Plataforma tecnológica para analítica, ciencia de grandes volúmenes de datos estadísticos, geográficos con privacidad.

Mayo 2025

Red de Transmisión del Conocimiento (RTC),
Instituto Nacional de Estadísticas

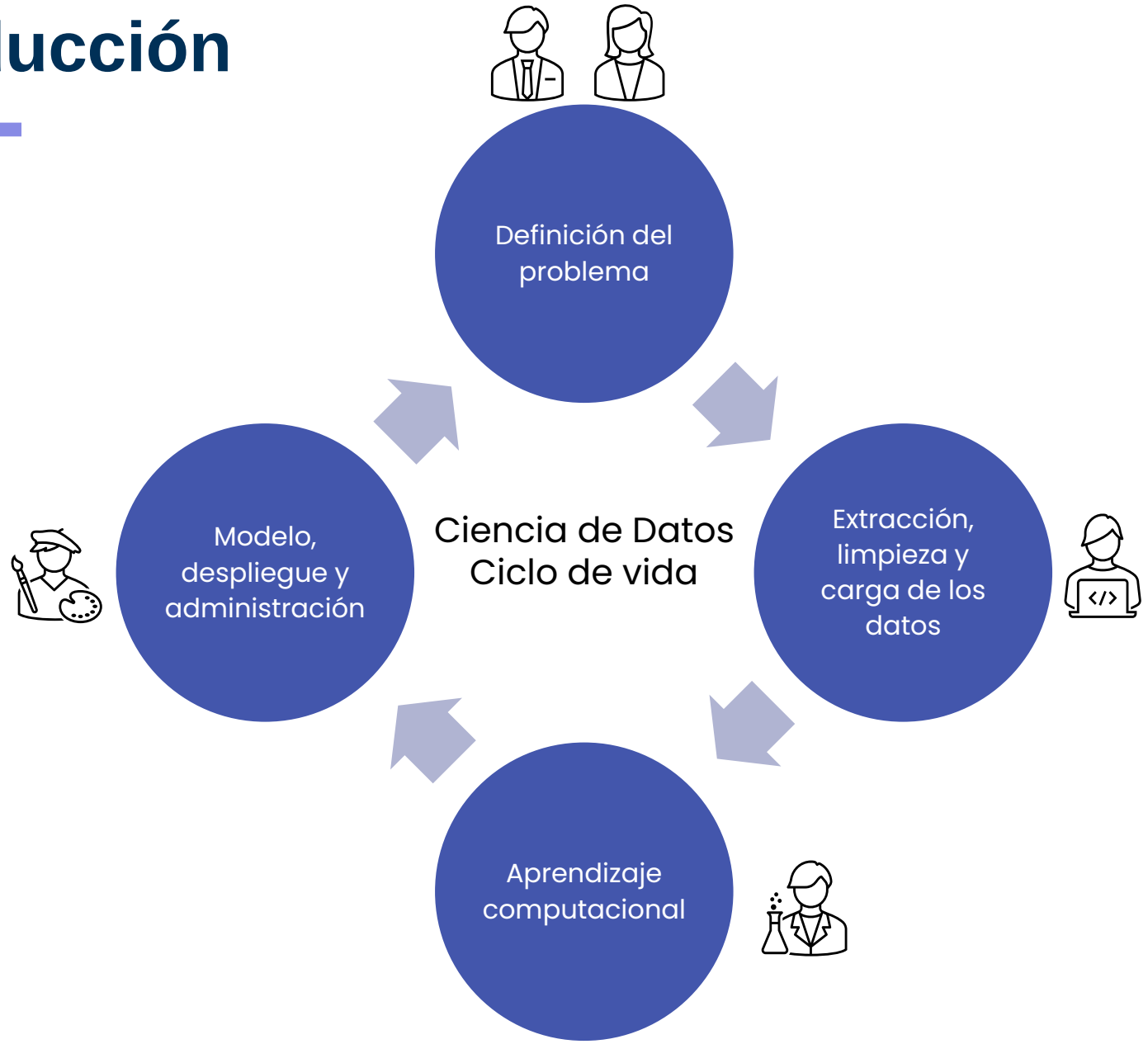
INEGI

inegi.org.mx

Objetivo

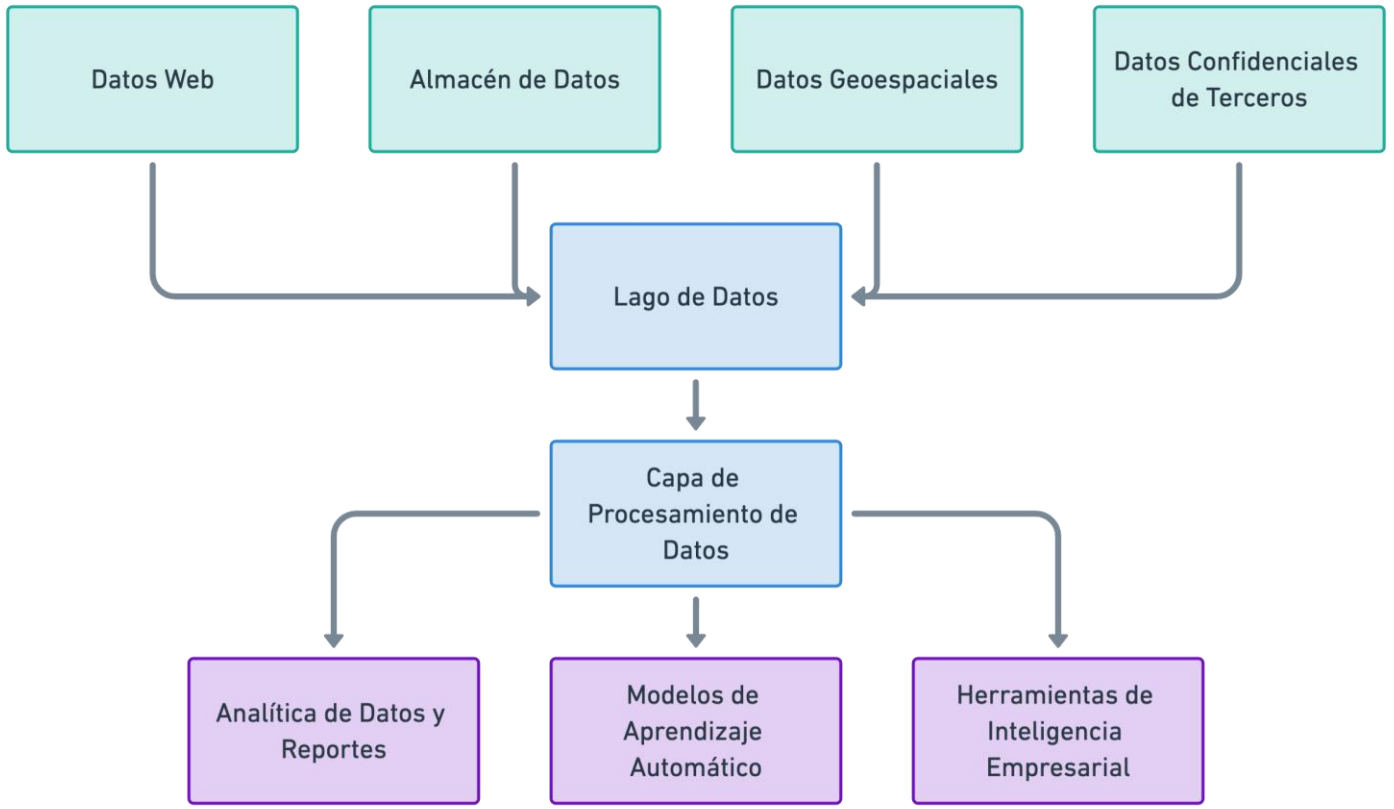
Intercambiar conocimiento, experiencia y buenas prácticas en torno al desarrollo e implementación de plataformas tecnológicas para la analítica avanzada de grandes volúmenes de datos estadísticos y geográficos, con un enfoque en la protección de la privacidad y la seguridad de la información.

Introducción



Descripción

La plataforma integra capacidades de procesamiento, análisis y gestión de datos, con el fin de fortalecer los procesos de investigación, generación de información estadística y geográfica, incorporación de métodos modernos de producción, y el acceso seguro a microdatos, promoviendo la colaboración institucional, la innovación metodológica y el cumplimiento de los objetivos estratégicos del INEGI y del SNIEG.



¿Qué puedo hacer en la plataforma?

- **Ingesta automatizada de datos multifuente**: Implementación de procesos para recolectar grandes volúmenes de datos con validación y estandarización automática.

- **Limpieza y transformación automática de datos:** Uso de herramientas especializadas para depurar, transformar y estructurar datos de manera continua, asegurando su calidad y consistencia.

- **Geocodificación y enriquecimiento geoespacial automatizado**: Aplicación de procesos automáticos para asociar coordenadas a registros, integrar capas geográficas y cruzar datos estadísticos con territoriales.

Referencia: E. A. Villaseñor García et al., "Data Lake Strategy for Data Science Workflows," 2022 11th International Conference On Software Process Improvement (CIMPS), Acapulco, Guerrero, Mexico, 2022, pp. 219-223, DOI: 10.1109/CIMPS57786.2022.10035694.

¿Qué puedo hacer en la plataforma?

- **Generación de tabulados y tableros de control dinámicos**: Automatización de reportes estadísticos y visualizaciones geográficas mediante herramientas especializadas, con actualizaciones basadas en eventos o programaciones periódicas.

- **Análisis exploratorio y descriptivo automatizado**: Generación de códigos especializados que ejecuten es estadísticos, correlaciones, análisis de dispersión, detección de valores atípicos de forma automática al integrarse nuevos datos.

- **Entrenamiento automático de modelos estadísticos y aprendizaje automático**: Implementación de flujos para tareas como predicción, clasificación, segmentación o imputación de datos, utilizando bibliotecas especializadas.

Referencia: E. A. Villaseñor García et al., "Data Lake Strategy for Data Science Workflows," 2022 11th International Conference On Software Process Improvement (CIMPS), Acapulco, Guerrero, Mexico, 2022, pp. 219-223, DOI: 10.1109/CIMPS57786.2022.10035694.

¿Qué puedo hacer en la plataforma?

- **Monitoreo y versionado de modelos analíticos**: Integración de herramientas especializadas para registrar experimentos, validar modelos de forma continua y gestionar versiones con métricas reproducibles.

- **Análisis de series de tiempo y detección de cambios**: Automatización de procesos para el análisis temporal de datos estadísticos y geográficos, incluyendo pronósticos, detección de rupturas estructurales y eventos anómalos con modelos estadísticos.

- **Publicación automatizada de datos abiertos y microdatos sintéticos**: Automatización de procesos para anonimización, generación de datos sintéticos con privacidad diferencial y publicación en portales de datos abiertos con metadatos normalizados.

Referencia: E. A. Villaseñor García et al., "Data Lake Strategy for Data Science Workflows," 2022 11th International Conference On Software Process Improvement (CIMPS), Acapulco, Guerrero, Mexico, 2022, pp. 219-223, DOI: 10.1109/CIMPS57786.2022.10035694.

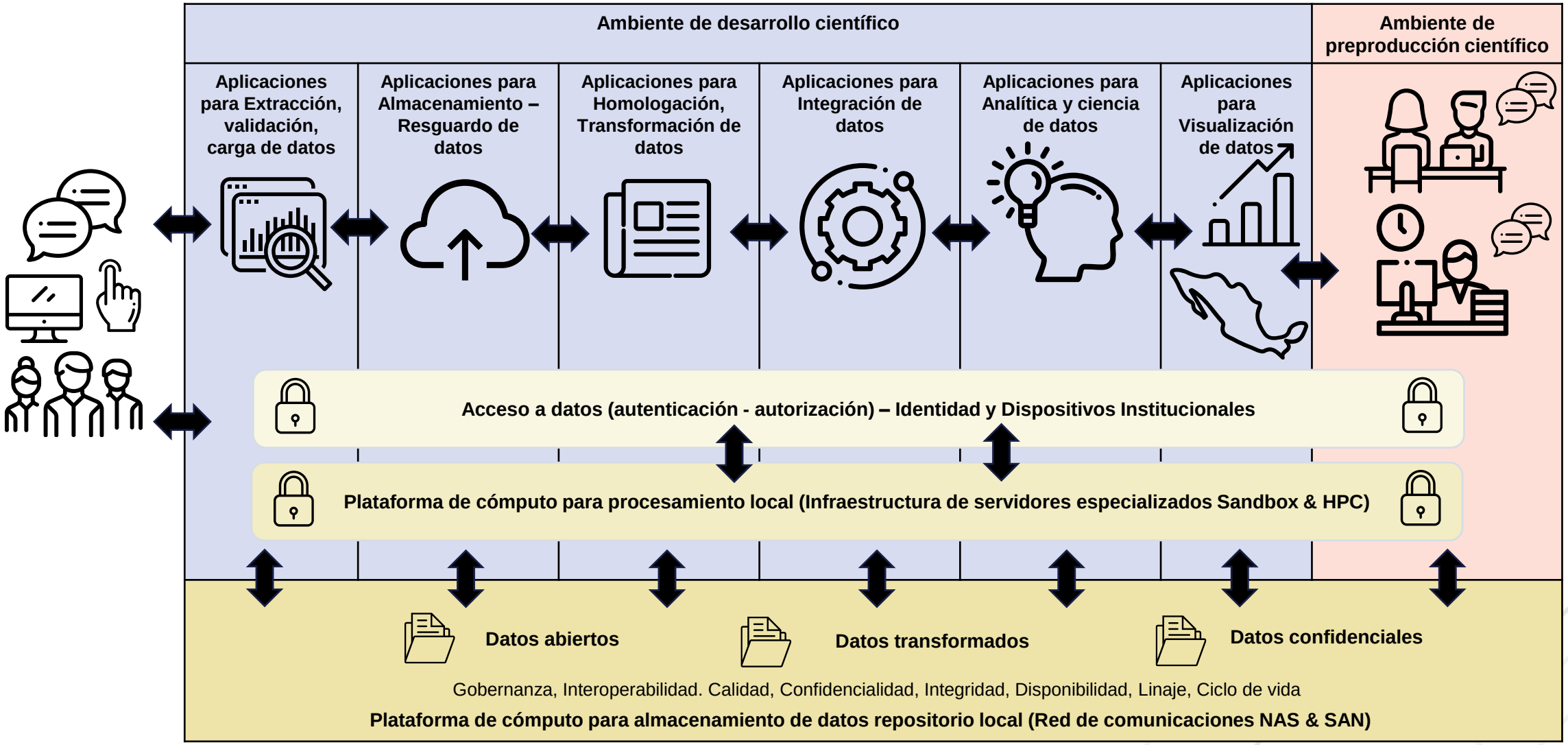
¿Qué puedo hacer en la plataforma?

Incorporación automatizada con grandes modelos de lenguaje, **combinados con técnicas de recuperación aumentada**, que permiten potenciar los procesos de analítica y ciencia de grandes volúmenes de datos estadísticos y geográficos al permitir la consulta inteligente de información, la automatización de análisis descriptivos e interpretativos, y la asistencia contextual en tareas técnicas complejas, mejorando así la eficiencia, escalabilidad y accesibilidad del conocimiento generado, por ejemplo:

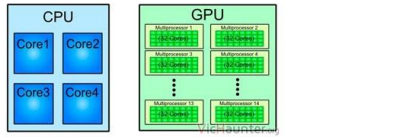
- Algoritmos entrenados especializados con conocimiento Institucional.
- Automatización avanzada de análisis e interpretación contextual de grandes volúmenes de datos.
- Aportación en la calidad del servicio para usuarios especializados.
- Asistencia técnica automatizada en procesos analíticos.
- Escalabilidad y personalización en la generación de conocimiento.

Referencia: E. A. Villaseñor García et al., "Data Lake Strategy for Data Science Workflows," 2022 11th International Conference On Software Process Improvement (CIMPS), Acapulco, Guerrero, Mexico, 2022, pp. 219-223, DOI: 10.1109/CIMPS57786.2022.10035694.

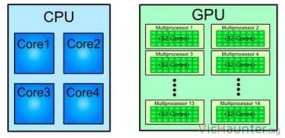
¿Cómo utilizar la plataforma?



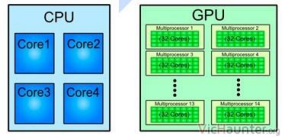
Infraestructura



Cluster and Grid Sandbox-ITo (Areneros Preproducción Capacitación – 4 nodos) Procesamiento 160 cores en cpu´s, Memoria Ram 1.5 TB, Almacenamiento 16 TB



Cluster and Grid Sandbox (LLM & SML – 4 nodos) Procesamiento 96 cores en cpu´s, Memoria Ram 2 TB, Almacenamiento 48 TB, GPU´s 1,152 núcleos Tensor (4 NVIDIA RTX 6000 ADA - 48 Gb).



Cluster and Grid HPC (High Performance Computing), Procesamiento 448 cores en cpu´s y 4 gpu´s [Tensor Core + TeraFlops], Memoria Ram 3 TB, Almacenamiento 30 TB, GPU´s 2,560 núcleos Tensor.

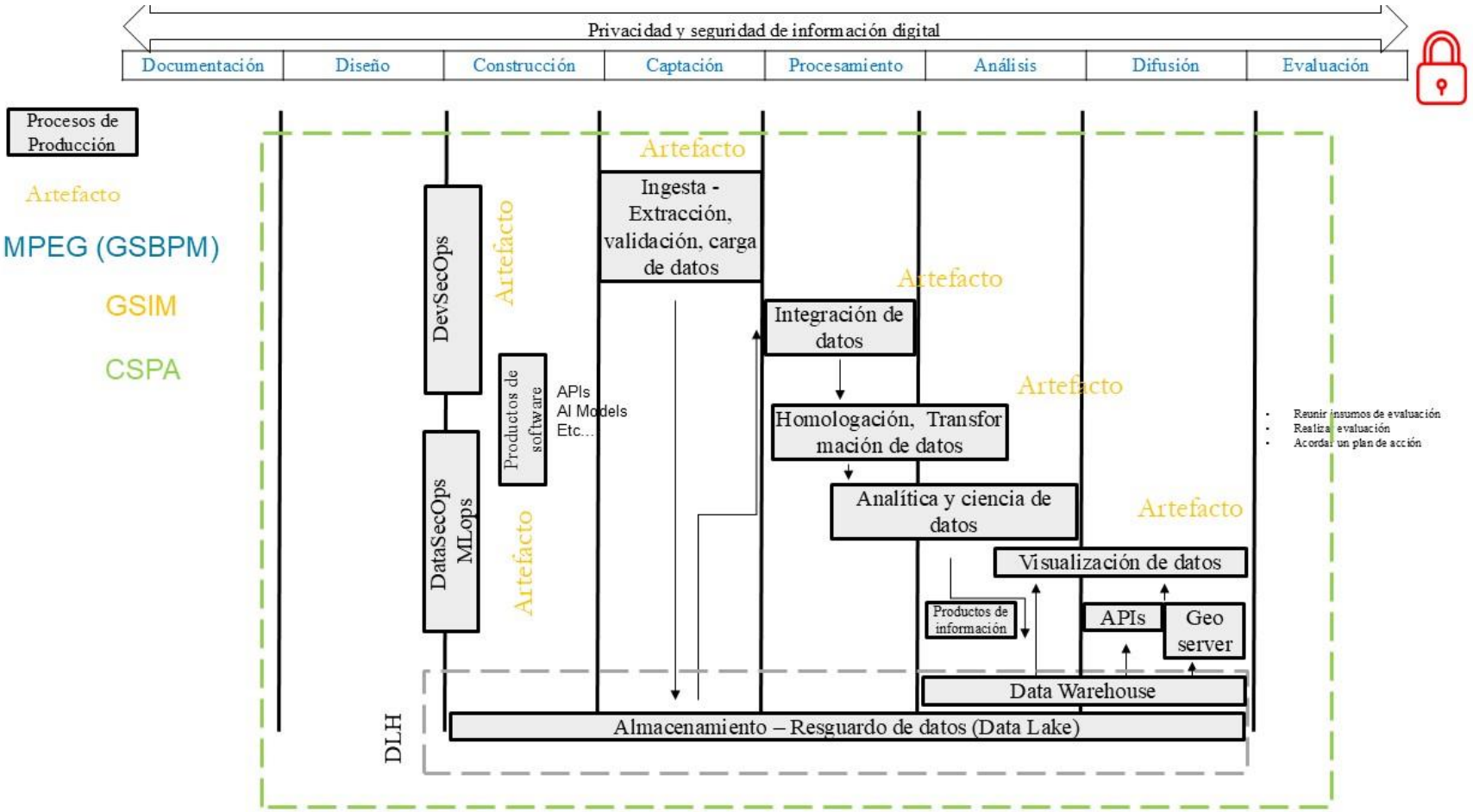
Grid Storage Raid (Data Lake | Lago de Datos)

NAS (Network Attached Storage)
Almacenamiento 50 TB

- Datos abiertos
- Datos transformados
- Datos confidenciales

SAN (Storage Area Network)
Almacenamiento 20 TB

Normatividad



Casos de uso para INE

Colaboración

chile-mexico / datalakehouse

datalakehouse

main | datalakehouse / | Find file | Edit | Code

Edit README.md
Oswaldo Díaz authored 2 months ago | 6b9502f4 | History

Name	Last commit	Last update
assets	Edit analitica.md	2 months ago
deployment	Herramienta para realizar pruebas de s...	2 months ago
notebooks	Upload New File	2 months ago
README.md	Edit README.md	2 months ago

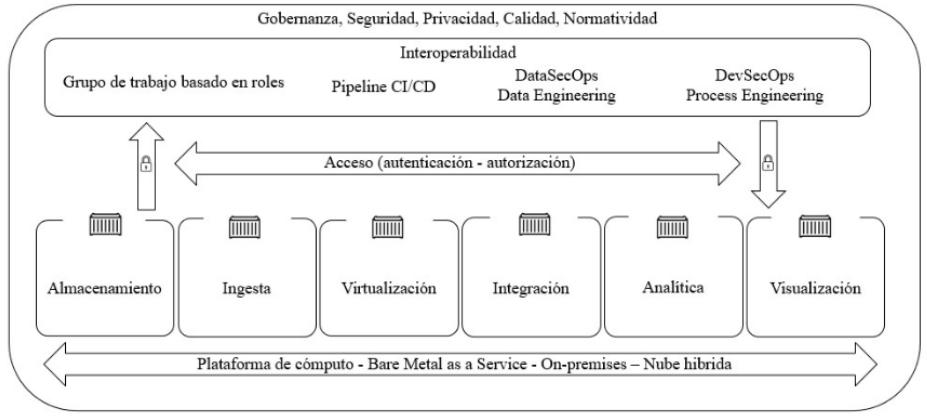
README.md

Desarrollo de una plataforma tecnológica de código abierto orientada a la integración de diversas fuentes de información para mejorar el aprovechamiento de los productos de las oficinas de estadística.

- ROL INE
- ROL INEGI

https://gitlab.com/chile-mexico/

Arquitectura



inegi.org.mx

Colaboración

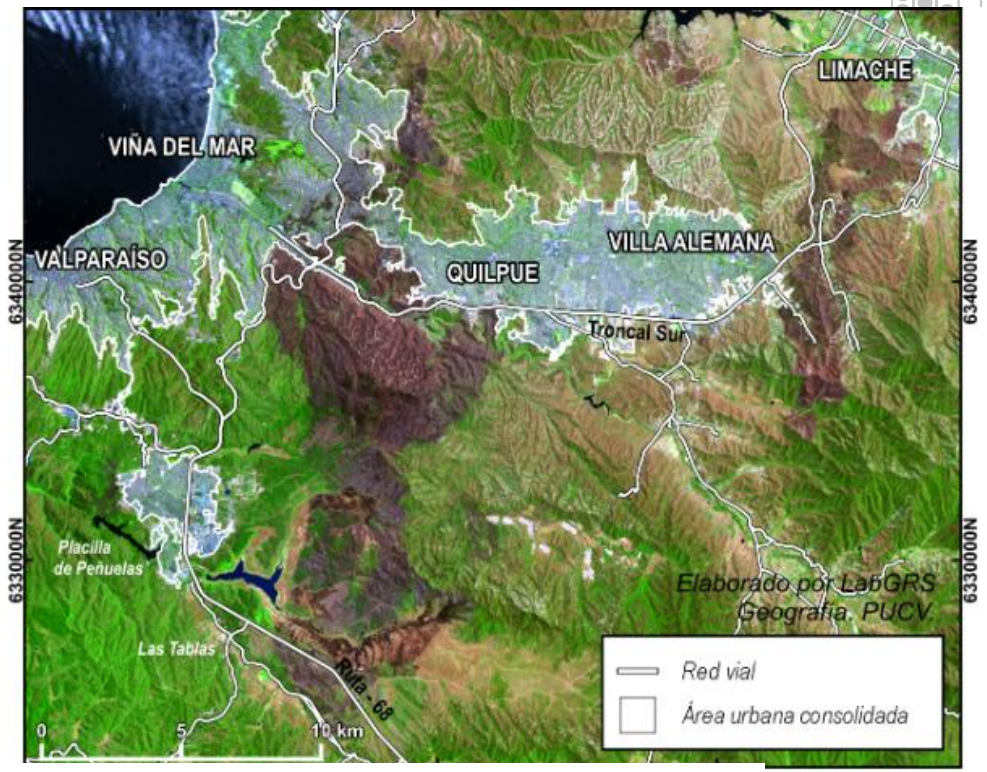
Wildfires Valparaiso

Demo on Integrating geospatial information for effective fore fire prevention

Tools | story map (on-line) | model (960 Mb) | shp (36 Mb)

Motivation

- Fires affect the environment, the population, and the economic activity of a territory.
- Early forecasting of forest fire danger and unraveling the relevant mechanisms affecting their occurrence are essentia efficient management of available resources.
- The availability of satellite image data and machine learning models presents a significant opportunity to address this challenge.



5. Model training

The hypothesis behind the model training is that a fire occurring on day t at the location x_t could be predicted based on a set of observations on previous days at the same location $x_{t'<t}$, where $t' = t-1, t-2, ..., t-10$ days.

Following the work proposed by Kondylatos et al. (GRL, 2022), we formulate the problem as a supervised binary classification problem, training LSTM model F that aims to predict the fire danger as the probability $y_t = p(F_t|x_{t'<t}) \in 0, 1$ of a fire (F) occurring on day t.

To reproduce the training of the LSTM model, please refer to the file model/README.md

https://gitlab.com/chile-mexico/

Colaboración

chile-mexico / clasificacion-delitos / modelo-enusc

modelo-enusc

main | modelo-enusc / | + | Find file | Edit | Code

add learning_rate beto
Javiera Magaluna Preuss Araya authored 1 month ago

3918df4e | History

Name	Last commit	Last update
assests	edit docs	2 months ago
deployment	edit docs	2 months ago
models	add LSTM model	2 months ago
notebooks	estructura incial	2 months ago
scr	add learning_rate beto	1 month ago
.gitignore	estructura incial	2 months ago

https://gitlab.com/chile-mexico/

Colaboración

chile-mexico / Data Privacy

Data Privacy

main | data-privacy / | + | Find file | Edit | Code

Delete assets
Oswaldo Diaz authored 1 week ago — 729d73d4 | History

Name	Last commit	Last update
assets	Delete assets	1 week ago
notebooks	Datos Sintéticos	2 weeks ago
README.md	Edit README.md	2 weeks ago

README.md

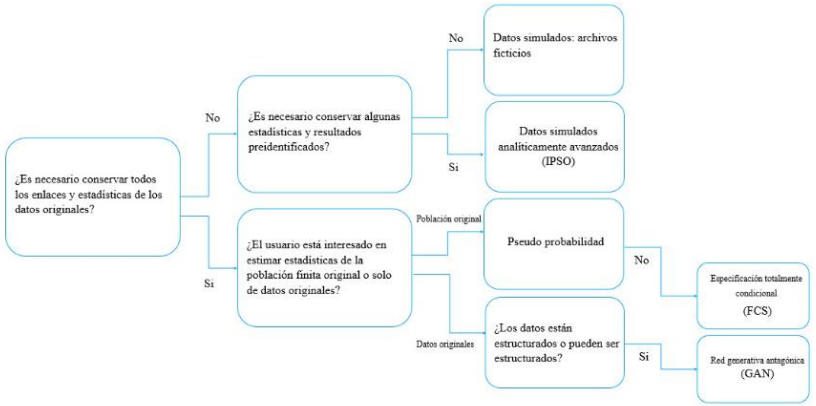
Marco experimental de datos sintéticos

¿Por qué usar datos sintéticos?

- Contar con una alternativa para cubrir la demanda de información con alta desagregación (microdatos).
- Habilitar la portabilidad de la información para automatizar procesos.
- Contribuir al marco de protección en la privacidad y confidencialidad estadística.
- Administrar el acceso a la información, dentro del entorno de gobernanza.

https://gitlab.com/chile-mexico/

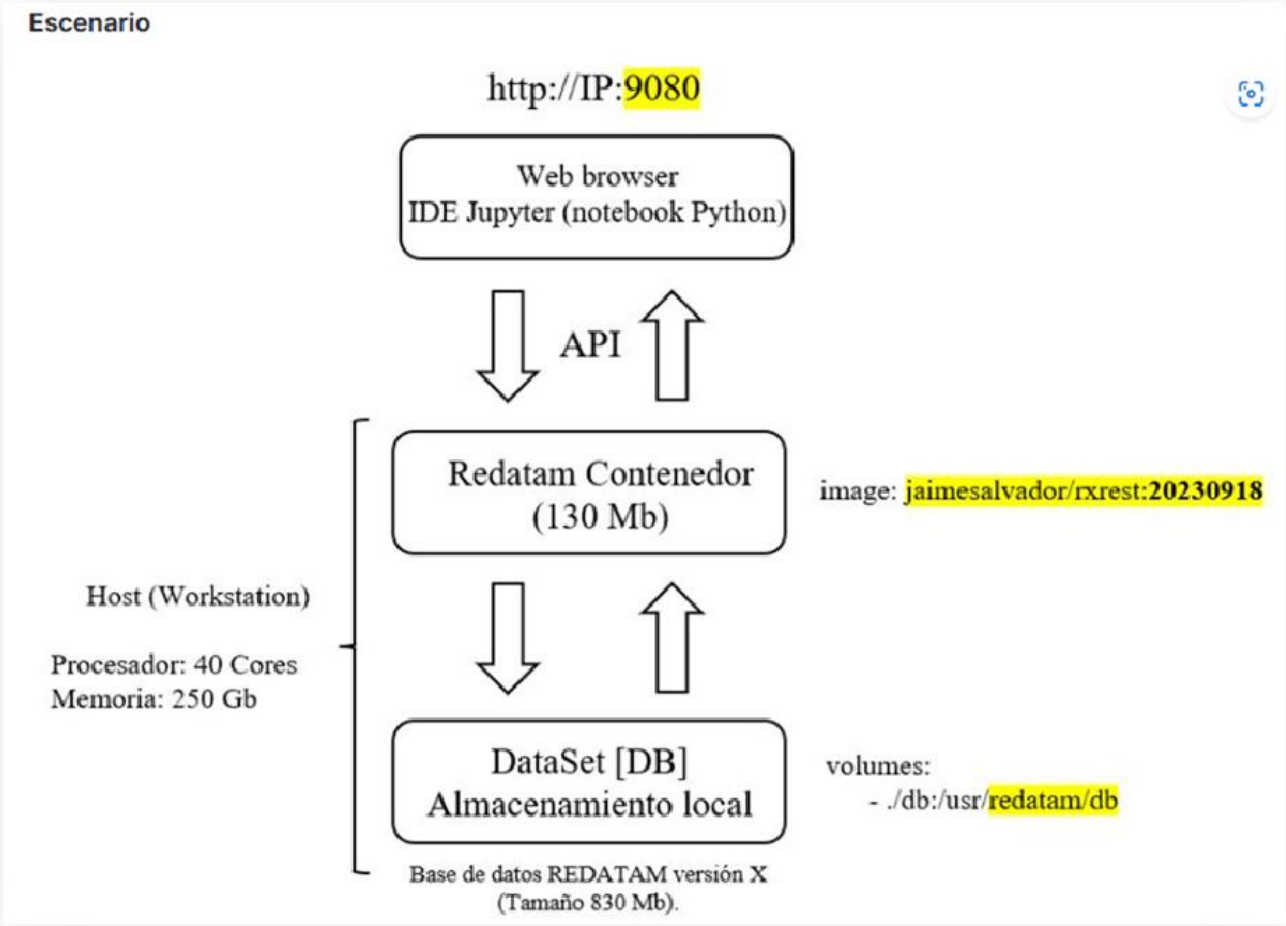
Diagrama para seleccionar que método definir para el proyecto



Casos de uso para INEGI

Proceso de registros administrativos sociodemográficos

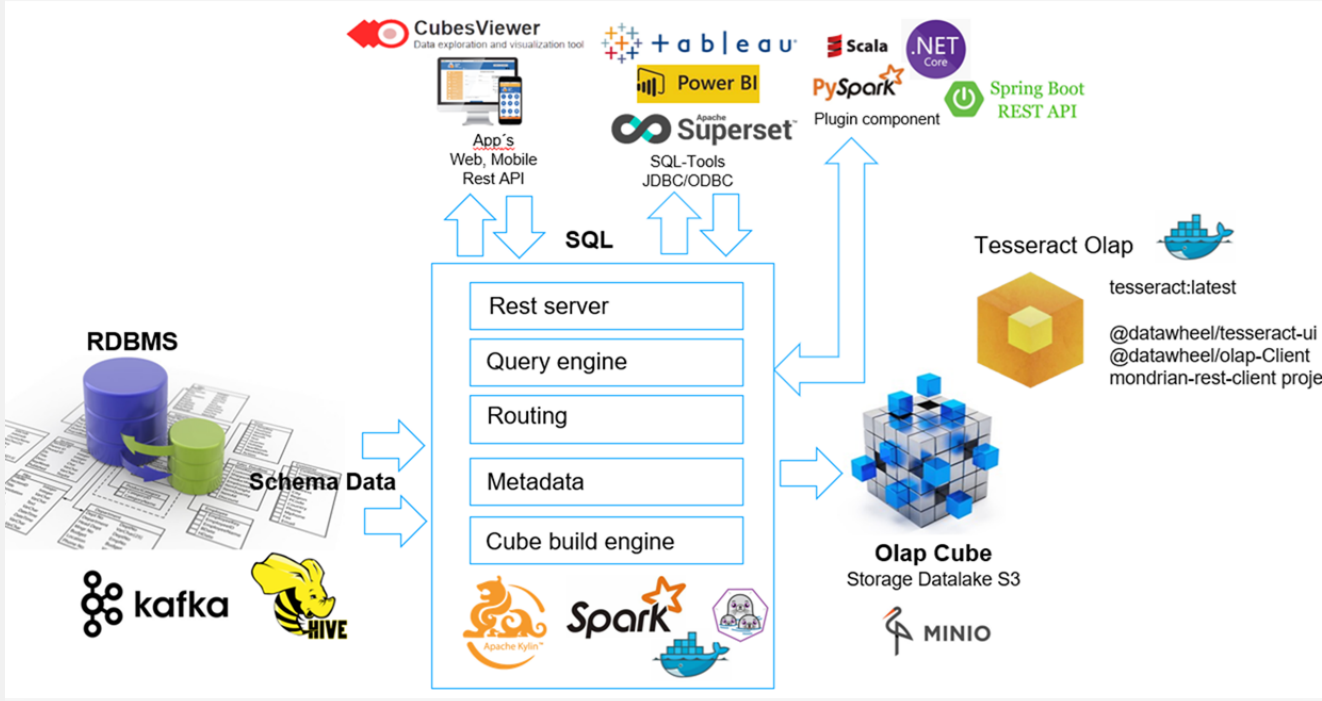
- Recepción masiva de datos sociodemográficos en bóveda digital (resguardo inicial).
- Selección de variables para transformación en Lago de Datos.
- Análisis y generación de respuestas a consultas planificadas.
- Uso de herramientas CEPAL – REDATAM.



INEGI

Procesamiento de información del almacén de datos.

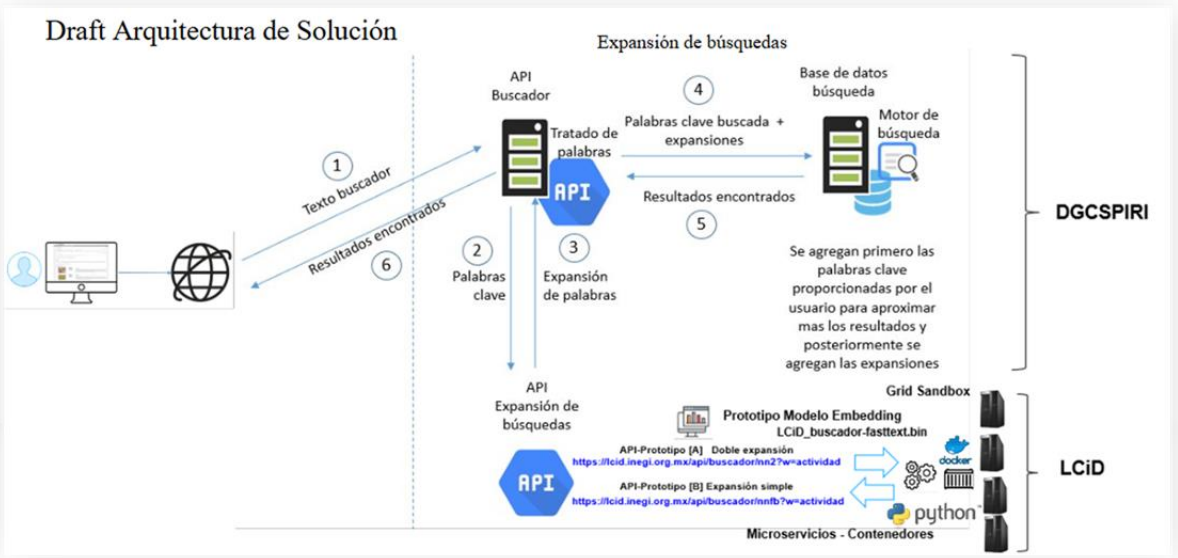
- Procesamiento de información del almacén de datos.
- Generación automatizada de tablas estadísticas desde el Lago de Datos.
- Publicación de resultados en el sitio web del INEGI.
- Uso de herramientas de desarrollo interno.



INEGI

Desarrollo de un algoritmo para búsquedas de información interés.

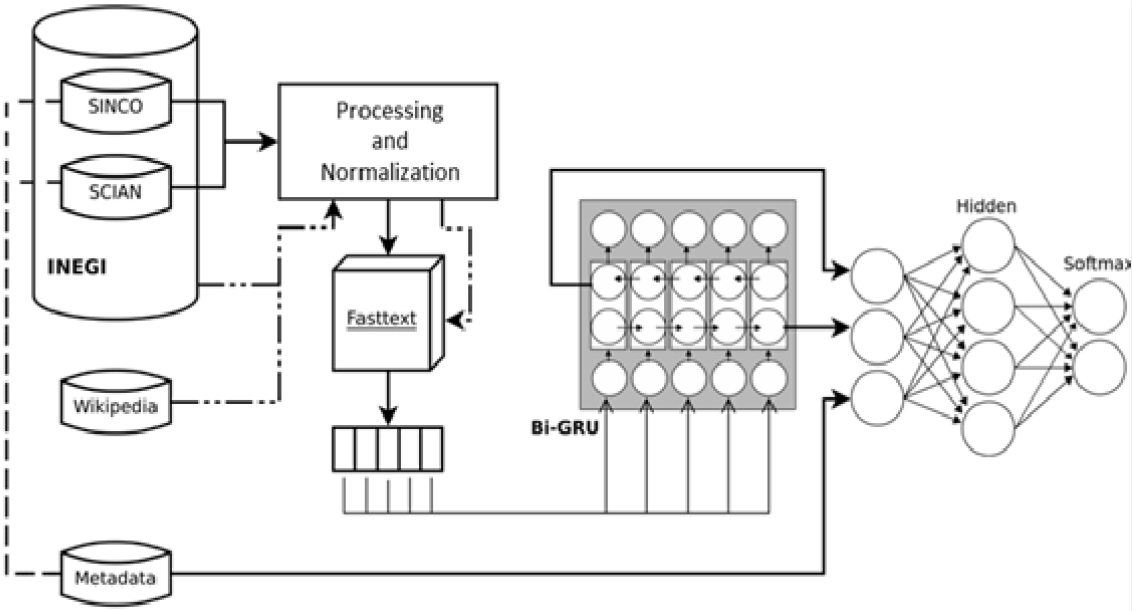
- Proyecto Mejoras de jerarquización para resultados del buscador del sitio del INEGI.
- Uso de herramientas de código abierto en el Lago de Datos.
- Aplicación de mejores prácticas internacionales.



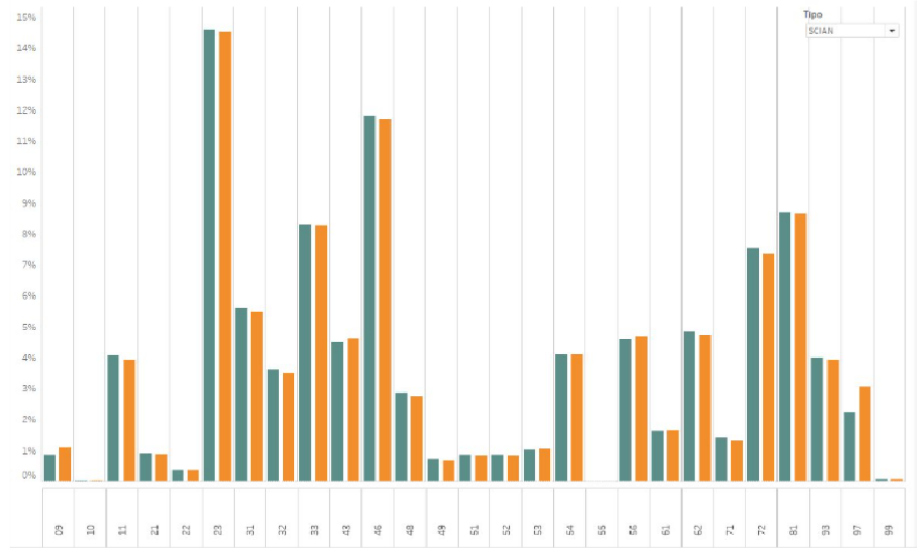
INEGI

Desarrollo de un algoritmo automatizado para codificación con Inteligencia Artificial.

- Aplicación en encuestas especiales y tradicionales sociodemográficas.
- Utilización de datos no estructurados de catálogos SCIAN y SINCO.
- Empleo de herramientas computacionales de código abierto adaptadas al Lago de Datos.



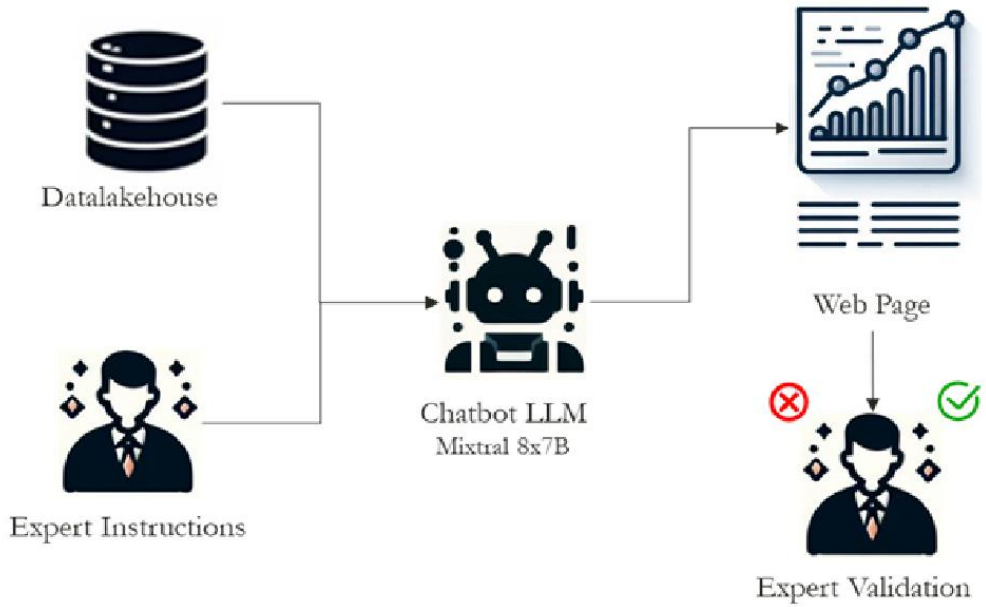
Estructura SCIAN original vs combinada.

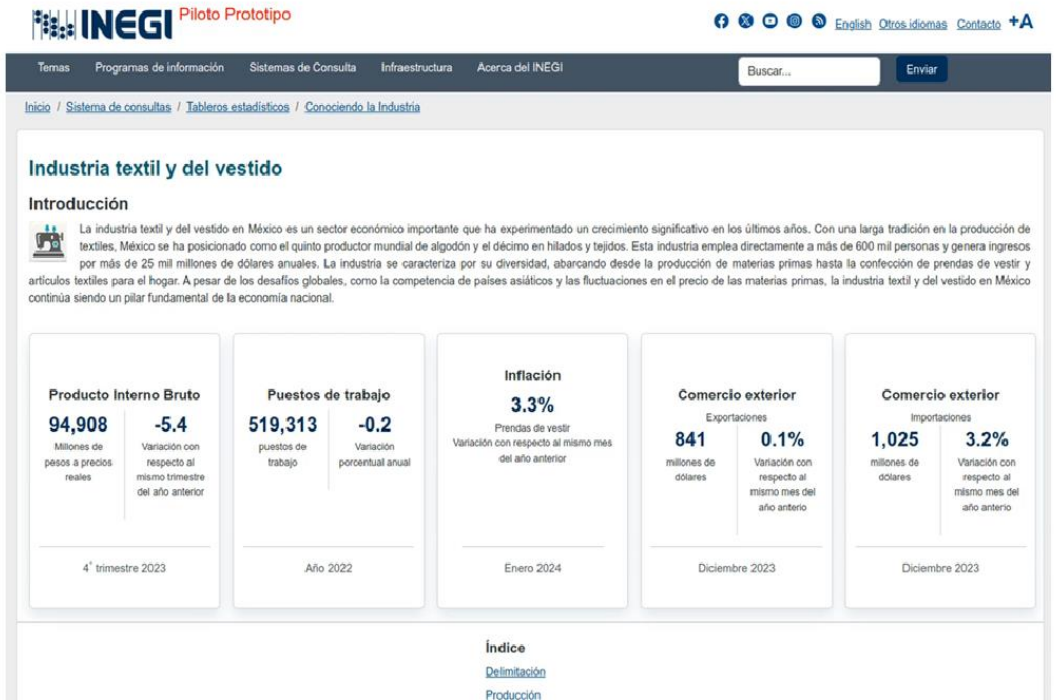


INEGI

Proceso de generación automatizada sobre la industria en México.

- Basado en procesamiento de datos masivos no estructurados (almacén a Lago de Datos).
- Integración en herramientas de código abierto para generar gráficas y tablas (cuantitativas y cualitativas).
- Empleo de grandes modelos de lenguaje (Inteligencia Artificial).



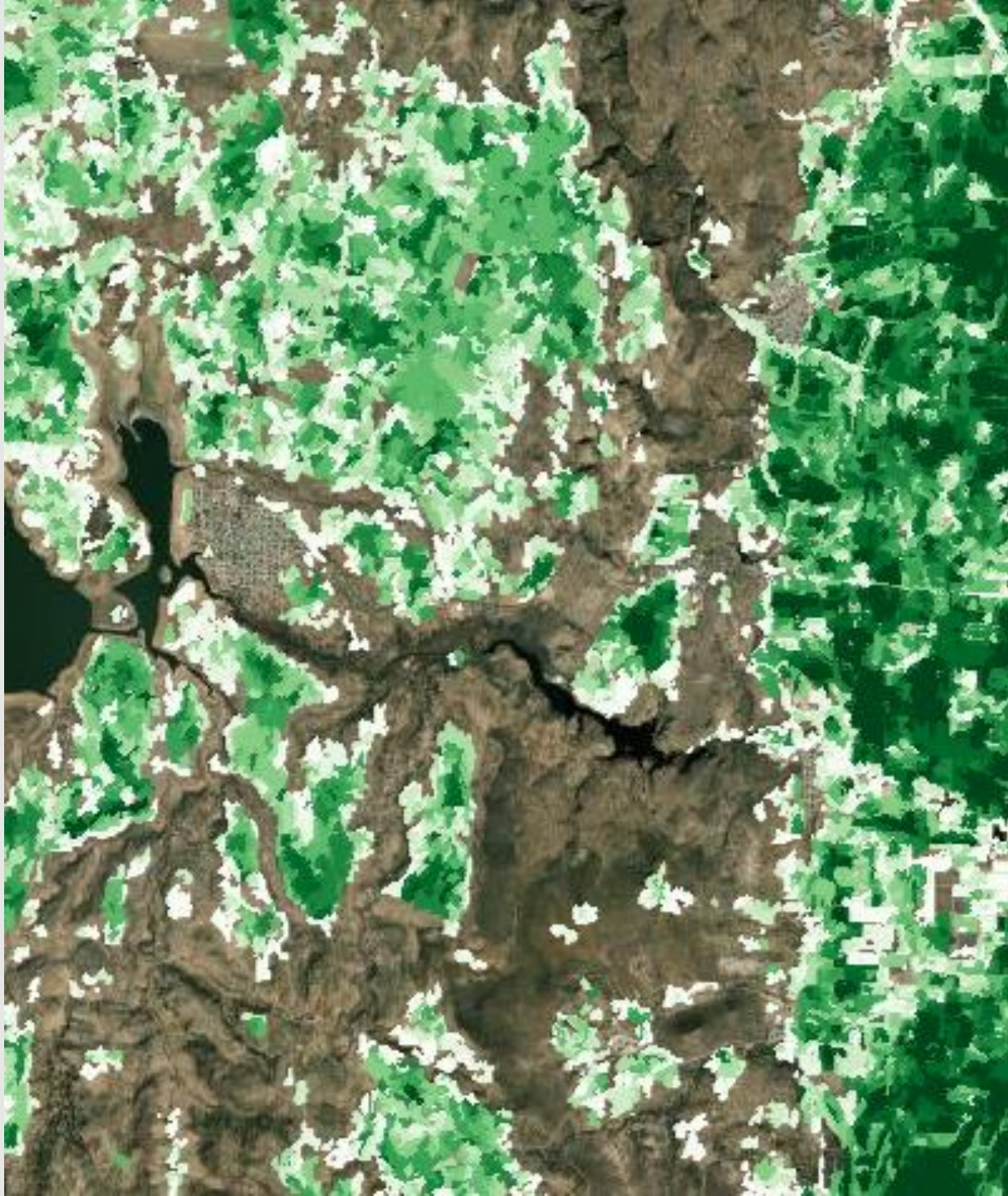


inegi.org.mx

INEGI

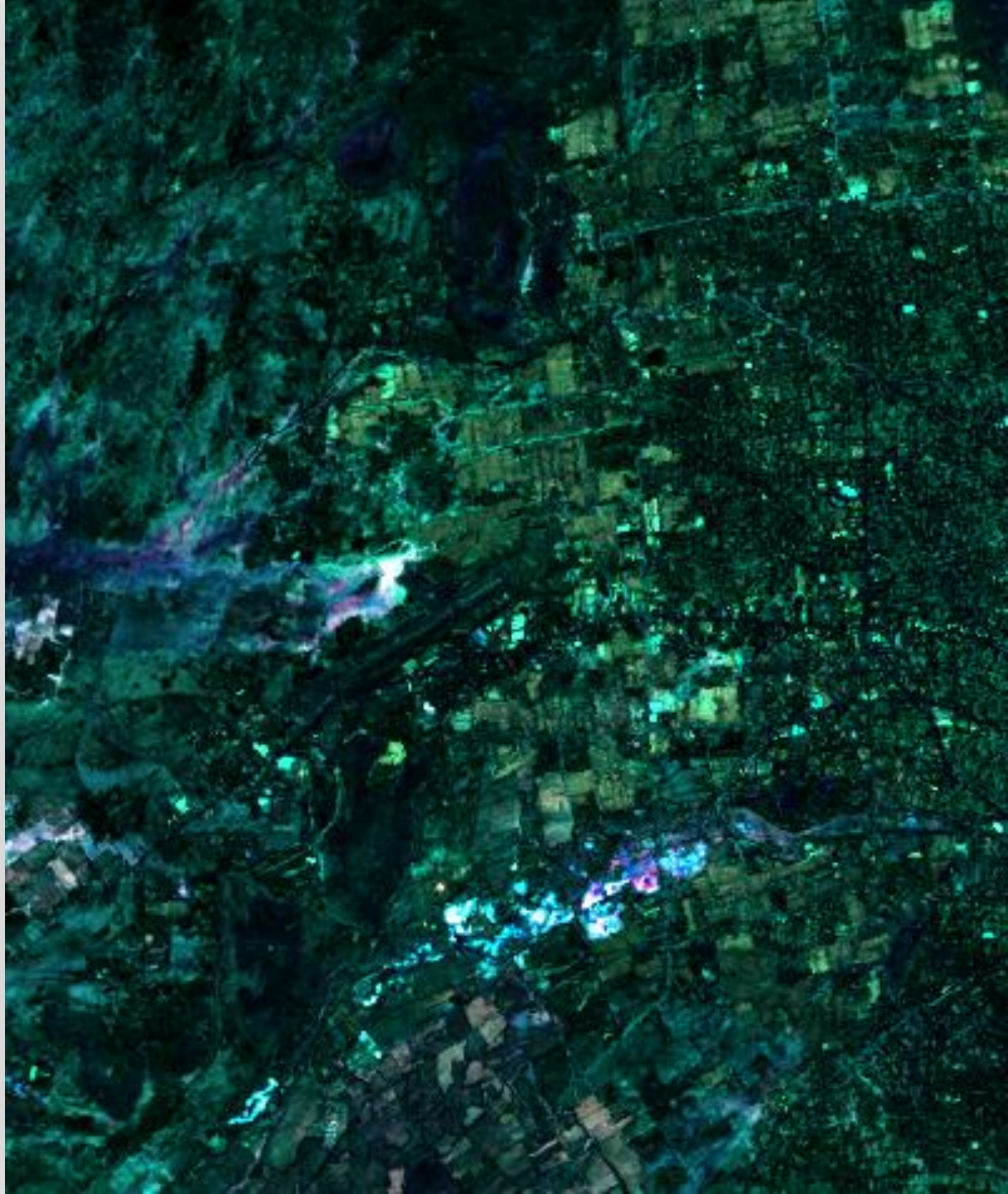
Desarrollo de algoritmo para actividades de la Frontera Agrícola Nacional.

- Aprovechamiento de datos geoespaciales no estructurados (formato Landsat).
- Realización de análisis computacional avanzado con grandes volúmenes de información en el Lago de Datos.
- Uso de herramientas de código abierto.
- Aplicación de mejores prácticas internacionales.



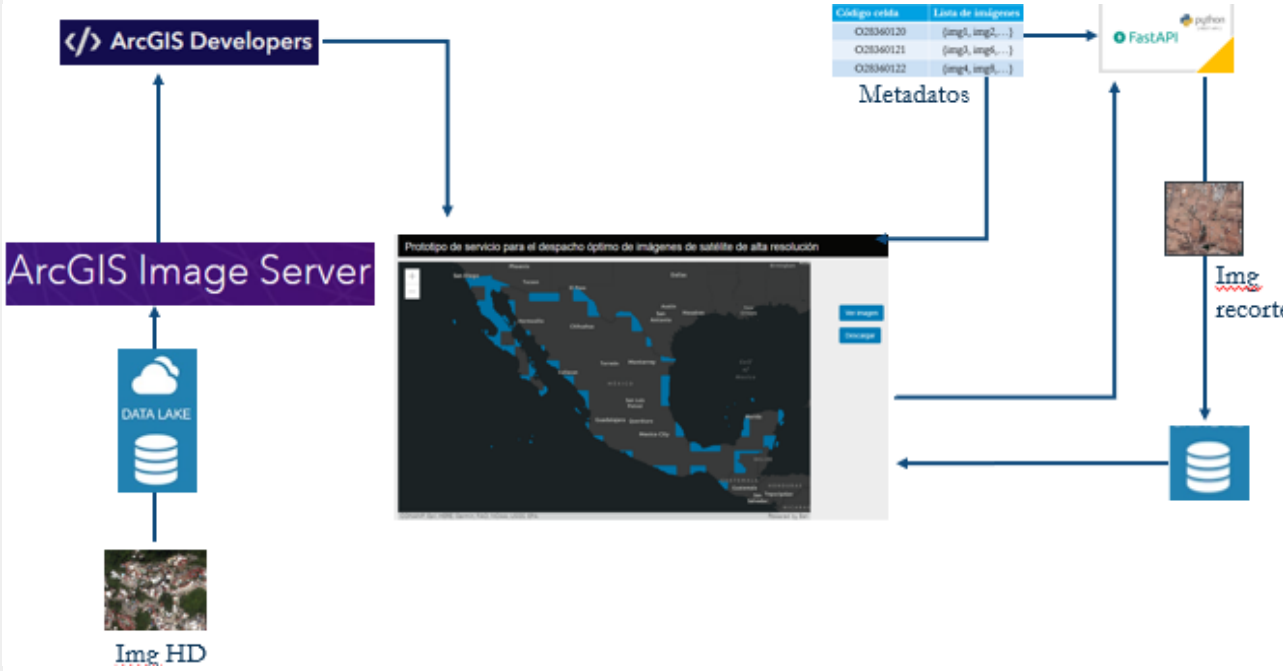
Desarrollo de un algoritmo para identificar cambios en objetos geográficos.

- Implementado en ambiente preproductivo para actualización cartográfica.
- Empleo de técnicas de Inteligencia Artificial (IA) para detección e identificación de cambios espaciales.
- Utilización de componentes de visualización geoespacial del Lago de Datos.



Procesamiento de imágenes geoespaciales de alta resolución.

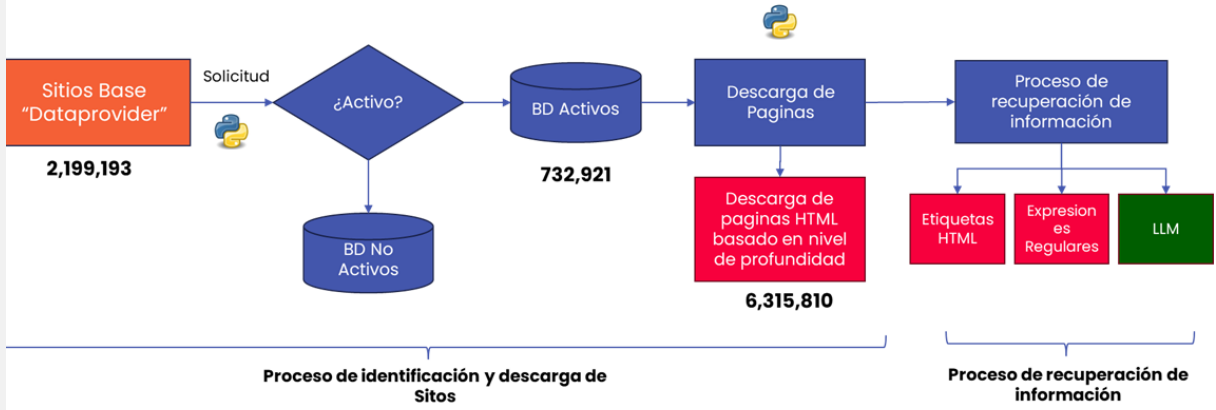
- Uso para analítica y ciencia de datos.
- Almacenamiento de datos en el Lago de Datos.
- Vinculación a herramientas GIS especializadas (ej. Quantum GIS).



INEGI

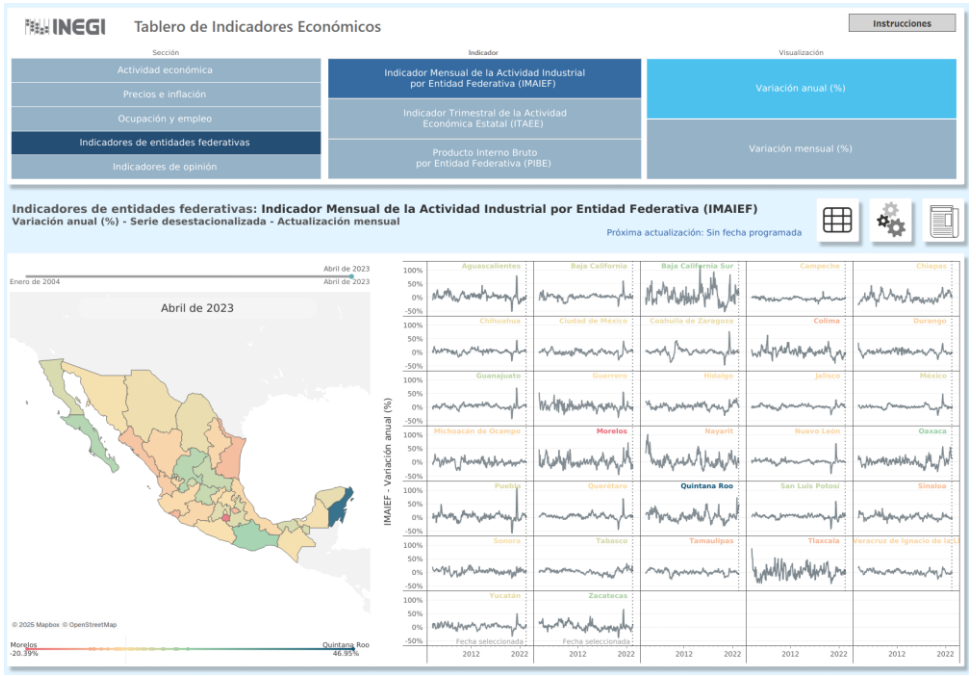
Desarrollo de proceso de extracción de contenidos y datos de sitios web.

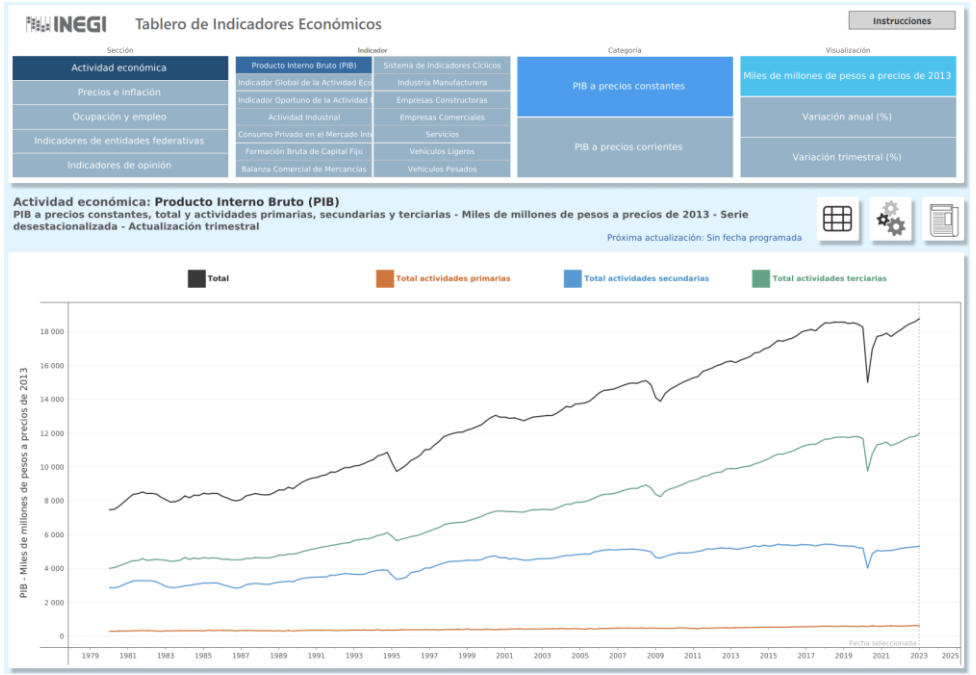
- Objetivo: Medición de negocios en la economía digital.
- Generación de archivos con relación de sitios web (.mx).
- Incluye grupos de datos y variables (sin especificar en el texto).
- Empleo de herramientas de código abierto conectadas al Lago de Datos.



Campos extraídos	Registros Asociados
Business Registry number	RFC
Company Name	Nombre del Establecimiento
Company Name	Razón social
Company Name y Address	Razón Social y Domicilio
Hostname	Sitio web
Hostname	Nombre del Establecimiento
Hostname	Razón Social
Phone Number	Teléfono

Generación de un Tablero de Indicadores de Actividad Económica para el seguimiento de la **información generada por INEGI** sobre todas las **actividades económicas**, dentro de la **plataforma del Lago de Datos**, optimizada para **visualización con herramientas especializadas**.

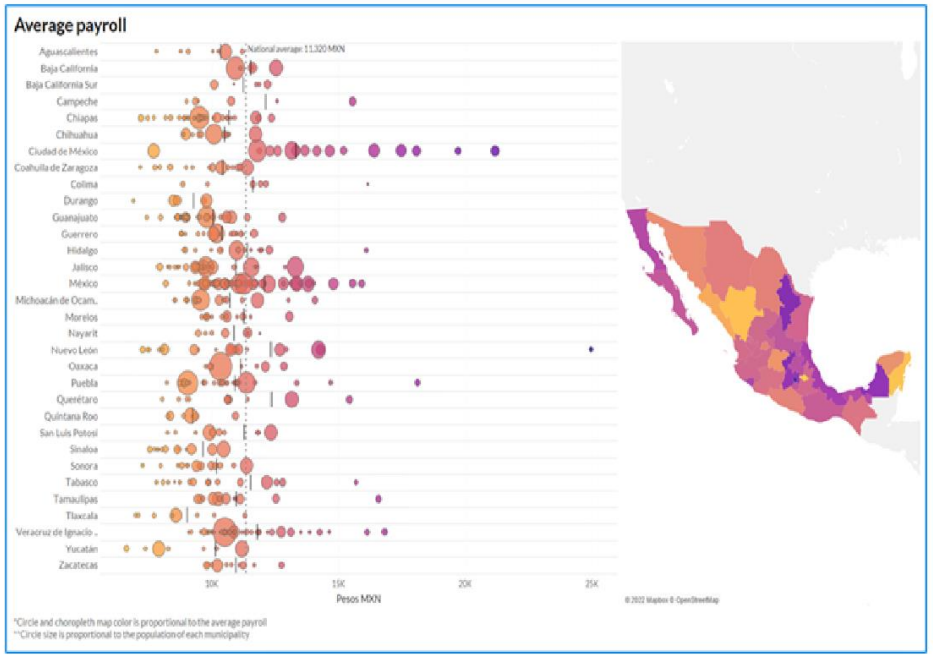




inegi.org.mx

INEGI

Proceso de generación de indicadores de actividad económica a partir de **información bancaria**, recibida en **grandes volúmenes** mediante **convenios** y almacenada en la **bóveda digital**. Posteriormente, ciertas **variables** son transformadas para **procesos de analítica y ciencia de datos** dentro del **Lago de Datos**, utilizando **herramientas computacionales de código abierto**, optimizadas según los requerimientos del proyecto.



INEGI

Preguntas

¡GRACIAS!

CONOCIENDO MÉXICO

800 111 46 34
www.inegi.org.mx
atencion.usuarios@inegi.org.mx

INEGIINFORMA