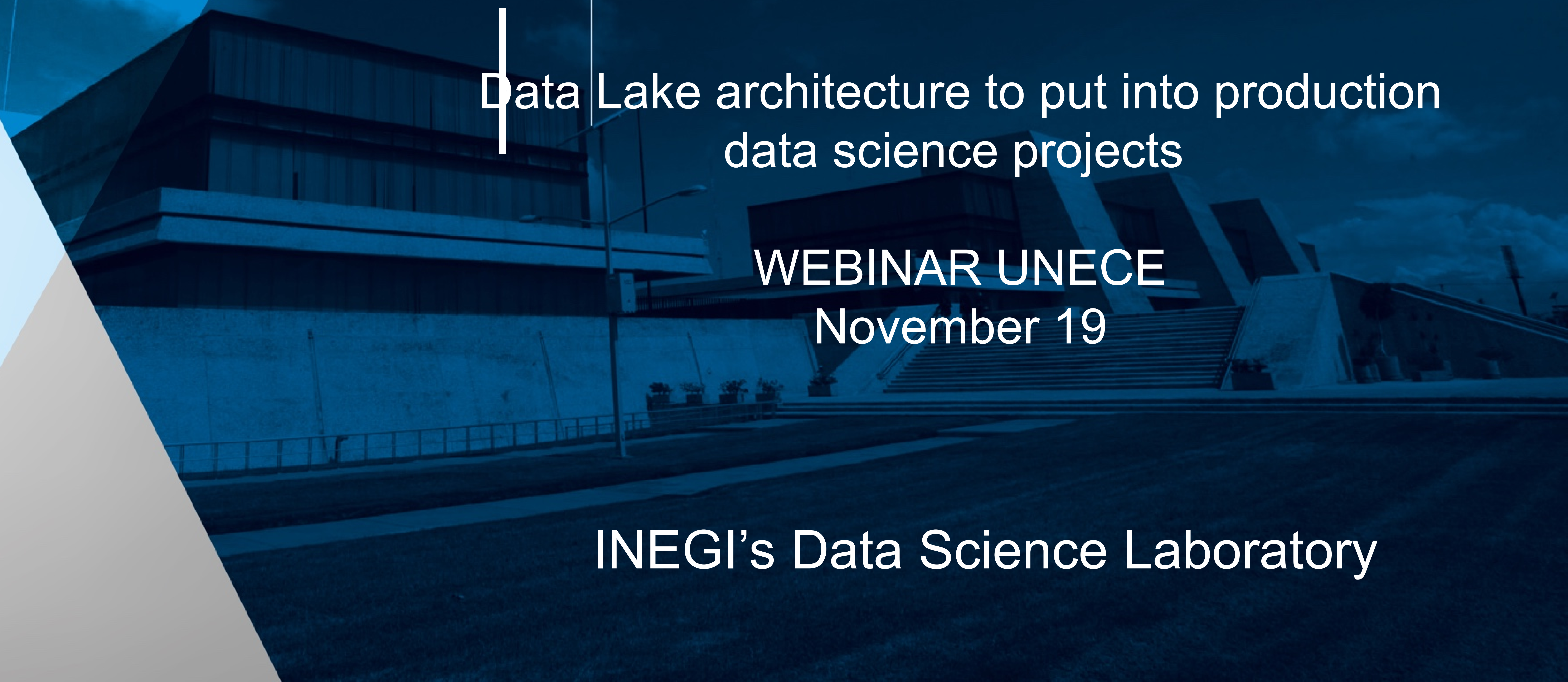
INEGI

Data Lake

Data Lake architecture to put into production data science projects

WEBINAR UNECE
November 19

INEGI’s Data Science Laboratory



What is the purpose of a **data lake**?

- Store all the data that an organization produces.

- Allowing data incorporation with the least possible friction:
  - Data without modeling
    - CSV
  - Semi-structured data
    - JSON
  - Unstructured data
    - Text
    - Images
- Data is accessible for analysis as soon as it is incorporated

What is the purpose of a **data lake?**

Why INEGI needs a data lake?

Prototype

Objective

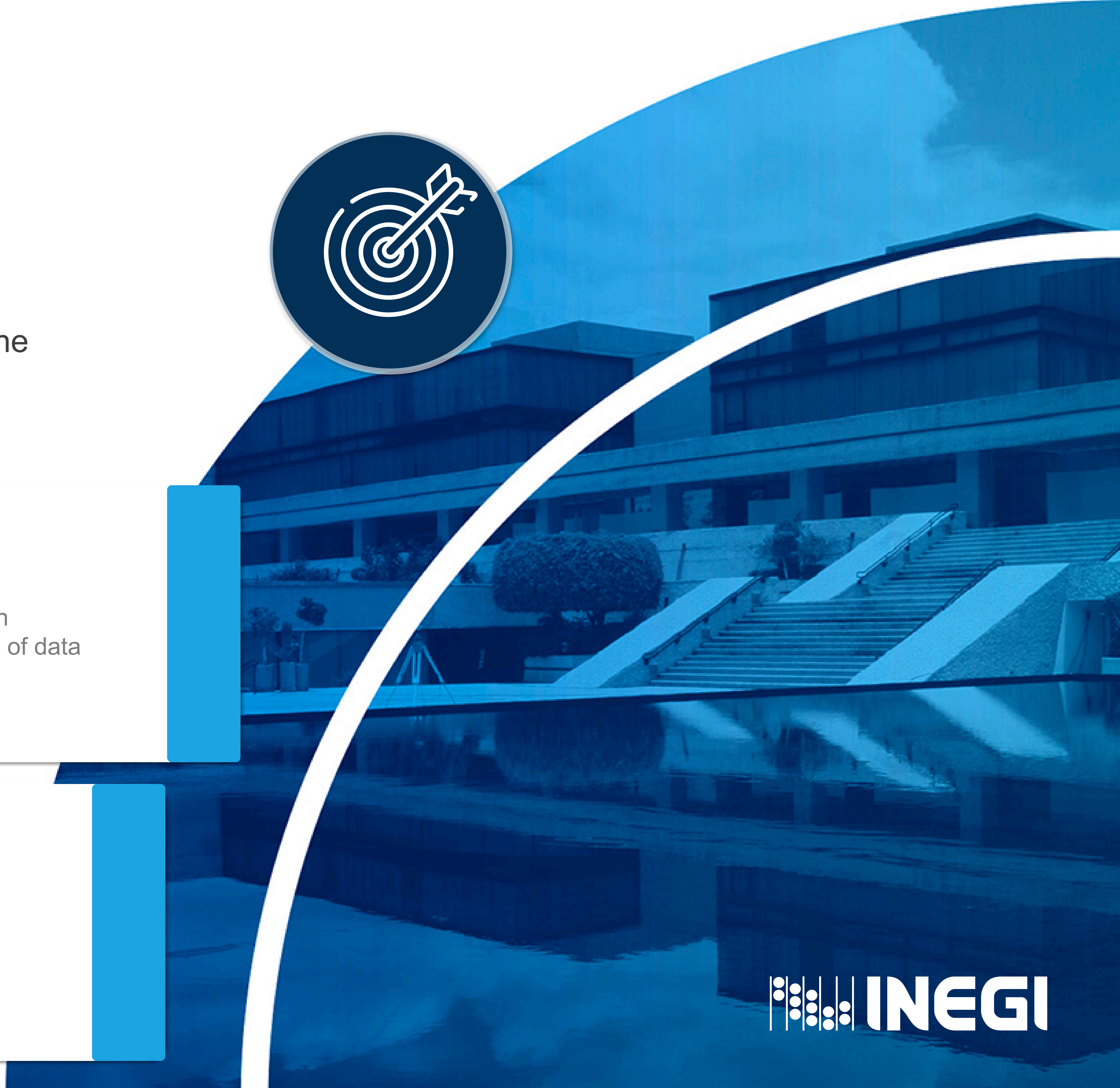
Generate an institutional data lake that allows all the diversity of the data produced by INEGI to "live" there.

For **Data Dissemination**

Connect data dissemination workflows to information deposited in the lake so that there is a single source of data for dissemination.

For analysis (**Laboratory**)

Have the information, both from INEGI and external sources, ready for analysis from a single environment.



INEGI

- To have all the data produced by INEGI in one place.
    - Statistical data
    - Geographical data
        - Cartography and Satellite Images
    - Unstructured data
        - Texts of the searches in INEGIs web site
        - Tweets collected for natural language processing.

- Give data scientists access to data, so they can generate new products.

- To Allow the data silos to talk to each other.

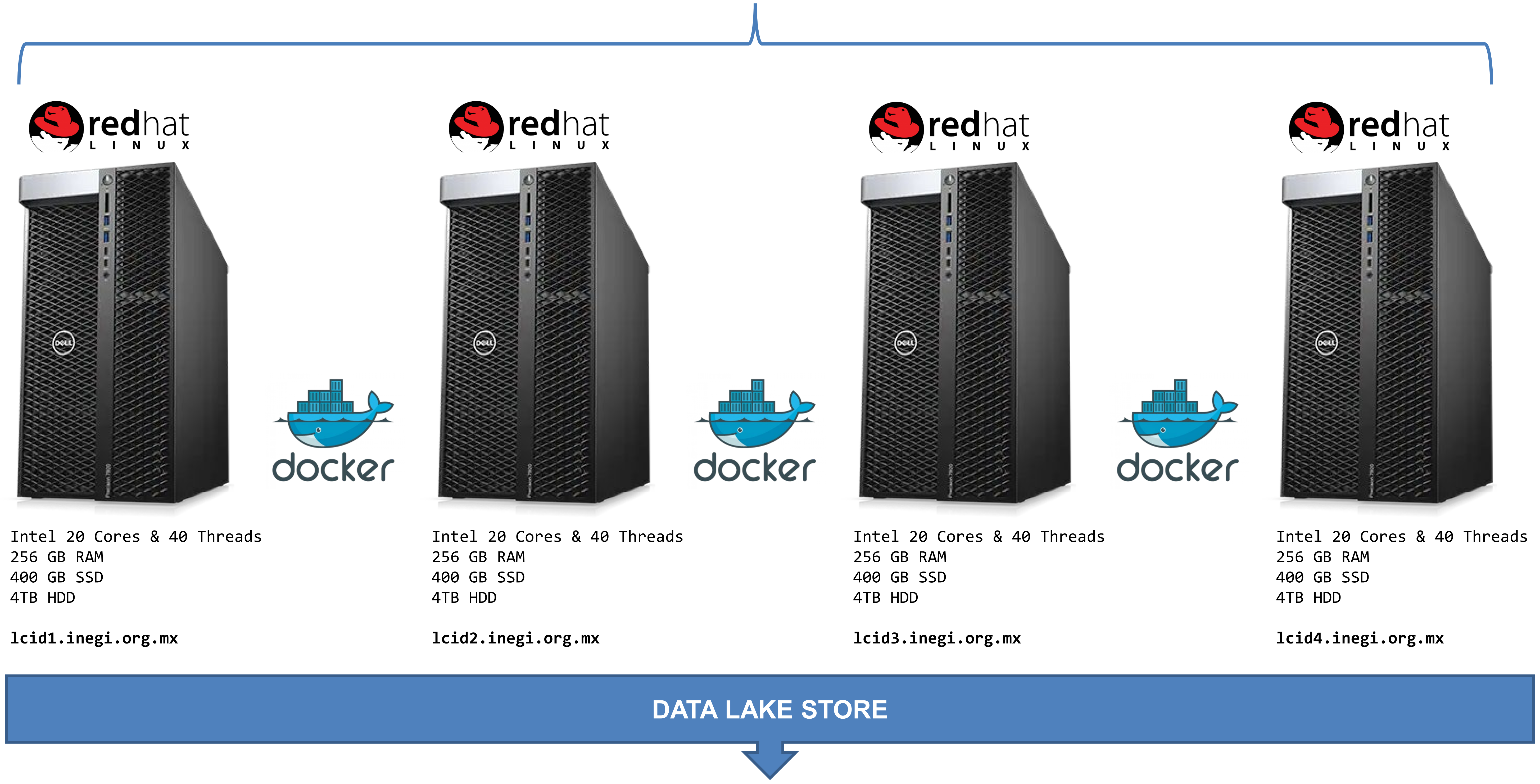
INEGI

Why INEGI needs a data lake?

Prototype Infrastructure

SANDBOX
cluster of workstations

80 Cores & 160 Threads
1 TB RAM



Intel 20 Cores & 40 Threads
256 GB RAM
400 GB SSD
4TB HDD

lcid1.inegi.org.mx

Intel 20 Cores & 40 Threads
256 GB RAM
400 GB SSD
4TB HDD

lcid2.inegi.org.mx

Intel 20 Cores & 40 Threads
256 GB RAM
400 GB SSD
4TB HDD

lcid3.inegi.org.mx

Intel 20 Cores & 40 Threads
256 GB RAM
400 GB SSD
4TB HDD

lcid4.inegi.org.mx

DATA LAKE STORE

nas-inegi.org.mx

The use cases

Technology Landscape

INEGI Open Data

CSV

Microsoft SQL Server

ORACLE DATABASE

TIFF

Data Ingestion

Data Storage

Data Virtualization

Data Integration

Data Analytics

Data Visualization

```
wget
ftp
s3
```

Linux Console

MINIO

trino

Formerly Presto

jupyter

jupyter

Apache Superset

Use Case 1
Data Science Lab

Java JDBC

S3 Buckets

dbt

dbt

Data-Driven Documents

Use Case 2
Data Publishing

Apache Kylin™

Apache Kylin™

jupyter

ESRI ArcGIS

Use Case 3
Data Warehouse

NAS

Microsoft Power BI

```
wget
ftp
s3
```

Linux Console

PostgreSQL

OPEN DATA CUBE

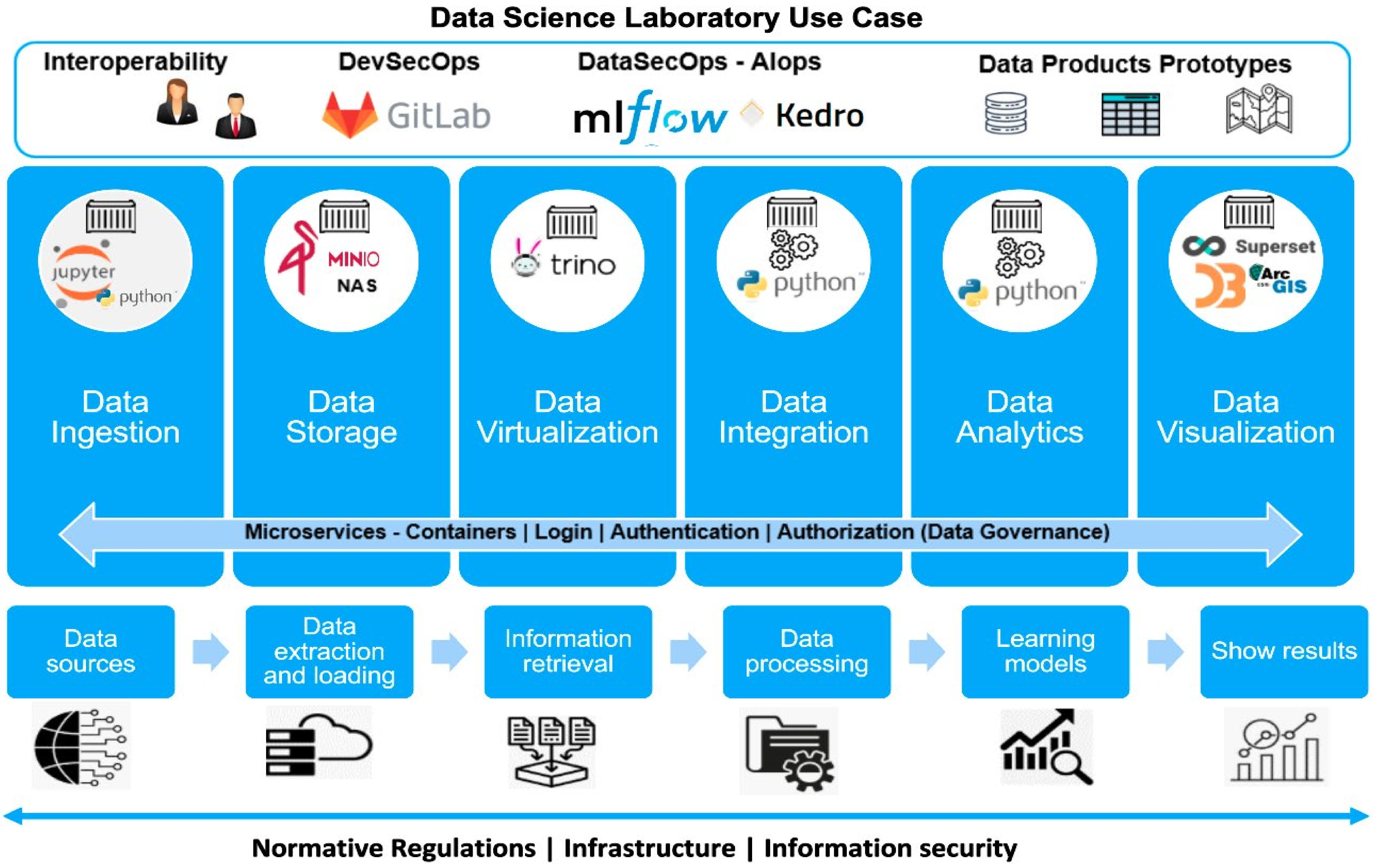
QGIS

Use Case 4
Open Data Cube

docker

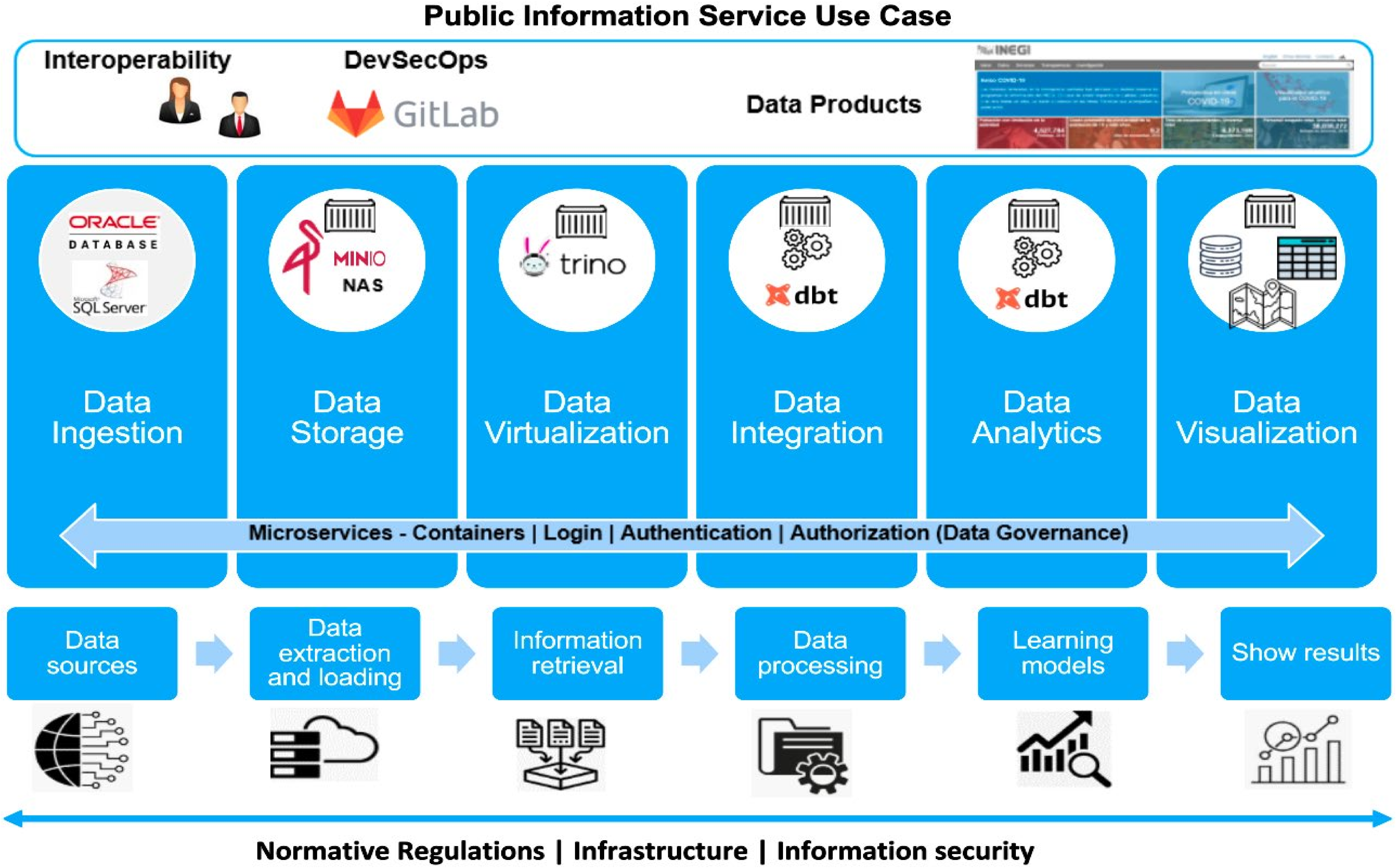
GitLab

DataSecOps



People involved in the use case

9



Public Information Service Use Case

Interoperability

DevSecOps

GitLab

Data Products

Data Ingestion

Data Storage

Data Virtualization

Data Integration

Data Analytics

Data Visualization

Microservices - Containers | Login | Authentication | Authorization (Data Governance)

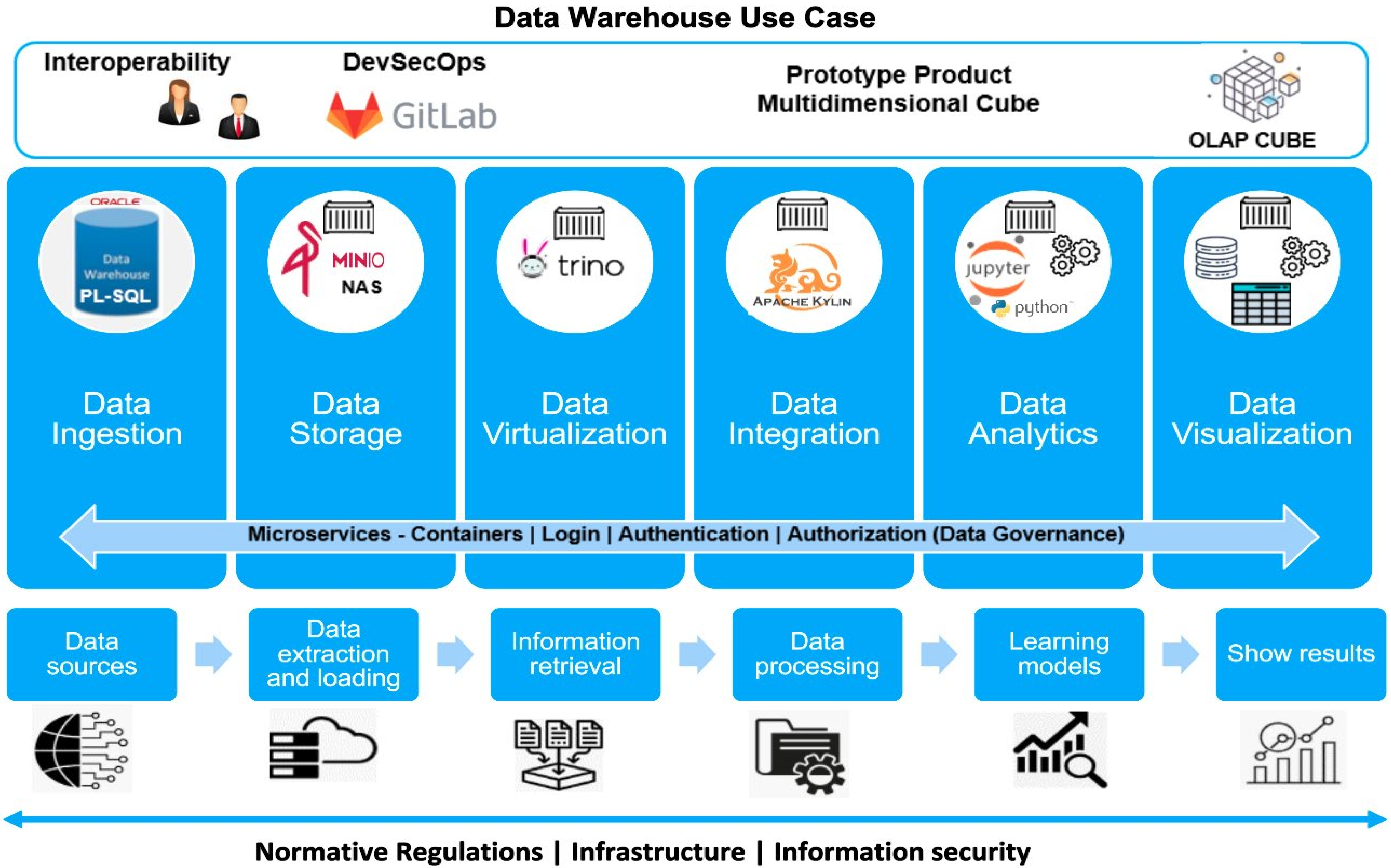
Data sources → Data extraction and loading → Information retrieval → Data processing → Learning models → Show results

Normative Regulations | Infrastructure | Information security

People involved in the use case

8

INEGI | Public Information Service Use Case

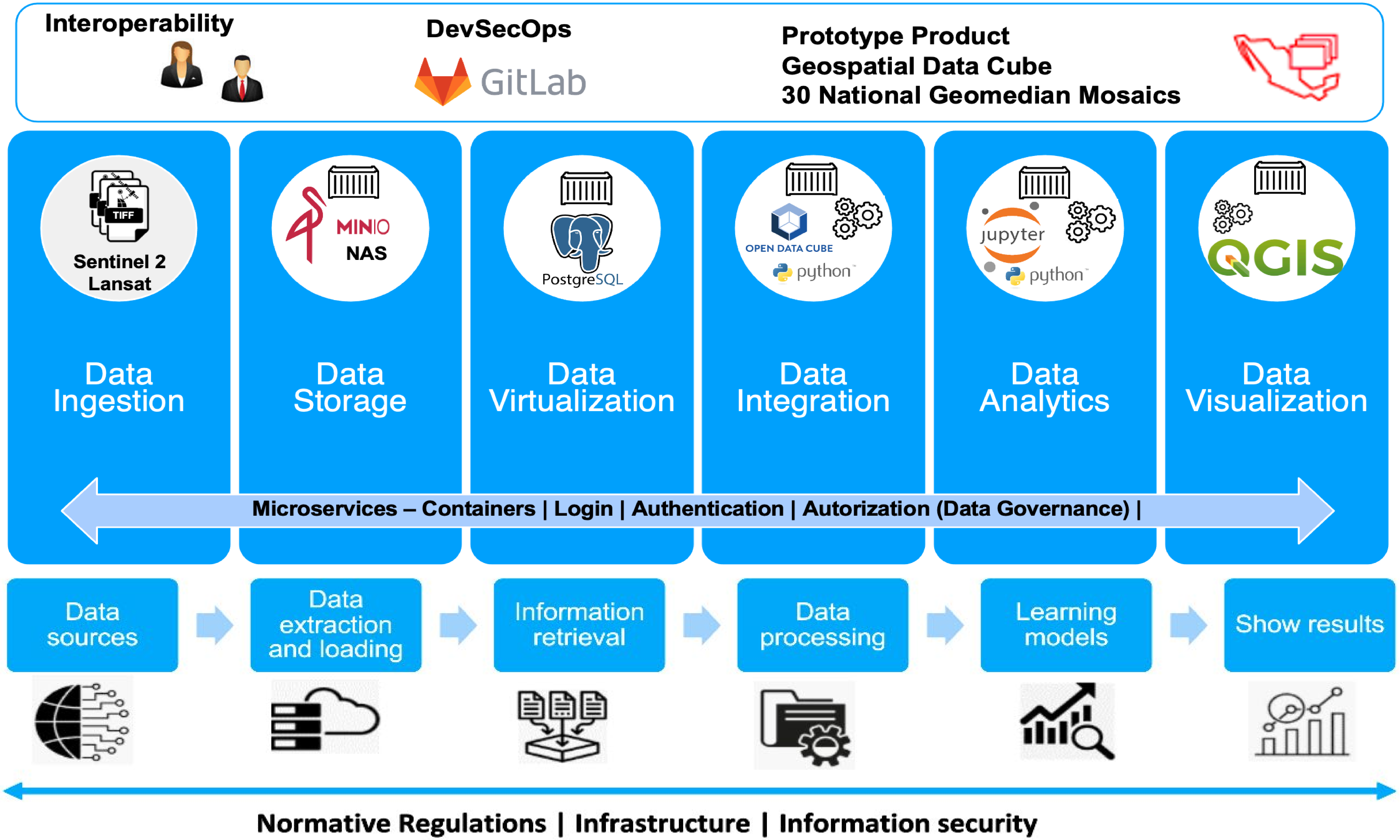


People involved in the use case

4

INEGI | Data Warehouse Use Case

Data Warehouse Use Case



Interoperability

DevSecOps

GitLab

Prototype Product
Geospatial Data Cube
30 National Geomedian Mosaics

Sentinel 2 Lansat

Data Ingestion

MINIO NAS

Data Storage

PostgreSQL

Data Virtualization

OPEN DATA CUBE python

Data Integration

jupyter python

Data Analytics

QGIS

Data Visualization

Microservices – Containers | Login | Authentication | Autorization (Data Governance) |

Data sources → Data extraction and loading → Information retrieval → Data processing → Learning models → Show results

Normative Regulations | Infrastructure | Information security

People involved in the use case

4

INEGI | Geospatial Data Cube Use Case

Next steps

- It is estimated that in December the infrastructure of the Data Science Laboratory will be updated

Server A

224 Threads
4X Nvidia Tesla V100
1 TB RAM
15 TB Local Storage

Server B

224 Threads
1 TB RAM
15 TB Local Storage

- Work in permission management and lake administration roles.
- Definition of elements to make this prototype productive:
  - Security
  - Infrastructure
  - User attention
- Improve data governance
- Capacity building for a larger audience within INEGI
- Explore alternatives for incorporating metadata and data lineage.

GRACIAS