UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE GESTÃO E NEGÓCIOS
GRADUAÇÃO EM GESTÃO DA INFORMAÇÃO

LUCIANO DE OLIVEIRA

Data lake na Gestão de Barragens de Usinas Hidrelétricas

ORIENTADOR: PROF. DR. JOSÉ EDUARDO FERREIRA LOPES

UBERLÂNDIA – MG
2024

LUCIANO DE OLIVEIRA

Data lake na gestão de barragens Usinas Hidrelétricas

Monografia apresentada ao Curso de Graduação em Gestão da Informação, da Universidade Federal de Uberlândia, como exigência parcial para a obtenção do título de Bacharel.

Orientador: Prof. Dr. José Eduardo Ferreira Lopes

UBERLÂNDIA – MG

2024

RESUMO

Objetivou-se com este relato tecnológico apresentar a implantação de um *Data Lake* na nuvem para unificar e automatizar os processos de coleta, armazenamento e análise de dados, visando otimizar a tomada de decisões operacionais e estratégicas da empresa. Como situação problema, destaca-se a fragmentação dos sistemas de coleta de dados e a ausência de automação, comprometendo a eficiência na gestão das informações. Como solução, implantou-se um *Data Lake* na plataforma *Azure*, utilizando ferramentas como *Azure Data Factory, Azure Datab*ricks e *Power BI* para centralizar e automatizar os processos de coleta e análise de dados. Como resultados alcançados, destacam-se a redução do tempo de processamento, a melhoria na qualidade dos dados e a centralização das informações para apoio às decisões.

Palavras-chave: *Data Lake*; *Azure*; Armazenamento.

SUMÁRIO

1 INTRODUÇÃO

No contexto industrial contemporâneo, onde a quantidade de dados gerados tem aumentado exponencialmente (Somasundaram; Shrivastava, 2011), empresas de grande porte do setor produção de energia enfrentam desafios significativos para coletar, transformar e analisar esses dados. Tal dificuldade decorre da dispersão dos dados, que são capturados a partir de diferentes origens, apresentando estruturas e formatos diversos, além de serem armazenados em múltiplos sistemas.

Inicialmente, as empresas utilizavam bancos de dados para armazenar e analisar os dados de forma simples e confiável. Contudo, com o crescimento exponencial do volume de dados, muitos desses bancos tornaram-se segmentados, criando "silos de dados" que dificultava a visão integrada das informações. Para solucionar esse problema, surgiram os *Data Warehouses*, centralizando dados de diferentes fontes em um repositório único. No entanto, com a evolução das demandas, especialmente a necessidade de armazenamento e análise de dados não estruturados, a extração de informações em tempo real e o uso de *Machine Learning*, os *Data Warehouses* apresentaram limitações importantes (Alura, 2023).

Dessa forma, torna-se necessário uma solução tecnológica que permita a unificação dos dados em uma plataforma centralizada. Nesse contexto, o conceito de *Data Lake*, proposto por James Dixon (Dixon, 2010) apresenta-se como uma solução robusta, permitindo que os dados sejam armazenados de forma segura e escalável em um *Data Lake*. Essa abordagem facilita a consolidação dos dados provenientes de diferentes sistemas, criando um repositório único que suporta análises avançadas e uma exploração eficiente das informações, contribuindo para uma gestão mais precisa e tomada de decisão baseada em dados.

Assim, este relato tecnológico tem como objetivo apresentar a implantação de um *data lake* na nuvem para unificar e automatizar os processos de coleta, armazenamento e análise de dados, visando otimizar a tomada de decisões operacionais e estratégicas da empresa.

2 FUNDAMENTAÇÃO TEÓRICA

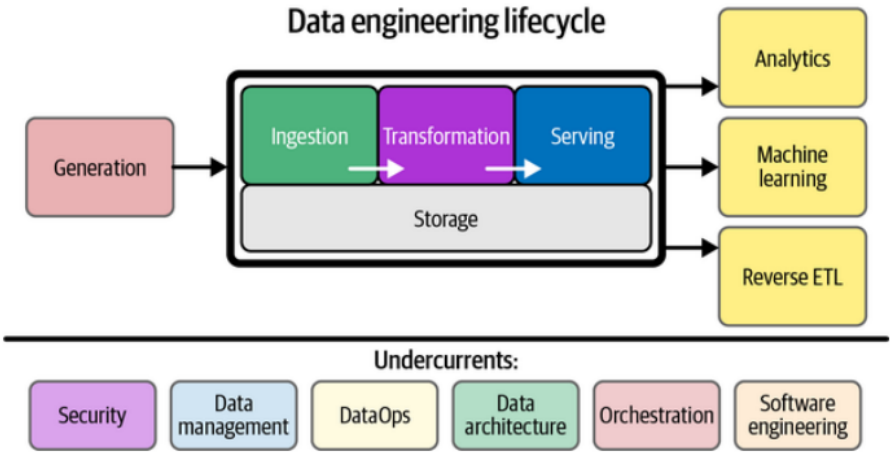
Para fundamentar este estudo, foi necessário organizar e explorar conceitos avançados de engenharia de dados, *Data Lake* e *Microsoft Azure Cloud*, além de ferramentas específicas como *Azure Data Factory*, *Azure Databricks*, *Azure Synapse* e *Power BI* empregadas no projeto.

2.1 ENGENHARIA DE DADOS

A engenharia de dados consiste no processo de desenvolver, implementar e gerenciar sistemas e procedimentos que transformam dados brutos em informações relevantes. A jornada da engenharia de dados abrange as etapas de geração, armazenamento, provisão, transformação e entrega dos dados visto na Figura 1 (Reis;Housley, 2022, *apud* Fernandes,2023). No entanto, algumas dessas etapas, como a transformação e o carregamento de dados, são muito repetitivas e consomem tempo. Nessa perspectiva os pipelines de dados surgem como solução para esses desafios, automatizando as tarefas de transformação e movimentação dos dados.

> Assim como uma tubulação de água move a água do reservatório para suas torneiras, um pipeline de dados move os dados do ponto de coleta para o armazenamento. Um pipeline de dados extrai dados de uma fonte, faz alterações e os salva em um destino específico (Amazon Web Services, 2024).

Figura 1 - *Data Engineering Lifecycle*



Fonte:(Reis;Housley, 2022, *apud* Fernandes,2023)

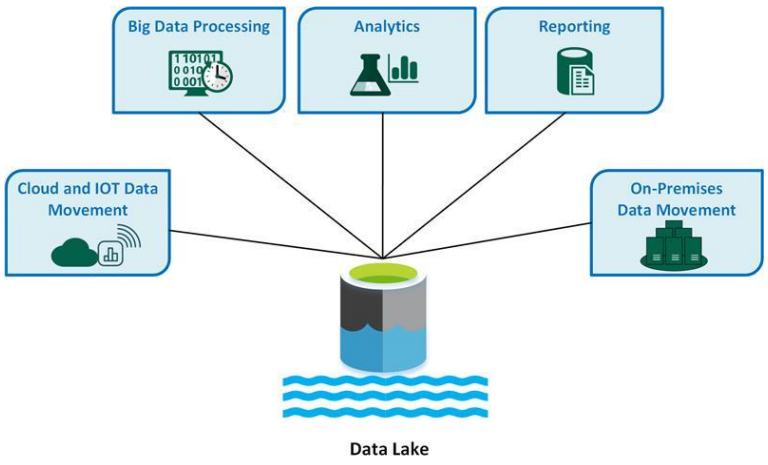
2.2 DATA LAKE

O conceito de D*ata Lake* foi inicialmente introduzido por James Dixon, CEO da *Pentaho*, e comparado a um lago que armazena dados de forma integral e "crua". Segundo Dixon:

> Se você pensar em um *datamart* como uma água engarrafada – limpa, empacotada e estruturada, fácil para consumo – o *Data Lake* é um imenso corpo de água num estado mais natural O conteúdo do *Data Lake* flui de uma fonte para encher o lago, e vários usuários do lago podem vir para examinar, mergulhar ou coletar amostras. (Dixon, 2010).

Um *Data Lake* é um repositório central de armazenamento que contém grandes volumes de dados em seu formato original, otimizando a escalabilidade para suportar *terabytes* a *petabytes*. Os dados, provenientes de diversas fontes, conforme ilustrado na Figura 1, podem ser estruturados, semiestruturados ou não estruturados, sendo armazenados sem transformações prévias. Diferentemente do *data warehouse*, que transforma e processa os dados durante a ingestão por meio de *ETL (Extract, Transform, Load)*, o *Data Lake* utiliza a abordagem *ELT (Extract, Load, Transform)*, na qual os dados são extraídos e carregados em seu formato bruto, e a transformação ocorre apenas durante a consulta, quando necessário (Microsoft, 2024a). Vale destacar que, na ausência de uma gestão e organização adequadas, o *Data Lake* corre o risco de se tornar um "*Data Swamp*" (pântano de dados), onde a desorganização e a falta de estrutura resultam em dados que perdem valor e dificultam a extração de insights (Alura, 2024b). Portanto, políticas de governança, processos rigorosos de qualidade de dados e estratégias de segurança são essenciais para garantir que o Data Lake atue como um repositório confiável e útil de informações para a organização.

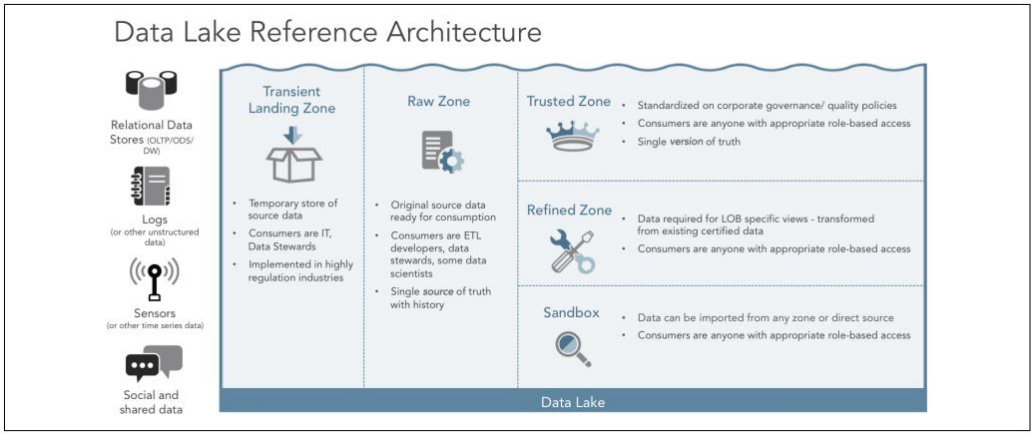
Figura 2 - *Data Lake*



Fonte: Microsoft (2024a)

A arquitetura de um *Data Lake* é composta por múltiplas camadas que desempenham um papel fundamental na organização e no processamento dos dados, sendo segmentadas em diferentes zonas de acordo com o nível de refinamento deles. Tal segmentação possibilita uma gestão mais eficiente, abrangendo desde a ingestão de dados brutos até a disponibilização de dados refinados e prontos para análise. A arquitetura de referência proposta por Zaloni, que exemplifica esse modelo, é apresentada na Figura 2, conforme descrito por Sharma (2018).

Figura 3 - Arquitetura *Data Lake*



Fonte: Sharma (2018, p 16)

A primeira zona, chamada de *Transient Landing Zone*, é destinada ao carregamento temporário dos dados, onde são realizadas verificações básicas de qualidade e medidas de segurança antes de seu armazenamento definitivo. Em seguida, os dados são movidos para a *Raw Zone*, onde permanecem em seu formato original, funcionando como a "fonte única da verdade". A *Trusted Zone* é responsável pela aplicação de políticas de conformidade e validação, transformando os dados brutos em um formato padronizado, pronto para consumo por meio de análises e relatórios. Já a *Refined Zone* refina os dados conforme as necessidades específicas de cada linha de negócio, garantindo sua adequação para o uso final. Por fim, o *Sandbox* oferece um ambiente flexível para exploração e experimentação de dados, permitindo que cientistas de dados e gestores conduzam análises sem comprometer o ambiente de produção.

2.3 FERRAMENTAS

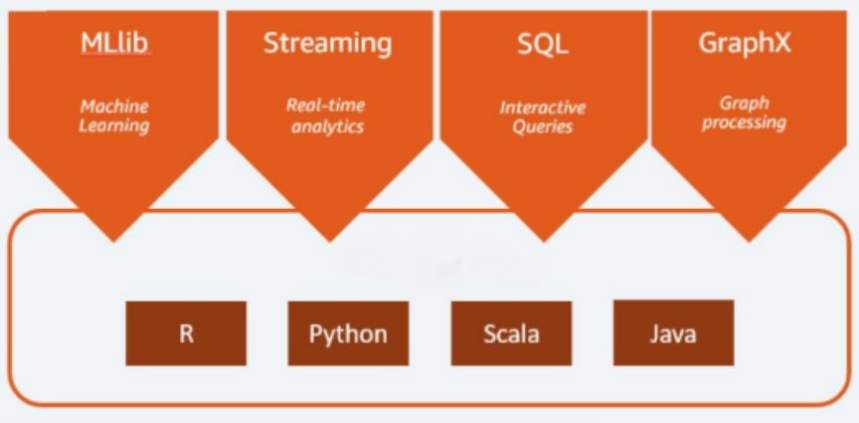
2.3.1 MICROSOFT AZURE

A *Microsoft Azure* é uma plataforma de nuvem que oferece uma ampla gama de serviços para o desenvolvimento e gestão de operações digitais por empresas e organizações, incluindo a hospedagem de máquinas virtuais, armazenamento de dados, análise de *Big Data*, inteligência artificial e aprendizado de máquina. Com alta escalabilidade, flexibilidade e segurança, a plataforma permite o ajuste dos recursos conforme a demanda, garantindo disponibilidade e continuidade dos serviços. Dessa forma, a *Microsoft Azure* facilita a transformação digital, otimizando processos, reduzindo custos e promovendo inovação tecnológica, contribuindo significativamente para a eficiência operacional e a competitividade organizacional (Microsoft, 2024b).

Além do mais disponibiliza uma variedade de ferramentas e serviços para a criação e implementação de infraestruturas de *Data Lake*, com o objetivo de viabilizar a ingestão, armazenamento, processamento e análise eficiente de grandes volumes de dados. Dentre as principais ferramentas disponíveis na plataforma, destacam-se *Azure Databricks, Azure Data Factory, Azure Synapse* e *Power BI.*

- ***Azure Databricks*** é uma plataforma de processamento de dados em larga escala, desenvolvida sobre o *framework Apache Spark*. Essa plataforma é amplamente utilizada para processar e transformar grandes volumes de dados de forma distribuída e em tempo real (Microsoft, 2024c). O *Apache Spark* fornece uma abstração de programação sofisticada, que possibilita o processamento paralelo e distribuído em clusters de computadores, além de oferecer suporte a diversas linguagens de programação, como *Scala, Java, Python, SQL* e *R*. Ademais, o Spark conta com bibliotecas específicas para aprendizado de máquina (*MLlib*), processamento de fluxos de dados (*Spark Streaming*) e processamento gráfico (*GraphX*), conforme ilustrado na Figura 3, tornando o *Databricks* uma solução eficaz para análises complexas e o processamento de grandes volumes de dados (Databricks, 2024).

Figura 4 - *Spark Core*

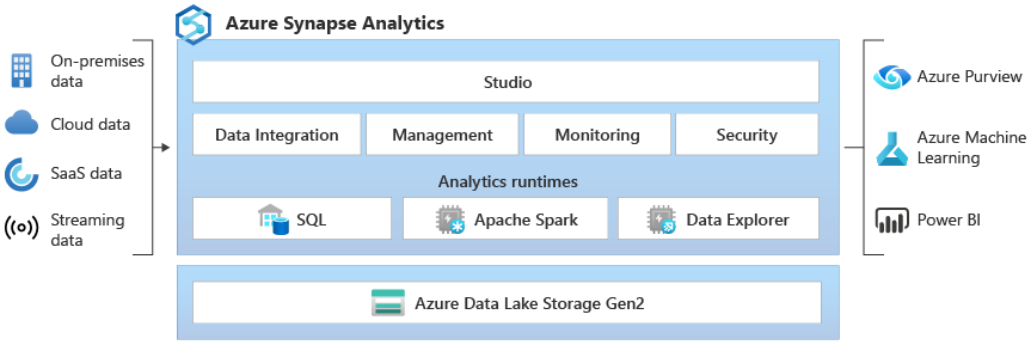


Fonte: Databricks (2024)

- ***Azure Data Factory*** é um serviço de integração de dados em nuvem que permite a orquestração e automação de *pipelines* para a movimentação e transformação de dados em larga escala. Ele possibilita a execução de processos complexos de extração, transformação e carregamento (ETL), com integração a serviços como *Azure HDInsight*, *Azure Databricks* e o *Banco de Dados SQL do Azure* (Microsoft, 2024d).

- ***Azure Synapse*** é uma plataforma de análise unificada que visa acelerar a obtenção de insights por meio da combinação de tecnologias avançadas para processamento de grandes volumes de dados. A plataforma possibilita a execução de consultas *SQL*, análise distribuída e processamento de dados em tempo real e em *batch*, integrando-se serviços do *Azure*, como *Power BI*, *CosmosDB* e *Azure Machine Learning*, conforme ilustrado na Figura 4 (Microsoft, 2024e).

Figura 5 - *Arquitetura do Azure Synapse Analytics*



Fonte: Microsoft (2024e)

- ***Power BI*** é uma plataforma abrangente de visualização de dados composta por um conjunto de serviços de *software*, aplicativos e conectores. Ela possibilita a criação de dashboards interativos e relatórios customizados, que funcionam em conjunto para transformar fontes de dados heterogêneas em informações consistentes, visuais e interativas. Esses dados podem estar armazenados em planilhas de Excel ou em *data warehouses* híbridos, sejam locais ou baseados na nuvem (Microsoft, 2024f).

3 CONTEXTO INVESTIGADO E SITUAÇÃO PROBLEMA

A empresa estudada é uma organização de grande porte, pertencente ao setor de geração e distribuição de energia elétrica, com operações espalhadas por diversas regiões do Brasil. Com mais de duas décadas de atuação, a companhia é responsável pela gestão de várias usinas hidrelétricas, que variam em porte e capacidade de geração. Entre essas usinas estão grandes Usinas Hidrelétricas de Energia (UHEs) com capacidade de produção acima de 30 MW/h, Pequenas Centrais Hidrelétricas (PCHs) com capacidade de produção entre 05 e 30 MW/h e Centrais Geradoras Hidrelétricas (CGHs) com potencial de produção de até 05 MW/h. Cada uma dessas unidades opera com sistemas específicos para coleta e análise de dados, resultando em uma estrutura de *TI* complexa e fragmentada, o que torna desafiador o gerenciamento integrado de informações.

A crescente demanda por energia e a expansão das operações da empresa contribuíram para um aumento significativo no volume de dados gerados. No entanto, a arquitetura atual de integração de dados, baseada em processos manuais e descentralizados, mostrou-se insuficiente para lidar com o fluxo intenso e variado de informações necessárias ao suporte da tomada de decisões estratégicas e operacionais. A falta de automação e de um sistema unificado comprometem a eficiência e a eficácia na análise dos dados.

Durante a etapa de entendimento do processo de extração, disponibilização e análise de dados, foram identificados quatro grandes grupos de dados essenciais para a operação: dados de meteorologia, dados de telemetria, dados operacionais e dados de monitoramento das estruturas civis. Cada um desses grupos de dados possui uma combinação de métodos e ferramentas para coletar informações, incluindo *APIs*, rotinas de raspagem de dados *(scrapers)*, consultas a bancos de dados relacionais e planilhas eletrônicas. A coleta de dados meteorológicos é realizada manualmente no site *Windy.com*, através de um scraper desenvolvido em *Python*, que coleta informações de

precipitação acumulada para o dia atual e os seis dias seguintes, gerando 54 arquivos .csv. Esses arquivos são salvos em uma pasta do *SharePoint* e usados como base para cálculos e geração de boletins meteorológicos em *PDF* e dashboards no *Power BI*, acessíveis internamente. Os dados de telemetria das usinas são descentralizados e variam conforme o porte das usinas. UHEs e PCHs coletam e fornecem dados para a Agência Nacional de Águas (ANA), via Hidroweb, enquanto CGHs utilizam a *API* da Construserv. Existe redundância na coleta de dados, sendo necessário acessar tanto a *API* da Construserv quanto a do Hidroweb, além de inconsistências nos dados históricos que prejudicam as análises de desempenho e segurança.

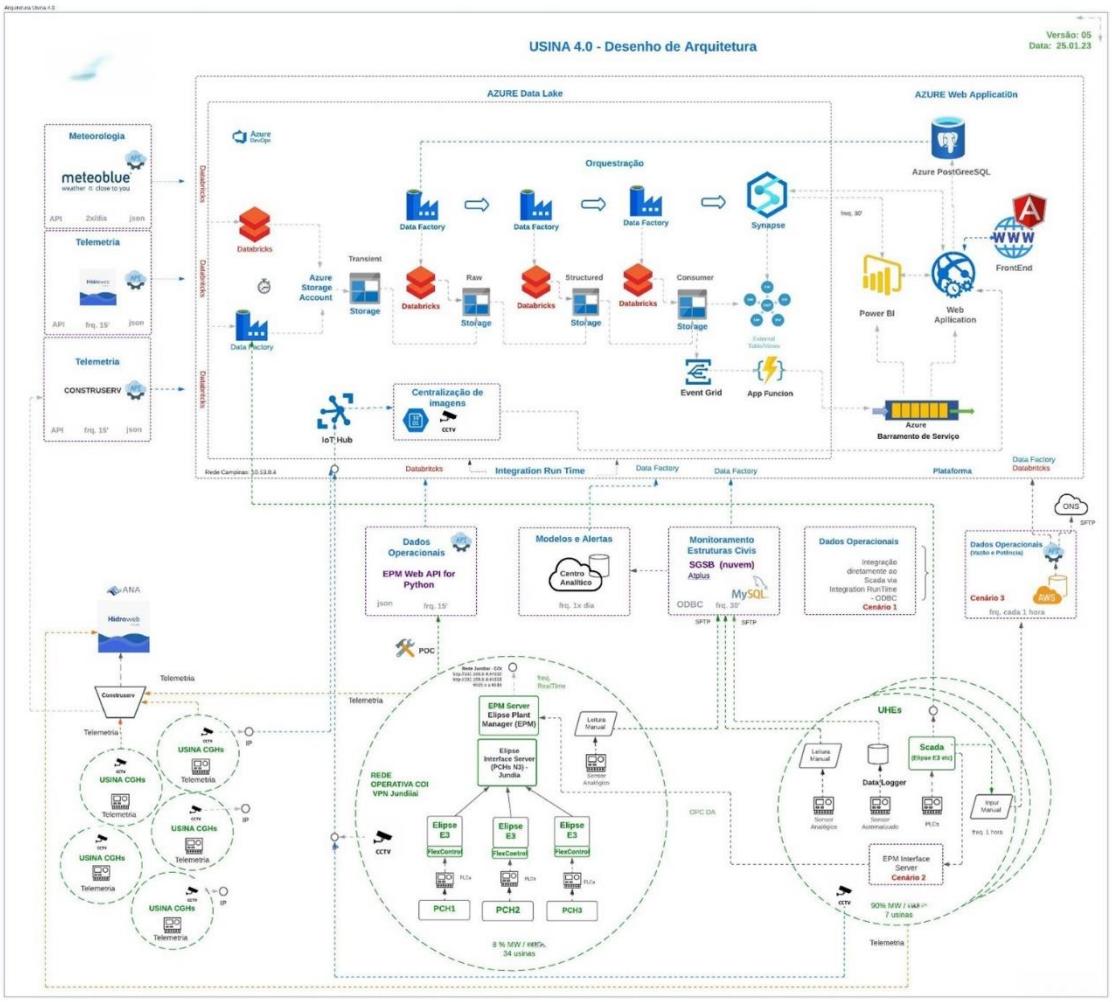
A falta de padronização e centralização dificulta a obtenção de uma visão integrada, essencial para decisões informadas em situações críticas. Para os dados operacionais, a empresa utiliza o sistema SCADA Elipse 3.0, que é implementado centralmente nas PCHs e de forma descentralizada nas UHEs, gerando diferentes níveis de integração. CGHs não utilizam SCADA devido ao custo e à falta de obrigatoriedade legal.

Os dados de monitoramento das estruturas civis tem como origem o Sistema Geral de Segurança de Barragens (SGSB), que está hospedado na nuvem e recebe dados via FTP/SFTP, o que aumenta a complexidade da integração. A falta de padronização dos dados históricos gera divergências e lacunas, prejudicando a avaliação precisa dos indicadores de desempenho e segurança. Além disso, a ausência de uma plataforma centralizada impede a realização de análises em tempo real, essenciais para a gestão eficiente das operações das usinas. Sem uma visão integrada e atualizada, a capacidade da empresa de tomar decisões rápidas e informadas é prejudicada, principalmente em situações críticas que exigem respostas imediatas.

4 INTERVENÇÃO ADOTADA

Para solucionar os desafios apresentados, foi proposta a implementação de um *Data Lake*, baseado na plataforma *Azure*, com o intuito de centralizar, automatizar e padronizar o processo de coleta, armazenamento e análise dos dados gerados pelas diferentes usinas e sistemas da empresa. Essa abordagem visa promover uma integração eficiente entre os distintos sistemas de coleta de dados, facilitando a gestão e o processamento das informações de maneira centralizada e segura. A Figura 5 ilustra a arquitetura do projeto, destacando os componentes e suas respectivas interações.

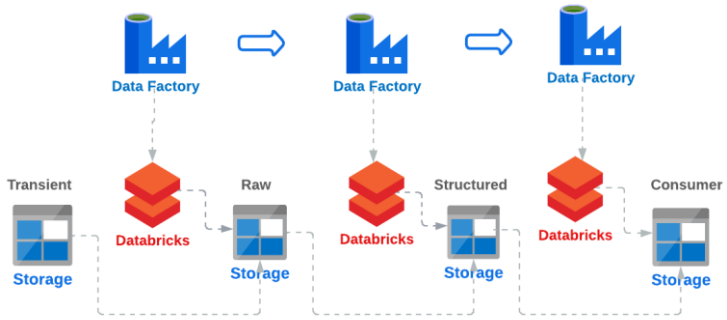
Figura 6 - Arquitetura do projeto



Fonte: Empresa Estudada

Os dados serão coletados de diversas fontes, incluindo sistemas internos, como SGSB, SCADA/Elipse, Telemetria Construserv e aplicações Web, bem como fontes externas, como Hidroweb e Meteorologia-Meteoblue. A ingestão dos dados será realizada por meio do *Azure Data Factory*, que fará a conexão com as bases de dados e *APIs*, agendando a atualização dos dados em intervalos regulares (15 minutos ou conforme a necessidade). *O Azure Databricks* será utilizado para processar e transformar esses dados, que serão armazenados no *Data Lake* em diferentes camadas (*raw*, *structured* e *consumer*). A Figura 7 mostra as camadas da malha de dados, detalhando como os dados são transformados e organizados ao longo do processo de ingestão.

Figura 7 - Malha de dados



Fonte: Empresa Estudada

A orquestração e a malha de ingestão dos dados serão realizadas utilizando o *Azure Data Factory*. A malha de ingestão será responsável pela execução automática, ordenada e programada dos *pipelines*, por meio da criação de *triggers* que definem a frequência de execução desses *pipelines*, movimentando os dados pelas diferentes camadas do *Data Lake*. Os *pipelines* executarão *notebooks* desenvolvidos em *PySpark* no Azure *Databricks*, aplicando as transformações necessárias conforme os dados transitam entre as camadas:

- **Ingestão e integração de fontes de dados (Camada *Raw*)**: Nesta etapa, os dados são ingeridos na camada *raw* de forma incremental, exceto pela primeira carga, que abrange dados históricos. A ingestão incremental garante a eficiência no armazenamento e no processamento dos dados ao longo do tempo.

- **Higienização e padronização (Camada *Structured*)**: Nesta fase, os dados passam por processos de limpeza, padronização e enriquecimento, sendo então armazenados na camada *structured*. Esse processamento incremental assegura que os dados estejam prontos para análises detalhadas e reduz o risco de inconsistências.

- **Transformação e agregação (Camada *Consumer*)**: Na última etapa, os dados são transformados em informações úteis por meio de agregações, cálculos de indicadores e cruzamentos de dados. A camada *consumer* é reprocessada a cada nova ingestão, garantindo a consistência e a integridade dos dados ao longo do tempo, proporcionando uma base confiável para análises avançadas.

Após a execução da malha de ingestão, os dados estarão disponíveis no *Azure Synapse* para consumo. Será utilizado o *Azure Synapse Serverless* para disponibilizar os dados estruturados no formato *Delta*, armazenados na camada consumer. Por fim, o *Synapse* deverá ter acesso de leitura à *Azure Storage Account*.

Nenhum processamento será realizado no *Synapse*, que será empregado exclusivamente para a integração com o *Power BI*. Para facilitar o acesso aos dados, serão criadas tabelas externas *(External Tables)* e *views*. As tabelas externas no *Synapse* permitem acessar dados em armazenamento externo sem a necessidade de carregá-los para o ambiente. Já as *views* são consultas que estruturam os dados para simplificar o acesso, sem armazenamento.

Essas tabelas e views apontarão para os arquivos *Delta (parquet)* processados na camada *consumer*, permitindo que as informações sejam acessadas e visualizadas de forma eficiente por meio de dashboards no *Power BI*.

5 RESULTADOS ALCANÇADOS (OU ESPERADOS)

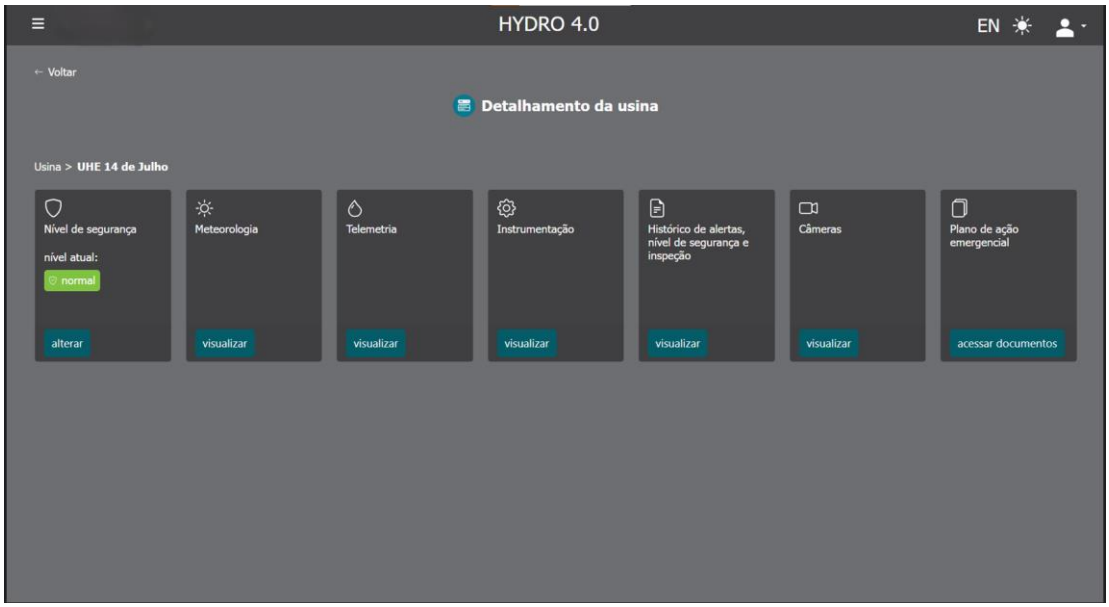
Com a implementação do *Data Lake*, como primeiro passo para estabelecer uma governança de informações, espera-se alcançar resultados como a centralização dos dados das usinas, promovendo uma visão integrada e facilitando a tomada de decisões. Além disso, foram desenvolvidos *dashboards* no *Power BI* que oferecem análises detalhadas das diversas fontes de dados, como Meteorologia, Telemetria e Dados Operacionais, aumentando a eficiência no monitoramento das operações.

A automação da coleta e integração de dados reduziu a dependência de processos manuais, minimizando falhas e inconsistências. Ademais, o uso de tecnologias como *Azure Data Factory* e *Azure Databricks* possibilitou o processamento eficiente dos dados, garantindo que as informações fossem atualizadas e disponibilizadas em tempo real para os gestores.

Os *dashboards* desenvolvidos, como o de detalhamento das usinas (Figura 8) e o mapa meteorológico (Figura 9), fornecem uma visualização clara das condições operacionais e climáticas, permitindo a adoção de ações preditivas e preventivas. Tais *dashboards* reúnem indicadores fundamentais relacionados à saúde das barragens, além

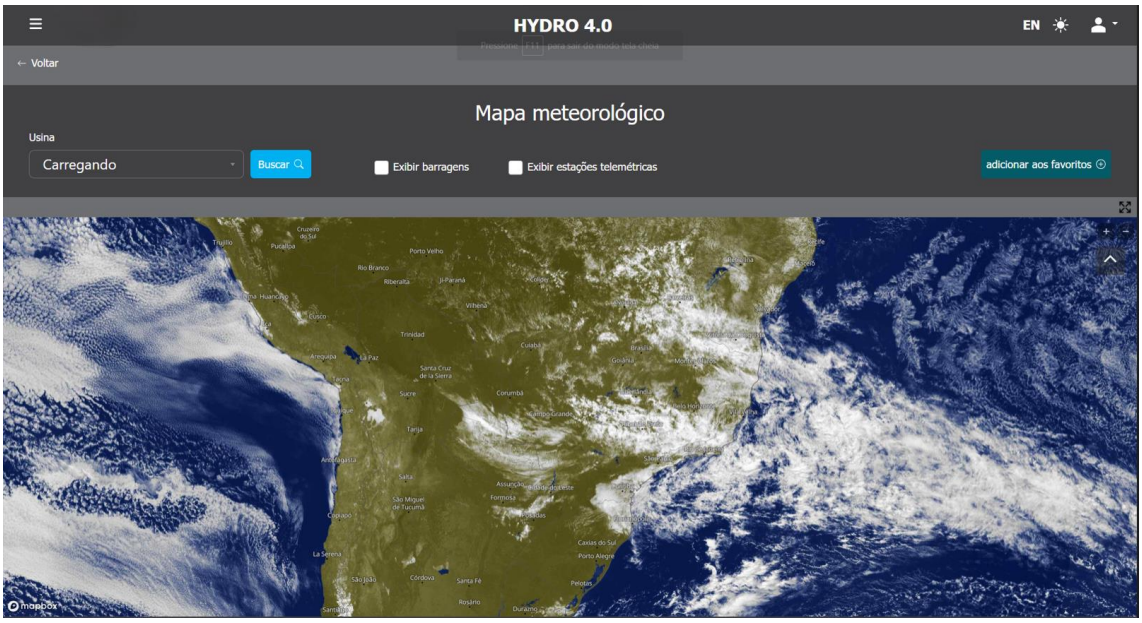
de análises gráficas meteorológicas com amplitude nacional e monitorar automaticamente a leitura dos equipamentos de medição.

Figura 8 - Detalhamento das Usinas



Fonte: Empresa Estudada

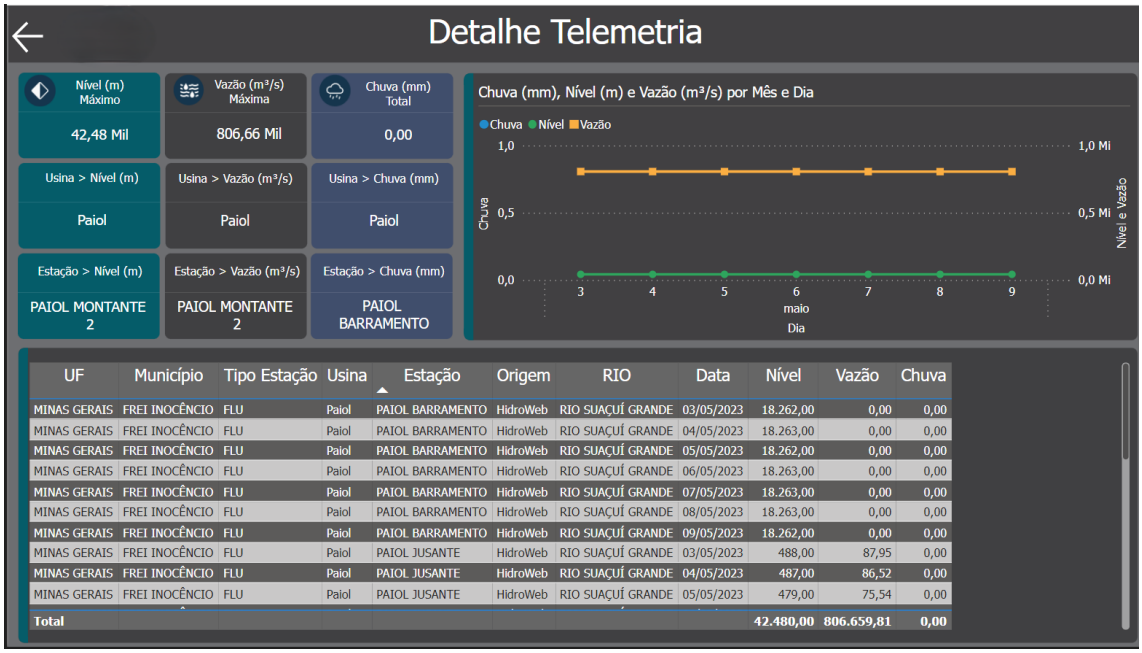
Figura 9 - Mapa Meteorológico



Fonte: Empresa Estudada

Também foram criados *dashboards* de telemetria (Figura 10) que fornece informações sobre nível, vazão e chuva em diferentes estações, bem como uma tabela detalhada contendo informações como município, tipo de estação, data, nível, vazão e chuva e instrumentação, que viabilizam o acompanhamento contínuo dos equipamentos e instrumentos das usinas.

Figura 10 - Telemetria



Fonte: Empresa Estudada

Esses *dashboards* permitiu simplificar e agilizar a compreensão de dados complexos, favorecendo uma tomada de decisão mais ágil, bem como a análise de tendências, padrões e *insights* de gestão. Um dos principais destaques do projeto são as interfaces visuais intuitivas e centralizadas, que foram desenvolvidas para proporcionar essa facilidade de uso, conforme ilustrado na Figura 11.

Figura 11 - *Dashboards* Barragem

Fonte: Empresa Estudada

Com isso, a empresa passou a utilizar os dados de forma estratégica, aumentando a confiabilidade das informações e viabilizando análises mais aprofundadas, com vistas a melhorar a segurança e a eficiência das operações. A padronização dos dados e a centralização em um repositório único proporcionaram uma estrutura escalável para futuras melhorias contínuas.

Dessa forma, os resultados esperados foram alcançados, evidenciando a eficácia da solução proposta para proporcionar uma gestão integrada e eficiente dos dados das usinas.

6 CONSIDERAÇÕES FINAIS

Objetivou-se, com este relato tecnológico, apresentar a implantação de um *Data Lake* na nuvem para unificar e automatizar os processos de coleta, armazenamento e análise de dados, visando otimizar a tomada de decisões operacionais e estratégicas da empresa. Os resultados esperados foram alcançados, promovendo a centralização dos dados, a redução de inconsistências e a maior eficácia na tomada de decisões. No entanto, desafios relacionados à padronização dos dados históricos e à integração de diferentes sistemas se mostraram complexos, requerendo soluções criativas e ajustes ao longo do

processo. A implementação do *Data Lake* também abriu novas oportunidades para a exploração de dados em tempo real e análises preditivas, trazendo aprendizados significativos em relação à gestão e à segurança das operações da empresa. Com base nesse projeto, evidencia-se a importância da adoção de soluções tecnológicas para a melhoria contínua e a inovação na gestão de dados industriais.

7 REFERÊNCIAS BIBLIOGRÁFICAS

ALURA. **Data Lake vs Data Warehouse: Qual a diferença?** .Alura Artigos, 2023. Disponível em: https://www.alura.com.br/artigos/data-lake-vs-data-warehouse. Acesso em: 5 out. 2024.

ALURA. **Data Lake: conceitos, vantagens e desafios**. Alura Artigos, 2024. Disponível em: https://www.alura.com.br/artigos/data-lake-conceitos-vantagens-desafios. Acesso em: 5 out. 2024.

AMAZON WEB SERVICES. **What is AWS Data Pipeline?**. AWS, 2024. Disponível em: https://aws.amazon.com/pt/what-is/data-pipeline/. Acesso em: 3 set. 2024.

DATABRICKS. **What is Apache Spark?.** Databricks, 2024. Disponível em: https://www.databricks.com/br/glossary/what-is-apache-spark. Acesso em: 5 set. 2024.

DIXON, J. **Pentaho, Hadoop, and Data Lakes**. 2010. Disponível em: https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/. Acesso em: 3 set. 2024.

FERNANDES, Mariana Ramos. **Contributos de projetos de data engineering na indústria de serviços financeiros**. 2023. Tese de Doutorado. Instituto Superior de Economia e Gestão.

MICROSOFT. **Data Lake**. Learn Microsoft, 2024a. Disponível em: https://learn.microsoft.com/pt-br/azure/architecture/data-guide/scenarios/data-lake. Acesso em: 3 set. 2024.

MICROSOFT. **Introduction to Azure Data Factory**. Microsoft Learn, 2024d. Disponível em: https://learn.microsoft.com/pt-br/azure/data-factory/introduction. Acesso em: 3 set. 2024.

MICROSOFT**. Introduction to Azure Databricks**. Microsoft Learn, 2024c. Disponível em: https://learn.microsoft.com/pt-br/azure/databricks/introduction/. Acesso em: 5 out. 2024.

MICROSOFT. **Overview of Power BI**. Microsoft Learn, 2024f. Disponível em: https://learn.microsoft.com/pt-br/power-bi/fundamentals/power-bi-overview. Acesso em: 3 set. 2024.

MICROSOFT. **What is Azure Synapse Analytics?.** Microsoft Learn, 2024e. Disponível em: https://learn.microsoft.com/pt-br/azure/synapse-analytics/overview-what-is. Acesso em: 3 set. 2024.

MICROSOFT. **What is Azure?**. Learn Microsoft, 2024b. Disponível em: https://azure.microsoft.com/pt-br/resources/cloud-computingdictionary/what-is-azure/.. Acesso em: 5 set. 2024.

SHARMA, B. **Architecting Data Lakes Data Management Architectures for Advanced Business Use Cases**. O'Reilly Media, Inc., 2018. Disponível em: https://github.com/ffisk/books/blob/master/architecting-data-lakes.pdf. Acesso em: 20 ago. 2024.

SOMASUNDARAM, G.; SHRIVASTAVA, A. **Armazenamento e gerenciamento de informações: como armazenar, gerenciar e proteger informações digitais.** Porto Alegre: Bookman, 2011. 460-461 p.