Data Analytics Platform

or How to Make Data Science in a Box Possible

Krzysztof Adamski ING WBAA

Moscow, 11th October 2018

thinkforward

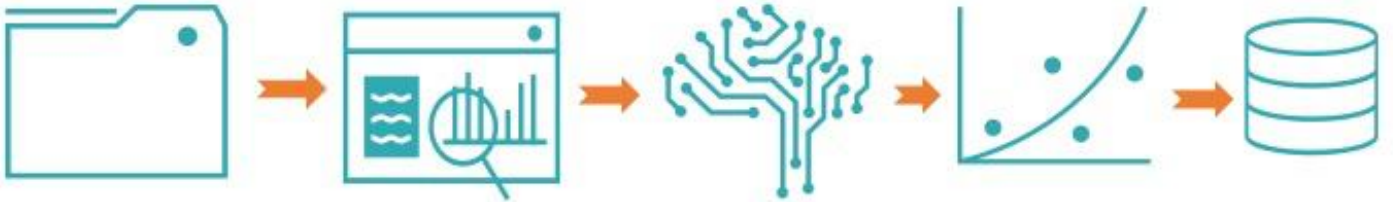
ING

What am I doing here

WHOLESALE BANKING
advanced
analytics

- Data Infrastructure Architect
- Focused on enabling data-driven decisions
- Main areas:
  - Large scale data storage and processing – this has to run somewhere
  - Data platforms security – we are a bank and trust is the biggest asset
  - Sharing the knowledge – as a nature is complex

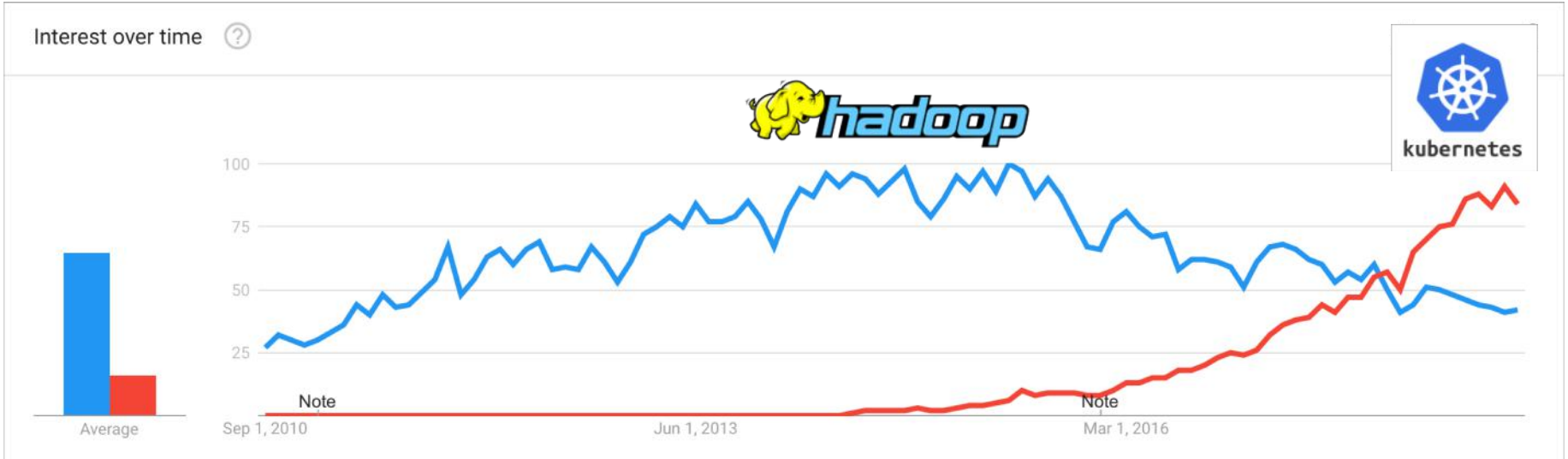
Data Engineering Pipeline



databricks.com

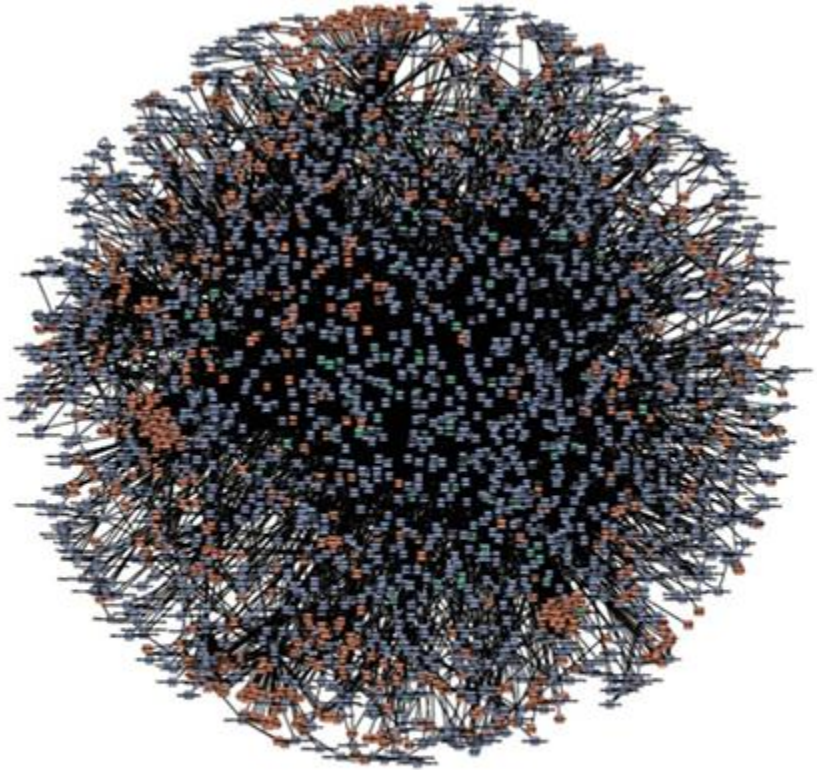
The new post Hadoop world

- Container based
- Need a more agile adoption of a growing set of technologies
- Stronger isolation is required
- Cloud-Readiness as an integral part of analytics frameworks



Hadoop as an isolated island

And the world nowadays looks like this



amazon.com

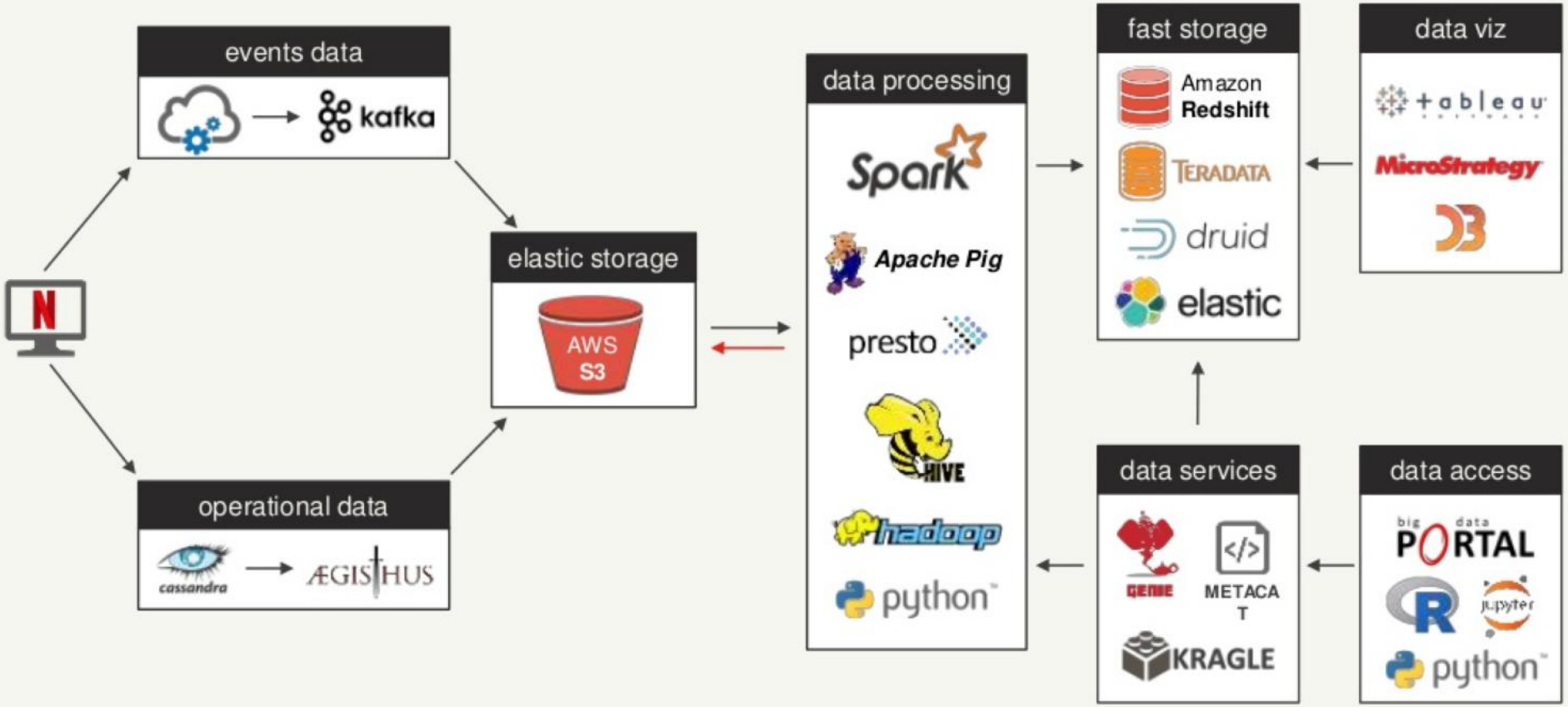
NETFLIX

Netflix scale

Around 700 microservices to control each of the many parts of what makes up the entire Netflix service

https://medium.com/refraction-tech-everything/how-netflix-works-the-hugely-simplified-complex-stuff-that-happens-every-time-you-hit-play-3a40c9be254b

Big Data Platform




Michelangelo
Jeremy Hermann, Machine Learning Platform @ Uber
UBER


aws


cloudera


dA
PLATFORM


Azure


HORTONWORKS


MAPR


Lightbend
Reactive Platform
A Unified Platform for Reactive Application Development


ORACLE
Data Cloud


TERADATA


dremio

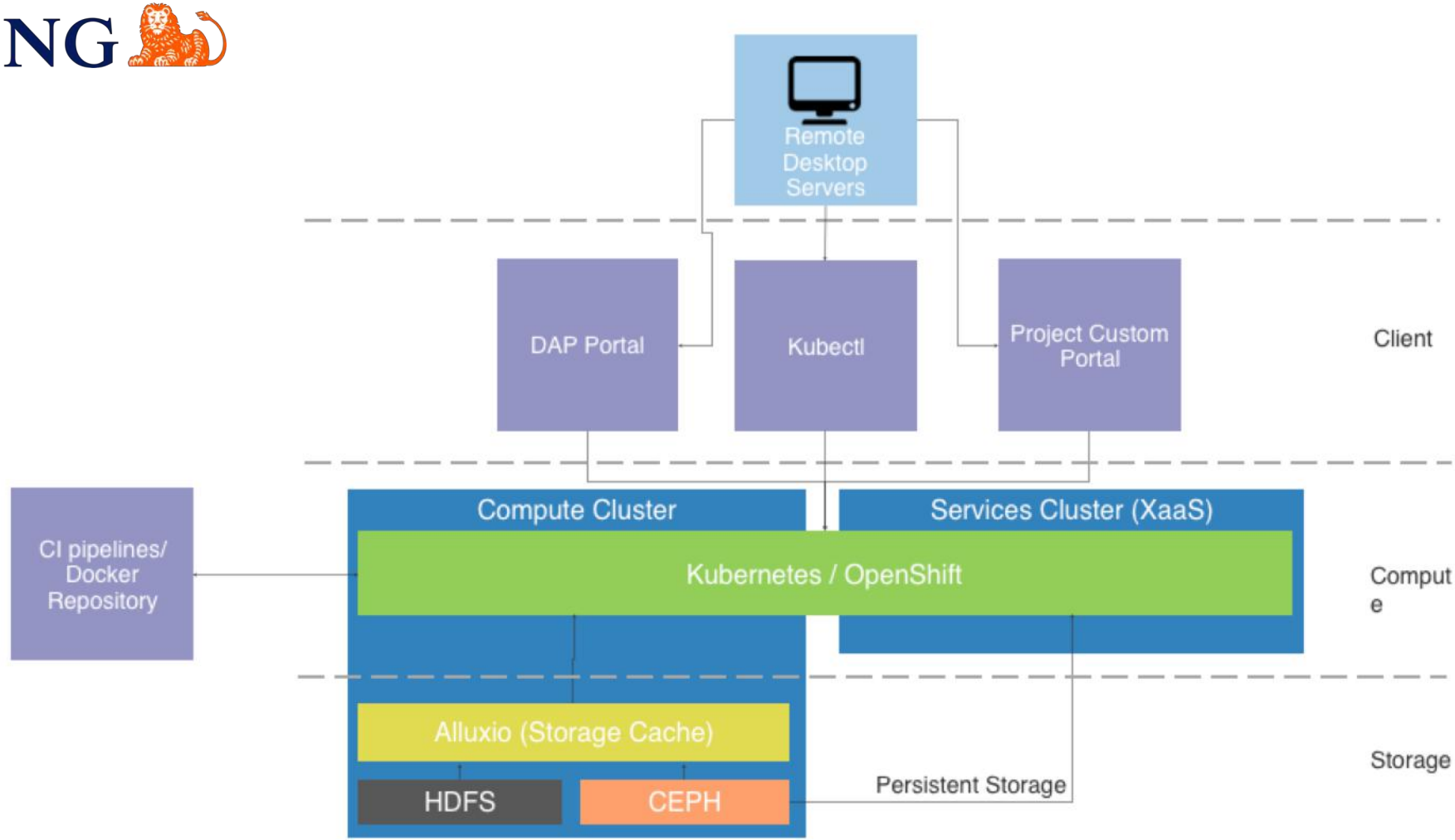

Google Cloud

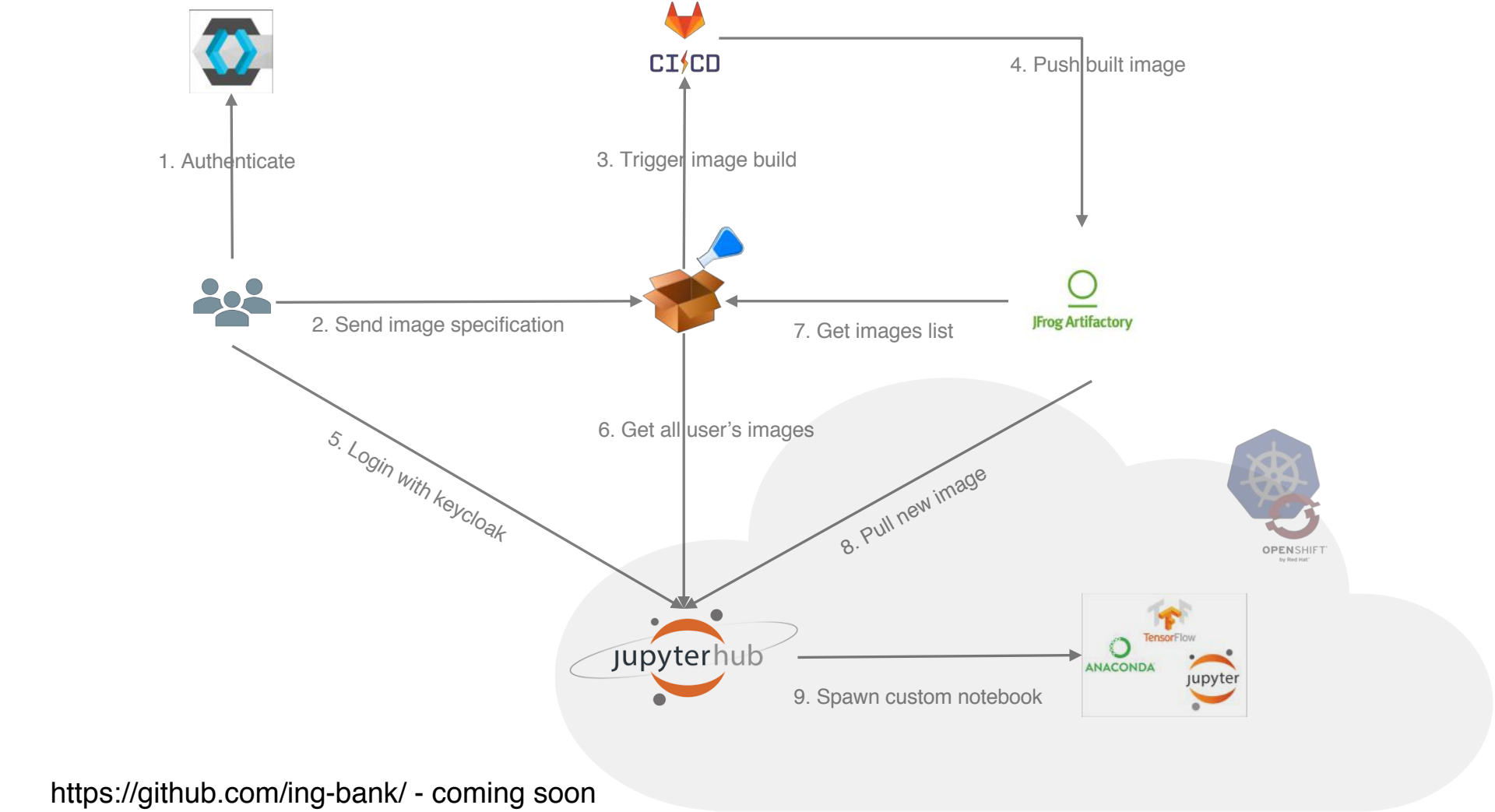

databricks


IBM WATSON

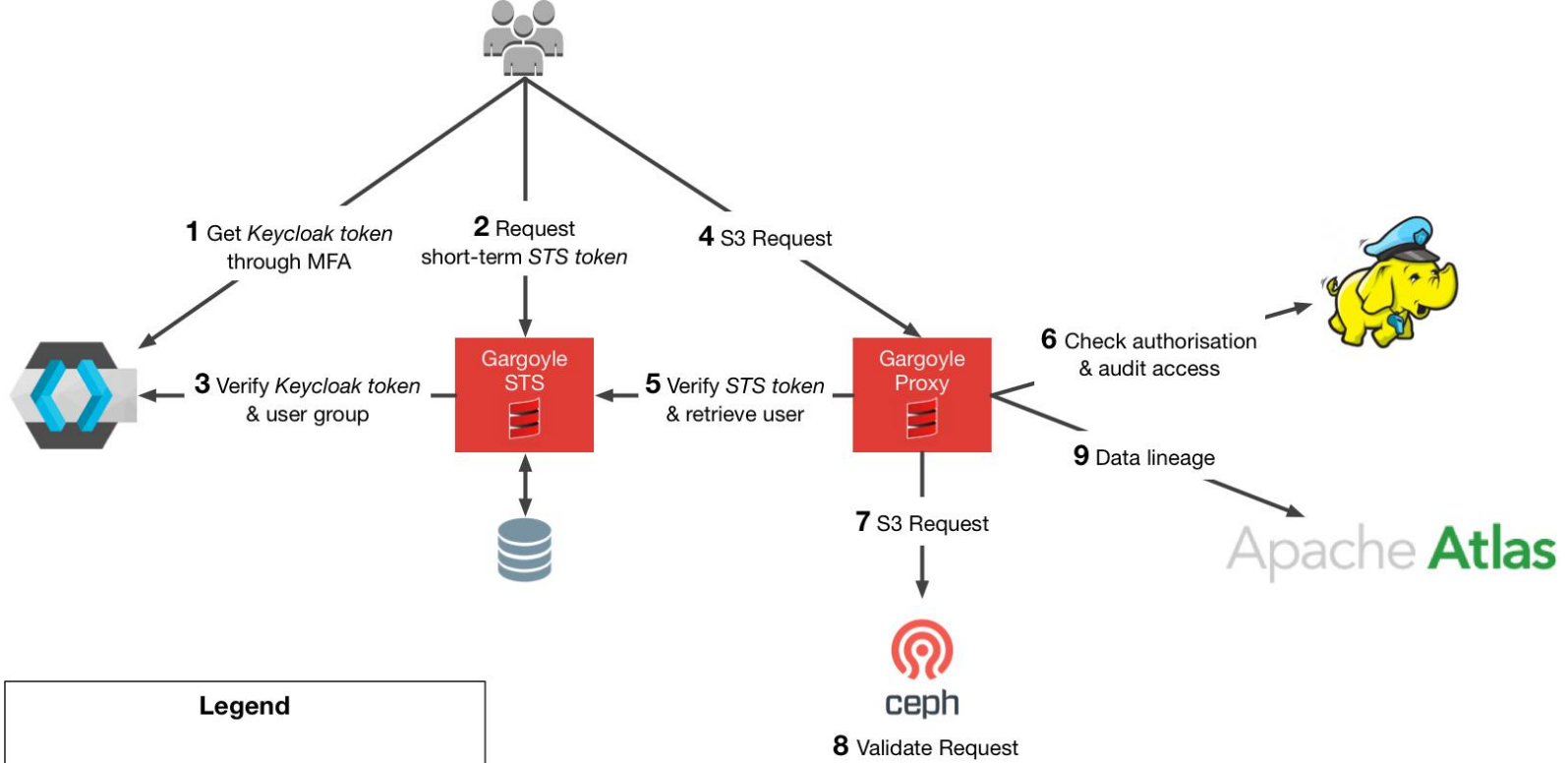
I see platforms everywhere

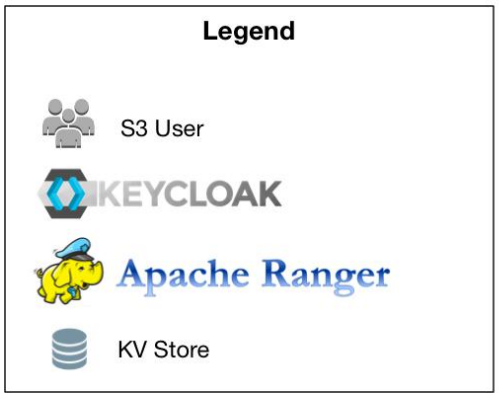

Team & Expansion
On average we receive one application every 15 minutes
Business Developers
User Experience Designers
Data Engineers
Data Scientists
Software Engineers
16 nationalities
100+ people
Madrid
Southern European Hub
Where our people come from
Data Talents
65 NL
9 PL
5 (23) LDN
kaggle MASTER
Google
Thought Leadership
Cancer research hackathon
Erasmus MC
Inspiration workshops and translator training
WAR child
Eneco
WORLD SUMMIT AI AMSTERDAM
Experimentation days
umcu
bugaboo
Code contributions
Apache Airflow
Apache Spark
10 commandments
AI Master Thesis Supervisors
Data Ethics
ENTITY MATCHING
PRIVITAR
Safe sharing data with 3rd parties
PPI
Possible Private Individual Protection
PEER ANALYSIS
DATA SCIENCE IN A BOX
IADO
Realtime pricing in emerging market bonds (Katana)
Core Tech & Algorithms
WHOLESALE BANKING
advanced analytics
Building algorithmic data-driven products with a 10x impact.
Anchor Products
Katana Lens
Trading signals for asset management companies.
Katana Internal
Pricing analytics for bond traders.
450+ users
Domino
Network analytics for payments and loans.
Experimentation days
New Products
Intellichain
Automated supply chain analytics.
Beyond Banking
Mandoline
R.I.P.
Celebrate failures!
KYC
AXYON.AI
Loan syndication prediction.
INTELMATCH
Smart knowledge management system.
MENTORING
Startupbootcamp
Owlin
News analytics.
SPAZIODATI
Knowledge graph for credit packs.
With FinTechs


Remote Desktop Servers
DAP Portal
Kubectl
Project Custom Portal
Client
CI pipelines/ Docker Repository
Compute Cluster
Services Cluster (XaaS)
Kubernetes / OpenShift
Compute
Alluxio (Storage Cache)
HDFS
CEPH
Persistent Storage
Storage



https://github.com/ing-bank/ - coming soon





https://github.com/ing-bank/airlock

Heads up for Spark 2.4 and beyond

- Pyspark
- R
- Client mode (finally you can have your notebooks running)
- Kerberos support is almost there
- Dynamic resource allocation and external shuffle service – still pending
- Spark streaming still missing driver resillience
- Highly recommended series of articles
  https://banzaicloud.github.io/tags/spark/

KEEP CALM IT IS DEMO TIME

...and keep fingers crossed