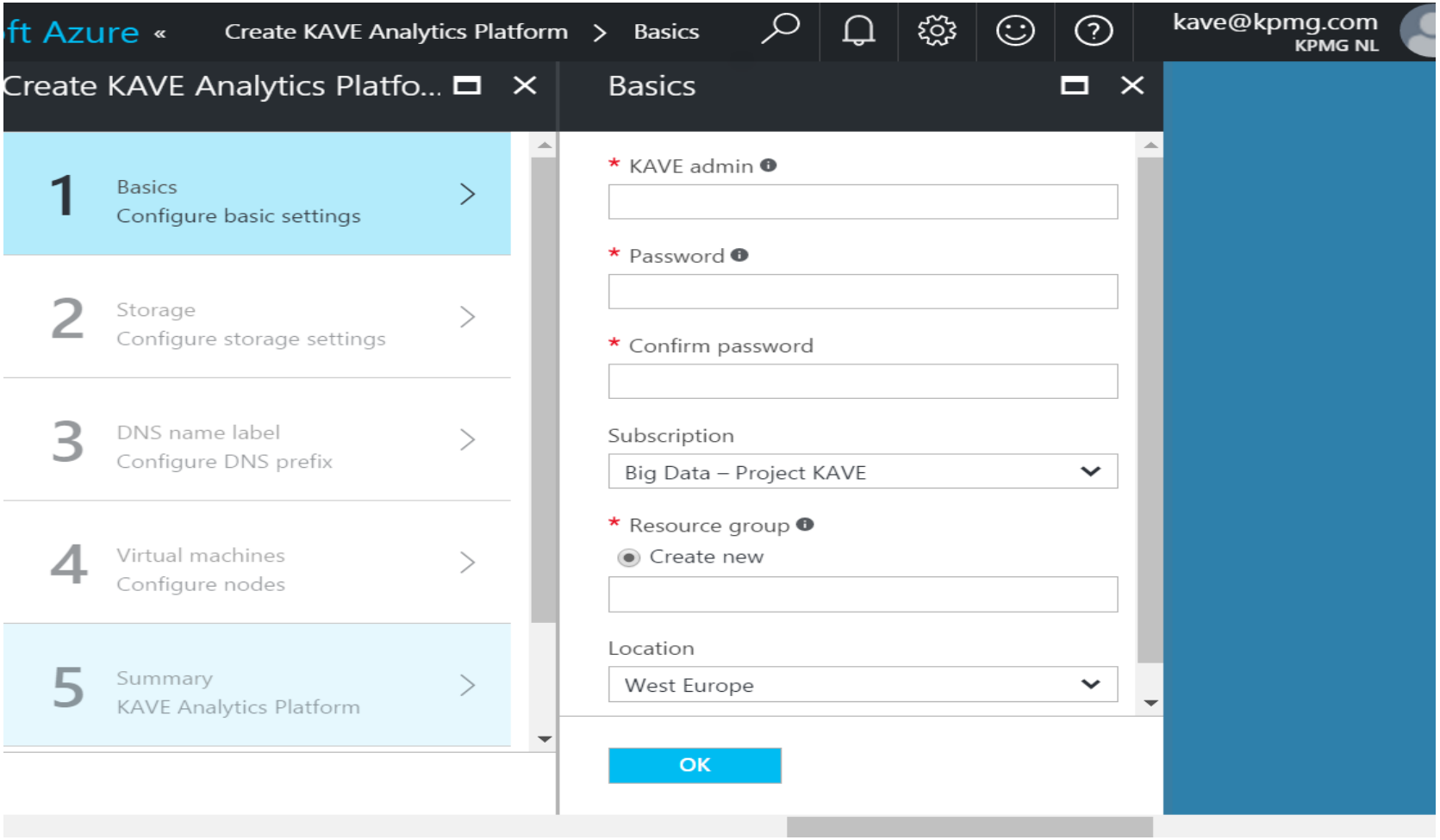
The idea of Data Lake as a service with KAVE



Try a fully working KAVE instance on the Azure marketplace!



Try a fully working KAVE instance on the Azure marketplace!



KPMG
cutting through complexity

Open source: contribute & extend to fit your needs!

KAVE

http://beta.kave.io

KPMG

Thanks !