LONDON METROPOLITAN UNIVERSITY

Data Platforms, Clouds and Spaces: Trends in Contemporary Data Processing

Prof. Vassil Vassilev & Team

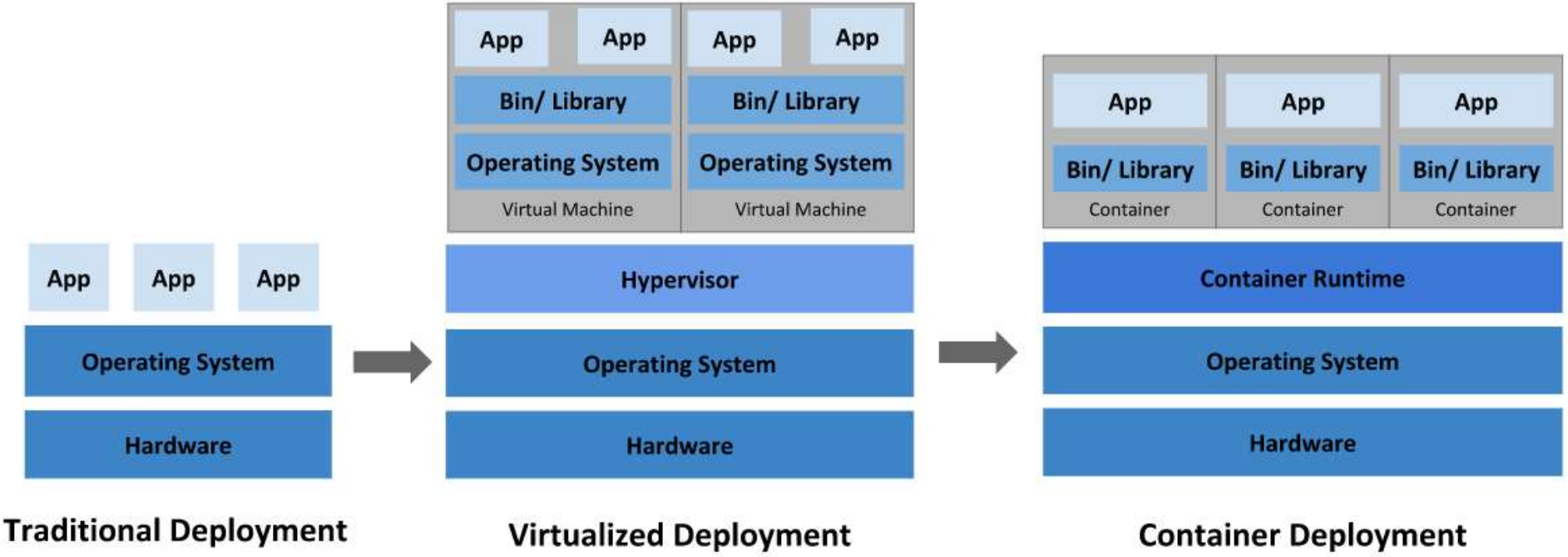
Cyber Security Research Centre

SCDM, 28 Sep 2022

Content

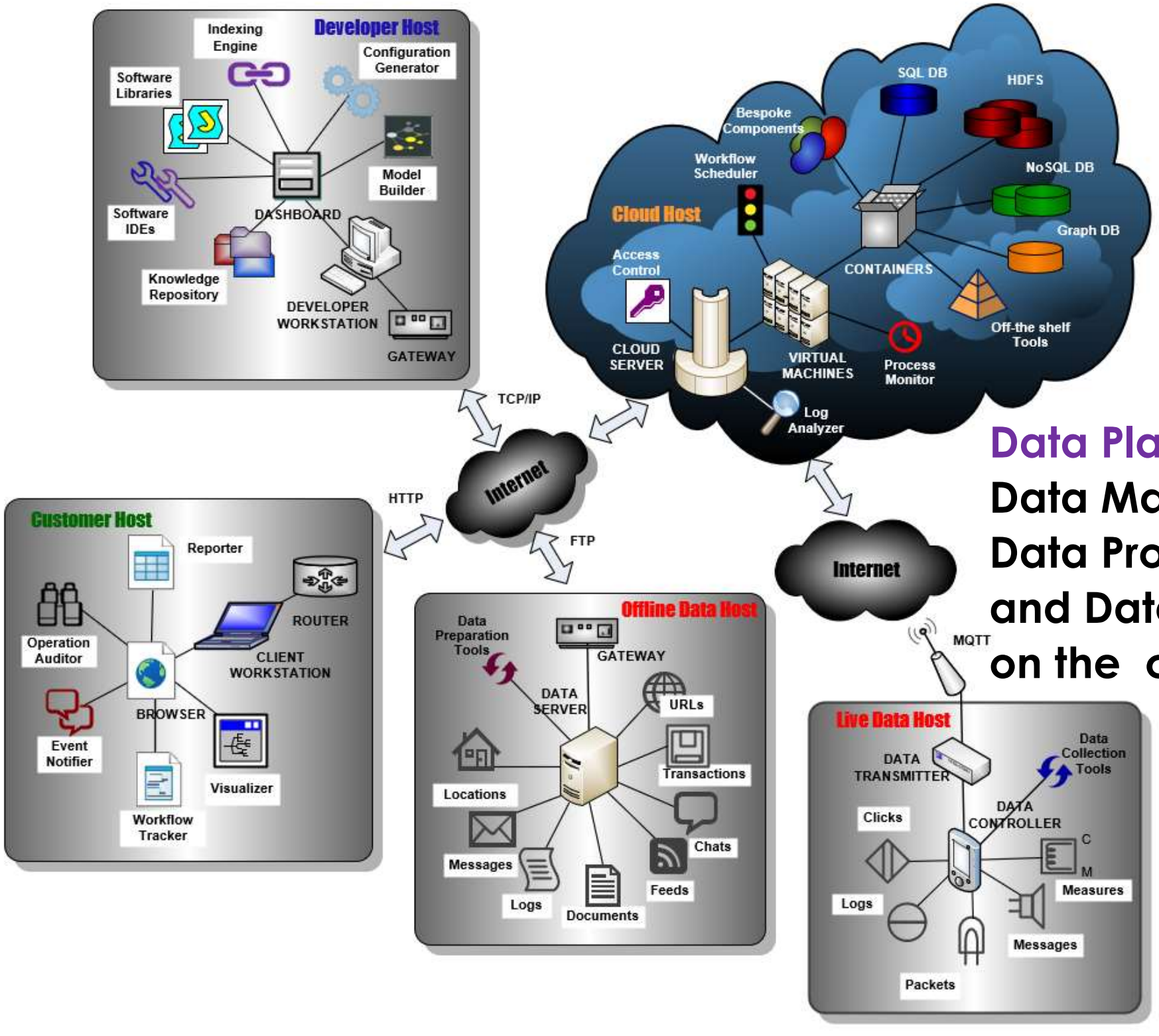
1. Contemporary Data Processing
2. Data Platform of Cyber Security Research Centre
3. Alternatives and Choices
   - The Data
   - The Metadata
   - The Technologies
   - The Tools
4. The Lifecycle
   - Design
   - Development
   - Deployment
   - Orchestration
   - Monitoring
   - Auditing
5. Where to go from here?

1 Contemporary Data Processing:
From the Desktop to the Cloud



Source: https://kubernetes.io/docs/concepts/overview/what-is-kubernetes/

- ✓ Hardware infrastructure
- ✓ Operating system
- ✓ Middleware software
- ✓ Execution environment



Data Platforms:

Data Management, Data Processing and Data Analysis on the cloud

Data Spaces: Decentralized Supply and Consumption of Data Services

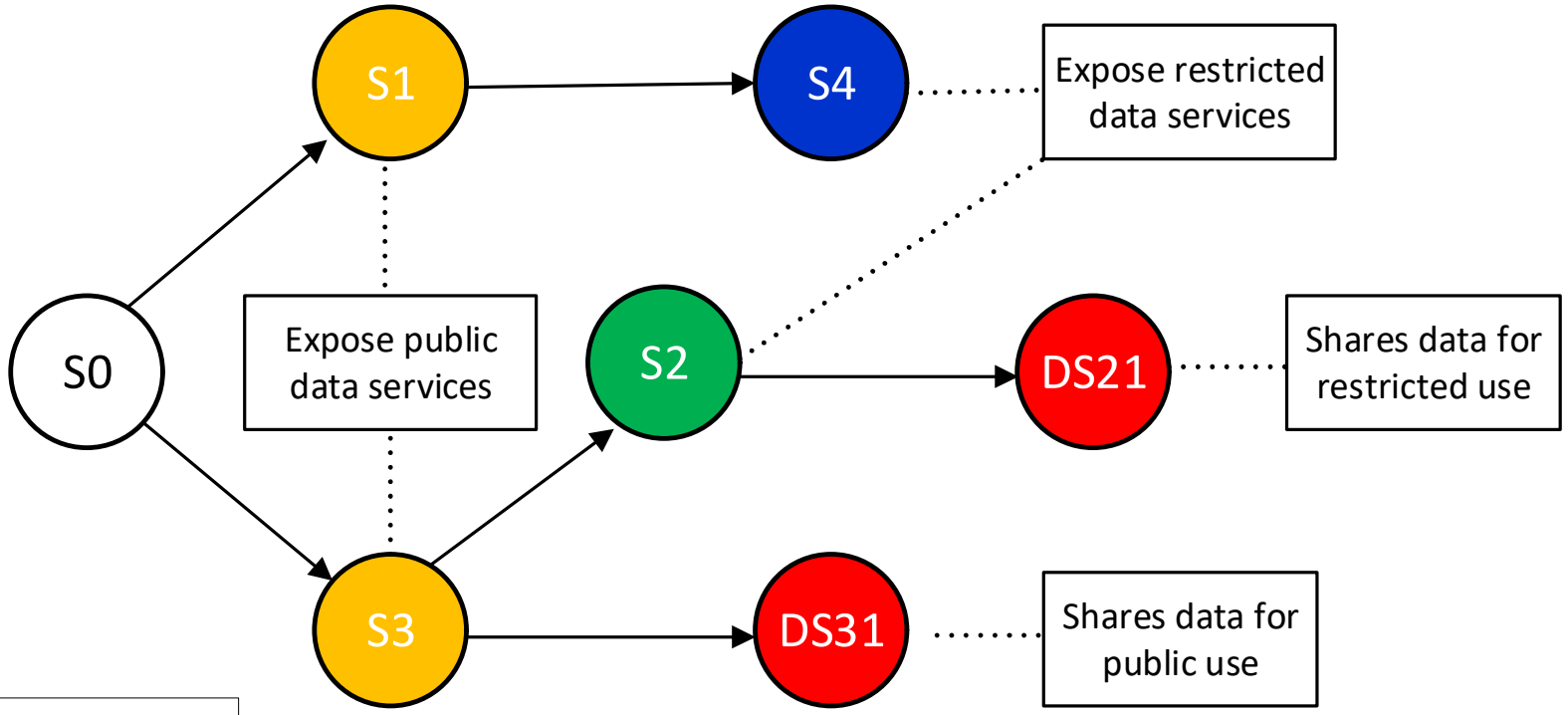
- Initiative of Fraunhofer Institute in Germany to serve the needs of Germany's industry, heavily based on B2B
  - Addresses the need for information from external sources in digital format over the Internet (***accessibility***)
  - Does not require the data to be available at the place of its use (***distribution***)
  - Retain the ownership while sharing the data and services (***control***)
  - Requires only two participants: provider (***sharing*** the data it owns and/or ***exposing*** some data services) and consumer (***consuming*** the shared data and/or ***utilizing*** the exposed data service for its own use); they can be combined in the same service-oriented architecture (client/server+server/server+peer-to-peer)

Adopted across EU (International Data Space Association, IDSA)

Example: Urban Mobility Data Space

	City Mobility Centre	Environment Control Agency	Urban Planning Department
Data Shared	routes, places, vehicles, times	pollutions, standards, polluters	
Data Access Permissions	public (routes, places, times), restricted (vehicles, locations)	public (pollutions, standards), restricted (polluters)	
Operations Supported	locating, placing, timing	pollution, polluter determination	
Operations Rights	public (placing, timing), restricted (locating)	public (pollution), restricted (polluter)	
Data Consumed		routes, places, vehicles	vehicles, places, routes, pollutions, polluters
Operations Executed		placing, locating	place pollutions, vehicle pollutions, route pollutions

Data Space as Service-oriented Architecture



LEGEND

- S0 — **Data Service Consumer**
- S1 — **Data Service Consumer & Provider**
- S2 — **Data Service Consumer & Provider**
- S4 — **Data Service Provider**
- DS21 — **Data Source**

Data Platform support needed for Data Spaces

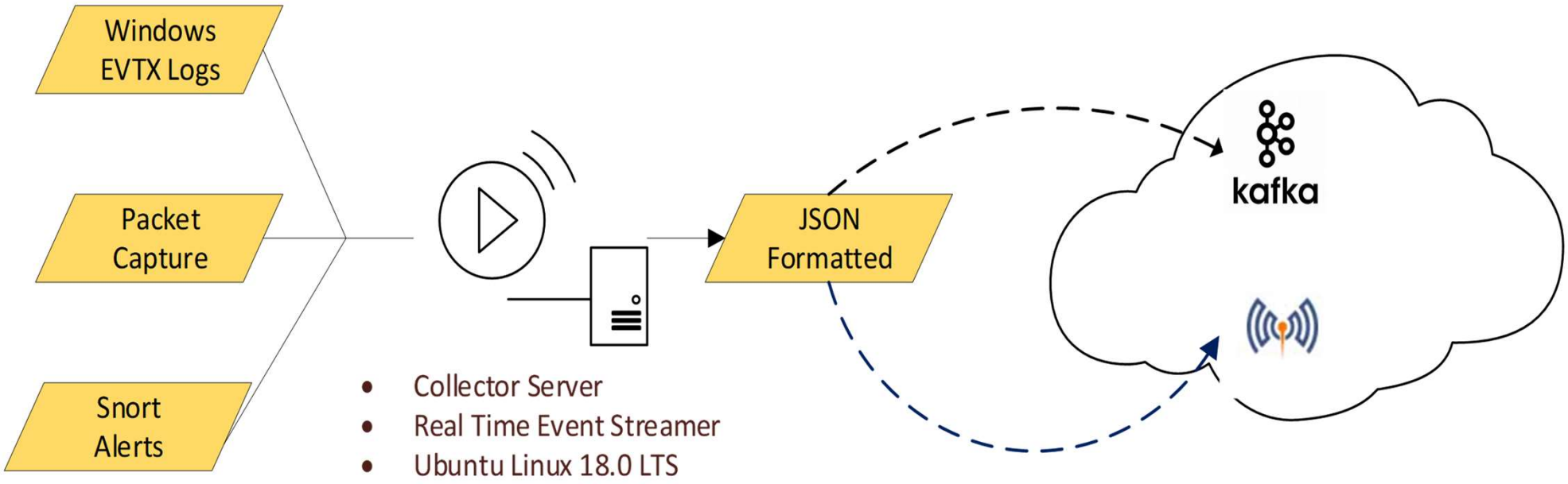
- ✓ **Data service consumption**
  - » Identification of the services (addressing)
  - » Identification of the consumers (authenticating)
  - » Requesting the services (consuming)
  - » Requesting the consumption (reporting)
- ✓ **Data service provision**
  - « Registration of providers and consumers (identity management)
  - « Assignment of responsibilities for data sharing and service provisioning to providers (provider profiling)
  - « Assignment of rights for data access and privileges for operation execution to consumers (consumer profiling)
  - « Identification of the communications (session tracking)
  - « Logging of the operations (event logging)
  - « Estimation of the consumptions (service reporting)
  - « Calculation of the cost of consumption (billing)

Data platforms certification under GAIA-X initiative of IDSA

2 Data Platform of Cyber Security Research Centre

- Piloted at the **Cyber Security Research Centre** of London Metropolitan University with funding from DCMS and HEIF
  - Private **cloud**, based on commodity server architecture
  - Public domain software for **virtualization, containerization** and **orchestration** of the server-side applications
  - Communal editions of enterprise software products for **data management** on the server
  - Free software for **application development**
  - Web-based interfaces for **service deployment** and **use**
- Replicated by the partner organization, **GATE Institute** of Sofia University
- Tested in three different scenarios for data processing in real-time: security analytics, outdoor air pollution in London and indoor air pollution factor analysis

Project 1: Real-time Security Data Analytics



Classification of Network Packets using Regression, NN and SVM

Model	Predicted regular packets:	Regular packets in test set:	Predicted ACK packets:	ACK packets in test set:	Predicted SYN packets:	SYN packets in test set:	Accuracy:
Neural Network	129	303	2023	1851	8	6	79%
Support Vector Machine	107	276	2050	1877	3	7	90%
Logistic Regression	417	258	1743	1893	0	9	71%
Linear Regression	791	255	1369	1897	0	8	60%

Legend

ACK - acknowledgement flag confirming normal exchange of packets between two sites

SYN - synchronization flag signaling initiation of normal communication between two sides

Detection of Potential Unauthorized Intrusions using CNN

Actual flags within the dataset		Predicted Suspicious Flags	
ACK packets in test set:	39913	Predicted ACK packets:	43647
SYN packets in test set:	13	Predicted SYN packets:	12
RST packets in test set:	165	Predicted RST packets:	164

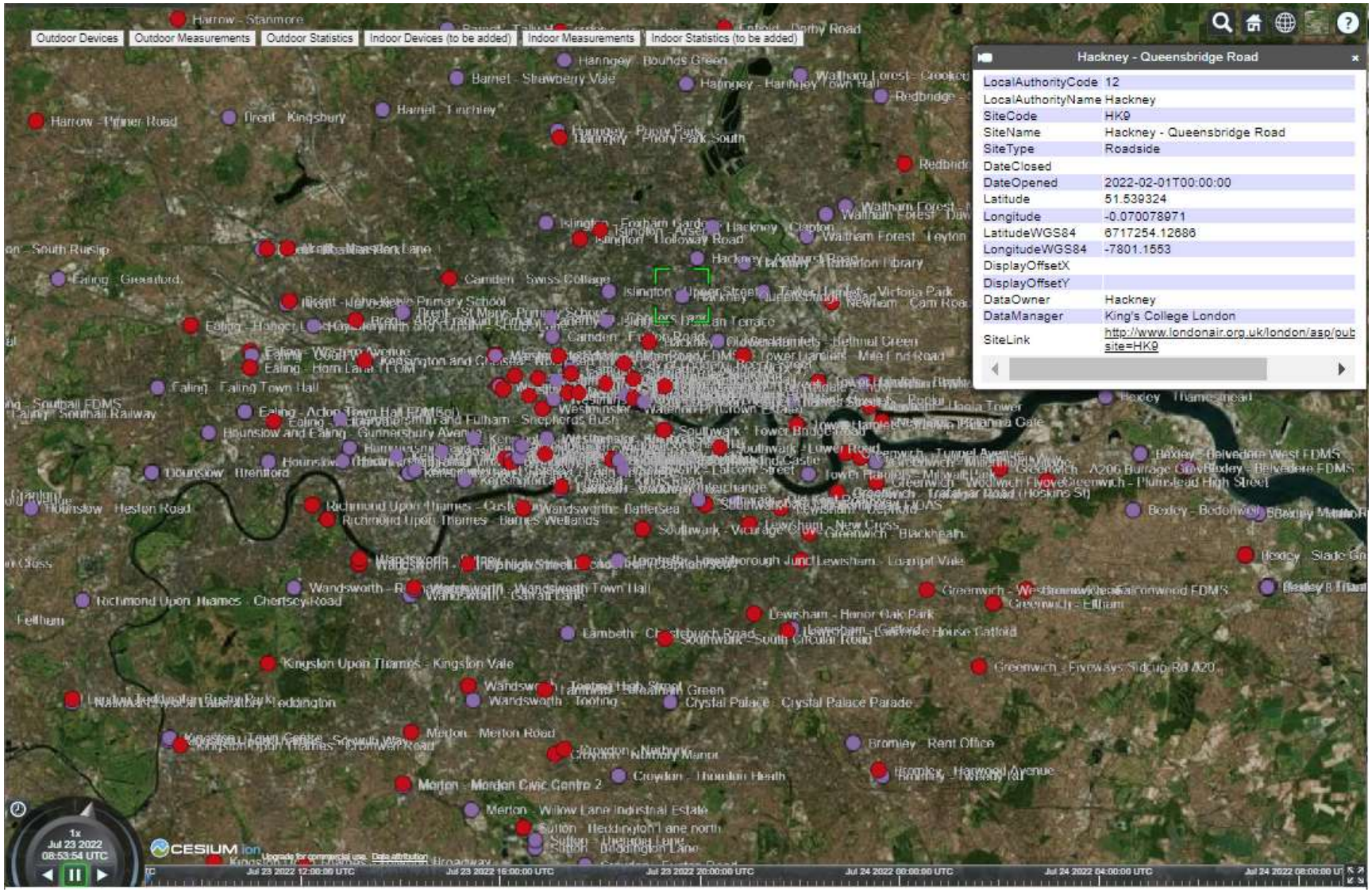
Legend

ACK - acknowledgement flag confirming normal exchange of packets between two sites

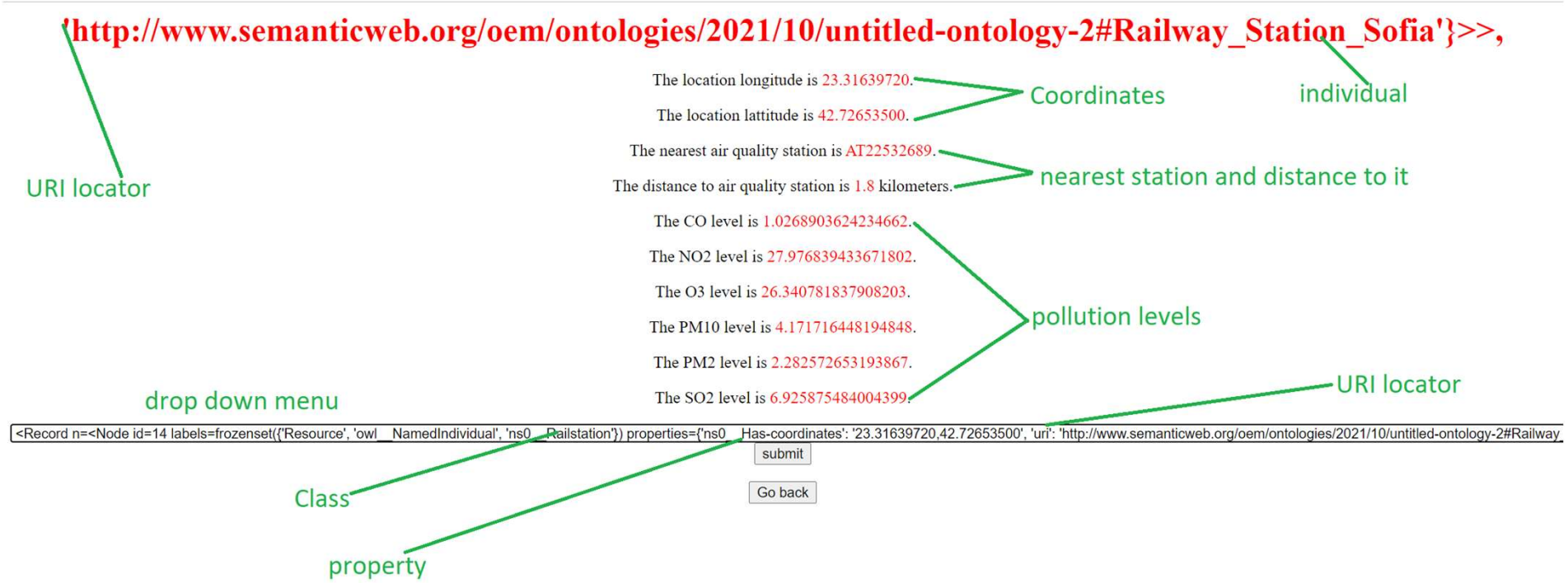
SYN - synchronization flag signaling initiation of normal communication between two sides

RST - warning flag sent after anomaly has been detected in the previous communication

Project 2: Outdoor Air Pollution in London

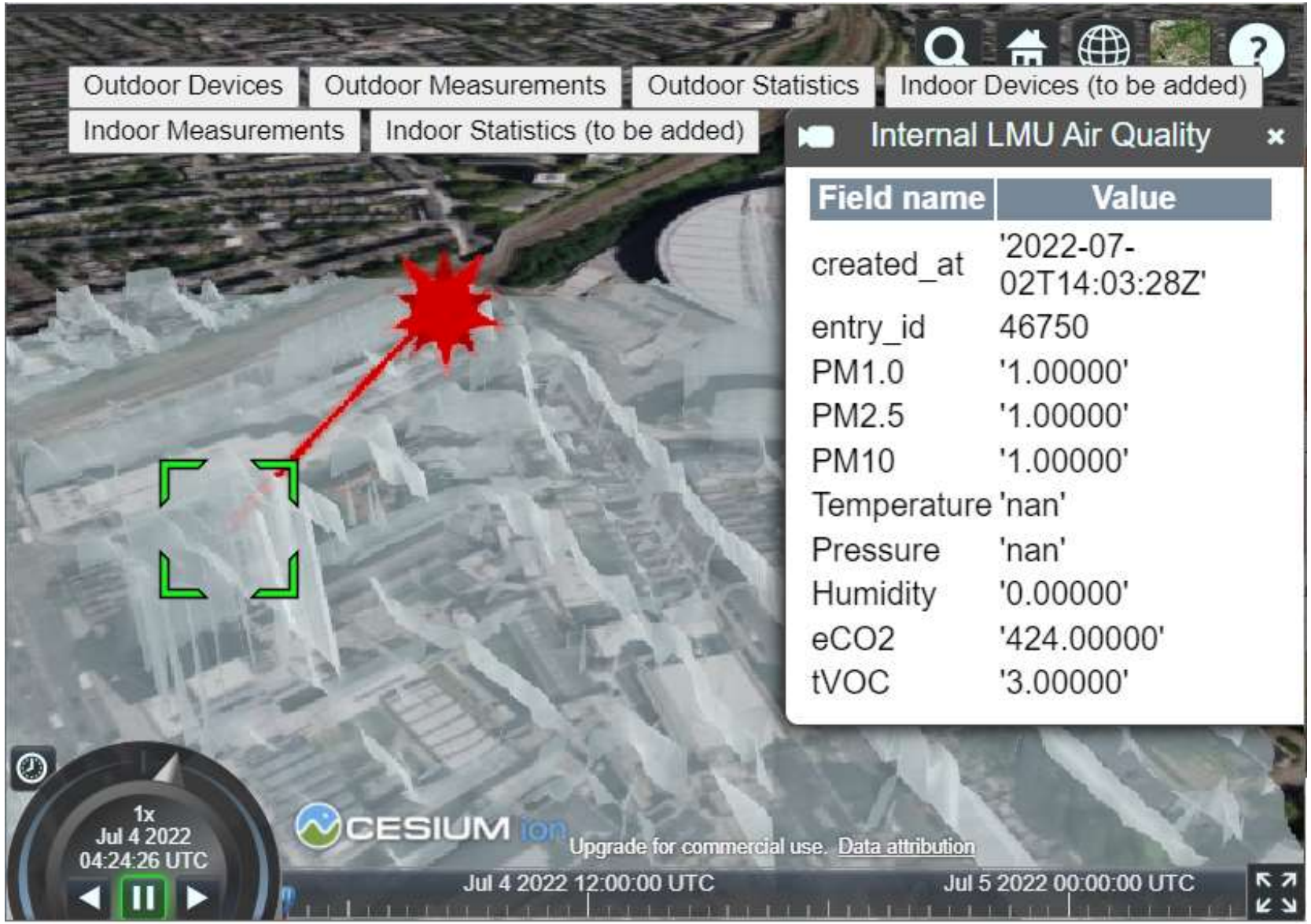


Logical Analysis of Air Pollution by combining Ontological and Sensor Data

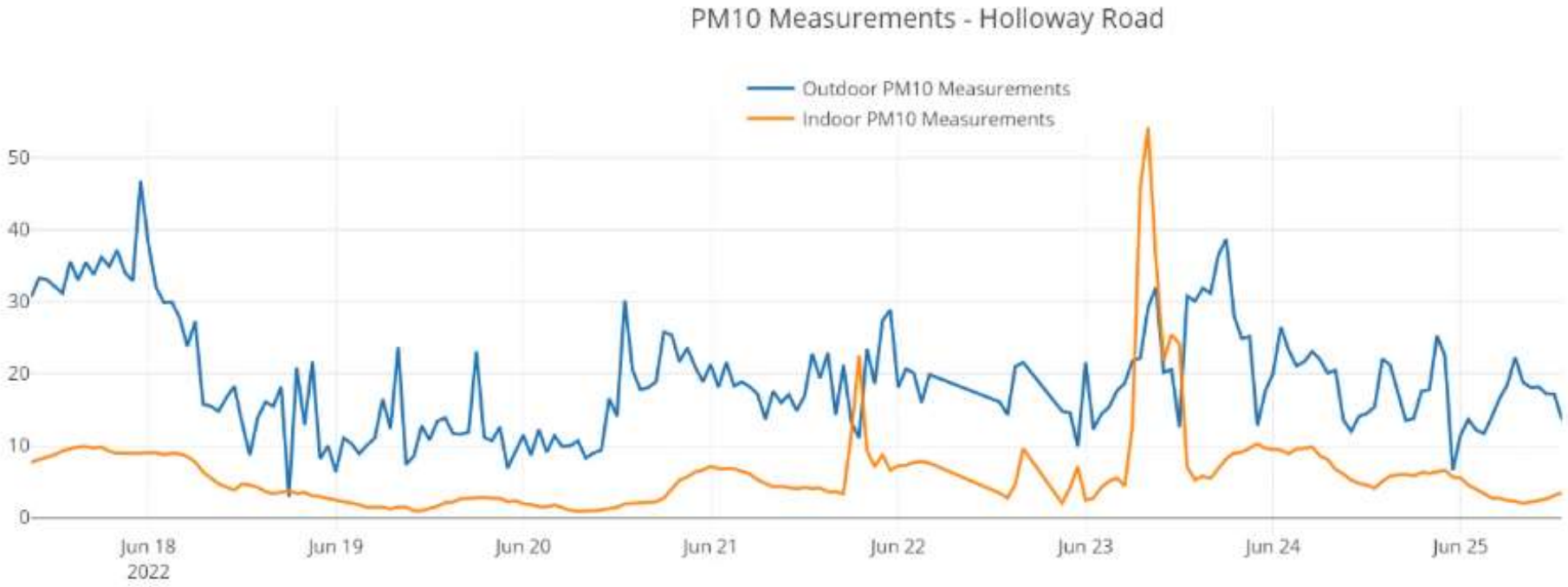


- Identifying location of the sensor station meta-data stored in **MongoDB** Database
- Loading individual descriptions of the objects from the city ontology stored in **Neo4J** Database
- Calculating the distance between air quality station and the objects in **Cypher** query language
- Analysing the air quality measurements at this location in **Python**

Project 3: Indoor Air Pollution Analysis



Correlation between Outdoor and Indoor levels of Particles in the Air



Correlation Matrix

Indoor Value	Outdoor Value
1	0.3634504520543431
0.3634504520543431	1

3 Alternatives and Choices: Data, Metadata, Technologies and Tools

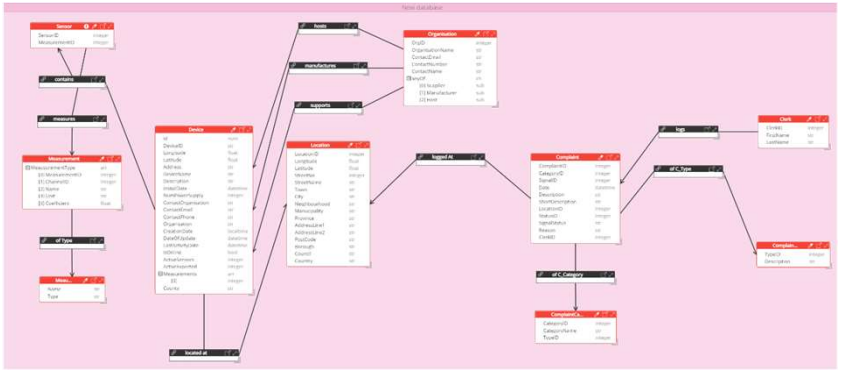
- The data is on **multiple scales** – different formats, granularity, volume, noise, location…
- The tasks for data analysis have **large diversity** – detection, recognition, classification, correlation, factorisation, prediction…
- There is a **variety of methods** with different applicability – temporal, structural, logical, model-driven, behavioural, hybrid
- Data analysis is performed as part of **complex workflows** – sampling, aggregation, buffering, feature selection, training, validation, analysis, merging, interpretation, explanation…
- The applications may require **significant resources** (both in terms of memory and computing power).
- AI technologies for data processing need to be comprehensive to reach **wide community** of users.

All about Data

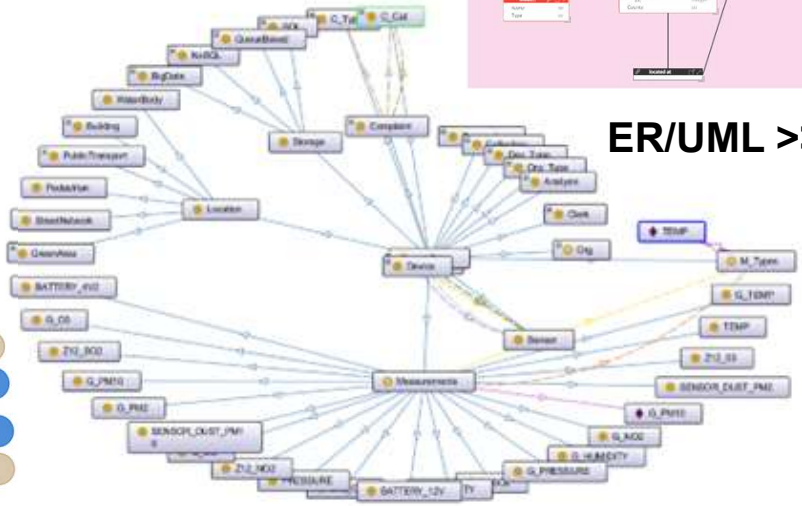
Data Types	Data Sources	Ingestion Methods	Transport Protocols
Samples	**Hardware** (external devices, infrastructure)	One-off	memory sharing, parameter passing
Files	**OS** (clients, suppliers)	One-off, Batch	FTP, HTTP, SCP, WebDAV, etc.
Messages	**Events** (Messengers, Listeners, Loggers)	One-off, Batch, Continuous	MQTT, AMQP, SMS/MMS, RCS, SOAP, etc.
Repository Collections	**Drivers** (databases, data warehouses, data lakes)	One-off, Batch	native to the repository
Streams	**APIs** (sensors, service providers)	Continuous	native to the streaming

The Metadata and its Utilization

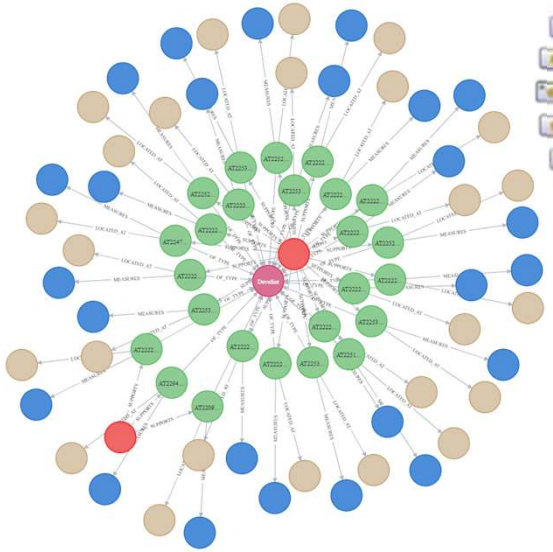
- ✓ Understand the data for **better design** of applications
- ✓ Enrich semantically the data for **more informative** and **more convenient handling**
- ✓ Prepare the data for **further** storage and processing



ER/UML >>> SQL DB, ER files



RDF/XML >>> Graph DB, XML files



JSON >>> NoSQL DB, JSON files

Technologies for Data Processing

Different Stages along the Data Processing Pipeline: At the source, Before transmission, During transmission, On arrival, Within repository, After retrieval, etc.

Different Structure and Formats of the Data: structured (CSV, SQL), semi-structured (JSON, XML, RDF, SVG, etc.) and unstructured (binary, text, graphics, video)

Different Preparation of the Rough Data: Filtering, Formatting, Anonymisation, Normalization, Enrichment, Aggregation, Reconciliation, Buffering, Accumulation, etc.

Different Methods for Data Analysis: Statistics, Regression, Correlation, Clustering, Graph based, Rule based, Neural, Genetic, Swarm, Deep Learning, Reinforcement Learning, etc.

Different Interpretations of the Results: Simple reporting, Blackbox explanation, Whitebox explanation, Causal explanation, Impact factor analysis, etc.

Software for Data Pipelines on the Cloud

Type	Software	Context
Virtual Machine	VMWare Workstation, Oracle VM, KVM Windows 10, MS Linux, etc.	OS or hypervisor
Hypervisor	VMWare vSphere, Oracle Virtual Box, MS Hyper-V, Linux KVM	OS
Container	Docker, LXC, Windows Containers, Portainer, Podman	OS, VM or container manager
Container Manager	Google Kubernetes, Apache Mesos, Docker Swarm, HashiCorp Nomad	OS
Engine	code interpreter (i.e., Python)	OS, VM or container
Server	off-the-shelf software (i.e., MongoDB)	OS, VM or container
Application	general server-side component (i.e., service registry)	Engine or server, deployed to OS
Service	domain-specific server-side component (i.e., sensor data filter)	Engine or server, deployed to VM
Microservice	application-specific server-side component (i.e., 2D city map)	Engine or server, deployed to container

Advantages of Cloud-based Data Pipelines

Containerization

- ✓ **Modularization** with no dependencies to set
- ✓ **Efficiency** in memory, CPU, and storage usage
- ✓ Application containers are **portable** across platforms without code changes
- ✓ Support for **configuration generation** through the use of parametrization and templates
- ✓ Full **traceability** of the operations for testing and debugging purposes

Orchestration

- ✓ **Model-driven** application development
- ✓ Support for **reusability** of existing solutions in the form of process workflows
- ✓ Support for **auditing** of data processing pipelines for monitoring, analysing and billing purposes
- ✓ Support for **reproducibility** by preserving data dependencies
- ✓ Possibility for process **automation** based on workflow models and planning heuristics

Take away: Always Informed Decisions with Optimal Choices

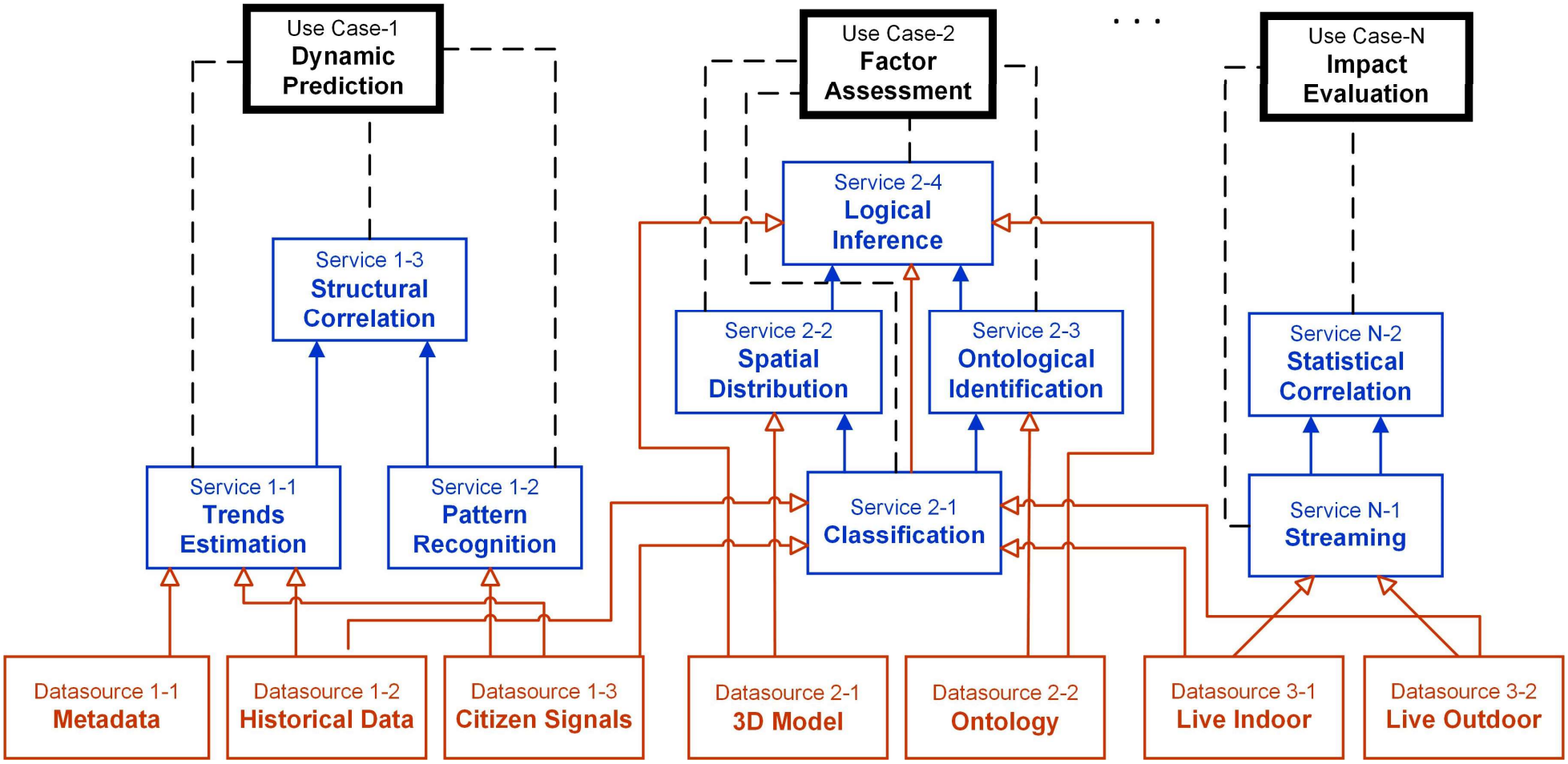
- ✓ **covering the full extent:** generation, pre-processing, transportation, post-processing, analysis, interpretation, reporting, etc.
- ✓ **constructing the richest models:** features, types, structures, constraints, annotations, indexes, etc.
- ✓ **selecting the most appropriate methods:** statistic, clustering, rule-based, graph-based, ML, RL, etc.
- ✓ **using the most convenient tools:** programming languages, engines, libraries, APIs, software tools, development methodologies
- ✓ **placing in the right context:** operating systems, virtual machines, containers, runtime engines
- ✓ **managing with an adequate organisation:** resources, policies, tasks, users

4 Data Services Lifecycle: DataOps in Action

DataOps: *"A set of practices, processes and technologies that combines an integrated and process-oriented perspective on data with automation and methods of agile software engineering to improve quality, speed, and collaboration and promote a culture of continuous improvement in the area of data analytics"*

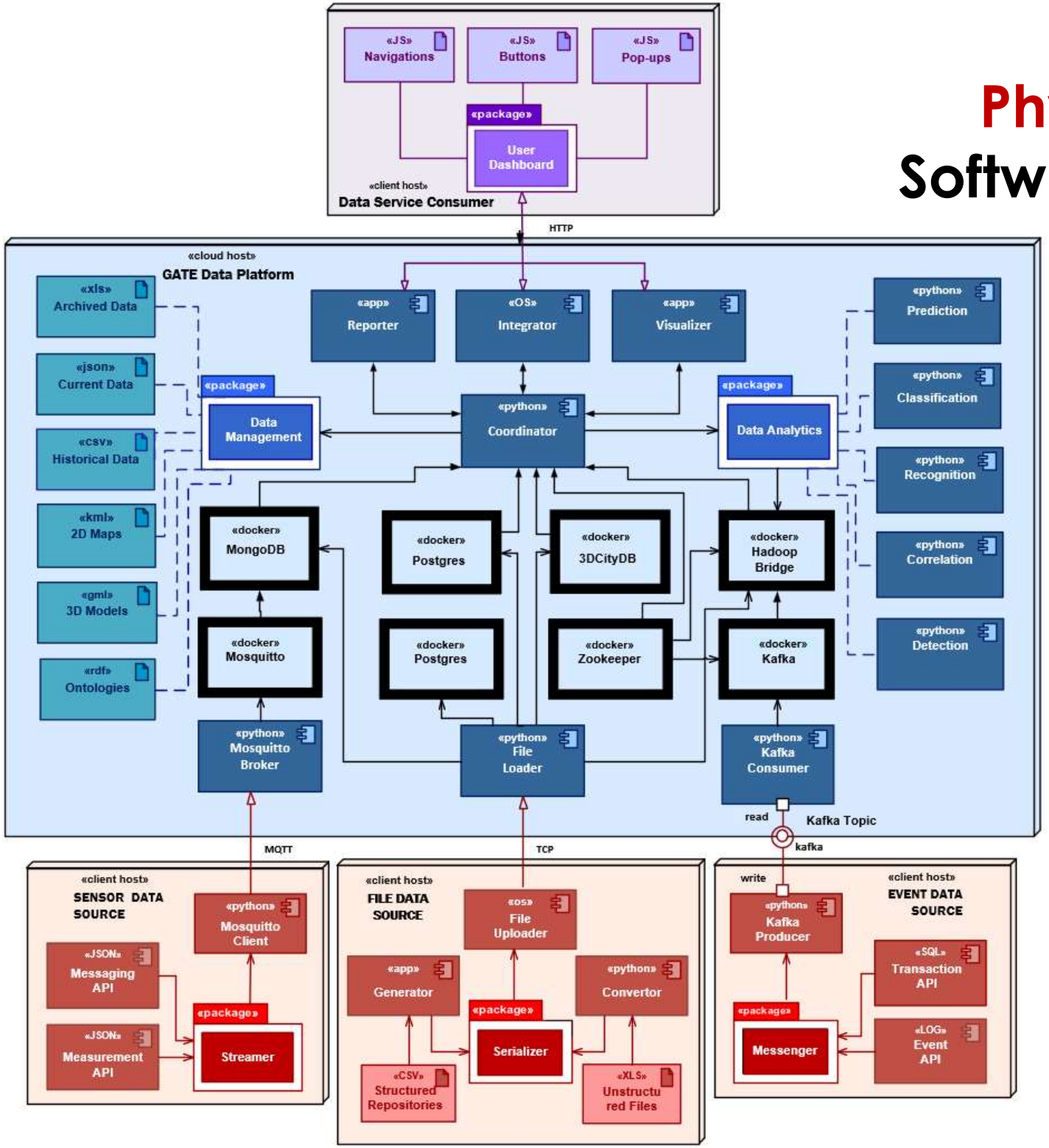
- Logical **specification**
- Physical **design**
- Software **development**
- Application **deployment**
- Service **orchestration**
- Process **monitoring**
- Operations **auditing**

Logical Specification: Data sources, Information Flows & Services



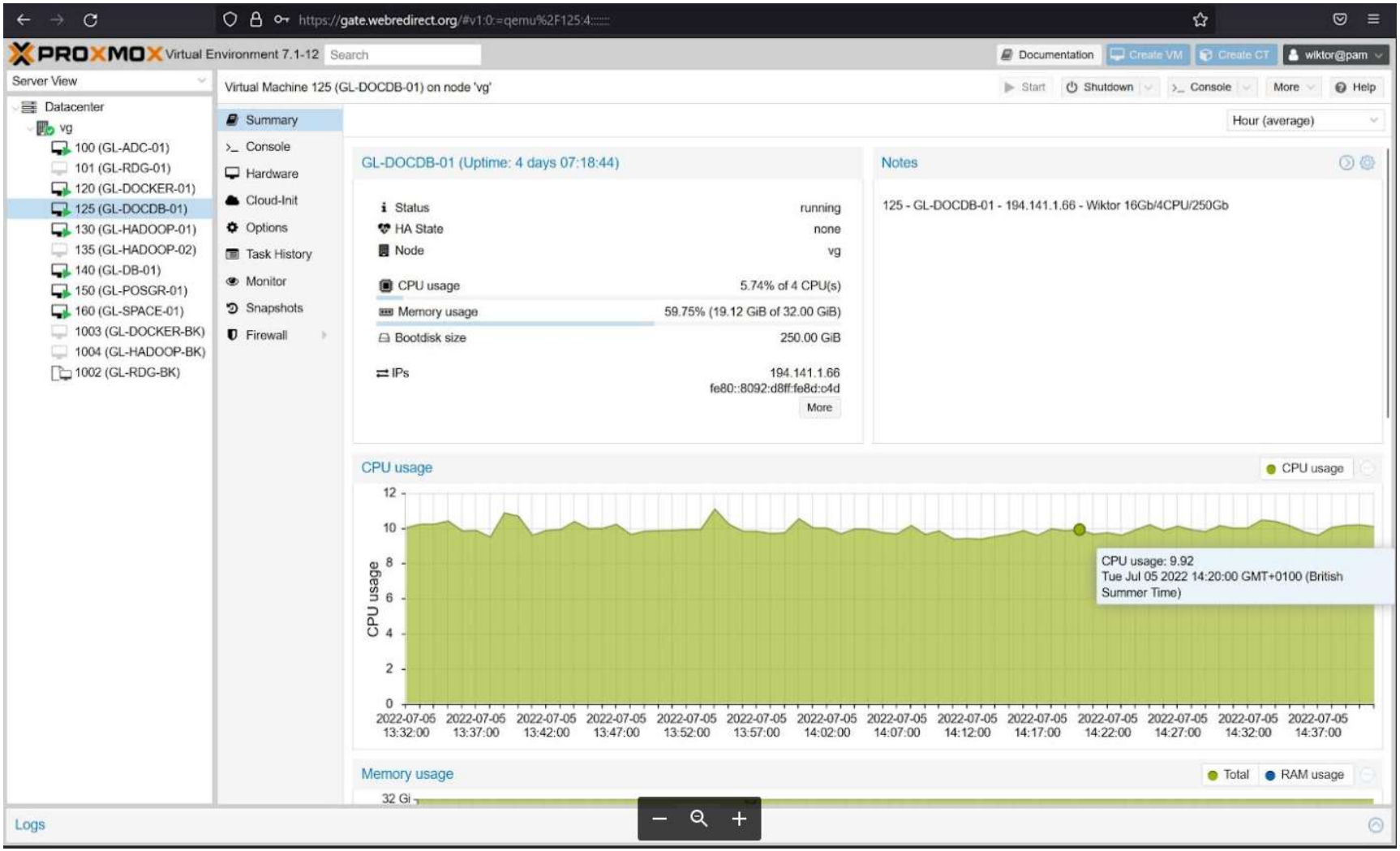
- ✓ Data sources
- ✓ Data processing services
- ✓ Parametric dependencies
- ✓ Information flows between services

Physical Design:
Software Components

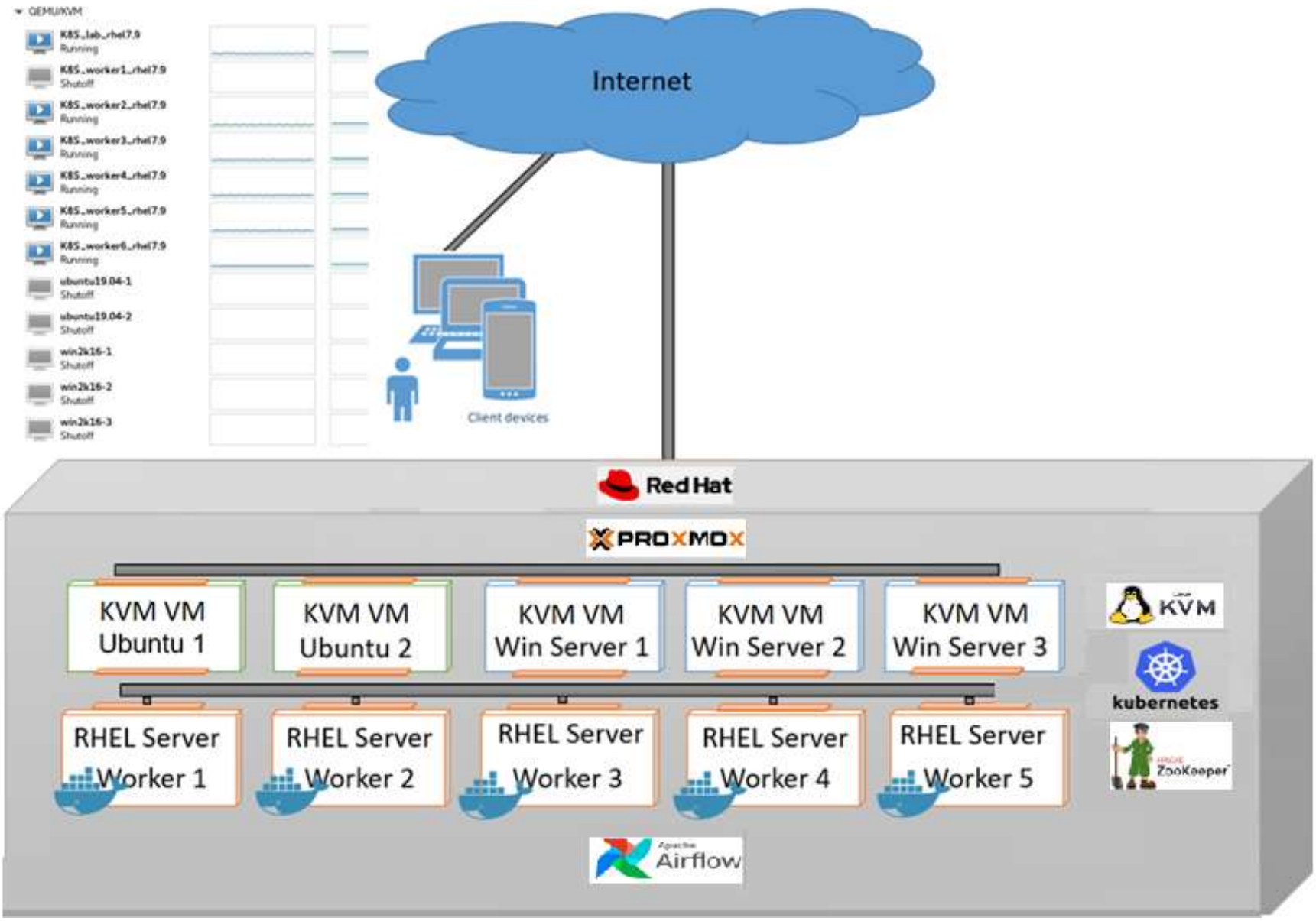


- Data formats, files and protocols
- Client-side and third-party provider APIs
- Server-side services and containerized microservices
- Language interpreters and runtime engines
- Configuration files, templates & scripts

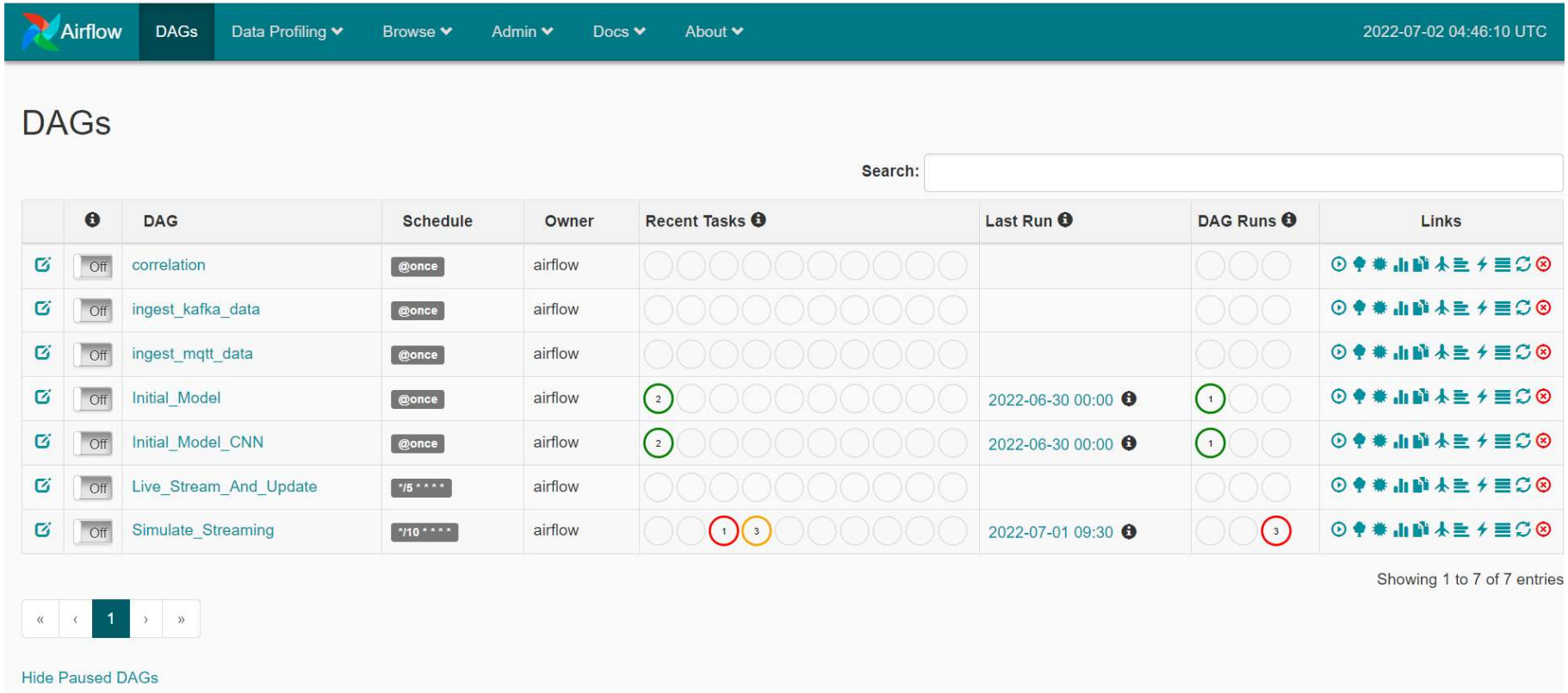
Software Development: Resource access using Proxmox VE



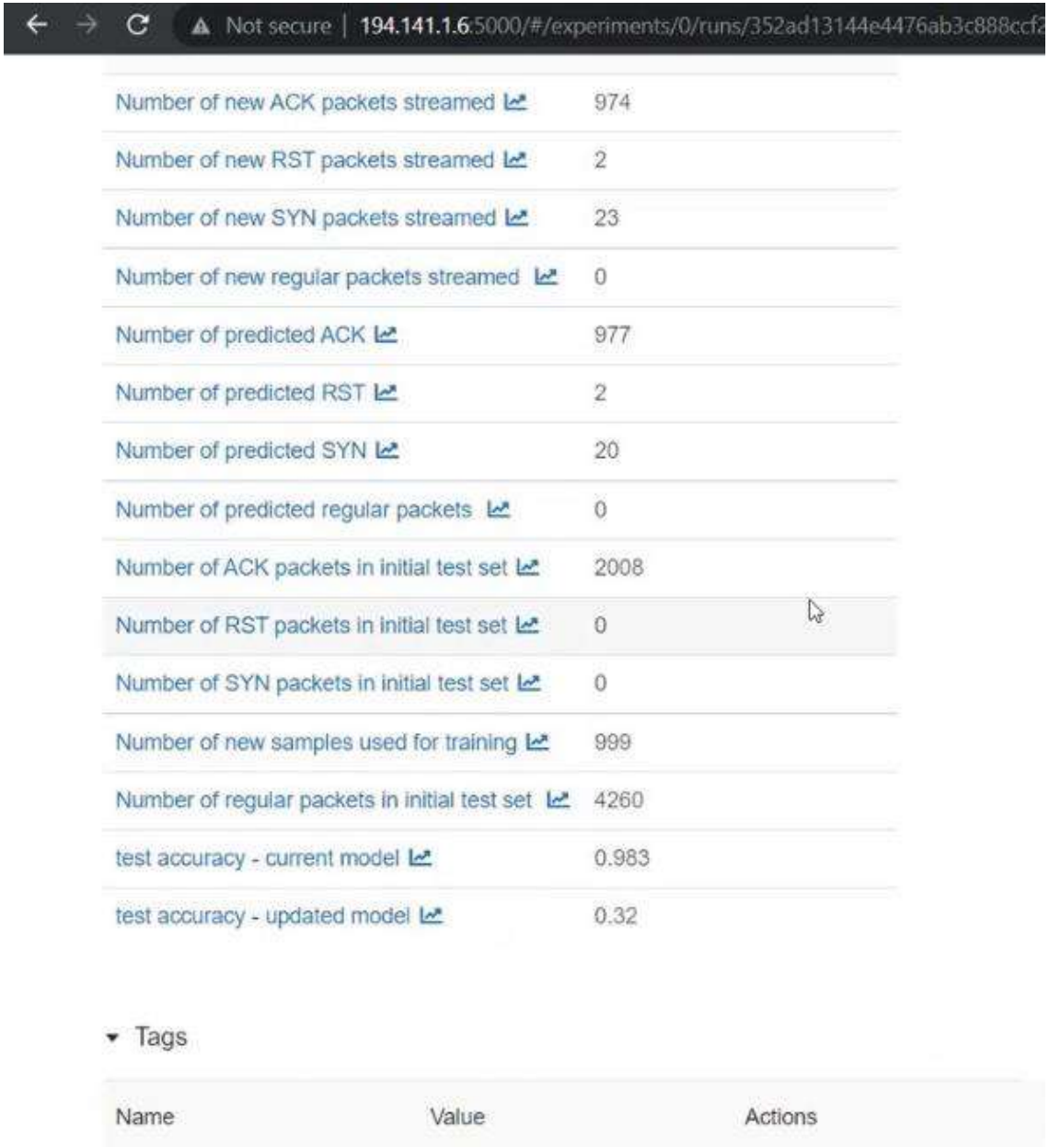
Application Deployment: Component Management using OS, VM and Docker



Service Orchestration: Workflow Scheduling and Execution using AirFlow



Process Monitoring: Execution Statistics gathering using MLFlow



Operations Auditing: Event Log Analysis and Reporting using Elasticsearch and Kibana

5 Where to go from here?

- **Cross-domain integration** of both data and analytics (i.e., environment and transport, environment and healthcare, healthcare and social services, environment and social services, etc.)
- Combining of real-time data with historical data for trends analysis and **investigation of process dynamics** (retrospective and predictive analytics)
- Localization of the data sources using rich data formats which allow **combining geolocation data with sensor data** (using richer representations such as semantic in GeoJSON and layering in KML)
- **Spatial navigation** to data sources by linking spatial ontologies and sensor data (using VR and games engines)
- Combining sensor data, symbolic meta-data and ontological information for **logical analysis** of the data beyond the pure data patterns (hybridization)

Publications

[1] **V. Vassilev, B. Virdee, K. Ouazzane, D. Maryanayagam, V. Sowinski-Mydlarz, et al.**, "Data Platform and Urban Data Services on Private Cloud", in: *Proc. Int. Conf. Smart Trends in Computing and Communications* (SmartCom2023), 24-25 Jan 2023, Jaipur, India, IEEE, 2023 (in print).

[2] **V. Sowinski-Mydlarz, V. Vassilev, K. Ouazzane, and A. Phipps**, "Security analytics framework validation based on threat intelligence", in: *Proc. 9th Annual Conf. Computational Science and Computational Intelligence* (CSCI2022), Dec 14-16, 2022, Las Vegas, USA, IEEE, 2023 (in print).

[3] **V. Vassilev**, **V. Sowinski-Mydlarz**, **D. Mariyanayagam**, **et al.**, "Towards first urban data space in Bulgaria", in: *Proc. Int. Smart Cities Conference* (ISC2), 26-29 Sep 2022, Paphos, Cyprus, IEEE, 2022.

[4] **V. Vassilev**, **S. Ilieva**, **D. Antonova, V. Sowinski-Mydlarz**, **et al.**, "AI-based Hybrid Data Platforms", in: *Curry, E., Scerri, S. and Tuikka, T. (eds.), Data Spaces: Design, Deployments and Future Directions*, pp. 147-172, Springer, 2021.

[5] **V. Vassilev**, **V. Sowinski-Mydlarz**, **P. Gasiorowski, K. Ouazzane, et al.**, "Intelligence Graphs for Threat Intelligence and Security Policy Validation of Cyber Systems", in: *P. Bansal et al. (eds.), Advances in Intelligent Systems and Computing*, Vol. 1164, pp. 125-140, Springer, 2020.

Questions?