The history and future of data processing by FP Data Solutions

https://datasolutions.future-processing.pl/

Speakers

Tomasz Krawczyk
Principal Data Architect
tkrawczyk@future-processing.com
https://www.linkedin.com/in/tomasz-krawczyk-1a531328/

Krzysztof Nykiel
Head of Data Solutions
knykiel@future-processing.com
https://www.linkedin.com/in/krzysztof-nykiel/
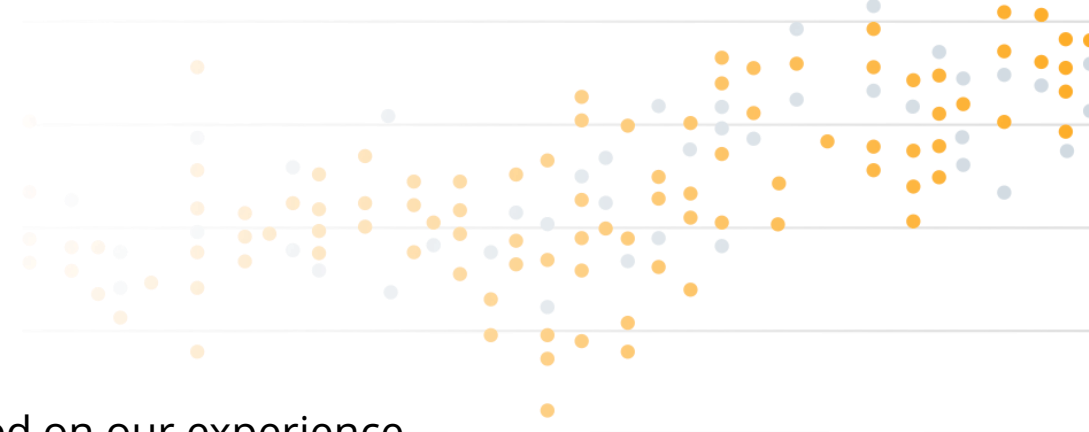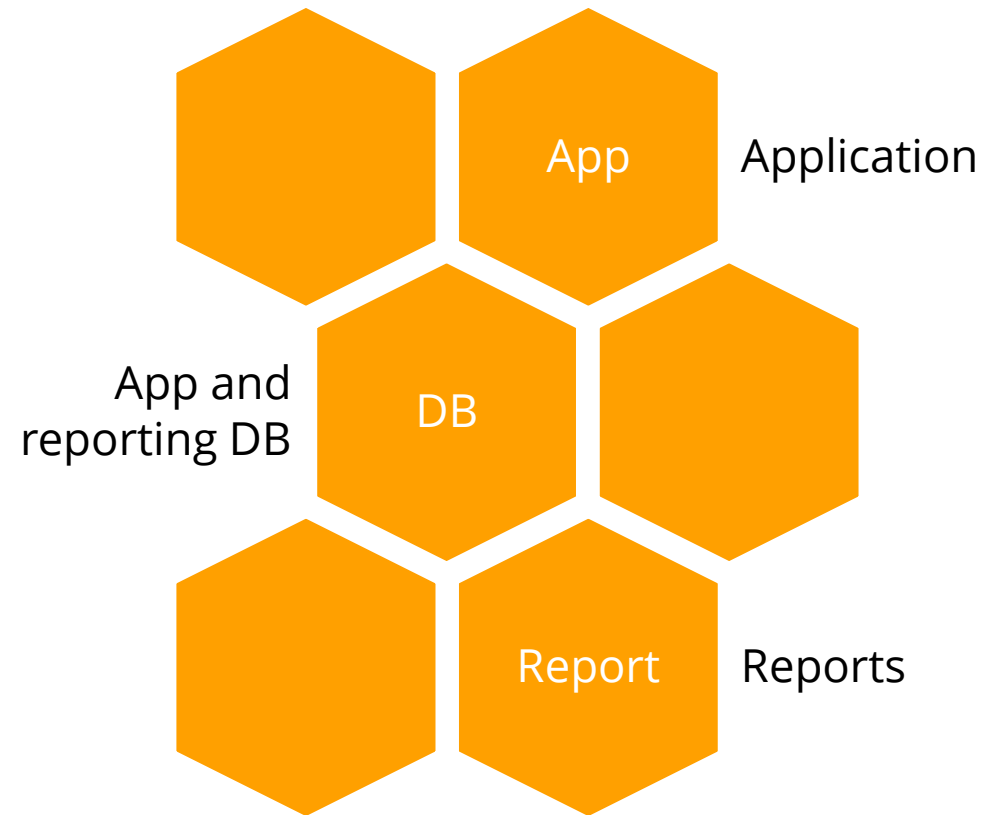
Concept for today's session

- Our subjective view on the data processing world based on our experience

- We encourage you to join the discussion and share your thoughts

- We'll go through a couple of concepts and a few examples and try to give you some takeaways
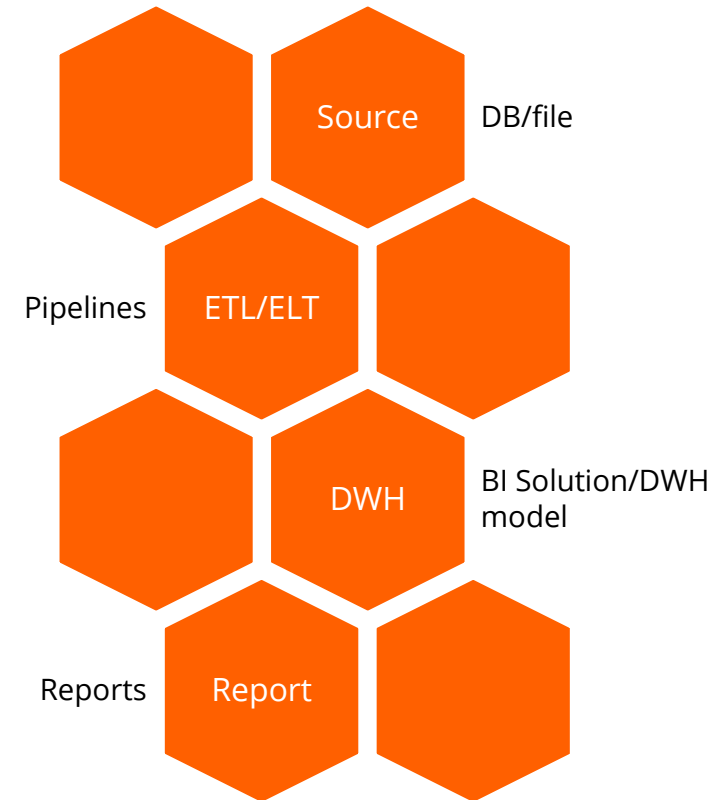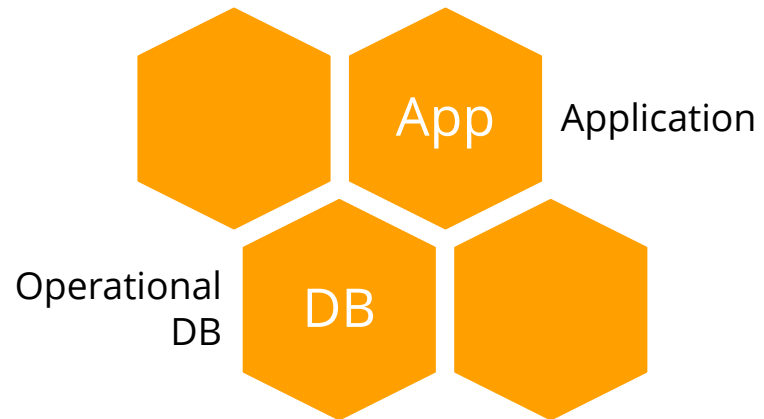
- Enjoy!☺

How it used to be – all in one

App — Application

App and reporting DB

DB

Report — Reports

How it used to be – operation/reporting division

App

Application

Operational
DB

DB

DWH

BI Solution

Reports

Report

How it used to be – operation/BI division

App — Application

Operational DB — DB

Source — DB/file

Pipelines — ETL/ELT

DWH — BI Solution/DWH model

Reports — Report

How it used to be – full division

App — Application

Operational DB — DB

Source — DB/file

Pipelines — ETL/ELT

Data model — BI Solution/DWH model

DWH engine — DWH

Data model — Data model for reporting

Presentaion and visualization — Report
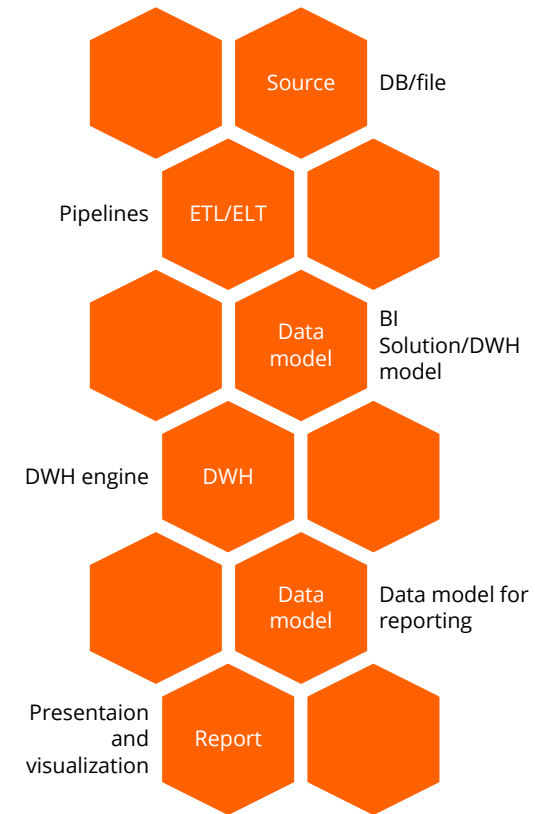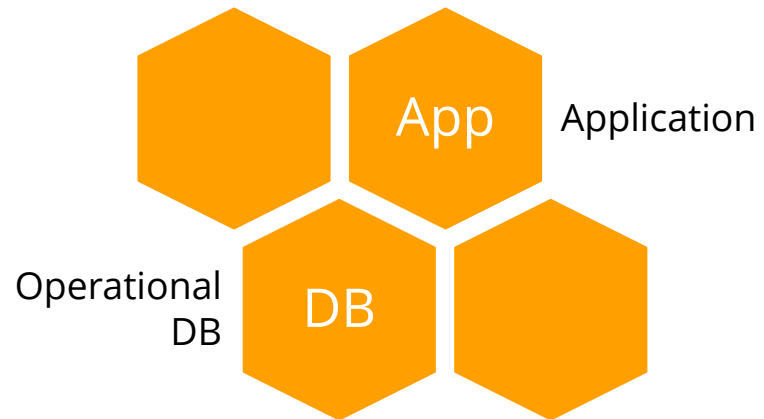
Takeaways

- Either the world is getting more complex or our development approach compicate things (or both)

- It's also possible that with time we get more experienced and knowledgeable and tackle more complex business problems, which have been there for a while, but we haven't seen them before (we tend to look locally, at out own playground)

Another perspective – DWH modelling

- Since 80s we observe evolution in the approach to modeling data for warehousing

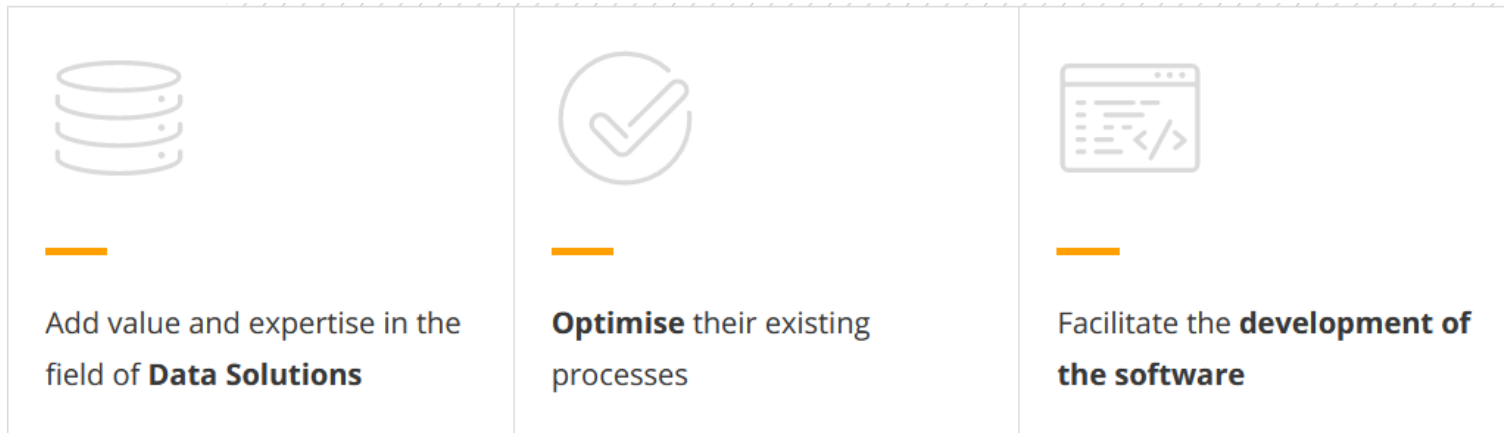**1990** Inmon → **1996** Kimball → **2000** Data Vault →
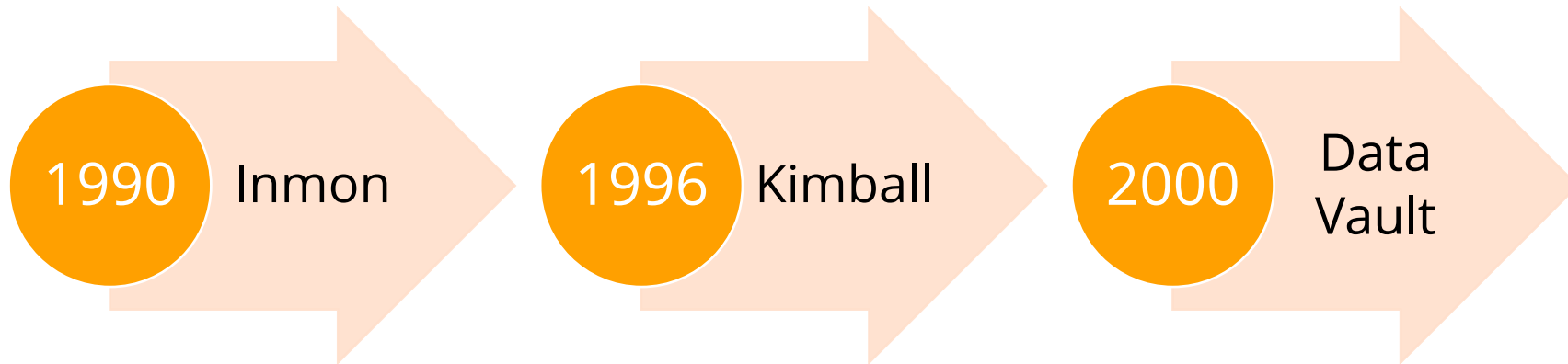
- According to our experience – in the end DWH model ends up to reflect Kimball's approach
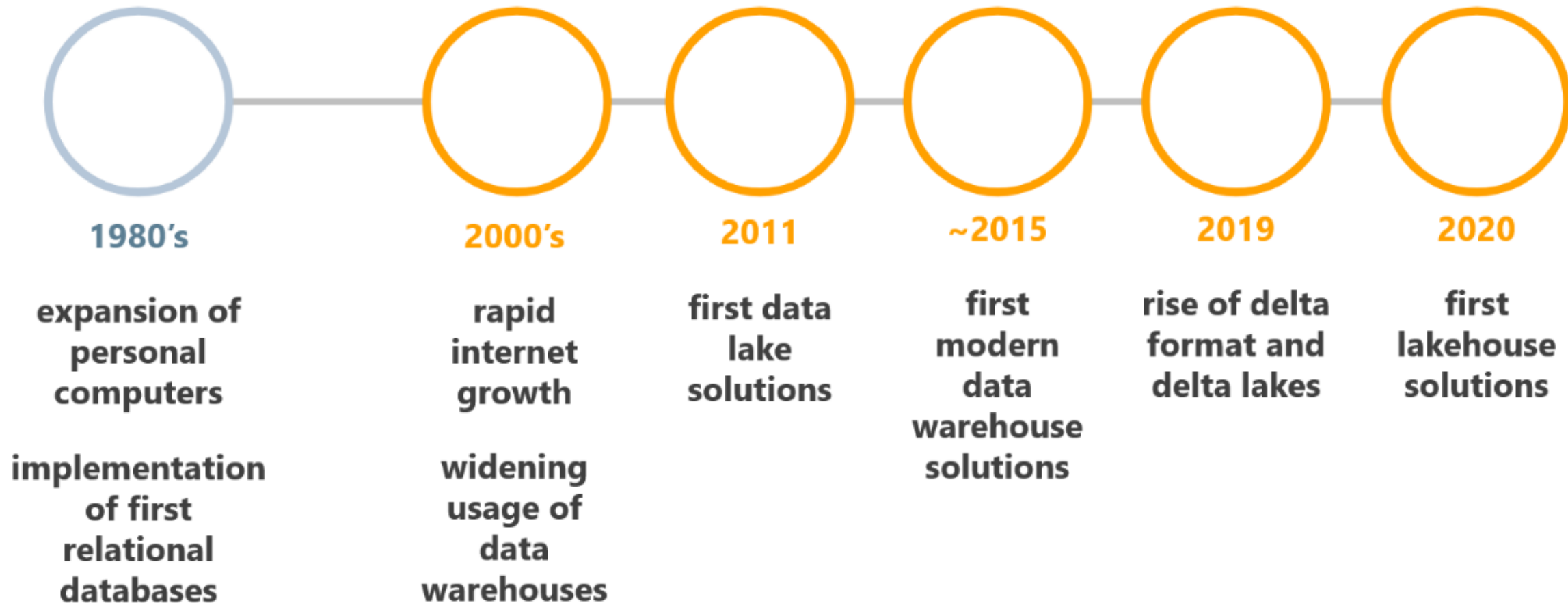
Classical DWH project example (PACT)

- Data warehouse supporting business processes in the Retail industry

- Designed in Kimball and built on MSSQL Server, SSIS and BIML

- Digital Transformation – this solution replaced manual processes and legacy BI tools

Add value and expertise in the field of **Data Solutions**	**Optimise** their existing processes	Facilitate the **development of the software**

Case study available here

Brief history of data processing concepts

1980's

expansion of personal computers

implementation of first relational databases

2000's

rapid internet growth

widening usage of data warehouses

2011

first data lake solutions

~2015

first modern data warehouse solutions

2019

rise of delta format and delta lakes

2020

first lakehouse solutions

Example of data lake (Big Data) project (Tauron)

- Manage large amounts of data coming from advanced metering infrastructure

- Advanced analysis of large sets of measurement data, which contributes to:

  - Increased observability / monitoring of the distribution network

  - Making more effective decisions in terms of maintaining and developing network infrastructure

  - Increasing operational efficiency and numerous business benefits for DSOs (Distribution System Operators)
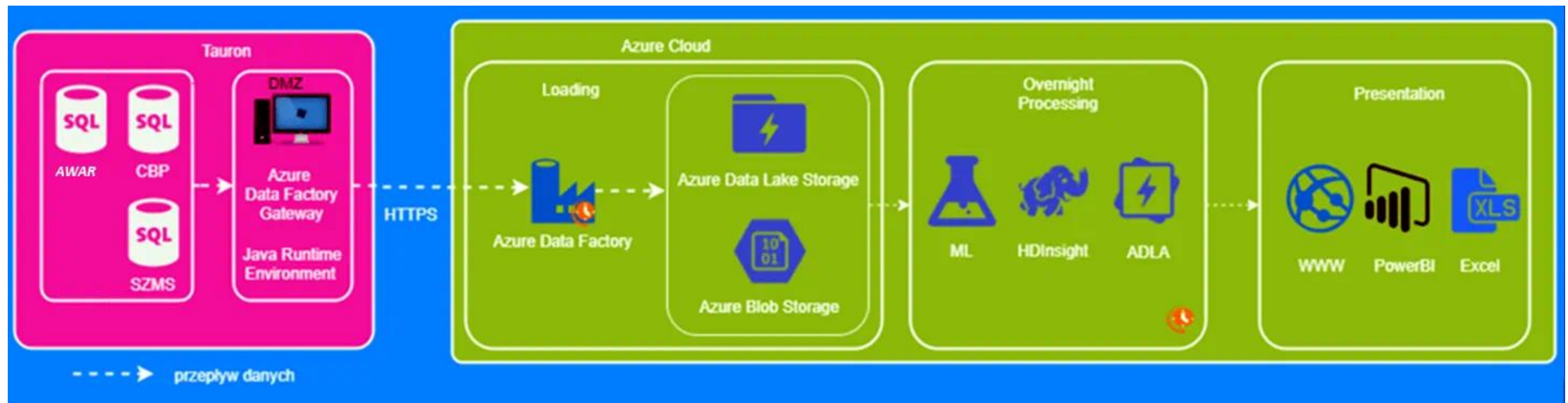
Case study available here

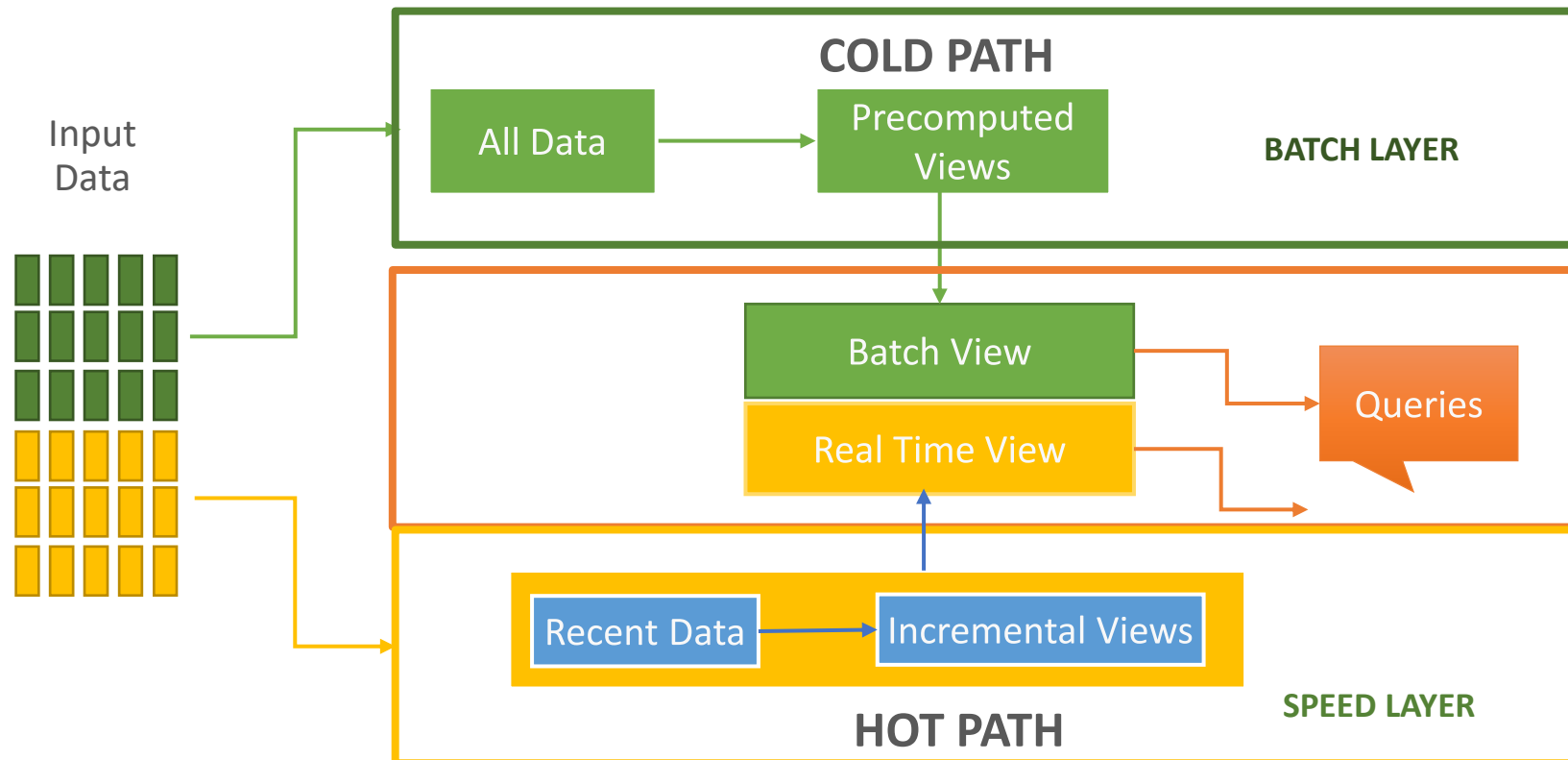# Example of data lake (Big Data) project (Tauron)

- Technologies:

  - Azure (ADF, Data Lake, Blob Storage)

  - Big Data (Hadoop, HDFS, Spark)

  - Machine Learning, Statistical Processing, IoT

  - Visualisation
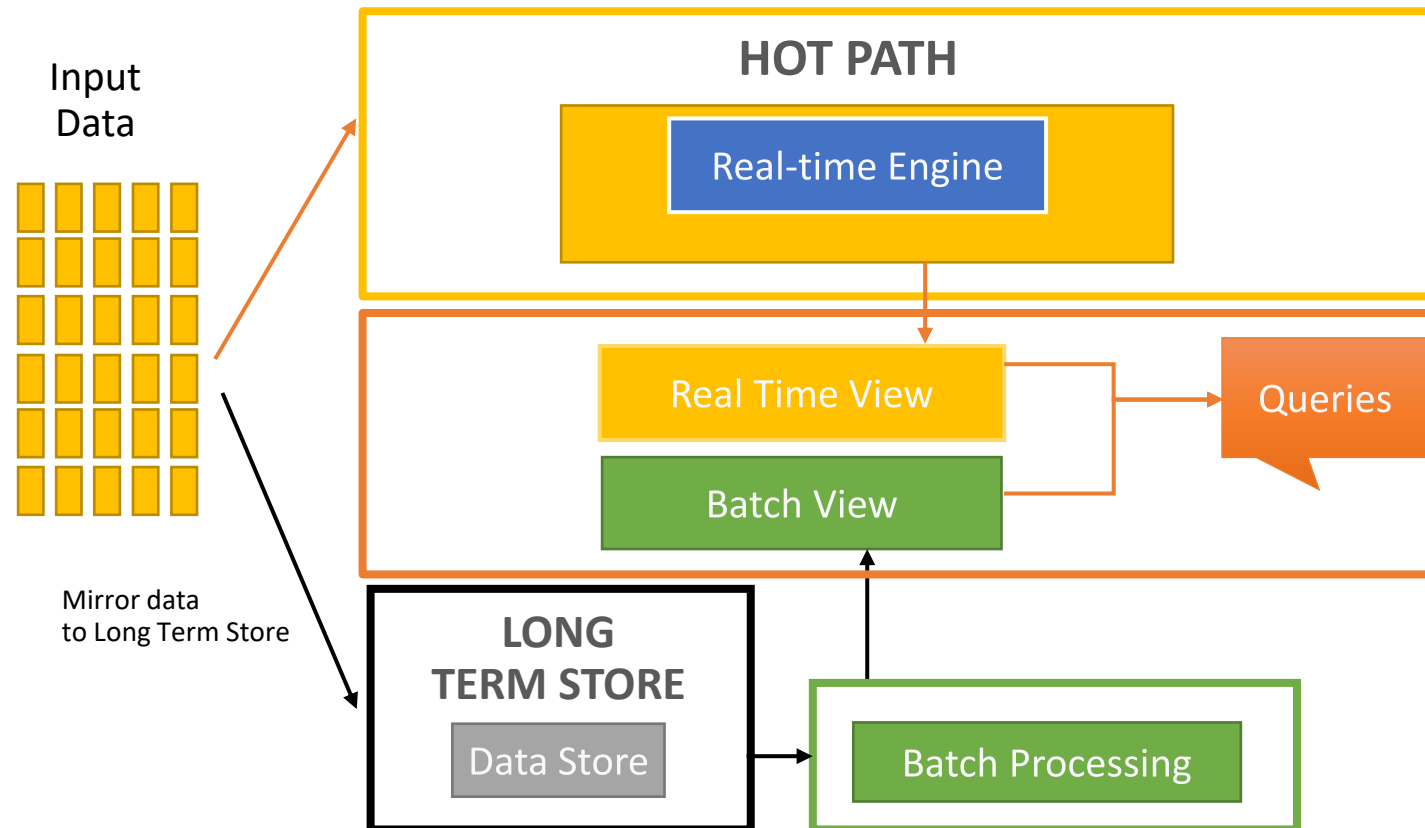
Three temperature concept to data lake

Lambda architecture

Three temperature concept to data lake

Kappa architecture

DWH vs DL vs DLH comparison

	Data Warehouse	Data Lake	Data Lakehouse
Data format	Closed, proprietary format	Open format	Open format
Types of data	Structured data, with limited support for semi structured data	All types: Structured data, semi-structured data, textual data, unstructured, (raw) data	All types: Structured data, semi-structured data, textual data, unstructured (raw) data
Data access	SQL-only	Open APIs for direct access to files with SQL, R, Python, and other languages	Open APIs for direct access to files with SQL, R, Python, and other languages
Reliability	High quality, reliable data with ACID transactions	Low quality, data swamp	**High quality, reliable data with ACID transactions**
Performance	High	Low**	High**
Scalability	Scaling becomes exponentially more expensive*	Scales to hold any amount of data at low cost, regardless of type **	Scales to hold any amount of data at low cost, regardless of type **
Use cases	Limited to BI, SQL applications, and decision support	Limited to machine learning *	**One data architecture for BI, SQL, and machine learning (and streaming)**

Novadays – division and staging - ISASA

- BI and data engineering are adopting best practices in software development (formely present mainly in application dev)

- Different stages may be realized by different data processing approaches:

  - Data lakes

  - Modern data warehouses

  - Delta lakes

  - Date lakehouses

ISASA defines the steps that must be taken to build a solution. These steps include:

I – Ingest
data collection

S – Store
data storage

A – Analyse
data analysis

S – Surface
presentation of the prepared data

A – Act
4xM Make Me More Money

FP Generic Metadata Framework

- The **Generic Metadata Framework** helps the solution creator in concentrating on its business aspects. What is more, it simplifies and automates

  - **processes of data loading** (supporting both full and incremental data loading),

  - **building data lakes** (defining the structure, data partitioning),

  - **initial data processing** (transformation of input data),

  - **building delta lakes** (defining the structure, data partitioning),

  - **creating data warehouses** (defining the model in the views).

The **Generic Metadata Framework** makes it possible to create solutions based on the **modern data warehouse** or **data lakehouse** concepts in the cloud, in a quick and agile way. It relies on services such as Azure Synapse Analytics and Azure Databricks.

The graph below presents the areas covered by the framework:
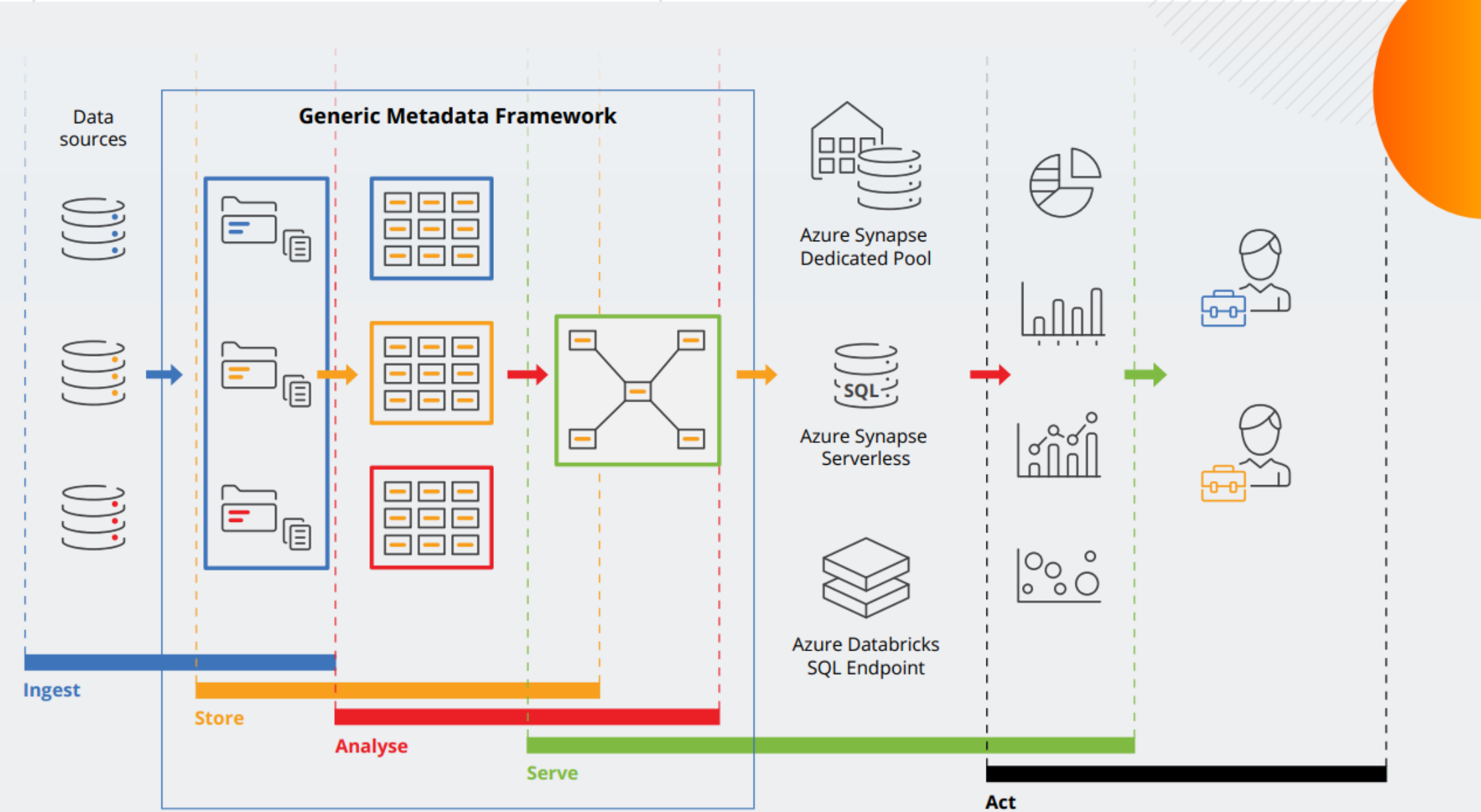
Cooperation with NGOs – Zielone Gliwice 3.0

- Project for a green change
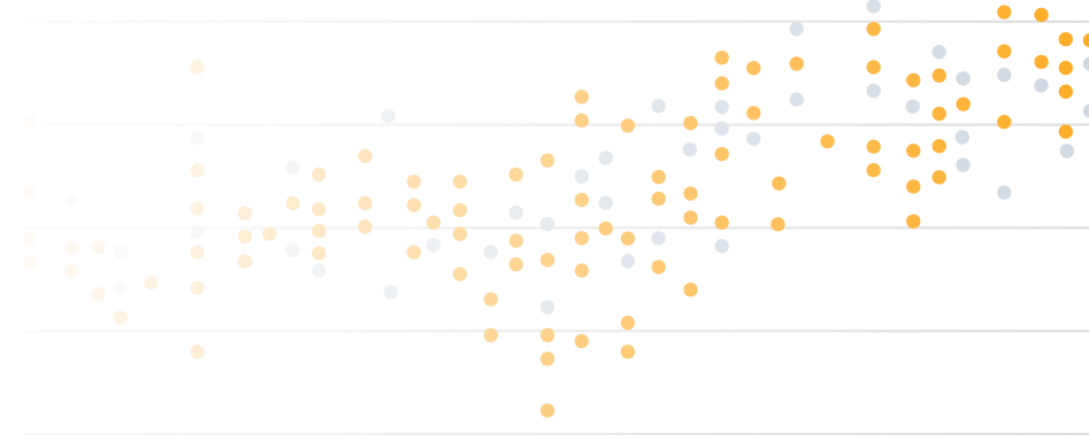
| Mobile phone | Picture of a tree | Endpoint | ML – tree recognition | Tree catalog |

Data mesh concept

- Engineer's challange:

    - We need to respond to business' requirements and follow new technological concepts and trend in order to deliver solutions according to best practices

    ~Zhamak Dehghani

- Data mesh is an interesting concept, which changes the approach to data serving

- It's not for everyone – the highest benefit will be gained by large organizations

Data mesh concept
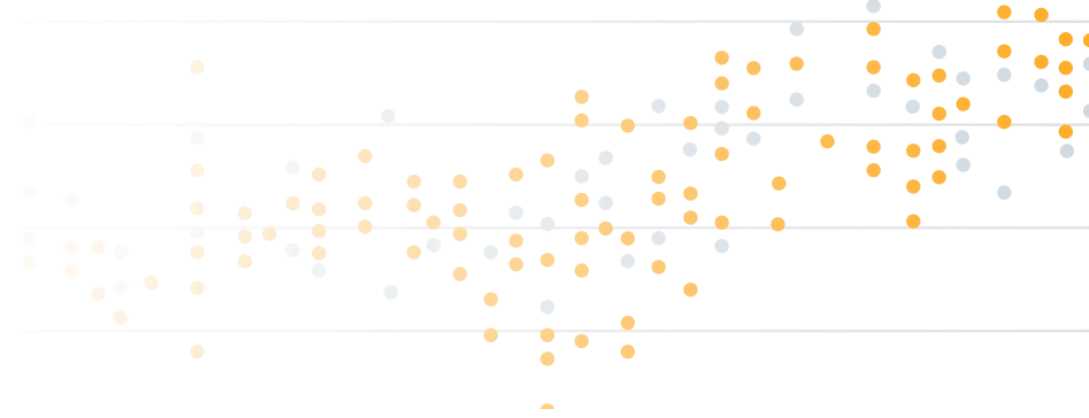
Cross-functional
Domain oriented source teams

Hyper-specialized
Data & ML Platform Engineers

Cross-functional
Domain oriented consumer teams

Takeaways

- Close cooperation between Data Engineers, Data Analysts and Software Developers is a key to understand that data is a product, not a side effect of app development

- Client's business problem is a key driving factor to select appropriate solution, not the opposite

- Our services mature and tend towards consulting approach in comparison to pure technical advisory and development
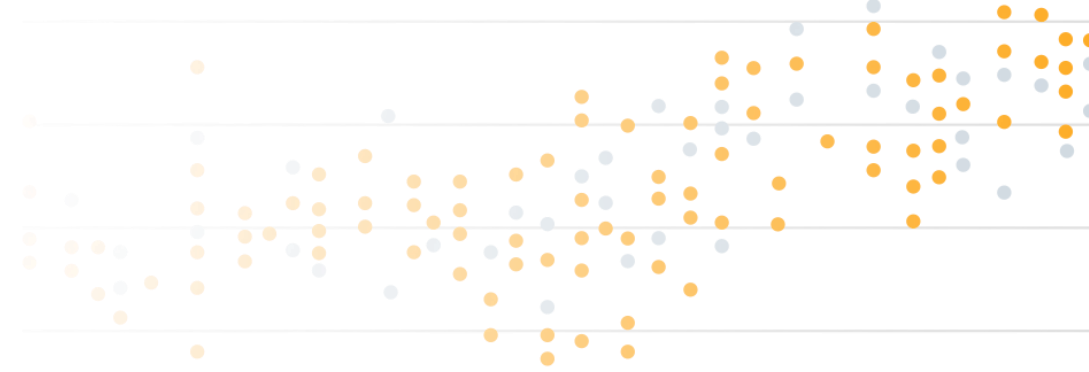
Last but not least – blockchain (Smart Steel)

SMARTSTEEL

Access to authenticated information about lifetime of a component.

Smart Steel

- Smart Steel is the response to the need for a reliable determination of the residual value of machinery in heavy industry.

- So far, this parameter has been determined based on simple premises (age, distance covered, appearance and current prices).

- Smart Steel ecosystem provides **comprehensive information about the analysed machine**, not only by wear or usage but also by monitoring the current technical condition, combustion, or just workload.

- After running a business analysis to understand the problem, we confirmed that the blockchain technology was actually a suitable solution and we were able to develop its potential in the project.

Blockchain	IoT	Machine learning
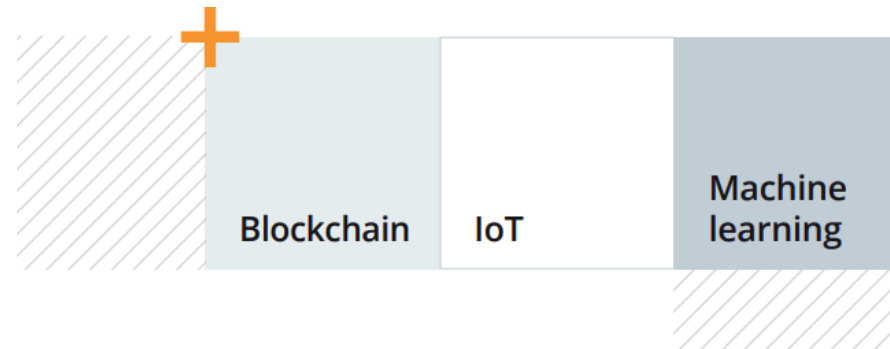
All these technologies are used in the project discussed here.

Future Processing

Q&A

https://datasolutions.future-processing.pl/

Dziękuję za uwagę!

ID Future Processing

ul. Bojkowska 37A

44-100 Gliwice

+48 32 461 23 00

datasolutions@future-processing.com

datasolutions.future-processing.pl

www.future-processing.pl