Data Lakes Lógicos Como Plataformas Para Dados Governamentais em Sociedades e Cidades Inteligentes

Geymerson S. Ramos, Danilo Fernandes[1], Jorge Artur P. de M. Coelho[2] Andre L. L. Aquino[1]

[1] Instituto de Computação – Universidade Federal de Alagoas (UFAL)

[2]Faculdade de Medicina – Universidade Federal de Alagoas (UFAL)
Maceió – AL – Brasil

{geymerson, dfc, jorge.coelho, alla}@laccan.ufal.br

Abstract. *Data lakes have received attention from corporate, academic, and government entities. This new approach to storing data has shown versatility for developing secure platforms, guaranteeing data privacy, quality, and governance. To take advantage of these characteristics and contribute to the value generation for government policies, we present a data lake architecture used to integrate government applications and data from Alagoas state, Brazil. As a preliminary result, we demonstrate a search result for the geographic distribution of the users in the systems integrated based on the proposed architecture.*

Resumo. *Data lakes têm recebido atenção de ambientes corporativos, acadêmicos e governamentais. A capacidade dessa nova abordagem de armazenamento de dados tem demonstrado versatilidade no desenvolvimento de plataformas seguras, garantidoras de privacidade, qualidade e governança. Para tirar proveito das referidas características e contribuir com a geração de valor à iniciativas governamentais, apresentamos uma arquitetura de data lake aplicada ao contexto de integração de sistemas e dados governamentais do estado de Alagoas. Como resultado preliminar, demonstramos uma consulta da distribuição geográfica dos usuários de cada um dos sistemas integrados em nossa aplicação com base na arquitetura.*

1. Introdução

Nas últimas décadas, o volume de dados gerados tem crescido consideravelmente. É previsto um número aproximado de 75 bilhões de dispositivos conectados mundialmente até 2025 [IRENA Group 2019], trazendo uma série de desafios para o gerenciamento e análise de dados. Estes desafios são representados em grande parte pelo volume, variedade e velocidade com que os dados trafegam na rede, caracterizando os três V's do conceito de *Big Data* [Marx 2013]. Neste contexto, *Data Lakes* [Zagan and Danubianu 2020] têm recebido considerável atenção do mercado e de entidades governamentais. Essa tecnologia tem como principal característica a sua capacidade de armazenar dados estruturados, semi-estruturados e não-estruturados em seu formato original, possibilitando a análise dos mesmos de forma integrada [Sawadogo and Darmont 2021]. A atenção recebida no meio governamental ocorre porque cresce o número de iniciativas sociais governamentais e é premente a necessidade de identificar a real demanda e o impacto das

mesmas. Isto é, compreender a magnitude do problema, quão são substanciais os benefícios da política governamental e se seus custos são mais administráveis que os custos de alternativas já existentes ou concorrentes.

Iniciativas de governança precisam ser concebidas como um sistema integrado que demanda avaliações antes, durante e depois de sua implementação. Trata-se de atividade essencial para a orientação e avaliação das políticas e suas formas de gestão. Nesse sentido, a avaliação de ações de gestão é o principal mecanismo de provimento de informações, de retroalimentação e de aperfeiçoamento contínuo, e envolve necessariamente algum sistema de coleta de dados que são a base para a tomada de decisão. Portanto, os esforços para modificar e melhorar o ambiente de decisão (por exemplo, *Business Intelligence*) são cada vez mais demandados. Trata-se de uma metodologia na qual se estabelecem ferramentas para obter, armazenar, organizar, analisar e prover acesso às informações necessárias aos tomadores de decisão [Božič and Dimovski 2019]. Em um ambiente governamental, a implantação de *Data Lakes* possibilita integrar os dados de diversas plataformas digitais sob a gestão do Estado, como sistemas tributários, de fiscalização, de assistência social e de infraestrutura. A integração traz consigo uma maior coesão das ações do governo e dos resultados das mesmas, permitindo deste modo realizar análises de dados, avaliar sua eficácia e eficiência e, por fim, auxiliar na tomada de novas decisões.

Há diversos desafios para a implantação de um *Data Lake*, a exemplo da contratação de infraestrutura de armazenamento e equipe de profissionais capacitados em diferentes áreas para ingestão, integração, processamento, análise de dados e tomada de decisão [Fang 2015]. Para sistemas existentes, é preciso migrar os dados para soluções escaláveis de armazenamento, o que gera custo adicional, ou utilizar uma abordagem de integração *ad hoc* para bases existentes, criando um *Data Lake* lógico [Gorelik 2019], esta será nossa abordagem. Neste trabalho, nossa proposta e principal contribuição é uma arquitetura de *Data Lake* lógico cuja finalidade é integrar bases de dados de alguns sistemas do governo estadual de Alagoas por meio de uma abordagem federada [Liaqat et al. 2017].

Dada a necessidade de organização e resgate dos dados para análises e entendimento dos resultados de iniciativas específicas, a arquitetura está sendo idealizada e desenvolvida para aplicação real em dados governamentais, e a abordagem federada tem sua justificativa em um requisito não-funcional de que os dados não poderiam ser copiados periodicamente de um sistema para outro. Os sistemas a serem integrados consistem em: 1. Uma plataforma dedicada à capacitação de profissionais de TI e divulgação de oportunidades ofertadas por empresas ou instituições de tecnologia; 2. Um sistema para auxiliar o incentivo da prática de esportes por meio da intermediação entre atletas e instituições esportivas; 3. Um sistema de cadastro para uma iniciativa governamental de assistência social à famílias em situação de pobreza e extrema pobreza.

Ademais, apresentamos como resultado preliminar uma análise da disposição geográfica dos usuários beneficiados pelos referidos sistemas do governo de Alagoas, justificada pelo potencial de identificar regiões mais ou menos beneficiadas. Os dados foram obtidos por meio de uma consulta delegada a um motor de busca distribuído da nossa arquitetura. Em termos de eficiência do *Data Lake* lógico proposto, o tempo médio de consulta foi de aproximadamente 4 segundos, com redução de 12.45% (3.54) e 4.4%(3.38)

segundos ao considerarmos mais máquinas na infraestrutura de consulta.

2. Trabalhos Relacionados

Com o barateamento e evolução de tecnologias de armazenamento, soluções de *Big Data* e Computação em Nuvem [Al-Ahmad and Kahtan 2018] têm se consolidado. De modo complementar, *Data Lakes* vêm ganhando notoriedade por seu poder de organização, documentação e governança de dados para agregação de valor a negócios e iniciativas governamentais. Diversos governos já aplicaram esforços de informatização de serviços e transparência de dados [Attard et al. 2015], e estudos mostram que o público tem recebido bem estes esforços [Welch et al. 2005], mas há poucos relatos de como integrar tais serviços, de modo a ter rápido acesso a análises e proposições de medidas que aumentem o valor de ações governamentais.

No trabalho de [Jetzek et al. 2014], podemos observar a proposta de um mecanismo de geração de valor a partir de dados abertos governamentais. O mecanismo tem base em quatro premissas: i) transparência governamental; ii) colaboração cidadã; iii) eficiência; iv) inovação. Todo processamento e análises são realizados por uma empresa privada. No desenvolvimento da solução, os autores mencionam desafios como dados inacessíveis, falta de políticas claras de uso, validação, precaridade de metadados e falta de interoperabilidade técnica e semântica. Tais problemas podem ser mitigados considerando o uso de um *Data Lake* lógico.

No trabalho de [Pereira et al. 2017], os autores definem e utilizam um modelo conceitual de análise para identificar quais são as melhorias obtidas em cidades inteligentes que aderiram ao movimento dos dados abertos governamentais. As melhorias identificadas impactam o setor da economia, saúde pública, segurança, qualidade de vida e bem estar. Como mencionado pelos autores, uma limitação do trabalho é a realização da análise para apenas um caso, que precisou aplicar esforços em integrar iniciativas relacionadas para avaliação de resultados.

Data Lakes tipicamente agregam dados de múltiplas fontes e requerem medidas eficazes de documentação e governança [Mehmood et al. 2019]. A existência de um ecossistema integrado possibilitaria a repetição em diferentes cenários para análises semelhantes e generalização dos resultados. A arquitetura proposta por [Li et al. 2018] permite rápida análise de dados de monitoramento de redes elétricas, um serviço essencial em centros urbanos. No trabalho, os autores apresentam funcionalidades de ingestão de dados, organização, flexibilidade à mudanças, análise e visualização.

Os estudos e esforços observados na literatura mostram que estruturas governamentais já vem obtendo resultados positivos do processamento e análise sistemática de dados [Stefanovic et al. 2016]. Mas é importante a adoção de plataformas confiáveis e robustas para organização, governança, análise e proposição de melhorias de ações sociais. Por alinhamento de propósito, fica evidente a importância de *Data Lakes* para sistematizar e agilizar o processo de geração de valor de ações governamentais.

Nos trabalhos citados, os autores se depararam com o acesso a dados nem sempre em boas condições, integração fraca ou manual, e desafio para análises e replicação de resultados, recorrendo inclusive à soluções do mercado, o que nem sempre é possível devido a exposição da privacidade dos usuários a terceiros. Como diferencial no contexto

governamental, nossa proposta visa mitigar esses desafios, criando condições de análise sistematizada para fácil replicação de resultados em uma plataforma integrada, considerando aplicações já existentes em um *Data Lake* lógico, o que evita replicação de dados e custo adicional com estrutura de armazenamento.

3. Data Lakes: Arquitetura e Tecnologias Usadas

Uma representação geral de arquitetura de *Data Lake* pode ser vista na Figura 1, a mesma é baseada no modelo de camadas (ou zonas) de [Sawadogo and Darmont 2021], permitindo Ingestão de dados, Destilamento, Processamento e Ideação. Transversalmente a todas as camadas, temos a camada de governança de dados para garantias de segurança, privacidade, qualidade e monitoramento. Para cada uma das camadas, temos diferentes níveis de governança, e quanto mais distante da camada de ingestão, maior o grau de governança dos dados.

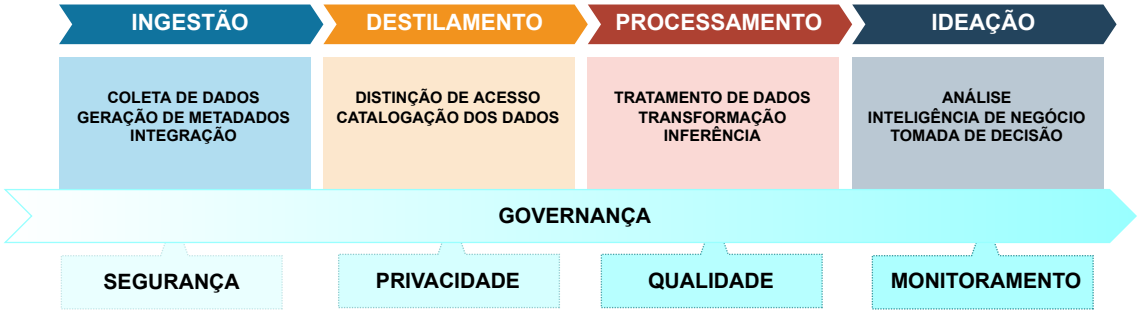

Figura 1. Arquitetura em zonas para *Data Lakes*.

Durante a etapa de ingestão, os dados são armazenados em sua forma crua, ou com processamento mínimo. Há integração de diferentes aplicações, caracterizando um meio de armazenamento não estruturado, volumoso, interconectado e enriquecido pela geração de metadados. Diferentemente do que é realizado em bases de dados estruturadas e organizadas, tipicamente conhecidas como armazéns (*Data Warehouse*), o não processamento em *Data Lakes* durante a ingestão visa evitar prejuízos de informação por descarte de dados. Todo tratamento e processamento é feito em estágios futuros para ganho máximo de informação.

Em geral, as arquiteturas propostas necessitam que os dados sejam copiados de suas respectivas fontes para um sistema de gerenciamento de arquivos distribuído como o *Hadoop Distributed File System* (HDFS) [Shvachko et al. 2010]. Por fonte de dados, é possível tomar como exemplo uma tabela em um banco de dados relacional, um arquivo no formato XML (*Extensible Markup Language*) ou texto puro. As principais desvantagens dessa metodologia é o uso adicional de armazenamento e a necessidade de inserções e atualizações periódicas de dados [Stefanowski et al. 2017]. As referidas desvantagens podem ser contornadas por meio de uma abordagem alternativa: o *Data Lake* lógico (ou federado). Neste, os dados são acessados diretamente em suas fontes. Contudo, as implementações dessa abordagem, em geral, apresentam uma perda de desempenho em relação à consulta dos dados, muito embora utilize-se mecanismos de *cache* para amenizar o problema [Stefanowski et al. 2017].

O destilamento dos dados consiste em segmentação e catalogação. Nesta etapa, é possível realizar separação por qualidade, permissão de acesso, propósito de uso, e outros

critérios relevantes. Após a separação, obtém-se, por exemplo, porções de dados de acesso livre, restrito ou anonimizados. Uma vez que metadados possibilitam a catalogação, um refinamento adicional da etapa de destilação seria uma separação por iniciativas de saúde, educação, segurança pública, etc. O ambiente integrado permite acesso a todos estes dados a depender das restrições. A granularidade da destilação pode se tornar tão fragmentada quanto necessário para atender às políticas de governança e ao mesmo tempo facilitar consultas.

Na camada de processamento, os usuários aplicam ou tem acesso a tratamentos para por exemplo, descartar dados faltantes, preencher lacunas ou realizar transformações. Conjuntos amplos de dados não processados permitem uma triagem mais seletiva e criteriosa, mostrando padrões antes ocultos pelo processamento precoce. Um maior poder de escolha também pode favorecer a engenharia de característica [Heaton 2016] para maximização de eficiência e o ganho de informação em algoritmos de inferência e avaliação de resultados das políticas ou modelos de negócio observados. Uma vez que *Data Lakes* integram múltiplas fontes de dados, os usuários podem se beneficiar de alto poder do processamento distribuído e análises horizontais para diferentes tipos de dados, de maneira rápida e sistemática para reprodução.

A camada de ideação contém um arcabouço de ferramentas que expõem resultados e eventos para gestores e interessados no sucesso de negócio ou aplicação específica, permitindo análise criteriosa e tomada de decisão com base em dados para melhoria contínua. A Tomada de Decisão Orientada por Dados (DOD) [Provost and Fawcett 2013] permite fundamentar decisões tendo como apoio a análise de dados, ao invés de intuição. Quanto mais orientada por dados, mais produtiva uma empresa é ou um órgão governamental. Os tipos de decisões que interessam se enquadram, mais especificamente, em duas categorias: (1) decisões para as quais "descobertas" precisam ser feitas nos dados e, (2) decisões que se repetem, principalmente em grande escala. Desta forma, a tomada de decisão pode se beneficiar até em condições de pequenos aumentos na precisão deste processo com base em análise de dados.

Para toda camada, há uma série de fatores constantemente presentes para garantir uma boa governança, evitando o surgimento de vulnerabilidades ou a transformação do *Data Lake* em um pântano de dados (*Data Swamp*) [Gorelik 2019]. Pântanos de dados são conjuntos de dados inutilizáveis que apenas geram despesas a seus mantenedores, seja por má gerência de metadados ou por inacessibilidade dos usuários. *Data Lakes* demandam minucioso cuidado e monitoramento, pois são soluções compostas e complexas de armazenamento, possuindo vantagens e desvantagens. Para aplicações mais simples, com baixa variedade e demanda de escalabilidade, *Data Warehouses* ainda são soluções mais viáveis.

Tabela 1. *Data Warehouse* vs. *Data Lake*: Vantagens e desvantagens.

	Data Warehouse	**Data Lake**
Vantagens	Fácil manutenção; Maturidade tecnológica; Consultas otimizadas; Usabilidade	Volume e variedade; Flexível à mudanças; Processamento em lotes; Análises avançadas; Governança
Desvantagens	Limitação de dados; Resistente à mudanças	Requer alta expertise técnica; Manutenção complexa

A Tabela 1 contém algumas das vantagens de se utilizar *Data Lakes* ou *Data Warehouses*, sendo as características de um sistema determinantes para a decisão de uso de

um, de outro, ou das duas soluções. Um *Data Warehouse* armazena dados previamente processados e estruturados, facilitando a consulta para modelos específicos de negócio. O amplo número sistemas de gerenciamento de bancos de dados (PostgreSQL, MongoDB, MySQL, etc.) desenvolvido ao longo dos últimos anos facilita a usabilidade e a manutenção dessas tecnologias.

Entre as limitações de *Data Warehouses*, podemos mencionar a resistência à mudanças e adaptação, que tipicamente afeta modelos de negócio, e também o acesso a um conjunto de dados limitado, que pode não fornecer informações suficientes para entendimento de eventos e tomadas e decisão. Por outro lado, *Data Lakes* oferecem mais flexibilidade, podendo armazenar e integrar diferentes tipos de dados, sejam estes estruturados, semi-estruturados ou sem qualquer tipo de estrutura. Considerando o baixo custo de armazenamento, o volume dos dados pode crescer mais do que o suficiente para proporcionar resultados ricos em informação. Também é possível realizar processamento e consultas em lote com auxílio de algoritmos avançados de Computação Distribuída [He and Da Xu 2012].

Como não há preocupação com estruturação de dados, *Data Lakes* são mais flexíveis à mudanças [Fang 2015], constantemente monitoradas e sujeitas à políticas de governança em todos os níveis de acesso. A garantia de consistência em múltiplas camadas pode tornar as atividades de manutenção mais complexas. Quanto maior o volume e variedade dos dados, maior a necessidade de uma expertise maior por parte dos usuários que realizam consultas e análises. A decisão de desenvolver ou não um *Data Lake* depende da quantidade de sistemas envolvidos e do grau de integração dos mesmos.

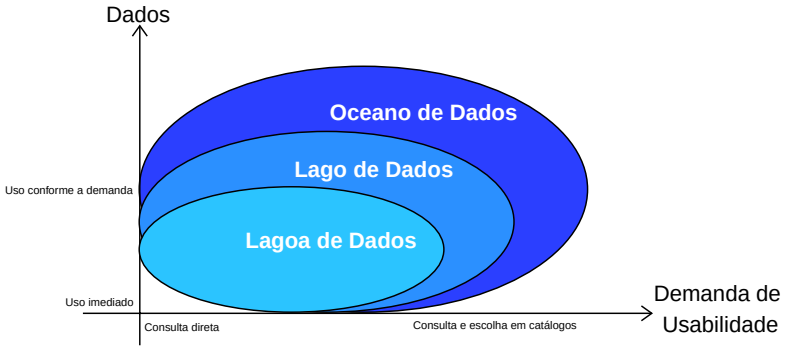

Figura 2. Comparativo de proporção entre lagoa, lago e oceano de dados. Adaptada de [Gorelik 2019].

Instituições que possuem um número reduzido de *Data Warehouses* e pretendem conectá-los, podem na verdade estar criando uma lagoa de dados (*Data Pond*), que como mostra a Figura 2, tem uma proporção menor quando comparada a um lago (*Data Lake*). A governança tipicamente não é boa nestas situações e a geração de valor é limitada. No outro extremo, temos os oceanos de dados (*Data Oceans*). Da perspectiva de uma aplicação que consome dados governamentais, o oceano de dados estaria vinculado a uma plataforma que integra todo e qualquer sistema em nível municipal, estadual e nacional. Validações cruzadas poderiam se beneficiar de dados de sistemas de saúde, educação, tributação e segurança pública, potencializando a detecção de fraudes intersetoriais ou a identificação de novas ações sociais para diferentes segmentos populacionais.

4. Data Lake Lógico para Dados de Aplicações Governamentais

Neste artigo propomos uma arquitetura de *Data Lake* lógico para integrar três bases de dados de sistemas governamentais. A arquitetura, que pode ser vista na Figura 3, utiliza unicamente tecnologias de código aberto para facilitar sua reprodutibilidade e acessibilidade a entidades governamentais que eventualmente queiram implementá-la. São elas: 1. Apache Spark [Zaharia et al. 2016]: uma ferramenta de processamento paralelo e distribuído em aplicações de *Big Data*; 2. PrestoDB [Sethi et al. 2019]: um motor de consultas distribuído; 3. Apache Ambari [Wadkar and Siddalingaiah 2014]: ferramenta desenvolvida para supervisionar e gerir o Apache HDFS; 4. Apache Hive [Huai et al. 2014]: uma interface para consultas de sistemas de arquivos utilizando linguagem estruturada; 5. Apache HDFS: sistema de arquivos distribuído.

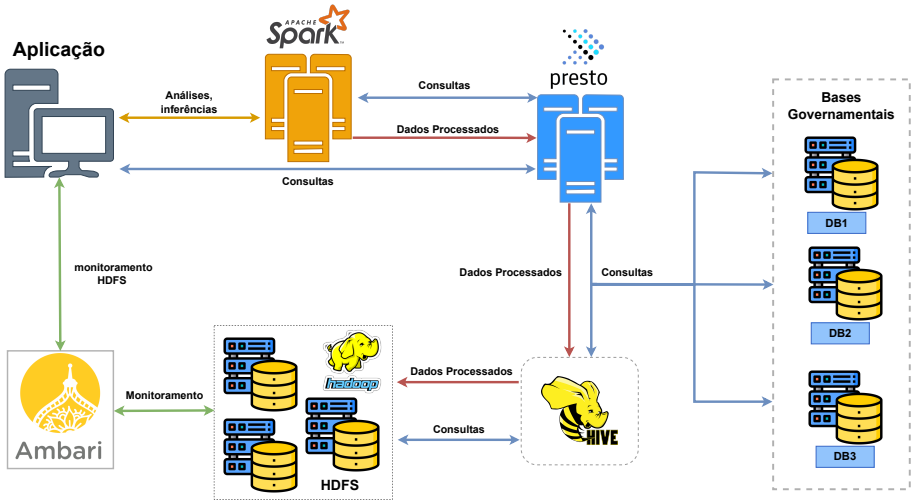

Figura 3. Arquitetura de *Data Lake* lógico.

O acesso às bases de dados vinculadas ao *Data Lake* é garantido por um motor de consultas distribuído. Este viabiliza a realização de buscas de maneira logicamente centralizada e por meio de uma linguagem única de consulta, ignorando a heterogeneidade da estrutura dos dados. O acesso centralizado é referente ao fato de que as consultas são destinadas a um único servidor, o nó central dessa ferramenta. Esse por sua vez é responsável por escalonar a tarefa de busca de forma distribuída entre os nós trabalhadores do sistema. Em nossa arquitetura, escolhemos o PrestoDB para condução das consultas, dada sua eficiência em comparação à soluções concorrentes [Mami et al. 2019]. Comumente, resultados de consultas em um *Data Lake* são submetidos a processamento para geração de gráficos ou modelos de inferência. As análises e manipulações dos dados são variadas e determinadas pelas necessidades dos gestores e aplicações, sendo limitadas apenas pela expertise técnica do time de desenvolvedores, analistas e cientistas de dados.

Conforme a arquitetura, utilizamos o Apache Spark em comunicação direta com o PrestoDB, requisitando dados e solicitando o salvamento dos resultados de processamentos. Esta é uma prática comum com o objetivo de evitar recálculos e enriquecer o *Data Lake* com mais dados [Sawadogo and Darmont 2021]. Os dados processados não são salvos nas bases de dados dos sistemas integrados pelo *Data Lake*. O PrestoDB solicita que o Apache Hive os armazene. O papel deste último é atuar como uma interface entre o

PrestoDB e o Hadoop, que é o verdadeiro responsável por guardar os dados produzidos pelo Spark. Além disso, após seu armazenamento, tais dados tornam-se disponíveis para serem consultados pelo PrestoDB através do Apache Hive. Finalmente, considerando a importância dos dados gerados, é relevante monitorar a saúde do *cluster* Hadoop onde os mesmos estão guardados. O monitoramento é realizado pelo Apache Ambari. O estado do HDFS é reportado ao usuário constantemente através de uma aplicação. Essa aplicação também é responsável por solicitar ao Spark a realização de procedimentos sobre os dados do *Data Lake* de acordo com as demandas do usuário.

5. Resultados e Discussão

Uma parte da arquitetura apresentada foi utilizada para analisar a disposição geográfica dos usuários beneficiados pelos referidos sistemas do governo de Alagoas. A escolha desta aplicação justifica-se no seu potencial em identificar as regiões mais beneficiadas por ações do estado, bem como a sua capilaridade e alcance. Além disso, permite aos gestores localizarem as regiões com maior carência de incentivos governamentais para o planejamento de iniciativas complementares. Para a execução desta análise, foi requisitada uma amostra dos valores únicos de CEP e suas respectivas frequências absolutas. Estas consultas foram delegadas ao motor de busca conforme o fluxo de execução da Figura 4, de modo que, através de uma aplicação, o usuário requisita os dados desejados, que são localizados em suas respectivas bases de dados pelo PrestoDB e retornados para a aplicação. O resultado da consulta pode ser visto na Figura 5.

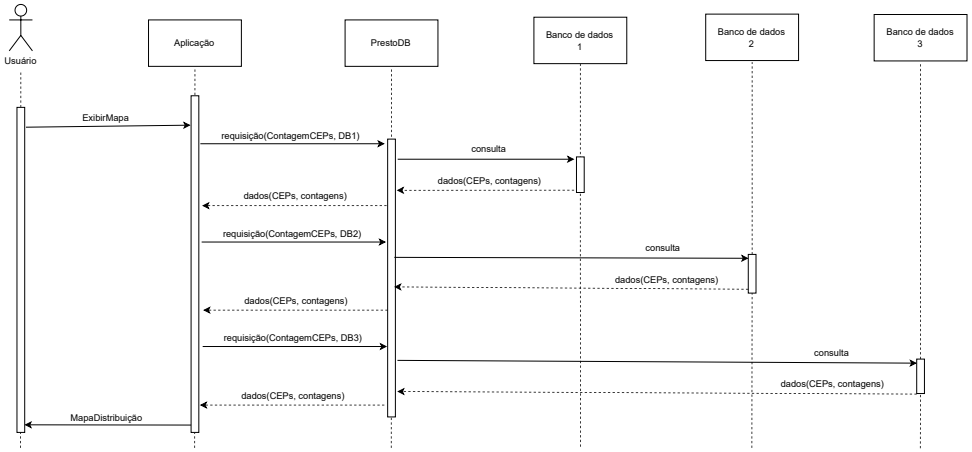

Figura 4. Fluxo de execução do *Data Lake* para a geração do mapa de usuários.

Em posse dos valores de CEP e suas contagens, a aplicação obtém coordenadas de latitude e longitude, que são utilizadas para desenhar um conjunto de pontos sobre um mapa. Os pontos tem diâmetro proporcional às respectivas frequências. As cores vermelha, azul e púrpura indicam dados do sistema de assistência social, promoção de práticas esportivas e capacitação em tecnologias da informação, respectivamente.

A aplicação Web utilizada para visualizar dados neste experimento foi desenvolvida utilizando o *framework* Streamlit. Além disso, por motivos de privacidade, os dados de CEP e suas frequências sofreram modificações para a geração da Figura 5, não refletindo a situação real da distribuição de usuários e dos sistemas que integram o *Data Lake*. Pelo volume de dados, consultas em *Data Lakes* são um grande desafio, principalmente se

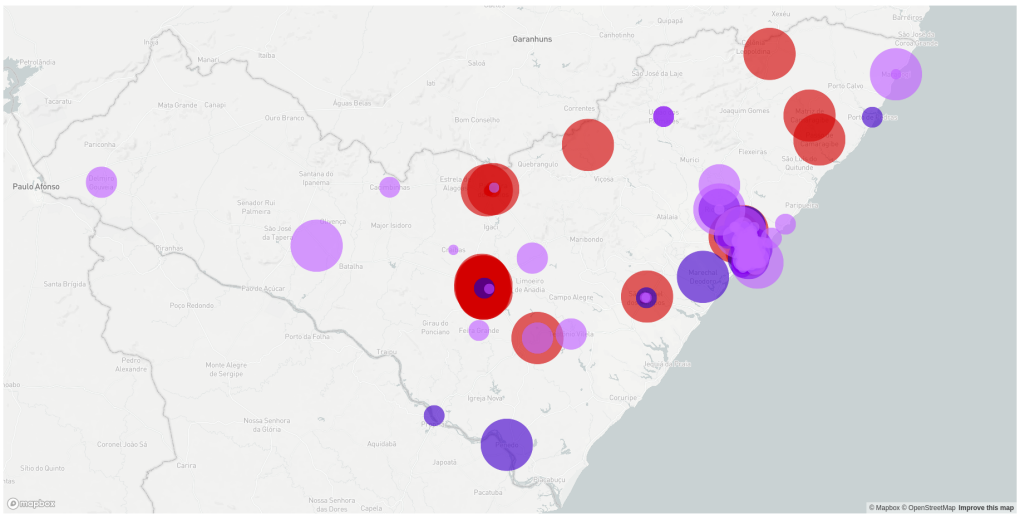

Figura 5. Posição geográfica e a proporção dos usuários de cada sistema.

os dados estiverem em diferentes sistemas. Mediante busca, os dados precisam ser encontrados, unidos e devolvidos. Na ausência de mecanismos de generalização dos termos de busca, os usuários devem inconvenientemente realizar a mesma consulta múltiplas vezes para diferentes bases de dados. Eficiência e desempenho são fatores importantes.

Na arquitetura proposta, esse mecanismo de generalização reside no motor de consultas distribuído. O mesmo é composto por um conjunto de um ou mais computadores conectados, cada um deles atuando como um nó coordenador ou trabalhador. O primeiro é responsável por receber as requisições de consultas, distribuí-las entre os nós trabalhadores e combinar os resultados retornados por estes em um resposta final. O último desempenha buscas nas bases de dados conectadas e retornam o resultado ao nó coordenador. Essa abordagem possibilita estabelecer múltiplas configurações a partir do número de nós envolvidos, o que impacta na performance do sistema e do *Data Lake*.

A fim de avaliar o desempenho dessa ferramenta de acesso à bases de dados, realizamos repetidas consultas com as seguintes configurações: 1. **Nó único**: O PrestoDB é instalado em um único computador, atuando como nó coordenador e trabalhador no sistema de buscas. O computador empregado dispõe de 32 núcleos de processamento e 64GB de memória; 2. **Um coordenador e um trabalhador**: O PrestoDB é instalado em dois computadores, onde um atua como nó coordenador e o outro como nó trabalhador. As máquinas encontram-se na mesma rede e ambas dispõem de 32 núcleos de processamento e 64GB de memória; 3. **Um coordenador e dois trabalhadores**: Na configuração anterior, adiciona-se um nó trabalhador na rede local. A terceira máquina possui 8 núcleos de processamento e 20GB de memória.

O resultado das consultas realizadas é apresentado no gráfico da Figura 5. Para cada uma das configurações, esse conjunto de consultas foi realizado 30 vezes, armazenando sempre o tempo decorrido para realizá-las e obter seus resultados. Os *boxplots* da variável de tempo decorrido por configuração do PrestoDB podem ser visualizados na Figura 6. É importante ressaltar que as consultas foram realizadas através de uma aplicação Web, executada em uma rede externa à rede do PrestoDB, isso significa que os resultados estão sujeitos a atrasos de conexão da Internet.

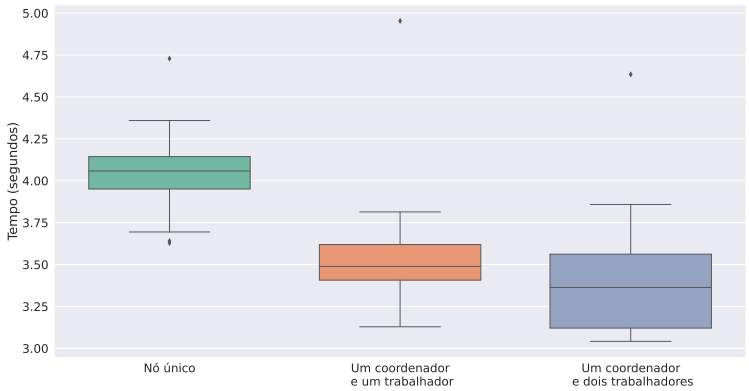

Figura 6. Tempos de consulta do PrestoDB com diferentes configurações.

Analisando o gráfico da Figura 6, é possível observar melhoria significativa na adição de um nó trabalhador à configuração de nó único, isso permitiu que o nó principal atuasse exclusivamente como coordenador. Para a adição de um segundo nó trabalhador, não verificamos melhoria significativa. As possíveis causas para esse fenômeno podem ser a baixa complexidade das consultas executadas, o pequeno número de bases de dados conectadas ou o menor poder computacional da terceira máquina. Na Tabela 2, pode-se observar o decréscimo dos valores de tempo de execução médio ($\hat{\mu}_t$), mediana, mínimo e máximo conforme acrescentam-se nós ao sistema. A melhoria é menos relevante entre as duas últimas configurações, visto que reduz-se $\hat{\mu}_t$ em 12.45% e em 4.4% com a inserção do segundo e terceiro nó, respectivamente. Adicionalmente, há um nítido crescimento do desvio padrão amostral ($\hat{\sigma}_t$) a medida que aumentamos o número de nós. Estas duas últimas observações corroboram a análise gráfica com indícios de que o terceiro nó foi pouco significativo para o ganho de desempenho.

Tabela 2. Estatísticas acerca de tempo de consulta do PrestoDB.

Configuração	em segundos				
	$\hat{\mu}_t$	$\hat{\sigma}_t$	min	$mediana$	max
Nó único	4.048	0.215	3.628	4.058	4.728
Um coordenador e um trabalhador	3.544	0.305	3.129	3.489	4.953
Um coordenador e dois trabalhadores	3.388	0.330	3.041	3.363	4.634

6. Conclusão

Este trabalho discutiu uma arquitetura para *Data Lakes* lógicos para uma aplicação real de sistemas e dados governamentais no estado de Alagoas. O arcabouço de processamento e análise de *Data Lakes* possibilita integração de aplicações diversas sob a gestão do estado, visando amplificar eficiência e coesão das ações do governo na tomada de decisão. Estamos continuamente refinando a arquitetura e sua implantação para facilitar a organização e análise dos dados. Preliminarmente, apresentamos uma análise da distribuição dos usuários de três sistemas governamentais, mas pretendemos construir um arcabouço de análises para, por exemplo, identificar perfis de usuário, observar impacto de políticas públicas e realizar projeções variadas de crescimento. Aplicação também pode ser estendida para englobar um número maior de sistemas. Ainda como trabalho futuro, pretendemos explorar e discutir em mais detalhes as políticas de governança e

qualidade dos dados a ser implantada, realizar um estudo sobre o desempenho dos mecanismos de busca e processamento, e propor soluções de busca semântica em múltiplas bases de dados, bem como a inserção de um sistema robusto de gestão de metadados.

Agradecimentos

O presente trabalho foi realizado com apoio da Fundação de Amparo à Pesquisa do Estado de Alagoas (FAPEAL), e da Secretaria de Estado da Ciência, Tecnologia e Inovação de Alagoas (SECTI-AL).

Referências

Al-Ahmad, A. S. and Kahtan, H. (2018). Cloud Computing Review: Features and Issues. In *International Conference on Smart Computing and Electronic Enterprise (ICSCEE'18)*.

Attard, J., Orlandi, F., Scerri, S., and Auer, S. (2015). A Systematic Review of Open Government Data Initiatives. *Government Information Quarterly*, 32(4):399–418.

Božič, K. and Dimovski, V. (2019). Business Intelligence and Analytics for Value Creation: The Role of Absorptive Capacity. *International Journal of Information Management*, 46:93–103.

Fang, H. (2015). Managing data lakes in big data era: What's a data lake and why has it became popular in data management ecosystem. In *2015 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER)*, pages 820–824. IEEE.

Gorelik, A. (2019). *The Enterprise Big Data Lake: Delivering the Promise of Big Data and Data Science*. O'Reilly Media.

He, W. and Da Xu, L. (2012). Integration of Distributed Enterprise Applications: A Survey. *IEEE Transactions on Industrial Informatics*, 10(1):35–42.

Heaton, J. (2016). An Empirical Analysis of Feature Engineering for Predictive Modeling. In *IEEE Region 3 South East Conference (SoutheastCon'16)*.

Huai, Y., Chauhan, A., Gates, A., Hagleitner, G., Hanson, E. N., O'Malley, O., Pandey, J., Yuan, Y., Lee, R., and Zhang, X. (2014). Major Technical Advancements in Apache Hive. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, pages 1235–1246.

IRENA Group (2019). Innovation Landscape Brief: Internet of Things. Book ISBN 978-92-9260-142-3, International Renewable Energy Agency, Abu Dhabi, United Arab Emirates.

Jetzek, T., Avital, M., and Bjorn-Andersen, N. (2014). Data-driven Innovation Through Open Government Data. *Journal of Theoretical and Applied Electronic Commerce Research*, 9(2):100–120.

Li, Y., Zhang, A., Zhang, X., and Wu, Z. (2018). A Data Lake Architecture for Monitoring and Diagnosis System of Power Grid. In *Artificial Intelligence and Cloud Computing Conference (AICC'18)*.

Liaqat, M., Chang, V., Gani, A., Ab Hamid, S. H., Toseef, M., Shoaib, U., and Ali, R. L. (2017). Federated Cloud Resource Management: Review and Discussion. *Journal of Network and Computer Applications*, 77:87–105.

Mami, M. N., Graux, D., Scerri, S., Jabeen, H., Auer, S., and Lehmann, J. (2019). Uniform Access to Multiform Data Lakes Using Semantic Technologies. In *21st International Conference on Information Integration and Web-based Applications & Services (IIWAS'19)*.

Marx, V. (2013). The Big Challenges of Big Data. *Nature*, 498(7453):255–260.

Mehmood, H., Gilman, E., Cortes, M., Kostakos, P., Byrne, A., Valta, K., Tekes, S., and Riekki, J. (2019). Implementing Big Data Lake for Heterogeneous Data Sources. In *IEEE 35th International Conference on Data Engineering Workshops (ICDEW'19)*.

Pereira, G. V., Macadar, M. A., Luciano, E. M., and Testa, M. G. (2017). Delivering Public Value Through Open Government Data Initiatives in a Smart City Context. *Information Systems Frontiers*, 19(2):213–229.

Provost, F. and Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-driven Decision Making. *Big data*, 1(1):51–59.

Sawadogo, P. and Darmont, J. (2021). On Data Lake Architectures and Metadata Management. *Journal of Intelligent Information Systems*, 56(1):97–120.

Sethi, R., Traverso, M., Sundstrom, D., Phillips, D., Xie, W., Sun, Y., Yegitbasi, N., Jin, H., Hwang, E., Shingte, N., et al. (2019). Presto: SQL on Everything. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 1802–1813. IEEE.

Shvachko, K., Kuang, H., Radia, S., and Chansler, R. (2010). The Hadoop Distributed File System. In *IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST'10)*.

Stefanovic, D., Marjanovic, U., Delić, M., Culibrk, D., and Lalic, B. (2016). Assessing the Success of E-government Systems: An Employee Perspective. *Information & Management*, 53(6):717–726.

Stefanowski, J., Krawiec, K., and Wrembel, R. (2017). Exploring Complex and Big Data. *International Journal of Applied Mathematics and Computer Science*, 27(4):669–679.

Wadkar, S. and Siddalingaiah, M. (2014). Apache Ambari. In *Pro Apache Hadoop*, pages 399–401. Springer.

Welch, E. W., Hinnant, C. C., and Moon, M. J. (2005). Linking Citizen Satisfaction With E-Government and Trust in Government. *Journal of Public Administration Research and Theory*, 15(3):371–391.

Zagan, E. and Danubianu, M. (2020). Data Lake Approaches: A Survey. In *International Conference on Development and Application Systems (DAS'20)*.

Zaharia, M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A., Meng, X., Rosen, J., Venkataraman, S., Franklin, M. J., et al. (2016). Apache Spark: A Unified Engine For Big Data Processing. *Communications of the ACM*, 59(11):56–65.