Data Science and Engineering

Stijn Meijers
Data Engineer

ORTEC
OPTIMIZE YOUR WORLD

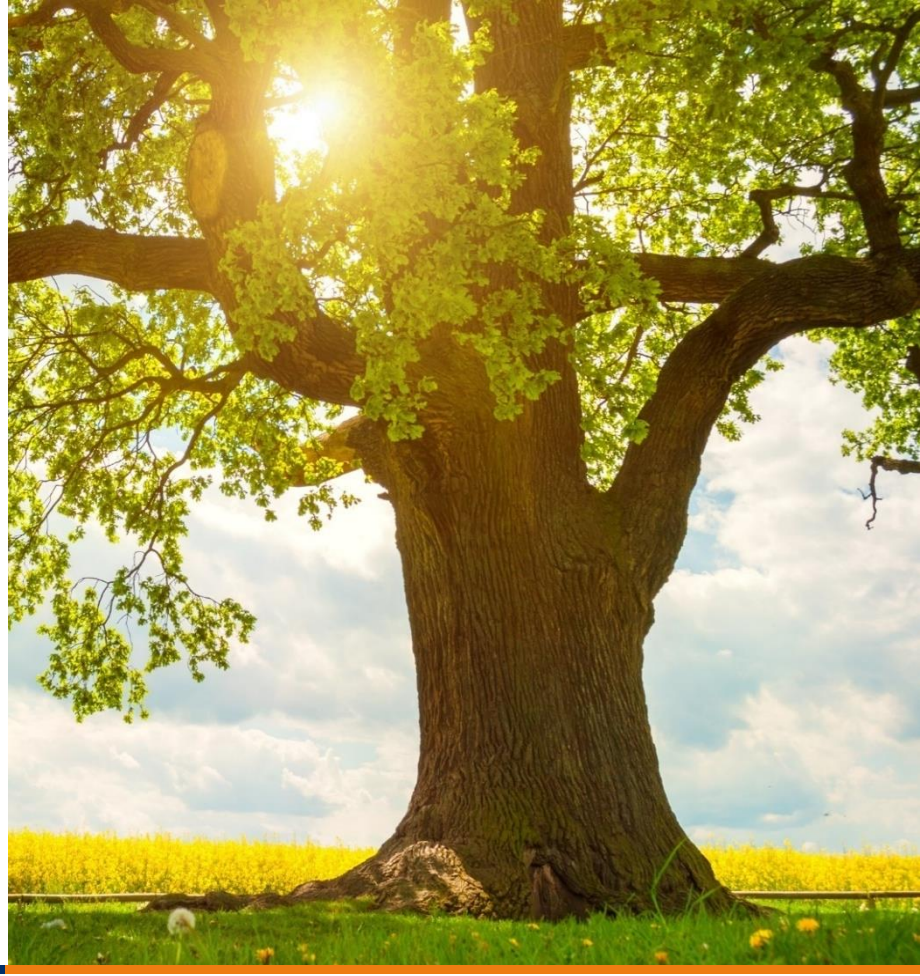
ORTEC Introduction

Data science & data engineering

VUmc collaboration: RNA Splicing

CRISP-DCW & Location based ad-serving

TOPICS

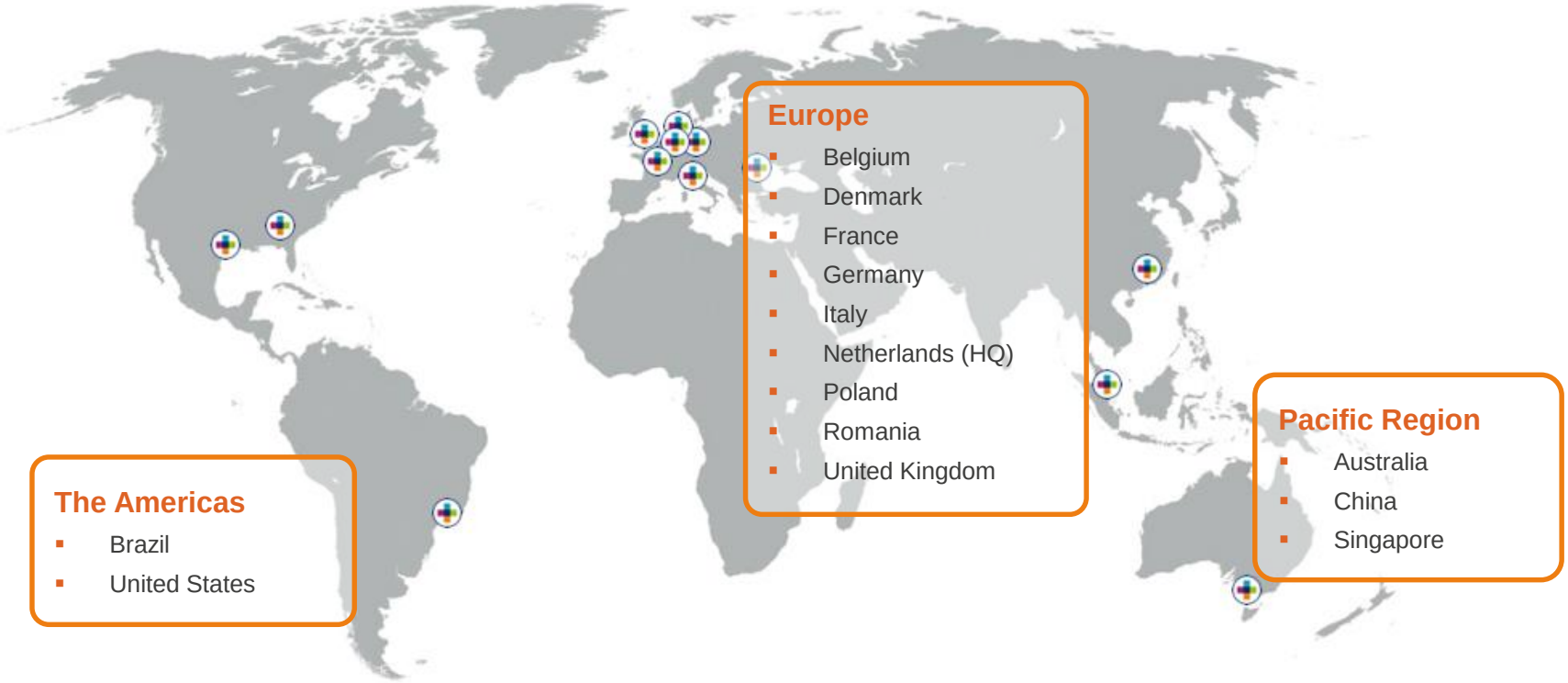


ORTEC

Who are we?

ORTEC

A Global Analytics & Optimization Company



Europe
- Belgium
- Denmark
- France
- Germany
- Italy
- Netherlands (HQ)
- Poland
- Romania
- United Kingdom

The Americas
- Brazil
- United States

Pacific Region
- Australia
- China
- Singapore

ORTEC

Our Core Competences

Data driven program

ORTEC

Empowering our Industry Leading Customers

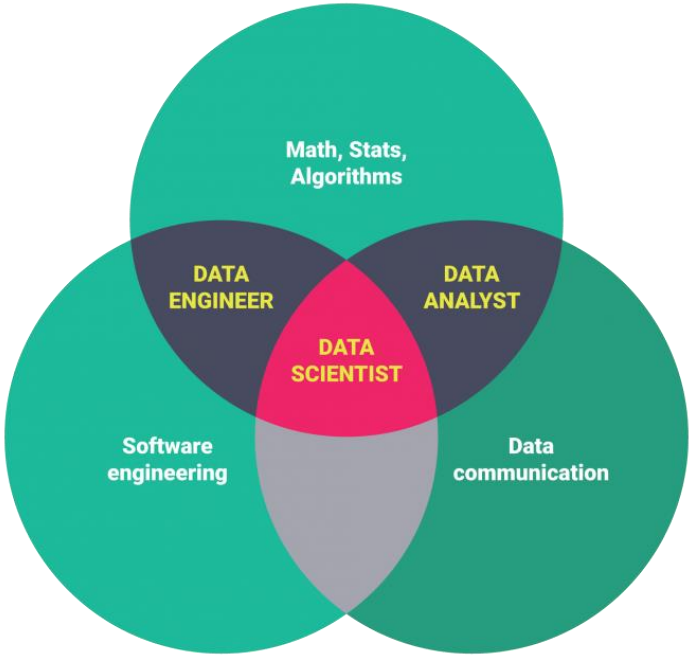
ORTEC

Data Science and Data Engineering

Differences in data science and data engineering

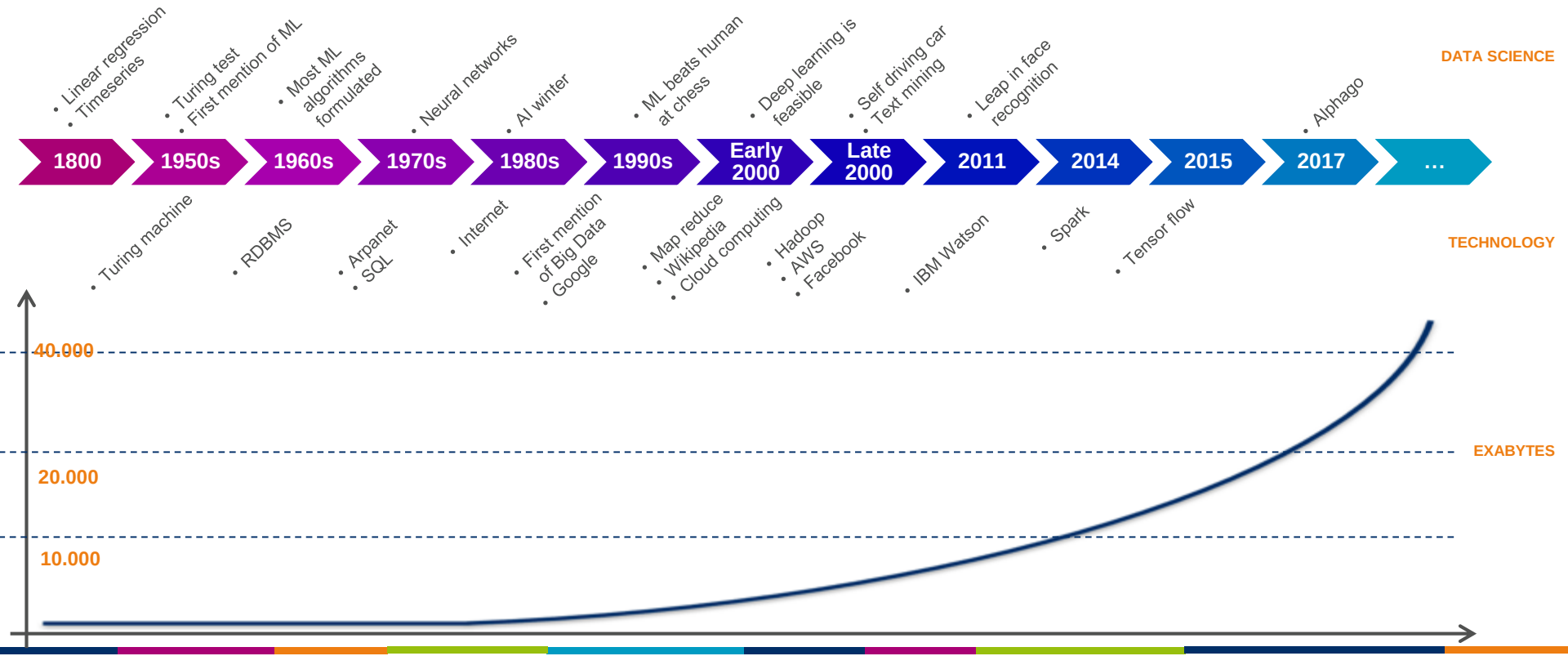
- Data scientists: create models and extracts business information using data

- Data engineers: Extract, transform and load (or; Extract, Load Transform) the data for analysis. Implements models for production.



ORTEC

History of Big Data

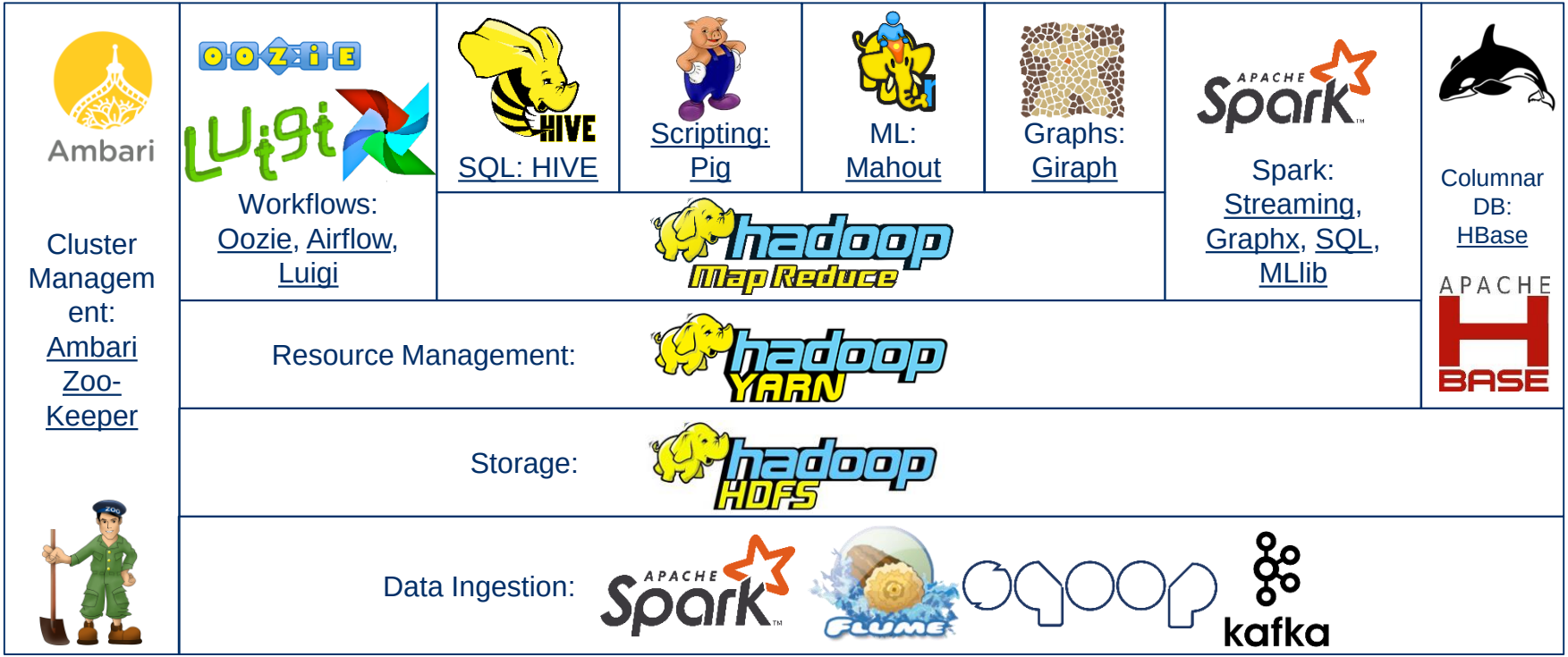


ORTEC

APACHE DRILL
Hortonworks
APACHE H BASE
hadoop YARN
aws
hadoop HDFS
Google Cloud Platform
APACHE GIRAPH
HIVE
Azure
APACHE Spark
mahout
FLUME
jupyter
kafka
hadoop Map Reduce
OOZIE
Ambari

ORTEC

Hadoop Ecosystem



ORTEC

Azure

aws

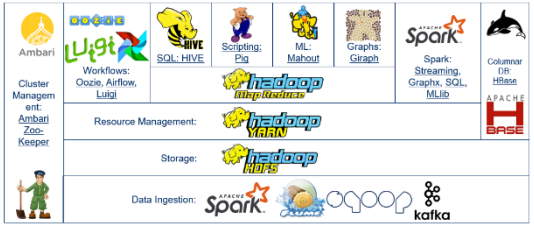
Cloud

Google Cloud Platform

cloudera

Hortonworks

Tech stacks



jupyter

UI / Notebook

ORTEC

VU medical center: RNA splicing research project

RNA Splicing project

- Performance comparison of RNA splicing tool when using distributed computing
- Microsoft Azure
- Part of a larger research project where identifying certain RNA sequences will be identified that are more prevalent to rheumatism
- Final testing will take place on a very large data set provided by Vumc
- Currently building the architecture and distribution of the tools from the ground up

Distributed computing workflow creation

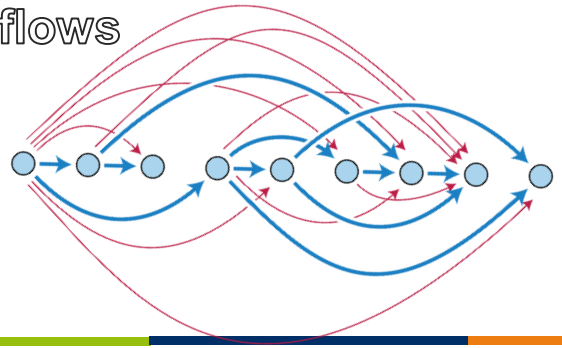
Big Data

Processing Platforms

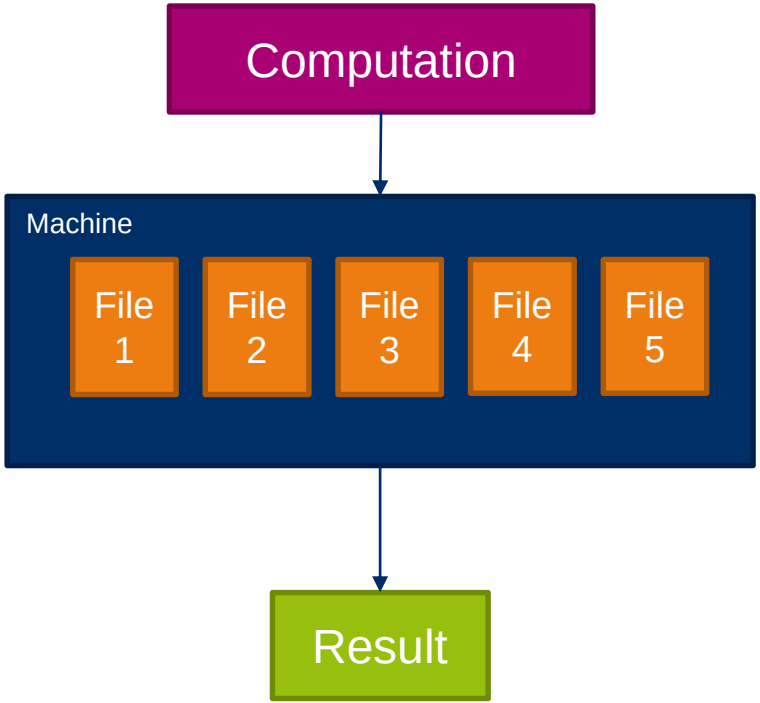
Introduction

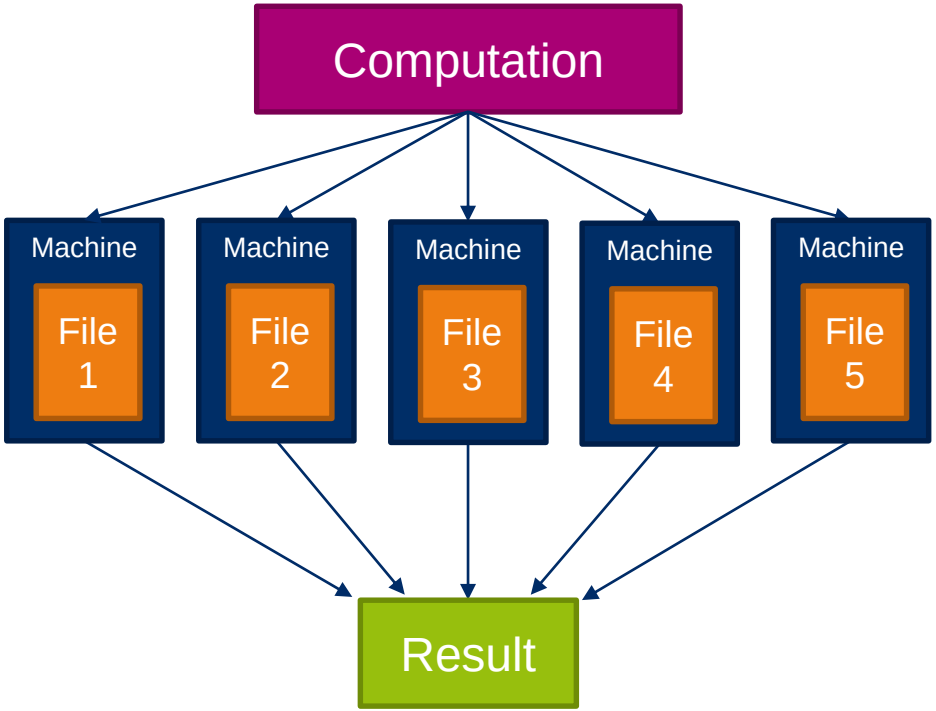
Cloud Computing

Scientific Workflows



Distributed Computing





ORTEC

Artifact Design

Problem Context

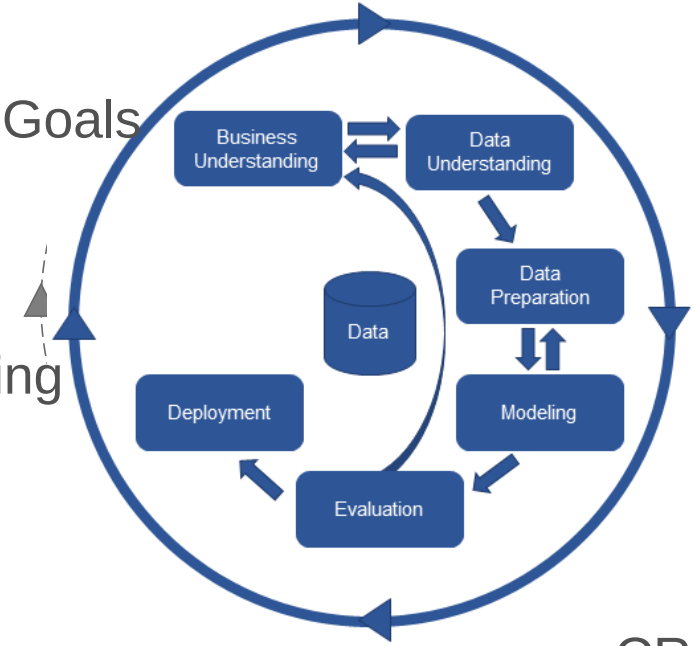
1. Define Context and Goals

Design

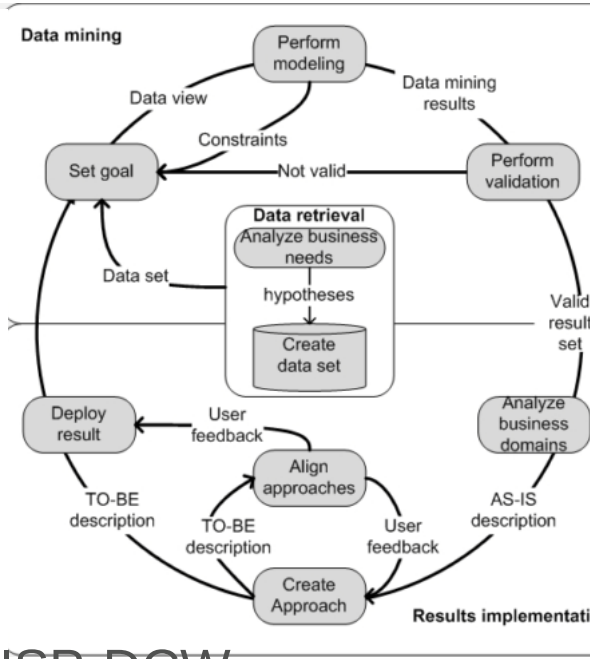
2. Determine Input
3. Estimate Output
4. Determine Processing

Implementation

5. Evaluate Results
6. Deploy Outcomes



CRISP-DM

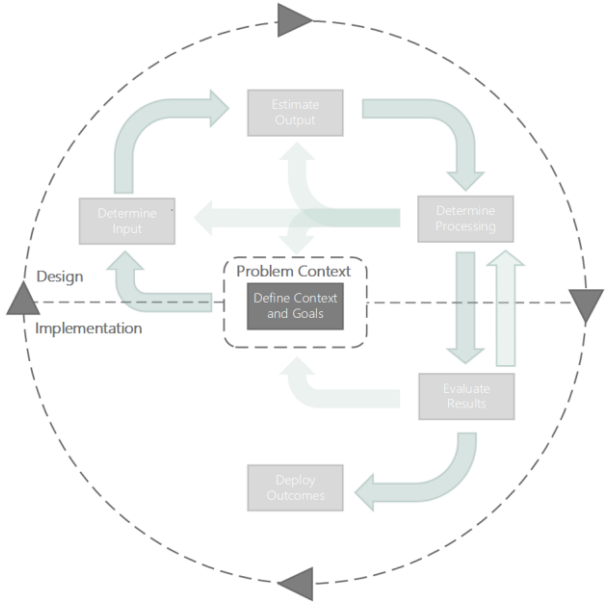


CRISP-DCW
Three Phases Method

ORTEC

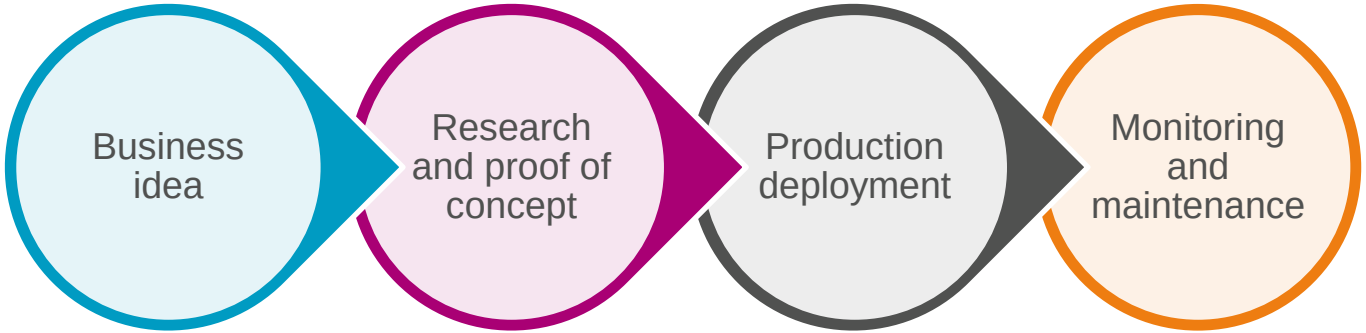
Artifact Design

- Define Context & Goals
  - Data Processing Objectives
  - Available Resources
  - Model Specifications



CRISP-DCW

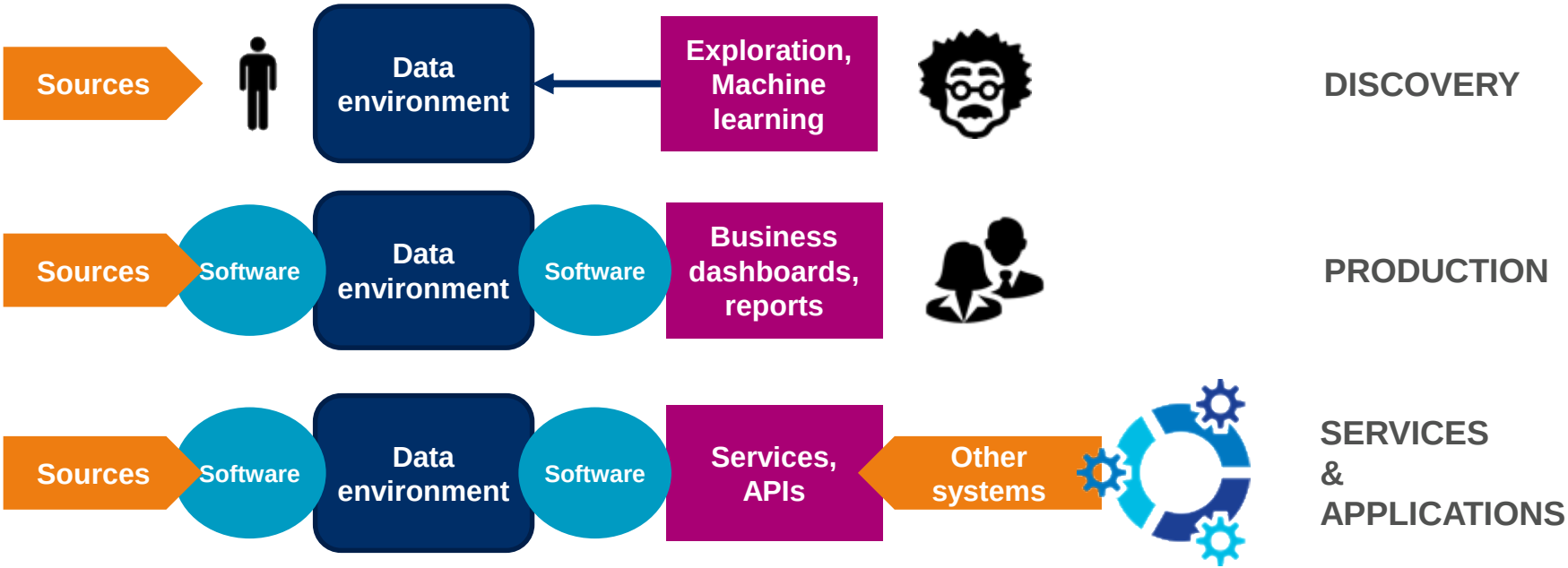
Lifecycle of data science products



ORTEC

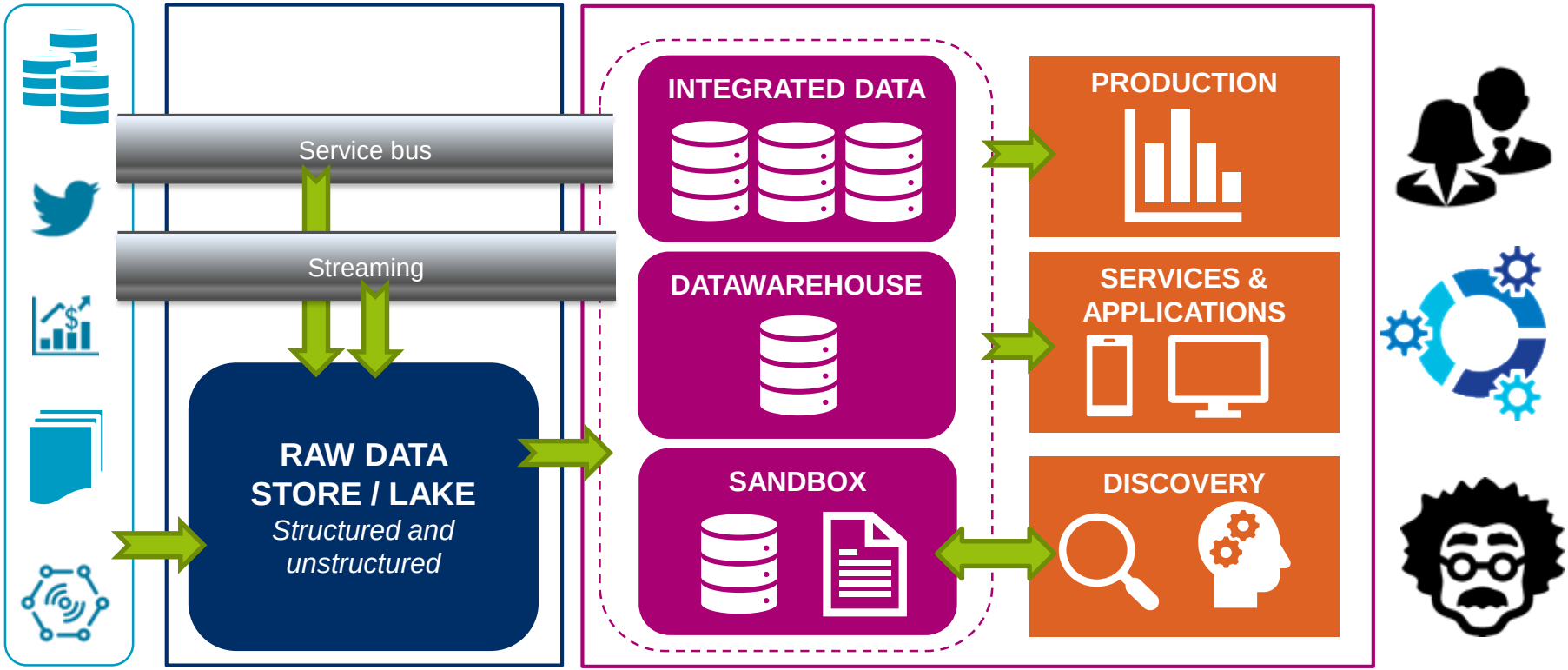
Type of application

Different flow, expectations, requirements



ORTEC

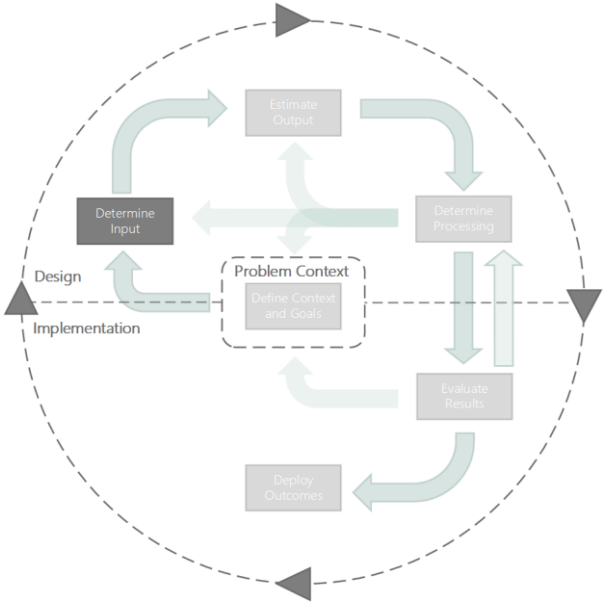
Data infrastructure blue print



ORTEC

Artifact Design

- Determine Input
  - Volume: A lot of data!
  - Velocity: Fast or slow?
  - Variety: Different forms?
  - Variability: Does it change?
  - Veracity: Messy data?
  - Value: What is it worth?

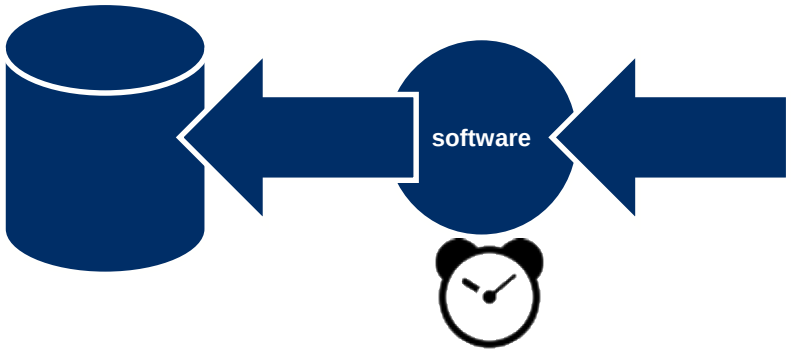


CRISP-DCW

ORTEC

Getting the data - PULL

(automated)



Sources:

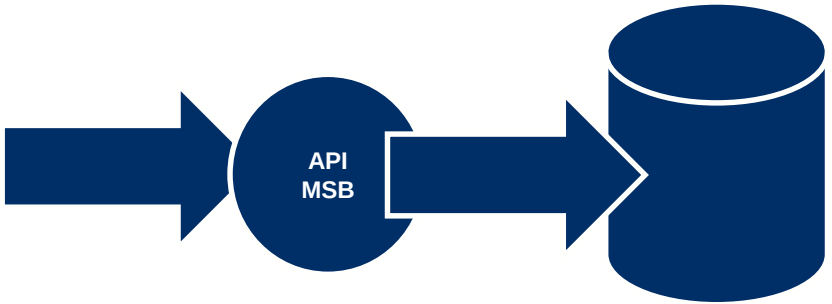
- Structured file (CSV, XLS, etc) → Download (FTP, HTTP, etc)
- Direct DB access (querying, synchronization/replication)
- API (XML/REST/SOAP/e.d.)
- Web scraping

ORTEC

Getting the data - PUSH

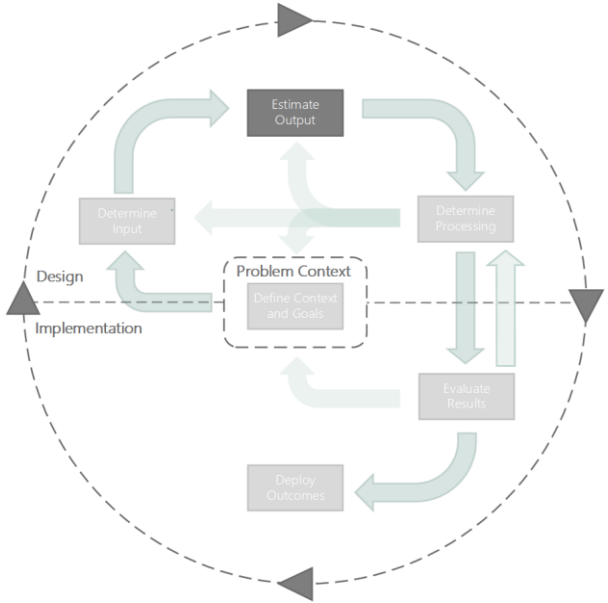
(automated)

- Streaming or message bus
- API (XML/REST/SOAP/e.d.)



Artifact Design

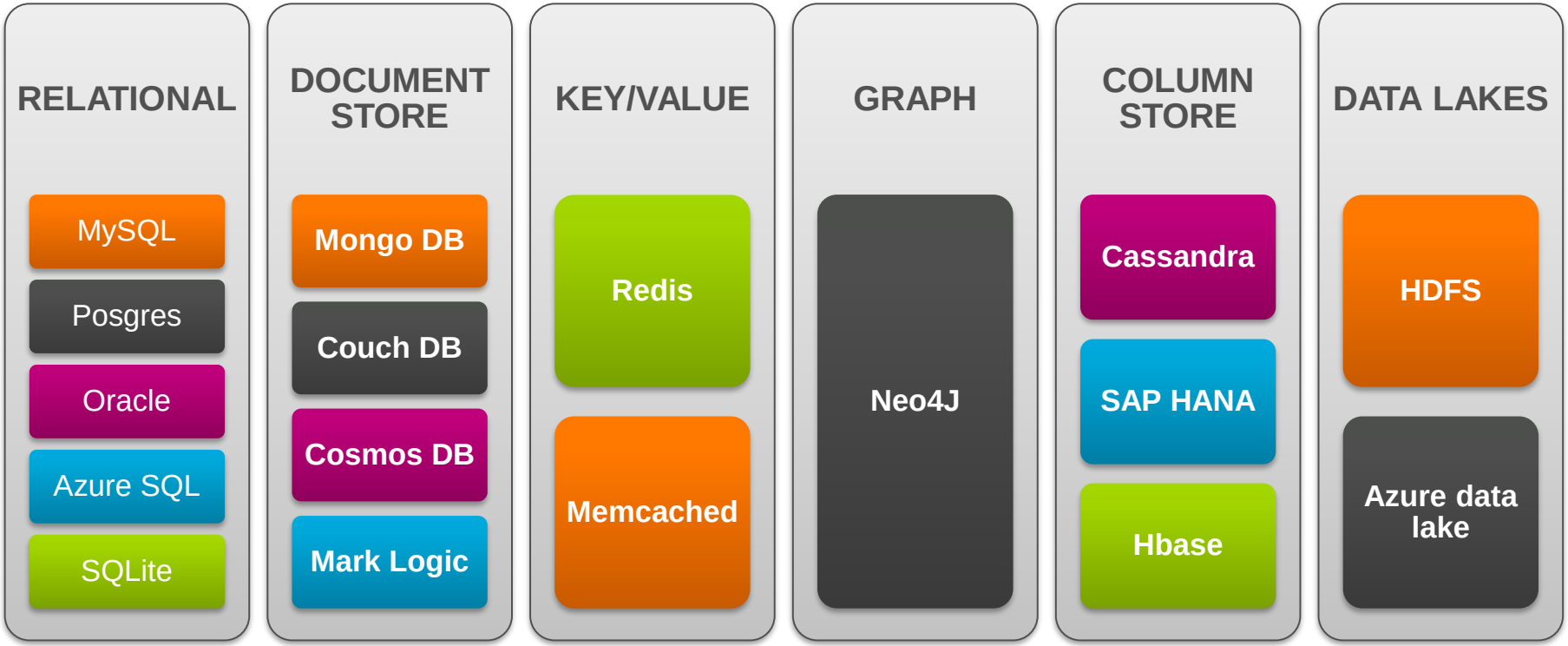
- Estimate Output
  - Volume: A lot of data!
  - Velocity: Fast or slow?
  - Variety: Different forms?
  - Variability: Does it change?
  - Veracity: Messy data?
  - Results & Visualizations



CRISP-DCW

ORTEC

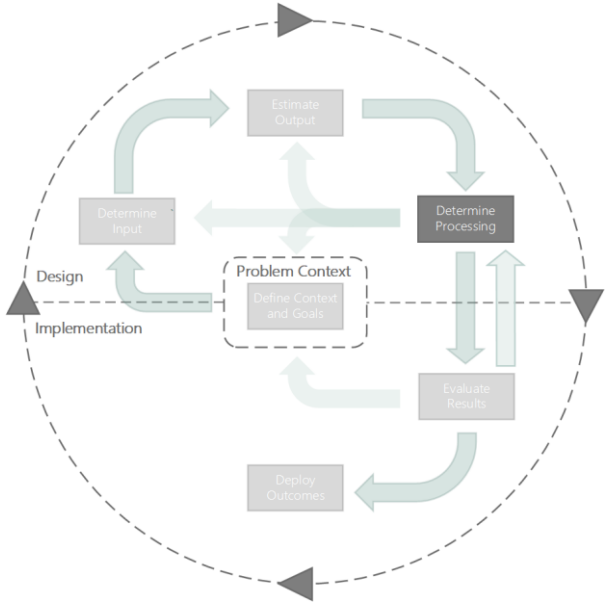
SQL and NoSQL databases categorization



ORTEC

Artifact Design

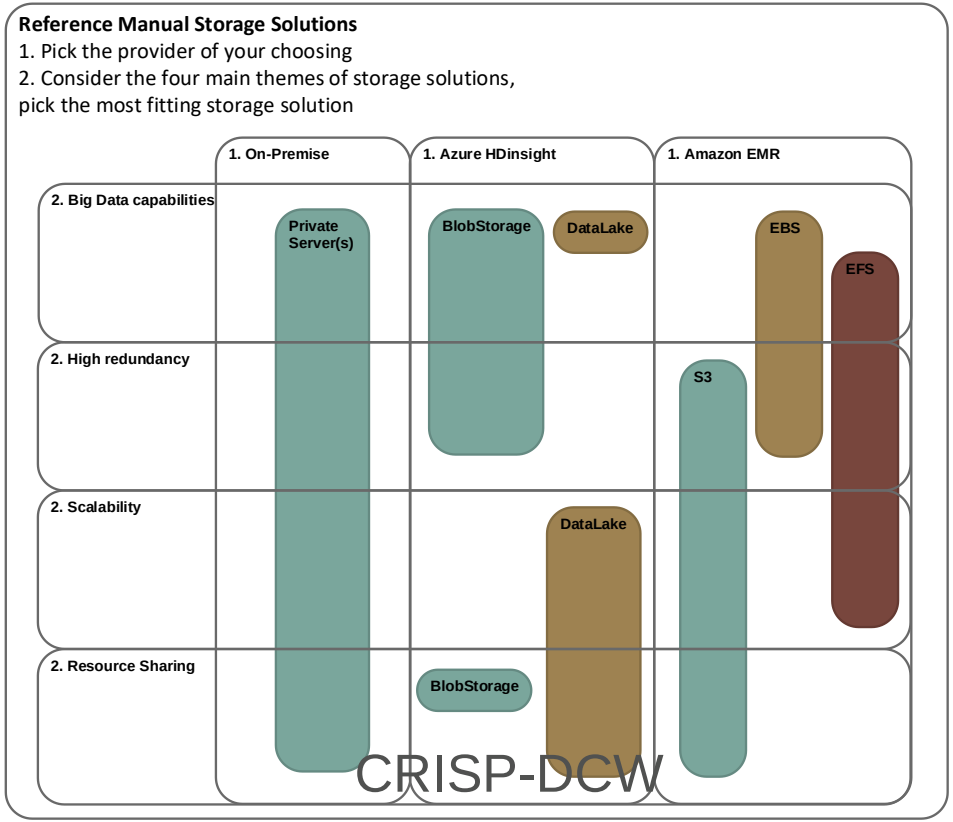
- Determine Processing
  - Input Processing
  - Data Transformations
  - Output Processing
  - Set up Workflow



CRISP-DCW

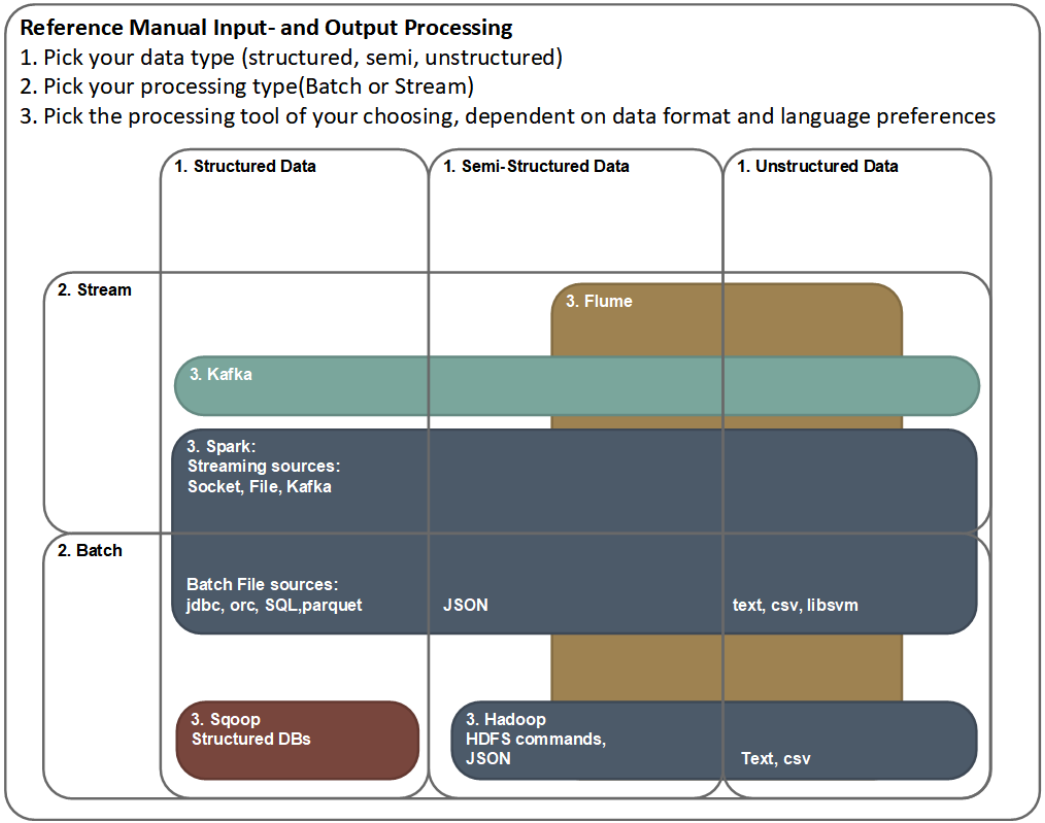
Artifact Design

- Determine Processing
  - **Input Processing**
  - Data Transformations
  - Output Processing
  - Set up Workflow



Artifact Design

- Determine Processing
  - **Input Processing**
  - Data Transformations
  - Output Processing
  - Set up Workflow



ORTEC

Data Ingestion

- Sqoop
- Flume
- Kafka
- Spark

- Move data from different data sources to your cluster, for example relational DBs, or streaming sources

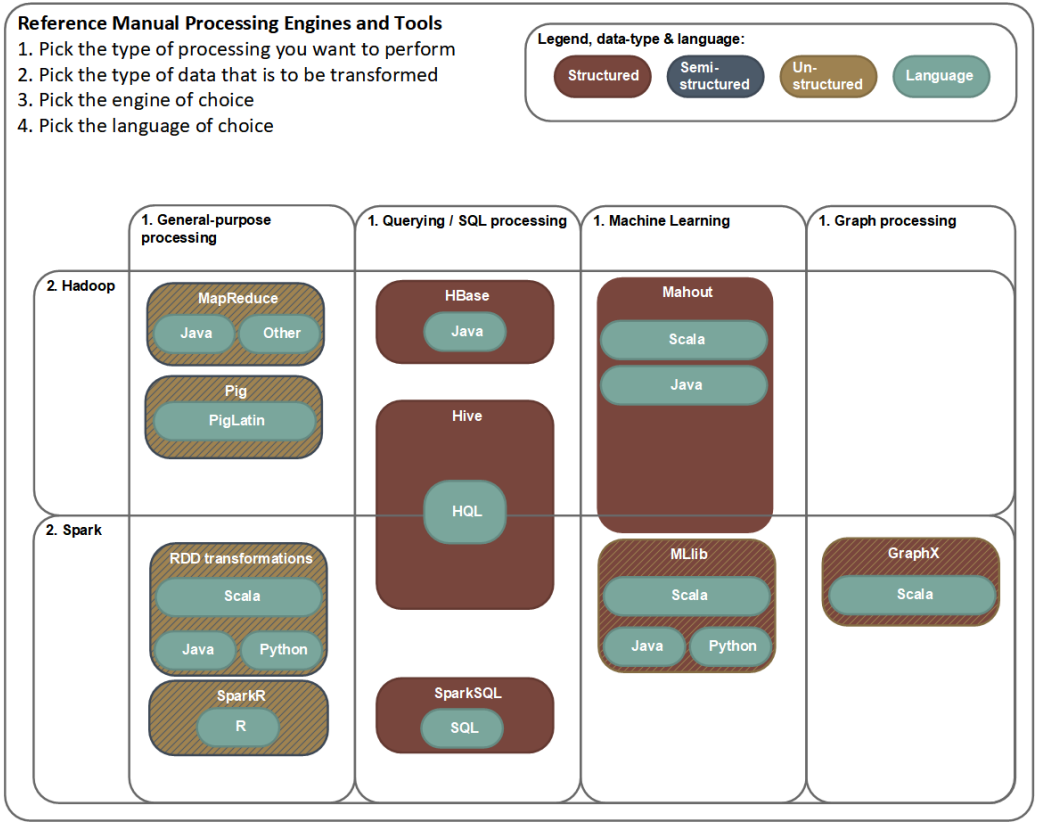
- **When to use:** Moving data from storage to HDFS, or the cluster.

kafka

ORTEC

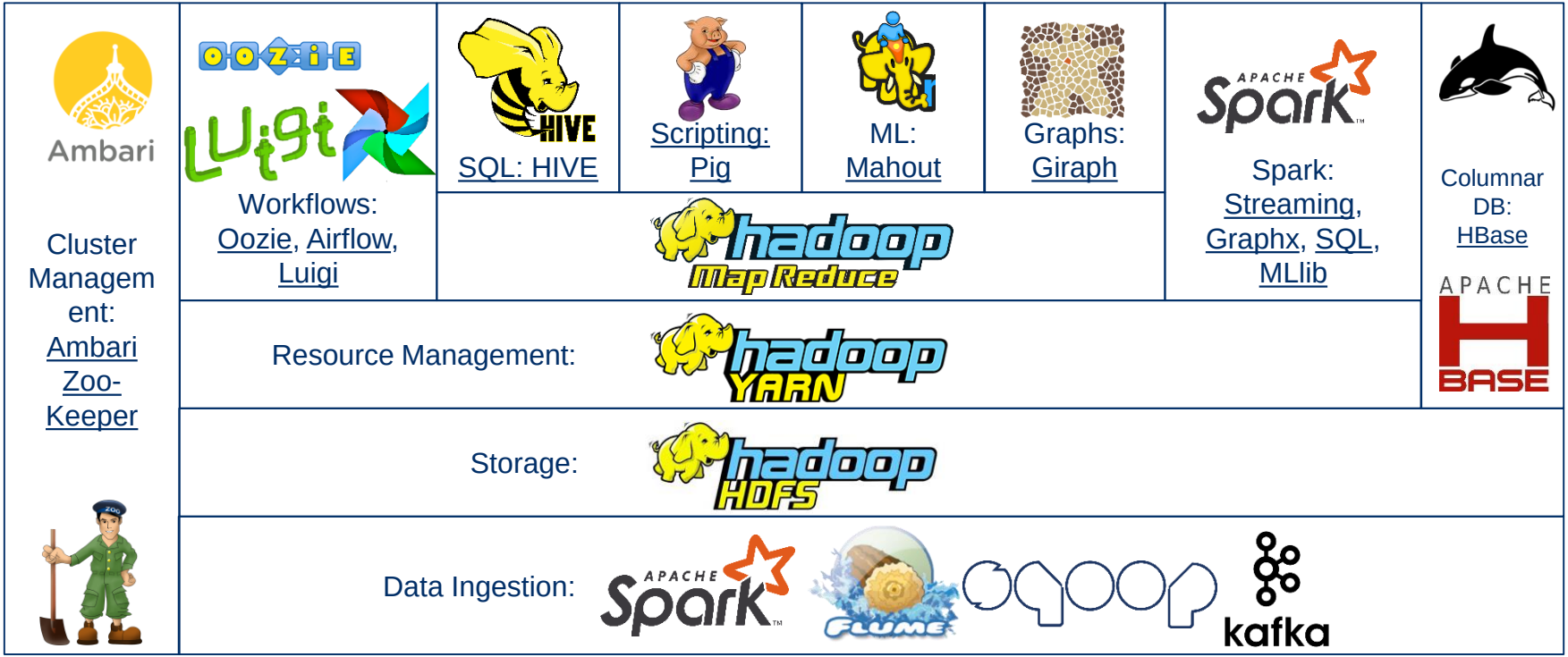
Artifact Design

- Determine Processing
  - Input Processing
  - **Data Transformations**
  - Output Processing
  - Set up Workflow



ORTEC

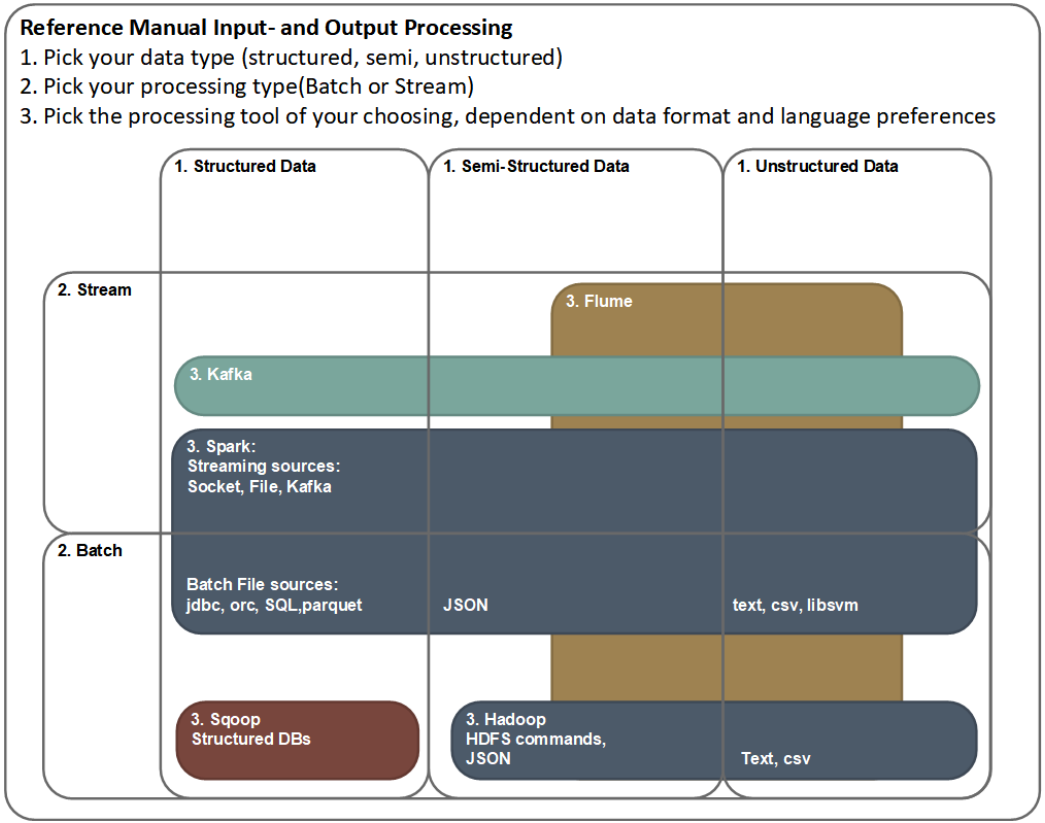
Hadoop Ecosystem



ORTEC

Artifact Design

- Determine Processing
  - Input Processing
  - Data Transformations
  - **Output Processing**
  - Set up Workflow



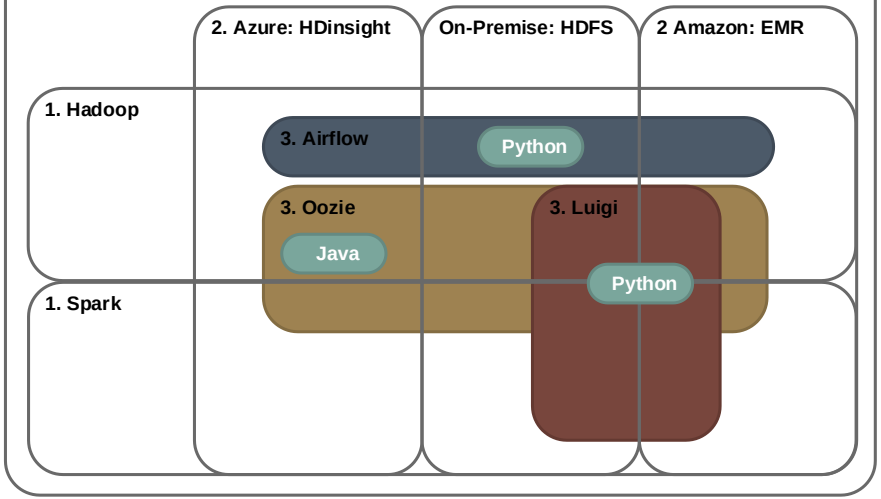
ORTEC

Artifact Design

- Determine Processing
  - Input Processing
  - Data Transformations
  - Output Processing
  - **Set up Workflow**

Reference Manual Workflows:

1. Pick your processing engine; What kind of jobs do you run(Hadoop, Spark, or both)?
2. Pick your storage; Do you use local machines, Azure, or Amazon?
3. Pick your workflow. Depended on the first two steps, and language preferences.



ORTEC

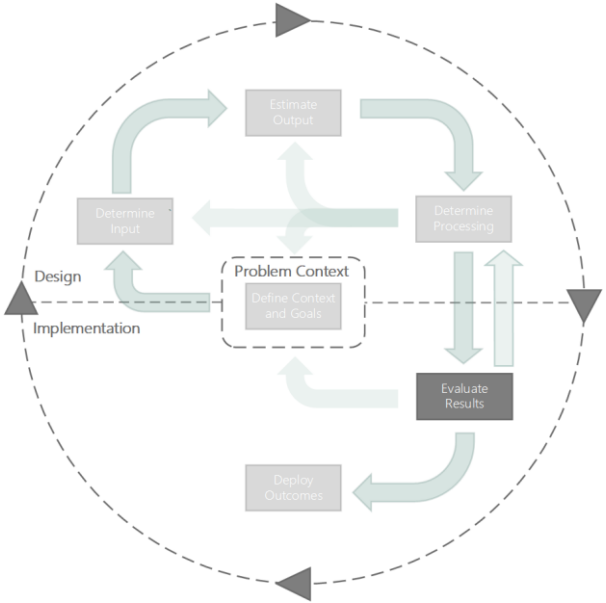
Workflow Management

- Oozie, Airflow, Luigi
- What jobs need to be executed when?
- Set timers, or execute jobs based on arrival of data
- **When to use:** Many different concurrent jobs, with specific dependencies on each other

ORTEC

Artifact Design

- Evaluate Results
  - Performance:
    - Results
    - Timing & Latency
    - Costs
  - Distributed System
    - Scalability
    - Distribution Transparency
    - Resource Sharing

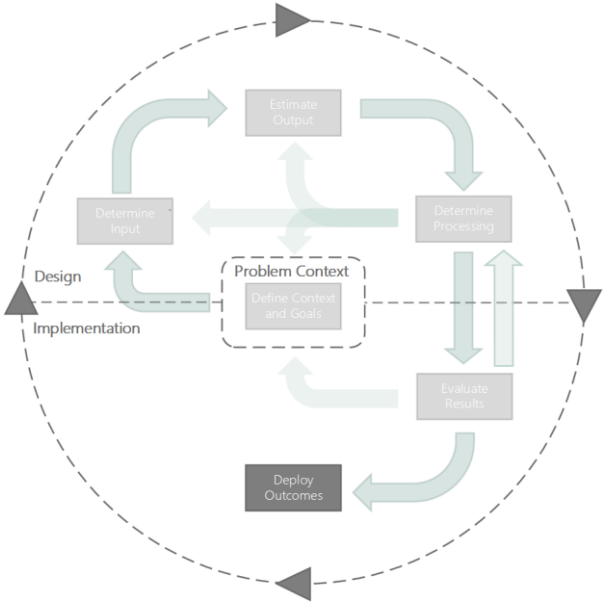


CRISP-DCW

ORTEC

Artifact Design

- Deploy Outcomes
  - Deployment Plan
  - Conclusions & Lessons Learned

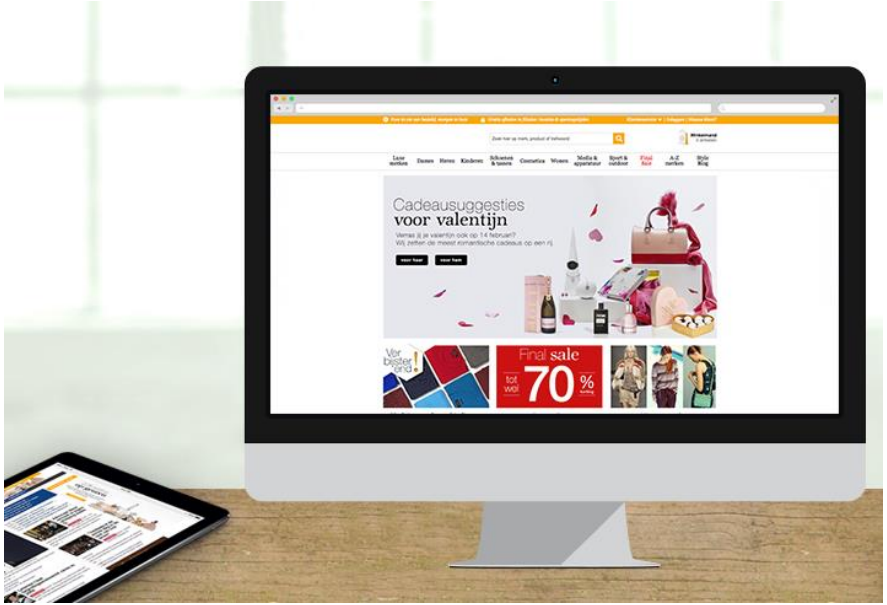


CRISP-DCW

ORTEC

Real-Time Bidding (with AdScience)

Real-time bidding for performance display advertising



BUSINESS CHALLENGE

- Decide which add to publish on a website

ANALYTICS & OPTIMIZATION TECHNIQUES

- Naïve Bayes
- Scala, C++, MongoDB, Druid

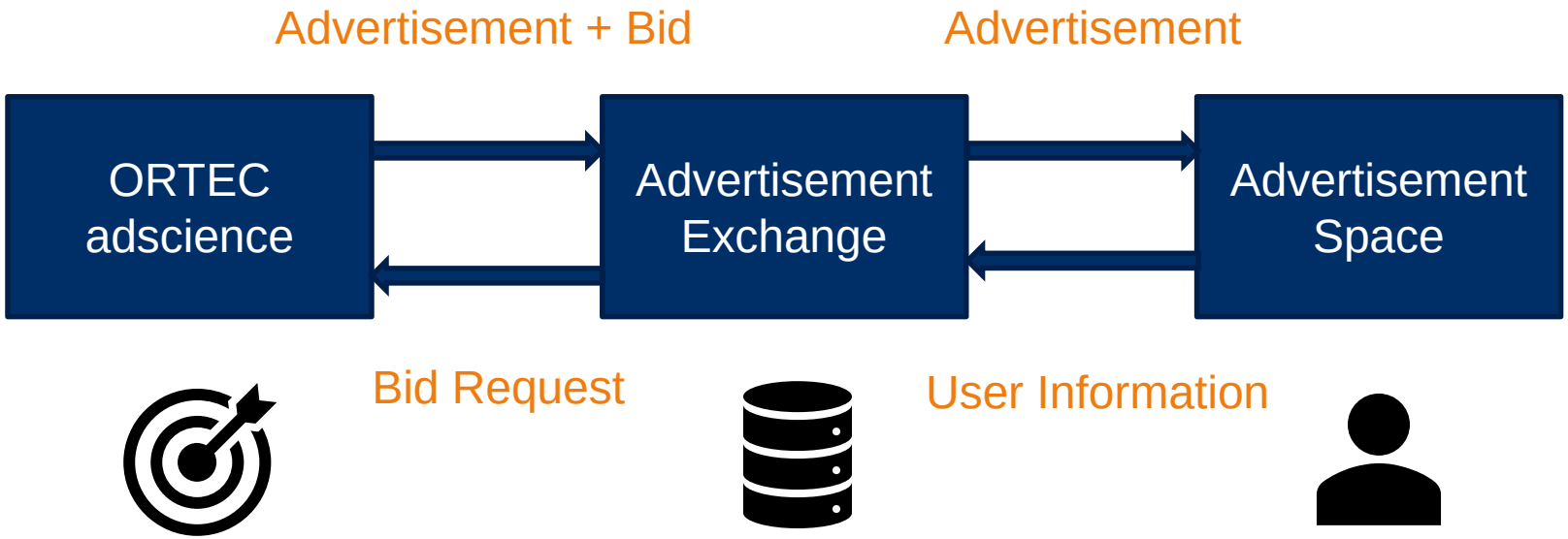
BUSINESS VALUE

- Improved CTR (CPA) and conversions by showing more relevant banners
- Transparency, real-time insights
- Smart media buying decisions

ORTEC

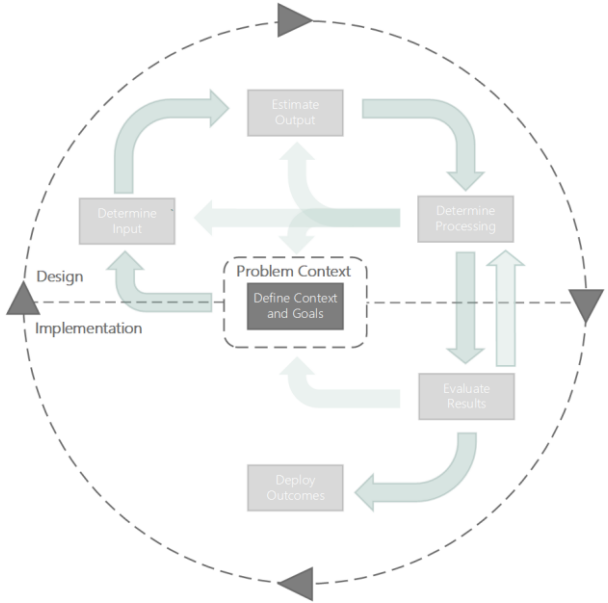
Putting the method into practice: Location Based Adserving

- ORTEC Adscience: Demand Side Platform



Design Validation: Define Context and Goals

- Geographical based targeting:
  Approximate locations of potential viewers, using data in Bid Requests
- Creation of dataset containing IP addresses and their approximate location
- Querying of dataset using unknown IP addresses
- Low budget, needs to be understandable, Azure has preference

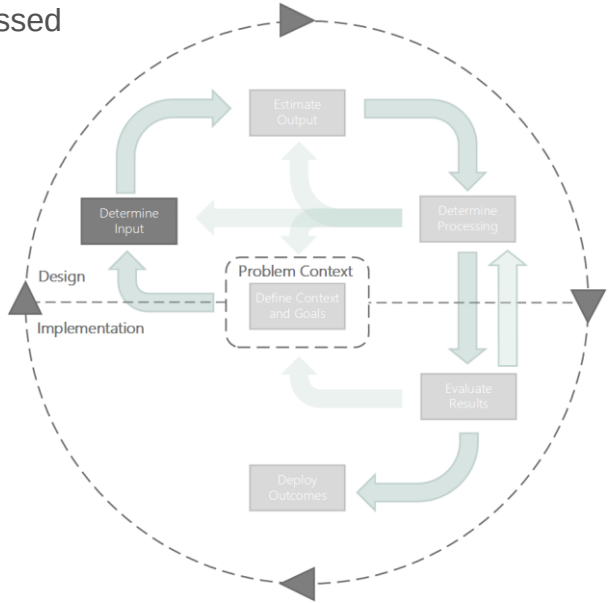


CRISP-DCW

Design Validation: Determine Input

- Volume: 500 GB compressed data per week, 5.5 TB uncompressed
  3000 files, 500 million rows
- Velocity: Batch analysis, completed every week,
  at the start of the week
- Variety: JSON files containing bid-request data, Gzip
  compressed, 66 variables
- Variability: Little change expected

- Value: Low current value, potential to create valuable insights

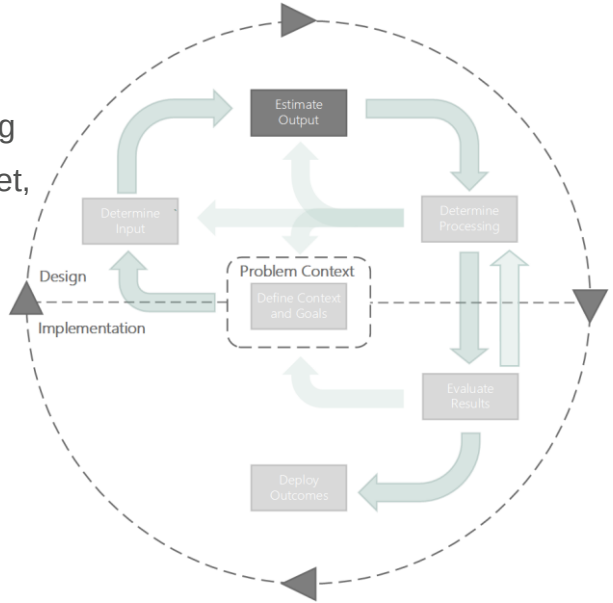


CRISP-DCW

ORTEC

Design Validation: Estimate Output

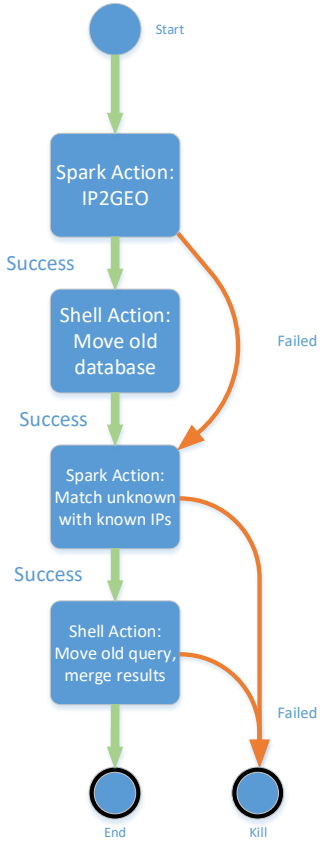
- Volume: Query results of about 10 MB & Dataset of about 10 GB
- Velocity: Set times for arrival of output data, batch processing
- Variety: Compressed files, preferably .parquet for the dataset, and JSON, CSV or TXT for the query results
- Variability: File size increases weekly, due to larger amount of raw data
- Results & Visualization: Human readable match results, further processing in Hadoop environment needed. Visualizations are nice-to-have.



CRISP-DCW

Design Validation: Determine Processing

- Cluster and Storage: Microsoft Azure Blob Storage & HDInsight
- Two Spark & two Hadoop jobs:
  1. Spark: Dataset creation of IPs and locations, using Cadaster data and bid-request data
  2. Hadoop: Moving old dataset
  3. Spark: Matching unknown IPs from a list, with IPs in the created dataset
  4. Hadoop: Merging files and removing old query
- Oozie workflow engine: Hadoop and Spark support, available on Azure



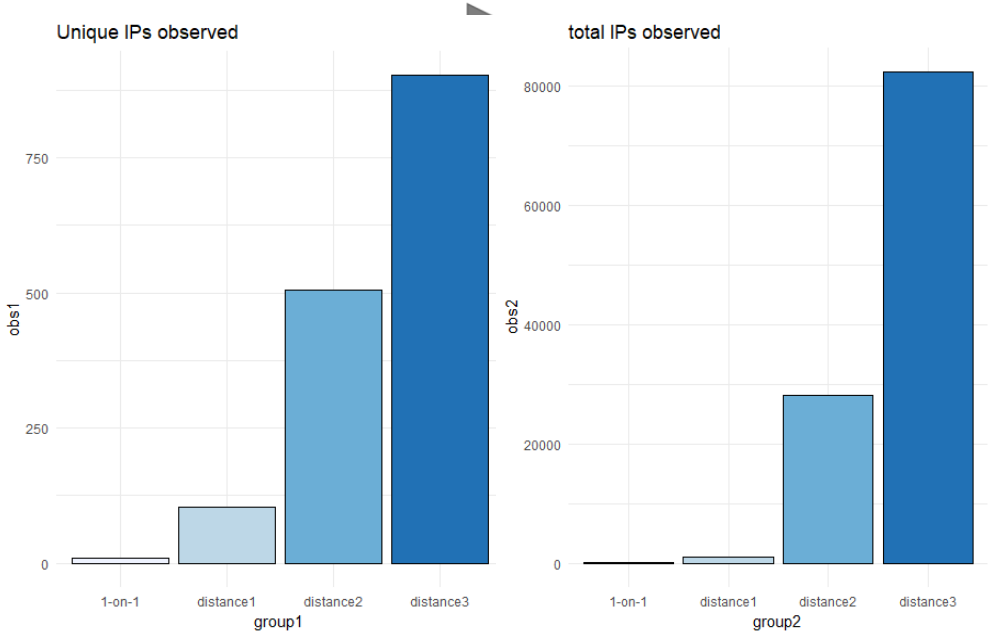
Apache Spark: Unified Big Data Processing

ORTEC

Design Validation: Evaluate Results

- **Performance: Results**

Result	Unique IPs observed	Total complete IPs
Geographical dataset	1785504	673220791
Query List	4141	
1-on-1 match	10	56
Distance 1	103	1051
Distance 2	506	28170
Distance 3	902	82326
Results total	902	108585

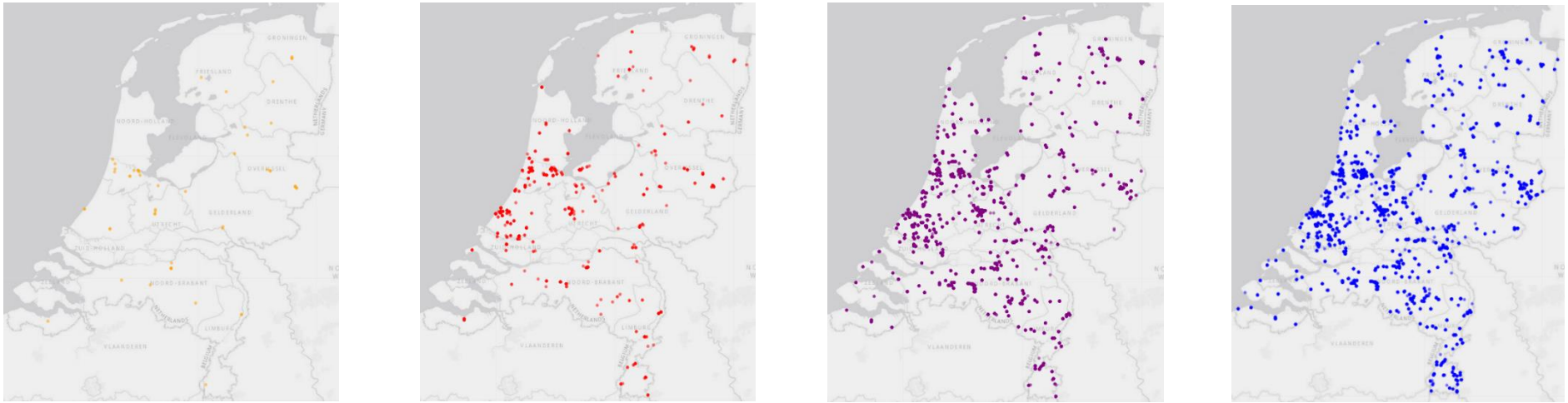


CRISP-DCW

ORTEC

Design Validation: Evaluate Results

- **Performance: Results**



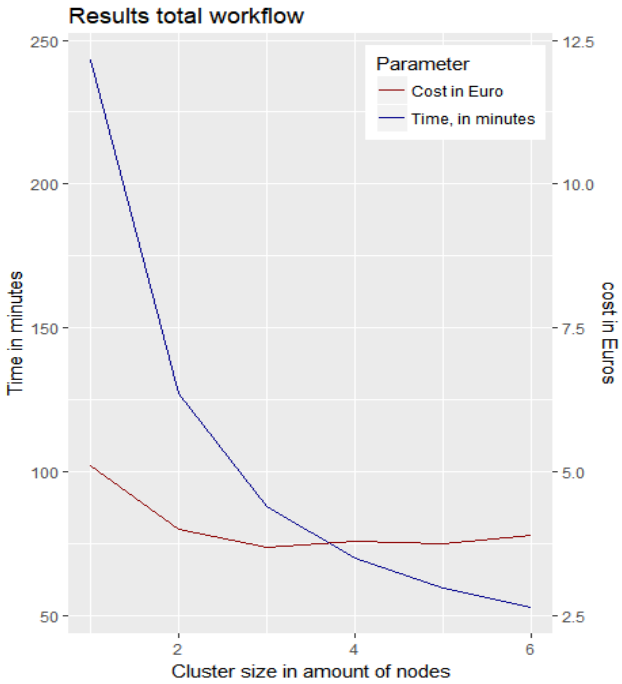
1-on-1 matches

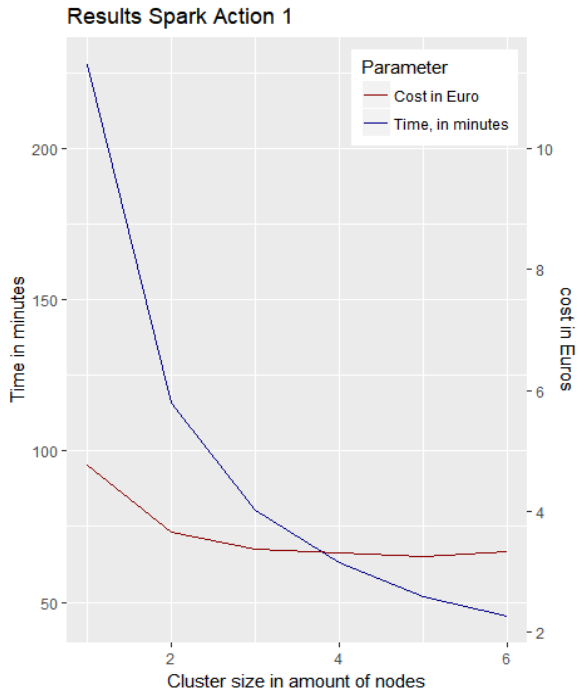
10.000 addresses difference

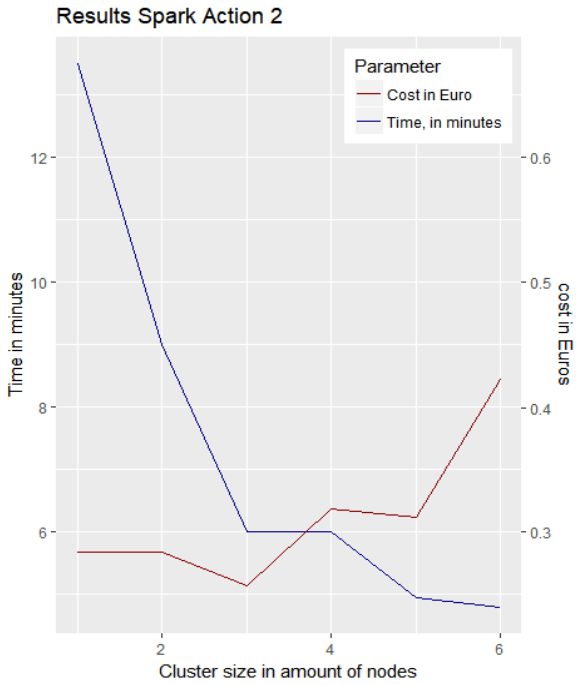
ORTEC

Design Validation: Evaluate Results

- **Performance: Timing and costs**

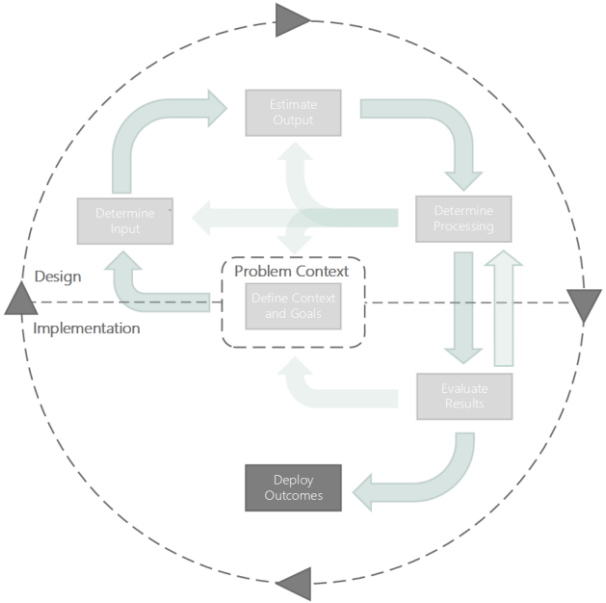






Design Validation: Deploy Outcomes

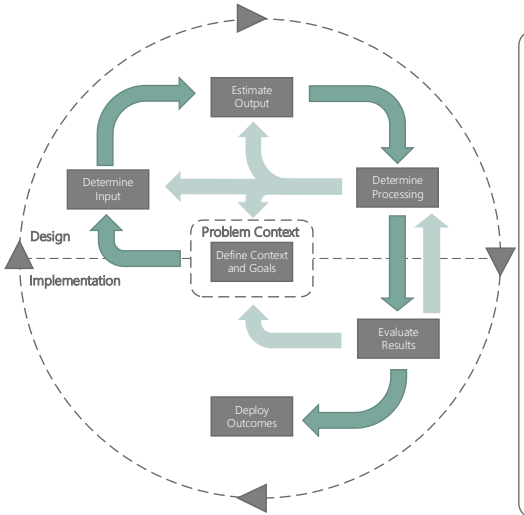
- Manual deployment during first month
- Upgrade to next versions of Spark, look into other workflow engines.
- Results deemed to be satisfactory by stakeholders, and final report is delivered to stakeholders

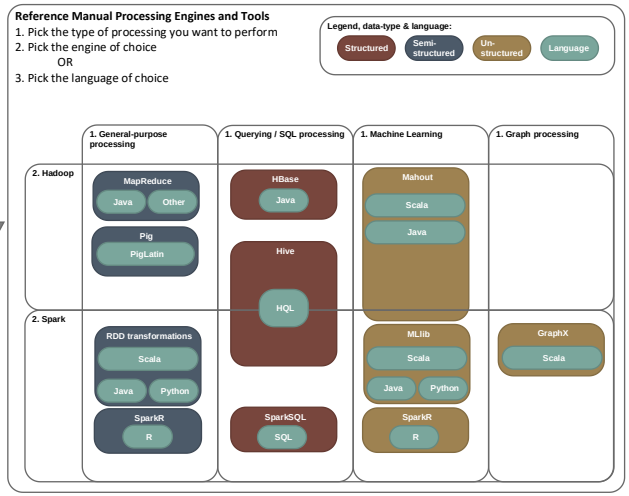


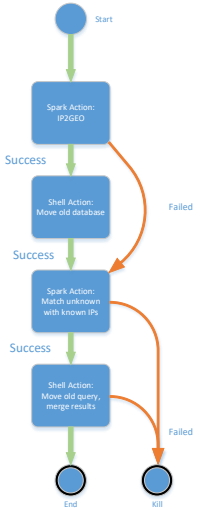
CRISP-DCW

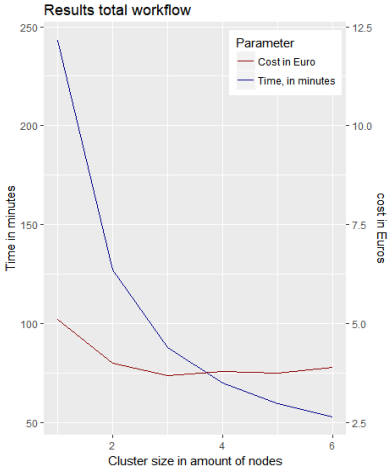
ORTEC

Conclusions



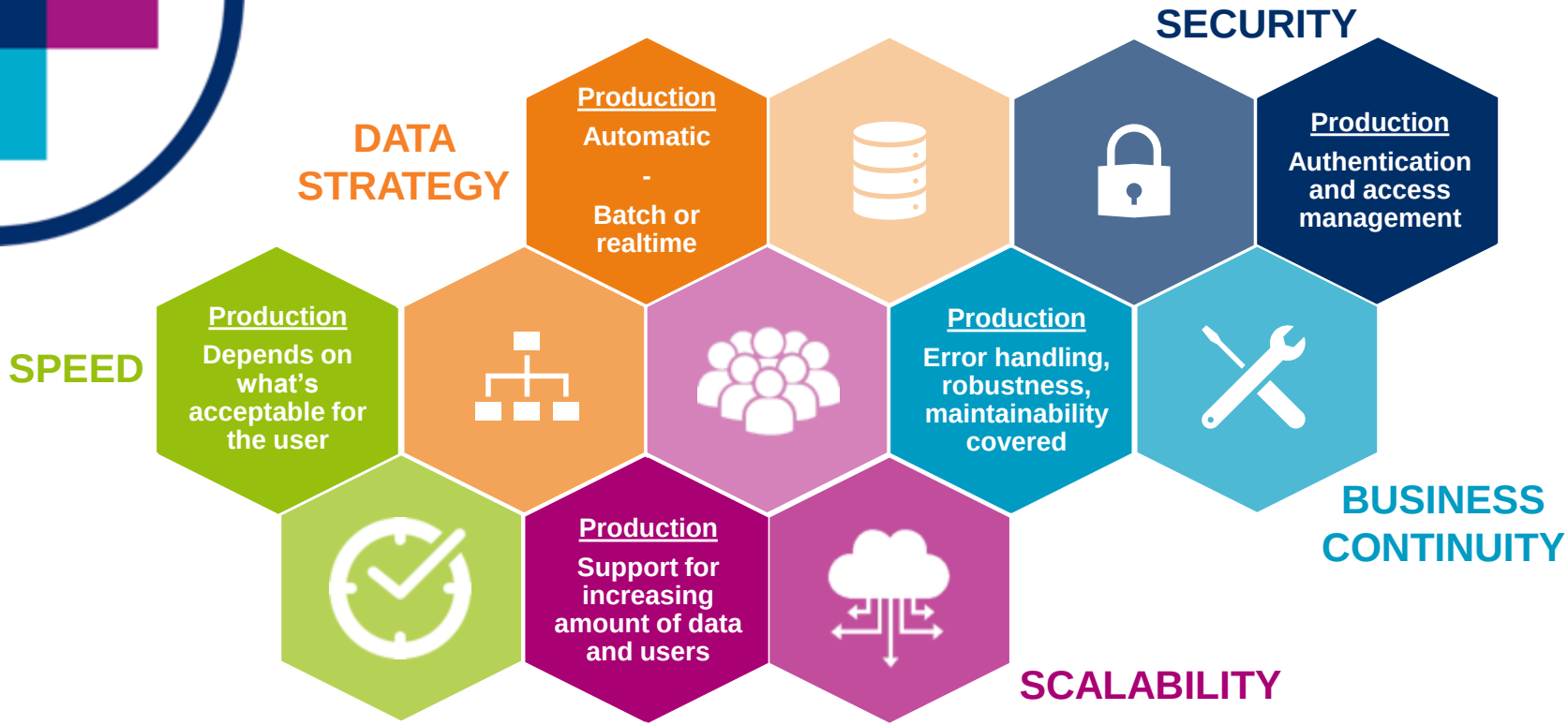






ORTEC

Proof of concept vs production



ORTEC

Some final remarks…

Remember to…

- Always make sure you have a goal in mind, and document that goal: Too many data science projects never make it to production, because the project never got past the proof of concept phase. Making those is expensive!

- See data engineers and data scientists as a team, in the world of big data one cannot live without the other.

- Experiment and learn from the different tools and techniques available! There is a very high demand for well-rounded data scientists and engineers (Go beyond R)

- Most importantly, have some **fun**; there are many opportunities in the industry right now and find something you really like☺

Interested in collaboration / thesis subjects / a job?

Contact me!

Stijn Meijers

Data Engineer

ORTEC

@ORTEC

www.ortec.com

stijn.meijers@ortec.com

Let’s optimize your world

ORTEC