Implantação de Data Lake e Visualização de Dados para auditoria pública na Controladoria-Geral do Estado de Mato Grosso

Eduardo William Alves Osti
Universidade Federal de Mato Grosso
Controladoria-Geral do Estado de Mato Grosso
Cuiabá-MT, Brasil
eduardo.osti@hotmail.com
ORCID: 0009-0001-4469-0387

Roberto Benedito de Oliveira Pereira
Universidade Federal de Mato Grosso
Cuiabá-MT, Brasil
roberto@ic.ufmt.br
ORCID: 0009-0000-8361-0071

Jonathas Eide Fujii
Universidade Federal de Mato Grosso
Controladoria-Geral do Estado de Mato Grosso
Cuiabá-MT, Brasil
jonathasfujii@gmail.com
ORCID: 0009-0006-6239-9530

Josiel Maimone de Figueiredo
Universidade Federal de Mato Grosso
Cuiabá-MT, Brasil
josiel@ic.ufmt.br
ORCID: 0000-0001-8569-7684

***Resumo*—Este trabalho teve como objetivo apresentar a implementação e avaliação de um *Data Lake* na Controladoria-Geral do Estado de Mato Grosso (CGE-MT), visando otimizar os processos de auditoria e análise de dados. Inicialmente, foi apresentado o conjunto de sistemas de código aberto que compõem o ambiente (como Apache HDFS, Spark e Trino) e, posteriormente, foram utilizados dois critérios para avaliação: técnico e operacional. Os resultados demonstraram que a infraestrutura implantada se mostrou eficiente para as atividades de análise de dados, proporcionando um ambiente seguro para o armazenamento e processamento das informações, garantindo a integridade dos dados. Além disso, com base no *Data Lake*, a CGE desenvolveu o sistema "CGE Alerta", que possibilitou a redução de 51% nas irregularidades de inassiduidade nas Secretarias do Estado de Mato Grosso e automatizou os processos de monitoramento. Também foi possível demonstrar a viabilidade da solução a longo prazo, uma vez que, considerando o espaço de armazenamento disponível, seria possível armazenar aproximadamente 15 anos de dados sem a necessidade de investimentos imediatos.**

Keywords*— *data lake, auditoria pública, análise de dados

I. INTRODUÇÃO

Com o avanço tecnológico proporcionado pelos meios computacionais, tem-se gerado uma quantidade massiva de dados, que vão desde registros de comportamentos de compras até transações em bolsa de valores. Esse crescimento exponencial do volume de dados transformou a forma como as organizações analisam e tomam suas decisões estratégicas. Conforme Rydning, Reinsel e Gantz [13], a projeção é que essa quantidade continue crescendo globalmente, o que torna essencial o uso de ferramentas que auxiliem na análise e auditoria das informações. Nesse contexto, o *Data Lake* surge como uma solução para armazenar e processar grandes volumes de dados de forma que auxilie nas análises das informações. Além disso, ferramentas de visualização de dados, como o *Tableau*, tornam-se fundamentais para extrair *insights* dessas análises e apoiar a auditoria.

Como consequência dessa evolução tecnológica e do aumento exponencial dos dados, a auditoria está passando por mudanças para se adaptar aos dias atuais, o que também se aplica à auditoria governamental. A auditoria é um processo cujo objetivo é verificar a veracidade dos dados informados. Para atingir esse propósito, são realizados procedimentos de comparação de dados a fim de identificar divergências e inconsistências [14].

Com essas novas tecnologias e processos, é possível processar, analisar e executar procedimentos em grandes volumes de dados de maneira eficiente e estratégica. Uma dessas abordagens é a análise de negócios (*Business Analytics*). Segundo Appelbaum et al. [17], essa prática consiste no uso de dados, tecnologia da informação e modelos estatísticos para ajudar os gestores a obterem uma melhor compreensão de suas operações e a tomarem decisões baseadas em fatos. No contexto específico da auditoria, essa abordagem tecnológica é frequentemente consolidada sob o termo *Data & Analytics* (*D&A*).

Santos [15] afirma que a utilização de *D&A* na auditoria proporciona resultados mais eficientes, tanto na obtenção dos dados quanto em sua análise. Isso permite processar grandes volumes de dados, em vez de utilizar amostragens, como é comum em análises manuais, identificando as exceções com precisão. Como resultado, é possível fornecer resultados assertivos aos clientes e ao governo em tempo hábil.

Portanto, com as vantagens apresentadas pela *D&A* e com o objetivo de otimizar as atividades de auditoria, tornando-as mais assertivas e eficientes nas análises, a Controladoria-Geral do Estado de Mato Grosso (CGE) viu a necessidade de preparar um ambiente, ou plataforma de dados, capaz de fornecer as ferramentas necessárias para o desenvolvimento de suas atividades internas. Esse ambiente foi projetado com a capacidade de integrar dados de diferentes fontes de origem, consolidando-os em um *Data Lake* e apresentando-os de forma dinâmica por meio de dashboards.

Há diversas ferramentas de visualização de dados disponíveis no mercado, como Apache Superset, Qlik Sense e Power BI. No entanto, a ferramenta escolhida para a construção dos dashboards neste estudo foi o *Tableau*, pois já havia sido previamente adotada pela CGE, facilitando sua integração com os processos internos e reduzindo a necessidade de adaptação dos usuários.

Dessa forma, este projeto propõe a investigação do impacto da implementação de um *Data Lake* na auditoria governamental da CGE, considerando sua integração com ferramentas como o *Apache HDFS*, *Trino* e o *Tableau*. A pergunta que este trabalho visa responder é: como a implementação de um *Data Lake*, aliada a essas tecnologias, pode otimizar os processos de auditoria na Controladoria-Geral do Estado de Mato Grosso?

Sendo assim, neste trabalho foram abordados e desenvolvidos os seguintes objetivos: 1 - A configuração de um ambiente de *Data Lake*; 2 - A coleta e o processamento dos dados de um dos órgãos governamentais; 3 - A construção de um dashboard com o *Tableau*, simulando, assim, as atividades dos auditores e 4 - A avaliação de critérios técnicos e operacionais do *Data Lake*.

II. FUNDAMENTAÇÃO TEÓRICA

O crescimento exponencial dos dados nos últimos anos tem impulsionado a necessidade de novas tecnologias capazes de processar e analisar grandes volumes de informações em tempo hábil [9]. Esse cenário é especialmente relevante no setor público, onde a auditoria governamental se beneficia do uso de técnicas de *D&A* para aumentar a precisão e a transparência na análise dos dados, superando as limitações das tradicionais auditorias amostrais [14], [15]. Segundo Boscov [10], o governo de Mato Grosso foi destaque no SECOP 2023 devido à excelência em governo digital e recebeu prêmios por seus projetos de transformação digital durante o evento nacional.

Nesse contexto, o conceito de maturidade digital surge como um indicador do nível de preparação e integração tecnológica das organizações. A maturidade digital de uma empresa ou órgão público refere-se ao grau de desenvolvimento das capacidades digitais, incluindo o uso de tecnologias, a integração de dados e uma cultura orientada por *insights* analíticos. Essa definição destaca a importância de adotar tecnologias digitais para otimizar processos e melhorar a eficiência [12]. Conforme ressaltado por Kafel et al. [12], o avanço na maturidade digital do setor público está diretamente ligado ao aumento da eficiência e à melhoria na prestação de serviços ao cidadão.

Esse nível de maturidade é fundamental para transformar o volume de dados disponíveis em informações valiosas, especialmente em auditorias governamentais, que exigem precisão e transparência. Ao ser capaz de processar grandes volumes de dados com mais eficiência, o setor público será capaz de identificar desvios e padrões que não seriam localizados facilmente em uma análise manual, possibilitando a evolução de técnicas baseadas em amostragem para análises mais completas do conjunto de transações [15].

Dessa forma, para trabalhar com esse tipo de dados, é necessário um conjunto de infraestrutura e sistemas especializados que facilite o armazenamento, processamento e a análise de dados. Além disso, a implementação dessas infraestruturas no setor público envolve a adoção de práticas de governança de dados e de segurança da informação. A governança garante a integridade e rastreabilidade necessárias para validar provas de auditoria [14]. Além disso, a literatura recente aponta para a evolução do conceito para arquiteturas de *Data Lakehouse*, que buscam unir a flexibilidade do *Data Lake* com as garantias de transações seguras e confiáveis típicas dos tradicionais *Data Warehouses* [5]. Existem diversas ferramentas *open-source* que são gratuitas, possuem suporte contínuo da comunidade e se destacam como líderes nesse segmento. A seguir são abordados os principais conceitos que servirão como apoio para o desenvolvimento do trabalho.

A. Data Lake

Segundo Da Silva *et al.* [6], o *Data Lake* é um repositório centralizado de dados projetado para armazenar dados heterogêneos, que podem ser estruturados, semiestruturados ou não estruturados. Devido a essa característica, as organizações são capazes de coletar e armazenar esses dados diretamente da fonte, sem a necessidade de realizar algum pré-processamento. Com isso, é possível manter os dados originais, garantindo a sua integridade.

Outro ponto vantajoso desse repositório é sua flexibilidade. Diferentemente do método tradicional, o *Data Lake* não exige a realização de uma modelagem prévia dos dados, como a modelagem relacional. Isso possibilita que os dados sejam organizados e processados conforme a necessidade da organização, proporcionando mais agilidade e menos restrições.

Os dados armazenados no *Data Lake* podem ser categorizados em três formas: dados estruturados, que possuem uma estrutura pré-definida, como tabelas de bancos de dados relacionais, arquivos Excel e formulários; dados semiestruturados, que combinam características de dados estruturados e não estruturados, não possuindo uma estrutura rígida, mas podendo ser organizados em agrupamentos, como arquivos JSON, XML e HTML; e dados não estruturados, que não seguem nenhum modelo ou formato pré-definido, sendo mais complexos de analisar e processar, incluindo arquivos de imagem, vídeo, sites, entre outros.

Há diversas formas de utilização e implantação de um *Data Lake*. Ele não fica restrito a uma ferramenta específica, podendo ser utilizado em um banco de dados relacional, Hadoop ou até mesmo em uma ferramenta de armazenamento em nuvem [5]. Essa flexibilidade no armazenamento é fundamental para os processos de auditoria da Controladoria-Geral do Estado de Mato Grosso, visto que é necessário coletar e analisar os dados de diversos órgãos em um único ambiente. A utilização desse ambiente possibilita aos auditores o acesso a uma ampla variedade de dados, garantindo uma auditoria mais precisa e detalhada.

B. *Camadas e Ferramentas na Arquitetura de Data Lake*

Para possibilitar a implantação de um *Data Lake* funcional, é necessário configurar diferentes camadas que executam tarefas específicas dentro de um ecossistema de dados, como podemos visualizar na Figura 1. Cada uma dessas camadas é composta por conjuntos de ferramentas que possibilitam o armazenamento, processamento, organização e visualização dos dados.

Fig. 1. Arquitetura do *Data Lake*

Fonte: Elaborado pelo autor, 2025

As principais funcionalidades e ferramentas das camadas da Arquitetura de um *Data Lake* (Figura 1) iniciam-se pela orquestração. Para que haja dados constantemente atualizados, deve-se contar com um fluxo contínuo de ingestão de dados. Uma das formas de alcançar esse resultado é a utilização de ferramentas de orquestração, como o *Apache Airflow*, que se tornam essenciais. O *Apache Airflow* é uma ferramenta de orquestração de pipelines que permite desenvolver, agendar e monitorar tarefas (denominadas DAGs) de coleta e processamento de dados. Uma de suas características é a capacidade de se conectar a diversas fontes de dados. Além disso, o desenvolvimento do fluxo de trabalho é bastante flexível [1].

Para lidar com o processamento de uma grande massa de dados proveniente da orquestração, é necessária uma engine capaz de realizar processamentos distribuídos e em memória. O *Apache Spark* se destaca nessa área. A vantagem de utilizá-lo é a sua capacidade de manter os dados em memória sempre que possível, ao contrário do *Apache Hadoop*, que armazena os resultados em disco ao final de cada etapa do processamento. Isso permite que, ao iniciar a próxima etapa, o tempo de recuperação das informações seja reduzido, tornando o desempenho mais eficiente [4].

Na base desse ecossistema, o armazenamento distribuído suporta as principais ferramentas de *Data Lake*, pois é possível garantir resiliência, escalabilidade e alta disponibilidade. Existem algumas ferramentas no mercado que possuem essas características, como o GlusterFS, *Apache HDFS* e o Ceph. Uma das que mais se destaca, e é utilizada com esse objetivo, é o *Apache HDFS*. Ele consegue atender a esses requisitos (resiliência, escalabilidade e alta disponibilidade) devido à forma como armazena os dados. Cada arquivo não é salvo de forma inteira; ele é fragmentado em blocos e distribuído entre os nós do cluster. Dessa forma, é possível ganhar velocidade na recuperação dos arquivos e proporcionar maior durabilidade, pois, mesmo em caso de falha em algum dos nós, os dados estarão disponíveis [2].

A catalogação dos metadados também é fundamental em um *Data Lake*, visto que os dados podem facilmente se tornar desorganizados e inconsistentes. Por isso, o gerenciamento dos dados como tabelas, além de possibilitar maior organização, auxilia na integridade, versionamento e disponibilidade das informações. O *Iceberg* é uma solução de gerenciamento de dados em larga escala, projetada para lidar com grandes volumes de dados organizados em formato de tabela. O *Iceberg* também oferece suporte a operações como versionamento, gerenciamento de esquemas e controle de transações ACID (Atomicidade, Consistência, Isolamento e Durabilidade). Além disso, ele possibilita o uso de SQL para manipulação dos dados e integração com várias ferramentas de processamento, como *Apache Spark*, *Trino*, Apache Flink, entre outras [3]. Para complementar, o *Project Nessie* faz o gerenciamento dos metadados, onde as soluções de dados armazenam informações sobre os formatos das tabelas. Essa estrutura permite que sistemas como *Apache Spark* e *Trino* acessem as tabelas, garantindo integridade e consistência dos dados ao longo de diversas operações. Além disso, o *Project Nessie* também tem a capacidade de gerenciar e rastrear as versões dos dados, semelhantemente ao Git [8].

O acesso aos dados armazenados no *Data Lake* é facilitado pelas engines SQL, sendo o *Trino* uma das ferramentas utilizadas para essa função. O *Trino* é um motor de consulta distribuído, projetado para atuar como uma interface entre os *Data Lakes* e os usuários, utilizando a linguagem SQL. Além disso, ele possui a capacidade de realizar consultas em grandes volumes de dados de forma eficiente e se integra com diversos catálogos de metadados, como o *Project Nessie*, por exemplo. Isso possibilita que o *Trino* acesse e integre diferentes *Data Lakes* e sistemas de banco de dados, proporcionando um ambiente ideal para o processamento de grandes conjuntos de dados em ambientes distribuídos [18].

Por fim, o processo de auditoria exige a análise e exploração contínua dos dados coletados, e o compartilhamento dos resultados das análises é essencial para um trabalho colaborativo e interdisciplinar. O *JupyterLab* oferece uma interface interativa para visualização e manipulação de dados, gráficos, scripts e documentos [11], proporcionando um ambiente adequado para fluxos de trabalho em ciência de dados e auditoria. Complementando essa abordagem, o *Tableau Server* é utilizado para a criação e disponibilização de relatórios interativos, permitindo que os auditores consolidem e compartilhem informações por meio de dashboards dinâmicos, facilitando a análise dos dados armazenados no *Data Lake*.

Embora a descrição dessas ferramentas delineie a arquitetura, é importante destacar a justificativa de suas escolhas frente a alternativas de mercado. Optou-se pelo *Apache HDFS* em ambiente *on-premise* em detrimento de soluções em nuvem pública para preservar a soberania de dados exigida pela CGE. Quanto à formatação e catalogação, a escolha por padrões como *Iceberg* e *Project Nessie* se deve à sua capacidade de operar com diferentes motores de processamento. Ao contrário do *Delta Lake*, mais associado ao ecossistema *Apache Spark*, o *Iceberg* foi desenvolvido com foco em interoperabilidade, permitindo integração com diversos motores de consulta SQL [3].

Por fim, o uso do *Trino* foi justificado perante tecnologias em lote (como o *Apache Hive*) por entregar uma performance superior de baixa latência em consultas interativas *ad-hoc* via *Tableau*. Conforme demonstrado por Sethi et al. [19], a arquitetura de processamento em memória e pipelining do *Trino* (anteriormente Presto) apresenta tempos de resposta significativamente inferiores em cenários analíticos quando comparada aos modelos baseados em *MapReduce* adotados por ferramentas tradicionais. Ao implementar o ambiente com as ferramentas mencionadas, o objetivo é disponibilizar uma arquitetura integrada, composta por soluções de código aberto, servindo como uma fonte centralizada e estruturada para a execução eficiente dos trabalhos de auditoria.

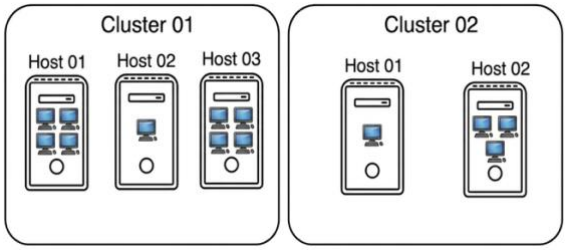
III. MATERIAIS E MÉTODOS

O presente capítulo descreve os materiais e métodos utilizados para o desenvolvimento deste trabalho, abordando os recursos computacionais seguidos da metodologia empregada na implantação e avaliação do Data Lake.

A. Recursos Computacionais

Para a sustentação do ambiente, foram utilizadas 13 máquinas virtuais (VMs) hospedadas em dois clusters VMware, na modalidade *on-premise*, conforme apresentado na Figura 2. O Cluster 01 é composto por três servidores físicos rodando o VMware ESXi, versão 6.7.0, com cada nó possuindo 256 GB de Memória RAM e 6 núcleos físicos operando a 3,7 GHz. Já o Cluster 02 conta com dois servidores físicos rodando o VMware ESXi, versão 6.0.0, sendo que cada nó possui 256 GB de memória RAM, 8 núcleos físicos e 16 *threads*. Além disso, cada nó conta com dois sockets, suportando dois processadores, totalizando 16 núcleos físicos e 32 threads, operando a 2,13 GHz.

Fig. 2. Desenho da infraestrutura



Fonte: Elaborado pelo autor, 2025

Devido às limitações de recursos nos clusters VMware, a distribuição das VMs foi realizada levando em conta tanto as funções dos serviços quanto a disponibilidade de recursos computacionais em cada cluster, garantindo que os sistemas tivessem a capacidade de executar suas funcionalidades.

Para estruturar o armazenamento do *Data Lake*, as primeiras quatro VMs foram configuradas no cluster 02 para a execução do sistema Apache HDFS. Cada VM foi provisionada com 12 GB de Memória RAM, 5 vCPUs e 2.5 TB de armazenamento, rodando a distribuição Oracle Linux 7.9 e a versão 3.3.4 do Apache HDFS.

Para o processamento distribuído de dados, foi configurado um cluster Apache Spark no Cluster 01, composto por cinco nós, sendo um coordenador e quatro *workers*. Cada *worker* possui 4 vCPUs, 16 GB de memória RAM e 150 GB de armazenamento, enquanto o nó coordenador foi configurado com 4 vCPUs, 8 GB de memória RAM e 100 GB de armazenamento. Todas as VMs utilizam a distribuição Oracle Linux 8.9, com a versão 3.5.1 do Apache Spark.

Para a orquestração de pipeline de dados, foi configurado o Apache Airflow no cluster 01, utilizando três VMs, cada uma com 2 vCPUs, 8 GB de Memória RAM e 100 GB de armazenamento, rodando com a distribuição Oracle Linux 8.9 e a versão 2.9.1 do Apache Airflow, permitindo o desenvolvimento, agendamento e monitoramento de pipelines de dados (DAGs).

Com o propósito de facilitar a exploração dos dados, foi configurado o JupyterLab em uma única VM no Cluster 01, com 4 vCPUs, 16 GB de Memória RAM e 100 GB de armazenamento, utilizando Oracle Linux 8.9 e a versão 4.2.5 do JupyterLab.

Devido à limitação de recursos computacionais nos clusters VMware, o Trino foi configurado nas mesmas VMs do Apache HDFS, alocadas no Cluster 02. Dessa forma, o cluster do Trino foi estruturado com um nó coordenador e três nós *workers*, permitindo a execução de consultas SQL distribuídas diretamente no *Data Lake*.

Para o gerenciamento de metadados, foi configurado o Project Nessie. Essa configuração foi realizada em formato de contêiner devido à falta de recursos nos clusters VMware. Como o *Project Nessie* atua exclusivamente no gerenciamento e versionamento de metadados, sem processar ou armazenar grandes volumes de dados diretamente, essa configuração foi viável. Seu objetivo é fornecer metadados consultáveis para o Apache Spark e Trino, o que não demanda muitos recursos computacionais.

Por fim, o Tableau Server foi utilizado para a criação e compartilhamento de relatórios. Este sistema já era utilizado antes da construção do ambiente de Data Lake, e foi reaproveitado e integrado à nova arquitetura. O Tableau Server foi configurado em uma VM no Cluster 02 com 24 vCPUs, 64 GB de memória RAM, 2 TB de armazenamento, utilizando o sistema operacional Microsoft Windows Server 2016.

B. Metodologia

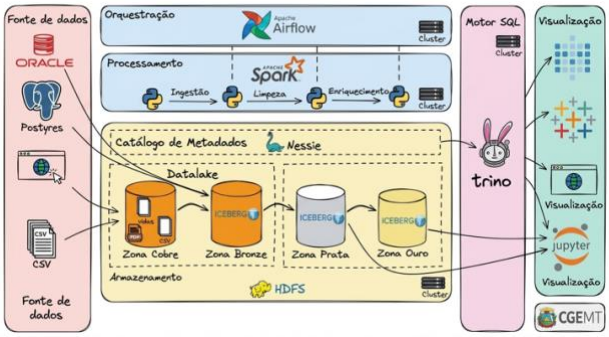
A metodologia utilizada para avaliação do ambiente foi dividida em duas partes: a primeira corresponde à avaliação de critérios técnicos, e a segunda, à avaliação de critérios operacionais. Na avaliação dos critérios técnicos, foram analisados a qualidade e integridade dos dados, a performance e o tempo de resposta, além da integração com o Tableau. Já na avaliação dos critérios operacionais, foram considerados a eficiência no suporte à auditoria e o custo-benefício.

A implantação teve como objetivo configurar, integrar e otimizar todos os sistemas que fazem parte do ecossistema do *Data Lake*. Com a implantação realizada, o processo de avaliação dos critérios técnicos ocorreu da seguinte forma: a qualidade e integridade dos dados foram avaliadas estruturalmente para garantir que as informações

foram armazenadas de maneira fidedigna e estavam aptas a serem consumidas sem anomalias estruturais. A performance do ambiente foi analisada por meio da cronometragem sistêmica do tempo de carregamento no *Tableau*, o que também permitiu atestar a eficácia da sua integração com o motor SQL. O critério metodológico adotado definiu que a performance seria considerada satisfatória se a arquitetura fosse capaz de processar os dados e exibi-los em um tempo de latência aceitável para a exploração interativa.

A arquitetura apresentada na Figura 3 ilustra as principais etapas de um pipeline de dados: ingestão, processamento e disponibilização das informações para consumo pelos usuários finais. A ingestão dos dados é realizada por meio de scripts desenvolvidos em Python com *Apache Spark* e orquestrados pelo *Apache Airflow*, por meio de DAGs (Directed Acyclic Graphs), que programam a execução periódica dessas tarefas. As fontes de dados incluem arquivos CSV, bancos de dados Oracle e PostgreSQL, além de páginas web, entre outras.

Fig. 3. Arquitetura e fluxo de processamento do Data Lake na CGE-MT



Fonte: Elaborado pelo autor, 2025

Na etapa de processamento, os dados passam por tarefas de limpeza, como a remoção de registros nulos, padronização de formatos de datas e renomeação de colunas, e de enriquecimento, com regras de negócio definidas com base nas demandas de auditoria, simuladas neste projeto para fins de validação do ambiente. Após o tratamento, os dados são organizados em camadas de armazenamento, denominadas zonas Cobre, Bronze, Prata e Ouro, e catalogados no metastore gerenciado pelo *Project Nessie*.

Além disso, o monitoramento de todo o fluxo dos dados é realizado pelo próprio *Apache Airflow*, dessa forma é possível acompanhar a execução das DAGs e identificar falhas em tempo real. Isso facilita o processo de detecção e correção de problemas, garantindo a continuidade do processamento dos dados.

Por fim, os dados tratados e organizados podem ser consultados via *Trino* e consumidos por ferramentas de visualização, como o *Tableau* e o *JupyterLab*.

IV. Resultados e Discussões

Para a avaliação empírica do ambiente, foram estabelecidas duas vertentes principais: a avaliação de critérios técnicos e a avaliação de critérios operacionais. A dimensão técnica abrange aspectos como a qualidade e integridade dos dados, a performance de processamento e a integração com o *Tableau*. Por sua vez, a dimensão operacional analisa a eficiência no suporte prático à auditoria governamental e o custo-benefício da solução implementada.

A primeira dimensão técnica avaliada foi a qualidade e integridade dos dados. Para atingir esse objetivo, procedeu-se à ingestão e seleção dos dados dentro do *Apache HDFS*. Os registros foram alocados na zona Bronze utilizando scripts PySpark. Nesta etapa, selecionaram-se todas as tabelas inerentes ao desenvolvimento do dashboard analítico. Consequentemente, foram ingeridas diversas tabelas provenientes do sistema de controle de acesso de Catracas da CGE, gerando múltiplos arquivos dentro do *HDFS*, de modo que cada arquivo corresponde a uma tabela distinta, conforme ilustrado na Figura 4.

Fig. 4. Lista de diretórios no HDFS

Fonte: Elaborado pelo autor, 2025

Dentro de cada diretório, os dados encontram-se armazenados no formato Parquet, acompanhados de metadados em JSON, gerados e catalogados pelo *Apache Spark* em conjunto com o *Apache Iceberg*. Estes arquivos ficam disponíveis para leitura por meio do motor *Trino* ou de processamentos paralelos em PySpark. Ao inspecionar o arquivo Parquet 00000-0-56b0c52e-be30-47e2-bb8c-16f52c62975f-00001.parquet, correspondente à tabela ev_evento_catraca, foi possível constatar a integridade estrutural e tipológica dos dados, conforme apresentado na Figura 5.

Fig. 5. Recorte da leitura do arquivo Parquet

id	nome	evento	hora
1486392	EDUARDO WILLIAM ALVES OSTI	ENTRADA	2022-07-01 09:51:19
1486461	EDUARDO WILLIAM ALVES OSTI	SAIDA	2022-07-01 10:26:03
1487785	EDUARDO WILLIAM ALVES OSTI	ENTRADA	2022-07-04 07:31:22
1487793	EDUARDO WILLIAM ALVES OSTI	SAIDA	2022-07-04 07:36:42
1487867	EDUARDO WILLIAM ALVES OSTI	ENTRADA	2022-07-04 07:58:52
1488656	EDUARDO WILLIAM ALVES OSTI	SAIDA	2022-07-04 12:23:27
1488707	EDUARDO WILLIAM ALVES OSTI	ENTRADA	2022-07-04 12:40:14
1489418	EDUARDO WILLIAM ALVES OSTI	SAIDA	2022-07-04 17:24:47
1489778	EDUARDO WILLIAM ALVES OSTI	ENTRADA	2022-07-05 07:53:52
1490581	EDUARDO WILLIAM ALVES OSTI	SAIDA	2022-07-05 12:00:25
1490683	EDUARDO WILLIAM ALVES OSTI	ENTRADA	2022-07-05 12:33:34
1494218	EDUARDO WILLIAM ALVES OSTI	SAIDA	2022-07-07 12:07:49
1494294	EDUARDO WILLIAM ALVES OSTI	ENTRADA	2022-07-07 12:36:51
1495352	EDUARDO WILLIAM ALVES OSTI	SAIDA	2022-07-07 17:19:02
1495482	EDUARDO WILLIAM ALVES OSTI	ENTRADA	2022-07-08 07:55:29
1496860	EDUARDO WILLIAM ALVES OSTI	SAIDA	2022-07-08 17:05:58

Fonte: Elaborado pelo autor, 2025

Adicionalmente, a Figura 6 demonstra que as mesmas informações podem ser consultadas nativamente por meio do catálogo *Project Nessie* utilizando um cliente SQL padronizado. A validação cruzada entre estes dois testes de leitura, mediante a comparação dos conjuntos de dados retornados, permitiu concluir que os dados estão plenamente disponíveis e íntegros, sem indícios de anomalias de codificação, perda de pacotes ou corrupção.

Fig. 6. Leitura da tabela ev_evento_catraca diretamente do *Data Lake*

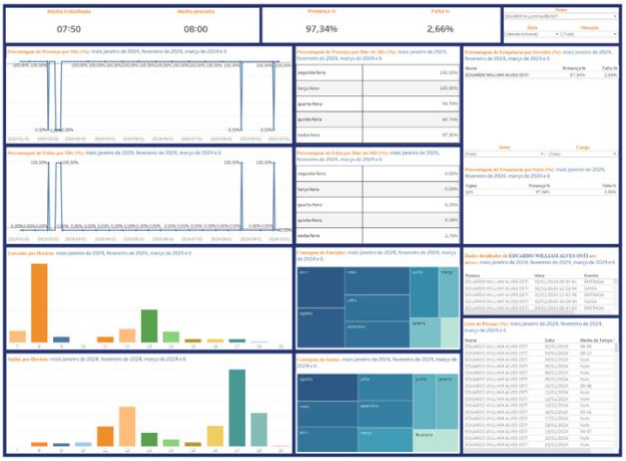
id	nome	evento	hora
1.486.392	EDUARDO WILLIAM ALVES OSTI	ENTRADA	2022-07-01 09:51:19
1.486.461	EDUARDO WILLIAM ALVES OSTI	SAIDA	2022-07-01 10:26:03
1.487.785	EDUARDO WILLIAM ALVES OSTI	ENTRADA	2022-07-04 07:31:22
1.487.793	EDUARDO WILLIAM ALVES OSTI	SAIDA	2022-07-04 07:36:42
1.487.867	EDUARDO WILLIAM ALVES OSTI	ENTRADA	2022-07-04 07:58:52
1.488.656	EDUARDO WILLIAM ALVES OSTI	SAIDA	2022-07-04 12:23:27
1.488.707	EDUARDO WILLIAM ALVES OSTI	ENTRADA	2022-07-04 12:40:14
1.489.418	EDUARDO WILLIAM ALVES OSTI	SAIDA	2022-07-04 17:24:47
1.489.778	EDUARDO WILLIAM ALVES OSTI	ENTRADA	2022-07-05 07:53:52
1.490.581	EDUARDO WILLIAM ALVES OSTI	SAIDA	2022-07-05 12:00:25
1.490.683	EDUARDO WILLIAM ALVES OSTI	ENTRADA	2022-07-05 12:33:34
1.494.218	EDUARDO WILLIAM ALVES OSTI	SAIDA	2022-07-07 12:07:49
1.494.294	EDUARDO WILLIAM ALVES OSTI	ENTRADA	2022-07-07 12:36:51
1.495.352	EDUARDO WILLIAM ALVES OSTI	SAIDA	2022-07-07 17:19:02
1.495.482	EDUARDO WILLIAM ALVES OSTI	ENTRADA	2022-07-08 07:55:29
1.496.860	EDUARDO WILLIAM ALVES OSTI	SAIDA	2022-07-08 17:05:58
1.496.951	EDUARDO WILLIAM ALVES OSTI	ENTRADA	2022-07-11 07:44:28
1.498.572	EDUARDO WILLIAM ALVES OSTI	SAIDA	2022-07-11 17:10:24
1.498.857	EDUARDO WILLIAM ALVES OSTI	ENTRADA	2022-07-12 07:43:46
1.500.445	EDUARDO WILLIAM ALVES OSTI	SAIDA	2022-07-12 17:08:13

Fonte: Elaborado pelo autor, 2025

Com a integridade dos dados validada, a etapa subsequente consistiu na análise de performance, mensurada pelo tempo de resposta no carregamento dos dados no dashboard. Esta fase também serviu para atestar a interoperabilidade do *Tableau* com o ecossistema do *Data Lake*. Tal integração viabiliza-se pelo fato de o *Tableau* conectar-se diretamente ao *Trino* para recuperar os dados distribuídos no *HDFS*, permitindo quantificar não apenas a latência da infraestrutura, mas o desempenho da ferramenta de visualização na ponta.

Para a avaliação empírica, aferiu-se o tempo de carregamento completo do dashboard, o qual registrou um tempo de resposta de 1 minuto e 20 segundos para realizar uma consulta abrangente em tempo real (*live connection*) no *Trino*, carregando o histórico de todos os servidores da CGE no período de janeiro a setembro de 2024. Contudo, ao aplicar filtros paramétricos por nome (buscando um servidor específico), o tempo de resposta reduziu-se para 10 segundos. A Figura 7 ilustra o dashboard de monitoramento de jornada de trabalho implementado no *Tableau* para esta atividade.

Fig. 7. Dashboard de monitoramento de presença e jornada de trabalho



Fonte: Elaborado pelo autor, 2025

A arquitetura também permite a utilização de extrações nativas do *Tableau*, onde os dados da fonte (*Trino*) são alocados no repositório em memória da própria ferramenta de visualização, otimizando a performance analítica. Ao ativar a extração e recarregar o dashboard com a totalidade dos dados dos servidores da CGE, o tempo de carregamento foi reduzido para 1 minuto e 6 segundos. Aplicando-se o filtro nominal, o tempo de carregamento foi de 7 segundos. Os testes de performance evidenciam que a arquitetura atende aos requisitos de escalabilidade e velocidade necessários para a CGE, fornecendo respostas quase instantâneas em cenários de consultas parametrizadas (*ad-hoc*).

Para além dos critérios estritamente técnicos, a avaliação operacional consolidou o impacto institucional da solução. Antes da implementação do *Data Lake*, a auditoria enfrentava barreiras estruturais ligadas à fragmentação informacional, visto que cada Secretaria detinha sistemas isolados. O auditor necessitava acessar múltiplos bancos de dados, compreender modelagens distintas e executar consultas diretamente em ambientes transacionais (de produção). Tal fluxo, além de oneroso e predominantemente manual, impunha riscos de degradação de performance aos sistemas operacionais do Estado.

Para mitigar esse cenário, a Unidade de Inteligência da CGE desenvolveu o sistema "CGE Alerta" apoiado na nova infraestrutura, visando o monitoramento contínuo das Secretarias e a identificação tempestiva de não conformidades [16]. O sistema consome os dados centralizados no *Data Lake*, viabilizando a automação de análises preditivas e preventivas.

Os resultados dessa integração foram expressivos. A partir do monitoramento automatizado, a CGE conseguiu reduzir em 51% as irregularidades relacionadas à categoria de inassiduidade habitual [7]. Para fins de rigor metodológico, cabe ressaltar que este percentual foi calculado comparando o volume absoluto de notificações geradas no período pós-implantação do "CGE Alerta" com o período equivalente anterior, em que o controle era puramente amostral e manual. Em categorias de auditoria mais críticas, observou-se a eliminação de 100% das inconformidades registradas, evidenciando que a infraestrutura analítica induz a correção do problema antes da necessidade de ações sancionatórias tardias [16]. O tempo de auditoria foi significativamente otimizado com a eliminação de cruzamentos manuais de planilhas.

Por fim, a avaliação do custo-benefício comprovou a sustentabilidade do projeto. Para o armazenamento do histórico de 9 meses (janeiro a setembro de 2024), compreendendo 220.712 registros de catracas, o espaço físico consumido foi de meros 462 MB no HDFS. Extrapolando este consumo linearmente para a capacidade bruta de 9 TB disponíveis no cluster, estima-se que a infraestrutura atual possua capacidade analítica para comportar aproximadamente 184 meses de dados (mais de 15 anos de histórico contínuo). Esta projeção técnica demonstra a viabilidade e longevidade do modelo *on-premise* adotado, garantindo a governança das informações sem a necessidade de dispêndios imediatos para ampliação de armazenamento.

Conclusões e Trabalhos Futuros

Os resultados deste estudo evidenciam empiricamente que a implantação da arquitetura de *Data Lake*, em sinergia com ferramentas analíticas modernas, proporcionou benefícios institucionais expressivos à Controladoria-Geral

do Estado (CGE). O ambiente estruturado viabilizou o armazenamento centralizado e a organização dos dados com governança, provenientes das diversas Secretarias do Estado de Mato Grosso, mitigando as barreiras operacionais anteriormente associadas à fragmentação informacional.

No escopo dos critérios técnicos, a infraestrutura desenhada demonstrou maturidade para atender aos requisitos de escalabilidade e integração. A disponibilização das informações em baixa latência (*ad-hoc*) foi validada, evidenciando a eficácia do uso das extrações do Tableau. A adoção do Trino como motor distribuído de acesso ao HDFS conferiu aos auditores uma interface de exploração amigável e de alta performance, consolidando o pipeline desde a ingestão bruta até a visualização analítica.

Sob a ótica dos critérios operacionais, o ambiente gerou impacto positivo nas rotinas de auditoria. A redução expressiva de 51% nos índices de inassiduidade habitual e a mitigação completa (100%) das irregularidades em outras categorias atestam a efetividade do modelo. Esses resultados decorrem da transição de um modelo de auditoria reativo e amostral para um monitoramento contínuo e preventivo, viabilizado pela alimentação ininterrupta do sistema "CGE Alerta" a partir do repositório unificado. A automação analítica eliminou a necessidade de cruzamentos manuais de planilhas, conferindo maior agilidade aos processos e permitindo que o auditor concentre seus esforços na análise crítica. Como consequência, essa visão também passa a orientar decisões dos gestores públicos, favorecendo estratégias mais preventivas e a alocação de recursos em áreas de maior risco.

Ademais, a análise de custo-benefício comprova a viabilidade do ecossistema a longo prazo. A sustentabilidade do ambiente baseia-se na capacidade de armazenamento excedente (estimada para suportar mais de 15 anos de carga transacional sem necessidade de novos investimentos) e na adoção de tecnologias *open-source*. O suporte das comunidades de software livre contribui para a correção de falhas, bem como para a atualização contínua do ambiente tecnológico. Nesse sentido, a arquitetura técnica elaborada neste estudo vai além do contexto local e pode ser entendida como um modelo passível de replicação, aplicável a outras controladorias e instituições públicas que busquem modernizar seus processos de auditoria com baixo custo e autonomia informacional.

Por fim, embora a análise aponte para a longevidade do modelo *on-premise* em termos de capacidade de armazenamento, observam-se limitações relacionadas à elasticidade dos clusters físicos diante de variações na demanda computacional e do crescimento contínuo do volume de dados processados. A partir dessa distinção, emergem problemas de pesquisa para investigações futuras: (1) a migração do *Data Lake* para uma arquitetura nativa em nuvem resulta em melhor relação custo-desempenho para a auditoria governamental em comparação ao modelo local? e (2) de que forma os custos de egressão de dados impactam a viabilidade econômica dessa migração? Para responder a essas questões, propõe-se a realização de experimentos de escalabilidade e desempenho, permitindo uma análise comparativa entre infraestruturas locais e ambientes de nuvem.

REFERENCES

[1] Apache Airflow, "What is Airflow?," Apache Airflow Documentation, 2024. [Online]. Available: https://airflow.apache.org/docs/apache-airflow/stable/. Accessed: Nov. 6, 2024.

[2] K. Shvachko, H. Kuang, S. Radia and R. Chansler, "The Hadoop Distributed File System," 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), Incline Village, NV, USA, 2010, pp. 1-10.

[3] J. Schneider, C. Gröger, A. Lutsch, et al., "The Lakehouse: State of the Art on Concepts and Technologies," SN Computer Science, vol. 5, p. 449, 2024.

[4] M. Zaharia et al., "Resilient distributed datasets: A Fault-Tolerant abstraction for In-Memory cluster computing," in 9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12), San Jose, CA, USA, 2012, pp. 15-28.

[5] C. Avci, B. Tekinerdogan, and I. N. Athanasiadis, "Software architectures for big data: a systematic literature review," Big Data Analytics, vol. 5, no. 1, p. 5, 2020.

[6] A. R. E. Da Silva et al., "Análise da relevância da arquitetura de implementação de Delta Lake para banco de dados empresariais," 2024.

[7] D. Borges, "CGE Alerta transforma gestão pública em 2024 e reduz pendências em até 51%," 2025. [Online]. Available: https://www.cge.mt.gov.br/w/cge-alerta-transforma-gest%C3%A3o-p%C3%BAblica-em-2024-e-reduz-pend%C3%AAncias-em-at%C3%A9-51-/. Accessed: Mar. 3, 2025.

[8] Dremio, "Project Nessie," 2024. [Online]. Available: https://www.dremio.com/open-source/nessie/. Accessed: Nov. 6, 2024.

[9] S. Fanelli et al., "Big data analysis for decision-making processes: challenges and opportunities for the management of health-care organizations," Management Research Review, vol. 46, no. 3, pp. 369–389, 2023.

[10] G. Boscov, "Mato Grosso é destaque no SECOP 2023: Excelência em Governo Digital," Secretaria de Estado de Meio Ambiente, Desenvolvimento Sustentável e Turismo (MTI), Cuiabá, MT, 4 set. 2023. Online. Available: https://www.mti.mt.gov.br/-/mato-grosso-é-destaque-no-secop-2023-excelência-em-governo-digital. Accessed: Apr. 14, 2026.

[11] T. Kluyver et al., "Jupyter Notebooks-a publishing format for reproducible computational workflows," in Positioning and Power in Academic Publishing: Players, Agents and Agendas, 20th International Conference on Electronic Publishing, IOS Press, 2016, pp. 87-90.

[12] T. Kafel, A. Wodecka-Hyjek, and R. Kusa, "Multidimensional public sector organizations' digital maturity model," Administration & Public Management Review, vol. 37, pp. 64-82, 2021.

[13] D. D. H. Ameen, S. W. Kareem, and S. B. Hasan, "A Big Data, Bigger Impact: A Comprehensive Review of Machine Learning Advancements," in 2024 International Conference on Electrical Engineering and Computer Science (ICECOS), IEEE, 2024, pp. 1-6.

[14] O. M. Ribeiro and J. M. R. Coelho, Auditoria fácil, 2. ed. São Paulo: Saraiva, 2013.

[15] M. Santos, "O impacto das novas tecnologias na profissão do auditor," KPMG Business Magazine, vol. 46, pp. 16-21, 2019.

[16] L. Silveira, "CGE lança sistema que permite monitoramento e correção proativa de questões administrativas," 2024. [Online]. Available: https://www.mti.mt.gov.br/-/cge-lan%C3%A7a-sistema-que-permite-monitoramento-e-corre%C3%A7%C3%A3o-proativa-de-quest%C3%B5es-administrativas/. Accessed: Mar. 3, 2025.

[17] D. Appelbaum et al., "Impact of business analytics and enterprise systems on managerial accounting," International Journal of Accounting Information Systems, vol. 25, pp. 29-44, 2017.

[18] Trino, "Trino 464 Documentation." [Online]. Available: https://trino.io/docs/current/overview/use-cases.html/. Accessed: Nov. 6, 2024.

[19] R. Sethi et al., "Presto: SQL on everything," in 2019 IEEE 35th International Conference on Data Engineering (ICDE), IEEE, 2019, pp. 1802-1813.