



EC Data Platform State of play and way forward external presentation

DIGIT B1

8 December 2023

DIGIT.B.1

Data, Artificial Intelligence & Web

Unit Mission

...focus on the following areas:

- data, information and knowledge management;
- data analytical processes and tools;
- artificial intelligence;
- business intelligence; digital innovation

^ Unit Services

✓ Data Analytics

✓ Data Centre File Storage (DCFS)

✓ Data Virtualization with Denodo

✓ EC Data Platform

✓ Europa Web Platform

✓ Qlik Sense

✓ QlikView

✓ SAP BusinessObjects Business Intelligence

✓ SAP Data Services

✓ Single Integrated Framework for Collaboration

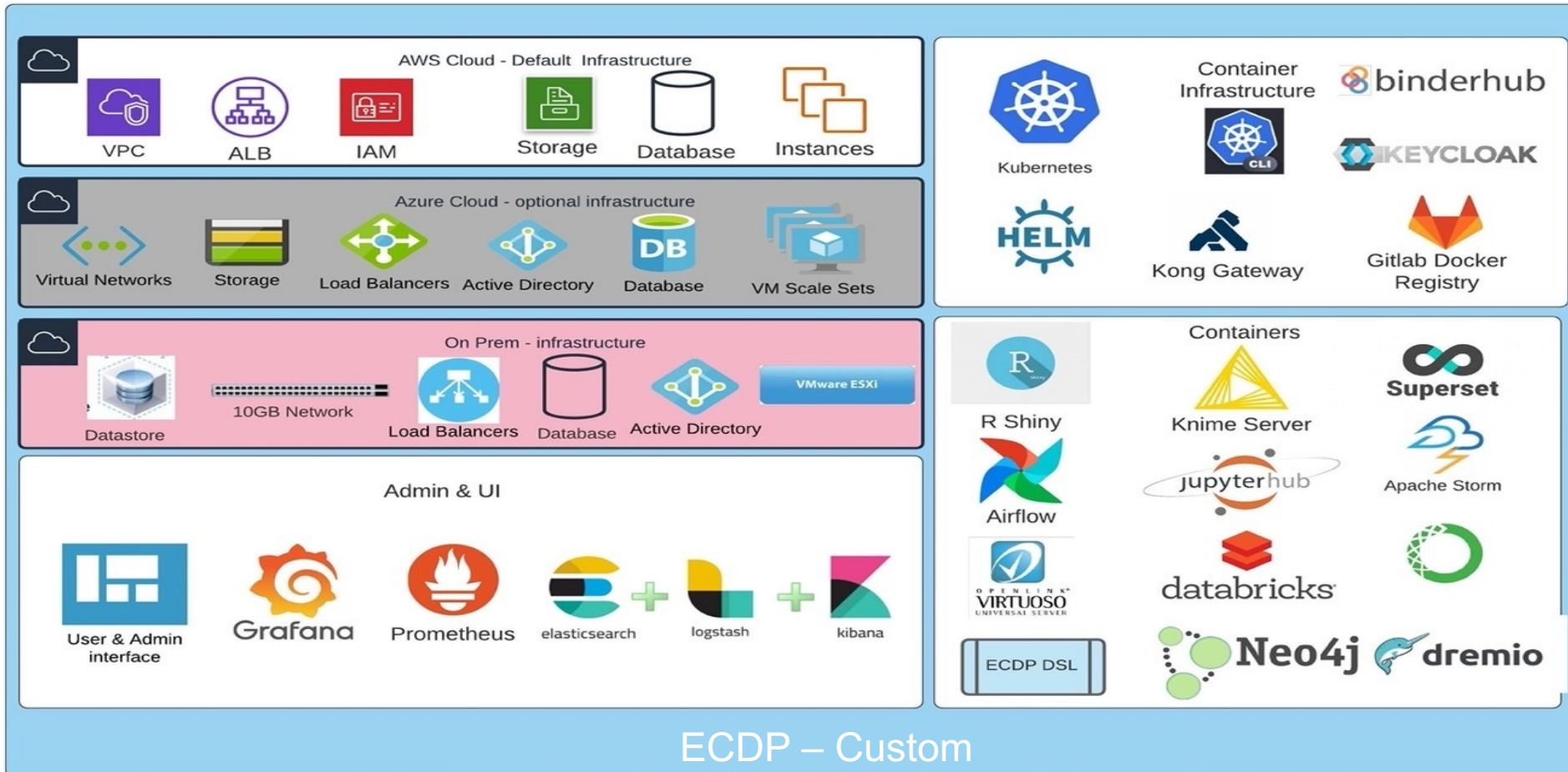
✓ Wikis platform

Data Analytics and ECDP

- **Detecting** analytics solutions available in the Commission
- **Assess** whether the solutions can become corporate services
- Address **specific needs** of the DGs on certain business areas
- Promote **prototyping/piloting** in view of creating corporate services
- **Create flagship projects** to pioneer solutions for certain business areas
- **Use the ECDP** as a building block for the solutions
- **Reuse the developed components** to feed the ECDP.

EC Data Platform

The EC Data Platform is a corporate data environment where EC staff can share and re-use data from different policy areas and perform analytics to extract insights



- Architecture
- Infrastructure
- Tools
- Integration
- Services
- Governance
- Security
- Data access
- Reusable components
- ...

Addressing @EC DGs and users



Implementation

- Design service offering to **answer most EC needs / target most EC persona.**



- DSL Azure-based (*standard*) offering is targeting mainstream usage, *e.g.* policy analysts.
- DSL AWS-based (*custom*) offering can be customised by users for own needs, *e.g.* data scientists.

- Support implementation of **ad-hoc projects requested by DGs with specific needs.**



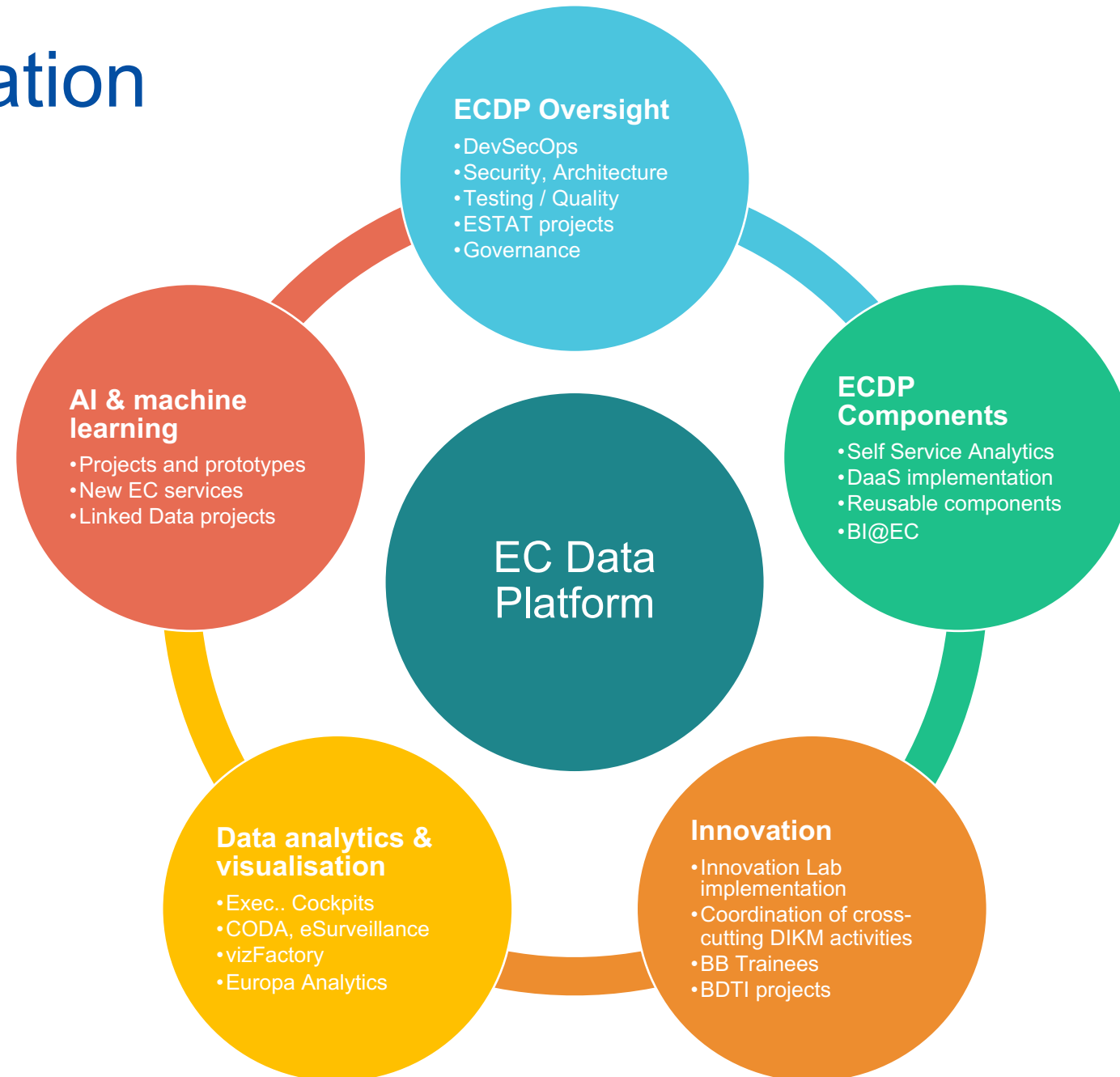
- Analytics PoCs and pilots.
- Dedicated flagship projects: ESTAT WIH, GROW eProcurement, JUST eLab, ...

- **Build communities, empower users, share best practices and collect feedback.**



- Identify most common data-related use cases and create synergies across DGs.
- Create / support CoPs on Connected, Yammer, ...
- Organise data events and share data stories.

Organisation



ECDP Overview

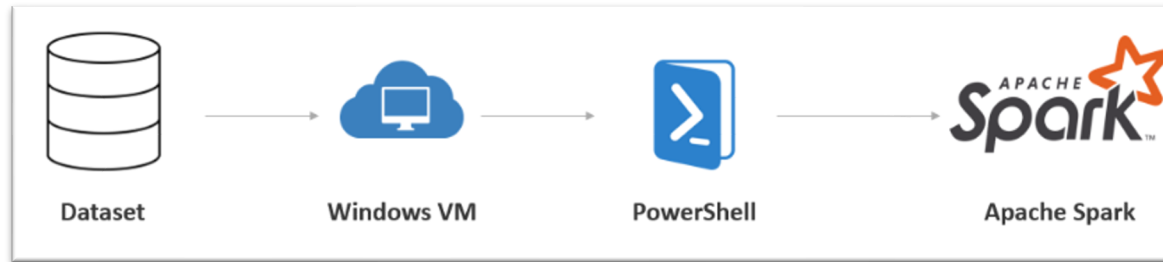


What the EC Data Platform is:

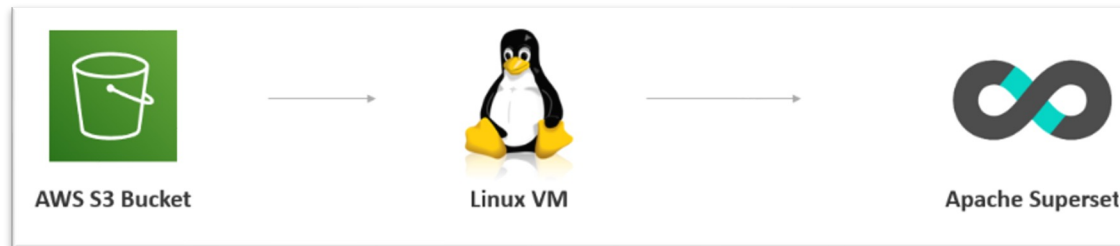
- ❖ Sandbox environment
- ❖ 3 flavours: AWS, Azure, Open Source
- ❖ Accessible to EC and Member States*
 - ❖ Big Data Test Infrastructure

What can you use it for? (examples)

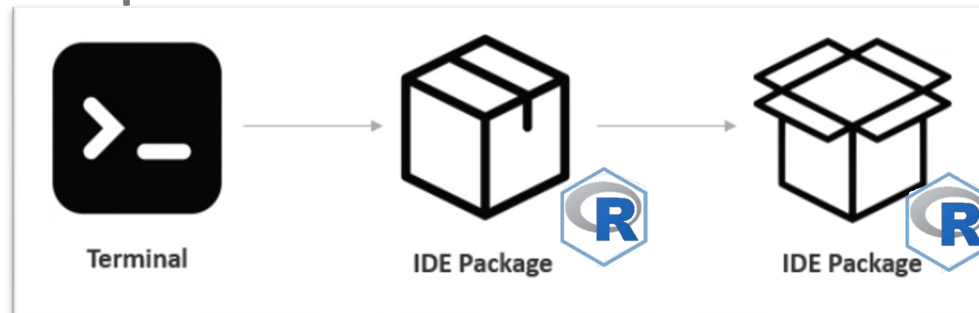
- I need to train Machine Learning models on a large amount of data stored on my Windows Virtual Machine.



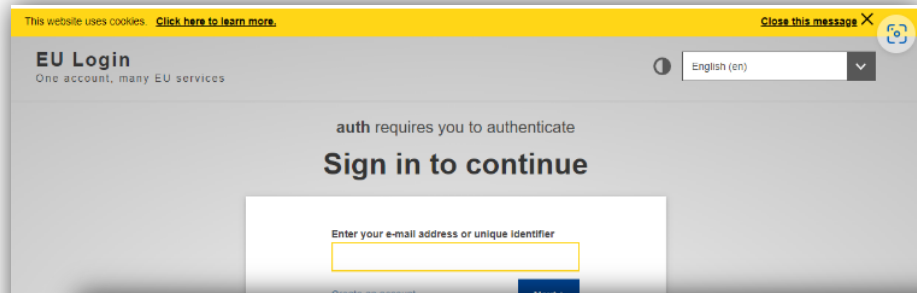
- I need to extract insights from a large data set and create the related visualizations.



- I need to setup a development environment in R on shared machines to build R scripts.



ECDP – Open-Source portal (1/3)



This screenshot displays the user account dashboard. On the left is a navigation menu with the following items: Home, My Account (highlighted), Service Catalog, My Services, My Data, and Admin. The main content area is divided into two sections:

User Info

Username: [REDACTED]
First Name: [REDACTED]
Last Name: [REDACTED]
Email: [REDACTED]
Groups: DSL0001
Created: Mon Jul 11 2022

DSL Info

DSL0001	Used	Total	Max per user
CPU	8.051 Cores	20 Cores	1 Cores
RAM	9.888 GBs	20 GBs	1 GBs
Storage	93 GBs	200 GBs	

ECDP – Open-Source portal (2/3)

- Home
- My Account
- Service Catalog**
- My Services
- My Data
- Admin
- Logout

Service Catalog

Airflow - v2.3.0

Description
Airflow is a platform created by the community to programmatically author, schedule and monitor workflows.

Apache Superset - v1.0

Description
Apache Superset is a modern data exploration and visualization platform. It is fast, lightweight, intuitive, and loaded with options that make it easy for users of all skill sets to explore and visualize

Apache Superset v2.1

Description
Apache Superset is a modern data exploration and visualization platform. It is fast, lightweight, intuitive, and loaded with options that make it easy for users of all skill sets to explore and visualize highly detailed geospatial

ElasticSearch - v7.17.3

Description
Elasticsearch is the distributed, RESTful search and analytics engine at the heart of the Elastic Stack.

H2o-3 - v36.1.1

Description
H2O is an in-memory platform for distributed, scalable machine learning. H2O uses familiar interfaces like R, Python, Scala, Java, JSON and the Flow notebook/web interface, and works seamlessly with big data technologies like Hadoop and Spark.

Knime - v5.1.0

Name [?]

Required (a-z,A-Z,0-9,-,_,.)

Sharing Status

Select Status

Required

Group

Select Group

Required

Launch

data science

Description
The Jupyter documents used for much more

KNIME

Description
KNIME A data science development data science everyone

PgAdmin

Description
PgAdmin is the most popular and feature rich Open Source administration and development platform for PostgreSQL, the most advanced Open Source database in the world.

Description
PostgreSQL is a powerful, open source object-relational database system with over 30 years of active development that has earned it a strong reputation for reliability, feature robustness, and performance.

Description
An integrated development environment for R and Python, with a console, syntax-highlighting editor that supports direct code execution, and tools for plotting, history, debugging and workspace management.

0.4 - all-spark-

on for creating and sharing ons, and text. It can be g, machine learning, and

Jupyterlab - lab-4.0.4 - datascience-notebook

Description
The Jupyter Notebook is a web application for creating and sharing documents that contain code, visualizations, and text. It can be used for data science, statistical modeling, machine learning, and much more.

MinIO - RELEASE.2022-06-20T23-13-45Z

Description
MinIO offers high-performance, S3 compatible object storage. Native to Kubernetes, MinIO is the only object storage suite available on every public cloud, every Kubernetes distribution, the private cloud and the edge. MinIO is software-defined and is 100% open source under GNU AGPL v3.

Spark - v3.2.1

Description
Apache Spark is an open-source unified analytics engine for large-scale data processing. Spark provides an interface for programming clusters with implicit data parallelism and fault tolerance.

Kibana - v7.17.3

Description
Kibana is your window into the Elastic Stack. Specifically, it is a browser-based analytics and search dashboard for Elasticsearch.

MongoDB - v4.4.13

Description
MongoDB® is a relational open source NoSQL database. Easy to use, it stores data in JSON-like documents. Automated scalability and high-performance. Ideal for developing cloud native applications.

Virtuoso - v7.2.7

Description
OpenLink Virtuoso is a next-generation Universal Server that facilitates the development and deployment of a new generation of Enterprise-wide, Internet, Intranet, and Extranet-based solutions, transcending prevalent enterprise challenge areas such as Disparate Databases and Data Sources, Web Service

ECDP – Open-Source portal (3/3)

Service Deployments

Name	Group	Status	Type	Date	Sharing	
Airflow_Demo	DSL0001	ACTIVE	AIRFLOW	Fri Oct 07 2022	PRIVATE	Terminate Open
Jupyterlab_Demo	DSL0001	ACTIVE	JUPYTERLAB	Fri Oct 07 2022	PRIVATE	Terminate Open
Metabase_Demo	DSL0001	ACTIVE	METABASE	Fri Oct 14 2022	PRIVATE	Terminate Open
Superset_Demo	DSL0001	ACTIVE	SUPERSET	Fri Oct 14 2022	PRIVATE	Terminate Open
Postgresql_Demo	DSL0001	ACTIVE	POSTGRESQL	Fri Oct 07 2022	PRIVATE	Terminate Copy

Secrets

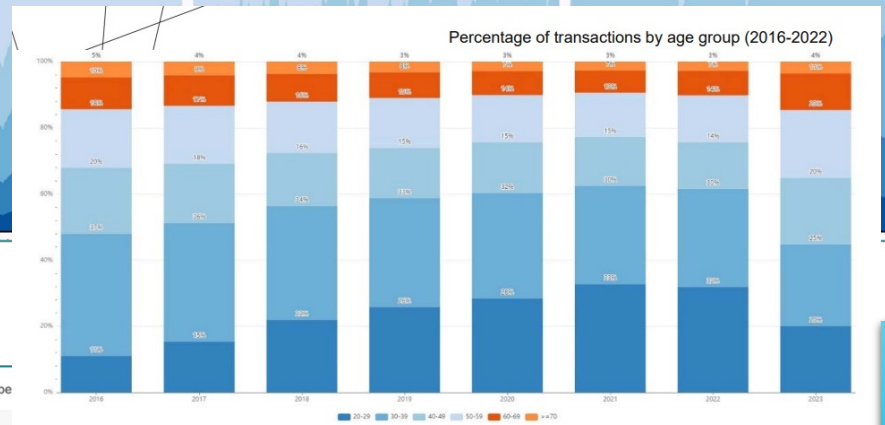
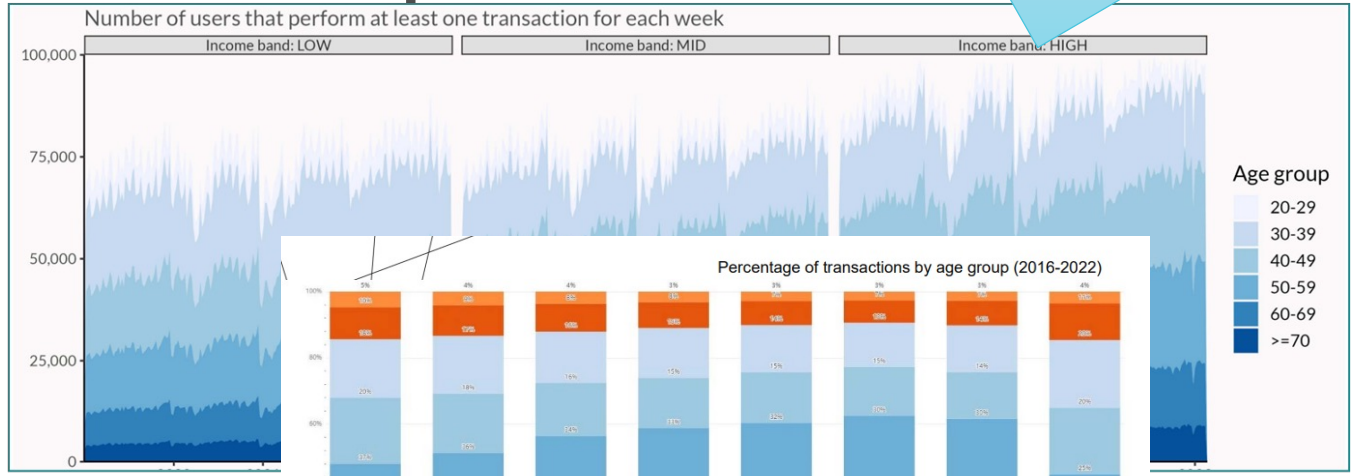
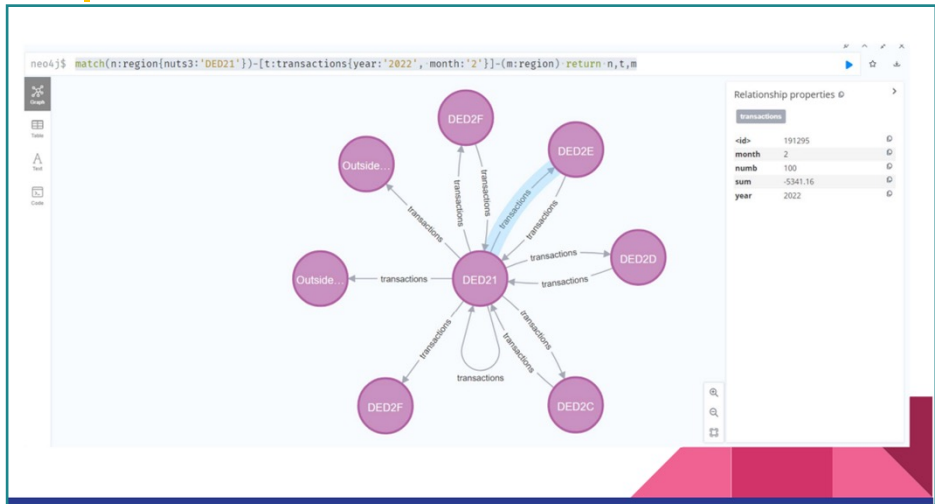
Name	Type	DSL
dsl0001-airflow_demo-airflow-password	PRIVATE	
dsl0001-jupyterlab_demo-password	PRIVATE	
dsl0001-postgresql_demo-admin-password	PRIVATE	
dsl0001-superset_demo-admin-password	PRIVATE	

Storage

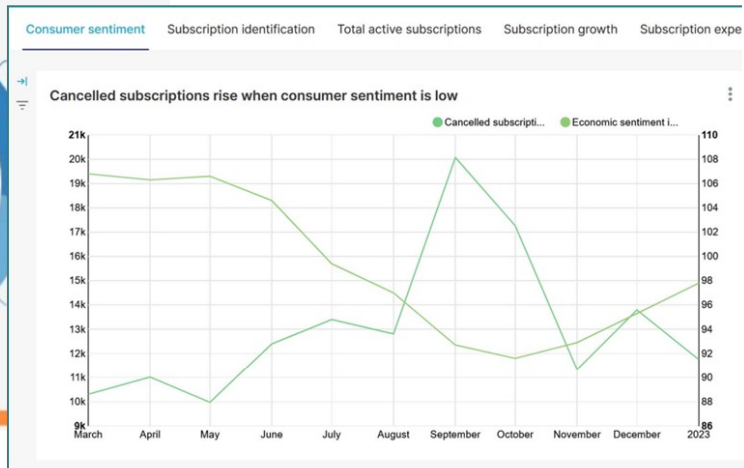
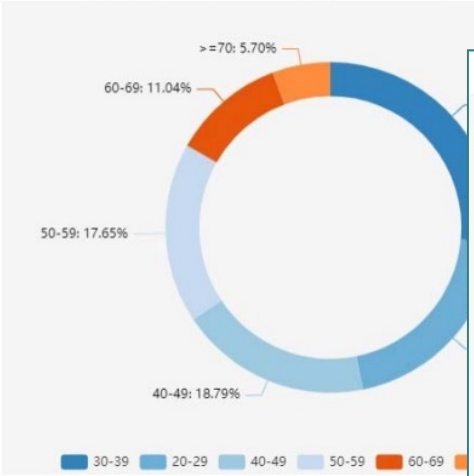
Name	Group	Type
dsl0001-ofs	DSL0001	ofs

ECDP data analysis examples

Consumer spending as early warning for financial crisis

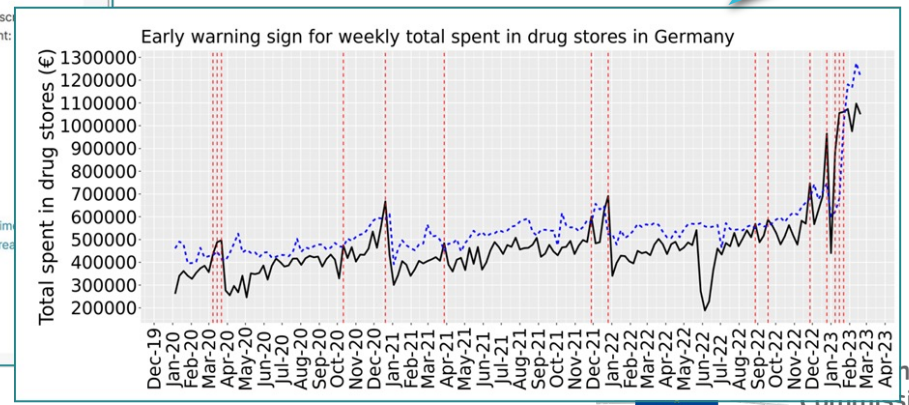


Percentage of registered users by age group



Correlation between cancelled subscriptions and consumer sentiment: **0.60** in 2022, **0.55** since 2016

Pharmacy spending to track disease outbreak



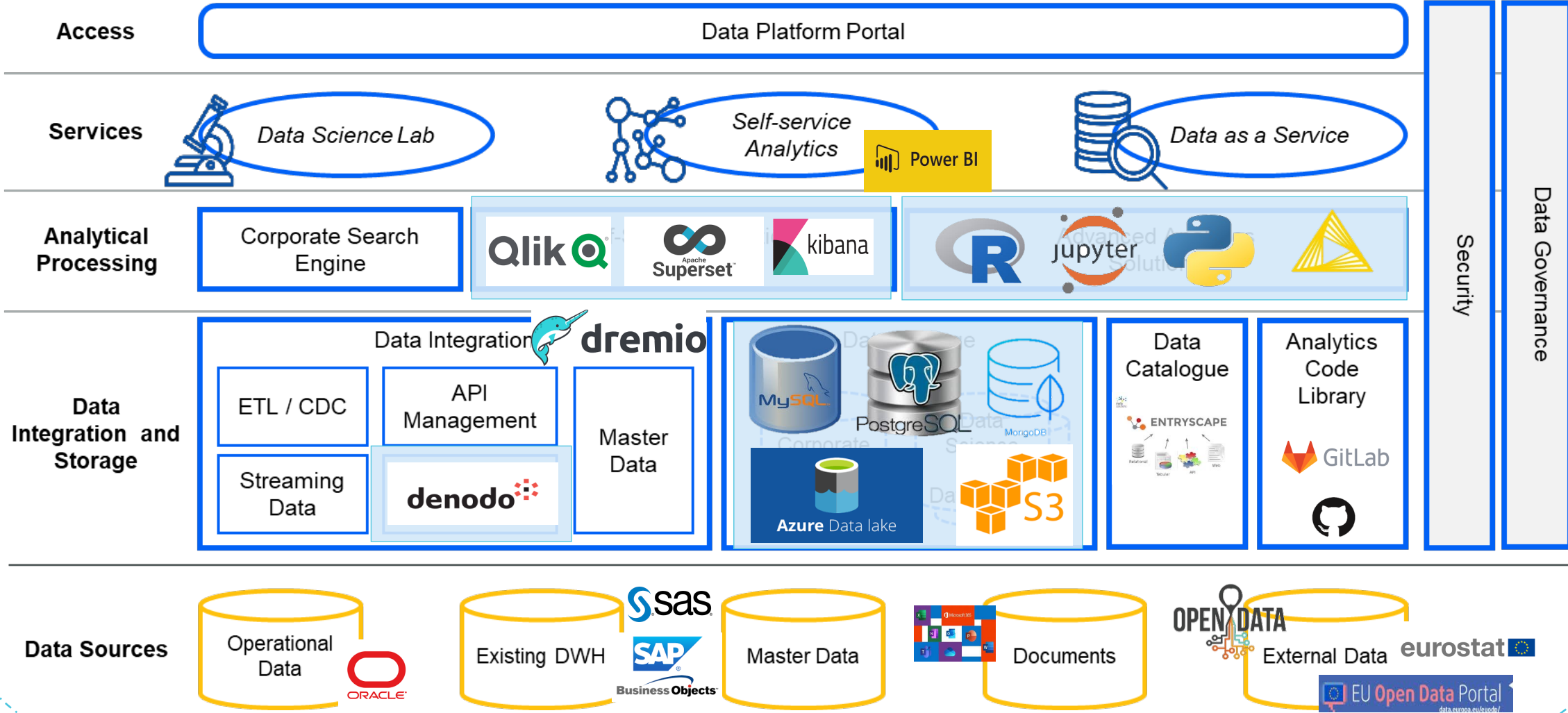
Subscription cancellation to track consumer sentiment

ECDP offerings

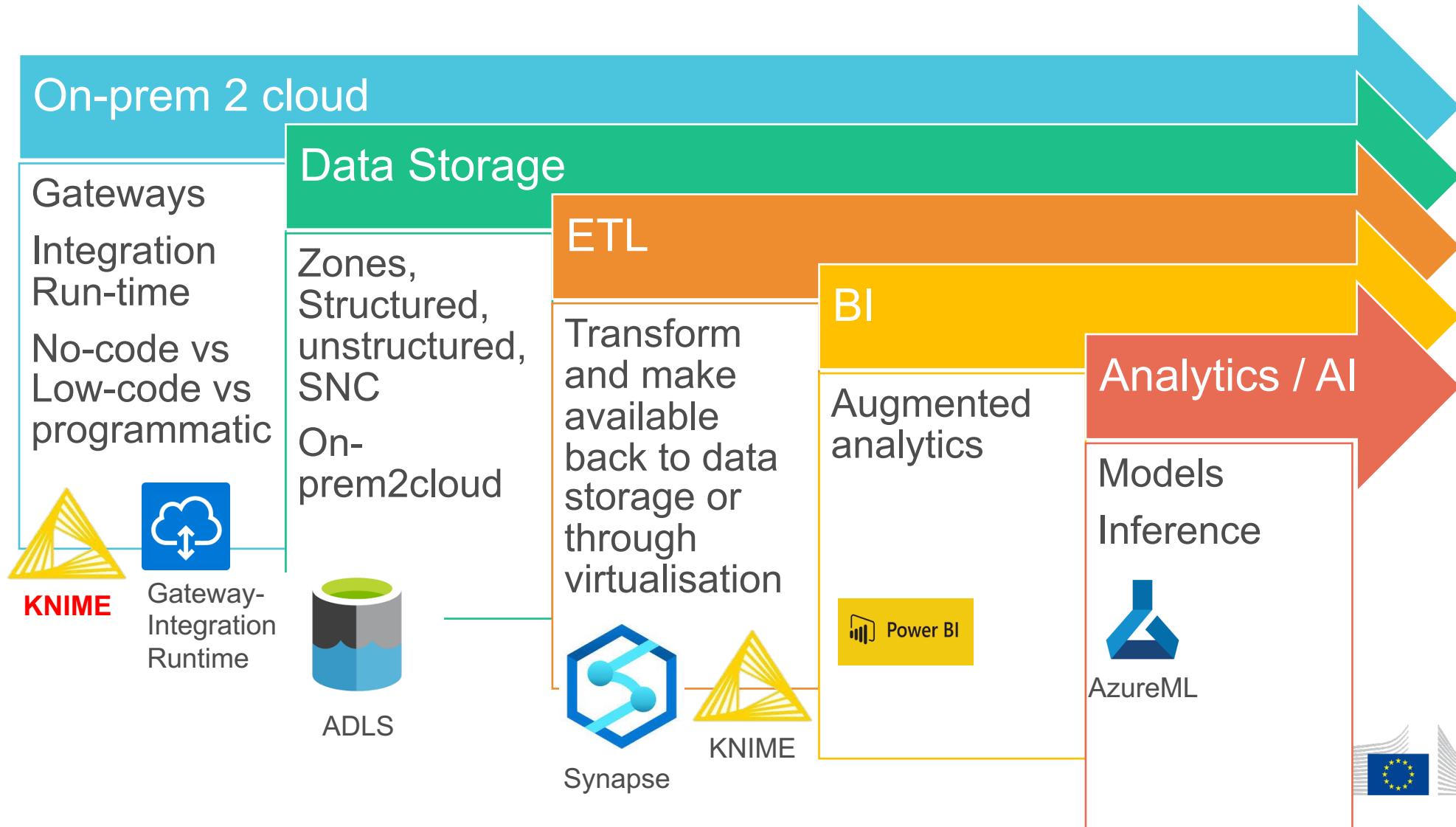


Part of the Data Platform

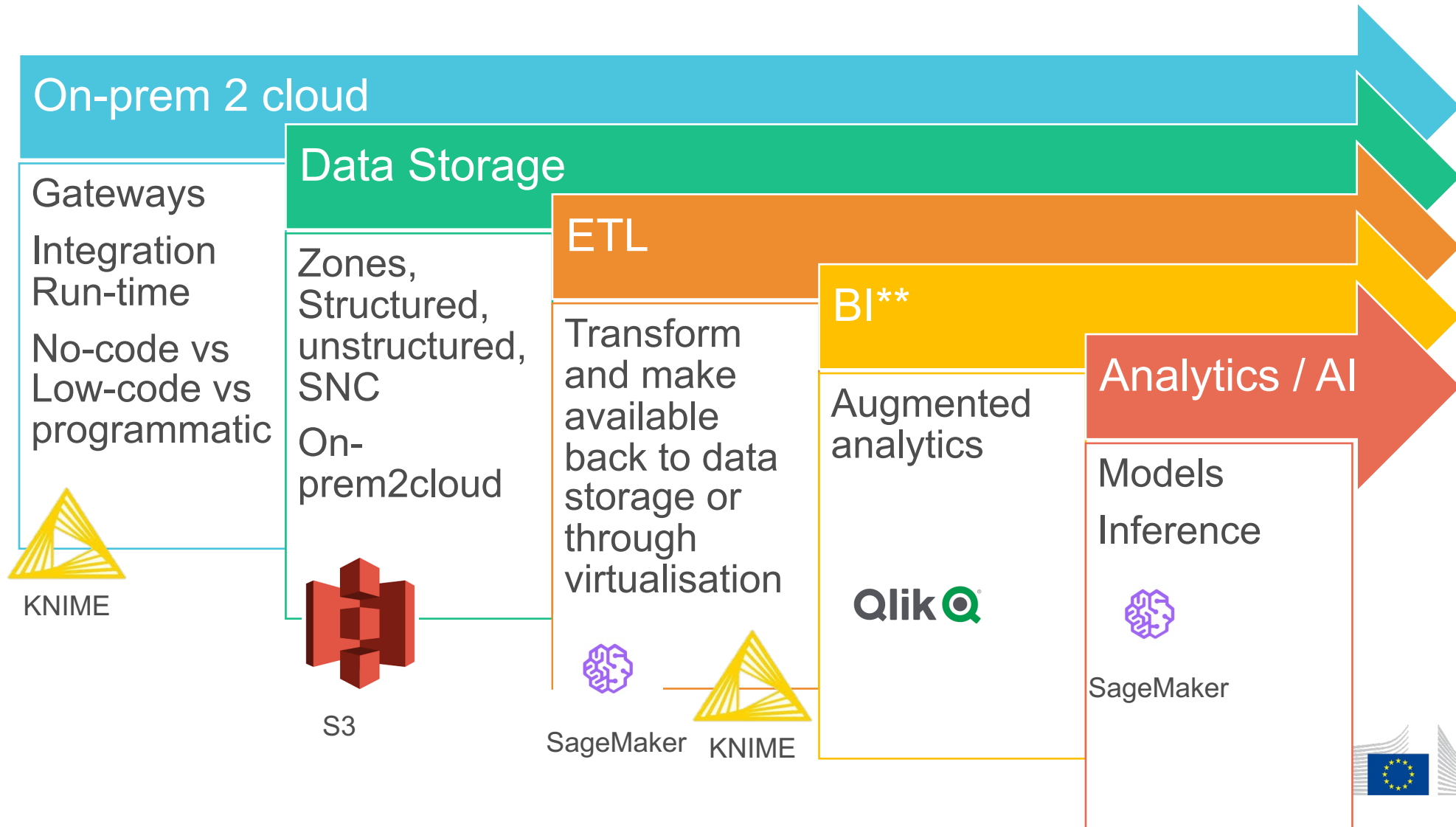
Not Part of the Data Platform



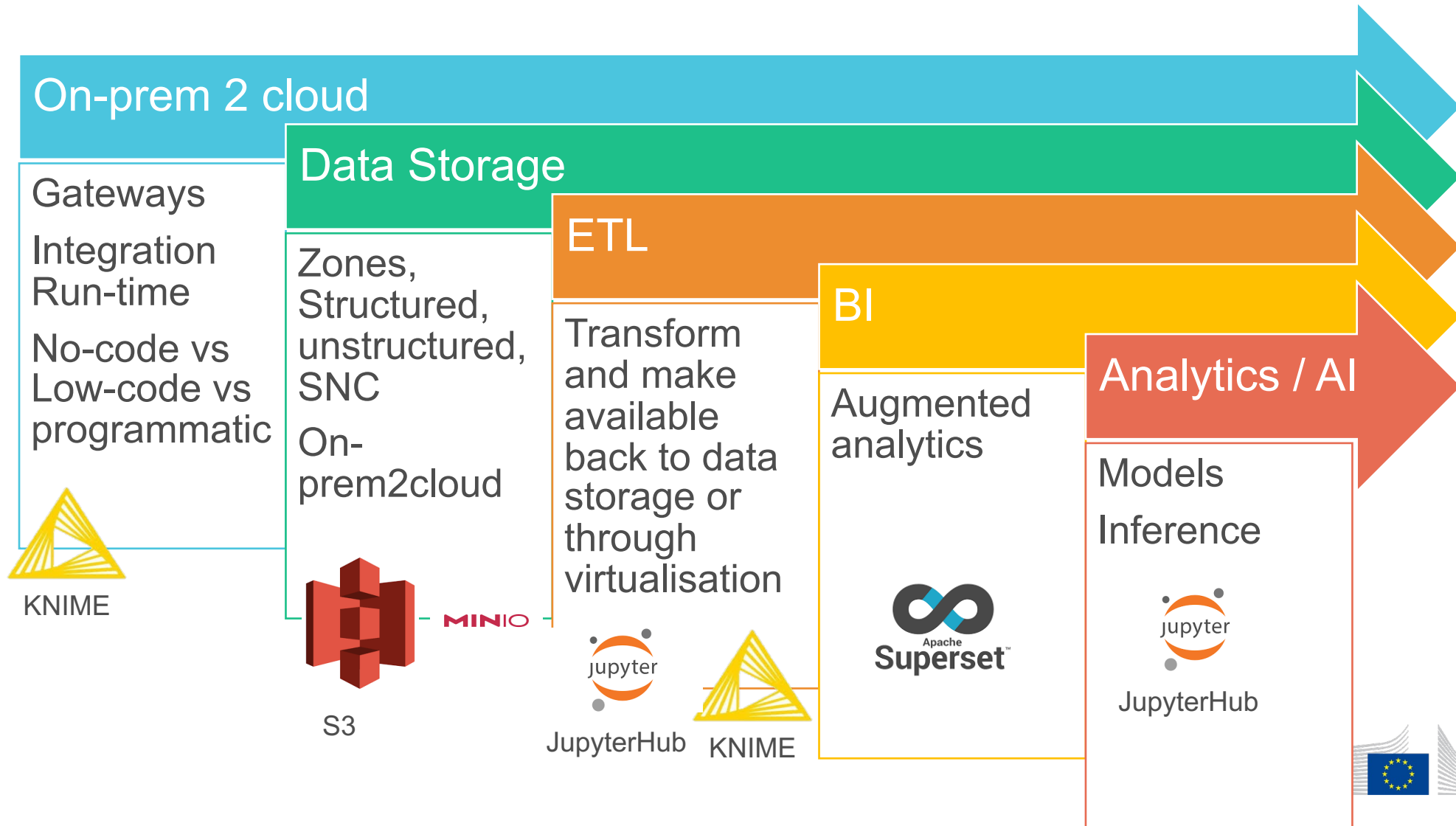
Journey: ECDP @Azure



Journey: ECDP @AWS



Journey: ECDP @Agnostic

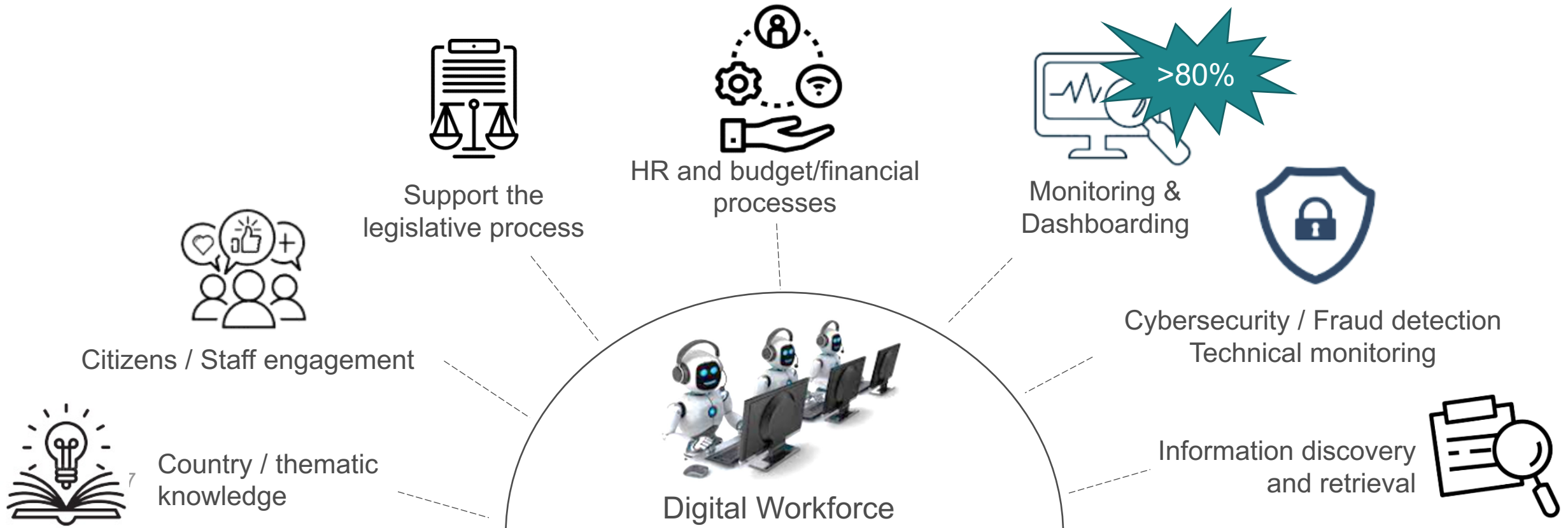


Data Analytics

Identifying and supporting **business priority areas** with services and solutions to tackle specific needs using data analytical tools.



- optimize workload / workforce
- increase quality
- work more effectively



Executive Cockpits



Commissioner Hahn's cockpit

Breaking news

- EU government leaders and heads of state are discussing the bloc's future long-term budget and a multibillion-euro post-coronavirus recovery plan during a video summit aimed at paving the way for a compromise later this summer
- Hungary likely to support EU's economic rescue plan, PM Orban says
- The EU's coronavirus recovery plan: What's at stake?

More

Selected news

HUNGARY PM ORBAN SAYS HUNGARY IS LIKELY TO SUPPORT EU... STATE RADIO

implementation of EU funds yet in 2021

an agreement on the recovery plan for the EU

Cabinet to-do-list

Task Name	Due Date



Ideation Design Implementation Finalised

Status Project

Who Cabinet Hahn, DIGIT

Type Info hub / Analytics / Dashboards

Challenge Tailor made information for EC executives, exploiting internal and external information sources, collaboration solutions, dashboards, news.

Solution Visualisation techniques, collaboration solutions and data dashboarding based on modern BI and interactivity

- Benefits
- Tailor made information
 - Flexibility and agility
 - Channelling different strands of information

eSurveillance – SafetyGate (DG JUST)

Ideation Design **Implementation** Finalised

Status

Project

Who

JUST, DIGIT

Type

Search / Web-scraping / ML / Analytics / Dashboards

Challenge

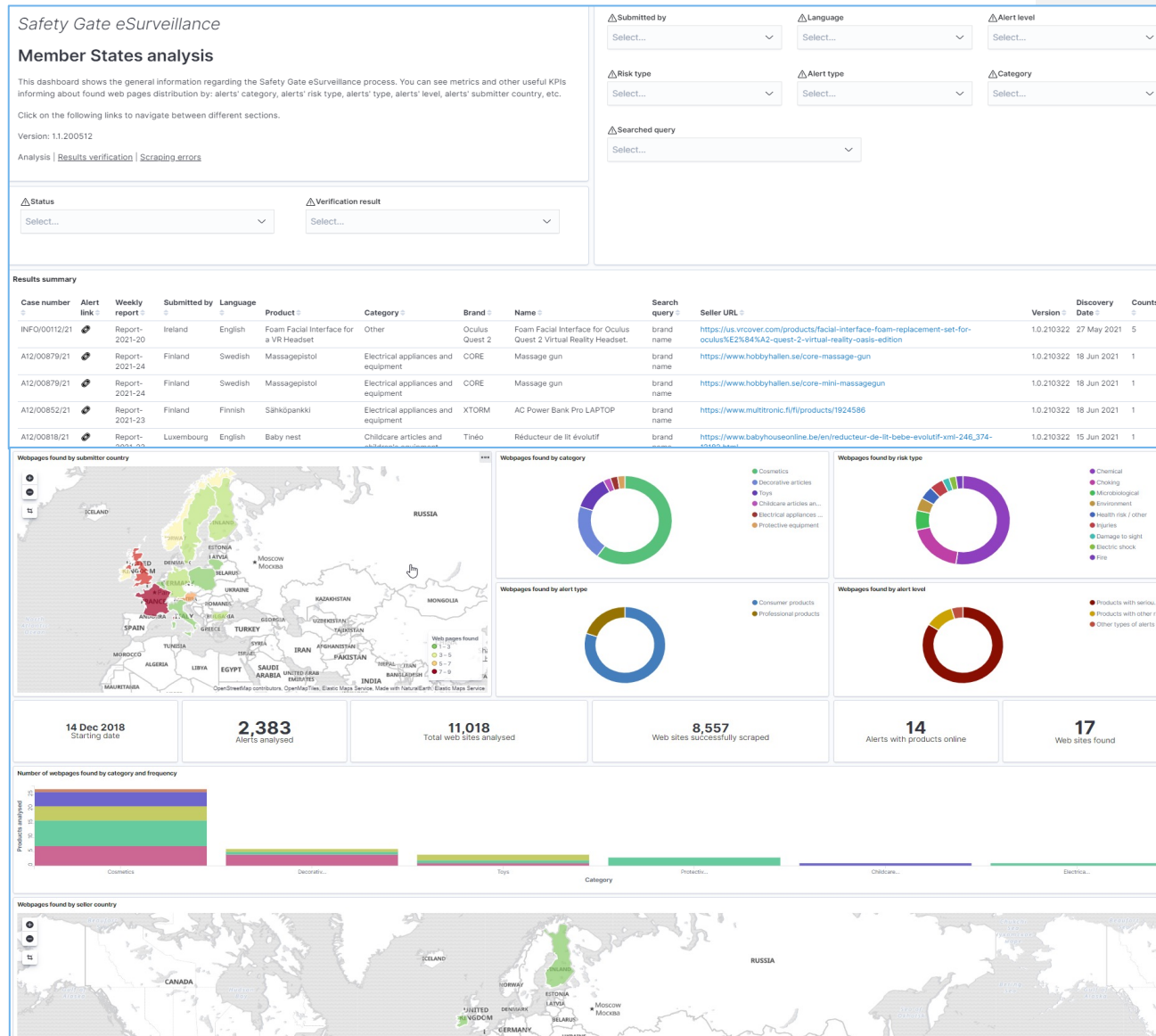
Detection of unauthorised online offers of unsafe products signalled in the EU rapid alert system for dangerous non-food products: Safety Gate.

Solution

Component system built entirely on standard EC technologies for searching, scraping and data analytics

Benefits

- Enhancing of the control capacity of Market Surveillance Authorities (MSAs)
- Harmonisation of fragmented online market surveillance approaches in MS



Thank you!