LONDON METROPOLITAN UNIVERSITY

Parameterization, Containerization & Orchestration of Data Services on Private Cloud

Prof. Vassil Vassilev
Head of the Cyber Security Research Centre
London Metropolitan University

CONFLUENCE 2022, Noida, India 27-28 Jan 2022

Content

- Some History and Pre-History
- Design Principles behind Cloud-based Data Platforms
- Threat Intelligence Framework of Cyber Security Research Centre (CSRC)
- Smart City Framework of GATE Institute
- Towards European Data Spaces

1 Some history and pre-history

- Some cloud-based security services are already available from the leading cloud providers (Amazon, Google, Microsoft)
- The big organisations which are the natural consumers (fintech, manufacturing and other industrial companies) are still reluctant to send data to the cloud
- The SMEs are unable to afford such services due to the high running costs
- The managed data centres are unable to provide such services due to the high investment costs

CSRC of London Metropolitan University

- ✓ **CISCO** and **Palo Alto** academies, running both standard university and professional training courses using materials of the two companies
- ✓ Operates **secure network** directly connected to the Internet for teaching, research and innovation projects
- ✓ Runs **private cloud** based on Kubernetes and created its own security framework for Tread Intelligence and Security Services
- ✓ **Funding** from Cyber Security division of Lloyds banking Group and private companies in London, UK Government research agencies and EU

Covid-19, Working from Home and the Partnership with GATE Institute

- **Threat Intelligence Framework** of CSRC (Cyber Security division of Lloyds banking Group, private companies in London)
- **The Void** (21 March 2021-31 Sep 2022)
  - Plan B: Sofia
  - Follow up: London
- **Urban Data Space** project
  - Collaboration with Communications Technologies Centre for sensor data capture, security of audio signals and localization within enclosed spaces
  - City Digital Twin project of GATE Institute of Sofia University
  - Indoor air quality in London

2 Design Principles behind Cloud-based Data Platforms

Vertical Layering: Multi-level AI problem

- **Knowledge Representation:** domain concepts, metadata, problem description, task specification, user profiling
- **Problem Solving:** representation, planning and heuristic policies
- **Decision Making:** planning strategies, concurrency resolution
- **Machine Learning:** data analysis
- **Rational Explanation:** the causes for selecting models, methods and algorithms, the solutions and results (including NLP generation)

Horizontal Integration: Service-oriented solution

- **Data Processing** (generation, aggregation, filtering, marshalling, transportation, accumulation, storing, analysis and interpretation)
- **Containerization** (component parameterization, container configuration and cloud deployment)
- **Orchestration** (workflow configuration, execution and monitoring)
- **Synchronization** (access control, isolation and operation ordering)

Technological backbone: Private cloud infrastructure, Service-oriented Architecture

Data Unification: Meta-data descriptions to separate physical and logical data types, abstract and specific operations

Operation Hybridization: Combining domain-specific ontology with domain-independent logic of data processing

Program Code Parametrization: Explicit descriptions of data processing operations parameters, both in isolation and in their mutual dependence

Software Component Containerization: Automatic generation of container configurations for deploying the software components to the cloud server

Service Orchestration: Executing the components in container workflows for solving specific tasks

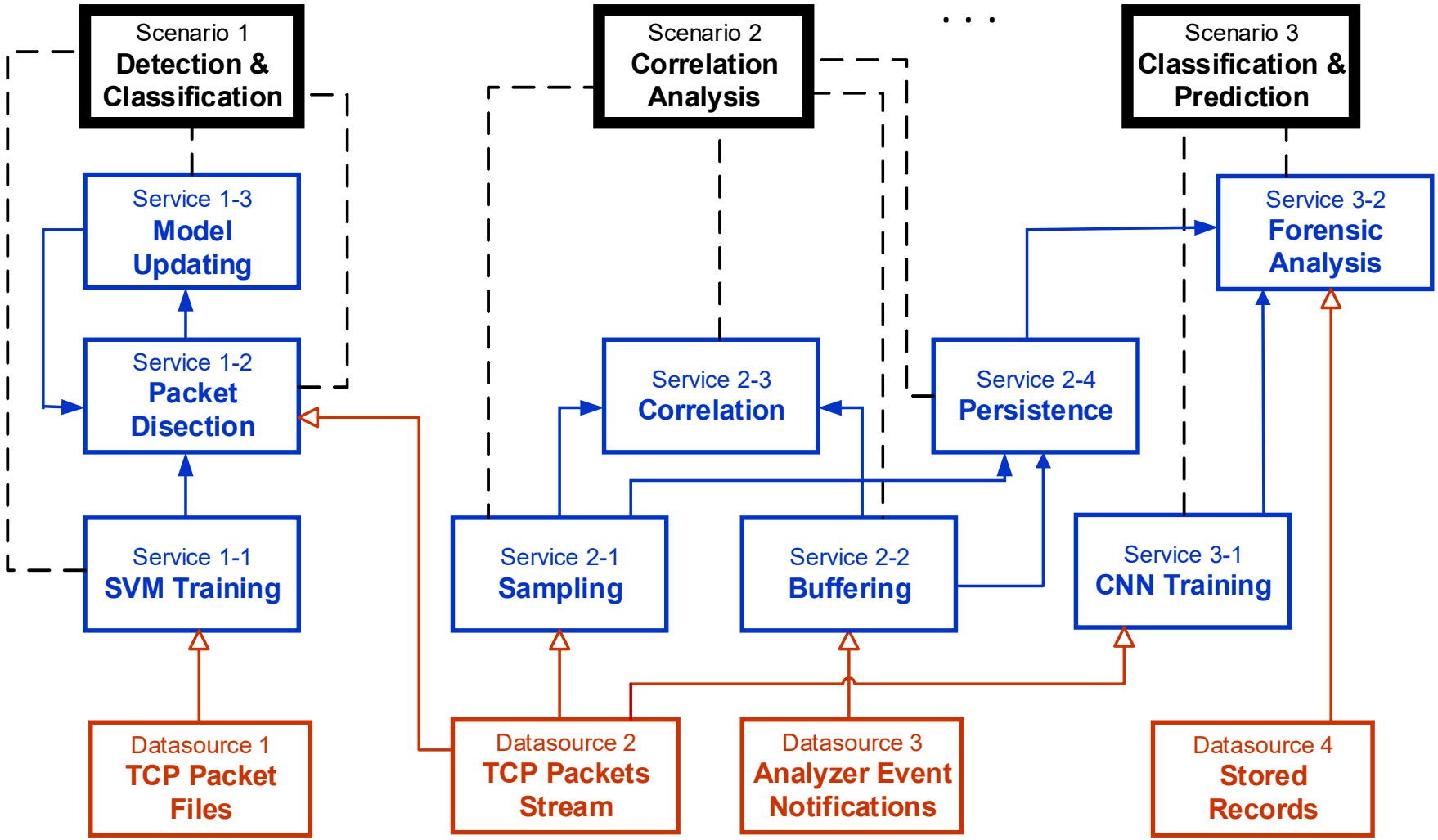
---

Modularization, Robustness, Interoperability, Reusability, Extendibility, Scalability, Maintainability, Explainability, etc.

Technological Inventory: Variety of standards

- ✓ Standard ***protocols*** for data transportation (TCP/IP, HTTP, MQTT, Kafka)
- ✓ Multiple ***repositories*** for storing data (HDFS, NoSQL), meta-data (SQL) and knowledge graphs (Graph DB)
- ✓ Variety of different ***modes*** of data processing (At-the-edge, On-the-fly, In-storage, In-memory)
- ✓ Different ***times*** for processing the data (Completely offline, Real-time, Pseudo Real-time)
- ✓ Different ***algorithms*** for analysis (statistical, machine learning, deep learning, reinforcement learning)

Data Sources, Data Processors and Use cases



3 The Threat Intelligence Framework of Cyber Security Research Centre

Sensors, analyzers and data processing **on the edge** (ESP32, Arduino and Raspberry Pi + Cisco, Palo Alto)

Datasets and **transportation** on the net (Firebase, Mosquitto, Kafka + NiFi)

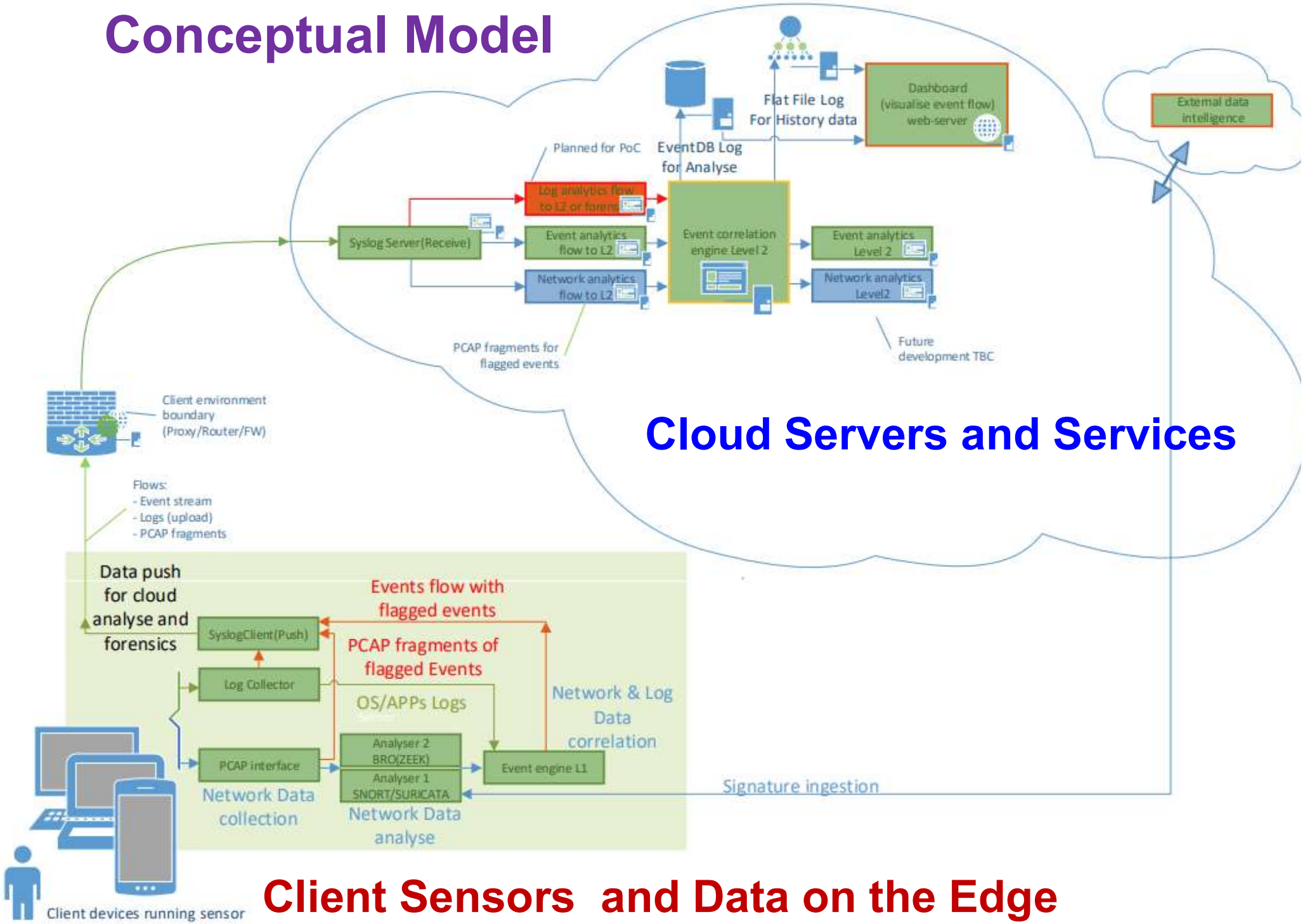
Databases, data lakes and repositories for **storing** data on the cloud (SQL, NoSQL, HDFS, Graph DB + Hadoop)

Machine Learning for **data analysis** on the way, during transmission, upon arrival and later (Python, Java + Spark)

Ontologies and knowledge graphs for **logical analysis** (RDF, OWL, SWRL)

Component containerization, service orchestration and monitoring of the **cloud operation** (Docker, Kubernetes, MLFlow, AirFlow)

Conceptual Model



Off-the-shelf software on the cloud server

Operating System: Linux

Application Management: Docker

Container Management: Kubernetes

Service Management: AirFlow, MLFlow, Zookeeper

Data Transportation: Mosquitto, Kafka, NiFi

Data Storage: MongoDB, Hadoop

Metadata and Ontologies: PostgreSQL, Neo4J

2D and 3D Models: 3DCityDB

Geospatial Visualization: Cesium Ion

Analytics: Python, Java, Spark

Service Ports of the GATE Platform

External Ports & Services

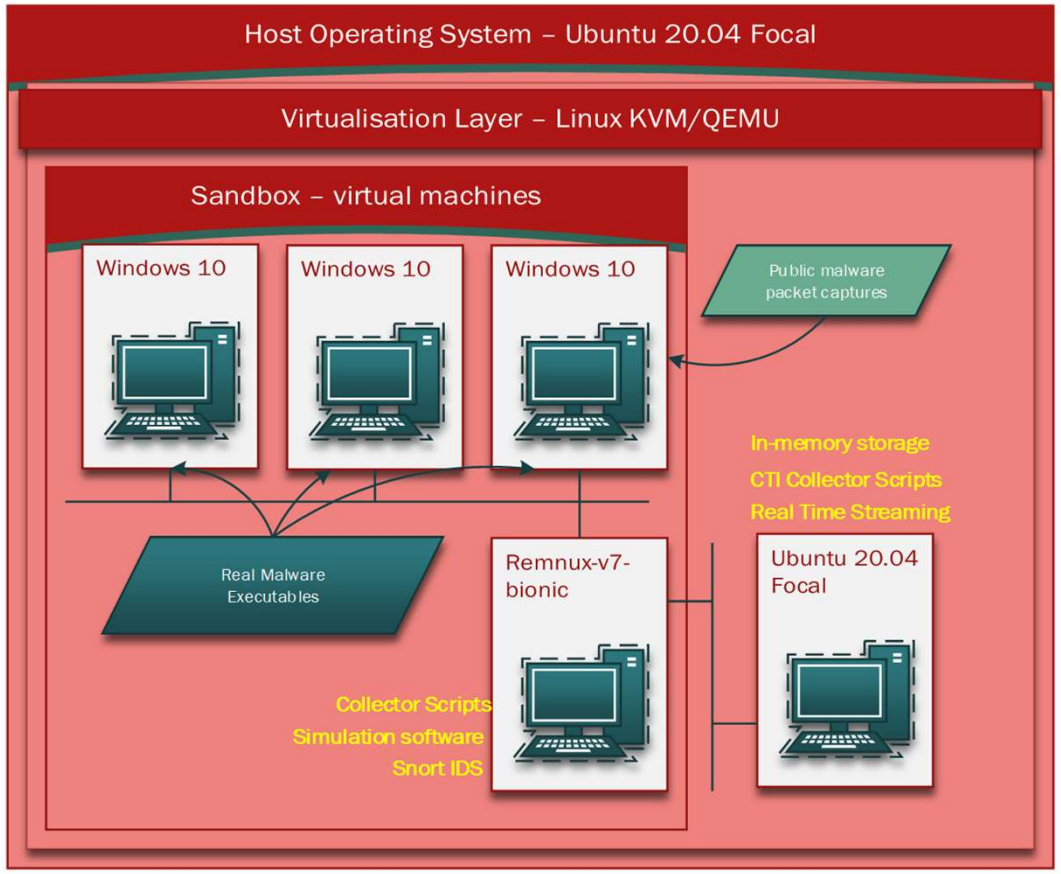
Port	Protocol	Service
5906	VNC	Server Admin
5907-11	VNC	Server-side
65522	SSH	Server Secure
49165	NiFi	Streaming
1883	MQTT	Messaging
65552	SCP	Uploading
30000	Kafka	Broadcasting
16000	Scoop	Importing
5000	HTTP	Configuration
8080	HTTP	Administration
8000	HTTP	Visualization

Internal Ports & Containers

Port	Protocol	Destination
2181	TCP	Zookeeper
5444	TCP	PostgreSQL
27017	TCP	MongoDB
7687	TCP	Neo4J
5432	TCP	3DCityDB
5432	TCP	Hive
8888	TCP	Hue
11000	TCP	Oozie
7070	TCP	Spark

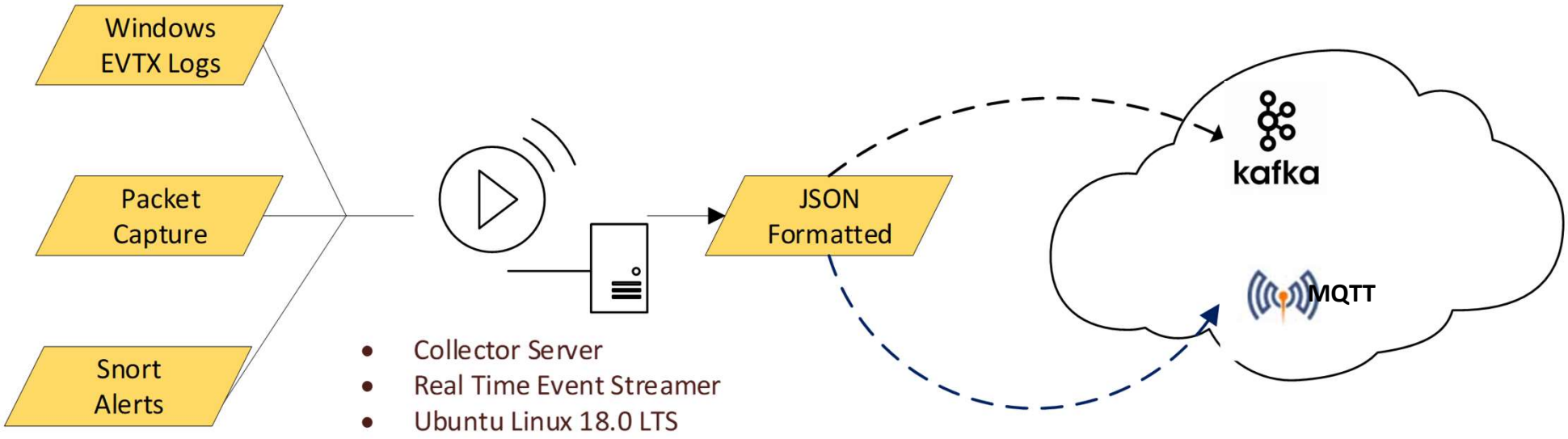
Data Collection

- Collects from multiple sources, both internal and external
- Generates own data from real malware samples in a sandbox environment
- Ingests intelligence data from public internet sources
- Uses virtualisation technology (kvm/qemu) on **Linux** host server
- Implements snort Intrusion detection
- Buffers events and packet streams in memory
- Streams data from memory to the cloud using real time streaming technologies (**Kafka** + **MQTT**)

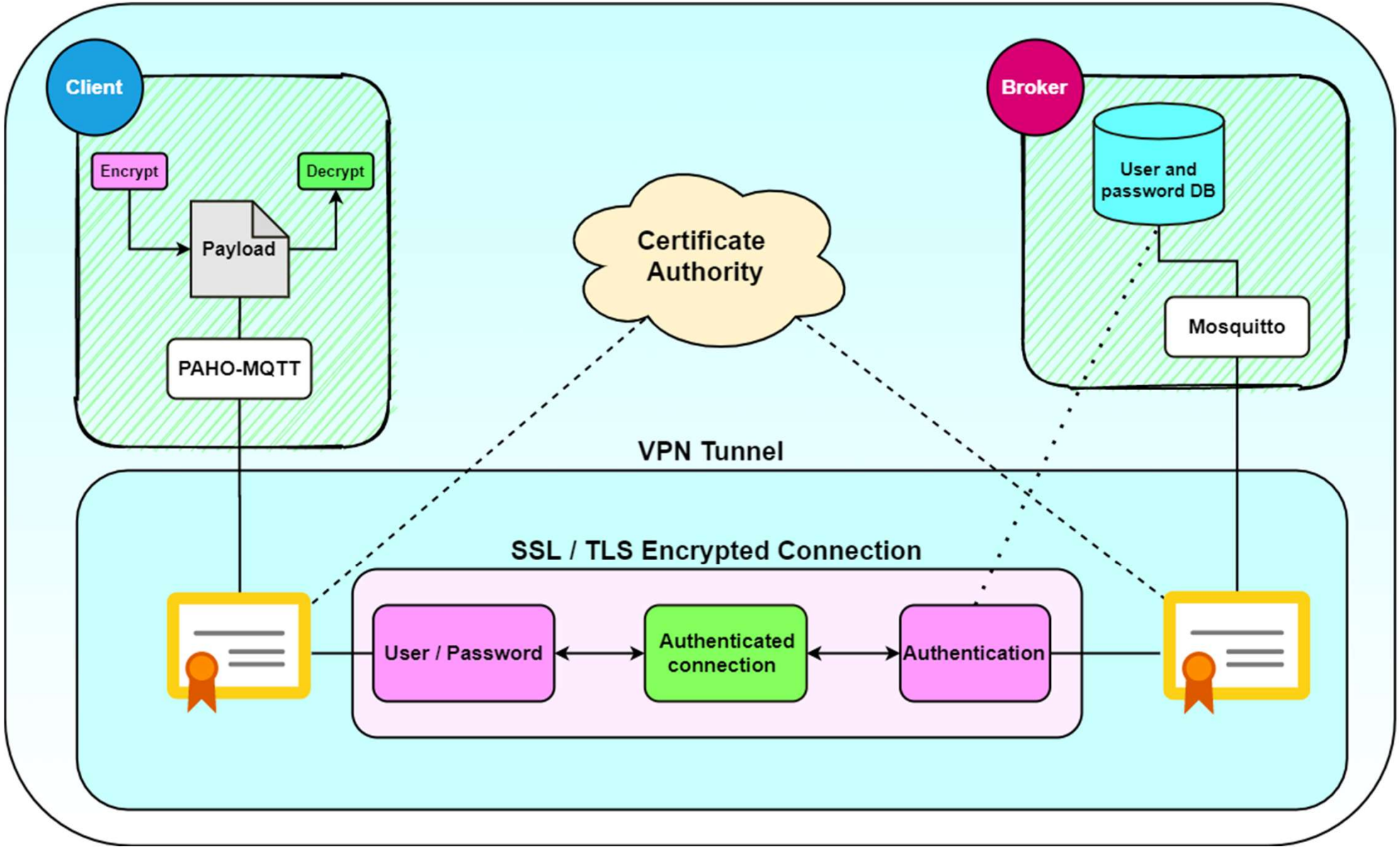


Real Time Streamer

- Receives **EVTX** Logs, **PCAP** network traffic captures and **Snort IDS** Alerts
- Implements a custom collector server application
- Streams data feeds using Apache **Kafka** and Confluent **MQTT** messaging as **JSON** formatted data



MQTT Encrypted Payload on Encrypted and authenticated connection over VPN



Data Analytics on the Cloud

Preliminary analysis of the network traffic by ***correlating*** packets and analyzer events using Pearson

Initial recognition of the suspicious packets in the network traffic based on ***classification*** using SVM

Deep forensic analysis of the trends and ***prediction*** of the appearance of malicious packets as a result of potential intrusion using CNN

SVM for Classification

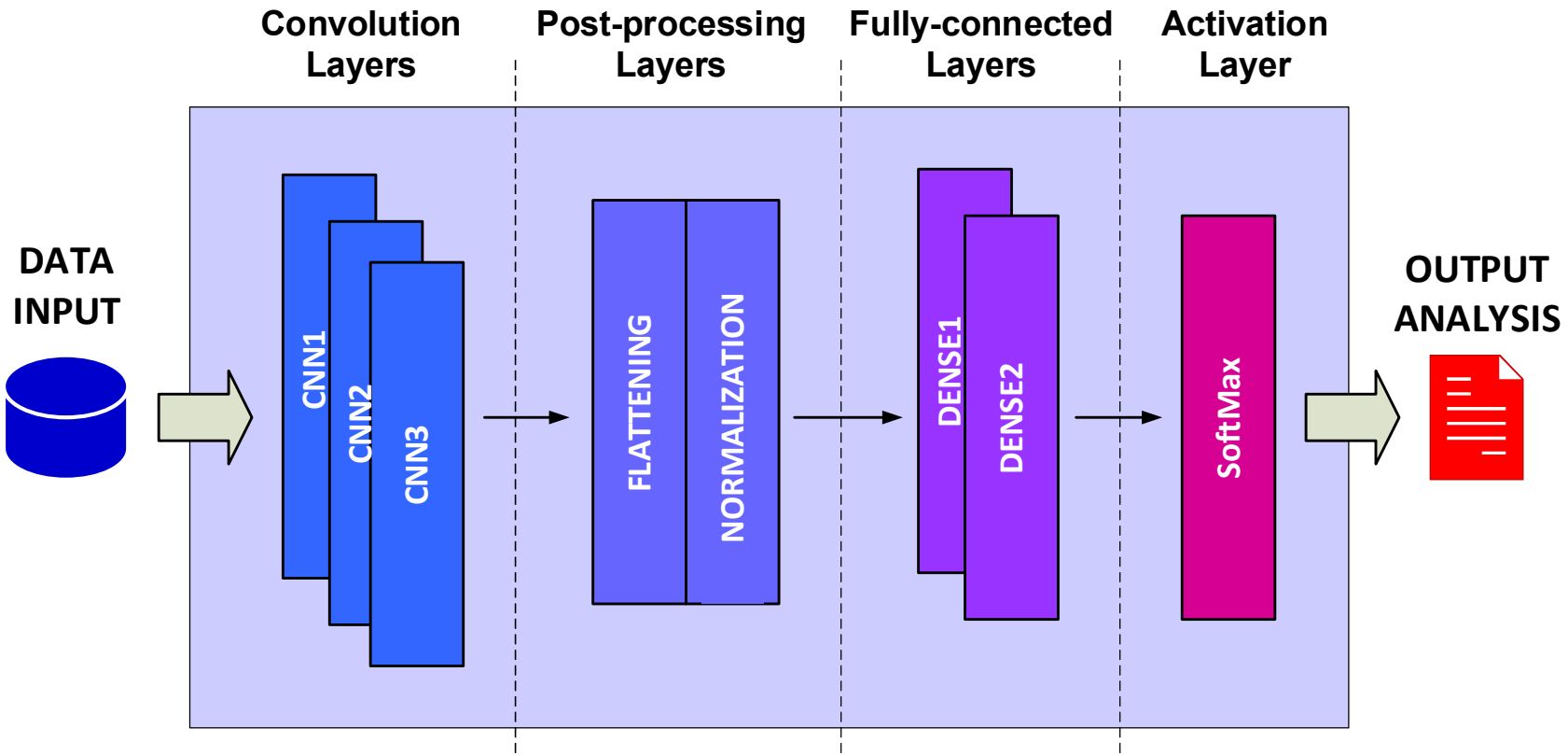
	Precision	Recall	F1-score	Support
0.0	1.00	0.94	0.97	4729
1.0	0.83	0.98	0.90	1376
2.0	0.96	0.93	0.94	164
accuracy			0.95	6269
macro avg	0.93	0.95	0.94	6269
weighted avg	0.96	0.95	0.95	6269

Legend

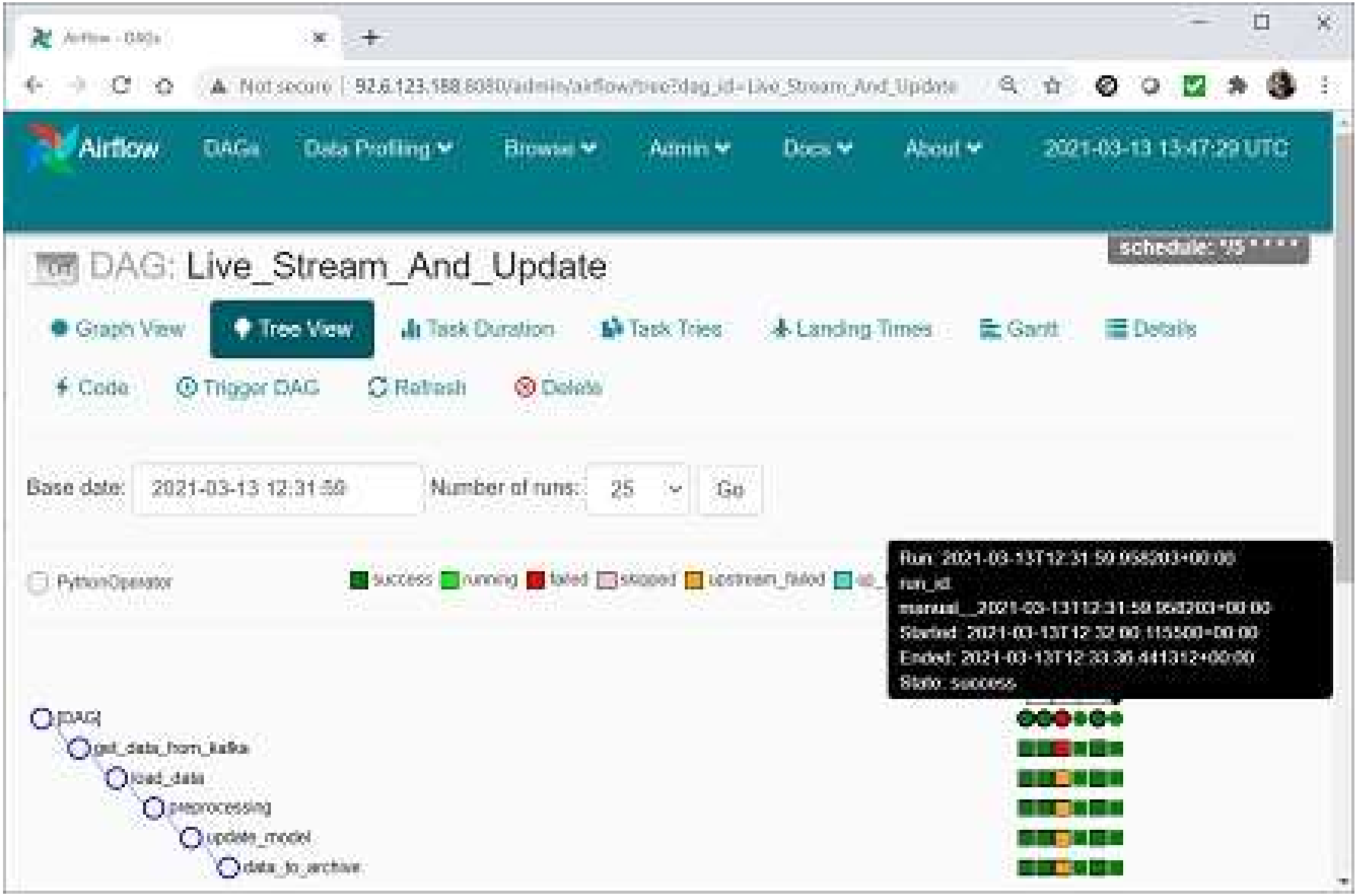
Precision - percentage of correct positives classification.
Recall - percentage of correctly classified true positives.
F1- score - weighted harmonic mean of precision and recall.
Support - the sum of actual instances of class.

CNN for prediction

- **Convolutional Neural Network (CNN):** The model is used on a time-series dataset in a predetermined time range to predict the suspicious flags within the data



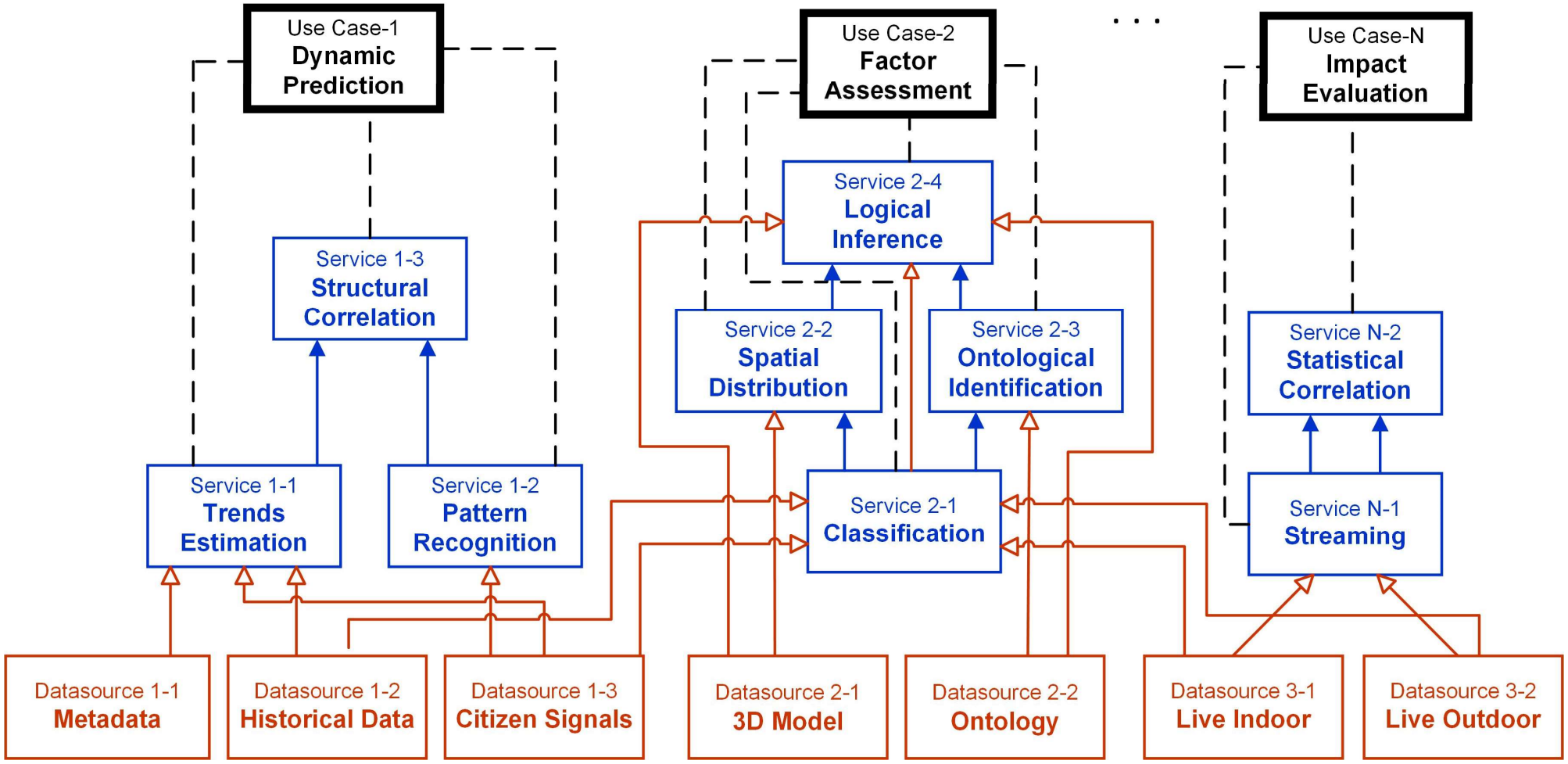
Orchestration of container workflows on Kubernetes under the control of AirFlow



4 Urban Data Space Project of GATE

- EU Horizon 2020 (15M EUR)
  - Chalmers University in Gothenburg (Sweden)
  - Knowledge Innovation Technologies Institute
- Local Funding (15M EUR)
  - Bulgarian Government
  - Rila Solutions
  - Ontotext
- International Partnership within Europe and America
  - EU: Norway, Italy, Austria, Poland, Estonia
  - Switzerland
  - USA, UK

Urban Data Space: Data Sources, Data Services and Use Cases



Seven different data sources:

- ***historical data*** about the environment pollution (CSV files)
- hourly updates of the ***outdoor conditions*** using real-time sensor measurements (JSON streams)
- ***citizen's complaints*** from a helpdesk database, currently available offline (Excel files)
- sensor measurements of the ***indoor conditions*** in specific buildings (JSON streams)
- 2D ***map of the city*** (KML files)
- 3D ***model of the urban area*** in focus (GML files)
- ***Ontological model*** of the metropolis (RDF files)

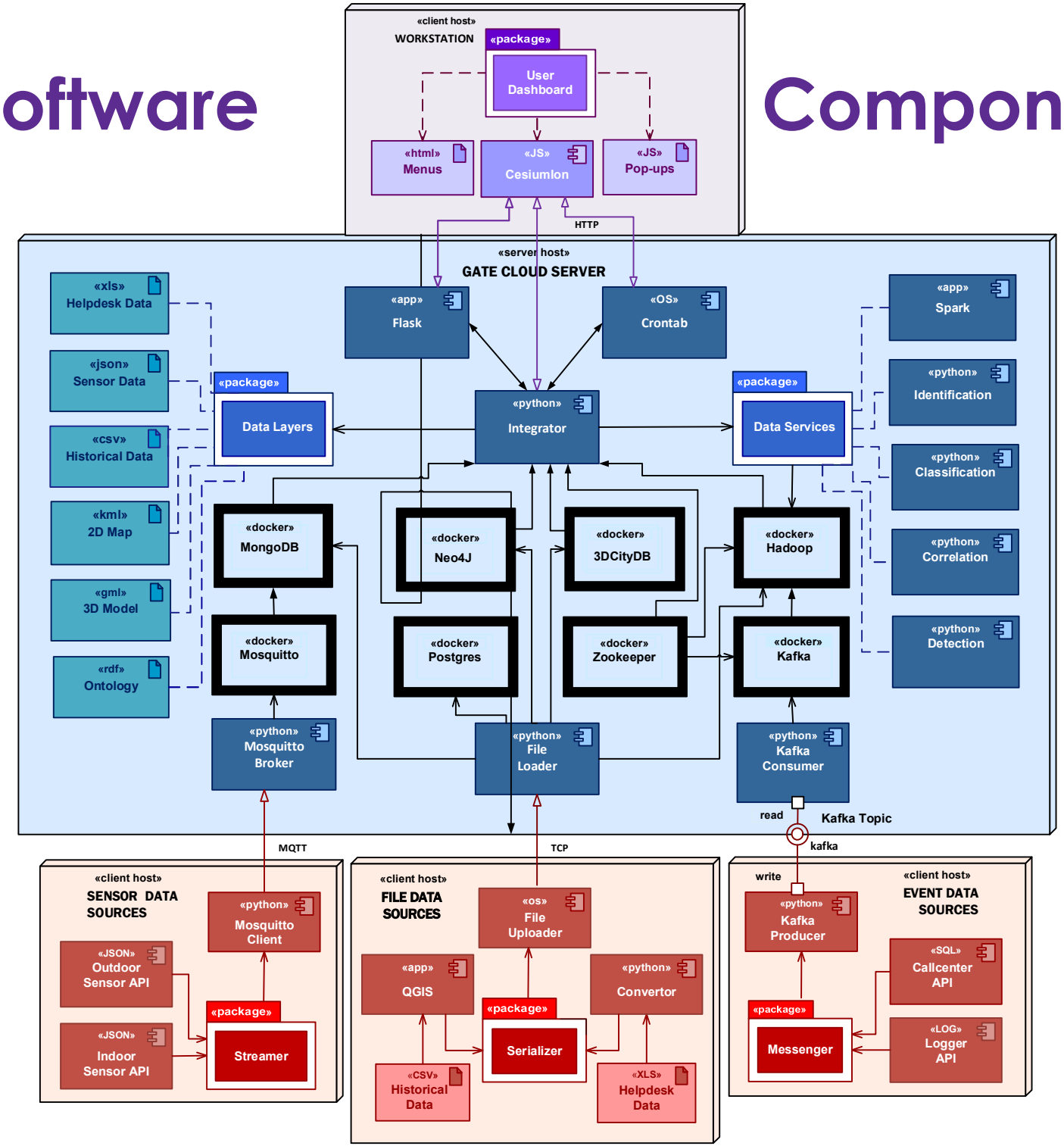
Eight main data services

- ✓ Data ***collection*** from the sources (sensors, files, databases)
- ✓ Data ***pre-processing*** at the source (formatting, labelling, annotating, etc.)
- ✓ Data ***transportation*** to the cloud (streaming, uploading, messaging)
- ✓ Data ***post-processing*** on the cloud (filtering, correlating, buffering)
- ✓ Data ***ingestion*** into the data storages
- ✓ Data ***visualization*** on the Web
- ✓ Data ***analysis*** (detection, classification, identification & prediction)
- ✓ Data ***migration*** (transferring the data to Hadoop for further analysis and support of the future **GATE Urban Data Space**)

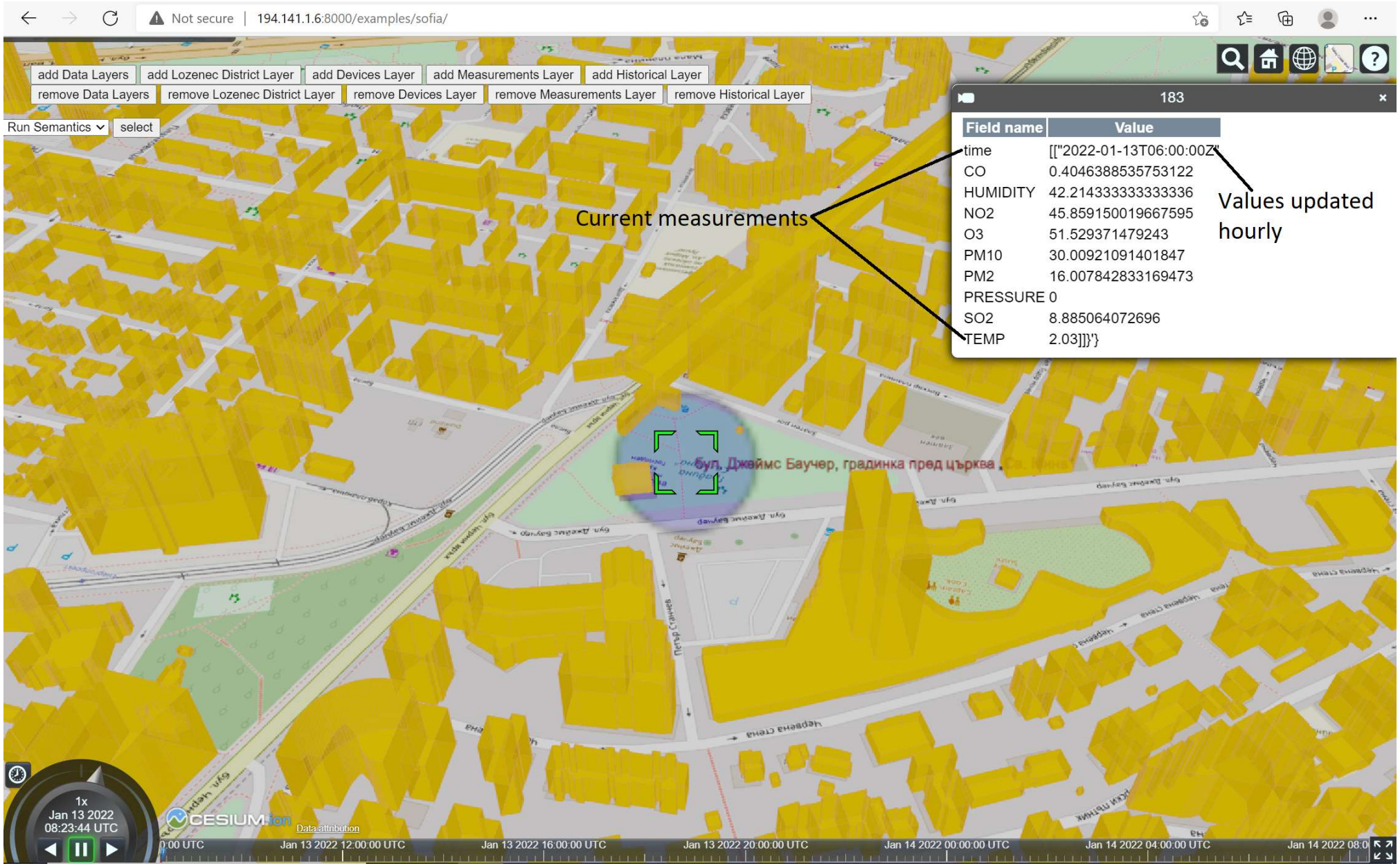
Five use cases in focus:

- ✓ Detecting the pollutions by statistical analysis of the sensor data about the air pollution - ***Statistical Data Analysis***
- ✓ Formulating the trends by analysis of the historical data about the air pollution – ***Offline Pattern Recognition***
- ✓ Identifying the polluters by searching the surrounding locations in the ontological model for possible emitters - ***Logical Data Analysis***
- ✓ Issuing warning concerning the air quality by correlating data streams containing measurements of relevant indoor factors – ***Real-time Correlation Analysis***
- ✓ Predicting the tendencies by matching historical trends and current trends – ***Offline Structural Analysis***

Software Components

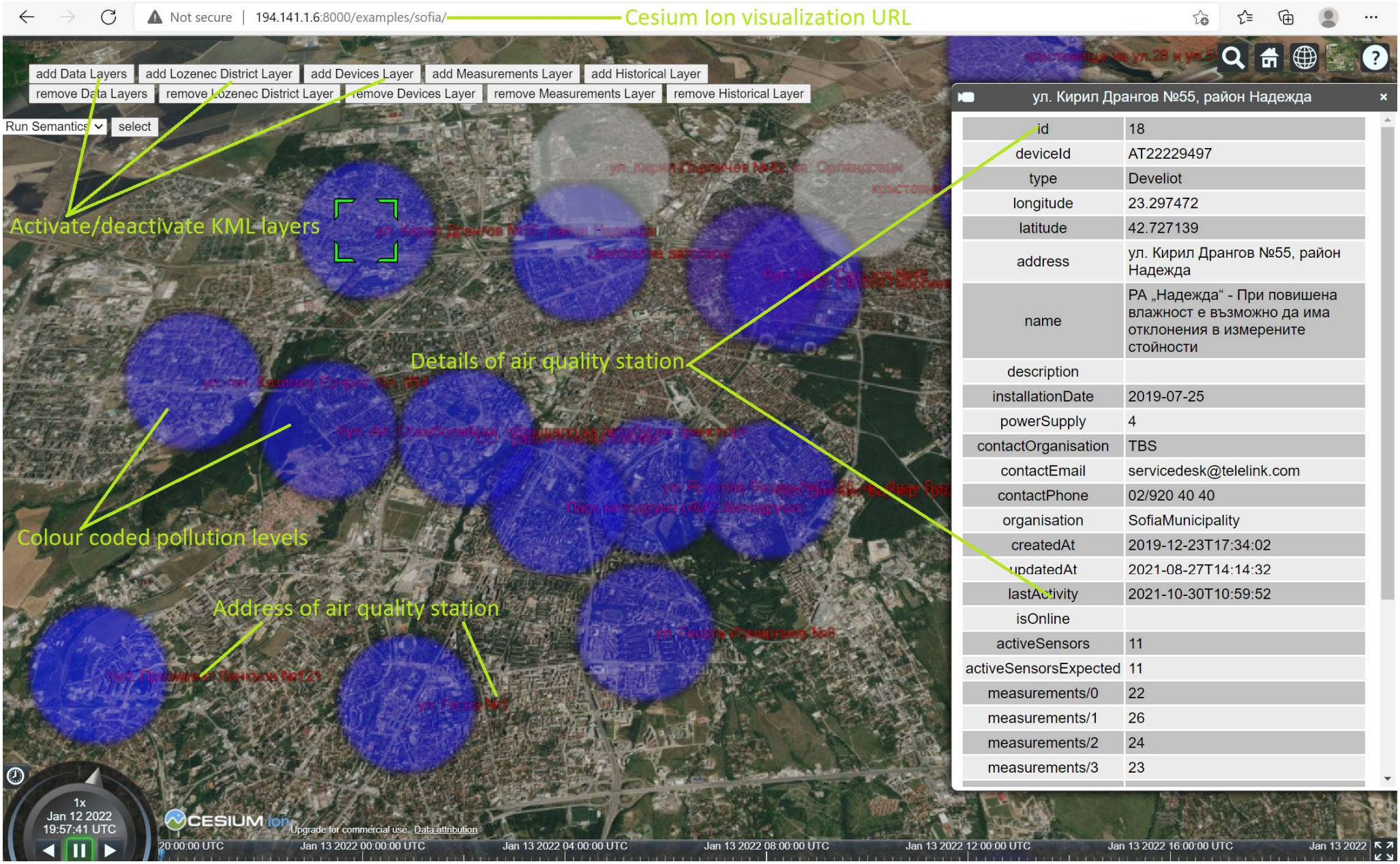


Visualization in 3D of Sofia District buildings with Open Street Map terrain layer



http://194.141.1.6:8000/examples/sofia/

Visualization of the air pollution in Sofia with Bing Map tarrain layer



http://194.141.1.6:8000/examples/sofia/

Analysis: Indoor Data vs Outdoor Data

- Indoor Data

	PM2.5	PM_2.5	PM_10	Temperature	Pressure	Altitude	Humidity	CO2	tVOC
0	1.00	4.00	4.00	31.12	1020.99	-64.24	51.45	612	32
1	2.00	4.00	4.00	31.32	1020.94	-63.86	51.06	1073486112	13
2	3.00	4.00	4.00	31.29	1020.98	-64.17	50.73	1073486112	13
3	1.00	3.00	3.00	31.04	1020.99	-64.26	51.46	574	26
4	3.00	3.00	3.00	31.09	1020.97	-64.04	51.30	1073486112	19
5	4.00	5.00	6.00	31.13	1020.90	-63.51	51.30	1073486112	33

```
corr, _ = pearsonr(df_indoor_data["Temperature"],df_indoor_data["Humidity"])
print('Pearson correlation for indoor data: %.3f' % corr)
```

```
Pearson correlation for indoor data: -0.856
```

⇒ There seems to be a negative correlation between temperature and humidity in the indoor sensor data.

Analysis: Indoor Data vs Outdoor Data

- Outdoor Data

	time	CO	HUMIDITY	NO2	O3	PM10	PM2	PRESSURE	SO2	TEMP
0	2021-10-30T10:00:00Z	0.503908	46.322000	15.733006	43.096059	8.291513	5.848458	955.855400	6.367629	16.202000
1	2021-10-30T10:00:00Z	0.503908	46.322000	15.733006	43.096059	8.291513	5.848458	955.855400	6.367629	16.202000
2	2021-10-30T10:00:00Z	0.496910	36.311800	38.971575	53.154892	15.797556	8.682095	933.432200	5.490842	16.578200
3	2021-10-30T10:00:00Z	0.496910	36.311800	38.971575	53.154892	15.797556	8.682095	933.432200	5.490842	16.578200
4	2021-10-30T10:00:00Z	0.501049	95.717667	11.779628	47.642170	5.965099	3.524747	959.594000	4.772792	20.416667
5	2021-10-30T10:00:00Z	0.501049	95.717667	11.779628	47.642170	5.965099	3.524747	959.594000	4.772792	20.416667

```
corr, _ = pearsonr(df_measurements_data["TEMP"],df_measurements_data["HUMIDITY"])
print('Pearson correlation for Outdoor data: %.3f' % corr)
```

```
Pearson correlation for Outdoor data: 0.972
```

⇒ There seems to be a positive correlation between temperature and humidity in the outdoor sensor data.

Next Phase: Towards European Data Spaces

- EBDVA
- IDSA and GaiaX
- Fiware
- GATE Institute
- London Metropolitan University

Hadoop & GATE Platform

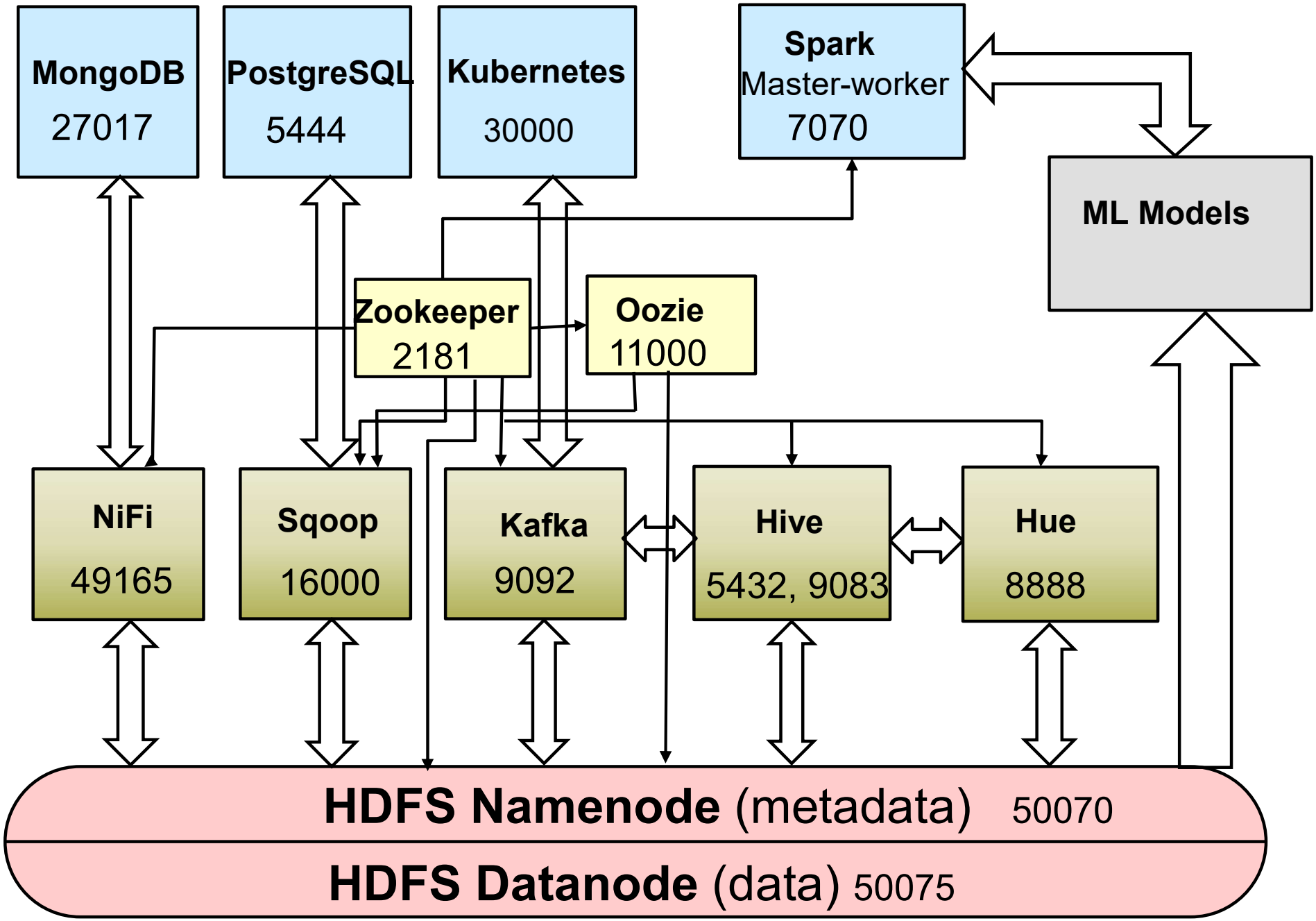
Hadoop distributions:

- Apache Hadoop – client/server, requires configuration
- HortonWorks distribution (HDP) – single container, not supported
- Cloudera's old distribution (CDH) – client/server, expiring support
- Cludera's (CDP) – cloud integration, not free (subscription only)

GATE Hadoop installation:

- Apache Hadoop 2.x.x was first tested in a local worker node of the GATE cloud server
- It was then migrated to the master node of the GATE cloud server (**194.141.1.6**) to be used for building data lakes in a client/server mode of operation
  - ✓ The standard components of Hadoop (**HDFS**, **MapReduce, YARN**) can be used on each via the OS, Hue or Oozie
  - ✓ The components from Hadoop ecosystem (**Kafka**, **NiFi**, **Scoop**, **Hive**) are installed in containers for data ingestion
  - ✓ Two additional public domain applications (**Spark and Hue**) were deployed in separate Docker containers for data analysis and ML

GATE Hadoop Ecosystem



Enabling IDSA Data Spaces

Identity management: Identification, Authentication and Authorization with guaranteed level of precision

Data management: Data Models, Serialization, Buffering and Combining with Metadata with guaranteed level of security

Operation Management: Access Control, Logging, Auditing

Service management: Data Measurement, Service Consumption, Trust

Summary

Parametrization of the analytics enables their reuse within ***service-oriented*** architectures and on the cloud

Containerization supports ***analytics on demand*** for business processing in industry, commerce, national grids and networks

- Security threats – intrusions, eavesdropping, fraud
- Safety issues – errors, malfunctions, breakdowns
- Decision problems – concurrency, scheduling, communication

Orchestrated cloud services can streamline the analytics by running ***workflows*** of data processing operations

- Domain ontology can be modelled using standard OWL format
- Policy rules can be encoded in SWRL format
- Intelligence graphs can represent scenarios on analytical level

Hybridization allows executing more complex ***use cases***

- Analyzing and correcting the logical vulnerabilities
- Detecting, recognition and classification of security/safety threats
- Mitigating risks caused by threats using correcting actions
- Optimization by reinforcement learning from historical data