Estrategia de un lago de datos para flujos de trabajo en ciencia de datos

Data Lake Strategy for Data Science Workflows

Grupo colaborativo
Dirección del Laboratorio de Ciencia de Datos y Métodos Modernos de Producción de Información, Instituto Nacional de Estadística y Geografía (INEGI)
Av. Héroe de Nacozari no 2301, Aguascalientes, México
lcid@inegi.org.mx

Resumen — **El presente documento detalla la investigación y estrategia tecnológica realizada para implementar un Lago de Datos y Areneros del Laboratorio de Ciencia de Datos en el Instituto Nacional de Estadística y Geografía (INEGI) México, este proyecto busca integrar información digital de diferentes repositorios, fuentes de datos internas y externas, que existen por parte de los diversos entes generadores de información estadística y geográfica, en diversos formatos para conjuntarlos en un ambiente unificado de almacenamiento (temporal o permanente), que permita llevar a cabo procesos avanzados con técnicas orientadas a la analítica y ciencia de datos.**

Palabras Clave – Lago de datos; areneros; ciencia de datos.

Abstract — **This paper details the research and technological strategy carried out to implement a Data Lake and Sandboxes of the Data Science Laboratory at the National Institute of Statistics and Geography (INEGI) Mexico, this project seeks to integrate digital information from different repositories, data sources internal and external, which exist by the various entities that generate statistical and geographic information, in various formats to combine them in a unified storage environment (temporary or permanent), which allows advanced processes to be carried out with techniques oriented towards analytics and data science.**

Keywords – data lake; sandbox; data science.

I. INTRODUCCIÓN

El planteamiento de la estrategia es implementar un entorno flexible que permita incorporar componentes basados en herramientas hardware y software, teniendo como objetivo la generación de un lago de datos que permita que diferentes formatos de grandes volúmenes de información digital estadística y geográfica, recolectadas, almacenadas y procesadas en un ambiente con acceso controlado para su integridad, disponibilidad y confidencialidad, así como también la automatización de procesos basados en flujos de integración y despliegue continuo agiles, para realizar modelos de información (datos, metadatos, microdatos), que sirven como insumo para la analítica y ciencia de datos.

Dirección General de Integración, Analisis e Investigación
Dirección General Adjunta de Investigación
Instituto Nacional de Estadística y Geografía (INEGI)
Av. Héroe de Nacozari no 2301, Aguascalientes, México
lcid@inegi.org.mx

II. ESTADO DEL ARTE

Un lago de datos, es un área de almacenamiento compartido de grandes volúmenes de datos de tipo estructurados y no estructurados con diferentes formatos, de solo lectura, para su interacción de componentes hardware y software, permitiendo generar un conjunto de estrategias en la creación de conocimiento a partir de minería, análisis y ciencia de datos recopilados, aportando al descubrimiento y explotación, de información ad hoc en tiempo real, para la generación de visualizaciones cuantitativas o cualitativas que apoyan a la toma de decisiones.

Un arenero o Sandbox, es un entorno controlado con base a tecnologías de información, escalable, para ejecutar herramientas especializadas, con niveles de seguridad y de manera independiente, permitiendo controlar los recursos hardware y software, que son utilizados en el acceso al lago de datos, teniendo la finalidad de apoyar en la generación de productos prototipos, orientados en analítica y ciencia de datos.

Algunos casos de éxito donde muestra la implantación de un lago de datos en oficinas de estadística, es el trabajo desarrollado por Llave (2018) [5], presenta un panorama general del estatus que se ha tenido al implementar la estrategia, en este estudio se exploran las tecnologías involucradas, los beneficios directos e indirectos, teniendo como conclusión importante que los lagos de datos no reemplazan a los almacenes de datos por sus siglas en inglés (Data Warehouses), sino que aumentan las capacidades, para el uso de grandes volúmenes de datos.

Anejionu et al. (2019) [1], presentan el SUDS o Sistema de Datos Urbano Espacial que se emplea en Reino Unido para realizar analítica de datos sociales y económicos en conjunción con un sistema geográfico, presenta de manera simplificada el flujo de la información dentro de un ambiente de un lago de datos, además de que presenta una serie de casos de éxito a raíz de la implementación del modelo referido.

Por otra parte, Sfaxi & Ben Aissa (2020) [7], presentan una propuesta para la gestión de los datos a gran escala para la toma de decisiones, en el particular contexto del piloto prototipo, aborda el plantear un modelo de cuatro fases: en la primera se presenta un lago de datos crudos, es decir, los datos en su estado natural, posteriormente se cuenta con un lago de datos refinados los cuales sirven a una tercera capa, en la cual se crean vistas de

2022 11th International Conference On Software Process Improvement (CIMPS) | 979-8-3503-9896-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/CIMPS57786.2022.10035686

la información a partir de los datos mencionados anteriormente, en la cuarta capa es la que persiste los resultados de la fase de análisis, accesible a través de herramientas de visualización. En lo que respecta a un caso de aplicación real en una oficina de estadística, CBS Statistics Netherlands (Países Bajos) ha implementado un lago de datos al interior de la institución, es capaz de captar los datos de entrada provenientes de diversas fuentes, conjuntarlos con otros datos generados al interior de CBS y producir una serie de salidas dirigidas a diferentes actores que pudieron ser producidas gracias al poder del análisis que se tiene al conjuntar las fuentes en un repositorio común.

III. FASES ESTRATEGICAS

En la Dirección del Laboratorio de Ciencia de Datos y Métodos Modernos de Producción de Información, adscrito a la Dirección Adjunta de Investigación dentro del INEGI, ha adoptado una metodología de trabajo ágil, basadas en ocho fases descritas a continuación.

Fase fuentes de datos: consiste en la selección de las fuentes de información que serán de relevancia para el proyecto de ciencia de datos en cuestión. El responsable de seleccionar las fuentes de información será el rol de Arquitecto de Big Data. En caso de que el proyecto se desarrolle por una solicitud externa, el Solicitante será responsable de proveer acceso a esta información cuando la misma no sea de acceso abierto. Estas fuentes pueden incluir datos, metadatos, microdatos de censos, encuestas, imágenes satelitales geoespaciales entre otras.

Fase extracción y carga de datos: consiste en la recolección de los datos insumo almacenados en el lago de datos, para implementar las técnicas en ciencia de datos con base al requerimiento inicial del proyecto, el rol responsable de esta fase el Ingeniero de Datos considerando implementar métodos para la extracción carga de datos, de forma automática o manual; acceso a repositorios internos y/o externos, donde está almacenados los datos; carga de datos extraídos a la infraestructura del lago de datos, y aplicación de las transformaciones de datos que sean requeridas.

Fase recuperación de información: consiste en implementar estrategias de búsqueda sobre los conjuntos de datos, metadatos, microdatos, extraídos y cargados en el lago de datos. Las estrategias de búsqueda se pueden hacer con base a estructuras identificadas y descritas por medio de diccionarios de datos con técnicas, que permiten generar grupos de información llamados "conjunto de datos". Esta etapa podrá llevarse a cabo en distintos momentos del proyecto de acuerdo con las necesidades que se presenten. El rol responsable es el Ingeniero de Datos.

Fase procesamiento de datos: consiste en la implementación y aplicación de técnicas estadísticas e informáticas en la cual se busca analizar los datos recolectados en las fases anteriores para tener un mayor entendimiento de la información, así como descubrir aspectos relevantes de los datos como las relaciones existentes entre las variables que se están analizando, el rol responsable de esta fase será el Científico de Datos Jr., considerando el análisis exploratorio de datos con base a las características del requerimiento; limpieza de datos, control de calidad de la información relacionada en el proyecto, y representación de datos que responda a las necesidades de la fase de construcción del modelo.

Fase construcción del modelo: consiste en realizar la selección de los modelos de aprendizaje computacional o modelos estadísticos, según sea el caso, y su implementación en un ambiente controlado. A lo largo del proyecto se podrá experimentar con distintos modelos, el rol responsable es el Científico de Datos Sr.

Fase evaluación y validación de resultados: la evaluación y validación de los resultados obtenidos (del procesamiento y análisis de los datos) se realizarán con base en las métricas de desempeño establecidos, así como procedimientos estadísticos basados en mejores prácticas internacionales. Se podrá realizar durante distintos momentos a lo largo de la ejecución del proyecto a fin de contar con criterios para seleccionar el modelo con el mejor desempeño, el rol responsable de esta fase será el Científico de Datos Sr.

Fase presentación de resultados: consiste en realizar componentes por medio de herramientas especializadas basadas en tecnologías de información, que permiten mostrar los avances del proyecto en ciencia de datos, el rol responsable de esta fase será el Científico de Datos Líder quien podrá considerar para la presentación de resultados la demostración de prototipos de productos de datos dentro del lago de datos y arenero; generar un reporte técnico que indica la metodología y el nivel de madurez; construcción de tableros de control u otras estrategias de visualización de datos, y opcional la producción de artículo de investigación científica.

Fase entrega de productos de datos: consiste en el proceso de implementación en la arquitectura con base a tecnologías de información que el solicitante disponga acompañado de su documentación, adicionalmente, como parte del proceso de entrega se podrá incluir una asesoría para el uso del prototipo y la estrategia para reducir riesgos en la operación con base al ciclo de vida del proyecto.

Para el cumplimiento de las anteriores fases son necesarios los siguientes roles descritos a continuación.

• Científico de datos líder: responsable de supervisar el desarrollo de los proyectos, así como de dirigir la investigación científica, decidir los métodos y procedimientos de ciencia de datos, establecer la estrategia de colaboración, dimensionar las necesidades de infraestructura y promover del desarrollo de nuevas capacidades;

• Científico de datos Sr.: responsable de dirigir la implementación de los métodos y procedimientos de ciencia de datos, y establecer las métricas de desempeño que se requerirán para la evaluación y validación de resultados;

• Científico de datos Jr.: responsable de implementar métodos y procedimientos, así como el desarrollo de artefactos (tableros de control, mapas históricos, visualizaciones, reportes) para la presentación de resultados;

• Arquitecto de flujo de trabajo: responsable de la especificación del flujo de trabajo de los proyectos de ciencia de datos, así como de su documentación;

• Arquitecto de infraestructura: responsable de la configuración y administración de la infraestructura asignada al LCiD;

• Arquitecto de big data: responsable de determinar la estrategia y plataforma tecnológica para la recolección, almacenamiento y procesamiento de grandes volúmenes de datos, requeridos para cumplir los objetivos particulares de los proyectos de ciencia de datos;

• Ingeniero de datos: responsable de llevar a cabo las tareas de extracción, carga y transformación de datos, así como implementar estrategias de recuperación de información que requieran los proyectos de ciencia de datos.

IV. ESTRATEGIA DE IMPLEMENTACIÓN

Para la implementación del lago de datos y el arenero son utilizadas herramientas tecnológicas (basadas en software con licenciamiento libre por sus siglas en ingles Open Source), que permitieran llevar a cabo las fases anteriores descritas, considerando las capas descritas a continuación.

Capa de Interoperabilidad: en esta capa están las herramientas transversales que permiten la integración continua, control de cambios, y otras características para proyectos basados en Ingeniería de Software de equipos de trabajo DevSecOps (Developer Security Operation) [3], así como también de equipos de trabajo para proyectos en Ciencia de Datos DataSecOps (Data Security Operation) [3], el acceso al lago de datos y areneros es restringido con niveles de seguridad.

Las herramientas utilizadas en esta capa son:

• GitLab: una plataforma web de desarrollo de software, la cual nos permite gestionar código fuente, planeación de proyectos, gestión de flujos automatizados de integración continua. [8]

• MLflow: herramienta para la gestión de proyectos de machine learning con la cual se puede registrar, empaquetar, distribuir y consultar experimentos, con la finalidad de obtener modelos reproducibles y robustos. [9]

• Kedro: es un marco para la creación de flujos de datos. Adopta las mejores prácticas en ingeniería de software para crear código de ciencia de datos que sea reproducible, mantenible y modular. [10]

Capa de ingestión de datos: la capa de ingestión de datos es la encargada de consumir y hacer acopio de prácticamente cualquier tipo de fuentes de datos estructurados y no estructurados interno y externos al Instituto, en este nivel son definidos los microservicios – contenedores que estarán implementados dentro del arenero, de tipo (controladores, conectores, entre otras herramientas o librerías de recolección de información digital).

Las herramientas utilizadas en esta capa son:

• Python: Es un lenguaje de programación de código abierto orientado a objetos. Su sintaxis sencilla y la cantidad de librerías disponibles facilitan la creación rápida de programas y flujos de información.[11]

• Jupyter Lab - Notebook: Es un entorno de desarrollo web interactivo con soporte para distintos lenguajes en los que destaca Python. Permite combinar ejecución de código con documentación y visualizaciones en un solo archivo, facilitando así la creación de prototipos y la difusión de resultados.[12]

Capa de almacenamiento de datos: la capa se encarga del almacenaje de datos define los servicios y/o microservicios en contenedores generados para el Arenero que permite por medio de protocolos de transferencia de datos, él envió de información digital, soportando varios formatos datos estructurados y no estructurados.

La herramienta utilizada en esta capa es, Minio: un software para servidores de almacenamiento distribuidos bajo el protocolo llamado “S3”, que permite crear nubes privadas de datos de alto desempeño, así como también realiza interacción estandarizada con otros servicios para acceder a sus archivos en cualquier formato. [13]

Capa de virtualización de datos: la virtualización de datos generada en el arenero, con herramientas que permiten hacer consultas de la información digital almacenada en el lago de datos, para establecer una estrategia en la generación de tablas de datos de solo lectura para que no pierdan el linaje, con la finalidad de estandarizar los nombres de campos y permita tener un grupo de datos homogéneo.

Las herramientas utilizadas en esta capa son:

• Trino: Motor de consultas SQL distribuido, el cual permite conectarse a otras fuentes de datos servidores que tienes gestores de bases de datos permitiendo interacción entre ellos dentro de una única interfaz transparente y estandarizada.[14]

• Hive: Este software permite leer, escribir y gestionar conjuntos de datos en almacenamientos distribuidos utilizando el lenguaje llamado SQL, permite asignar estructura a archivos ya existentes con lo cual es posible hacer consultas de información.[15]

Capa de integración de datos: los servicios o microservicios en contenedores generados en el arenero, permite integrar tecnologías de gestión de información de diversos proveedores (gestores de datos), permitiendo la administración de la información digital almacenada (real o virtual) en el lago de datos, en apoyo al gobierno y documentación, entre otras características, la herramienta utilizada son librerías especializadas basadas en Python [16] que permiten el agrupamiento de información para generar un grupo de datos.

Capa de analítica y ciencia de datos: la analítica de datos permite llevar a cabo estudios complejos sobre los datos empleando algoritmos emanados a proceso de ciencia de datos, en el arenero, permiten utilizar la información almacenada en el lago de datos, como insumo grupo de datos para el entrenamiento de modelos (aprendizaje automatizado, procesamiento de lenguaje natural, aprendizaje profundo), los resultados que generan las iteraciones también son almacenadas en el lago de datos, para su referencia en generar prototipos para revisión de niveles de madurez por parte del grupo de trabajo establecida por proyecto, la herramienta utilizada son en entornos de trabajo (frameworks) y librerías especializadas basados en Python [16] que permiten el procesamiento con técnicas de aprendizaje supervisado, no supervisado, profundo y de lenguaje natural para generar algoritmos que respondan a el requerimiento.

Capa de visualización de datos: para la visualización de datos es el medio mediante el cual con herramientas especializadas

pueden generar reportes, graficas o tableros de control que permitan mostrar los indicadores cuantitativos y cualitativos para tomar la decisión sobre el nivel de madurez del proyecto con ciencia de datos para posteriormente realizar la entrega del prototipo.

Las herramientas utilizadas en esta capa son:

• SuperSet: es una plataforma de exploración y visualización de datos que puede conectarse a cualquier fuente de datos basada en SQL a través del componente llamado SQLAlchemy, soportando grandes volúmenes de información basados en escala de petabytes, tiene variedad de gráficos y mapas para elegir dependiendo el tipo de la necesidad para mostrar los datos, a través de selectores se configura para ser presentados en un tablero de control basado en tecnología web permitiendo el acceso controlado de los usuarios. [17]

• D3.js (Data-Driven Documents): es una librería de JavaScript para manipular documentos basados en datos especializado en visualización con tecnologías web.[18]

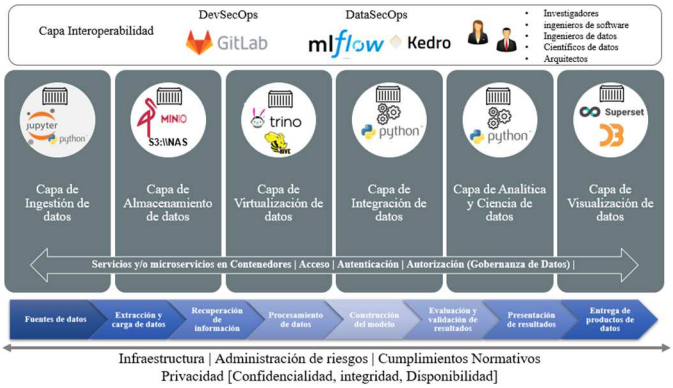

Figura 1. Diagrama de fases, capas y herramientas estratégicas utilizadas en la arquitectura del lago de datos y areneros.

Nota: la infraestructura requerida para la implementación de un lago de datos y arenero depende de las características del requerimiento por los proyectos y servicios que se generan, a lo cual el siguiente es un ejemplo del dimensionamiento básico, generar cuatro servidores físicos o lógicos, que cada servidor tiene 20 núcleos 40 procesadores lógicos, 256 Gb en memoria RAM, 400 Gb en almacenamiento tipo SSD, y 4 Tb en almacenamiento de tipo SATA, configurados sistema operativo multiusuario basado en el kernel de Linux para soportar las herramientas anteriores mencionadas, con tecnología que permita contenerización o tecnología de hipervisores, los métodos de conexión entre los servidores deberán considerar el uso de tecnologías de red de área de almacenamiento SAN (Storage Area Network) y/o Almacenamiento conectado en red NAS (Network Attached Storage), con la finalidad de tener una estrategia de alta disponibilidad, escalable, segura para reducir riesgos en la operación como parte de los procesos.

V. CONCLUSIONES

La experiencia en la generación, implementación y administración de un lago de datos y areneros dentro del Instituto genera valor agregado a los procesos basados en analítica y ciencia de datos considerando grandes volúmenes de información estadística y geográfica en diferentes formatos, así como también consolida un repositorio para generar productos y servicios orientados a visualizaciones cuantitativas y cualitativas que permitan la toma de decisiones.

AGRADECIMIENTOS

Grupo colaborativo Laboratorio de Ciencia de Datos INEGI

- VILLASEÑOR GARCIA ELIO ATENOGENES elio.villasenor@inegi.org.mx
- CORONADO IRUEGAS ABEL ALEJANDRO abel.coronado@inegi.org.mx
- PIMENTEL ALARCON ALEJANDRO ESTEBAN alejandro.pimentel@inegi.org.mx
- SUAREZ PONCE DE LEON RANYART RODRIGO ranyart.suarez@inegi.org.mx
- FIGUEROA MARTINEZ ALEJANDRA alejandra.figueroa@inegi.org.mx
- ESQUER MARTINEZ AMADO amado.esquer@inegi.org.mx
- SILVA CUEVAS VICTOR victor.silvac@inegi.org.mx
- CABRERA ZAMORA IRVING GIBRAN irving.cabrera@inegi.org.mx obra.
- DIAZ EDGAR OSWALDO oswaldo.diaz@inegi.org.mx

REFERÊNCIAS BIBLIOGRÁFICA

[1] Anejionu, O. C. D., Thakuriah, P. (Vonu), McHugh, A., Sun, Y., McArthur, D., Mason, P., & Walpole, R. (2019). Spatial urban data system: A cloud-enabled big data infrastructure for social and economic urban analytics. Future Generation Computer Systems, 98, 456–473. https://doi.org/10.1016/j.future.2019.03.052

[2] Ashofteh, A., & Bravo, J. M. (2021). Data science training for official statistics: A new scientific paradigm of information and knowledge development in national statistical systems. Statistical Journal of the IAOS, 37(3), 771–789. https://doi.org/10.3233/sji-210841

[3] Díaz, O., Muñoz, M., & Mejía, J. (2019). Responsive infrastructure with cybersecurity for automated high availability DevSecOps processes. 2019 8th International Conference On Software Process Improvement (CIMPS), 1–9. https://doi.org/10.1109/CIMPS49236.2019.9082439

[4] Jimenez-Marquez, J. L., Gonzalez-Carrasco, I., Lopez-Cuadrado, J. L., & Ruiz-Mezcua, B. (2019). Towards a big data framework for analyzing social media content. International Journal of Information Management, 44, 1–12. https://doi.org/https://doi.org/10.1016/j.ijinfomgt.2018.09.003

[5] Llave, M. R. (2018). Data lakes in business intelligence: reporting from the trenches. Procedia Computer Science, 138, 516–524. https://doi.org/10.1016/j.procs.2018.10.071

[6] Mankins, J. (1995). Technology Readiness Level – A White Paper.

[7] Sfaxi, L., & Ben Aissa, M. M. (2020). DECIDE: An Agile event-and-data driven design methodology for decisional Big Data projects. Data and Knowledge Engineering, 130(July 2019), 101862. https://doi.org/10.1016/j.datak.2020.101862

[8] Documentación para instalación de la herramienta en la capa de interoperabilidad llamada GitLab (https://docs.gitlab.com/ee/install/requirements.html)

[9] Documentación para instalación de la herramienta en la capa de interoperabilidad llamada MLflow (https://mlflow.org/docs/latest/quickstart.html#installing-mlflow)

[10] Documentación para la instalación de la herramienta en la capa de interoperabilidad llamada Kedro y que permite la conexión con MLflow (https://kedro-mlflow.readthedocs.io/en/stable/source/02_installation/index.html)

[11] Documentación para la instalación e implementación de los modulos para el lenguaje Python (https://docs.python.org/es/3/installing/index.html#installing-into-the-system-python-on-linux)

[12] Documentación para implementar el IDLE llamado Jupyter Lab utilizado como herramienta pars interoperar con python (https://docs.jupyter.org/en/latest/install.html)

[13] Documentación para implementar la herramienta de almacenamiento llamada Minio con el protocolo de transferencia S3 (https://docs.min.io/minio/baremetal/)

[14] Documentación para implementar el componente que permite la conectividad con fuentes de informacion relacionadas con gestores de datos llamada trino (https://trino.io/docs/current/installation/deployment.html)

[15] Documentación para implementar el componente que permite la virtualización de los datos llamada Hive (https://cwiki.apache.org/confluence/display/Hive//GettingStarted#GettingStarted-InstallingHivefromaStableRelease)

[16] Documentación para implementar las biblotecas de código que permiten la gestion de analítica y ciencia de datos (https://docs.python.org/es/3/library/)

[17] Documentación para instalar la herramienta que permite la visualización llamada Superset (https://superset.apache.org/docs/installation/installing-superset-from-scratch)

[18] Documentación para implementar código basado en javascript que permite la visualización de datos llamada D3js (https://d3js.org/)