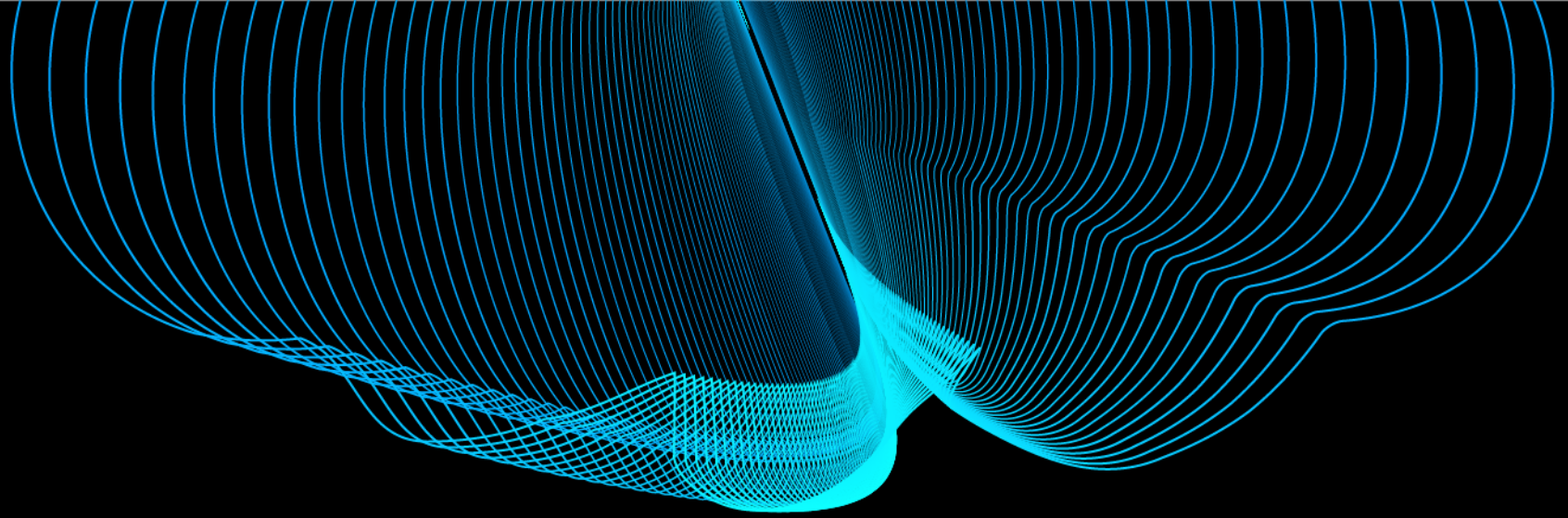


DataCloud

Intro and Highlights

Radu Prodan

University of Klagenfurt, Austria

EU Day @ CC Grid

Taormina, Italy

Outline

- Brief intro to the project
- Concept and Scenario
- Architecture
- Summary and outlook

ENABLING THE BIG DATA PIPELINE LIFECYCLE ON THE COMPUTING CONTINUUM

Fact Sheet | Results

Project description

DE | EN | ES | FR | IT | PL

Big Data for a lifetime

Nowadays, the increasing pervasiveness of data and computing results in the proliferation of edge applications for timely and effective processing of data and advanced analytics. However, as the available data grows, new solutions are needed to ensure a fluid integration of resources to support dynamic, data-driven application workflows. In that direction, the EU-funded DataCloud project introduces a groundbreaking paradigm with a complete life cycle managing Big Data pipelines through discovery, design, simulation, provisioning, deployment and adaptation across the computing continuum. It will allow Big Data pipelines to interconnect the end-to-end industrial operations from the preprocessing and collecting of data to the realisation of a business target. DataCloud will make Big Data advancements more accessible regardless of hardware.

Show the project objective

Fields of science

natural sciences › computer and information sciences › data science › **big data**

Programme(s)

H2020-EU.2.1.1. - INDUSTRIAL LEADERSHIP - Leadership in enabling and industrial technologies - Information and Communication Technologies (ICT)

Topic(s)

ICT-40-2020 - Cloud Computing: towards a smart cloud computing continuum

Call for proposal

H2020-ICT-2020-2

Project Information

DataCloud

Grant agreement ID: 101016835

Start date
1 January 2021

End date
31 December 2023

Funded under
H2020-EU.2.1.1.

Overall budget
€ 4 999 996,25

EU contribution
€ 4 999 996,25

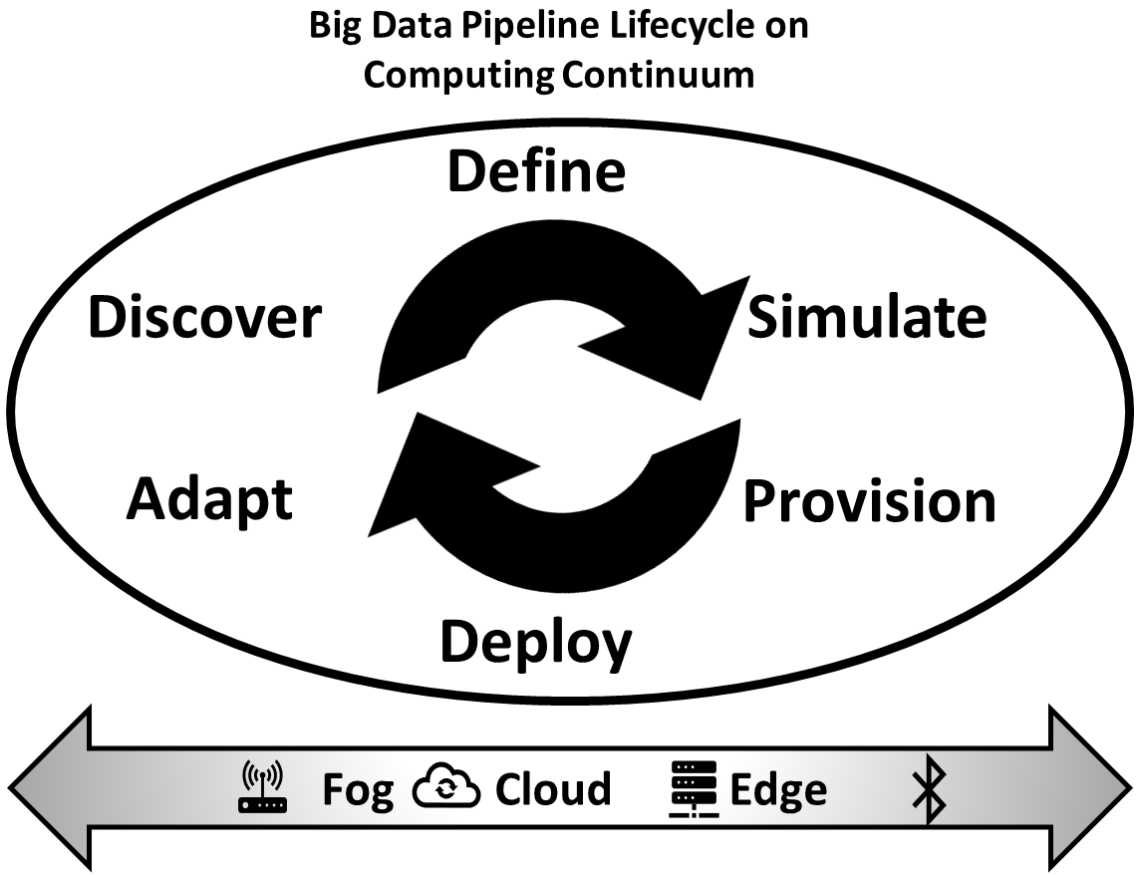
Coordinated by
SINTEF AS
Norway

DataCloud Scope

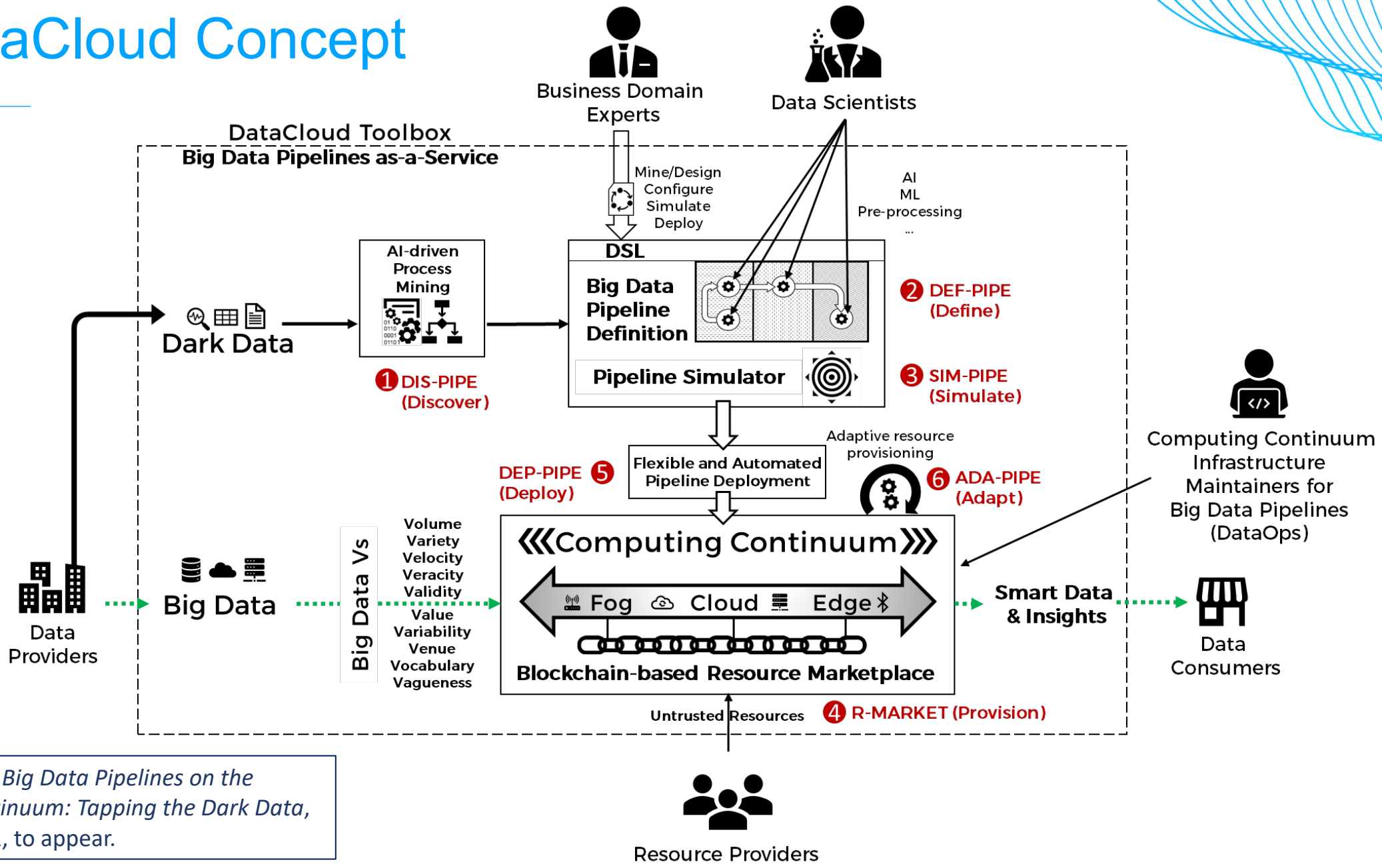
Big Data pipelines are composite pipelines for processing data with non-trivial properties and characteristics, commonly referred to as the *Vs of Big Data*.

Example: 1TB database with sensor measurements

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

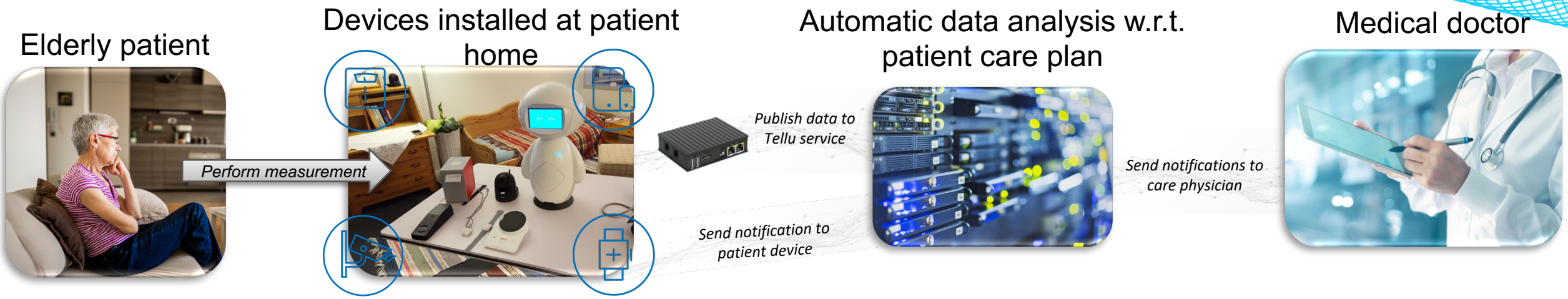


DataCloud Concept



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

Example eHealth scenario: remote monitoring of elderly patients



- Existing application 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 (need to build application version for each new type of sensor)
  - Difficult to find the optimal resource allocation for data processing steps

Use of DataCloud tools in the eHealth scenario

Task #1: Designing a data pipeline
- Configured pipeline description
- Validated pipeline
- Loosely-coupled steps and step-level scalability
- Containerization and programming language independence

Task #2: Distributed deployment of the data pipeline
- Provisioning a set of heterogenous Edge/Cloud resources
- Automated scheduling and deployment
- Runtime execution and monitoring of 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

Task#1 Completed

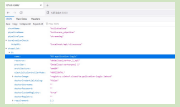
Configured pipeline model and run stats

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

- Display information about resources
- Add a new resource 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**

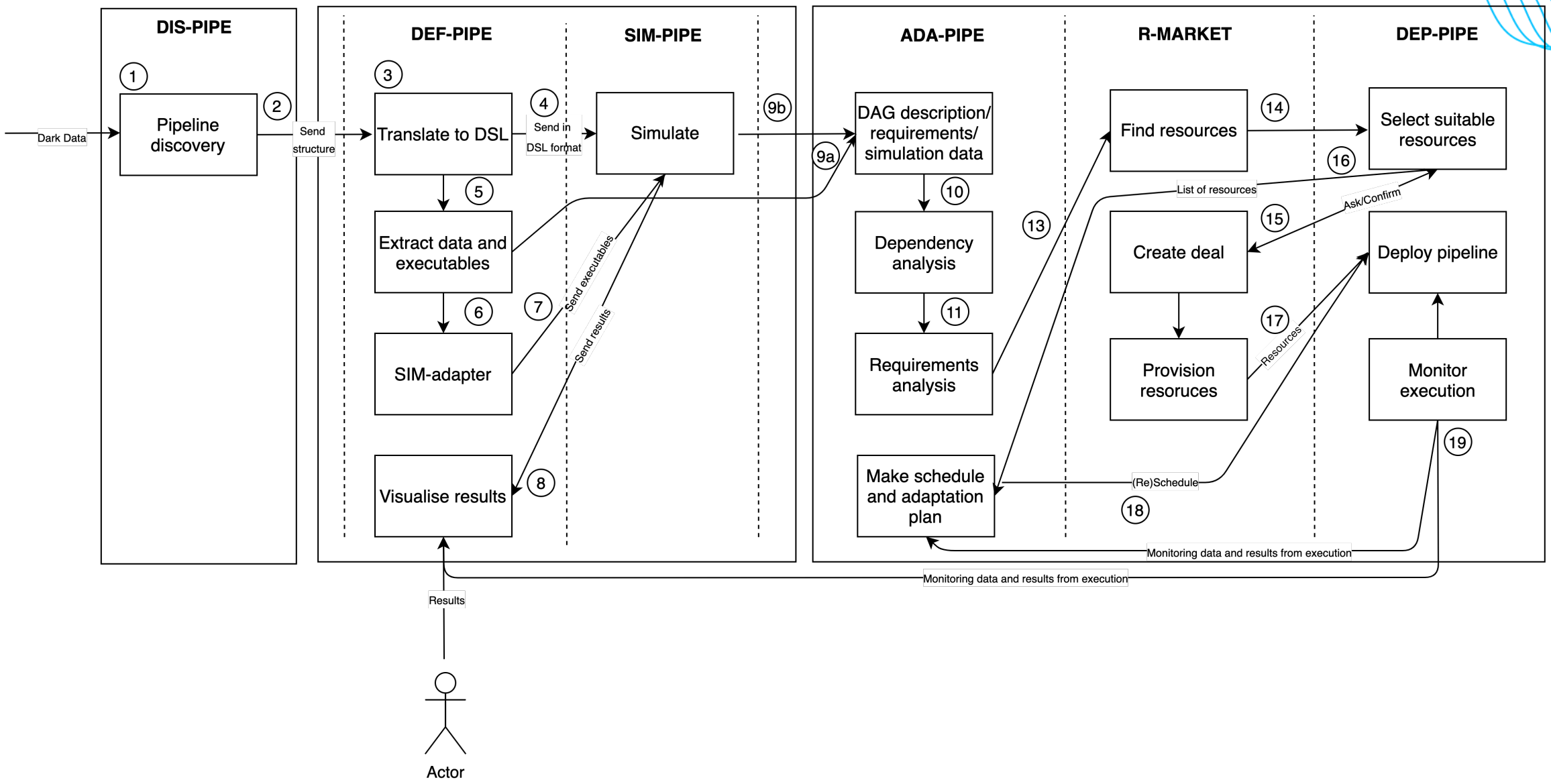
- Display available resources and providers
- Perform deployment of pipeline steps

7 **Pipeline execution**

Task #2 Completed

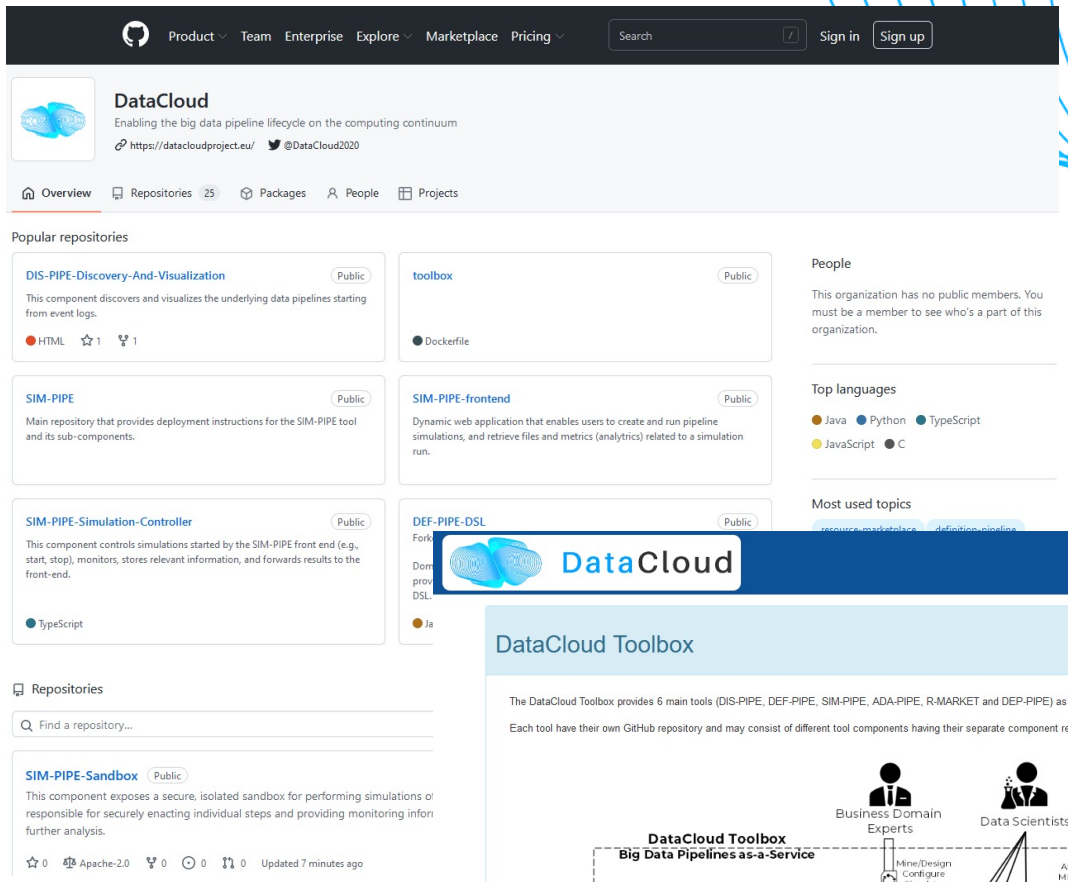
- Sensor data simulation
- Notifications

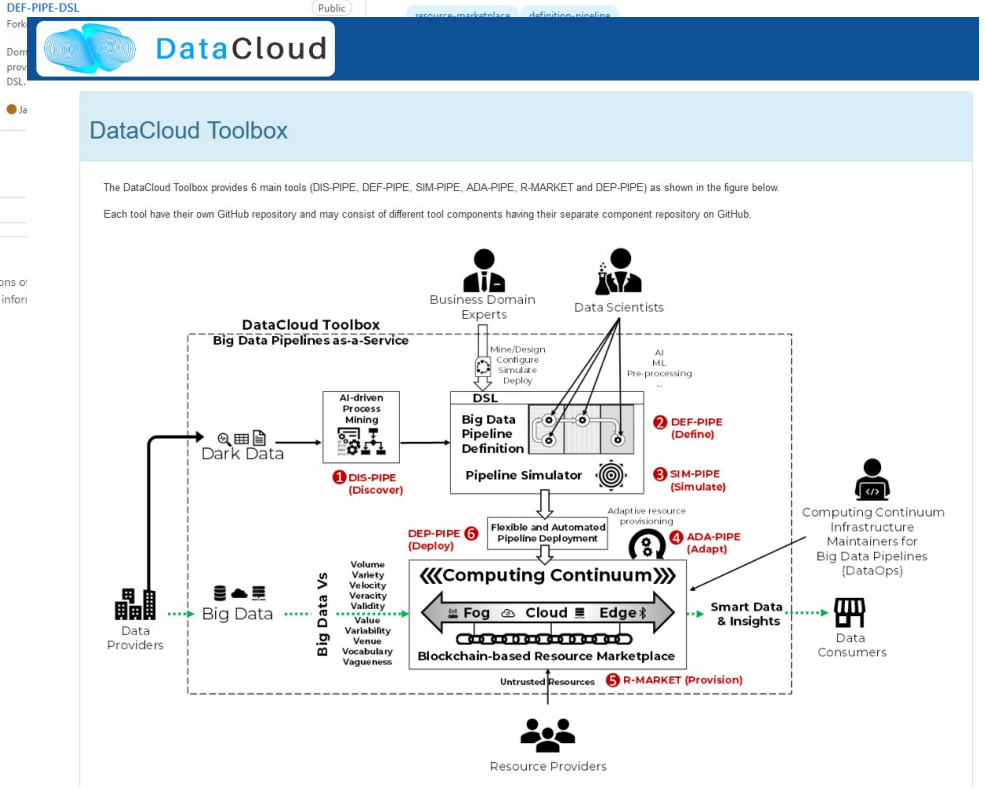
DataCloud Tool Interaction Workflow



DataCloud Toolbox on GitHub

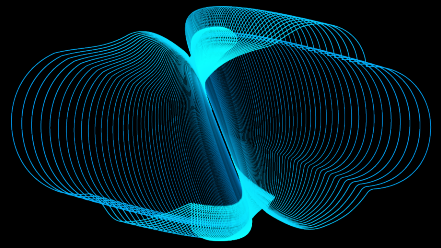
- DataCloud repository overview
  - https://github.com/DataCloud-project
  - DataCloud tool repositories
    - Documentation
    - License information
    - Source code
- DataCloud Toolbox overview
  - https://datacloud-project.github.io/toolbox
  - Section for each DataCloud tool
    - Main tool and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





Summary first year of the project and outlook

- Thorough state of the art review
- Detailed requirements specification for the toolbox and business cases, incl specific pipelines
- Architecture specification
- Software for all toolbox components
- Future work
  - Full implementation of tools integration using APIs
  - DataCloud toolbox as-a-Service
    - Integrated user-based access to toolkit assets; Security, authentication, authorization; Improved user experience for the common UI
  - Continuous maintenance and updates to tools as they become more mature
  - Application in project business cases



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/