Estratégia de data lake para fluxos de trabalho de ciência de dados

Estratégia de data lake para fluxos de trabalho de ciência de dados

Grupo Colaborador
Direcção do Laboratório de Ciência de Dados e Métodos Modernos de Produção de Informação, Instituto Nacional de Estatística e Geografia (INEGI)
Av. Héroe de Nacozari nº 2301, Aguascalientes, México lcid@inegi.org.mx

Direção-Geral de Integração, Análise e Investigação
Direção-Geral Adjunta de Investigação
Instituto Nacional de Estatística e Geografia (INEGI)
Av. Héroe de Nacozari nº 2301, Aguascalientes, México lcid@inegi.org.mx

***Resumo* — Este documento detalha a estratégia de pesquisa e tecnologia realizada para implementar um Data Lake e Sandboxes do Laboratório de Ciência de Dados do Instituto Nacional de Estatística e Geografia (INEGI) México, este projeto busca integrar informações digitais de diferentes repositórios, fontes de dados internos e externos, que existem pelas várias entidades que geram informação estatística e geográfica, em vários formatos para os combinar num ambiente de armazenamento unificado (temporário ou permanente), que permite a realização de processos avançados com técnicas orientadas para a análise e ciência de dados .**

Palavras-chave – Data lake; caixas de areia; ciência de dados.

***Resumo* — Este artigo detalha a estratégia de pesquisa e tecnologia realizada para implementar um Data Lake e Sandboxes do Laboratório de Ciência de Dados do Instituto Nacional de Estatística e Geografia (INEGI) México, este projeto busca integrar informações digitais de diferentes repositórios, fontes de dados internas e externas, que existem pelas várias entidades que geram informação estatística e geográfica, em vários formatos para as combinar num ambiente de armazenamento unificado (temporário ou permanente), que permite a realização de processos avançados com técnicas orientadas para analytics e data science .**

Palavras-chave – data lake; caixa de areia; ciência de dados.

I. INTRODUÇÃO

A abordagem da estratégia é implementar um ambiente flexível que permita a incorporação de componentes baseados em ferramentas de hardware e software, com o objetivo de gerar um data lake que permita diferentes formatos de grandes volumes de informações estatísticas e geográficas digitais, coletadas, armazenadas e processados em ambiente com acesso controlado quanto à sua integridade, disponibilidade e confidencialidade, bem como a automatização dos processos com base em integração ágil e fluxos contínuos de implantação, para a criação de modelos de informação (dados, metadados, microdados), que servem de input para análises e ciência de dados.

II. ESTADO DA ARTE

Um data lake é uma área de armazenamento compartilhado para grandes volumes de dados estruturados e não estruturados com diferentes formatos, somente leitura, para interação de componentes de hardware e software, permitindo a geração de um conjunto de estratégias na criação de conhecimento. análise e ciência dos dados coletados, contribuindo para a descoberta e exploração de informações ad hoc em tempo real, para a geração de visualizações quantitativas ou qualitativas que apoiem a tomada de decisão.

Um sandbox ou sandbox é um ambiente controlado baseado em tecnologia da informação, escalável, para executar ferramentas especializadas, com níveis de segurança e de forma independente, permitindo o controle dos recursos de hardware e software, que são utilizados para acessar o data lake. a geração de protótipos de produtos, orientados em analytics e ciência de dados.

Alguns casos de sucesso onde é mostrada a implementação de um data lake em escritórios de estatística, é o trabalho desenvolvido por Llave (2018) [5], apresenta uma visão geral do status alcançado ao implementar a estratégia, neste estudo é explorar as tecnologias envolvidas, os benefícios diretos e indiretos, tendo como importante conclusão que data lakes não substituem data warehouses por sua sigla em inglês (Data Warehouses), mas sim aumentam capacidades, para utilização de grandes volume

Anejionu et al. (2019) [1], apresentam o SUDS ou Spatial Urban Data System que é utilizado no Reino Unido para realizar análises de dados sociais e econômicos em conjunto com um sistema geográfico, apresenta de forma simplificada o fluxo de informações dentro de um ambiente . de um data lake, além de apresentar uma série de casos de sucesso como resultado da implementação do referido modelo.

Por outro lado, Sfaxi & Ben Aissa (2020) [7], apresentam uma proposta de gerenciamento de dados em larga escala para tomada de decisão, no contexto particular do protótipo piloto, abordando um modelo de quatro fases: Na primeira, é apresentado um lago de dados brutos, ou seja, os dados em seu estado natural; posteriormente, há um lago de dados refinados, que servem a uma terceira camada, na qual as visõ

979-8-3503-9896-0/22/$31,00 ©2022 IEEE

a informação dos dados acima referidos, na quarta camada é a que persiste os resultados da fase de análise, acessíveis através de ferramentas de visualização. Em relação a um caso real de aplicação em um escritório de estatística, a CBS Statistics Netherlands (Holanda) implementou um data lake dentro da instituição, capaz de capturar os dados de entrada de várias fontes, combiná-los com outros dados gerados na CBS e produzir uma série de saídas destinadas a diferentes atores que poderiam ser produzidas graças ao poder de análise que está disponível ao combinar as fontes em um repositório comum.

III. FASES ESTRATÉGICAS

A Direcção do Laboratório de Ciência de Dados e Métodos Modernos de Produção de Informação, adjunta à Direcção Adjunta de Investigação do INEGI, adoptou uma metodologia de trabalho ágil, assente em oito fases abaixo descritas.

Fase das fontes de dados: consiste em selecionar as fontes de informação que serão relevantes para o projeto de ciência de dados em questão. A pessoa responsável por selecionar as fontes de informação terá a função de Arquiteto de Big Data. Caso o projeto seja desenvolvido por solicitação externa, o Solicitante será responsável por fornecer acesso a essas informações quando não for de acesso aberto. Essas fontes podem incluir dados, metadados, microdados de censos, pesquisas, imagens geoespaciais de satélite, entre outros.

Fase de extração e carregamento de dados: consiste em coletar os dados de entrada armazenados no data lake, para implementar técnicas de ciência de dados com base no requisito inicial do projeto, o papel responsável por esta fase é o Engenheiro de Dados considerando a implementação de métodos para extração de dados, automaticamente ou manualmente; acesso a repositórios internos e/ou externos, onde os dados são armazenados; upload de dados extraídos para a infraestrutura do data lake e aplicação das transformações de dados necessárias.

Fase de recuperação da informação: consiste na implementação de estratégias de busca nos conjuntos de dados, metadados, microdados, extraídos e carregados no data lake. As estratégias de busca podem ser feitas com base em estruturas identificadas e descritas por meio de dicionários de dados com técnicas que permitem gerar grupos de informações denominados "conjuntos de dados". Esta etapa pode ser realizada em diferentes momentos do projeto de acordo com as necessidades que surgirem. A função responsável é o Engenheiro de Dados.

Fase de tratamento de dados: consiste na implementação e aplicação de técnicas estatísticas e informáticas em que se procura analisar os dados recolhidos nas fases anteriores para ter uma melhor compreensão da informação, bem como descobrir aspetos relevantes dos dados como as relações existentes .entre as variáveis que estão sendo analisadas, o responsável por esta fase será o Cientista de Dados Jr., considerando a análise exploratória dos dados com base nas características do requisito; limpeza de dados, controle de qualidade das informações relacionadas ao projeto e representação de dados que atendam às necessidades da fase de construção do modelo.

Fase de construção do modelo: consiste na seleção dos modelos de aprendizagem computacional ou modelos estatísticos, conforme o caso, e sua implementação em ambiente controlado. Ao longo do projeto será possível experimentar diferentes modelos, o responsável é o Data Scientist Sr.

Fase de avaliação e validação dos resultados: a avaliação e validação dos resultados obtidos (a partir do tratamento e análise dos dados) será realizada com base nas métricas de desempenho estabelecidas, bem como em procedimentos estatísticos baseados nas melhores práticas internacionais. Pode ser realizada em diferentes momentos ao longo da execução do projeto de forma a ter critérios para selecionar o modelo com melhor desempenho, o papel responsável por esta fase será o Data Scientist Sr.

Fase de apresentação de resultados: consiste em fazer componentes através de ferramentas especializadas baseadas em tecnologias de informação, que permitem mostrar o andamento do projeto em ciência de dados, o papel responsável por esta fase será o Lead Data Scientist que pode considerar para a apresentação de resultados o demonstração de protótipos de produtos de dados dentro do data lake e sandbox; gerar um relatório técnico indicando a metodologia e nível de maturidade; construção de dashboards ou outras estratégias de visualização de dados e produção opcional de um artigo de investigação científica.

Fase de entrega do produto de dados: consiste no processo de implementação na arquitetura baseada em tecnologias de informação que o solicitante tenha acompanhado de sua documentação, adicionalmente, como parte do processo de entrega, pode ser incluída a consultoria para o uso do protótipo e a estratégia para reduzir riscos na operação com base no ciclo de vida do projeto.

Para o cumprimento das fases anteriores, são necessárias as seguintes funções descritas abaixo.

- Lead Data Scientist: responsável por supervisionar o desenvolvimento dos projetos, bem como dirigir a pesquisa científica, decidir os métodos e procedimentos da ciência de dados, estabelecer a estratégia de colaboração, dimensionar as necessidades de infraestrutura e promover o desenvolvimento de novas capacidades;
- Data Scientist Sr.: responsável por direcionar a implementação de métodos e procedimentos de ciência de dados e estabelecer as métricas de desempenho que serão necessárias para a avaliação e validação dos resultados;
- Cientista de Dados Jr.: responsável pela implementação de métodos e procedimentos, bem como pelo desenvolvimento de artefatos (dashboards de controle, mapas históricos, visualizações, relatórios) para apresentação de resultados;
- Arquiteto de Workflow: Responsável pela especificação do workflow de projetos de ciência de dados, bem como sua documentação;
- Arquiteto de infraestrutura: responsável pela configuração e administração da infraestrutura atribuída ao LCD;

979-8-3503-9896-0/22/$31,00 ©2022 IEEE

• Arquiteto de big data: responsável por definir a estratégia e plataforma tecnológica para coleta, armazenamento e processamento de grandes volumes de dados, necessários para atender aos objetivos particulares dos projetos de ciência de dados;

• Engenheiro de Dados: responsável por realizar tarefas de extração, carregamento e transformação de dados, bem como implementar estratégias de recuperação de informações exigidas por projetos de ciência de dados.

4. ESTRATÉGIA DE IMPLEMENTAÇÃO

Para a implementação do data lake e do sandbox, são utilizadas ferramentas tecnológicas (baseadas em software livre licenciado por sua sigla em inglês Open Source), que permitem realizar as fases anteriores descritas, considerando as camadas descritas a seguir.

Camada de Interoperabilidade: nesta camada estão as ferramentas transversais que permitem integração contínua, controle de mudanças e outras características para projetos baseados em Engenharia de Software de equipes de trabalho DevSecOps (Developer Security Operation) [3], bem como equipes de trabalho para projetos em Data Science DataSecOps (Data Security Operation) [3], o acesso ao data lake e sandboxes é restrito com níveis de segurança.

As ferramentas usadas nesta camada são:

• GitLab: plataforma web de desenvolvimento de software, que permite gerenciar código-fonte, planejamento de projetos, gerenciamento de fluxos automatizados de integração contínua. [8]

• MLflow: ferramenta de gerenciamento de projetos de aprendizado de máquina com a qual experimentos podem ser registrados, empacotados, distribuídos e consultados, a fim de obter modelos reprodutíveis e robustos. [9]

• Kedro: é um framework para criação de fluxos de dados. Adote as melhores práticas em engenharia de software para criar código de ciência de dados que seja reproduzível, sustentável e modular. [10]

Camada de ingestão de dados: a camada de ingestão de dados se encarrega de consumir e coletar praticamente qualquer tipo de fonte de dados estruturada e não estruturada interna e externa ao Instituto, neste nível são definidos os microsserviços - containers que serão implementados dentro do sandbox, do tipo (controladores, conectores, entre outras ferramentas ou bibliotecas para coleta de informações digitais).

As ferramentas usadas nesta camada são:

• Python: É uma linguagem de programação de código aberto orientada a objetos. Sua sintaxe simples e o número de bibliotecas disponíveis facilitam a criação rápida de programas e fluxos de informação.[11]

• Jupyter Lab - Notebook: É um ambiente de desenvolvimento web interativo com suporte para diferentes linguagens em que se destaca o Python. Permite combinar execução de código com documentação e visualizações em um único arquivo, facilitando assim a criação de protótipos e a divulgação dos resultados.[12]

Camada de armazenamento de dados: a camada é responsável pelo armazenamento de dados, define os serviços e/ou microsserviços em contêineres gerados para o Sandbox que permite, por meio de protocolos de transferência de dados, o envio de informações digitais, suportando vários formatos de dados estruturados e não estruturados.

A ferramenta utilizada nesta camada é o Minio: um software para servidores de armazenamento distribuído sob o protocolo chamado “S3”, que permite criar nuvens de dados privados de alto desempenho, bem como realizar interação padronizada com outros serviços para acessar seus arquivos. qualquer formato. [13]

Camada de virtualização de dados: a virtualização dos dados gerados no sandbox, com ferramentas que permitem consultas das informações digitais armazenadas no data lake, para estabelecer uma estratégia na geração de tabelas de dados somente leitura para que não percam a linhagem, para padronizar os nomes dos campos e permitir um conjunto homogêneo de dados.

As ferramentas usadas nesta camada são:

• Trino: motor de consulta SQL distribuído, que permite conectar-se a outros servidores de fontes de dados que possuem gerenciadores de banco de dados permitindo a interação entre eles dentro de uma única interface transparente e padronizada.[14]

• Hive: Este software permite ler, escrever e gerenciar conjuntos de dados em armazenamento distribuído usando a linguagem chamada SQL, permite atribuir estrutura a arquivos já existentes com os quais é possível consultar informações.[15]

Camada de integração de dados: os serviços ou microsserviços em containers gerados no sandbox, permitem a integração de tecnologias de gestão da informação de diversos provedores (gerenciadores de dados), permitindo a administração da informação digital armazenada (real ou virtual) no data lake. suporte de governo e documentação, entre outras características, a ferramenta utilizada são bibliotecas especializadas baseadas em Python [16] que permitem o agrupamento de informações para gerar um conjunto de dados.

Análise de dados e camada de ciência: a análise de dados permite que estudos complexos sejam realizados sobre os dados usando algoritmos emanados do processo de ciência de dados, no sandbox, eles permitem que as informações armazenadas no data lake sejam usadas como um grupo de entrada de dados para treinamento de modelos (aprendizado automatizado, processamento de linguagem natural, aprendizado profundo), os resultados gerados pelas iterações também são armazenados no data lake, para referência na geração de protótipos para revisão dos níveis de maturidade pelo grupo estabelecido pelo projeto, as ferramentas utilizadas estão em ambientes de trabalho (frameworks) e bibliotecas especializadas baseadas em Python [16] que permitem o processamento com técnicas de aprendizado de linguagem supervisionado, não supervisionado, profundo e natural para gerar algoritmos que respondam ao requisito.

Camada de visualização de dados: para visualização de dados é o meio pelo qual ferramentas especializadas

979-8-3503-9896-0/22/$31,00 ©2022 IEEE

Podem gerar relatórios, gráficos ou quadros de controle que permitem mostrar os indicadores quantitativos e qualitativos para tomar a decisão sobre o nível de maturidade do projeto com ciência de dados para posteriormente realizar a entrega do protótipo.

As ferramentas usadas nesta camada são:

• SuperSet: é uma plataforma de exploração e visualização de dados que pode ser conectada a qualquer fonte de dados baseada em SQL através do componente chamado SQLAlchemy, suportando grandes volumes de informações com base em escala de petabytes, possui uma variedade de gráficos e mapas para escolher dependendo o tipo de necessidade de exibição dos dados, através de seletores é configurado para ser apresentado em um painel de controle baseado em tecnologia web permitindo o acesso controlado do usuário. [17]

• D3.js (Data-Driven Documents): é uma biblioteca JavaScript para manipulação de documentos baseados em dados especializados em visualização com tecnologias web.[18]

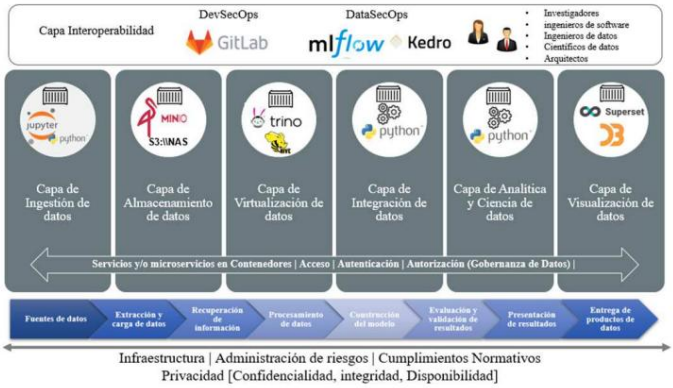

Figura 1. Diagrama de fases, camadas e ferramentas estratégicas utilizadas na arquitetura do data lake e sandboxes.

Nota: a infraestrutura necessária para a implementação de um data lake e sandbox depende das características dos requisitos para os projetos e serviços que são gerados, para os quais a seguir é um exemplo de dimensionamento básico, gerando quatro servidores físicos ou lógicos, cada um servidor possui 20 núcleos 40 processadores lógicos, 256 Gb de memória RAM, 400 Gb de armazenamento tipo SSD e 4 Tb de armazenamento tipo SATA, configurado como um sistema operacional multiusuário baseado no kernel Linux para suportar as ferramentas mencionadas, com tecnologia que permite tecnologia de conteinerização ou hypervisor, os métodos de conexão entre os servidores devem considerar o uso de tecnologias de rede de área de armazenamento SAN (Storage Area Network) e/ou Network Attached Storage NAS (Network Attached Storage), com o objetivo de ter uma alta disponibilidade , estratégia escalável e segura para reduzir riscos na operação como parte dos processos.

V. CONCLUSÕES

Experiência na geração, implementação e administração de um data lake e sandboxes dentro do Instituto gera valor agregado a processos baseados em analytics e data science considerando grandes volumes de informações estatísticas e geográficas em diferentes formatos, bem como consolidar um repositório para gerar produtos e serviços orientados a visualizações quantitativas e qualitativas que permitem a tomada de decisões.

OBRIGADO

Grupo Colaborativo Laboratório de Ciência de Dados do INEGI

• VILLASEÑOR GARCIA ELIO ATENOGENES
elio.villasenor@inegi.org.mx

• COROADO 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 trabalho.

• DIAZ EDGAR OSWALDO oswaldo.diaz@inegi.org.mx

REFERÊNCIAS BIBLIOGRÁFICAS

[1] Anejionu, OCD, Thakuriah, P. (Vonu), McHugh, A., Sun, Y., McArthur, D., Mason, P., & Walpole, R. (2019). Sistema de dados urbanos espaciais: uma infraestrutura de big data habilitada para nuvem para análise urbana social e econômica. Future Generation Computer Systems, 98, 456–473. https://doi.org/10.1016/j.future.2019.03.052 [2]

Ashofteh, A., & Bravo, JM (2021). Formação em ciência de dados para estatísticas oficiais: Um novo paradigma científico de desenvolvimento de informação e conhecimento nos sistemas estatísticos nacionais. Jornal Estatístico da IAOS, 37(3), 771–789. https://doi.org/10.3233/sji-210841 [3]

Díaz, O., Muñoz, M., & Mejía, J. (2019). Infraestrutura responsiva com segurança cibernética para processos DevSecOps automatizados de alta disponibilidade. 2019 8ª Conferência Internacional sobre Melhoria de Processos de Software (CIMPS), 1–9. https://doi.org/10.1109/CIMPS49236.2019.9082439 [4]

Jimenez-Marquez, JL, Gonzalez-Carrasco, I., Lopez-Cuadrado, JL, & Ruiz-Mezcua, B. (2019). Rumo a uma estrutura de big data para análise de conteúdo de mídia social. Jornal Internacional de Gerenciamento de Informações, 44, 1–12. https://doi.org/https://doi.org/10.1016/j.ijinfomgt.2018.09.003 [5] Key,

MR (2018). Data lakes em inteligência de negócios: relatórios das trincheiras. Prossegue Ciência da Computação, 138, 516–524. https://doi.org/10.1016/j.procs.2018.10.071

[6] Mankins, J. (1995). Nível de prontidão tecnológica – um white paper.

[7] Sfaxi, L., & Ben Aissa, MM (2020). DECIDE: Uma metodologia ágil de design orientada a eventos e dados para projetos de Big Data decisórios. Data and Knowledge Engineering, 130 (julho de 2019), 101862. https://doi.org/10.1016/j.datak.2020.101862 [8] Documentação para

instalação da ferramenta na camada de dados chamada de interoperabilidade GitLab (https://docs.gitlab.com/ee/install/requirements.html)

[9] Documentação para instalação da ferramenta na camada de interoperabilidade chamada MLflow (https://mlflow.org/docs/latest/quickstart.html#installing-mlflow)

[10] Documentação para a instalação da ferramenta na camada de interop chamado Kedro e permitindo conexão com MLflow (https://kedro

mlflow.readthedocs.io/en/stable/source/02_installation/index.html)

[11] Documentação para a instalação e implementação dos módulos para a linguagem Python (https://docs.python.org/es/3/installing/index.html#installing-into-the-system-python-on-linux)

[12] Documentação para implementar o IDLE chamado Jupyter Lab usado como ferramenta para interoperar com python (https://docs.jupyter.org/en/latest/install.html)

[13] Documentação para implementar a ferramenta de armazenamento chamada Minio com o protocolo de transferência S3 (https://docs.min.io/minio/baremetal/)

[14] Documentação para implementar o componente que permite o conectividade com fontes de informação relacionadas a gerenciadores de dados chamados trino (https://trino.io/docs/current/installation/deployment.html)

[15] Documentação para implementar o componente que permite a virtualização de dados chamado Hive (https://cwiki.apache.org/confluence/display/Hive//GettingStarted#GettingStarted-InstallingHivefromaStableRelease)

[16] Documentação para implementar as bibliotecas de código que permitem o gerenciamento de análise e ciência de dados (https://docs.python.org/es/3/library/)

[17] Documentação para instalação da ferramenta de visualização chamada Superset (https://superset.apache.org/docs/installation/installing superset-from-scratch)

[18] Documentação para implementar código baseado em javascript que permite a visualização de dados chamado D3js (https://d3js.org/)