FLAMI vs GADDS: Estudo Comparativo de Arquiteturas de Big Data Aderentes aos Princípios FAIR

Cecíla Nunes Sedenho[1], Thiago Zero Araujo[1], João Pedro de Carvalho Castro[1,2], Cristina Dutra Aguiar[1]

[1]Departamento de Ciências de Computação, Universidade de São Paulo, Brasil

[2]Diretoria de Tecnologia da Informação, Universidade Federal de Minas Gerais, Brasil

{ceciliasedenho, thzeroa}@usp.br, jpcarvalhocastro@ufmg.br, cdac@icmc.usp.br

Abstract. *This paper investigates two Software Reference Architectures (SRAs) for big data that adhere to the FAIR principles: FLAMI and GADDS. We describe these state-of-the-art SRAs using theoretical concepts, architectural designs, and implementation aspects of selected components to assess their practical feasibility. The compliance with the FAIR Principles is discussed based on structural and operational criteria. We also describe the data and metadata flow during an analytical query execution. The paper fills a gap in the literature by providing a theoretical and practical comparison between the investigated SRAs.*

Resumo. *Neste artigo são investigadas duas Arquiteturas de Referência de Software (SRAs) para big data que aderem aos princípios FAIR: FLAMI e GADDS. Essas SRAs, que representam o estado da arte, são descritas considerando conceitos teóricos, projetos arquiteturais e aspectos de implementação de componentes selecionados para avaliação da viabilidade prática. A conformidade de cada SRA com os Princípios FAIR é discutida com base em critérios estruturais e operacionais. Também é feita uma descrição do fluxo de dados e de metadados durante a execução de uma consulta analítica. O artigo contribui para preencher uma lacuna existente na literatura ao conduzir uma comparação teórica e prática entre as SRAs investigadas.*

1. Introdução

No cenário atual de *big data*, com os avanços em computação em nuvem e processamento distribuído, dados e metadados são coletados em taxas cada vez maiores. Esse volume impõe desafios de armazenamento e consulta, gerando complexidade de gestão de máquinas e recursos. Tais dificuldades podem ser mitigados pela terceirização da infraestrutura via computação em nuvem [Bhowmik 2017]. A abstração e o uso dessas tecnologias podem impulsionar o compartilhamento dos dados e dos *insights* derivados dos mesmos.

A Ciência Aberta permite a colaboração no meio científico, de forma a impulsionar o compartilhamento de dados e metadados entre pesquisadores. O objetivo é tornar a produção digital de objetos de pesquisa gratuitamente acessível [Medeiros et al. 2020].

Tal compartilhamento demanda uma infraestrutura dedicada, implementada de forma padronizada para evitar eventuais conflitos de compatibilidade. Uma alternativa de padronização refere-se aos Princípios FAIR (*Findability*, *Accessibility*, *Interoperability* e *Reusability*) [Wilkinson et al. 2016], os quais constituem uma série de requisitos que descrevem como repositórios devem ser implementados para dar suporte à exploração, compartilhamento e reutilização manual e automatizada.

Entretanto, os Princípios FAIR por si só não são suficientes para subsidiar a implementação de infraestruturas capazes de lidar com os desafios impostos pela Ciência Aberta. Logo, a adoção de uma Arquitetura de Referência de Software (SRA) pode suprir a complexidade de implementação em ambientes de *big data*. Uma SRA é um modelo que engloba o conhecimento sobre como projetar arquiteturas concretas de sistemas, adaptadas aos requisitos de contextos específicos [Angelov et al. 2012].

Apesar da necessidade de SRAs de *big data* aderentes aos Princípios FAIR em Ciência Aberta, a presença dessas na literatura é limitada. Existem somente quatro soluções, segundo [Castro et al. 2025b]: BigFAIR [Castro et al. 2022a], CloudFAIR [Castro et al. 2022b], GADDS [Vazquez et al. 2022] e FLAMI [Castro et al. 2025b]. O presente artigo se propõe a analisar as duas SRAs que representam o estado-da-arte: FLAMI (FAIR *Lakehouse Architecture for Multiple Infrastructures*) e GADDS (*Globally Accessible Distributed Data Sharing*).

O objetivo do artigo é comparar as SRAs FLAMI e GADDS em termos de componentes e da implementação dessas SRAs considerando instanciações no domínio da sismologia. Nas instanciações, visa-se apresentar uma visão de baixo nível dessas SRAs, investigando como seus componentes atuam para garantir a conformidade com os Princípios FAIR. O artigo também descreve o fluxo de dados e de metadados considerando a execução de uma mesma consulta analítica em ambas SRAs

O artigo está organizado como segue. Na seção 2 são descritos trabalhos relacionados. Na seção 3 são comparadas as SRAs FLAMI e GADDS. Na seção 4 são descritos os experimentos realizados. Na seção 5 são feitas as conclusões do artigo.

2. Trabalhos Relacionados

Os trabalhos relacionados a este artigo referem-se a dois estudos que realizam comparações entre arquiteturas. Em [Davoudian and Liu 2020], as SRAs de *big data* são comparadas sob a ótica da engenharia de *software*, descrevendo as arquiteturas Kappa, Lambda, Liquid, Solid e Bolster. No entanto, esse estudo não considera os Princípios FAIR e não realiza instanciações e consultas que avaliem o funcionamento das SRAs. A revisão sistemática de [Castro et al. 2025a] descreve BigFAIR, CloudFAIR e GADDS. Embora esse estudo discuta o grau de aderência dessas arquiteturas aos Princípios FAIR, o mesmo não aprofunda essa análise nos níveis de componentes e camadas, tampouco realiza instanciações ou execução de consultas.

A relevância do presente artigo reside na ausência de estudos comparativos que instanciem as SRAs comparadas a fim de analisar como as mesmas se comportam na execução de consultas. Outro diferencial refere-se à comparação de como cada SRA satisfaz aos Princípios FAIR. Portanto, o artigo oferece uma contribuição inédita para a comunidade científica, ao fornecer subsídios técnicos e metodológicos para o desenvolvimento e seleção de arquiteturas de *big data* aderentes ao Princípios FAIR.

3. Análise comparativa entre as SRAs FLAMI e GADDS

Esta seção faz uma comparação das SRAs GADDS e FLAMI em termos de aspectos lógicos (seção 3.1) e da instanciação das SRAs no domínio de sismologia (seção 3.2).

3.1. Visão lógica

A SRA FLAMI, voltada a cenários colaborativos e distribuídos, é uma evolução das SRAs BigFAIR e CloudFAIR, composta por seis camadas: (i) armazenamento externo; (ii) armazenamento de repositório; (iii) interação; (iv) recuperação de dados; (v) *insight* de dados; e (vi) mapeamento de conhecimento.

As camadas de armazenamento oferecem suporte a dois tipos de provedores: externos, que mantêm seus dados em repositórios fora da rede principal; e internos, que utilizam uma infraestrutura unificada com *data lake* e estrutura híbrida de *metadata lake* e *metadata warehouse*, com processamento dinâmico dos metadados entre os modelos. As camadas de mapeamento de conhecimento e *insight* de dados viabilizam a exploração semântica e o apoio à decisão. A camada de recuperação orquestra tarefas de consulta e processamento, atendendo aos requisitos de *big data*. Por fim, a camada de interação conecta os usuários à infraestrutura, centralizando o acesso e oferecendo mecanismos de controle de permissão e anonimização de dados.

Já GADDS, também voltada para ambientes colaborativos e distribuídos, prioriza a descentralização. Ela é composta por quatro camadas: (i) armazenamento de dados; (ii) armazenamento de metadados; (iii) sistema de versionamento; e (iv) interface de usuário.

O armazenamento de dados consiste em um repositório de armazenamento distribuído em nuvem, satisfazendo requisitos de *big data*, enquanto o armazenamento de metadados é uma rede *blockchain* que rastreia inserções e atualizações de metadados. A camada de versionamento interage com as duas anteriormente mencionadas ao manter um histórico de atualizações de metadados, bem como suas referências a objetos de dados. Por fim, a camada de interface consiste em uma aplicação *web* que centraliza o acesso de usuários ao sistema, abstraindo conceitos técnicos.

3.2. Instanciações no domínio de sismologia e comparação FAIR

As SRAs FLAMI e GADDS foram instanciadas no contexto de análises sismológicas, adotando estratégias distintas para garantir colaboração e disseminação de dados. Esta seção descreve essas instanciações e discute sua aderência aos princípios FAIR, analisando como suas camadas atendem aos critérios de encontrabilidade, acessibilidade, interoperabilidade e reusabilidade. As Figuras 1 e 2 ilustram as instanciações, cujas camadas são discutidas a seguir.

3.2.1. Instanciações das Camadas de Armazenamento de Dados

Os repositórios de armazenamento interno e externo da FLAMI foram implementados com HDFS e MongoDB, respectivamente. Como resultado, assegurou-se a recuperação de dados mesmo se excluídos de suas fontes originais, garantindo alta disponibilidade a longo prazo. Ademais, a adoção do formato padrão de arquivos SAC e Parquet impulsionou a reusabilidade. Entretanto, o acesso aos dados foi limitado apenas aos analistas responsáveis pela implementação, prejudicando a acessibilidade de FLAMI.

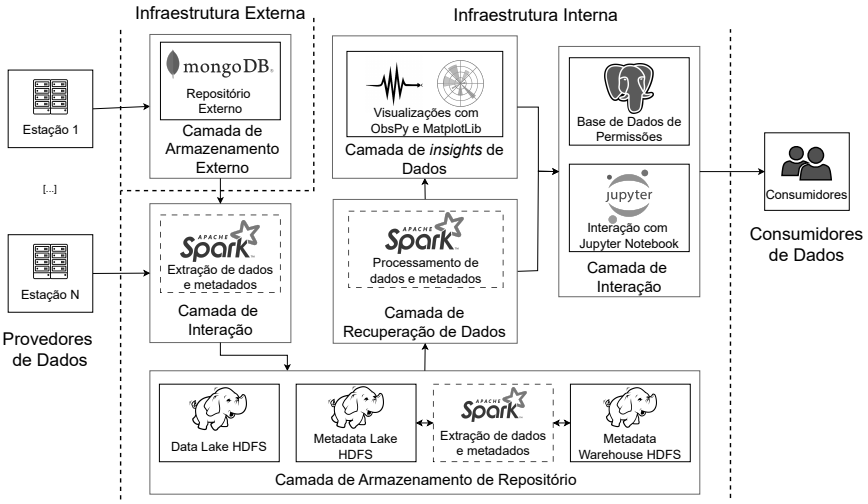

Figura 1. Instância da arquitetura FLAMI.

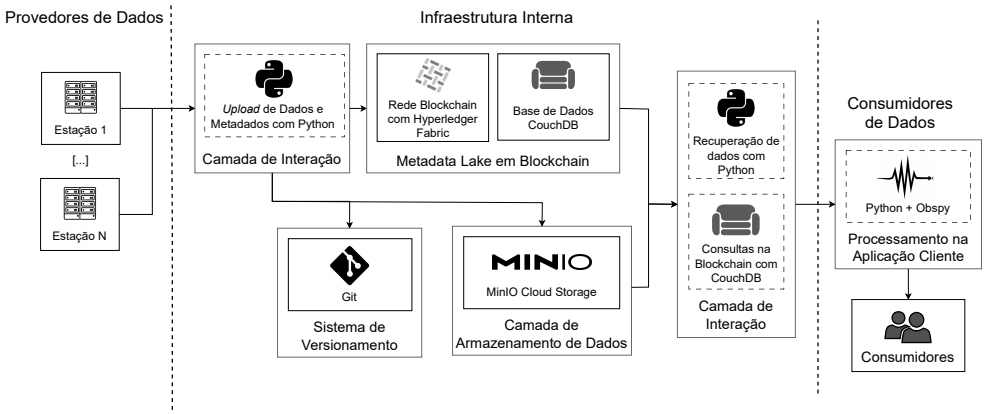

Figura 2. Instância da arquitetura GADDS.

O armazenamento de dados da GADDS foi implementado utilizando o MinIO como um sistema de *Cloud Storage* unificado, garantindo replicação e disponibilidade dos dados. Entretanto, GADDS não permite manter uma infraestrutura externa para o armazenamento. Essa SRA também não provê um formato padrão de armazenamento de dados, prejudicando sua reusabilidade. Além disso, por conta da natureza da camada de *blockchain*, somente membros participantes das organizações da rede conseguem acessar os dados, impactando negativamente na acessibilidade.

3.2.2. Instanciações das Camadas de Interação

A Camada de Interação da SRA FLAMI foi implementada oferecendo os recursos necessários para alimentar dados e metadados via *framework* Spark. A recuperação em consultas analíticas ocorre por meio da aplicação *web* Jupyter Notebook, permitindo que os usuários realizem análises nas linguagens de sua preferência, com controle de permissões utilizando uma base de dados com SGBD relacional PostgreSQL. A arquitetura suporta interfaces personalizáveis para múltiplos idiomas, incluindo APIs multilíngues,

ampliando sua acessibilidade. Além disso, a camada integra os processos de alimentação e recuperação, oferecendo suporte a consultas complexas e geração de *insights*, promovendo a encontrabilidade e interoperabilidade.

A camada de interação implementada para a GADDS usa *scripts* em Python para envio, consulta e recuperação de dados e metadados, com processamento realizado por meio das bibliotecas Pandas e ObsPy [OBSPY 2025]. Entretanto, a ausência de uma interface gráfica e multilíngue de GADDS compromete a interoperabilidade e a acessibilidade. Ainda assim, as tecnologias referenciadas para GADDS oferecem recursos eficazes de recuperação, favorecendo a encontrabilidade de dados e metadados.

3.2.3. Instanciação das Camadas de Armazenamento de Metadados

A camada de armazenamento de metadados da FLAMI é baseada no *Metadata Lake*, o qual foi implementado com HDFS, com transformações feitas via *framework* Spark, possibilitando o armazenamento no *Metadata Warehouse*. Os metadados referenciam os dados correspondentes em infraestrutura interna ou externa. Entretanto, a ausência de um identificador globalmente único na implementação prejudicou a encontrabilidade.

Na GADDS, os metadados foram armazenados em uma camada de *blockchain* implementada em Hyperledger Fabric, com núcleo em CouchDB. Apesar de conter identificadores únicos internos, a arquitetura não oferece suporte à criação de identificadores globalmente únicos, uma vez que a *blockchain* é acessível somente aos participantes da rede, o que limita a encontrabilidade. A natureza privada da camada de *blockchain* compromete tanto a publicização dos artefatos de pesquisa — afetando a acessibilidade — quanto a interoperabilidade da arquitetura, ao impor restrições quanto à utilização de repositórios externos.

4. Execução de uma consulta analítica

Nesta seção é feita a descrição da execução de uma mesma consulta analítica em cada uma das SRAs sob análise, FLAMI e GADDS. O objetivo consiste em mostrar os componentes acessados em cada SRA e como o resultado final é obtido. A consulta analítica é de um tipo de grande interesse dos usuários dessas SRAs, uma vez que requer a recuperação tanto de metadados quanto de dados de origem.

Para a definição da consulta, foi considerado o domínio de sismologia, composto por objetos de dados no formato *Seismic Analysis Code* (SAC), os quais descrevem as atividades sísmicas capturadas de estações do Nordeste do Brasil. Cada arquivo contém um vetor de pontos representando as medições sísmicas, acompanhadas de metadados como estação, canal e número de pontos de dados. Com base nesse conjunto, a consulta **Q** foi definida como "Recuperar, para cada estação, a amplitude relativa ao arquivo de maior tamanho, além do número de pontos e tamanho do arquivo". Essa consulta foi utilizada como base para a submissão da carga de trabalho sobre as SRAs. As Figuras 3 e 4 descrevem o *pipeline* da execução de **Q** considerando FLAMI e GADDS, respectivamente.

Na SRA FLAMI, o primeiro passo para executar **Q** consiste no acesso ao *Metadata Warehouse*, realizando as junções estrela entre as tabelas de fato e dimensões para obter as informações desejadas (número de pontos, tamanho e nome dos arquivos), além

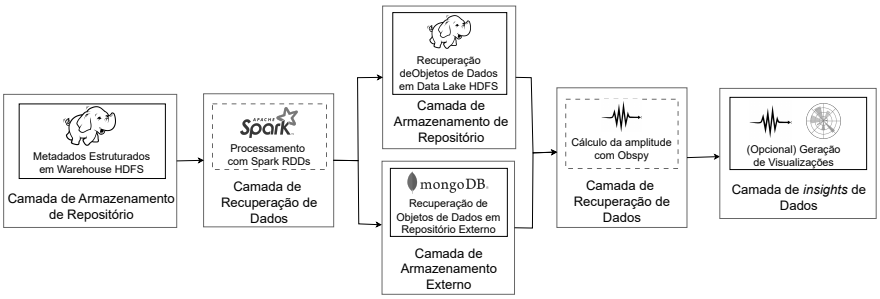

Figura 3. Pipeline de execução na SRA FLAMI.

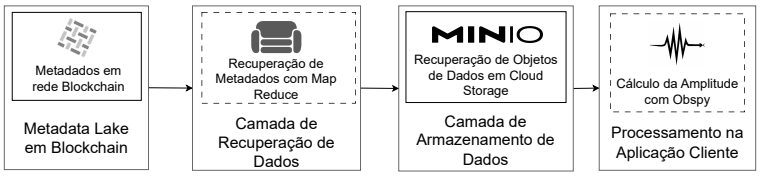

Figura 4. Pipeline de execução na SRA GADDS.

das informações necessárias para a conexão com os provedores de dados relativos a cada estação. Os dados obtidos são resultantes da agregação (função de maximização) sobre o atributo (fato) de tamanho do arquivo, agrupando pelo nome da estação correspondente. Como resultado, são recuperados os arquivos Parquet do *Data Lake* HDFS para os provedores internos e os arquivos SAC do MongoDB para os provedores externos. O processamento dos dados ocorre com as APIs de DataFrame e RDD do *framework* Spark em conjunto com a biblioteca Obspy para o tratamento dos dados contemplados. Após a obtenção dos resultados, os mesmos são escritos em um arquivo CSV e salvos no HDFS.

Na SRA GADDS, a aplicação cliente acessa o *world state* da *blockchain* usando a base de dados local em CouchDB para consultar os metadados, obtendo os mesmos dados anteriormente mencionadas no caso da SRA FLAMI. A consulta nos metadados é realizada com base em funções *map* e *reduce* nativas do CouchDB, como forma de reproduzir a agregação necessária. Com as informações de acesso aos dados de origem, os mesmos são recuperados dos *buckets* instanciados no MinIO. Contudo, devido ao armazenamento adotado pela GADDS, o processamento dos objetos de dados (arquivos SAC) e cálculo de suas amplitudes são realizados localmente com auxílio da biblioteca ObsPy. Por fim, os resultados são escritos em um arquivo CSV e salvos no MinIO.

5. Conclusão

Neste artigo, comparou-se duas SRAs para *big data* alinhadas aos Princípios FAIR: FLAMI e GADDS. FLAMI demonstra maior flexibilidade ao permitir o uso de infraestruturas internas e externas, equilibrando posse de dados e simplicidade de gestão, enquanto GADDS se destaca pelo maior controle dos metadados devido à *blockchain*. No contexto da Ciência Aberta, a adesão aos Princípios FAIR é essencial para incentivar o compartilhamento e fortalecer a colaboração científica. Mesmo sem todas as camadas implementadas, FLAMI apresentou maior conformidade com os princípios FAIR, consolidando-se como a SRA mais alinhada ao estado-da-arte. Como trabalhos futuros, estão previstas novas consultas analíticas e testes de desempenho para avaliação de tempos de resposta.

Referências

Angelov, S., Grefen, P., and Greefhorst, D. (2012). A framework for analysis and design of software reference architectures. *Inf. Softw. Technol.*, 54(4):417–431.

Bhowmik, S. (2017). *Cloud Computing*. Cambridge University Press, [S.l.].

Castro, J. P. C., De-Grandi, M. J., and Aguiar, C. D. (2025a). A systematic review of FAIR-compliant big data software reference architectures. *J. Inf. Data Manag.* Accepted for Publication.

Castro, J. P. C. et al. (2022a). FAIR Principles and Big Data: A software reference architecture for Open Science. In *Proc. ICEIS*, pages 27–38.

Castro, J. P. C. et al. (2022b). Open Science in the cloud: The CloudFAIR architecture for FAIR-compliant repositories. In *Proc. ADBIS*, pages 56–66.

Castro, J. P. C., Vasconcelos, F. X. G., Vargas-Solar, G., and Aguiar, C. D. (2025b). Building FAIR-compliant lakehouses with FLAMI. In *Proc. CAiSE*. Accepted for Publication.

Davoudian, A. and Liu, M. (2020). Big data systems: A software engineering perspective. *ACM Comput. Surv.*, 53(5):1–39.

Medeiros, C. B., Darboux, B. R., Sánchez, J. A., Tenkanen, H., Meneghetti, M. L., Shinwari, Z. K., Montoya, J. C., Smith, I., McCray, A. T., and Vermeir, K. (2020). *IAP input into the UNESCO Open Science Recommendation*. Available at https://www.interacademies.org/sites/default/files/2020-07/Open_Science_0.pdf.

OBSPY (2025). ObsPy. https://www.obspy.org/.

Vazquez, P. et al. (2022). Globally accessible distributed data sharing (GADDS): A decentralized FAIR platform to facilitate data sharing in the life sciences. *Bioinformatics*, 38:3812–3817.

Wilkinson, M. D. et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Sci. Data.*, 3(1):1–9.