Use Cases applied to a Data Lake Prototype in a National Statistical Office

Data Lake architecture to put into production data science projects

WEBINAR UNECE
November 19

INEGI’s Data Science Laboratory
What is a data lake?
A bucket where you can pour any type of digital object
● It stores all the data that an organization produces.
● It allows 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
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 (DS Laboratory)

Have a unified repository of information, both from INEGI and external sources, ready to be analyzed from a single environment.
• 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 INEGI's web site
    • Tweets collected for natural language processing

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

• To Allow the data silos to talk to each other.
What infrastructure is used for the prototype?
Prototype Infrastructure

SANDBOX

cluster of workstations

80 Cores & 160 Threads
1 TB RAM

INTEL 20 Cores
256 GB RAM
400 GB SSD
4TB HDD

lcid1.inegi.org.mx
lcid2.inegi.org.mx
lcid3.inegi.org.mx
lcid4.inegi.org.mx

DATA LAKE STORE

nas-inegi.org.mx
The use cases
People involved in the use case

9
Data Warehouse Use Case

People involved in the use case: 4
People involved in the use case 4
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
● Explore alternatives for incorporating metadata and data lineage.

● 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
GRACIAS