Centro Federal de Educação Tecnológica de Minas Gerais

Sistema Integrado de Gestão de Atividades Acadêmicas

Diretoria de Pesquisa e Pós-Graduação

Emitido em 17/09/2025 às 09:17

Projeto de Pesquisa

Dados do Projeto Pesquisa	
Código:	PIC1022-2024
Titulo do Projeto:	Integração de Componentes Open Source para Construção de Data Lakes
Tipo do Projeto:	INTERNO (Projeto Novo)
Natureza do Projeto:	Projeto de Desenvolvimento Científico e Tecnológico
Tipo de Pesquisa:	Pesquisa Aplicada
Situação do Projeto:	EM EXECUÇÃO
Unidade de Lotação do Coordenador:	DEPARTAMENTO DE COMPUTAÇÃO - NG (11.56.03)
Unidade de Execução:	DEPARTAMENTO DE COMPUTAÇÃO - NG (11.56.03)
Centro:	CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS (11.00)
Palavra-Chave:	data lakes, big data, compenentes open-source, HDFS, Hive, Spark
E-mail:	evandrino@cefetmg.br
Edital:	EDITAL DPPG Nº 73/2024 - PIBIC FAPEMIG
Cota:	PIBIC FAPEMIG 2025-2026 (28/01/2025 a 28/02/2026)
Objetivos de Desenvolvimento Sustentável	
9	Indústria, Inovação e Infraestrutura
Área de Conhecimento, Grupo e Linha de Pesquisa	
Área de Conhecimento:	Banco de Dados
Grupo de Pesquisa:	Não possui vínculo com grupo de pesquisa.
Linha de Pesquisa:	banco de dados
Comitê de Ética	
N° do Protocolo:	Não possui protocolo de pesquisa em Comitê de Ética.

Resumo

A necessidade de gerenciar, processar, armazenar e analisar grandes volumes de dados é cada vez mais comum em diversos setores. Soluções tradicionais, como bancos de dados escaláveis verticalmente, não são mais suficientes para lidar com o volume atual de dados. É essencial uma infraestrutura escalável e distribuída para atender às crescentes demandas de processamento, armazenamento e diversidade de dados. Esse cenário é representado pelas tecnologias conhecidas como big data (Hashem et al., 2015). Os data lakes desempenham um papel fundamental nessa tecnologia, fornecendo uma arquitetura flexível, centralizada e heterogênea para análises avançadas e uso em aprendizado de máquina (Khine e Wang, 2018; Giebler et al., 2019b). Diferentemente dos data warehouses tradicionais, os data lakes não aplicam esquemas na ingestão de dados, adotando uma abordagem schema-on-read, que acomoda dados estruturados, semiestruturados, não estruturados ou binários, permitindo consultas conforme a necessidade (Dixon, 2010). Embora as soluções proprietárias para data lakes ofereçam robustez e suporte técnico, elas frequentemente envolvem custos elevados e falta de transparência no controle de dados. Em contrapartida, ferramentas de código aberto, como Apache Hadoop, Spark, Hive, Ranger e Atlas, têm demonstrado ser alternativas viáveis, permitindo não apenas economia, mas também maior flexibilidade e personalização. No entanto, a integração desses componentes para formar uma solução coesa e otimizada continua sendo um desafio técnico significativo. Este projeto de pesquisa propõe explorar, avaliar e implementar um framework integrado de componentes open source para a construção de data lakes que atendam a três critérios fundamentais:
1) Segurança e Governança: Garantir o controle de acesso e a proteção de dados sensíveis.
2) Escalabilidade: Permitir o crescimento modular, adaptando-se a demandas crescentes de volume e complexidade de dados.
3) Desempenho: Maximizar a eficiência no processamento e nas análises, suportando cargas de trabalho diversificadas e consultas em tempo real.
O setor público e empresas privadas enfrentam barreiras para adotar soluções de código aberto devido à falta de conhecimento especializado e de frameworks que ofereçam integração simplificada. Este projeto visa preencher essa lacuna, fornecendo um modelo replicável que facilite a adoção de um framework para a construção de um data lake seguro e escalável, trazendo agilidade e eficiência às instituições.

Introdução/Justificativa

(incluindo os benefícios esperados no processo ensino-aprendizagem e o retorno para os cursos e para os professores da CEFET-MG em geral)

Avanços significativos têm sido feitos recentemente na infraestrutura de data lakes de código aberto, especialmente no contexto governamental. Com o aumento da quantidade de dados gerados por organizações governamentais, surge a necessidade urgente de desenvolver soluções para armazenar, processar e analisar esses dados de forma eficiente e transparente. As soluções open-source, como Apache Hadoop, Apache Spark, Apache Hive, Apache Ranger, Apache Knox e Apache Atlas, têm se destacado como alternativas poderosas para a implementação de data lakes. No entanto, a adoção dessas tecnologias exige uma análise cuidadosa das necessidades específicas, como segurança, governança e escalabilidade, além de considerações sobre orçamento e a especialização técnica necessária para sua implementação e manutenção. Pesquisadores exploraram várias metodologias e ferramentas para enfrentar os desafios associados ao uso de data lakes open-source, com ênfase em segurança, escalabilidade e governança de dados. Para empresas e go, surgem desafios adicionais, como restrições orçamentárias, falta de especialização e a necessidade de conformidade regulatória, aspectos críticos no Brasil, considerando a Lei Geral de Proteção de Dados (LGPD). Apesar dessas dificuldades, a implementação de data lakes com ferramentas open-source tem se mostrado uma solução viável. Este estudo busca explorar como essas tecnologias podem ser integradas para construir um data lake seguro e escalável em instituições públicas e privadas de pequeno e médio porte, tornando os processos decisórios mais ágeis e confiáveis, além de promover maior eficiência e escalabilidade.

Diversos estudos recentes exploraram os data lakes, abordando temas como arquitetura, desafios e métodos de implementação com foco nas soluções open-source. A pesquisa de Azzabi et al. (2024) e Giebler et al. (2019a) destaca os principais componentes usados em arquiteturas de data lakes, enquanto Hai et al. (2023) e Zagan e Danubianu (2020) analisam os principais desafios enfrentados, como a integração com sistemas legados e as questões de segurança e privacidade. O Gartner Group, em pesquisa com CIOs do setor público, revelou que uma das principais prioridades nas organizações governamentais é o aumento do investimento em análise de dados e cibersegurança, destacando a necessidade de soluções eficazes para gerenciar dados em grande escala e garantir a segurança e privacidade das informações (Guess, 2019). No entanto, soluções comerciais como Cloudera Data Platform (CDP), MapR, Amazon EMR e Microsoft Azure HDInsight apresentam custos elevados, tornando-se inacessíveis para instituições com orçamentos restritos. Essas soluções, embora robustas, frequentemente se baseiam em tecnologias open-source como Hadoop, Spark, Hive e Impala (Ozgur e Coto, 2022; Mathis, 2017; Singh e Kaur, 2016), levantando a questão: por que pagar por soluções comerciais quando as mesmas tecnologias

(Ozgur e Coto, 2022; Mathis, 2017; Singh e Kaur, 2016), levantando a questão: por que pagar por soluções comerciais quando as mesmas tecnologias open-source oferecem o mesmo nível de eficácia a um custo muito menor?

A adoção de soluções open-source oferece flexibilidade, autonomia e custo-benefício, essenciais no contexto de orçamentos limitados. Além disso, a utilização de ferramentas como Apache Knox, Apache Ranger e Apache Atlas permite a implementação de uma infraestrutura robusta de segurança e governança. Apache Knox é um gateway de segurança que fornece autenticação e autorização centralizada, protegendo interfaces de dados e APIs, essenciais para garantir que apenas usuários autorizados acessem dados sensíveis (Ozgur e Coto, 2022). O Apache Ranger, integrado com Knox, oferece um sistema detalhado de políticas de autorização, permitindo o controle de acesso em nível de tabela, coluna e linha (Singh e Kaur, 2016). O Apache Atlas proporciona governança de dados e rastreamento de linhagem, facilitando a auditoria e a conformidade legal, um requisito essencial para garantir que as organizações sigam legislações como a LGPD no Brasil (Camacho-Rodríguez et al., 2019). Além dessas ferramentas de segurança, Hive, Spark e MapReduce são componentes essenciais para o processamento e análise de grandes volumes de dados. O Hive, framework de data warehouse baseado em Hadoop, permite consultas SQL sobre dados distribuídos, facilitando o trabalho dos analistas de dados (Sharmila et al., 2019). Isso permite a análise de dados estruturados, fundamental para processos decisórios eficientes. O Apache Spark, por sua vez, oferece vantagens significativas sobre o MapReduce, especialmente para tarefas de análise interativa e em tempo real, utilizando fluxos de dados on-line. O Spark permite realizar análises mais rápidas e complexas, essenciais para operações de análise em tempo real. O MapReduce, embora mais lento, continua sendo eficaz para o processamento em lote, ideal para grandes volumes de dados que precisam ser processados de maneira distribuída e escalável.

Objetivos

O problema central envolve a construção de um data lake eficiente, seguro e escalável usando soluções open-source, atendendo às necessidades específicas de segurança e governança de dados em organizações públicas e privadas com orçamentos limitados.

Objetivo:
1)Explorar como integrar componentes open-source (Hadoop, Spark, Hive, Knox, Ranger, Atlas) para construir um data lake eficiente, seguro e escalável.
2) Avaliar a viabilidade dessas soluções para o contexto de organizações públicas e privadas no Brasil, com ênfase em segurança de dados e governança, envolvendo grandes volumes de dados.

Particularmente, este projeto de pesquisa tem esse objetivo, basendo-se em três critérios fundamentais:
1) Segurança e Governança: Garantir o controle de acesso, proteção de dados sensíveis e conformidade com regulamentações.
2) Escalabilidade: Permitir o crescimento modular, adaptando-se a demandas crescentes de volume e complexidade de dados.
3) Desempenho: Maximizar a eficiência no processamento e nas análises, suportando cargas de trabalho diversificadas e consultas em tempo real.

Contribuições esperadas:
1) Um framework detalhado para integração de ferramentas open source, com documentação de implementação e melhores práticas.
2) Avaliação comparativa de desempenho e segurança de diferentes configurações.
3) Um modelo replicável para implatanção de data lakes seguros e escaláveis, adequado para diversos contextos organizacionais.
4) Documentação para instalação e configuração do framework, assim como a divulgação em artigos científicos.

Metodologia

A metodologia deste projeto é apresentada a seguir em suas principais atividades.
1) Levantamento Bibliográfico: revisar estudos que destaquem a integração de ferramentas open source para data lakes, com foco em segurança, desempenho e escalabilidade (Dixon, 2010; Giebler et al., 2019a).

2) Validação dos ccomponentes Open-Souce: validar as ferramentas ou componentes do ecossistema Apache, como, por exemplo, Knox, Ranger, Atlas, Hadoop e Hive, bem como outras soluções open source, para garantir a integração dos componentes, visando segurança, governança e escalabilidade.

3) Projeto do framework: desenvolver uma arquitetura modular que integre componentes de armazenamento, processamento e segurança. O projeto incluirá uma camada de governança, baseada em Apache Ranger e Apache Atlas, uma camada de autenticação, centralizada com Apache Knox, uma camada de armazenamento de dados, baseada no Apache HDFS, uma camada de processamento de dados, baseada no Apache Hive, Apache MapReduce e Apache Spark.

4) Implementação e Testes: Implementar um protótipo do framework em um ambiente controlado, com testes de desempenho, escalabilidade e conformidade de segurança. A implementação será em ambiente de cloud computing, utilizando os recursos free-tiers de ambientes em nuvem computacionais, como o da Oracle Cloud.

5) Validação em Cenários Reais: Aplicar o framework em um caso de uso do setor público ou privado, avaliando sua eficácia e replicabilidade, considerando grande volume de dados.

6) Elaboração de Documentação e Artigos Científicos: Elaboração de documentação detalhada do projeto, de maneira que inclua como reproduzir a construção de um data lake, com relatos de configurações específicas necessárias à integração dos componentes a ser implementados. Elaboração de artigos científicos, cujos obetivos são a divulgação científica e reforço do CEFET-MG como entidade atuante no contexto do desenvolvimento técnico e tecnológico.

Referências

Azzabi, S., Alfughi, Z., Ouda, A. (2024). Data Lakes: A Survey of Concepts and Architectures. Computers, 13, 183.

Armbrust, M., Xin, R.S., Lian, C., et al. (2015). Spark SQL: Relational Data Processing in Spark. Proceedings of the 2015 ACM SIGMOD ICMD.

Camacho-Rodríguez, J., et al. (2019). Data Governance and Metadata Management with Apache Atlas. Proceedings of the 2019 ACM SIGMOD ICMD.

Guess, A. (2019). Gartner Survey Finds Government CIOs to Invest More in Data Analytics and Cybersecurity. Dataversity.

Giebler, C., Scherzinger, S., Habich, D. (2019). Big Data Systems Meet Data Lakes: Challenges, Open Issues, and Opportunities. Big Data Research, 18, 2033.

Hai, J., Wang, Y., et al. (2023). Challenges in Building Data Lakes for Big Data Analytics. Journal of Big Data (JBD).

Hukkeri, T.S., Kanoria, V., Shetty, J. (2020). A Study of Enterprise Data Lake Solutions. IRJET.

Mathis, C. (2017). Data Lakes: Technologie, Nutzen und Herausforderungen. Datenbank-Spektrum.

Mami, A., et al. (2019). Open-source Big Data Analytics and Integration in Data Lakes: A Review. JBD.

Nazari, E., Shahriari, M., Tabesh, H. (2019). Big Data Analysis in Healthcare: Apache Hadoop, Apache Spark, and Apache Flink. Frontiers in Health Informatics.

Ozgur, C., Coto, J. (2022). Usage of Hadoop and Microsoft Cloud in Big Data Analytics. AIMS IJM.

Sharmila, D., Somasundaram, K., Devi, R., Shanthi, C. (2019). Big Data Analysis Using Apache Hadoop and Spark. International Journal of Recent Technology and Engineering (IJRTE), 8, 167170.

Singh, R., Kaur, P. (2016). Analyzing Performance of Apache Tez and MapReduce with Hadoop Multinode Cluster on Amazon Cloud. JBD, 3.

Yang, C., Chen, C., Tsan, Y., Liu, P., et al. (2021). The Implementation of Data Storage and Analytics Platform for Big Data Lake of Electricity Usage with Apache Spark. The Journal of Supercomputing.

Zagan, J., Danubianu, M. (2020). A Review of Data Lakes and Their Implementation in Big Data Applications. IJDSA.

Membros do Projeto

CPF	Nome	Categoria	CH Dedicada	Tipo de Participação
###.###.913-##	EVANDRINO GOMES BARROS	DOCENTE	Não informada	Coordenador(a)

2025

Atividades	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
LEVANTAMENTO BIBLIOGRÁFICO												
VALIDAÇÃO DOS COMPENENTES OPEN-SOURCE												
PROJETO DO FRAMEWORK												
IMPLEMENTAÇÃO E TESTES												
VALIDAÇÃO EM CENÁRIOS REAIS												
ELABORAÇÃO DE DOCUMENTAÇÃO E ARTIGOS												

2026

Atividades	Jan	Fev
LEVANTAMENTO BIBLIOGRÁFICO		
VALIDAÇÃO DOS COMPENENTES OPEN-SOURCE		
PROJETO DO FRAMEWORK		
IMPLEMENTAÇÃO E TESTES		
VALIDAÇÃO EM CENÁRIOS REAIS		
ELABORAÇÃO DE DOCUMENTAÇÃO E ARTIGOS		

Avaliações do Projeto

Situação/Parecer	Data da Avaliação	Média
AVALIAÇÃO REALIZADA A proposta é aderente ao edital e apresenta todos os elementos adequadamente.	17/01/2025	25.0

Histórico do Projeto

Data	Situação	Usuário
09/12/2024	CADASTRO EM ANDAMENTO	EVANDRINO GOMES BARROS / ###.###.913-##
09/12/2024	SUBMETIDO	EVANDRINO GOMES BARROS / ###.###.913-##
17/12/2024	DISTRIBUÍDO PARA AVALIAÇÃO (AUTOMATICAMENTE)	DENIS EMANUEL DA COSTA VARGAS / ###.###.996-##
18/12/2024	DISTRIBUIÇÃO PARA AVALIAÇÃO (MANUALMENTE)	DENIS EMANUEL DA COSTA VARGAS / ###.###.996-##
31/01/2025	APROVADO	DENIS EMANUEL DA COSTA VARGAS / ###.###.996-##
28/04/2025	EM EXECUÇÃO	EVANDRINO GOMES BARROS / ###.###.913-##

Documento emitido por: EVANDRINO GOMES BARROS