ξ-DL: um Sistema de Gerência de *Data Lake* para Monitoramento de Dados da Saúde*

Lucas Tito[1], Cristina Motinha[1], Filipe Santiago[1]†, Kary Ocaña[2], Marcos Bedo[1], Daniel de Oliveira[1]

[1]Universidade Federal Fluminense (IC/UFF), Brasil

{lucastito, cmotinha, filipe_santiago, marcosbedo}@id.uff.br

danielcmo@ic.uff.br

[2]Laboratório Nacional de Computação Científica

karyann@lncc.br

Resumo. *Na última década, diversos domínios científicos vêm produzindo um grande volume de dados heterogêneos (i.e., estruturados e não-estruturados) e variantes ao longo do tempo. Apesar da popularidade, tecnologias como Data Warehouses têm se mostrado pouco adaptáveis a esses tipos de dados. Por outro lado, os Data Lakes se mostram flexíveis nesse cenário, uma vez que não necessitam de modelagem prévia (os dados são armazenados em seu formato bruto) e provem mecanismos de consulta. Apesar de existirem diversas soluções voltadas para Data Lakes (a maioria baseada no stack Hadoop), elas requerem determinada expertise em computação que nem todo cientista possui. Esse artigo apresenta o ξ-DL, um sistema de gerência de Data Lakes para dados científicos, que permite que cientistas sem conhecimento profundo em computação possam gerenciar seus Data Lakes. O ξ-DL foi avaliado por meio de um estudo de viabilidade com um dataset de COVID-19 no Brasil. A avaliação inicial com usuários do domínio mostrou que a abordagem é promissora.*

1. Introdução

Nos últimos anos, com o aumento crescente na produção de dados e o rápido desenvolvimento de técnicas de processamento massivo de dados, o valor dos dados foi identificado em diversos domínios, como a medicina e a biologia [Hey et al. 2009]. Entretanto, ainda existem barreiras para a disseminação de dados científicos entre os diversos pesquisadores, *e.g.*, formatos proprietários, heterogêneos, permissões, *etc*. Esse isolamento dos dados cria os chamados *silos de dados* [Nargesian et al. 2019]. O armazenamento não-integrado de dados pode fazer com que pesquisadores percam oportunidades de pesquisa e desenvolvimento. Tomemos como exemplo o cenário da área da saúde que utilizaremos consistentemente no restante do artigo. Nas últimas décadas, as comunidades da computação e da saúde têm despendido esforços para prover soluções computacionais na área da saúde, inclusive no Brasil [Silva et al. 2019]. Segundo [Shishvan et al. 2018], essas soluções variam desde o monitoramento de pacientes internados até a gerência de recursos. Um dos exemplos é o *Sistema Gerenciador de Ambiente Laboratorial* (GAL)[1] do SUS. O GAL tem como objetivo informatizar a rede laboratorial de saúde pública brasileira, *i.e.,* ele registra informações de amostras de origem humana e animal que

*O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior- Brasil (CAPES) - Código de Financiamento 001, CNPq e FAPERJ.

[1]https://gal.nacional.sus.gov.br/

possam ter sido expostas à doenças, possibilitando que profissionais da saúde possam consultar e extrair os dados para desempenhar vigilâncias epidemiológicas.

Apesar de serem importantes, os dados extraídos do GAL por si só podem não ser suficientes para que o gestor público tome uma decisão. No caso da recente pandemia de COVID-19 (causada pelo coronavírus SARS-CoV-2), os dados de exames laboratoriais são de suma importância para controle epidemiológico, porém devem ser enriquecidos com dados de análises filogenéticas (estudo da relação evolutiva entre grupos de organismos) realizadas [Li et al. 2020] para se descobrir se a cepa do SARS-CoV-2 difere de outras. Estudos dessa natureza já vem sendo realizados pelo LNCC, UFMG e UFRJ[2]. Em cenários como o anteriormente citado, é necessário que se disponibilize uma plataforma que possa impulsionar o compartilhamento de dados. Diversas tecnologias podem apoiar esse compartilhamento, *e.g., Data Warehouses* (DW) [Inmon 1996]. Entretanto, os DWs armazenam somente dados estruturados e sua modelagem pode ser complexa. Mais recentemente, o conceito de *Data Lake* [Nargesian et al. 2019] emergiu como uma abordagem para integração de dados estruturados, semi-estruturados e não-estruturados.

Um *Data Lake* é uma abordagem que consiste em um repositório de dados associado à uma *engine* para processamento de consultas e dados. A grande vantagem de um *Data Lake* é que o mesmo não necessita de uma modelagem prévia, e é capaz de armazenar dados em seu formato bruto, preservando o princípio de imutabilidade. Entretanto, para ser capaz de armazenar e consultar dados de diferentes formatos, o *Data Lake* deve possuir uma série de metadados de forma a facilitar a localização dos dados e sua análise *a posteriori*. Existem diversas soluções para *Data Lakes* no mercado, sendo as soluções do *stack* Hadoop as mais usadas. Entretanto, essas soluções apresentam diversos fatores limitadores quando tratamos de dados científicos. Muitos cientistas não possuem *expertise* em computação, e lidar com tecnologias como HDFS, Hadoop e Spark pode não ser trivial. Além disso, dados de proveniência [Freire et al. 2008] devem ser capturados, assim como questões de privacidade dos dados devem ser tratadas.

De forma a permitir que cientistas sem formação em computação possam se beneficiar de *Data Lakes*, esse artigo propõe o ξ-DL, um sistema que apoia a gerência de *Data Lakes* para dados científicos. Com o ξ-DL os cientistas podem: *(i)* importar dados de múltiplas fontes para o *Data Lake*, *(ii)* associar metadados de proveniência aos dados importados, e *(iii)* consultar e visualizar dados em diferentes formatos (*e.g.*, gráficos). O ξ-DL foi avaliado por meio de um estudo de viabilidade com *datasets* que contém dados dos casos de COVID-19 em diversos estados do Brasil [Mello et al. 2020], disponibilizados no Portal *COVID-19 Data Sharing/BR*[3] da FAPESP. O restante do artigo está organizado da seguinte forma. Na Seção 2 discutimos referencial teórico e trabalhos relacionados. Na Seção 3 apresentamos a abordagem proposta. Na Seção 4 discutimos a avaliação e, finalmente, na Seção 5 concluímos o artigo.

2. Referencial Teórico e Trabalhos Relacionados

Um *Data Lake* é um repositório centralizado que permite armazenar dados estruturados, semi-estruturados e não-estruturados em grande escala. O armazenamento é realizado no formato original dos dados [Nargesian et al. 2019]. O foco é armazenar todos os dados, sem perda, para posterior exploração. O fato de não obrigar o usuário a definir um *schema* para os dados é uma grande vantagem, uma vez que muitas iniciativas de DW não prosperam justamente por conta de dificuldades de modelagem, uma vez que DWs tendem a seguir a noção de um único

[2] `https://tinyurl.com/twbg2lf`
[3] `https://repositoriodatasharingfapesp.uspdigital.usp.br/handle/item/97`

esquema para todas as consultas analíticas a serem submetidas. Em diversos cenários, criar um único modelo de dados pode ser impraticável.

O principal desafio de um *Data Lake* é permitir a consulta sobre esses dados brutos que não possuem *schemas* associados (*i.e.*, *schema on read* ou *schemaless*). A operação de um *Data Lake* possui as seguintes etapas: (i) Ingestão de dados - permite importar qualquer quantidade de dados em tempo real de múltiplas fontes; (ii) Armazenamento dos dados no formato original - depois de recuperados, os dados precisam ser armazenados em um formato durável e facilmente acessível, sem transformações prévias; (iii) Processamento e Análise - nessa etapa, os dados são transformados de seu formato bruto para formatos que interessem ao usuário; e (iv) Exploração e visualização - a etapa final é a de conversão dos resultados da análise em um formato que facilite a visualização e a extração de informações.

Existem vários *frameworks* para implementação de um *Data Lake*, *e.g.*, Google Cloud Platform[4], o Athena da AWS[5], e o Apache Drill[6]. Essas abordagens oferecem o arcabouço necessário para a implementação de um *Data Lake*, porém, podem ser difíceis de usar por um usuário não especialista em computação. De forma similar ao ξ-DL, existem abordagens que focam em desenvolver sistemas de gerência de *Data Lakes*. [Chen et al. 2018] propõem uma nova estrutura para *Data Lakes* públicos para controlar e proteger a privacidade de dados do compartilhamento de dados. Assim como a abordagem de [Chen et al. 2018], o Kaiak [Maccioni and Torlone 2017] é um sistema com o foco na otimização de *pipelines* de preparação de dados em *Data Lakes*. O Kayak permite que os usuários especifiquem seus *pipelines* e consultem seus dados com restrições de qualidade.

3. Abordagem Proposta: ξ-DL

O ξ-DL se encontra entre os usuários e/ou aplicações e o sistema de arquivos distribuídos onde os dados brutos são armazenados. A arquitetura do ξ-DL é apresentada na Figura 1, e é composta de 4 camadas: (i) Fontes de Dados, (ii) Camada de Processamento, (iii) Camada de Dados, e (iv) Sistema de Arquivos Distribuído. As Fontes de Dados representam os dados que podem ser representados em múltiplos formatos (*e.g.*, BDs relacionais, arquivos CSV, *etc*) e que podem ser importados para o ambiente do *Data Lake*. A importação dos dados é realizada por meio do *Portal ξ-DL* da Camada de Processamento, que faz a interface com o usuário. O ξ-DL segue o conceito de coleção de dados (*dataset*). Essa estrutura de coleções é apresentada de maneira visual para o usuário por meio do *Portal ξ-DL*. Na Figura 2 é possível visualizar a tela que contém um *dataset* criado pelo usuário. O usuário pode importar tuplas em lote por meio do *upload* de arquivo *CSV*, verificar a estrutura atual do *dataset* (que é interpretada dinamicamente pelo ξ-DL), submeter consultas, visualizar os dados de toda a tabela e navegar pelo histórico de consultas já submetidas. É importante ressaltar que caso o usuário escolha um *dataset* e inclua novos atributos ao mesmo, a estrutura do *dataset* é dinamicamente adaptada para incluir novos atributos. A proveniência dos dados e o *schema* identificado são armazenados na Camada de Dados para posterior uso por parte do processador de consultas.

Os dados brutos importados sofrem uma modificação de formato para serem armazenados no Sistema de Arquivos Distribuídos. O sistema de arquivos usado na versão atual é o Amazon S3, cuja unidade de armazenamento básica é *objeto*, que possui um identificador e é armazenado em *buckets*. Os dados tabulares são convertidos para o formato *Parquet* já que

[4]`https://cloud.google.com/solutions/build-a-data-lake-on-gcp?hl=pt-br`
[5]`https://aws.amazon.com/pt/athena/`
[6]`https://drill.apache.org`

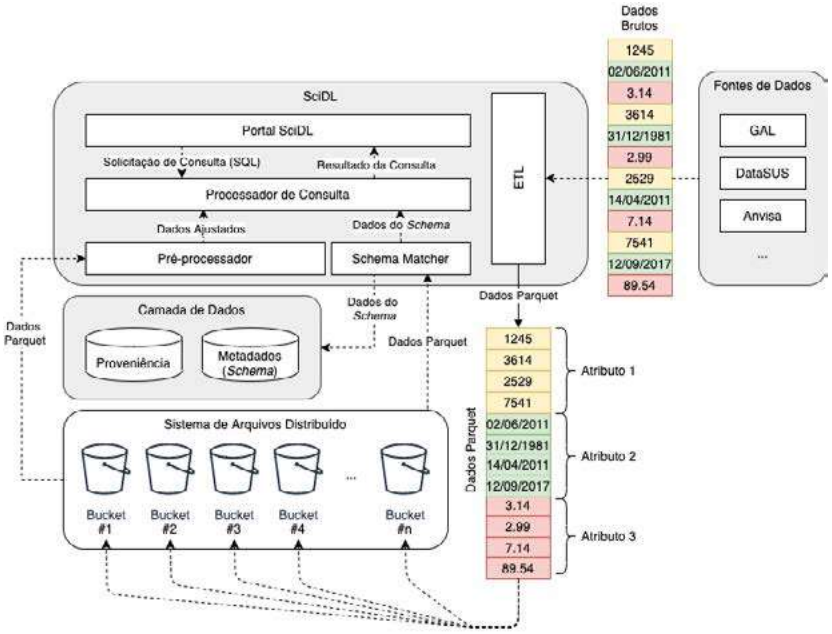

Figura 1: Arquitetura do ξ-DL.

o mesmo oferece estruturas complexas de dados aninhadas e foi desenvolvido para apoiar esquemas de compressão e codificação eficientes. Além disso, esse formato permite reduzir os custos de armazenamento de arquivos de dados e faz com que consultas sejam mais eficientes. O *Processador de consultas* é o componente responsável por receber consultas do usuário e processá-la sobre os dados brutos armazenados. Em sua versão atual, o ξ-DL utiliza o AWS Athena para processamento das consultas. O Athena é integrado ao AWS *Glue Data Catalog*, que funciona como um repositório de metadados unificado. Os dados da estrutura do *dataset* são enviados ao *Glue* pelo *Schema Matcher*. Dependendo do tipo de consulta, os dados podem sofrer ajustes para processamento pelo Athena. O componente *Pré-Processador* é responsável por ajustar o formato dos dados antes do processamento da consulta (caso necessário).

4. Estudo de Viabilidade

Para avaliar o ξ-DL foi utilizado um *dataset* do portal *COVID-19 Data Sharing/BR*. Foi selecionado o *dataset* do Hospital Sírio Libanês como estudo de caso, que contém: (i) dados anonimizados sobre pacientes que fizeram teste para o COVID-19; (ii) os resultados dos exames laboratoriais; e (iii) dados do atendimento de um paciente e o diagnóstico, quando existir. O *dataset* contém 2.732 tuplas e 4,5 MB. A Figura 2 apresenta uma sumarização do uso do ξ-DL para o estudo de caso. Na Figura 2(a) o *dataset* é importado e o usuário deve informar a proveniência dos dados, de forma a se realizar o rastreio da origem dos dados. Uma vez carregado o *dataset*, a estrutura dos dados é apresentada na Figura 2(b), onde cada atributo e valores associados podem ser visualizados. Na Figura 2(c) pode ser visualizada a interface para geração de gráficos sobre o *dataset*. No exemplo apresentado, dois gráficos foram gerados: (i) Número de Casos por sexo do paciente e (ii) Número de casos por UF. É importante ressaltar que o usuário não precisa conhecer a linguagem de consulta, já que o mesmo seleciona os atributos a serem apresentados nos eixos e a função de agregação a ser utilizada via Portal ξ-DL.

O *deploy* do ξ-DL foi realizado em uma máquina virtual do tipo *t2.micro* Amazon AWS que possui 2 vCPUS e apenas 1GB de RAM. Nessa configuração, o processo de carga, conversão e armazenamento do *dataset* consumiu 6,7 minutos. Apesar de ser um tempo não ne-

Figura 2: Interface do ξ-DL

gligenciável para um *dataset* relativamente pequeno ($<$ 5MB), acreditamos que as restrições na configuração do *hardware* influenciaram no tempo de processamento. Em relação ao tempo de geração dos gráficos, cada um foi gerado em 2,5 minutos, o que é um tempo aceitável. De forma a avaliar as funcionalidades do ξ-DL utilizamos o TAM *(Technology Acceptance Model)* [Davis 1989]. A ideia principal do TAM é avaliar a receptividade/comportamento de um usuário no que se refere à facilidade e utilidade da tecnologia/ferramenta que está sendo proposta. A utilidade refere-se ao quanto o usuário acredita que a abordagem proposta o auxiliará em suas tarefas e a facilidade se refere ao quão fácil/simples será utilizar tal abordagem. A avaliação contou com 6 usuários especialistas e ocorreu em 17 de junho de 2020 via Google Meet. As avaliações para as questões do TAM são apresentadas na Tabela 1. É possível perceber que os usuários tiveram uma boa percepção do ξ-DL, apesar de algumas correções de erro terem sido apontadas. Acreditamos que o ξ-DL se mostrou promissor no que tange a gerência de dados científicos, em especial nesse cenário de dados da área de saúde.

5. Conclusões

Compartilhar dados científicos pode não ser uma tarefa trivial. Dados científicos são naturalmente heterogêneos. Essa gerência ainda se torna mais complexa à medida que consultas devem ser realizadas sobre esses dados e se torna complicado definir um *schema*. Os *Data Lakes* têm mostrado ser uma boa escolha para compartilhamento e análise de grandes volumes de dados, já que permitem armazenar dados estruturados, semi-estruturados e não-estruturados. Porém, pode não ser simples para um usuário não especialista em computação gerenciar todas as tecnologias envolvidas em um *Data Lake*. Nesse artigo apresentamos o ξ-DL, um sistema gerência de *Data Lakes*, que apoia o usuário desde o *upload* de dados até a geração de visualizações. Avaliamos o sistema com dados reais de pacientes de COVID-19 e a avaliação experimental mostrou que a maioria dos avaliadores acredita que o uso do ξ-DL agrega valor ao seu trabalho, e possivelmente melhorará o compartilhamento e análise de dados científicos. Trabalhos futuros incluem uma avaliação experimental em uma escala maior.

Referências

Chen, Y., Chen, H., and Huang, P. (2018). Enhancing the data privacy for public data lakes. In *2018 IEEE ICASI*, pages 1065–1068.

Tabela 1: Resultado da Avaliação do ξ-DL com o TAM

Id	Pergunta	Muito Baixo	Baixo	Médio	Alto	Muito Alto
u1	Escolha o grau da aplicabilidade da ξ-DL no seu trabalho.	0%	0%	33,33%	66,67%	0%
u2	Escolha o grau do desempenho que seria obtido com a aplicação da ξ-DL no seu trabalho.	0%	33,33%	16,67%	50%	0%
u3	Escolha o grau da qualidade das informações exibidas pela ξ-DL.	0%	0%	33,33%	66,67%	0%
u4	Escolha o grau da qualidade dos resultados que você poderia obter ao usar ξ-DL.	0%	0%	16,67%	83,33%	0%
f1	Escolha o grau da facilidade de uso da ξ-DL.	0%	0%	50%	50%	0%
f2	Escolha o grau da facilidade de aprendizado com pouco ou nenhum treinamento da ξ-DL.	0%	16,67%	0%	83,33%	0%
f3	Escolha o grau da facilidade de lembrar de como se usa o Scξ-DLiDL.	0%	16,67%	0%	66,67%	16,67%
f4	Escolha o grau da facilidade de identificar quando ocorrem erros no ξ-DL.	16,67%	50%	16,67%	16,67%	0%

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Q.*, 13(3):319–340.

Freire, J., Koop, D., Santos, E., and Silva, C. T. (2008). Provenance for computational tasks: A survey. *Comput. Sci. Eng.*, 10(3):11–21.

Hey, T., Tansley, S., and Tolle, K., editors (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, Redmond, Washington.

Inmon, W. H. (1996). The data warehouse and data mining. *CACM*, 39(11):49–50.

Li, Y., Liu, B., Cui, J., Wang, Z., Shen, Y., Xu, Y., and Yao, K. (2020). Similarities and evolutionary relationships of COVID-19 and related viruses. *CoRR*, abs/2003.05580.

Maccioni, A. and Torlone, R. (2017). Crossing the finish line faster when paddling the data lake with kayak. *PVLDB*, 10(12):1853–1856.

Mello, L. E., Suman, A., and et al. (2020). Opening Brazilian COVID-19 patient data to support world research on pandemics.

Nargesian, F., Zhu, E., Miller, R. J., Pu, K. Q., and Arocena, P. C. (2019). Data lake management: Challenges and opportunities. *Proc. VLDB Endow.*, 12(12):1986–1989.

Shishvan, O. R., Zois, D., and Soyata, T. (2018). Machine intelligence in healthcare and medical cyber physical systems: A survey. *IEEE Access*, 6:46419–46494.

Silva, A. B., Guedes, A., Síndico, S., Vieira, E., and de Andrade Filha, I. (2019). Registro eletrônico de saúde em hospital de alta complexidade: um relato sobre o processo de implementação na perspectiva da telessaúde. *Ciência & Saúde Coletiva*, 24:1133–1142.