



KAVE: core platform for the Data Lake

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



KAVE

Table of contents

- Open Source for Analytics: an established yet evolving trend
- Closed-source D&A: drawbacks & risks
- Data Lakes
- What is the KAVE?
- KAVE & the fulfillment of the Data Lake evolution
 - Data Warehouse & Business Intelligence functionalities
 - Controlling the access and usage of the data
 - From experiments to production
 - The modern Cloud experience



Open Source for Analytics: an established yet evolving trend

A person with glasses and a red t-shirt is sitting cross-legged on a rooftop at night, working on a laptop. The background is a blurred city skyline with many lights, including a prominent building with a glowing dome. The text is overlaid on the right side of the image.

The most successful Fortune 500
companies run and grow on Open Source
Data&Analytics software

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)

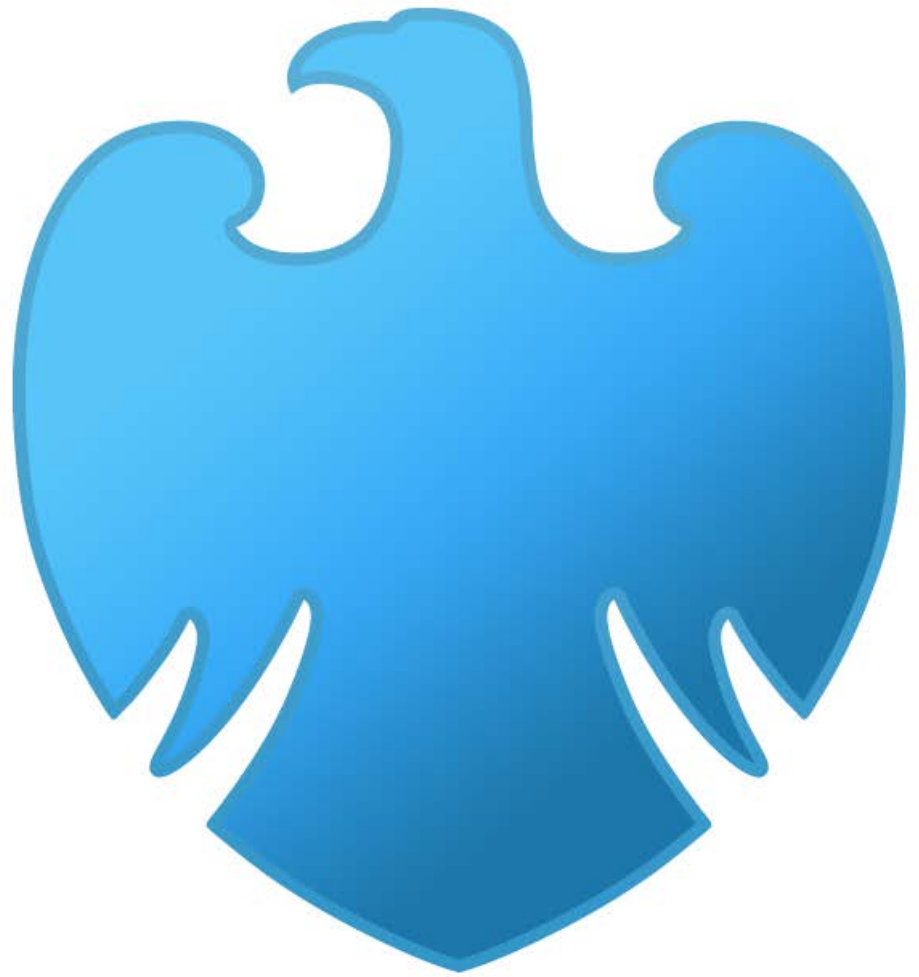


Google has released
over 900 open source
projects, totaling over
20M code lines.

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

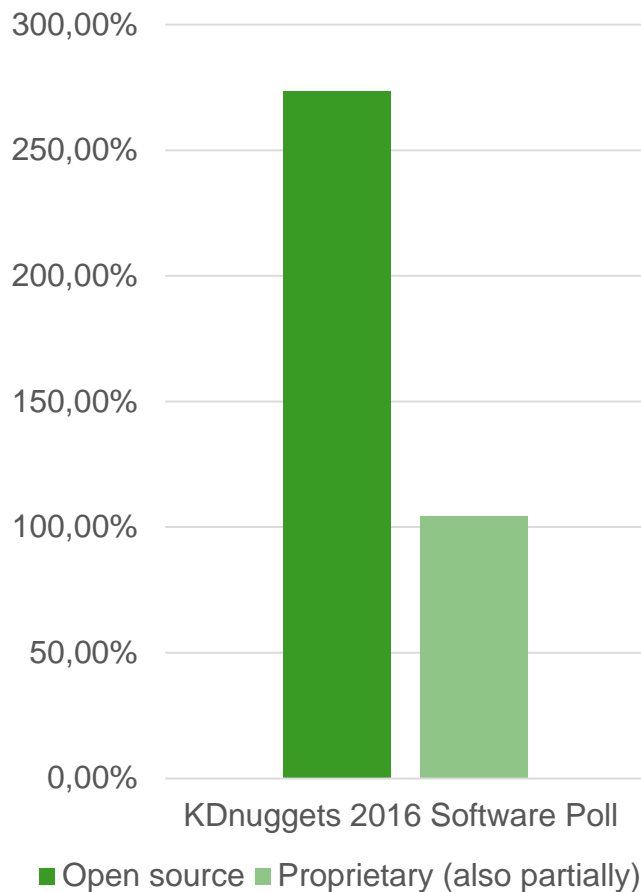


Closed-source D&A: drawbacks & risks

A man in a blue and white striped shirt is looking at a computer monitor in a modern office. The office has large windows in the background, showing a cityscape. The text is overlaid on the left side of the image.

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

In 2008 Nokia very successfully open-sourced its SymbianOS handset system, with an expense about 300M \$



It was too late: most hardware vendors had already moved to Android, the open source mobile system by Google



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

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





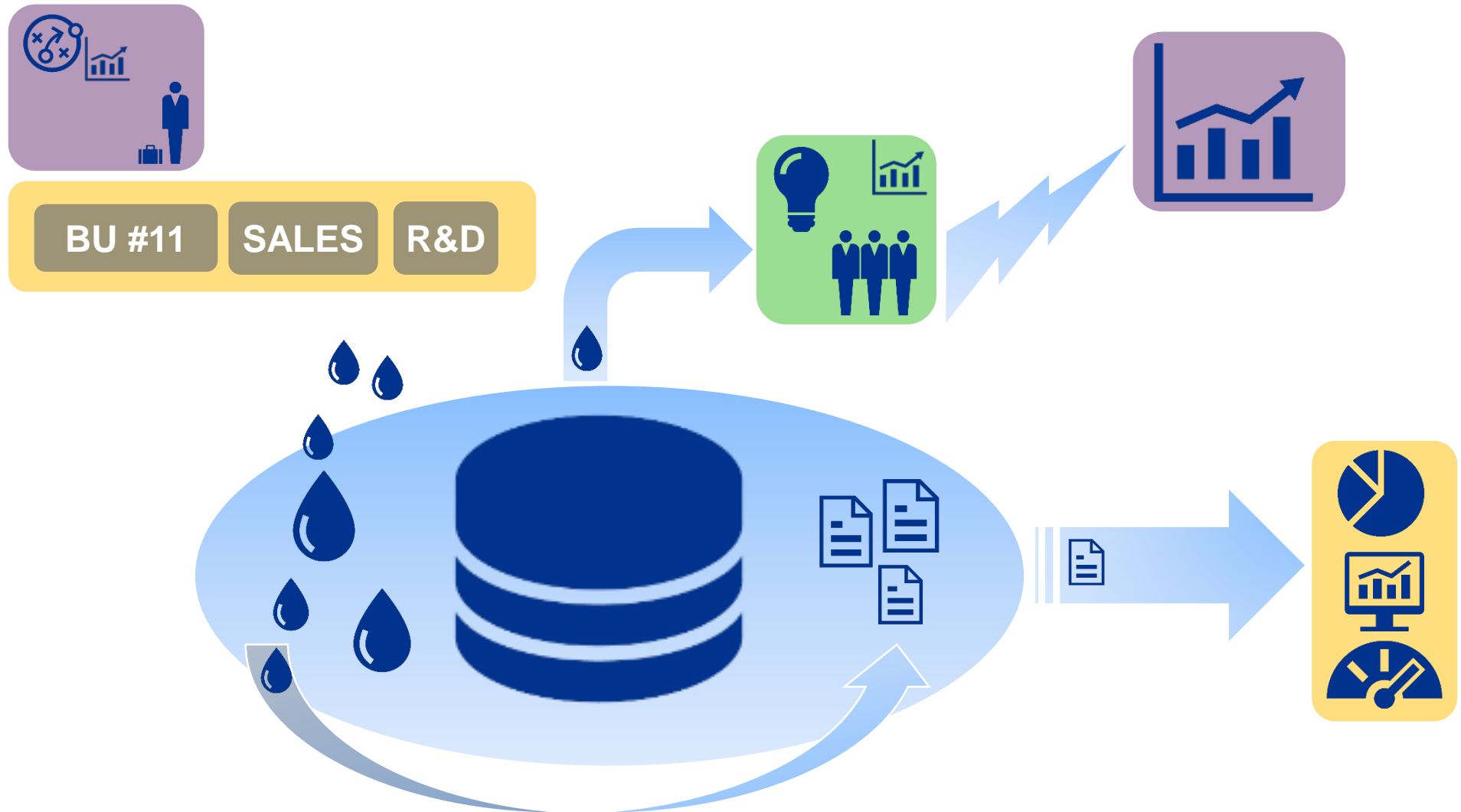
Data Lakes

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

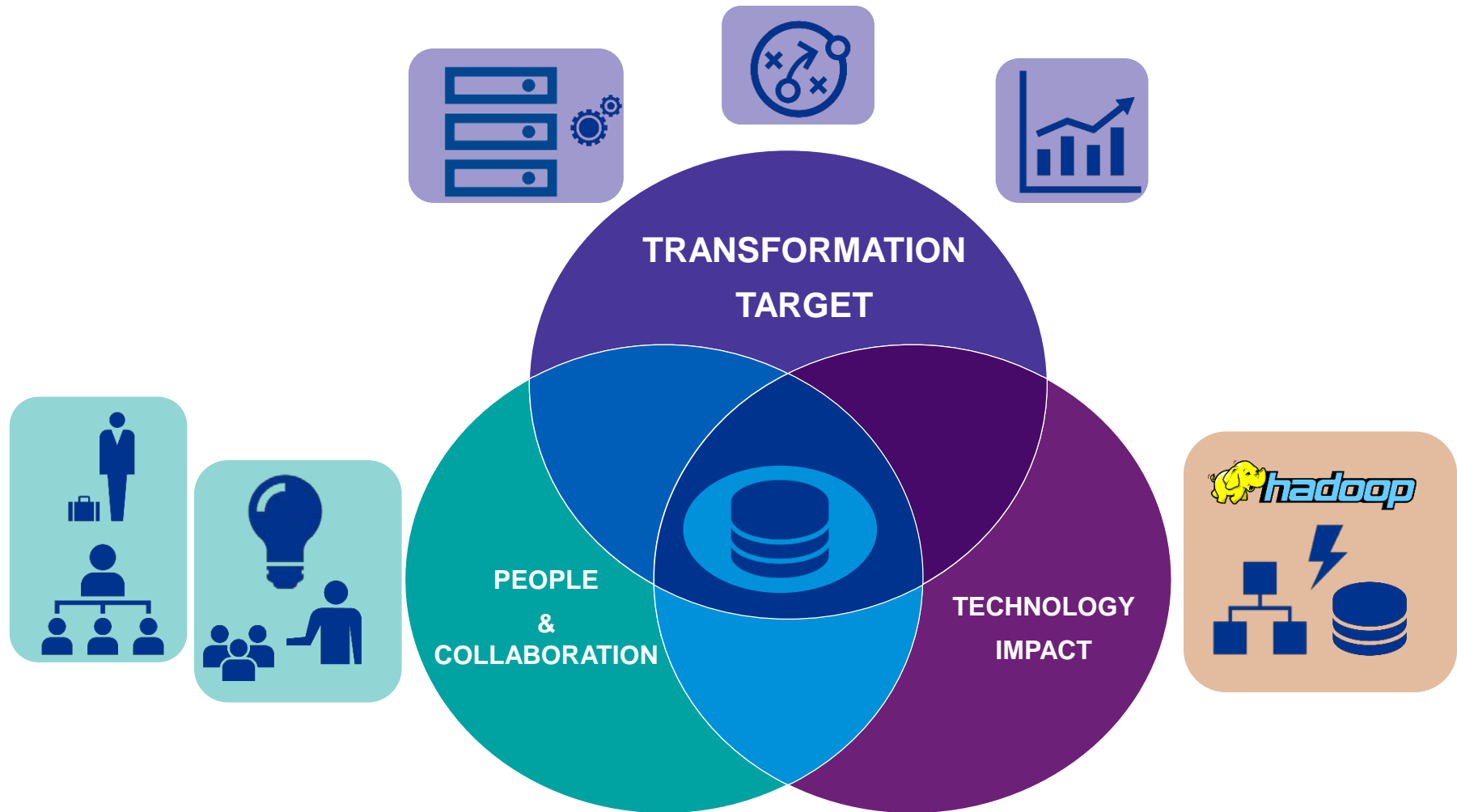
The Data Lake empowers a pan-organizational and holistic view on the information.

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

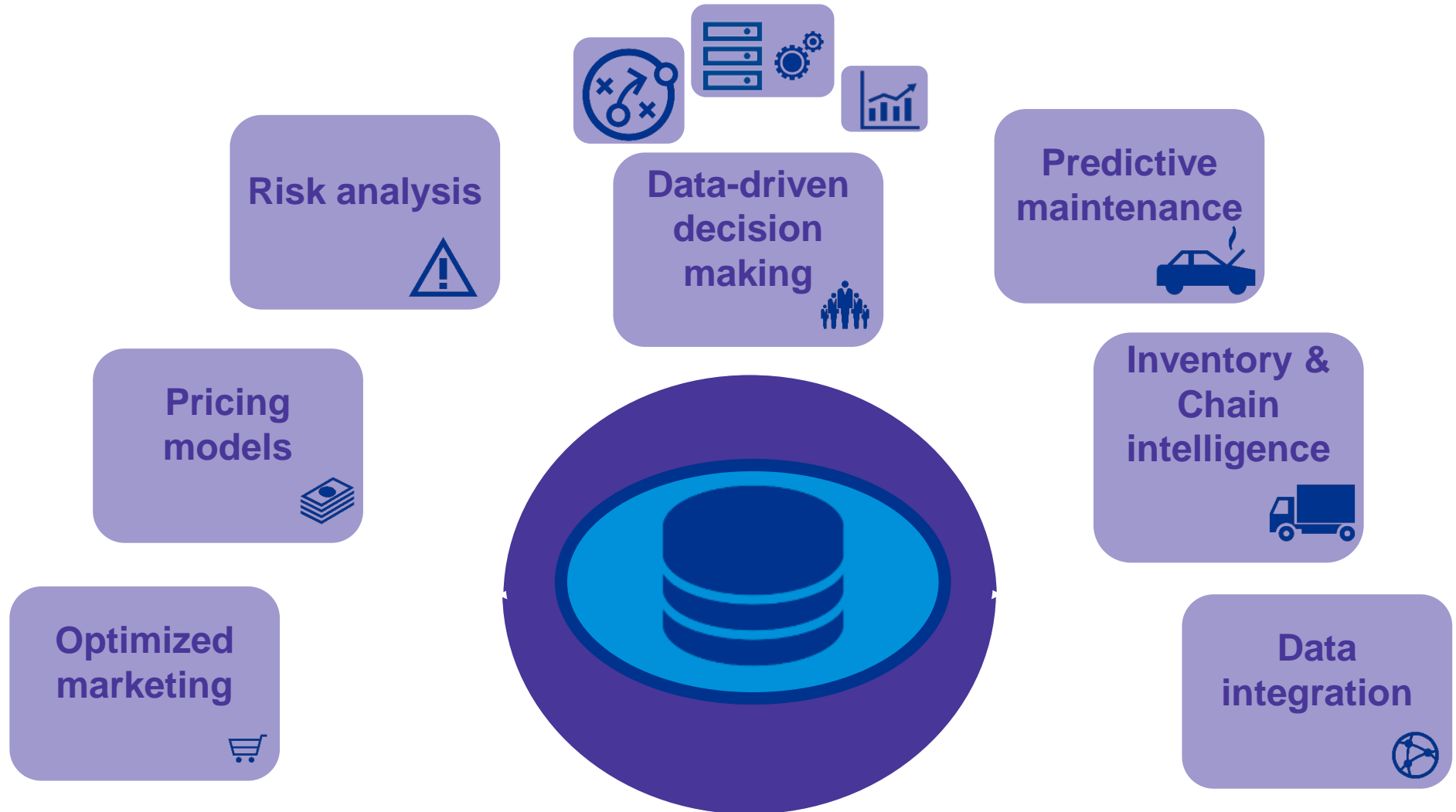
The Data Cycle



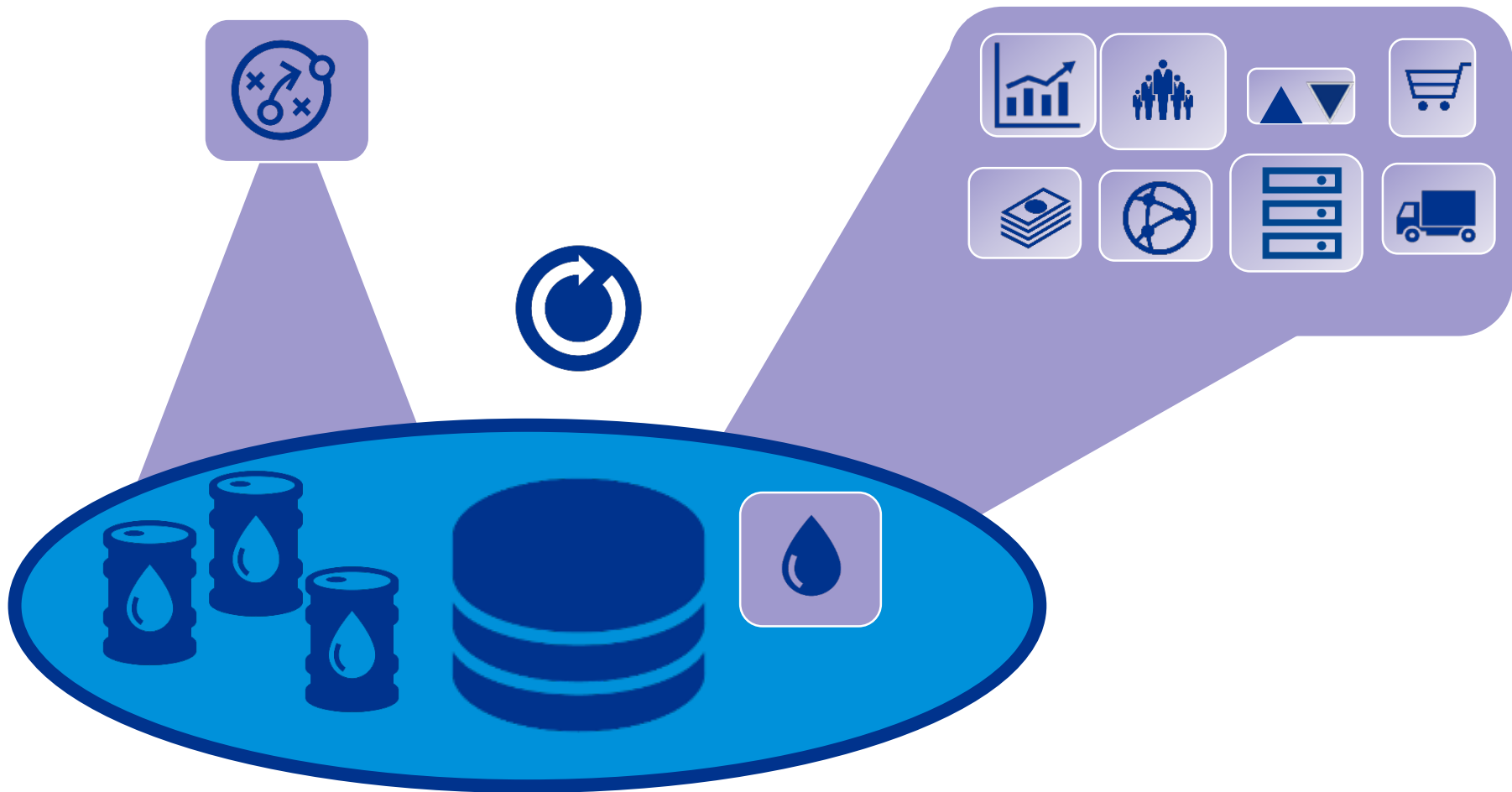
The Data Lake: driving the analytics evolution



The Data Lake: analytics processes & new strategies



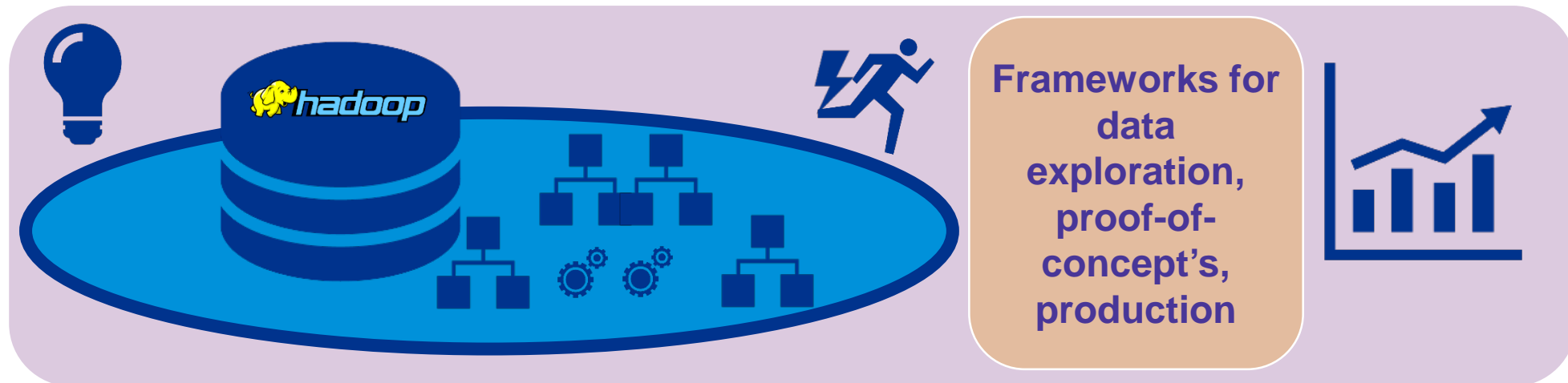
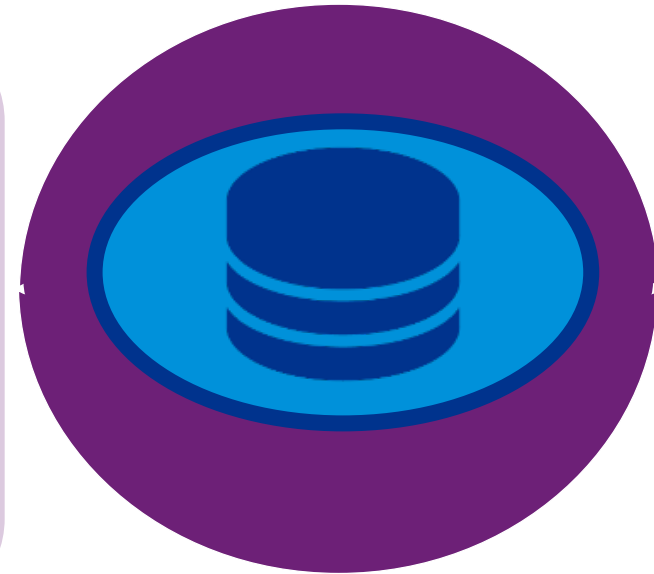
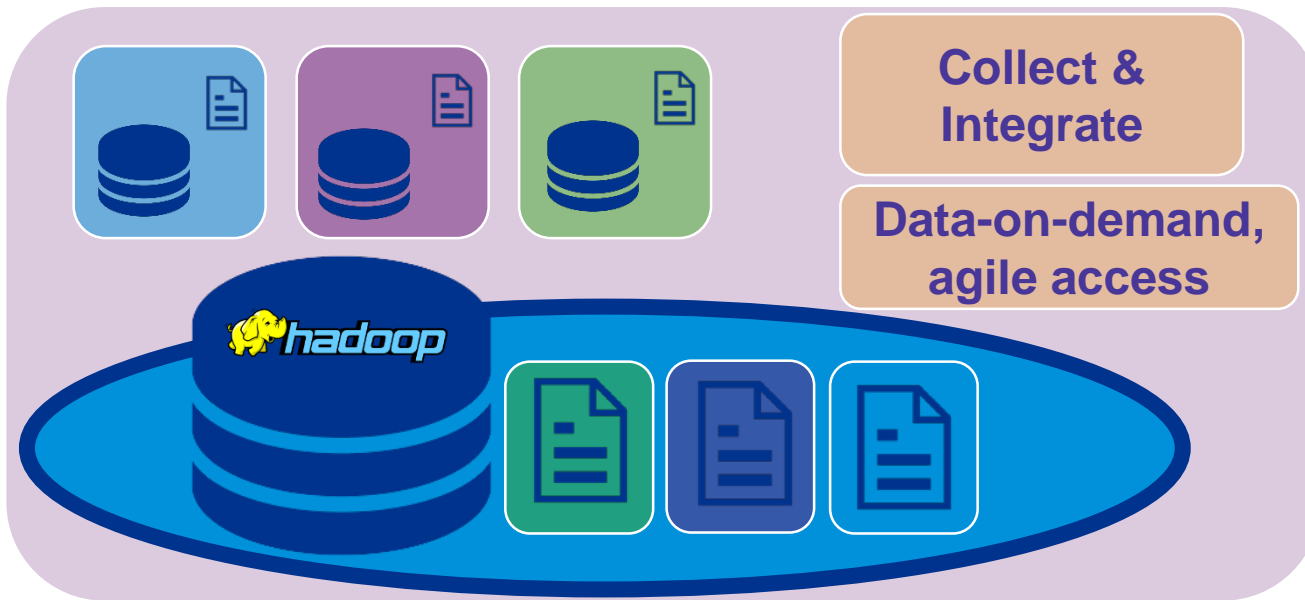
Data: not a by-product but a source of value



Enterprise Data Lake: analytics-driven organization



Data analytics: make value out of data

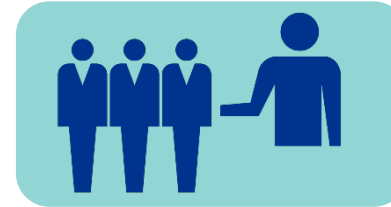


Focus: people



**Not just tech-trend,
real value for CIOs**

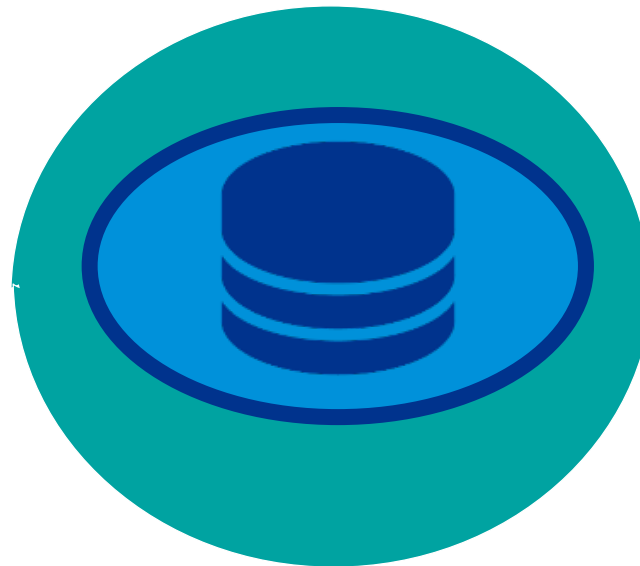
**Main reference for
CDOs**



**Enhanced customer
experience, ad-hoc
scenarios**



**Valorize your
analysts team,
attract new talent**



**Comply to the
organization
structure with
respect to data**

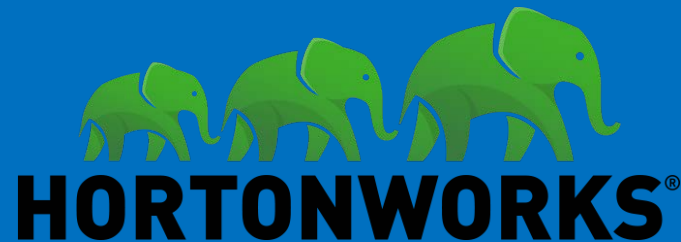


What is the KAVE?

KAVE: extension of the HortonWorks Hadoop distribution



KAVE extension

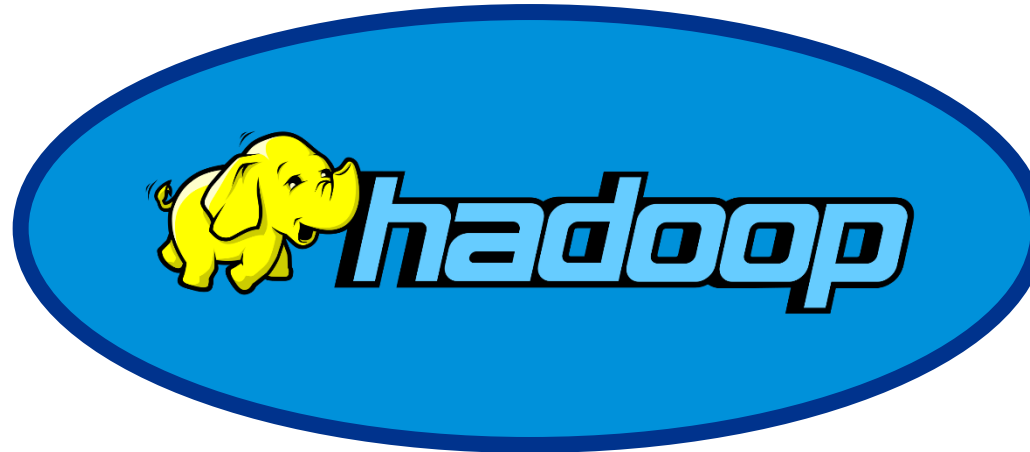


**HortonWorks
Data Platform
distribution**



**Hadoop core
software**

Data Lakes established technology ecosystem: Hadoop

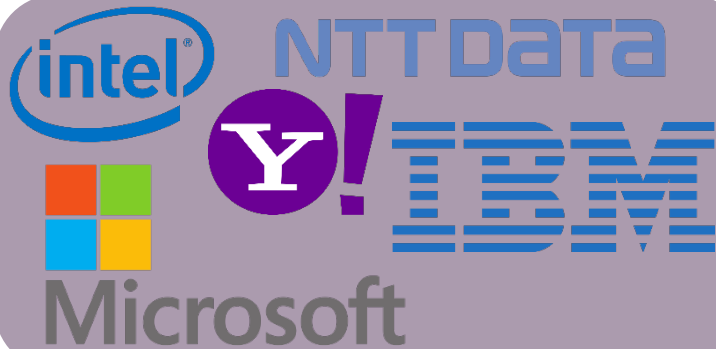
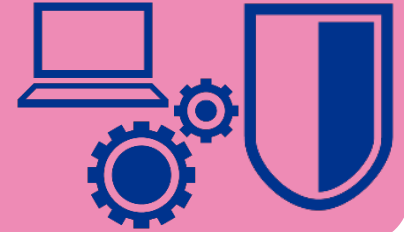


Data Lakes established technology ecosystem: Hadoop

De-facto industry
standard for Big Data



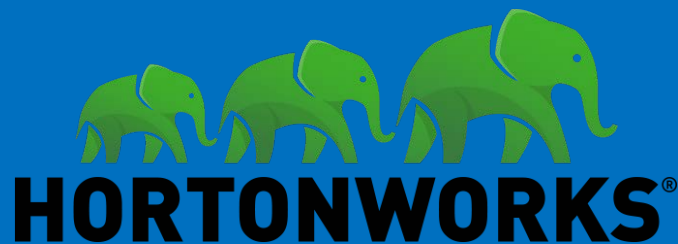
Modern
architecture &
service



Open source:

- 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

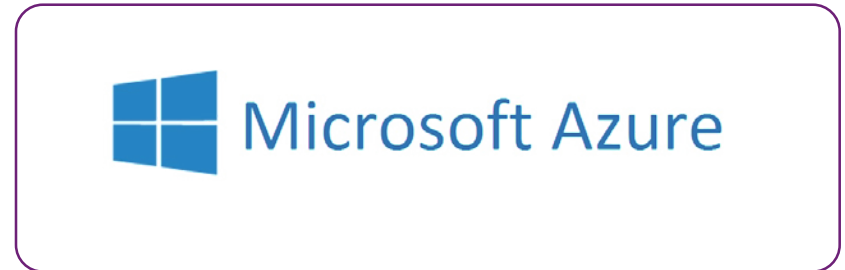
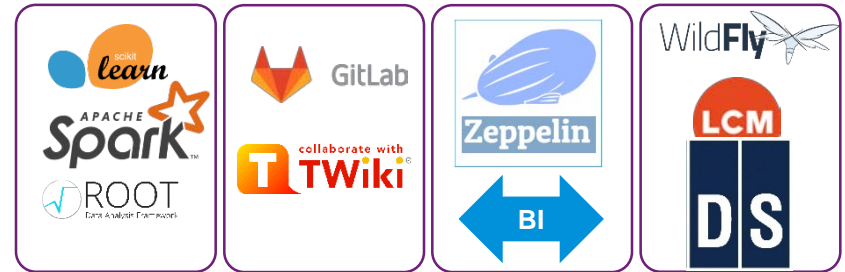
BI/visualization
integration

Web interfaces

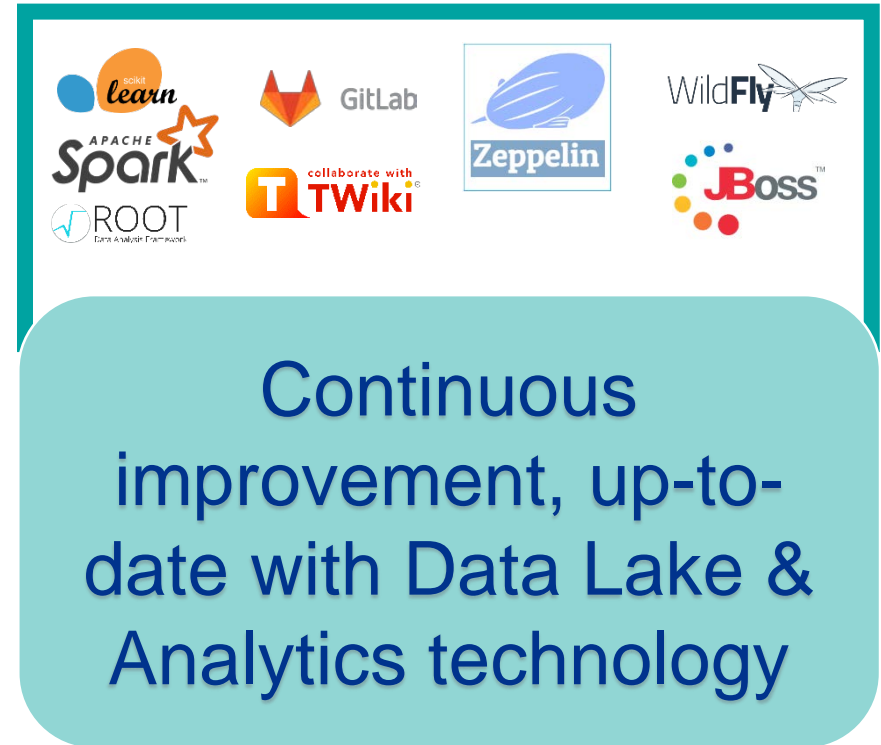
Integrated security
layer

Automated installation
on Microsoft Azure

KAVE: extension of the HortonWorks Hadoop distribution



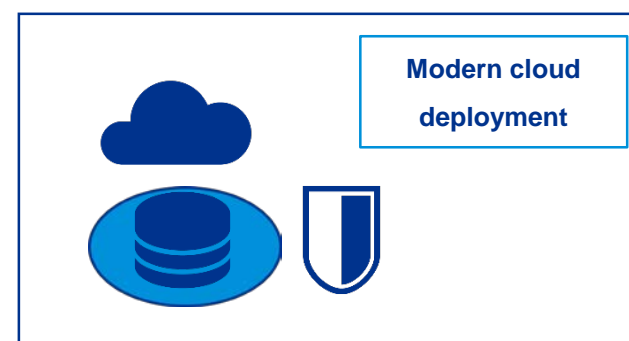
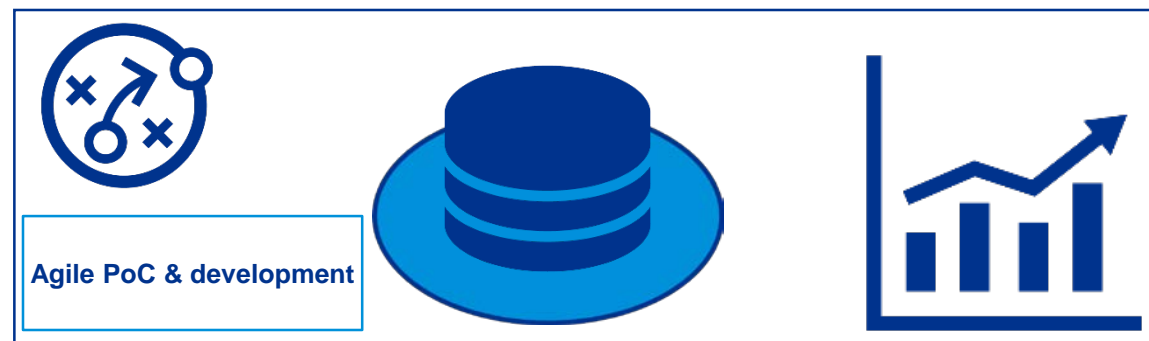
KAVE: extension of the HortonWorks Hadoop distribution





KAVE & the fulfillment of the Data Lake evolution

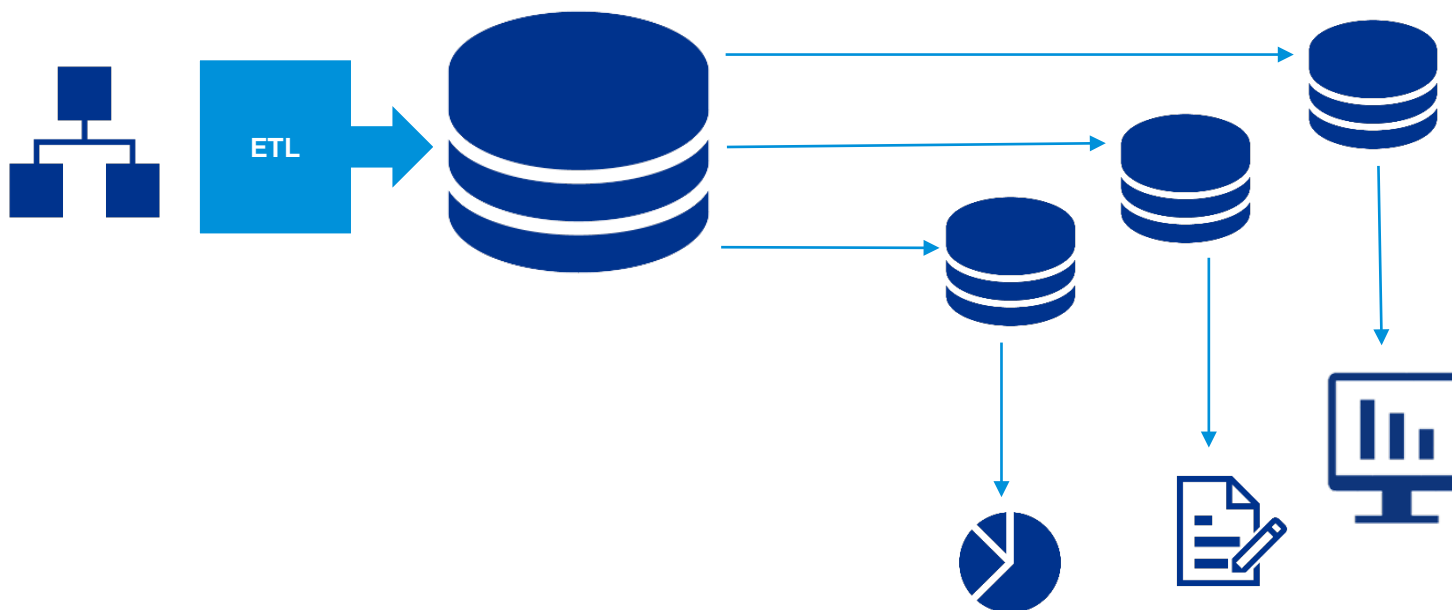
Enterprise Data Lake: topics & directions





Data Warehouse & Business Intelligence functionalities

The traditional DWH/BI stack



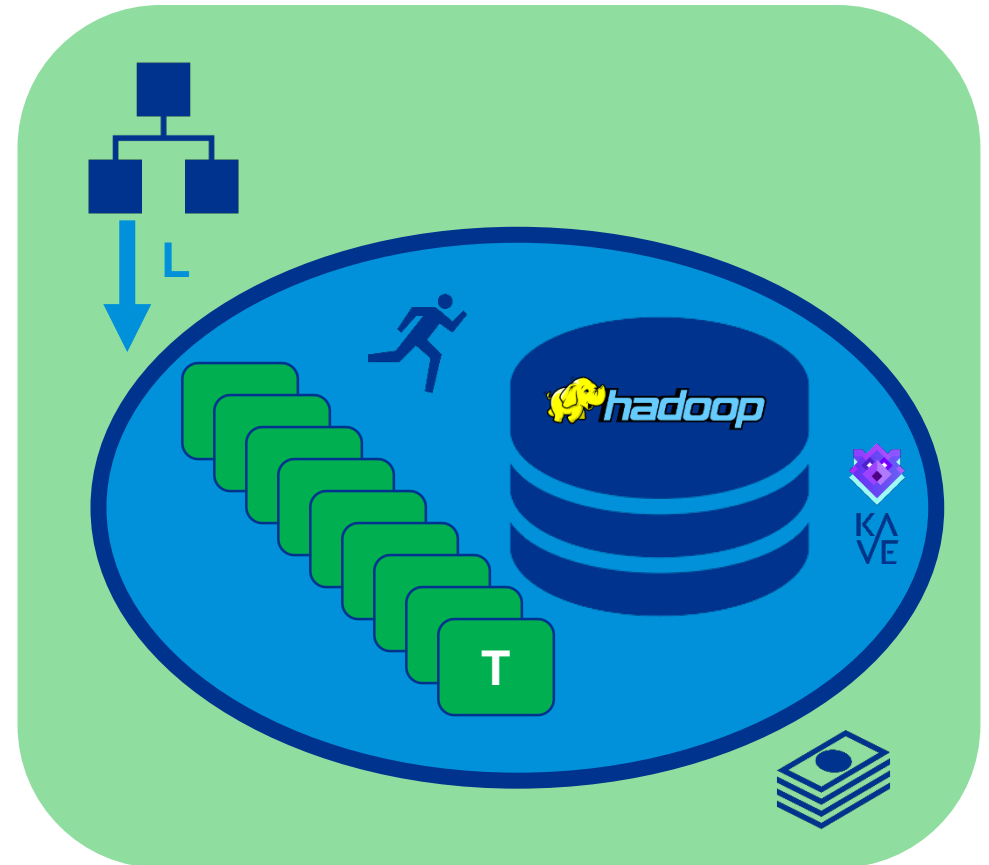
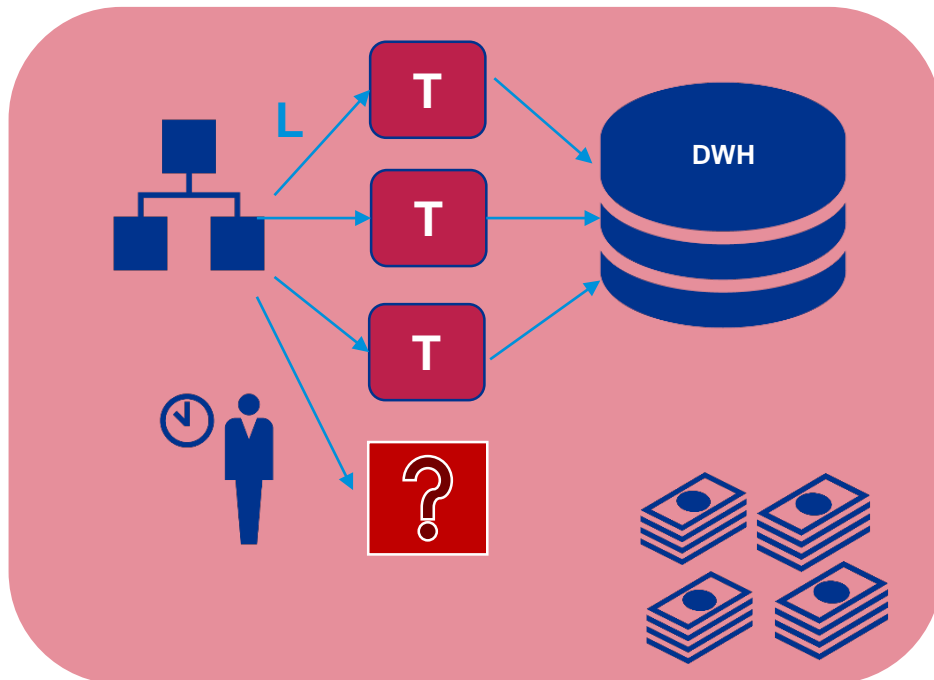
Traditional DWH/BI stack: capacity scale



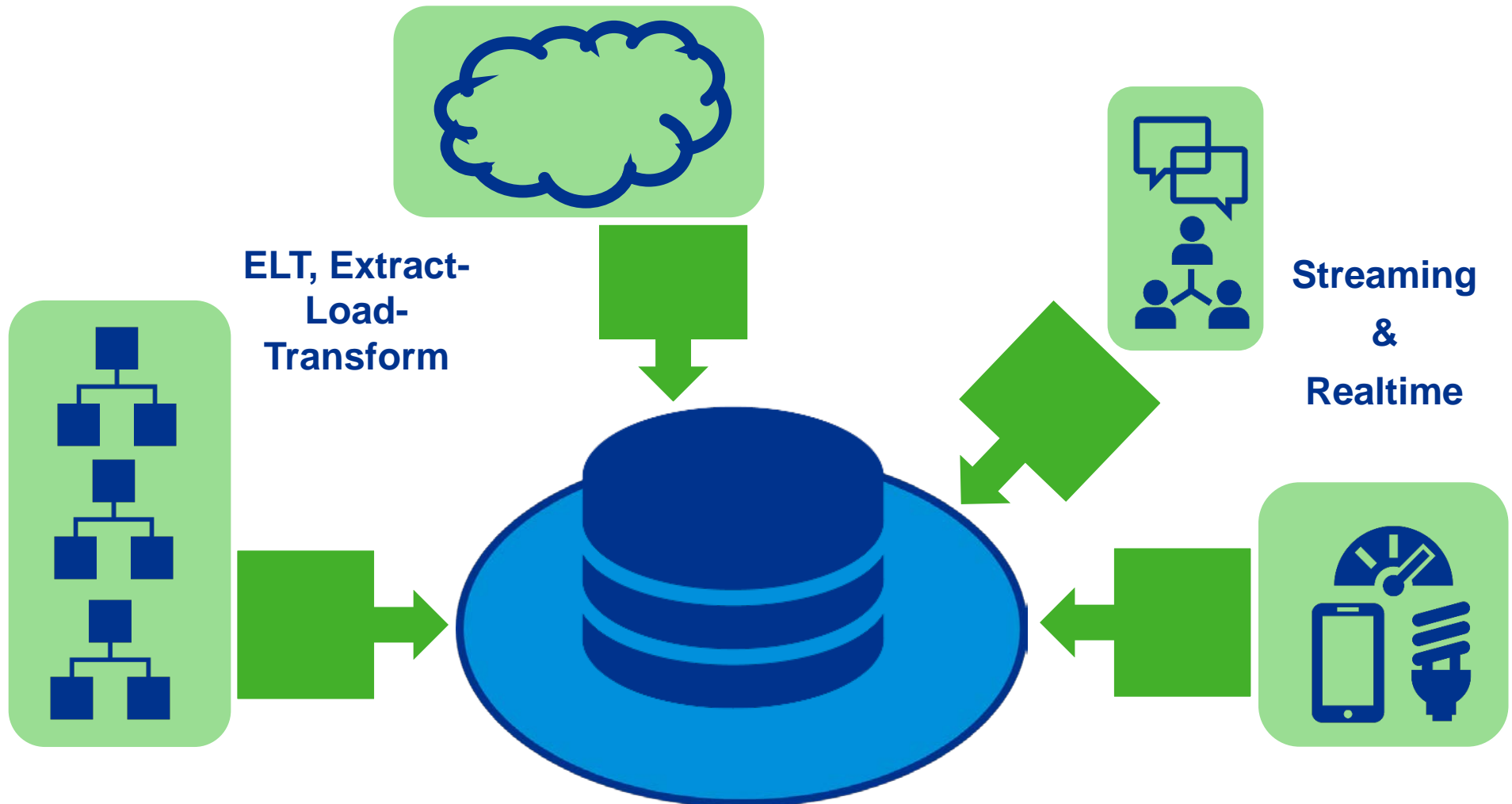
- **Costs ?**
- **Performance ?**
- **SQL-only ?**



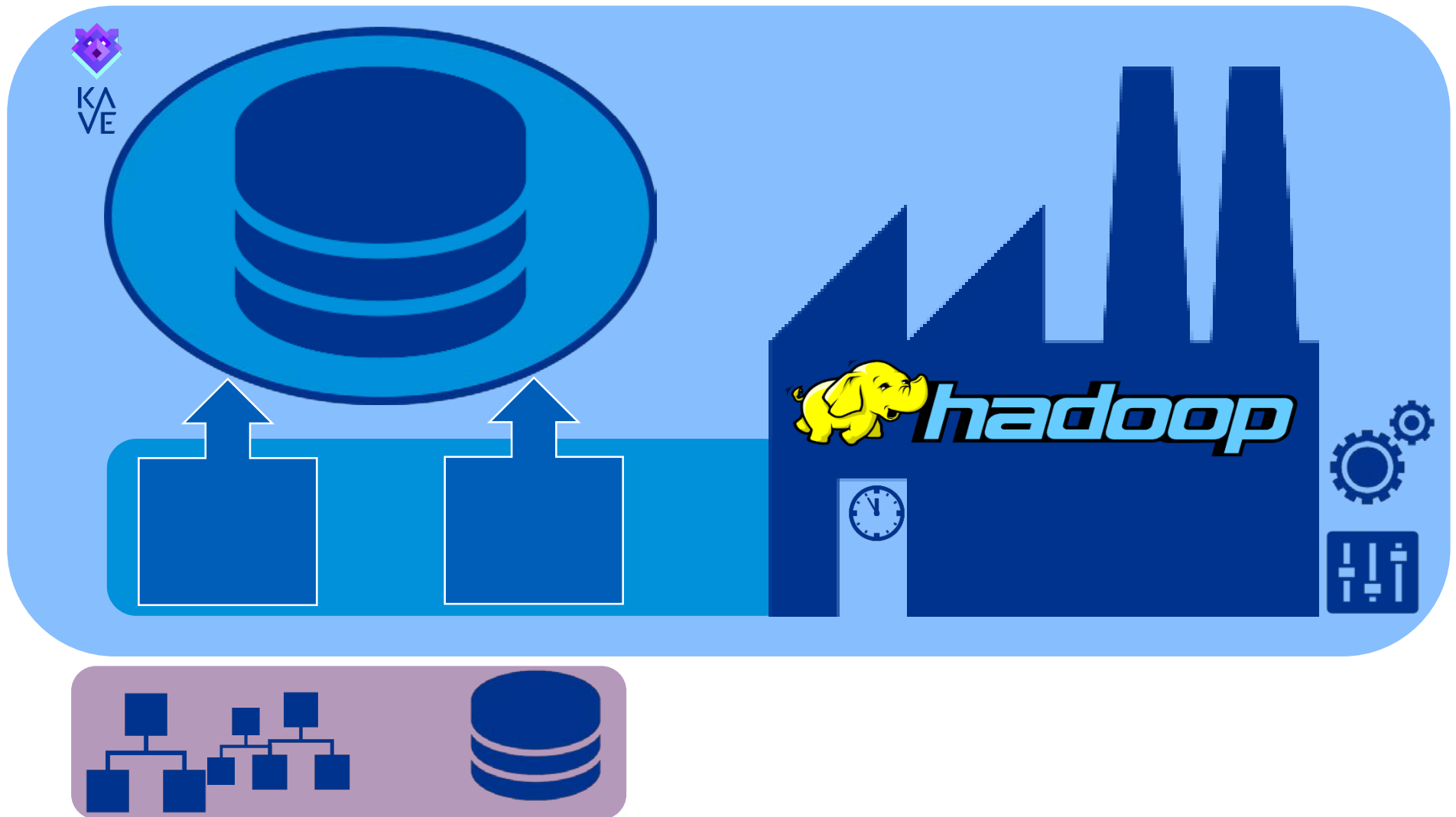
Enterprise Data Lake: ELT scaling in KAVE



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
heterogeneous data
sources (logs, queues,
files, webpages...)



Apache Atlas



Integrated and automatic metadata creation and management



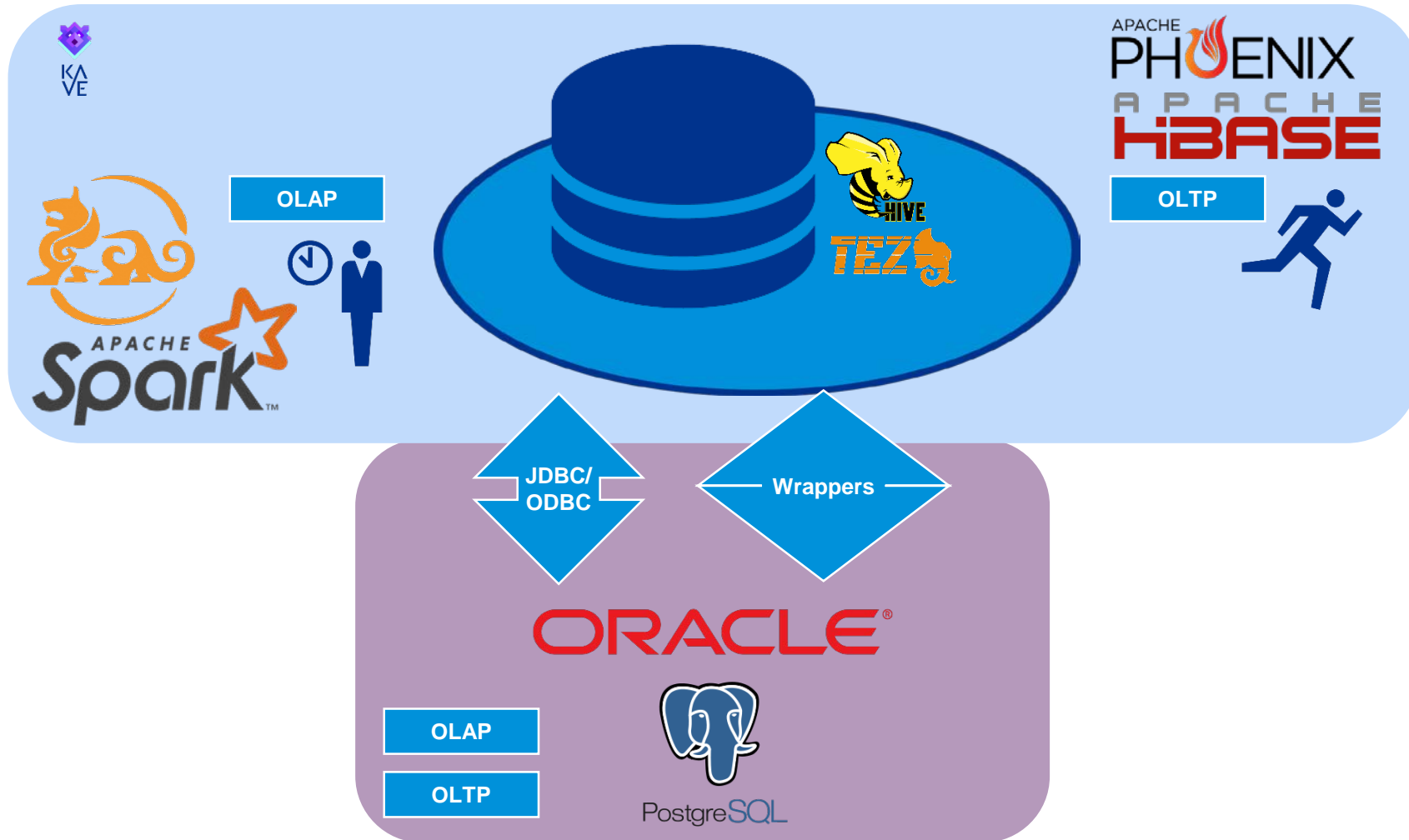
Apache ORC™



Parquet

Advanced and optimized Hadoop storage formats

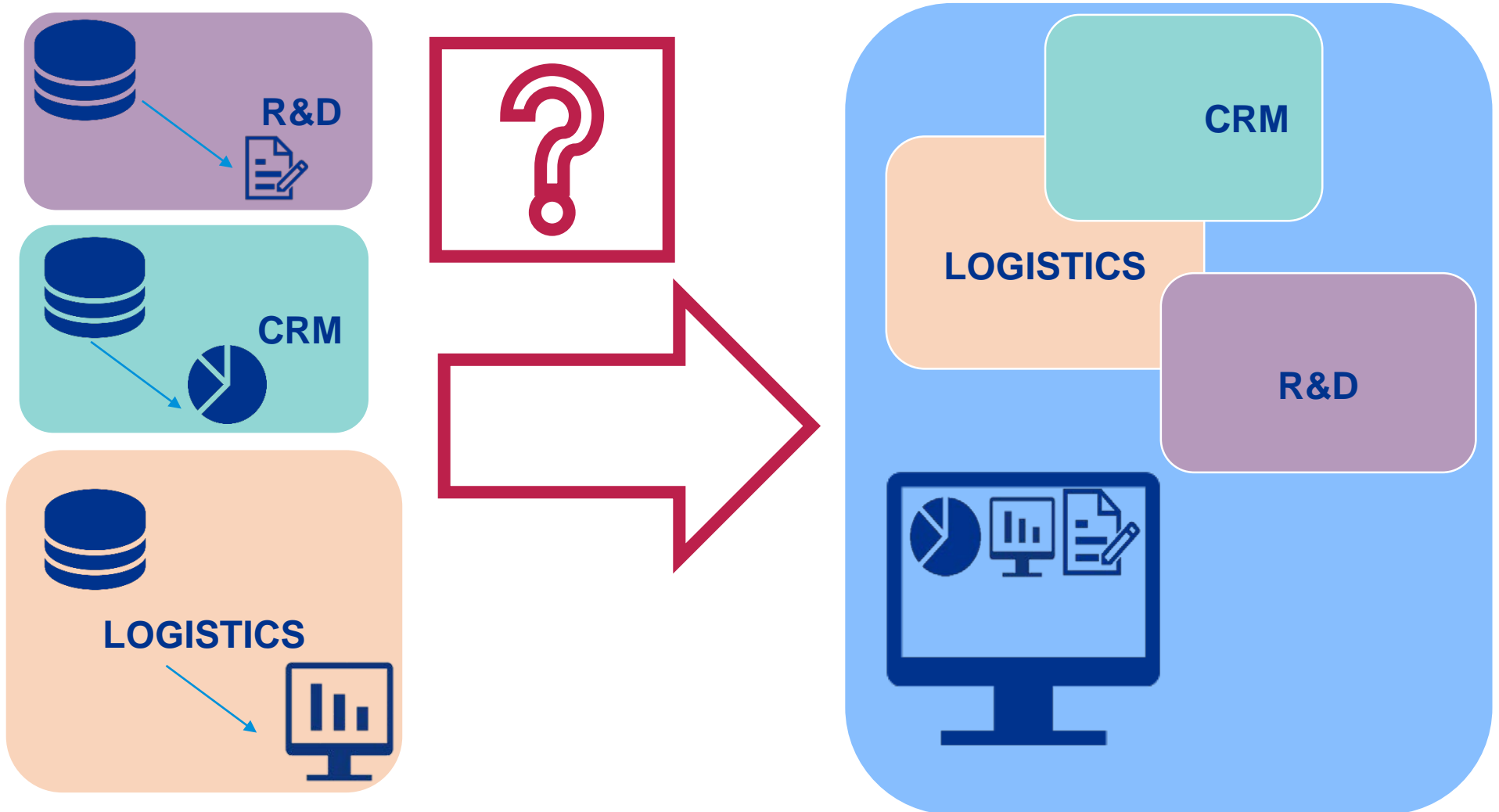
Enterprise Data Lake: OLAP & OLTP workloads



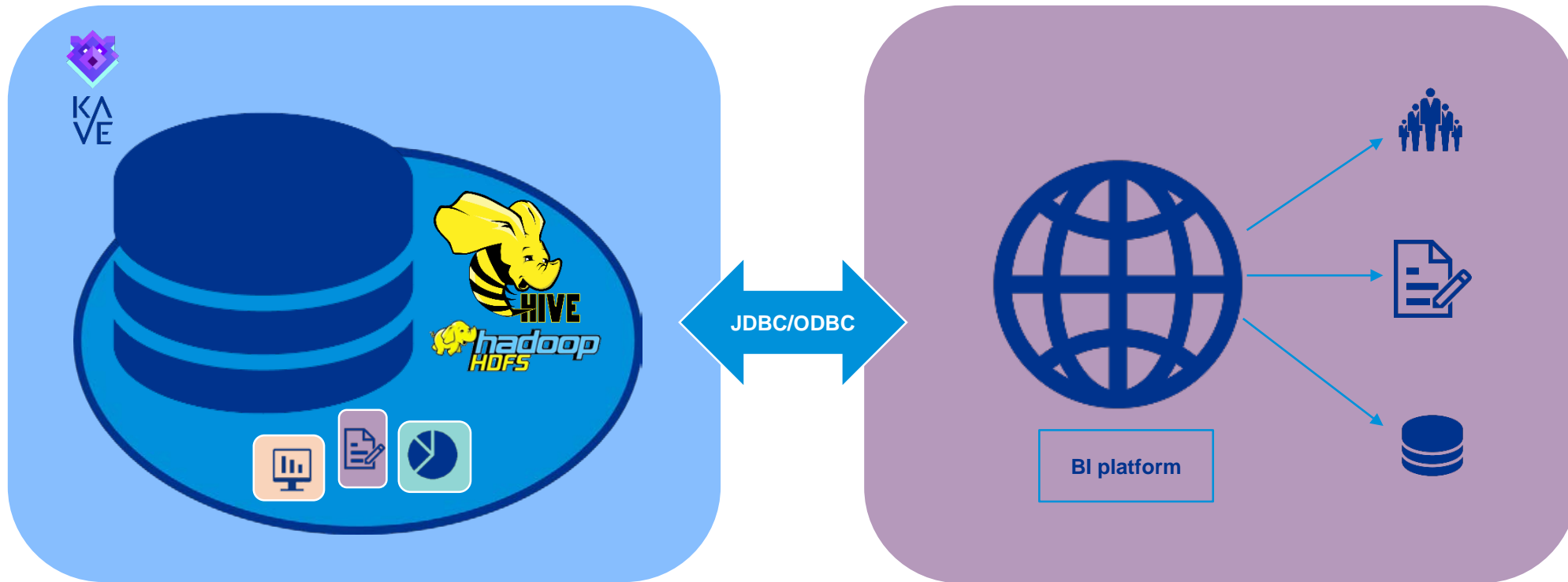
KAVE: OLAP & OLTP workloads



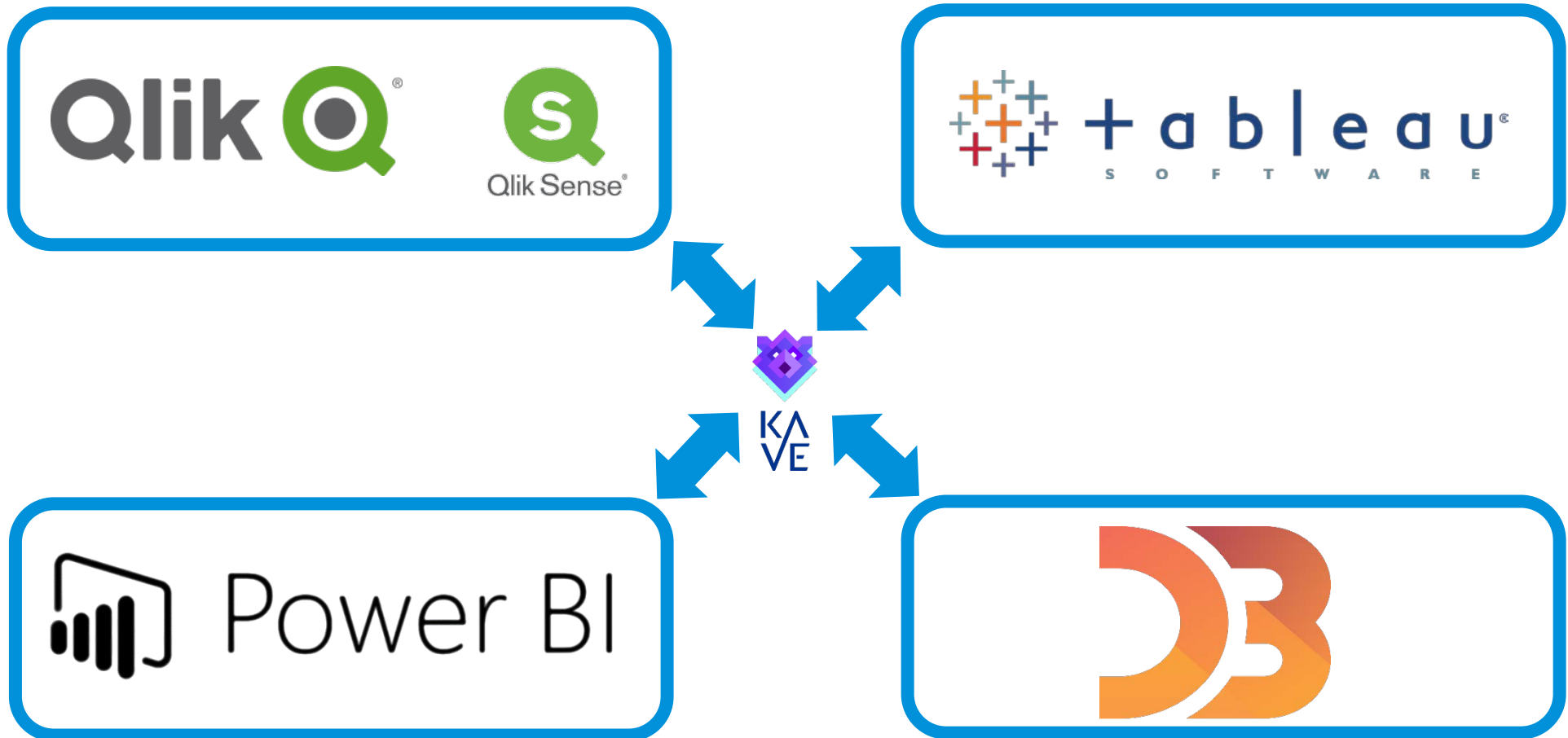
BI & BigData: are we there yet?



KAVE: reports & visualization



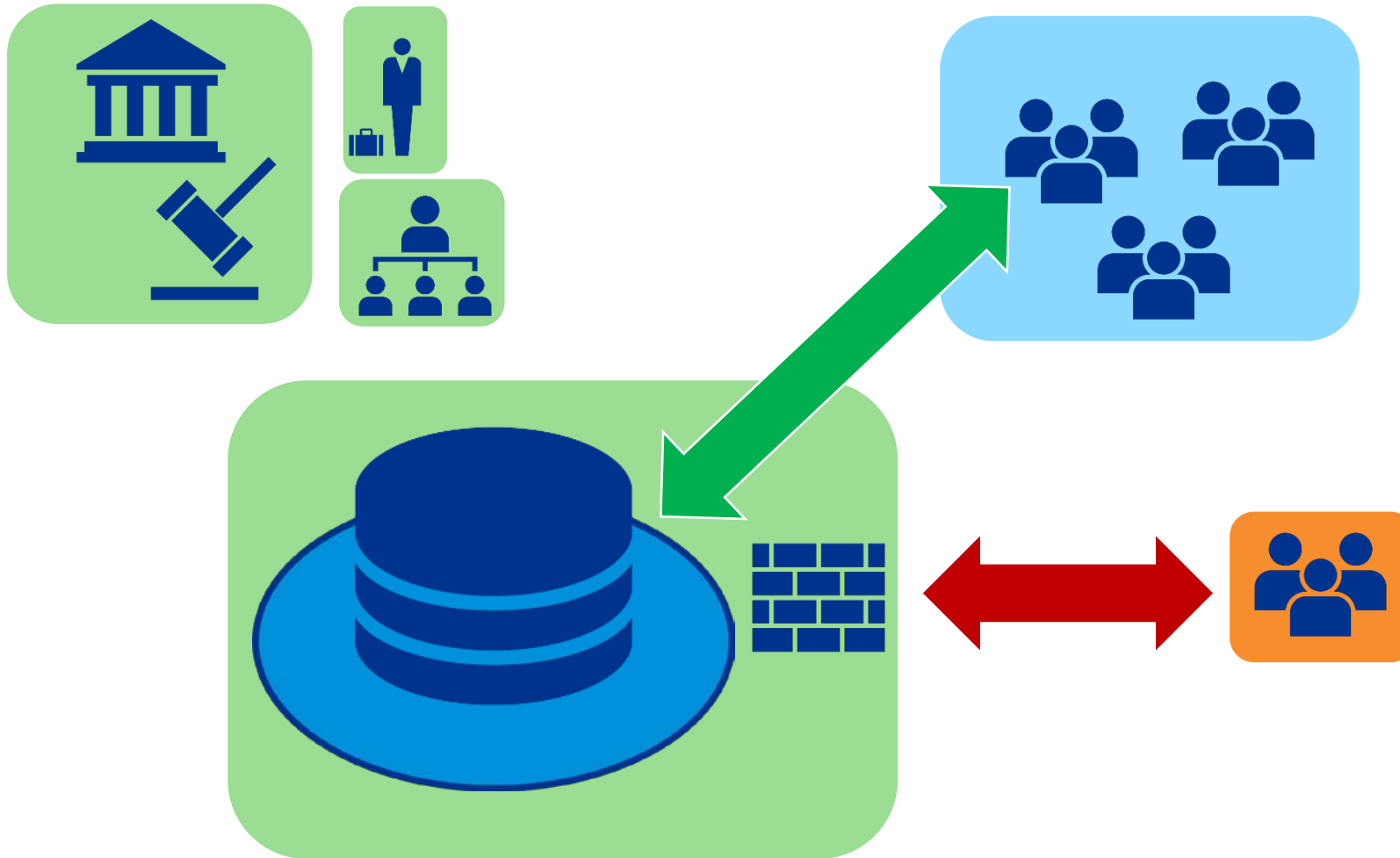
KAVE: reports & visualizations



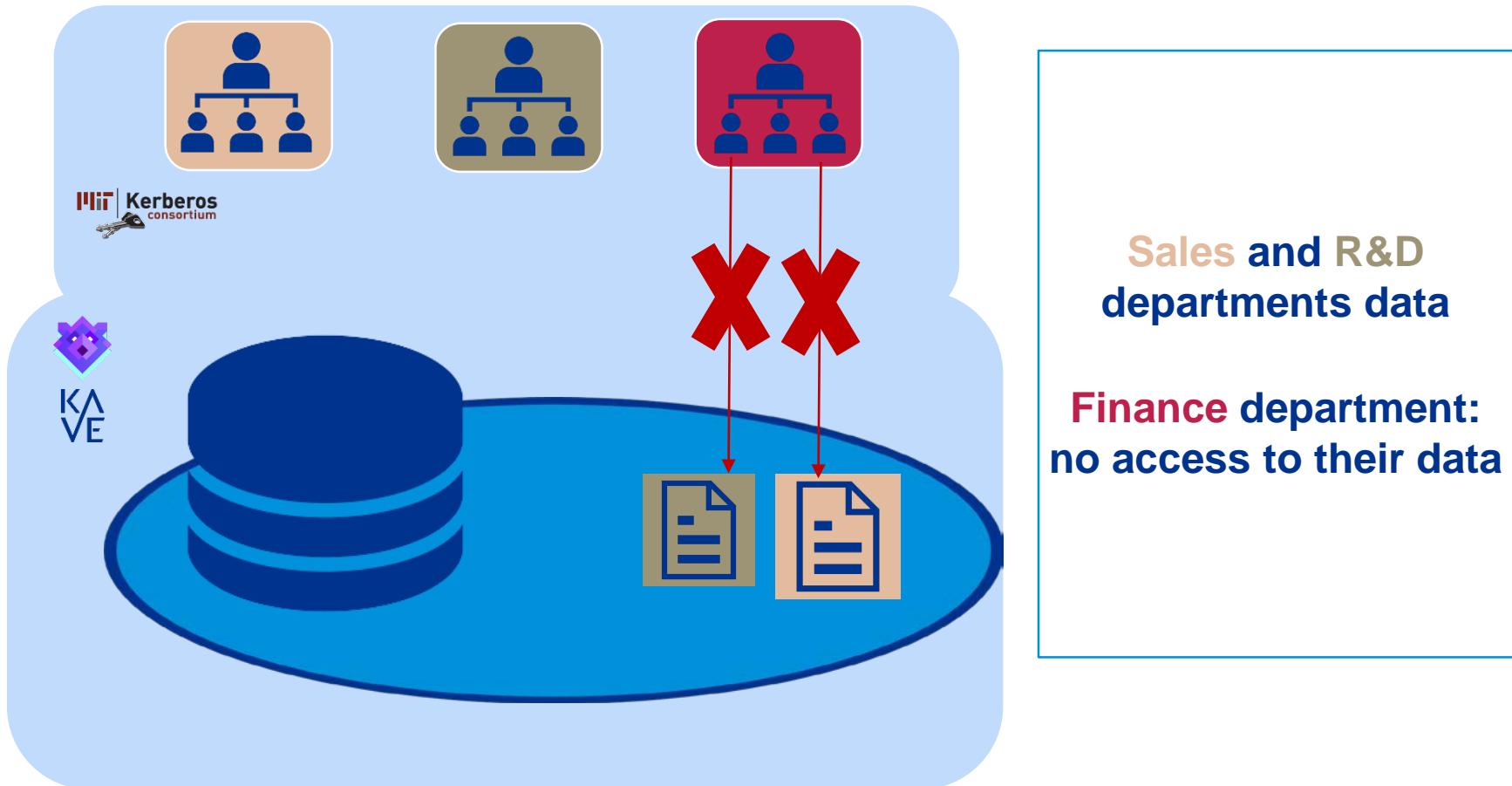


Controlling the access and usage of the data

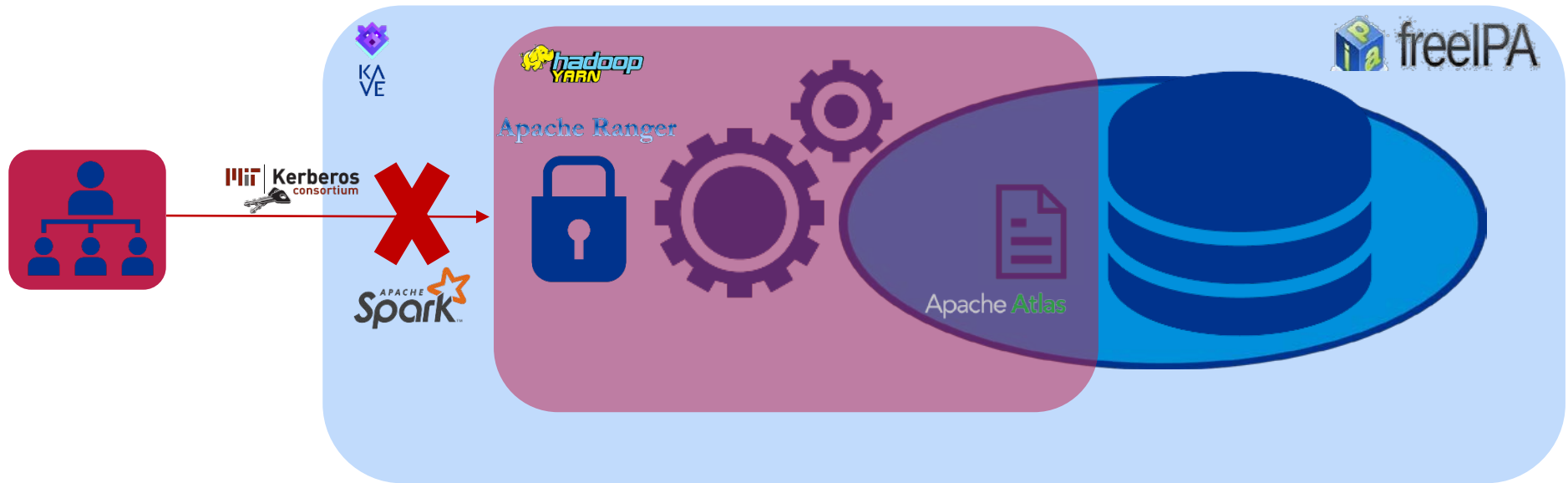
Enterprise Data Lake: security & governance



Enterprise Data Lake: security & governance



Enterprise Data Lake: security & governance



Finance department
cannot run Spark on
test cluster

KAVE: full data management on secured infrastructure

Apache Ranger

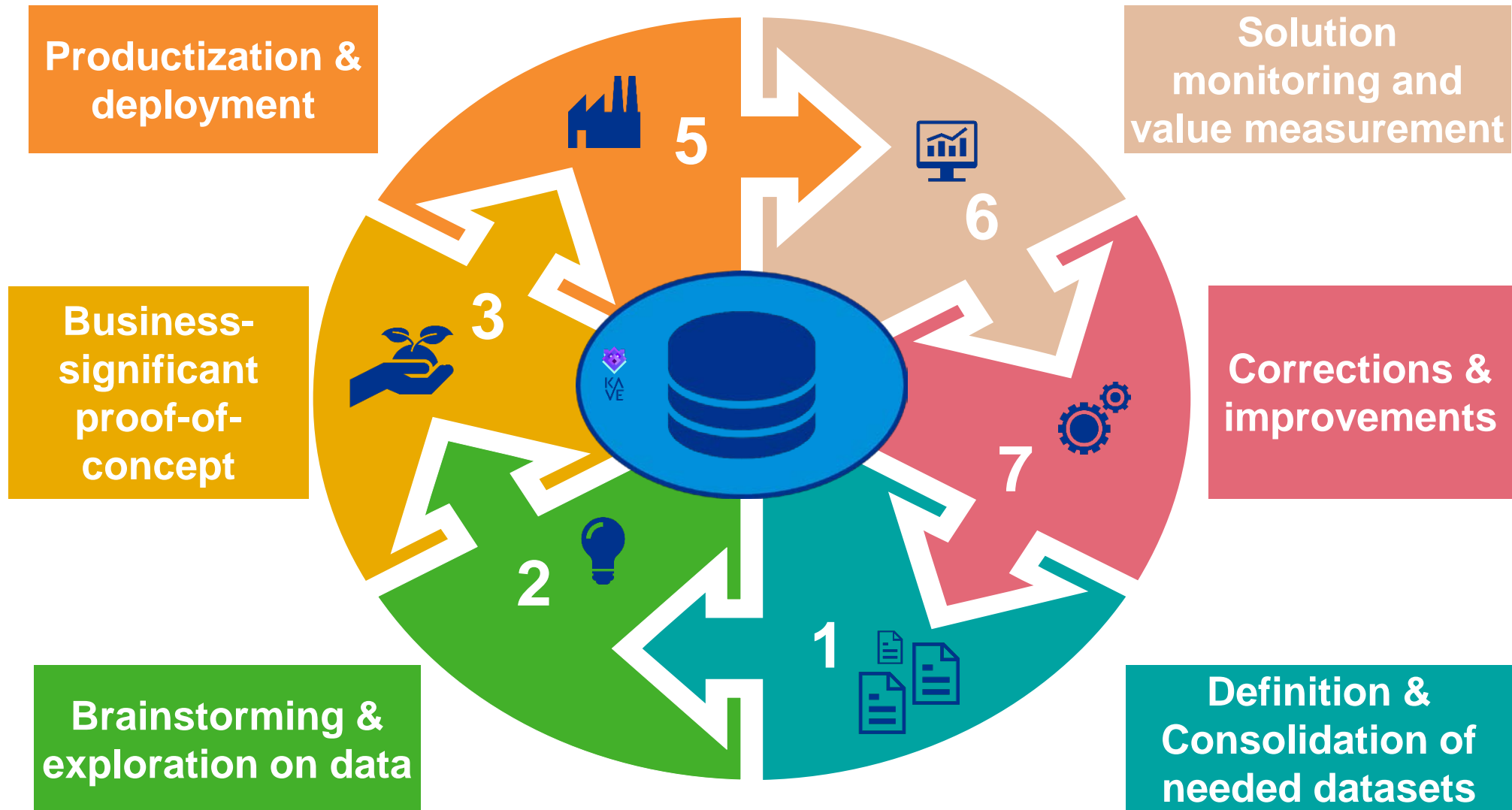
Apache Atlas



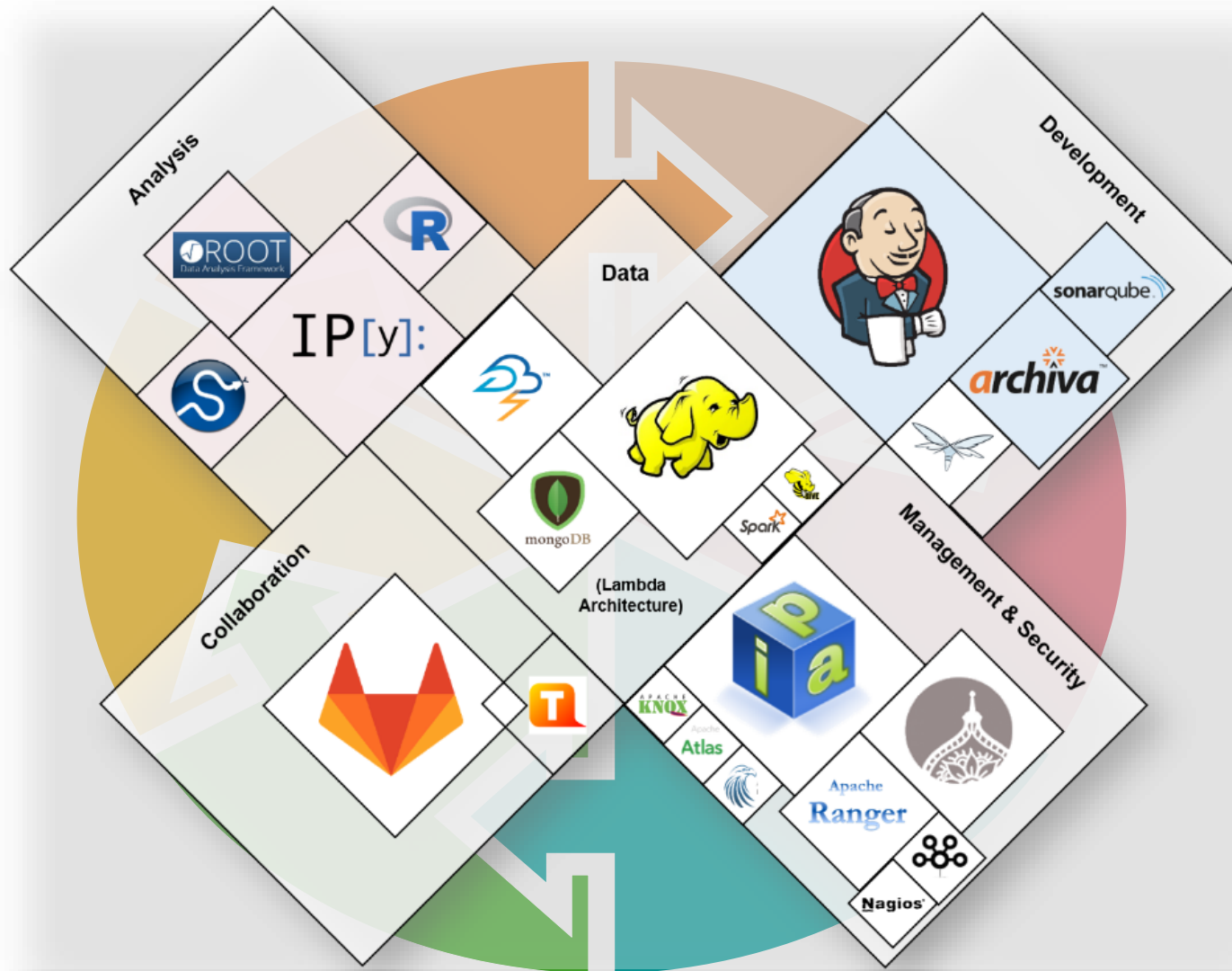


From experiments to production

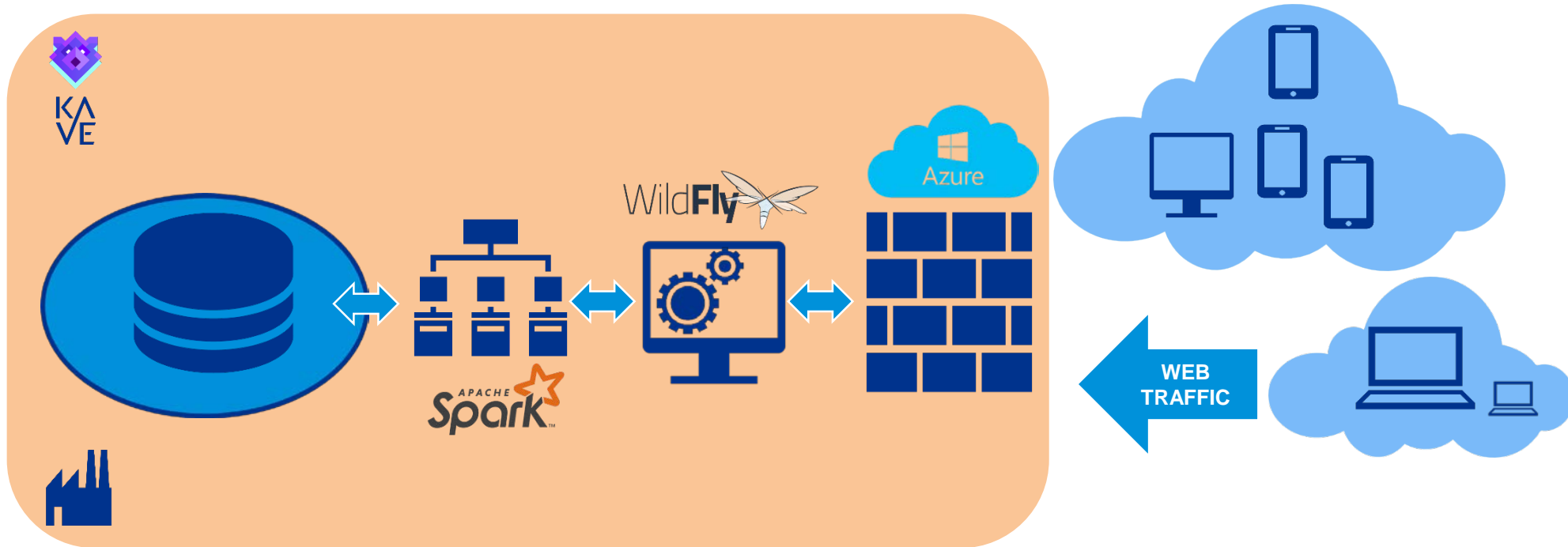
Data-centric software products with cycle optimization



KAVE & data-centric development model: a glance



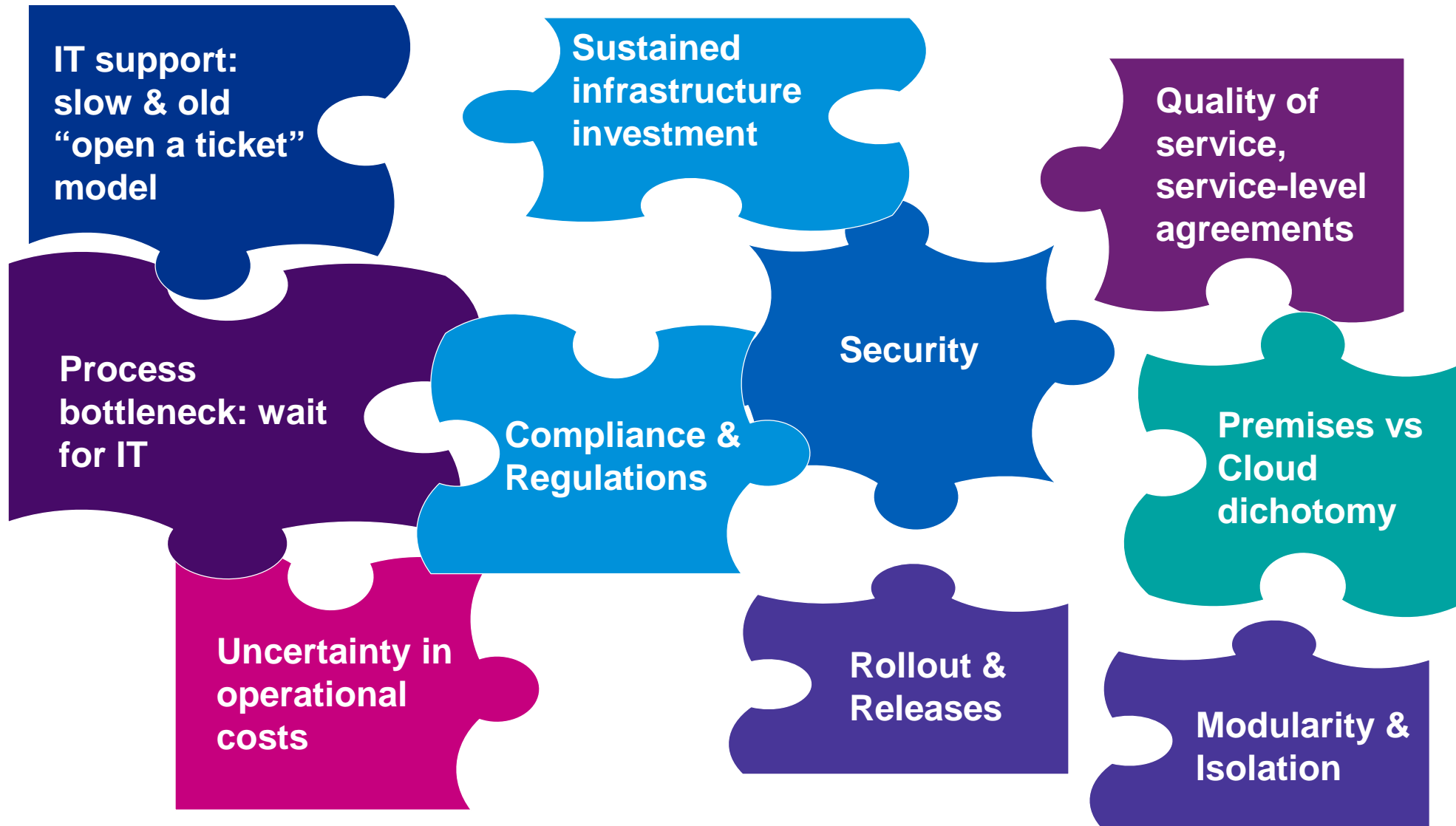
Prototype data product deployment for the web



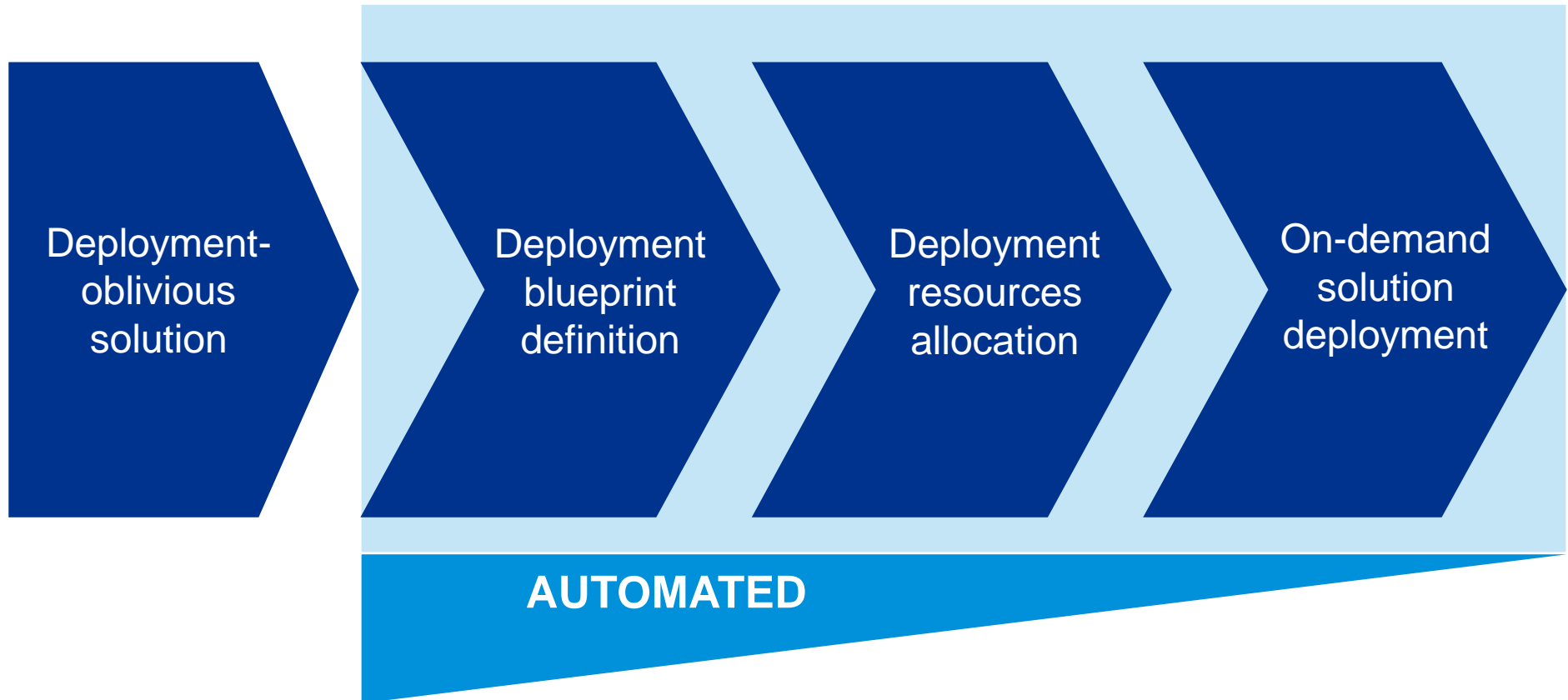


The modern Cloud experience

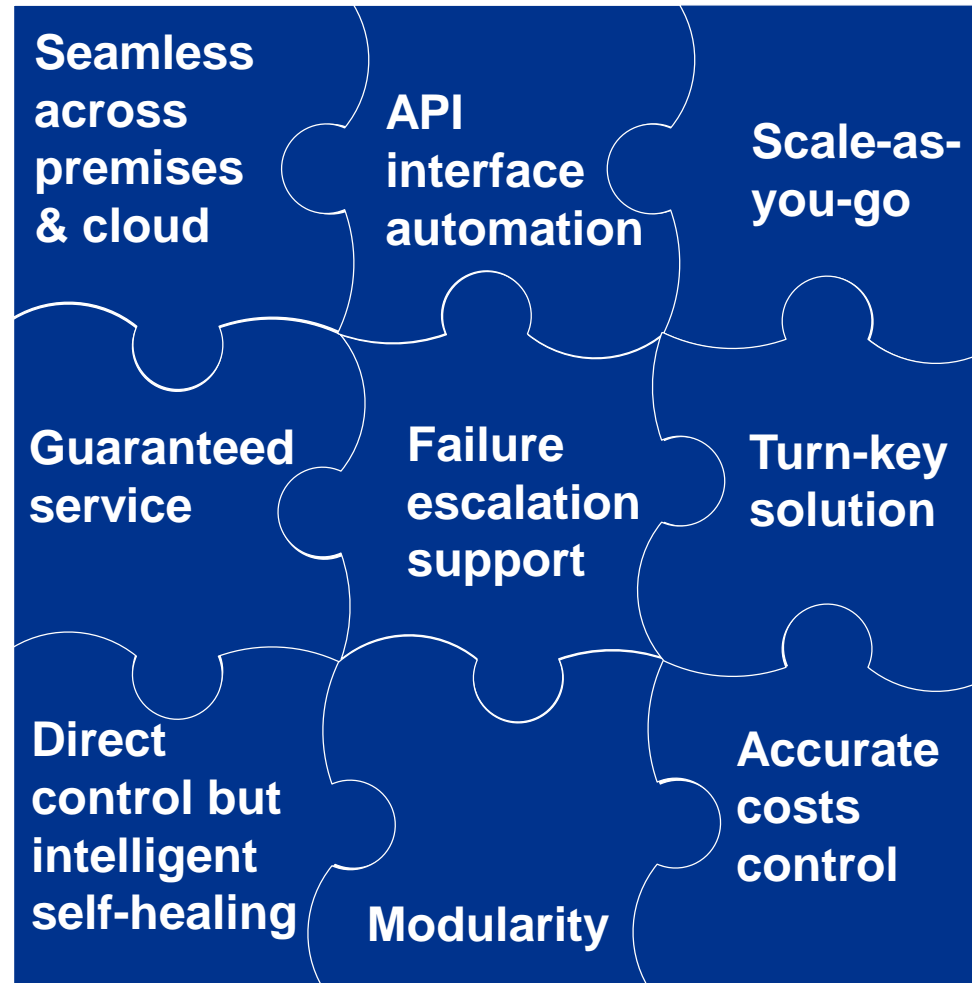
The many unknowns of low-automation infrastructure



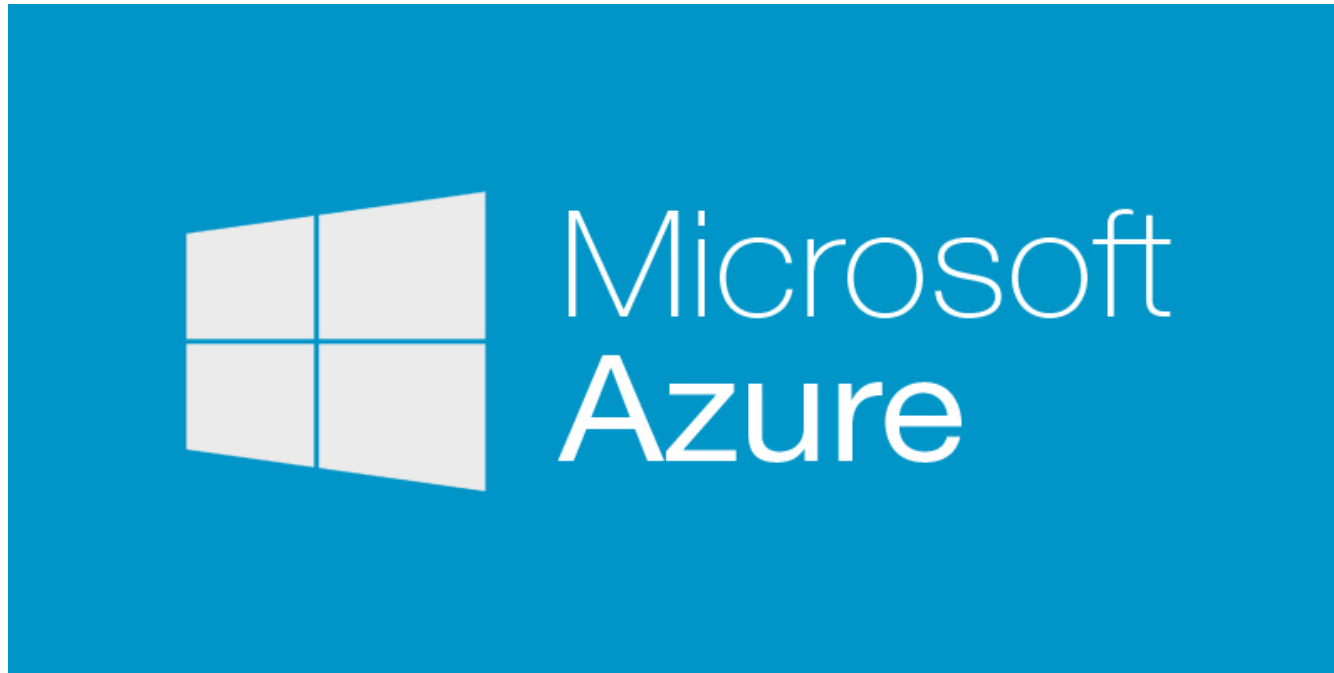
Satisfying the data product cycle: continuous delivery



The need for an integrated, dynamic and automated infrastructure



The preferred infrastructure model for KAVE: Azure





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

Microsoft Azure Resource groups > kaveassets

kave@kpmg.com KPMG NL

kaveassets Resource group

Search (Ctrl+ /)

Overview

Activity log

Access control (IAM)

Tags

SETTINGS

+ Add Columns Delete Refresh Move

Essentials ^

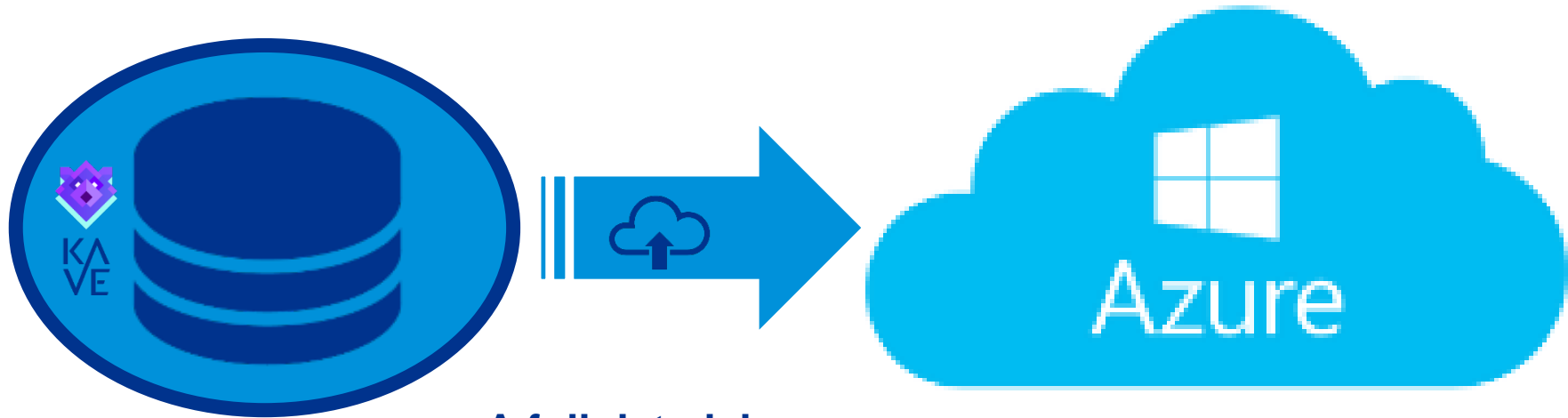
Subscription name (change)	Subscription ID
Big Data – Project KAVE	a31a9a22-ae1c-4dea-8086-7a498f64a4
Deployments	Location
1 Succeeded	West Europe

Filter by name...

29 items

NAME	TYPE	LOCATION
------	------	----------

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

**Deploy mixed on-
prem / remote
solutions**



**Scale the solution
in no time with a
few clicks**



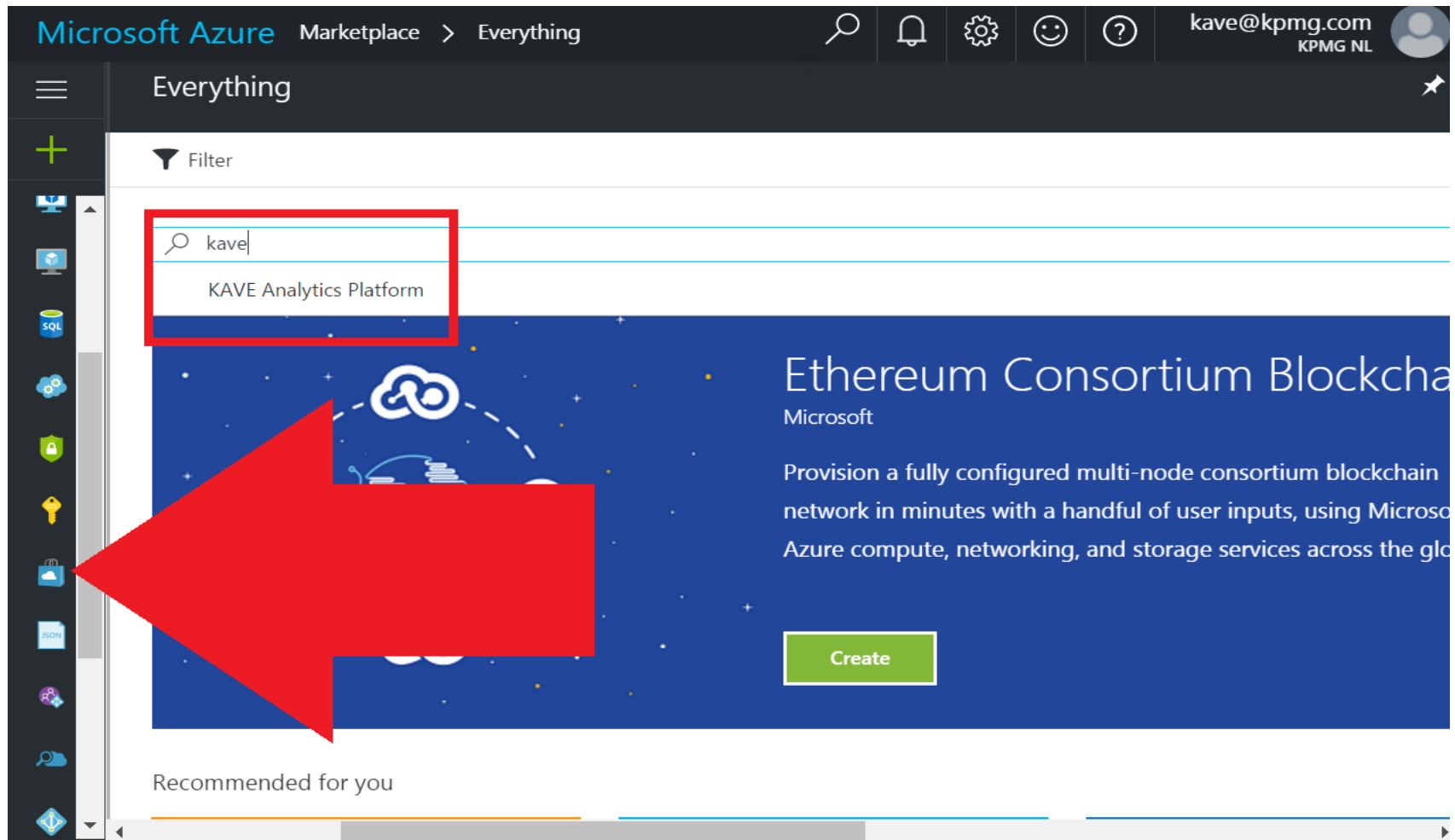
**Agile for solution
PoC, enterprise-
class for
production**



**Save budget with
exact billing and
direct systems
access**



Try a fully working KAVE instance on the Azure marketplace!



Try a fully working KAVE instance on the Azure marketplace!

The screenshot shows the 'Create KAVE Analytics Platform' wizard in the Azure portal, specifically the 'Basics' step. The left sidebar contains a navigation pane with five steps: 1. Basics (selected), 2. Storage, 3. DNS name label, 4. Virtual machines, and 5. Summary. The main area displays the 'Basics' configuration form. It includes fields for 'KAVE admin' (password), 'Password', and 'Confirm password', each with a red asterisk indicating it is required. Below these is a 'Subscription' dropdown menu set to 'Big Data – Project KAVE'. The 'Resource group' section has a radio button selected for 'Create new'. The 'Location' dropdown is set to 'West Europe'. A blue 'OK' button is at the bottom of the form. The top of the portal shows the 'ft Azure' logo, the breadcrumb 'Create KAVE Analytics Platform > Basics', and the user profile 'kave@kpmg.com KPMG NL'.

ft Azure « Create KAVE Analytics Platform > Basics

Create KAVE Analytics Platfo... Basics

1 Basics
Configure basic settings

2 Storage
Configure storage settings

3 DNS name label
Configure DNS prefix

4 Virtual machines
Configure nodes

5 Summary
KAVE Analytics Platform

* KAVE admin

* Password

* Confirm password

Subscription
Big Data – Project KAVE

* Resource group
☒ Create new

Location
West Europe

OK

Open source: contribute & extend to fit your needs!



<http://beta.kave.io>



Thanks !