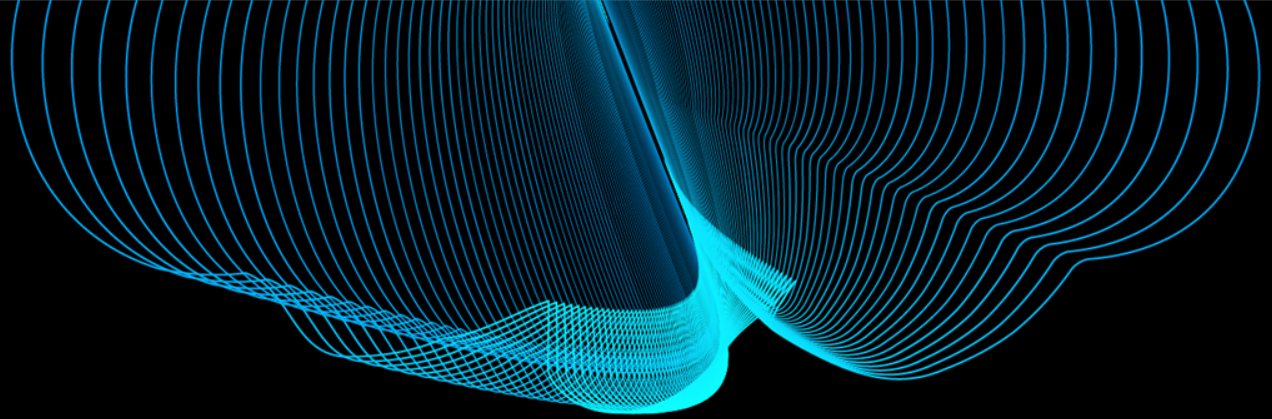


DataCloud

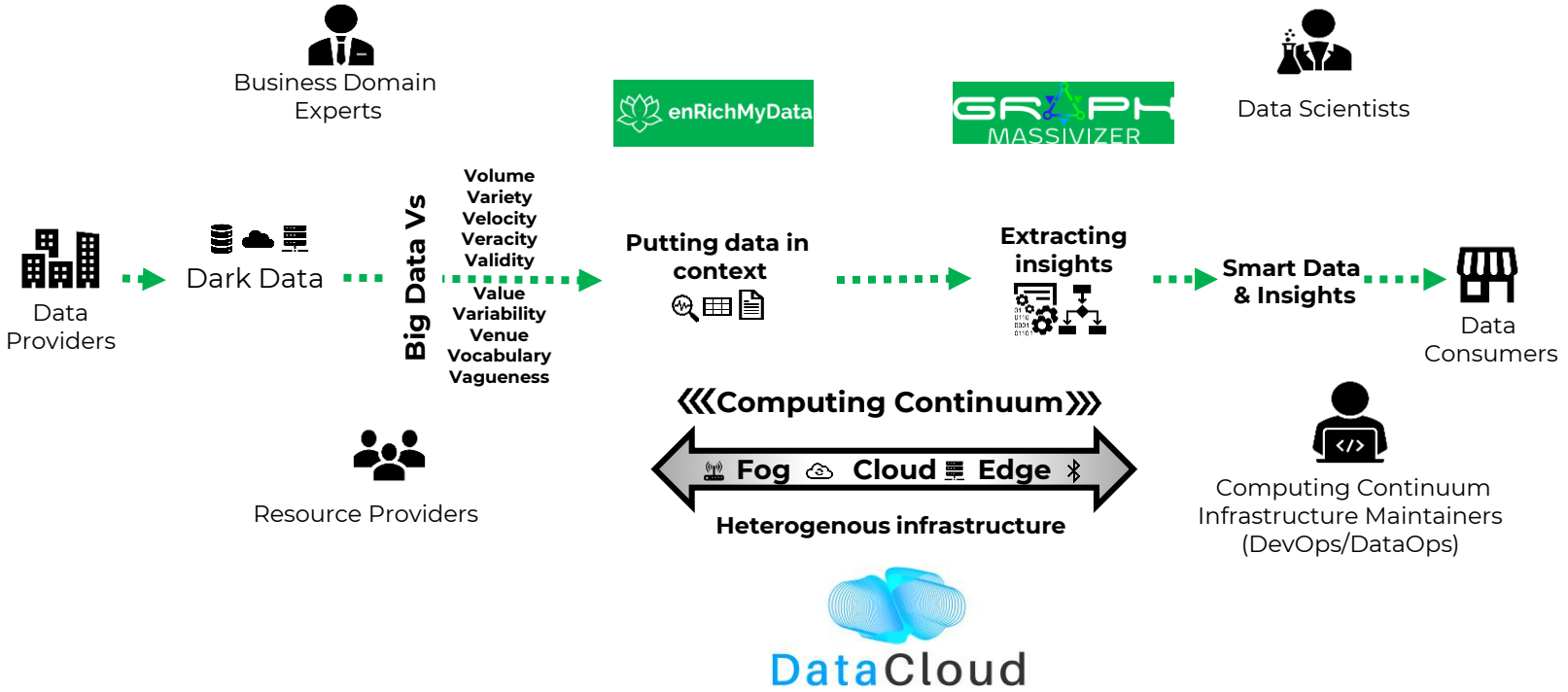
Enabling the Big Data Pipeline Lifecycle on the Computing Continuum

Dumitru Roman

SINTEF AS, Norway

DataWeek 2023 Luleå, Sweeden

A perspective on the overall context



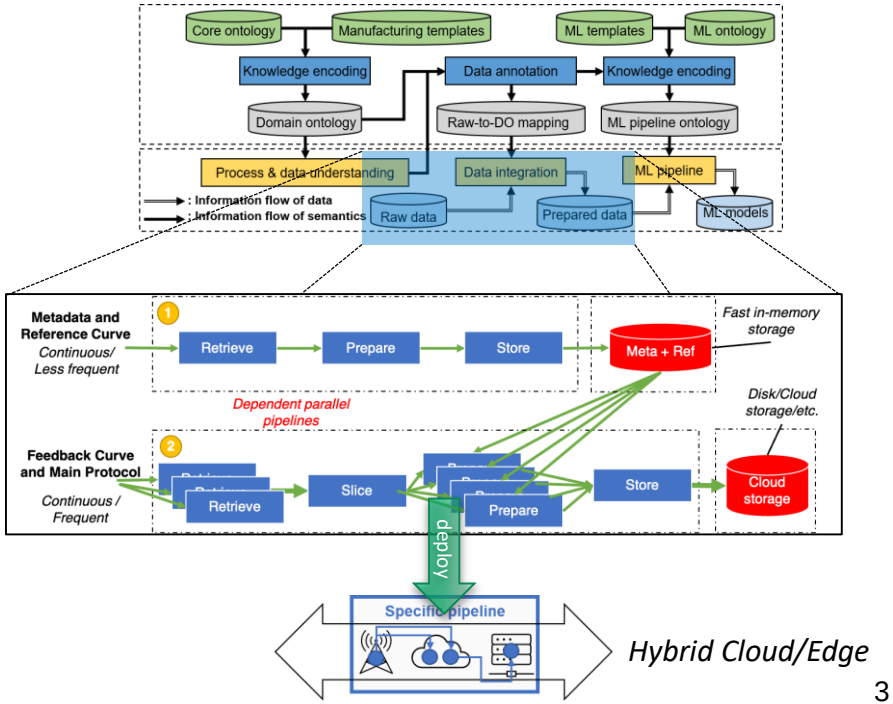
Example use case

Big Data pipelines are composite pipelines for processing data with non-trivial properties and characteristics, commonly referred to as the *Vs of Big Data*.

Industrial example: Data preprocessing for quality prediction of spot welding industrial process in car factories

Goal: To **ensure high quality** of welding spots, **predict failures** before they happen

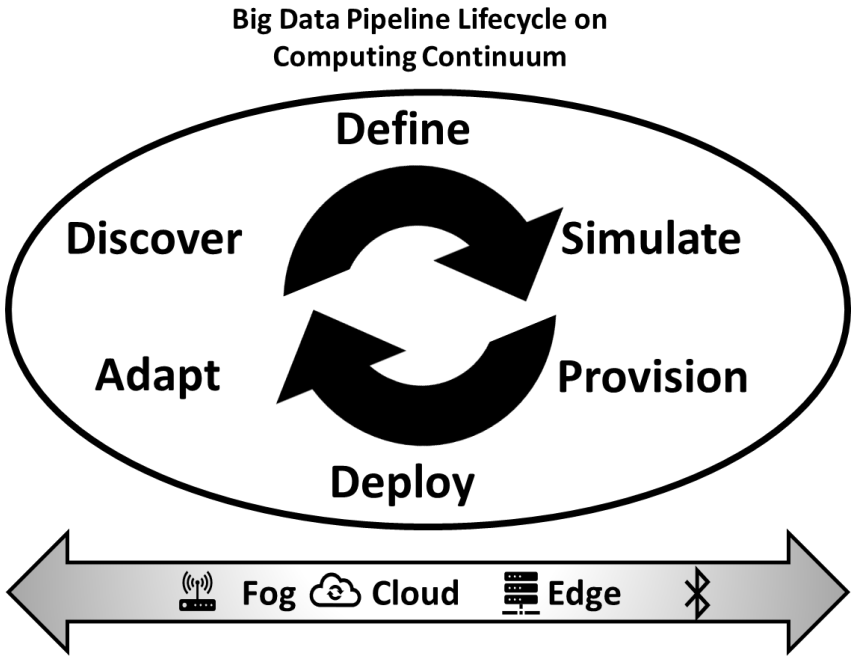
- 1 assembly line = 500-1500 car bodies/day
- 1 x car body ~ 6 000 welding spots
  - 1 factory = ~6 million welding spots per month
  - Each welding spot = at least stored in four types of datasets
- Heterogeneous infrastructure for deployment of pipeline (Cloud + Edge)



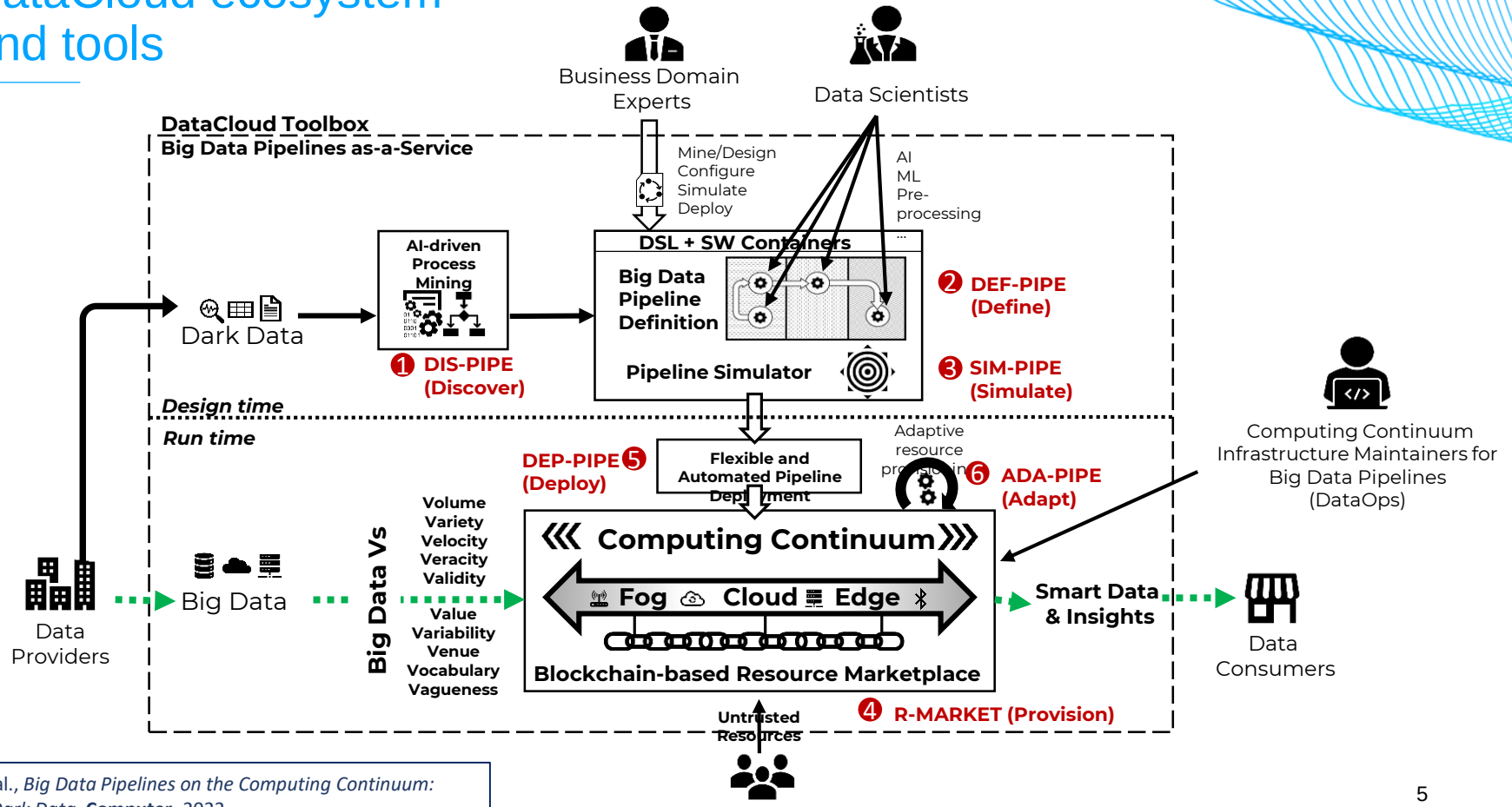
DataCloud Scope

Example: 1TB database with sensor measurements

Retrieve sensor data
Reformat data
Split data in chunks
Enrich split data
Perform ML training

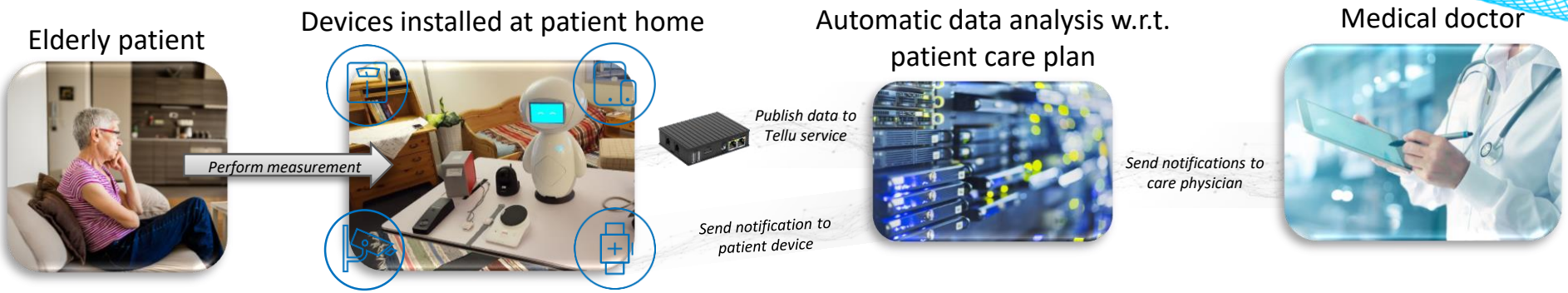


DataCloud ecosystem and tools



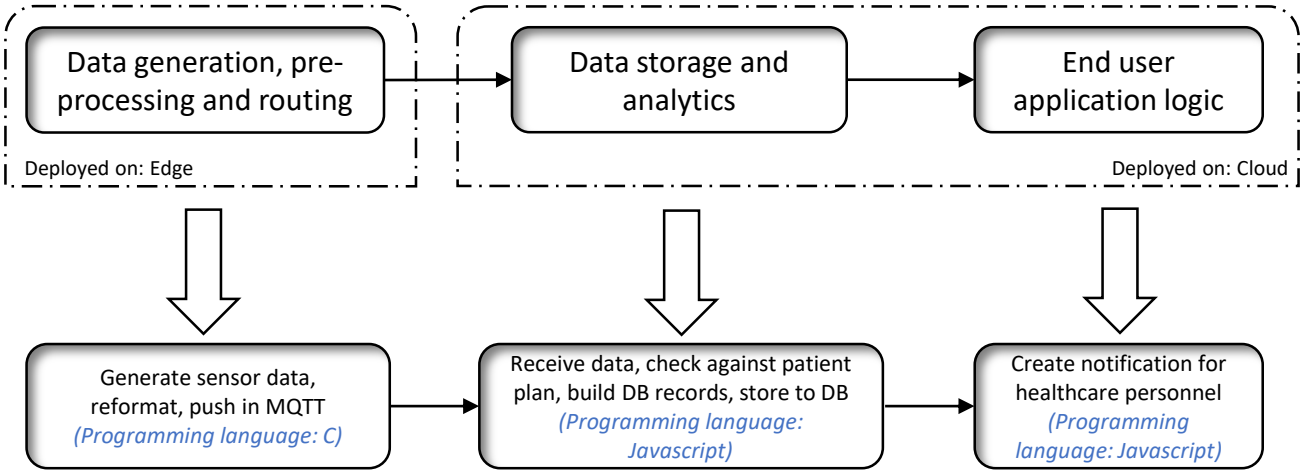
D. Roman et al., *Big Data Pipelines on the Computing Continuum: Tapping the Dark Data*, **Computer**, 2022

Application of pipelines for digital healthcare (PoC)



- Existing Tellu application/product is not designed as a data pipeline:
  - Difficult to scale individual steps/pipelines of the data processing
  - Cannot easily be replicated for new patients/devices
  - Difficult to find the optimal resource allocation for data processing steps

Proof-of-concept Data Pipeline



Design-time: Designing a data pipeline
- Configure pipeline description
- Validate pipeline
- Test configuration (before deployment)

Run-time: Distributed deployment of the data pipeline
- Provision heterogenous Edge/Cloud resources
- Automatically schedule and deploy
- Execute and monitor the pipeline

Application execution logs

1 Visually discover pipeline from logs using **DIS-PIPE**

- Import logs and manage complexity
- Display graph-based representation
- Generate DSL for discovered pipeline

Partial pipeline model

2 Visually design/edit pipeline using **DEF-PIPE**

- Edit steps of designed pipeline
- Add pipeline-specific parameters to steps
- Show DSL representation of pipeline

Configured pipeline model

3 Simulate executions of pipeline using **SIM-PIPE**

- Execute simulation of data pipeline
- Display information about runs

Configured pipeline model and run stats

4 Provision resource for pipeline steps to be executed using **R-MARKET**

- Display information about resources
- Add resources using SDK and wallet

Provisioned resources

5 Schedule execution of pipeline using **ADA-PIPE**

- Schedule resources according to defined requirements
- Generate deployment-ready schedule

Deployment schedule

6 Deployment of pipeline on Cloud/Edge using **DEP-PIPE**

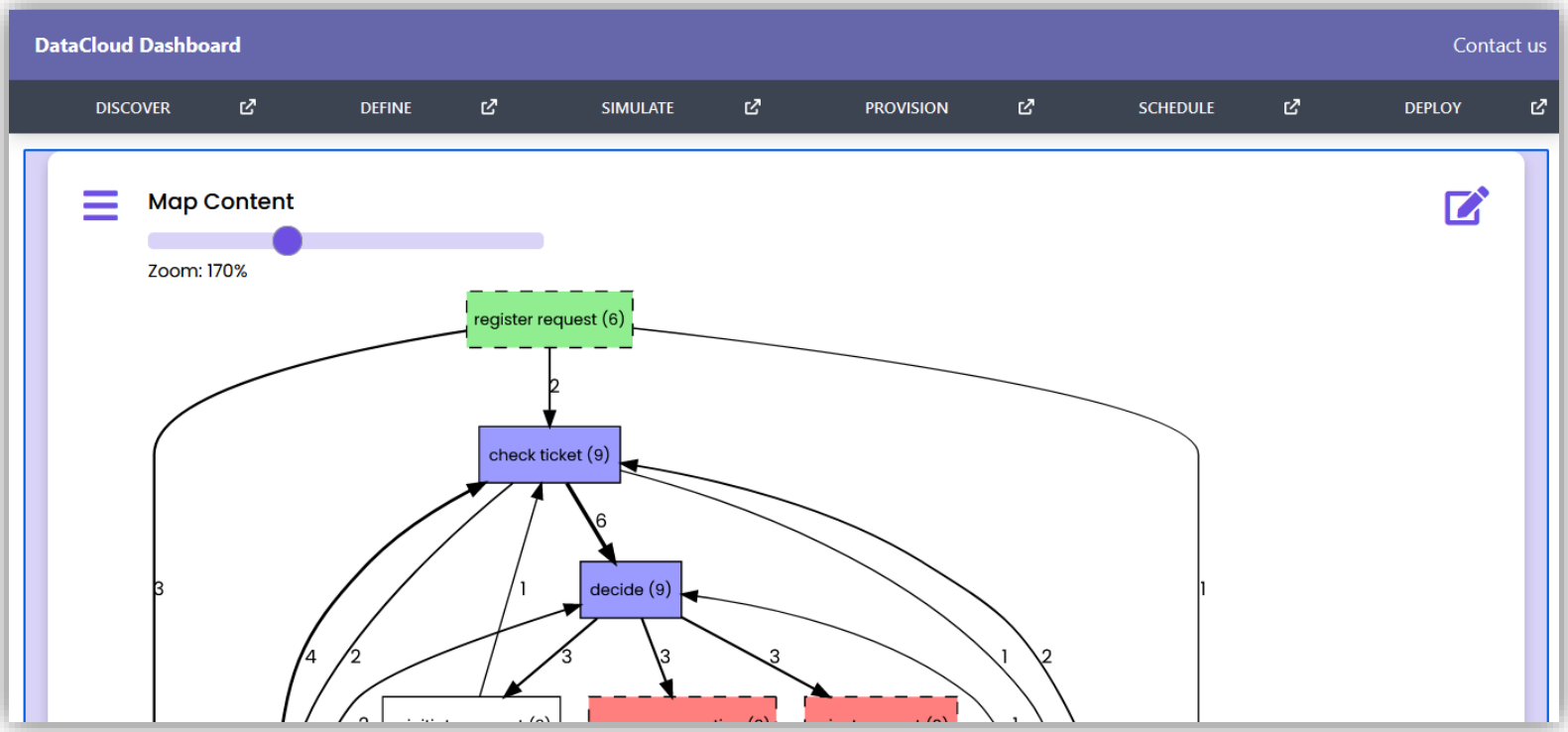
- Display available resources and providers
- Perform deployment of pipeline steps

7 **Pipeline execution**

- Sensor data simulation
- Notifications

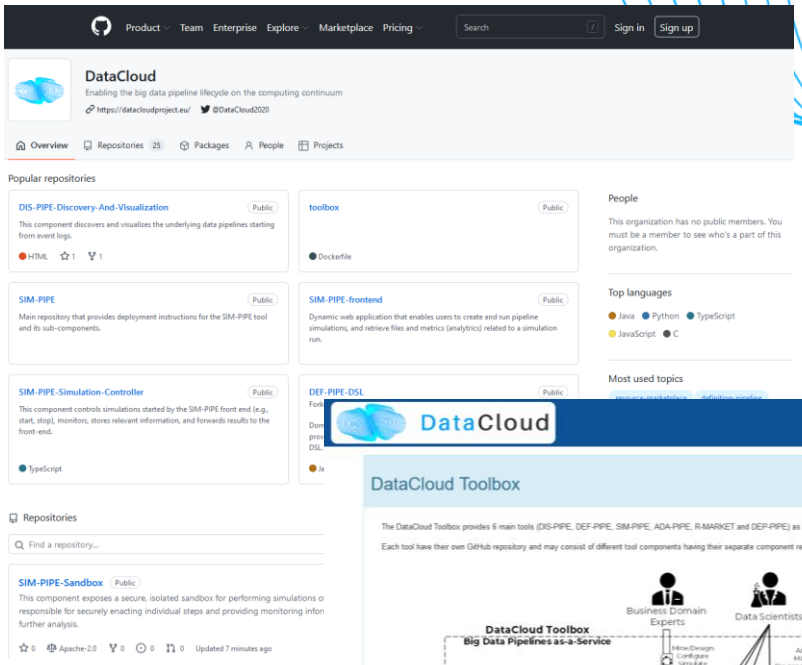
N. Nikolov et al., *Container-based Data Pipelines on the Computing Continuum for Remote Patient Monitoring*, **Computer**, Special Issue on Computing in Telemedicine2023, to appear.

DataCloud Toolbox Common Dashboard



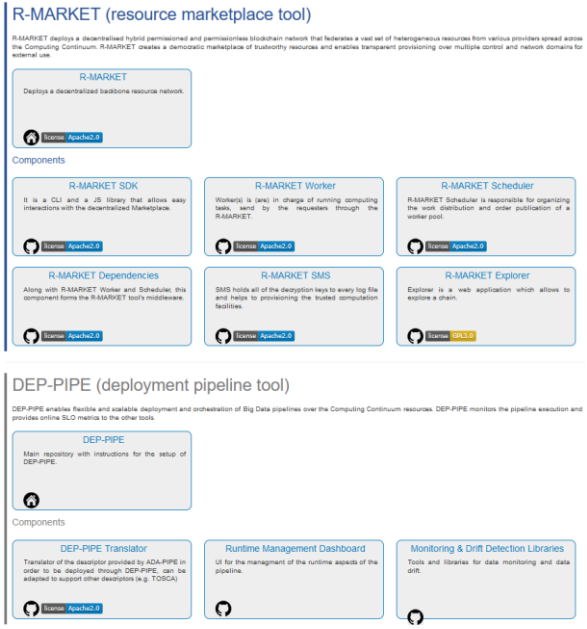
DataCloud Toolbox on GitHub

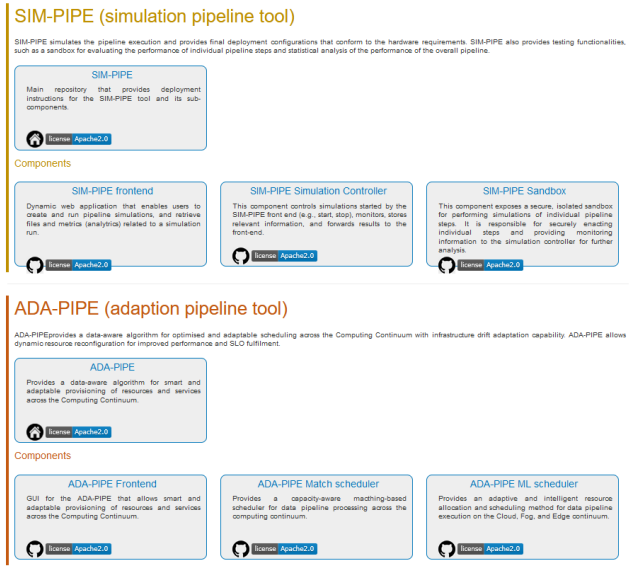
- DataCloud repository overview
  - https://github.com/DataCloud-project
  - DataCloud tool repositories
    - Documentation
    - License information
    - Source code
- DataCloud Toolbox overview
  - https://datacloud-project.github.io/toolbox
  - Section for each DataCloud tool
    - Main tool and components
    - Link to GitHub repositories
    - License icon



10

DataCloud Toolbox GitHub Overview

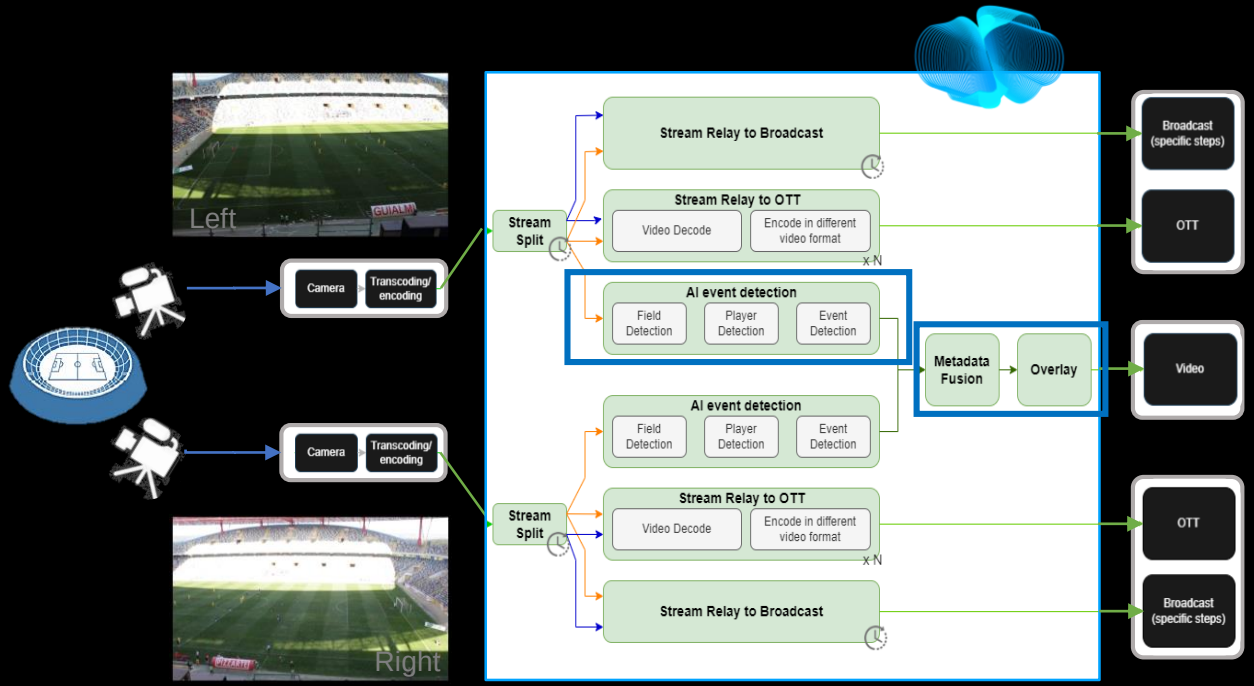




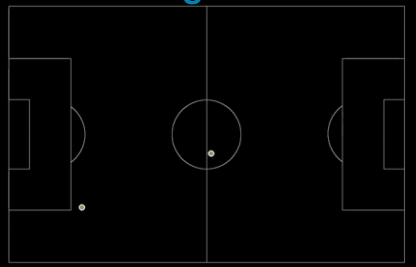
11

MOGSPORTS - Automatic Live Sports Content Annotation

of large quantities of media data

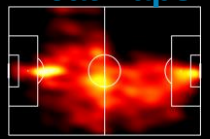


Team Segmentation



Player Tracking

Heatmaps



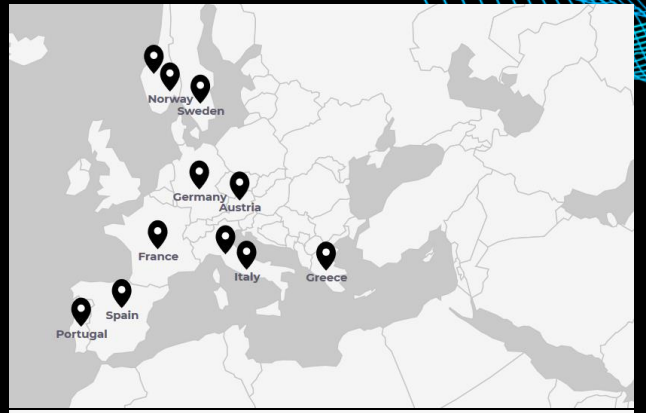
Overlays

- Lower production costs
- Enhance audience engagement

DataCloud project

- DataCloud – Enabling the Big Data Pipeline Lifecycle on the Computing Continuum
- H2020-ICT-40-2020 Cloud Computing
- Research and Innovation Action, 2021-2023
- https://datacloudproject.eu

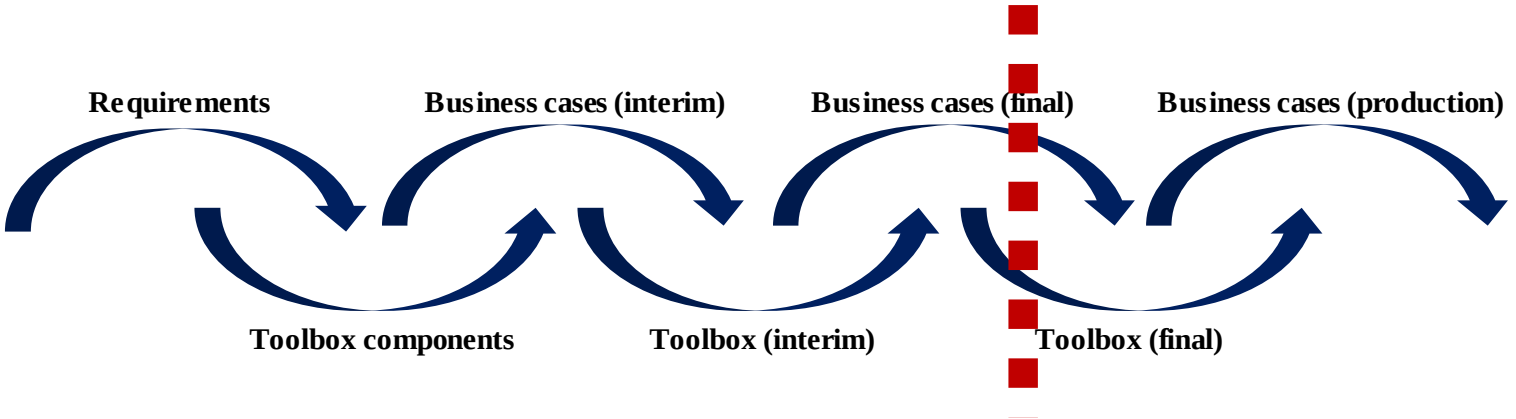
- Goals:
  - Software toolbox to support Big Data pipeline lifecycle on the Computing Continuum
  - Validation in five business cases (digital marketing, live media streaming, electronic healthcare, predictive maintenance, Industry 4.0)

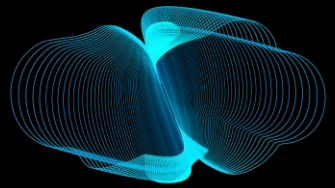


SINTEF · Sapienza Università di Roma · Universitat Klagenfurt · Kungl Tekniska Högskolan · iExec · UBITECH · JOT Internet Media · MOG Digital Media · Catalano · tell.u · Bosch

Current status

- Software for all technical components and their first integration in the toolbox
- Business cases prototypes
- Validation of the toolbox components in the business cases
- Dissemination (papers, etc.)





THANK YOU!

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016835, the DataCloud Project.

https://datacloudproject.eu/