Distributed Document Management

Brecht Loos

exellys

Tech talent integrators

KU LEUVEN

UHASSELT
KNOWLEDGE IN ACTION

ING

exellys
Tech talent integrators

About me

Table of Contents

Distributed Systems

CAP theorem

Security

Document management

Context

Document management @ ING

Architecture

Messaging

Kafka

RabbitMQ

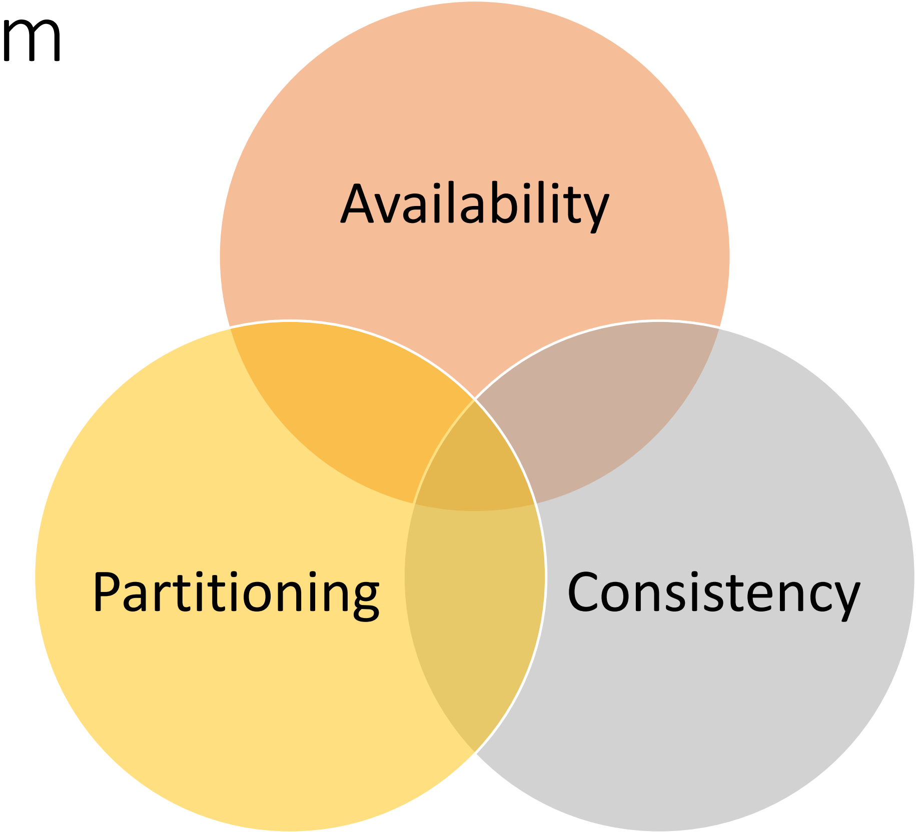
Future

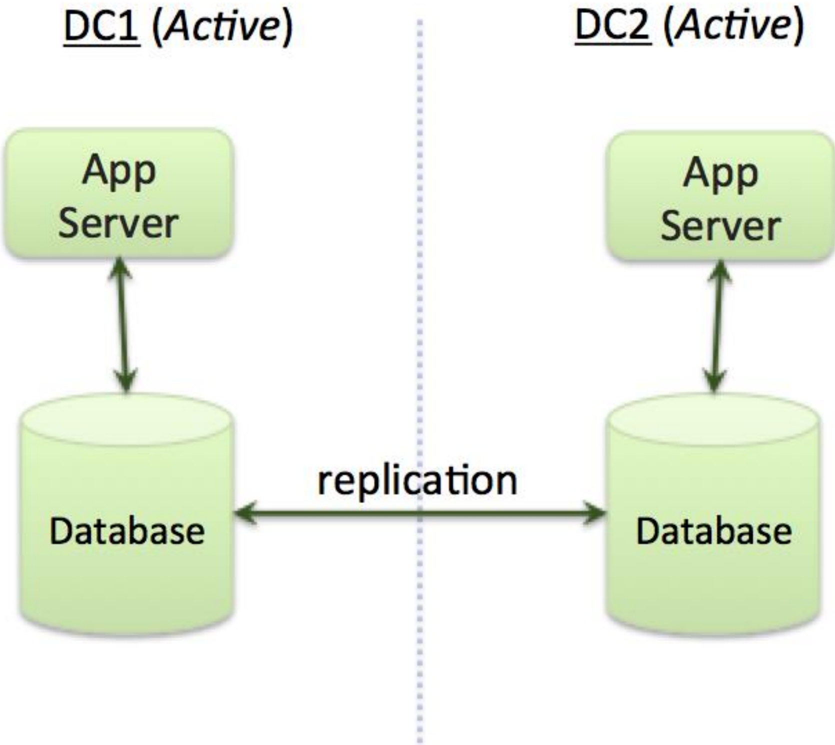
Distributed systems

Examples

CAP theorem

CAP theorem





DC1 (*Active*)

App Server

Database

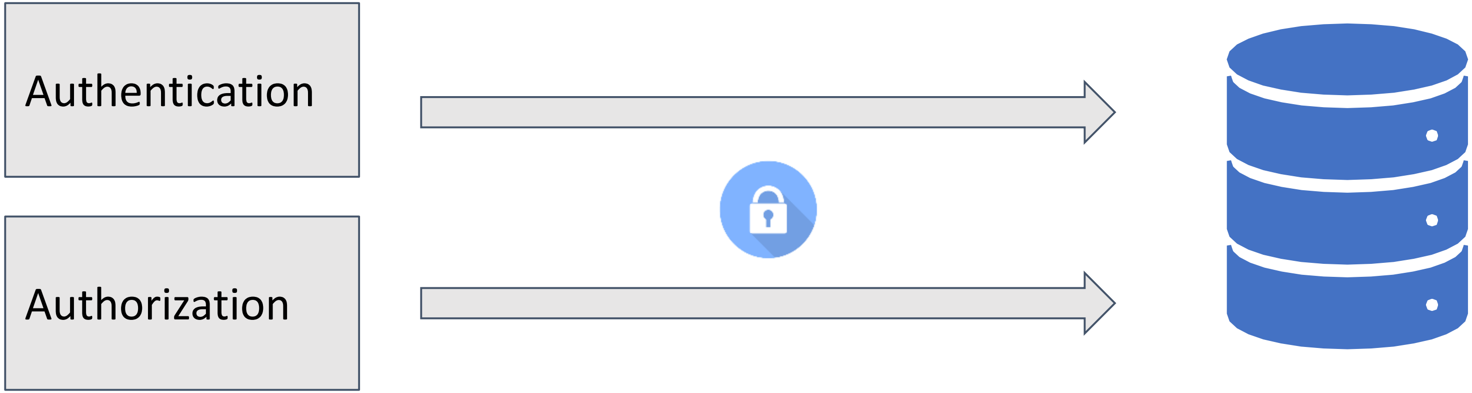
replication

DC2 (*Active*)

App Server

Database

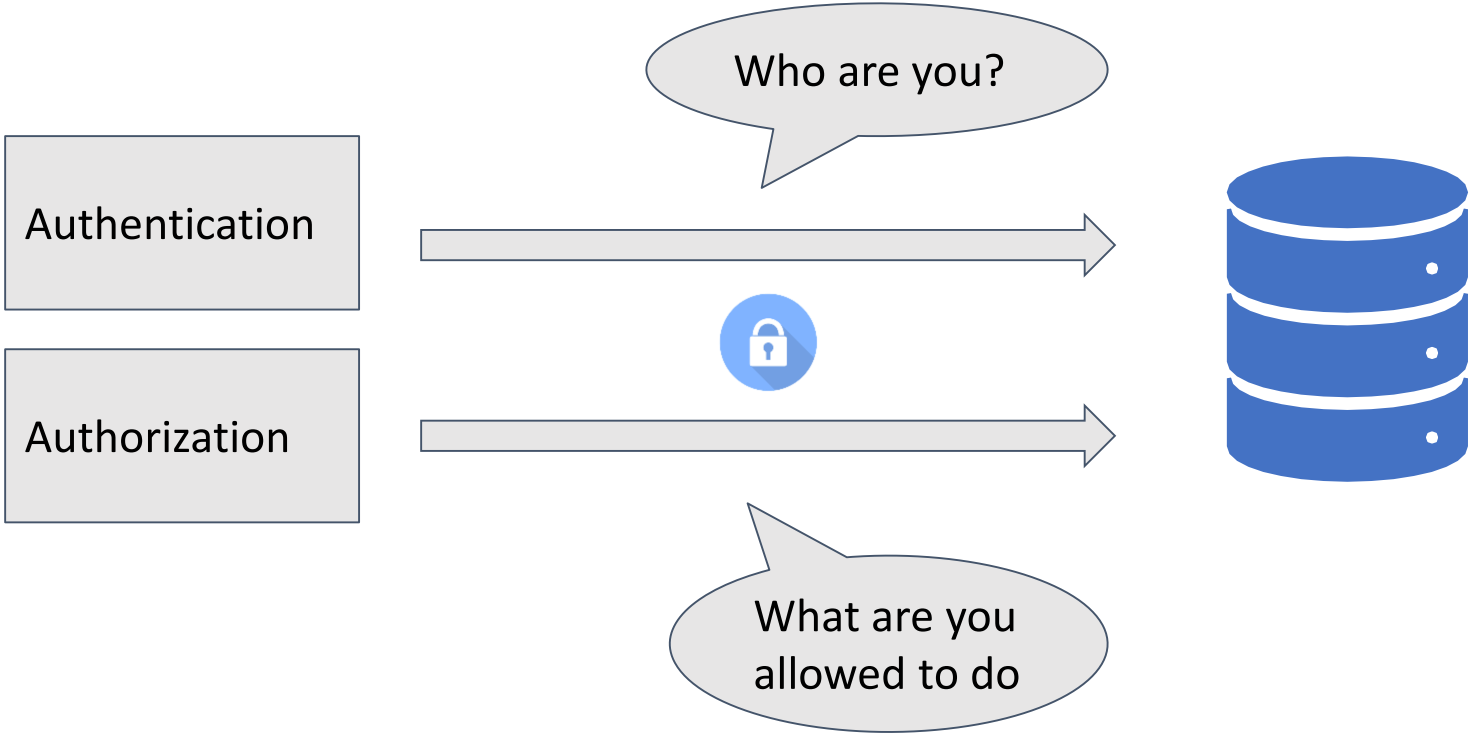
Security



Authentication

Authorization

Security



Authentication

Who are you?

Authorization

What are you allowed to do

Access Level

GROUPS & USERS, INHERIT ACCESS LEVELS

IDENTITY ACCESS MANAGEMENT

DIFFERENT ACCESS LEVELS
NONE
BROWSE
READ
RELATE
VERSION
WRITE
DELETE

VERY COMPLICATED! BUT SO IMPORTANT

Data integrity: how NOT to do it

ARNHEM
HAARLEM
AMSTERDAM
NETHERLANDS
THE HAGUE
UTRECHT
ROTTERDAM
ZEEBRUGGE
OSTEND
BRUGES
GHENT
ANTWERP
BRUSSELS
BELGIUM
LIÈGE
LUXEMBOURG CITY
LUXEMBOURG

ACCEPTING
MANY
CAPABILITY
INDEXING
TEXT
EXAMPLE
COMPLEX
PUBLISHED
FILES
MAY
FORM
PRINTED
USING
POINT
VERSIONS
TYPES
TOOLS
IMAGES
PAPER
ELECTRONIC
RETURN
RETRIEVE
COLLABORATION
ATTRIBUTES
SEARCHING
LATER
IMPORTED
PERFORM
REGULATED
ALTERED
DOCUMENT
LAW
QUERY
SAVE
CAPTURE
FILE
METADATA
STORAGE
BEGINNING
RULES
CREATED
RESULTING
RECOGNITION
WITHOUT
CONTENT
ADMINISTRATOR
VERSIONING
STORE
INTEGRATION
REPOSITORY
APPLICATIONS
STORED
TRACK
DISTRIBUTION
SUPPORTED
CONTENTS
SIMPLE
SEE
MULTIPLE
EDMS
CAPABILITIES
PROCESS
IMAGE
TERMS
MANAGEMENT
DMS
INVOLVES
USUALLY
PROVIDE
SPECIFY
ORGANIZATION
AMOUNT
USER
WORKFLOW
PEOPLE
SEARCH
EXPECTED
THINKING
MODULE
FORMS
EXTRACTED
ADVANCED
INDUSTRIES
INSTANCE
PREVIOUS
HISTORY
IMAGING
CHARACTER
SYSTEM
RETRIEVAL
SOFTWARE
SCANNED
OFTEN
ALSO
PROVIDING
TIME
PLANS
HTML
ALLOW
COMMONLY
USED
REQUIREMENTS
SECURITY
EDM
NUMBER
STORING
REQUIRE
ENVIRONMENT
DIFFERENT
USERS
USE
CERTAIN
FORMAT
TYPE
APPLICATION
COMPONENTS
ALLOWS
OPTICAL
MANAGE
SESSION

LAW & CRIME
PSYCHOLOGY
MEDICAL
LANGUAGES
THOMPSON

summit
IBM

Document management

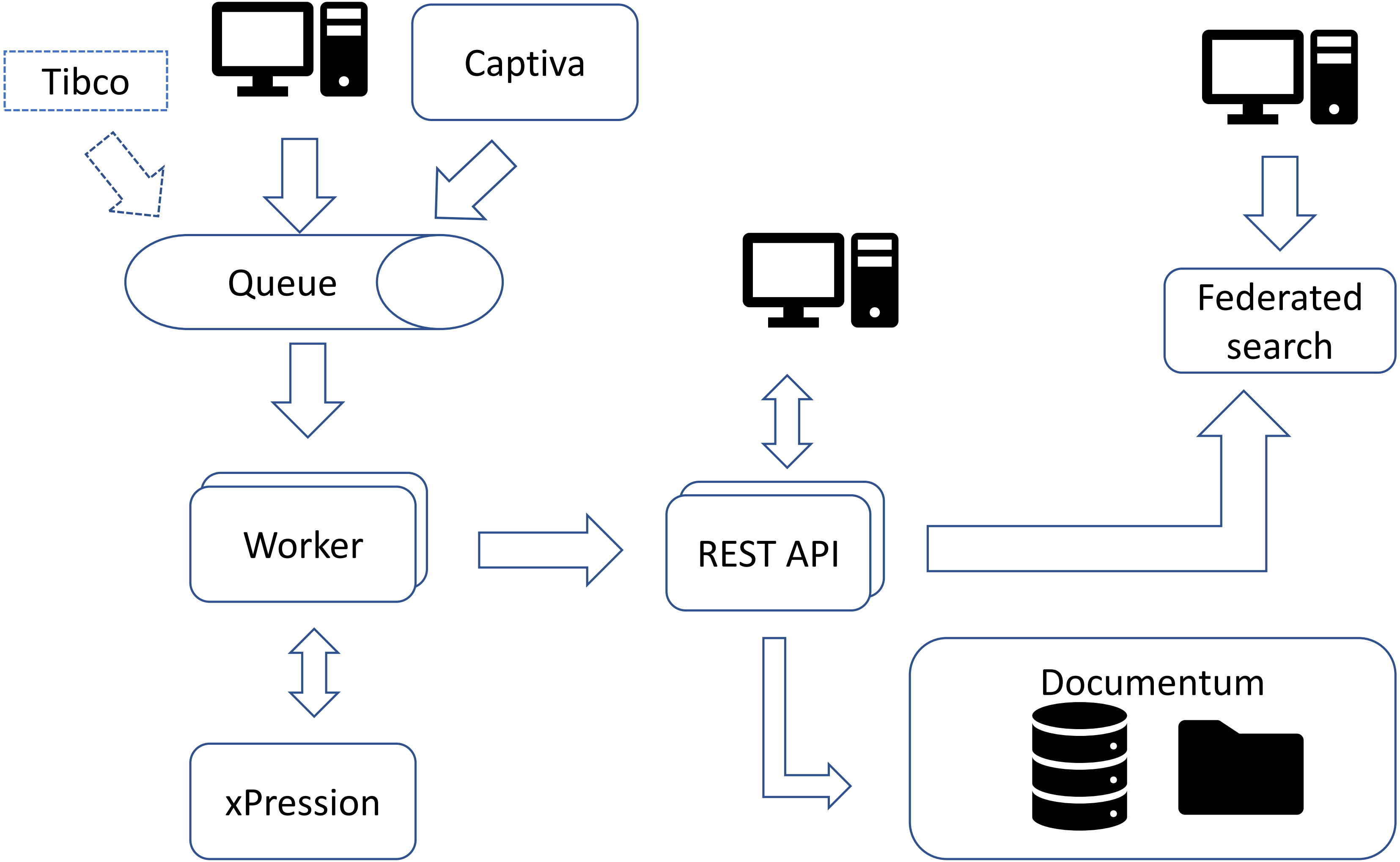
- Archiving
- Document generation
- Capturing
- CRUD operations
- Workflows
- Federated search
- Security: Authentication & Authorization

Document management @ ING

- Storing documents
- Authorization & Authentication
- Oracle database + Filesystem
- Complete system
  - UI management
  - Lot’s of extra features(that you have to buy)
- Integrated with complete JAVA REST API
- xPression
- Captiva

documentum is now opentext™

Architecture


Tibco
Captiva
Queue
Federated search
Worker
REST API
xPression
Documentum

REST API

- Access point for CRUD operations on documents
- Interface for working with documents
- Performant
- Reliable
- Stateless
- DFC – Documentum Java API

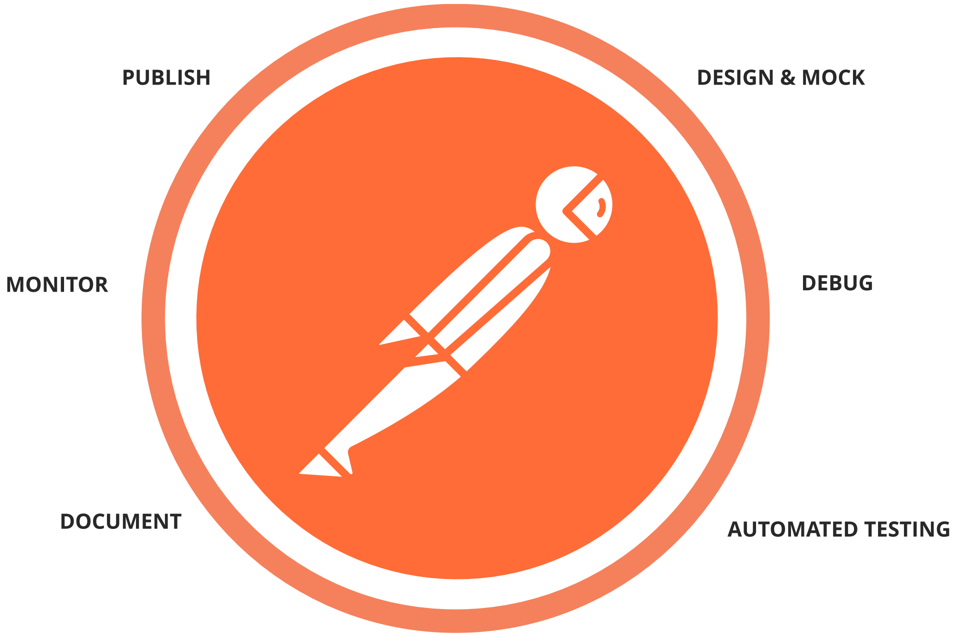
spring

Jersey

RESTful Web Services in Java.

```
@POST
@Path("/{repository}/{objecttype}")
@Produces("application/json")
@Consumes("multipart/*")
@ApiOperation(value = "[SID 19] Create one or more documents of a certain type")
@AuditedOperation(operation = "CREATE", arguments = { "body", "repository",
"objecttype" })
public Object createObjectWithContent(MultipartBody body, @PathParam("repository")
String repository,
@PathParam("objecttype") String objecttype) throws DfException {
```

- Design and model API’s
- Interactive UI
- API documentation
- Standarized api design



PUBLISH

DESIGN & MOCK

MONITOR

DEBUG

DOCUMENT

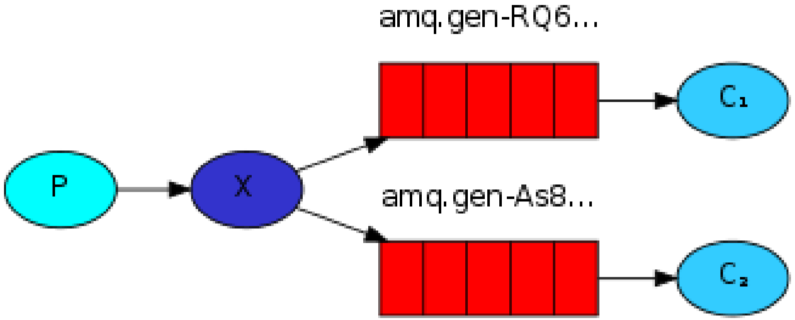
AUTOMATED TESTING

Postman

RabbitMQ by Pivotal

- Management for jobs
- Asynchronous operations
- Connection with lot of different systems

Worker applications



- Batch operations
- Archiving
- Document generation
- Extra processing steps

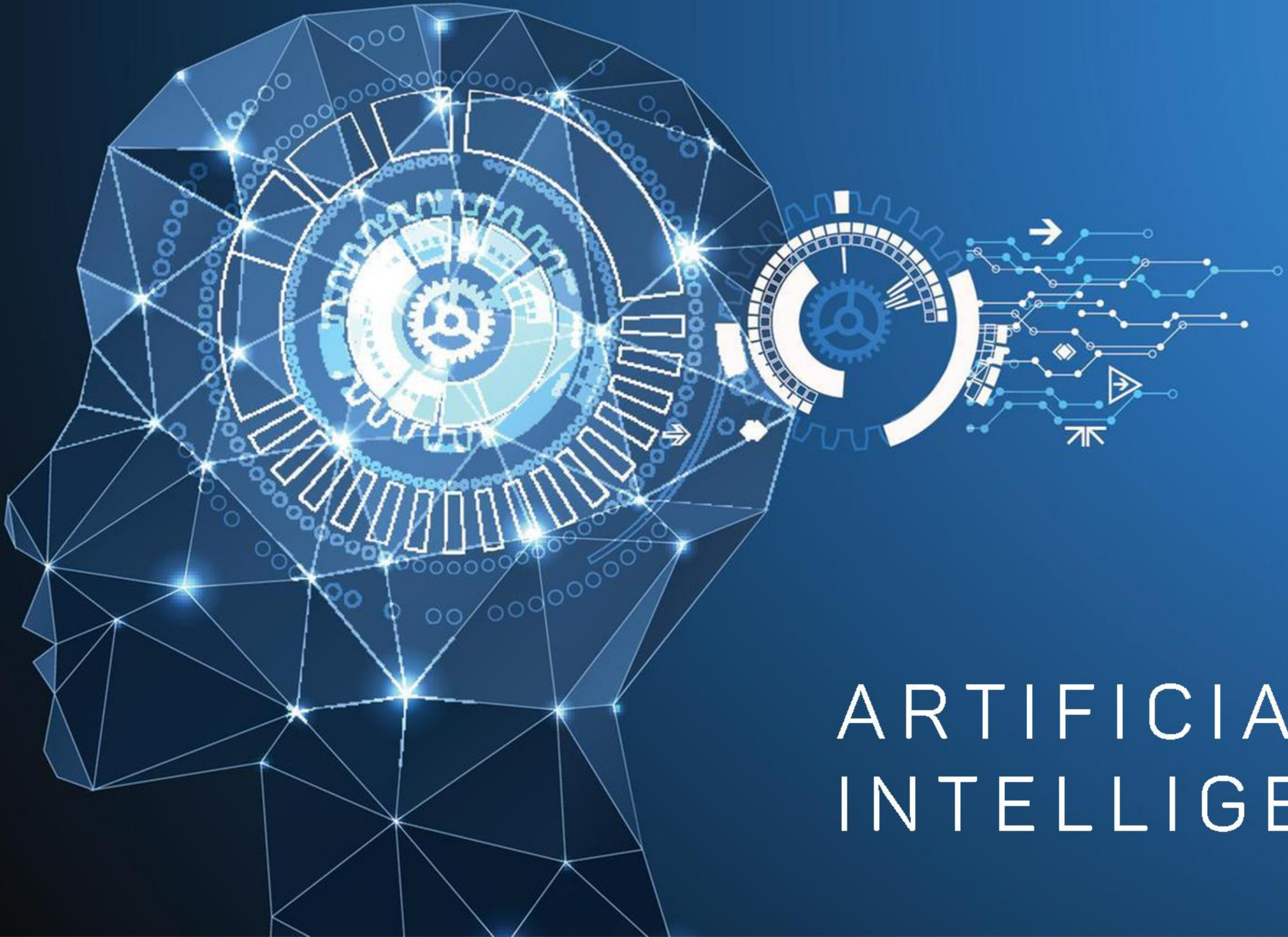
Clients

- Lot’s of different clients
- CRUD operations
- Process flows

Federated search

elasticsearch

- Distributed search engine
- Centralized coordination
- Indexed data
- Usable for different use cases
- Scalable



ARTIFICIAL INTELLIGENCE

Artificial Intelligence

- OCR
- Full-text analysis
- Classification of documents
- Contract checking
- …

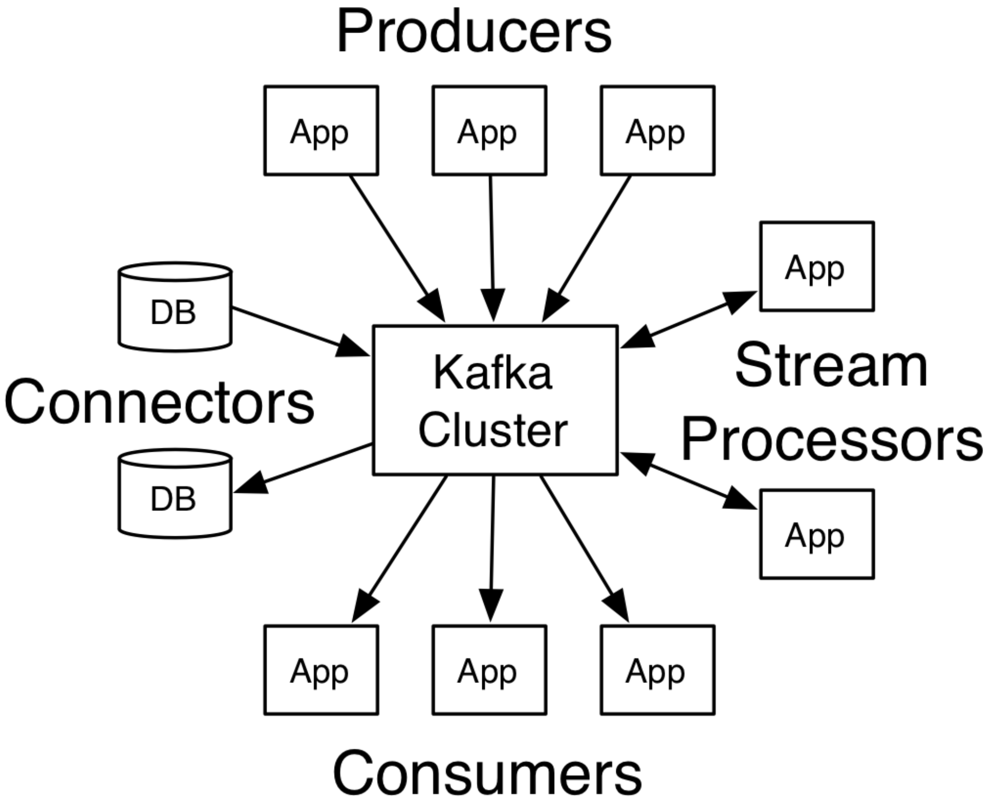
Messaging

- Real-time data pipeline & processing
- Streams
- distributed storage system
- Publish – Subscribe system
- Queue system

APACHE

kafka®

A distributed streaming platform



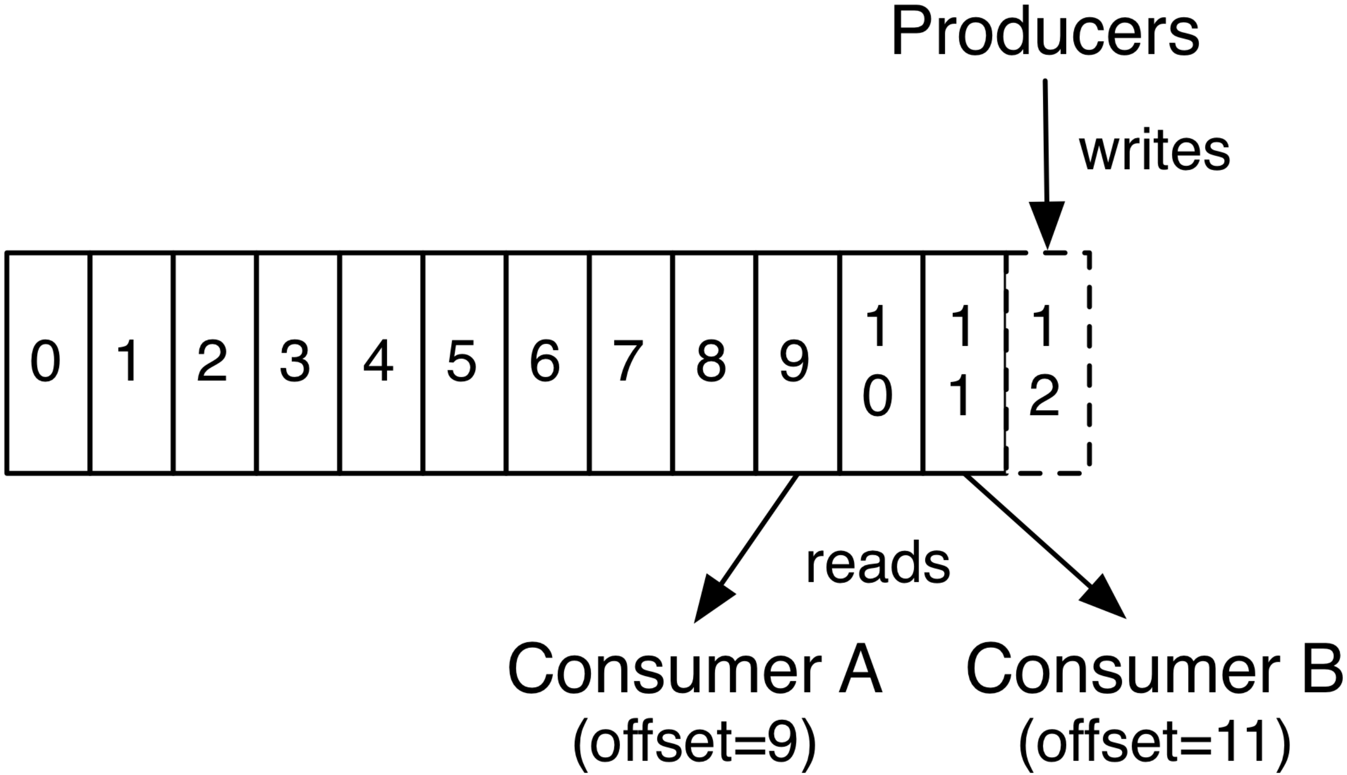
Producers

Connectors

Kafka Cluster

Stream Processors

Consumers



Producers

writes

0	1	2	3	4	5	6	7	8	9	10	11	12

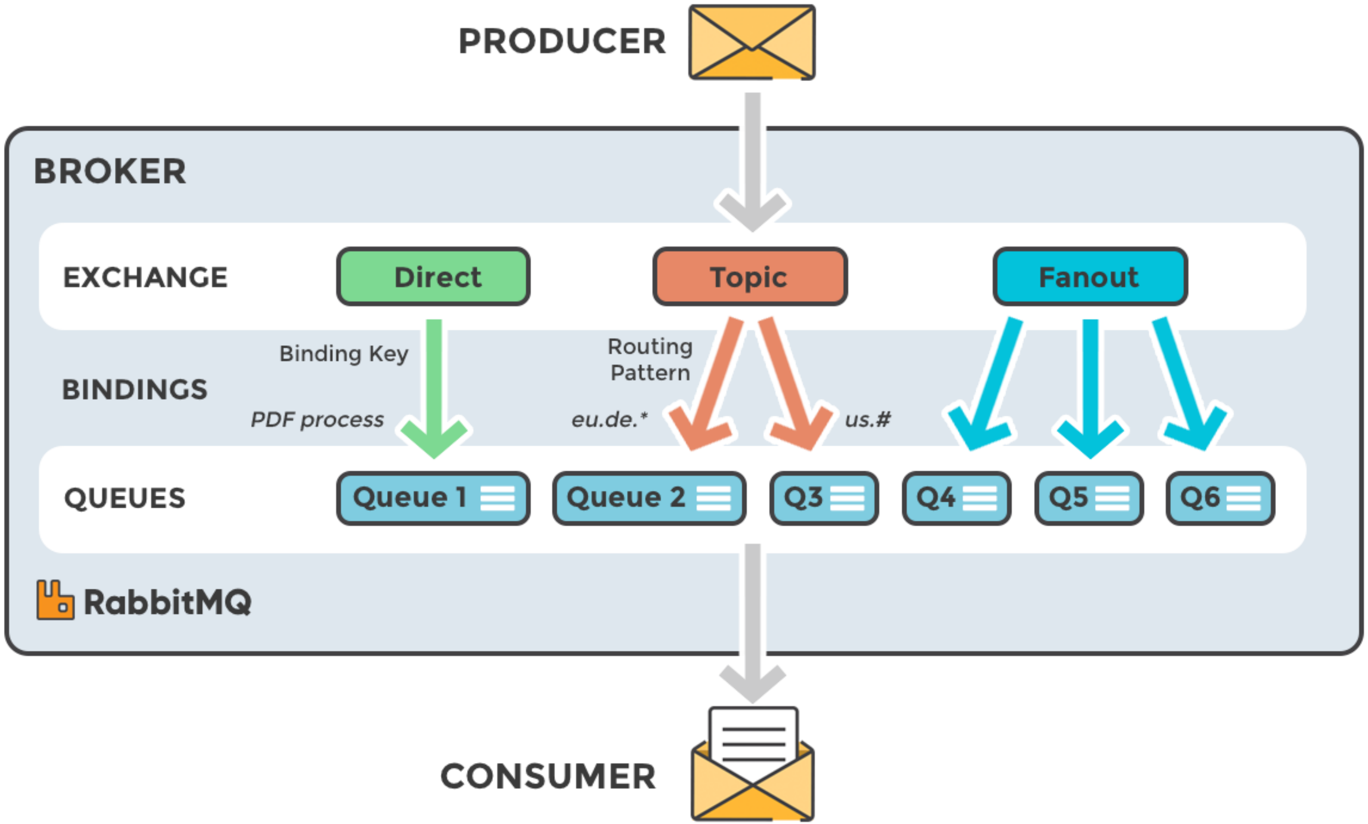
reads

Consumer A
(offset=9)

Consumer B
(offset=11)

RabbitMQ

- Asynchronous messaging
- Lightweight
- Easy to use
- Distributed environments
- High availability

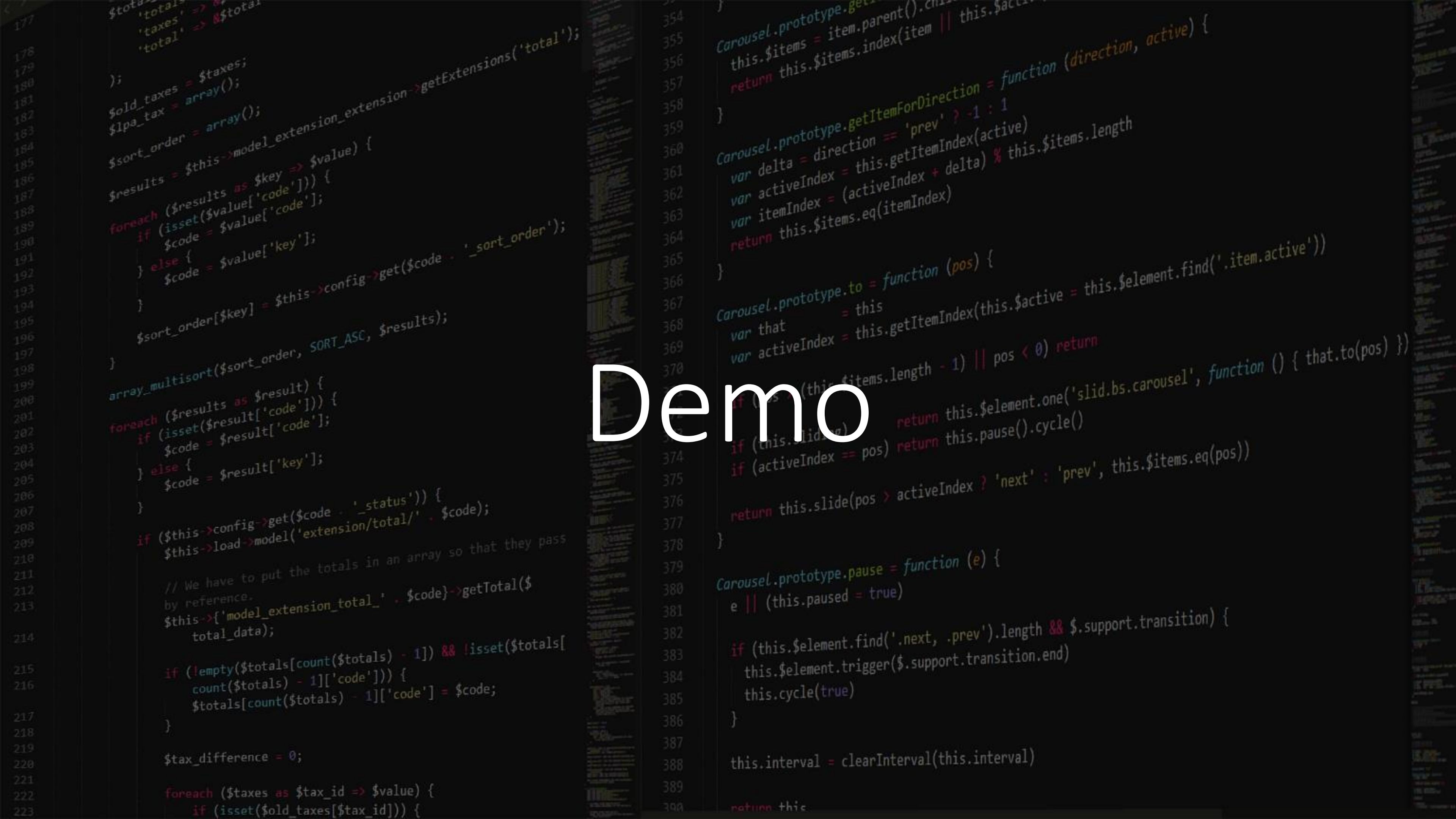

PRODUCER
BROKER
EXCHANGE
Direct
Topic
Fanout
Binding Key
Routing Pattern
BINDINGS
PDF process
eu.de.*
us.#
QUEUES
Queue 1
Queue 2
Q3
Q4
Q5
Q6
RabbitMQ
CONSUMER

Basic features

- API in all major languages
- CLI & UI management
- External plugins
  - Message delays
  - Management API over http
- Acknowledgements
- Durability and persistency

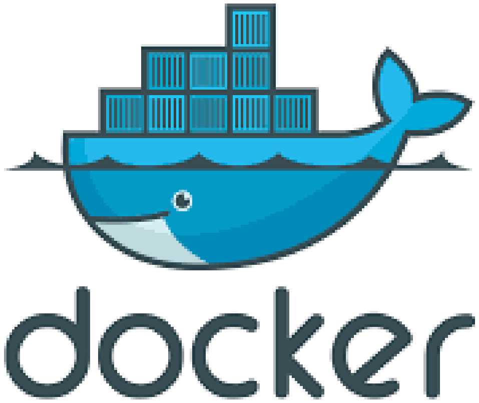
Distributed message broker: Cluster

- Different methods: cluster, federation, shovel
- Cluster: single logical broker
  - Collection of different nodes
  - exchanges and bindings are mirrored
  - Queues are optionally mirrored(master – slave)
- Everything accessable from any node



Demo

Go to the next level…

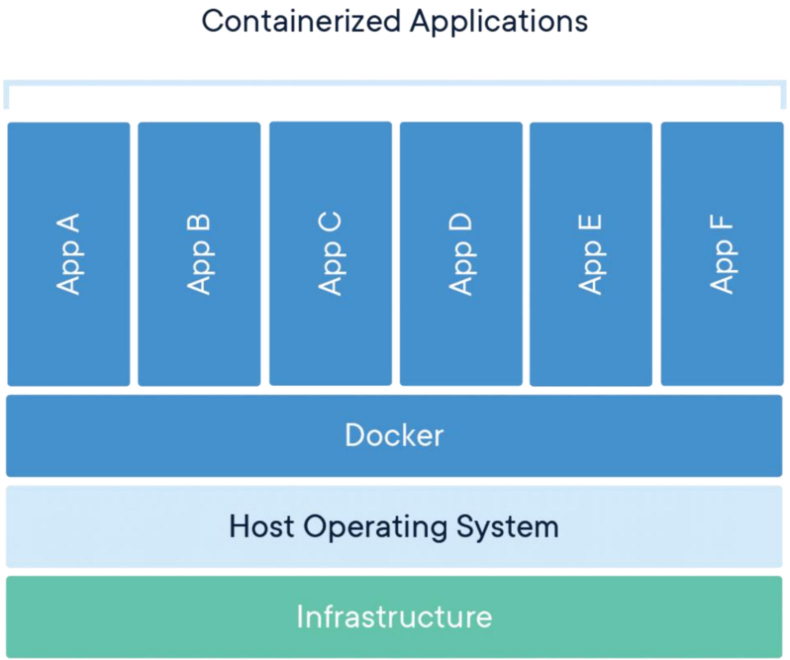


Docker

- Lightweight deployments
- Containers
- Independent isolated applications
- Supports CI/CD

Docker: Containers

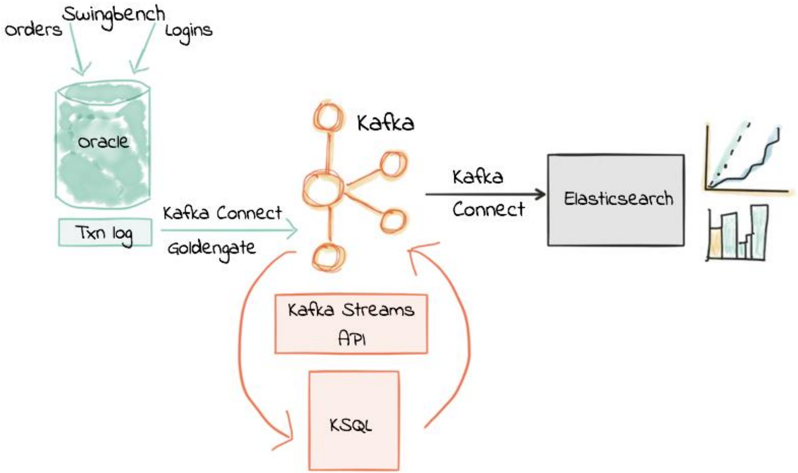
- Lightweight, standalone
- All dependencies for application and nothing more
- Separate the software from the environment
- Isolated applications
- Reduce time to bring app to production
- Less dependent on underlying technology



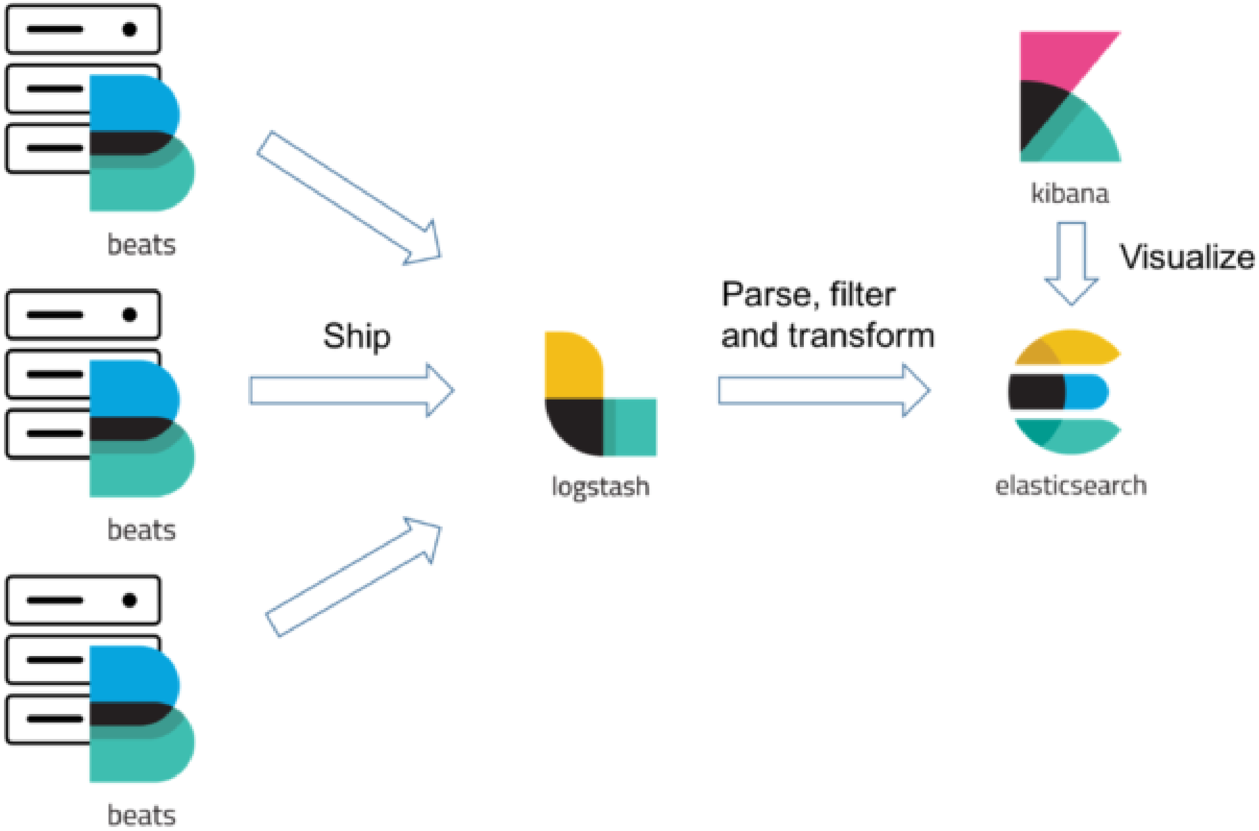
Kubernetes

- Used and created by Google
- Automatic deployments
- Desired state management
- Orchestrated update: rollout & rollback
- Horizontal scaling to Planet scale

Kafka: KSQL



- Confluent
- real-time, fault-tolerant stream processing
- Realtime stream: SQL queries
- Easy to use
- Transform data in topics
- Usage of streams and tables


beats
beats
beats
Ship
logstash
Parse, filter
and transform
elasticsearch
kibana
Visualize

Recap

- Document management is distributed on different levels
- Goal of ING is to be an it/tech company with a banking license
- AI is getting more important
- Future holds interesting challenges

Questions