Uma proposta de data lake para pesquisa em saúde a partir de data pools multicêntricos interoperáveis

Daniel M. Lima[ab], Ramon A. Moreno[a*], Fabio A. Pires[a], Marco A. Gutierrez[a]

[a]Laboratorio de Informatica Biomedica, Instituto do Coracao,
Hospital das Clinicas HCFMUSP, Faculdade de Medicina,
Universidade de Sao Paulo, Sao Paulo, SP.

[b]Instituto de Ciencias Matematicas e de Computacao (ICMC-USP),
Universidade de Sao Paulo, Sao Carlos, SP.

Abstract. *With the high demand in data science, the organization and preparation of databases became critical activities, consuming more than 80% of the project effort. In the medical domain, many hospitals already use a myriad of technologies and information systems for medical records and images, but they do not always adopt standards of uniform and interoperable data, and they seldom adopt analytics-oriented tools (data lakes and warehouses). In this article we propose the data pool, an intermediate data model to ease the organization of data lakes for health research. The data pool was implemented and adopted in real medical research, supporting computational learning workflows.*

Resumo. *Com a alta demanda em ciência de dados, a organização e preparo de bases de dados se tornaram atividades críticas, consumindo mais de 80% do esforço do projeto. No domínio de assistência ao paciente, muitos hospitais já utilizam uma miríade de tecnologias e sistemas informatizados para prontuários e imagens, mas nem sempre adotam padrões de dados uniformes e interoperáveis, e raramente adotam ferramentas voltadas à análise (*data lakes *e* warehouses*). Neste artigo é proposto o* data pool*, um modelo de dados intermediário para facilitar a organização de* data lakes *voltados à pesquisa em saúde. O* data pool *foi implementado e adotado em um ciclo completo de pesquisa médica real, dando suporte a fluxos de aprendizagem computacional.*

1. Introdução

A ciência de dados está em alta, abrindo várias iniciativas de cooperação de dados, metodologias e ferramentas. Em projetos dessa natureza as atividades mais cruciais são a organização, limpeza e manutenção das bases de dados [Segaran and Hammerbacher 2009]. Estas tarefas compõem a engenharia de dados e chegam a consumir mais de 80% do esforço do projeto[1]. Tal área é contemplada com ferramentas e métodos próprios, como os sistemas gerenciadores de bancos de dados (SGBD) e a modelagem de dados.

No domínio de assitência ao paciente, o gerenciamento de dados dentro de um hospital é voltado aos pacientes e acontece em 3 fases principais: (1) aplicações de tempo

Projeto apoiado pela Fundação Zerbini e FoxConn (AIMED-CATI 030/2007 FOXCONN-001/2019).

*Autor correspondente. Email: ramon.moreno@hc.fm.usp.br

[1]Nem Sempre se vê Mágica no Absurdo: Engenharia de Dados e Ciência de Dados, Prof. Altigran (UFAM), https://www.youtube.com/watch?v=N43528a23xo, acessado em 2021-06-25.

crítico, que envolvem a comunicação, processamento e armazenamento de dados na rede de equipamentos e sistemas hospitalares – *e.g.* protocolos de rede e formatos de dados médicos, arcabouços HL7 e OpenEHR [Benson 2012], sistemas *on-line transaction processing* (OLTP) e *picture archiving and communication systems* (PACS) [Furuie et al. 2003, de Azevedo-Marques and Salomão 2009]; (2) aplicações de acompanhamento contínuo do paciente – prontuários eletrônicos de pacientes [Canêo and Rondina 2014] com consultas, exames, diagnósticos, internações e outros eventos ao longo da vida; e (3) aplicações voltadas à gestão e pesquisa, que visam obter um apanhado geral das atividades, doenças comuns e tratamentos mais eficazes, historicamente – usando sistemas *on-line analytical processing* (OLAP) e mineração de dados [de Amo 2004]. Segundo o OCEBM[2] e [DiCenso et al. 2009], a pesquisa médica de alto nível envolve estudos multicêntricos. Em estudos assim, os dados estão presentes em múltiplas instituições e são consolidados em infraestruturas baseadas em *data lakes* (DL) e *data warehouses* (DW) [Miller 2018].

Contudo, tais infraestruturas não são amplamente utilizadas em instituições individuais [Tito et al. 2020], mesmo no caso de bases multicêntricas como a COVID-19 Data Sharing/BR [FAPESP 2020]. Nessa base os dados foram exportados como tabelas textuais desnormalizadas e seguindo padrões exclusivos de cada instituição de saúde. Objetivando motivar a adoção de padrões de dados orientados à pesquisa médica multicêntrica, neste artigo é apresentada uma arquitetura mínima, nomeada como *data pool* (*i.e.* uma piscina ou tanque de dados). Assim, um DL próprio à pesquisa médica com dados de prontuário e PACS pode ser construído através da integração de múltiplos *data pools*.

2. Trabalhos relacionados

De maneira geral, as abordagens para integrar dados podem ser organizadas em três fases, segundo [Miller 2018]. Inicialmente, os sistemas objetivavam a construção de DWs rigidamente estruturados através de fluxos *extract-transform-load* (ETL) programados, *e.g.* [Furuie et al. 2003]. A segunda fase foi a organização de DLs onde os dados são carregados brutos de cada instituição (zona), organizados de forma livre, e um processo *extract-load-transform* (ELT) mapeia os esquemas de dados entre si. Um dos projetos que auxilia nesse sentido é o OMOP *common data model* (CDM), um modelo comum de dados para pesquisas observacionais que vem sendo continuamente melhorado para trabalhar com grandes volumes de dados, dados genômicos e imagens médicas [Kang et al. 2021]. As ferramentas OMOP traçam um caminho desde a captura de dados brutos, organização em DLs e então em um DW padronizado e interoperável [Voss et al. 2015]. Tais ferramentas são utilizadas por parcerias multicêntricas como a National COVID Cohort Collaborative (N3C)[3]. A fase mais recente traz algoritmos para mapeamento e vínculo automático de esquemas usando metadados e dados, *e.g.* [Tito et al. 2020] que executou uma consolidação de dados de COVID-19 de múltiplos centros. Contudo, enquanto a modelagem de um DW pode ser complexa e demorada, a análise automática é rápida mas sem conhecimento de quais dados críticos podem ter sido omitidos nas fontes do DL. A proposta do *data pool* sugere uma organização mínima para facilitar o projeto de algoritmos da terceira fase, balanceando assim a facilidade de integração do DL sem perder o conhecimento do projeto das aplicações, e com atenção para a consistência e eficiência durante o processamento.

[2]OCEBM Levels of Evidence, https://www.cebm.ox.ac.uk/resources/levels-of-evidence/oxford-centre-for-evidence-based-medicine-levels-of-evidence-march-2009, acessado em 2021-06-28.

[3]National COVID Cohort Collaborative, https://ncats.nih.gov/n3c, acessado em 2021-07-18.

3. Metodologia

A metodologia proposta pode ser organizada em três etapas: (1) ingestão dos dados no *data pool*; (2) seleção e preparo dos dados para as aplicações; (3) implantação e avaliação das aplicações por casos de uso. A pipeline completa é ilustrada na Figura 1.

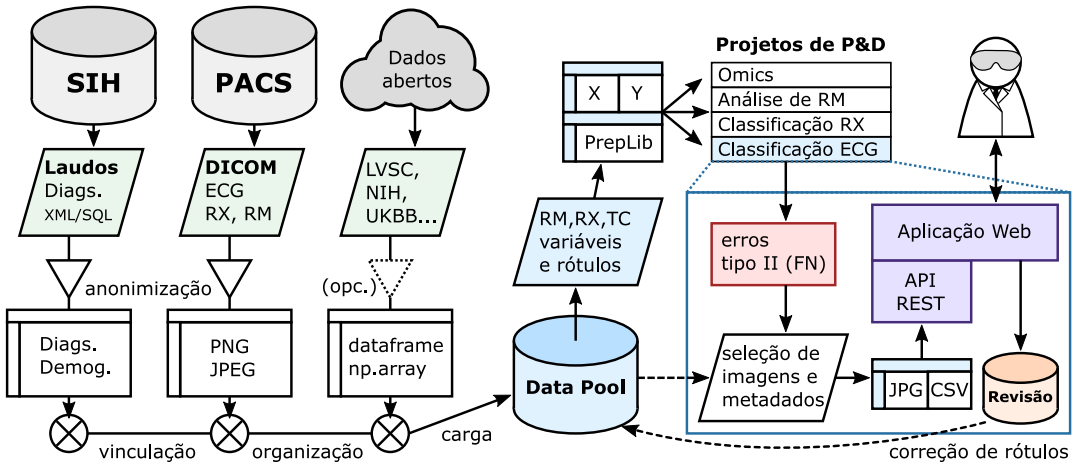

Figura 1. Pipeline completa desde a ingestão dos dados para o datapool, seleção e preparo de dados para um modelo, e revisão dos resultados do modelo.

3.1. Ingestão e armazenamento de dados

O processo de ingestão de dados se inicia nas fontes. As fontes de dados usuais são os sistemas de prontuário eletrônico ou de informações hospitalares (PEP ou SIH) com informações dos pacientes, exames e laudos; os sistemas de arquivamento de imagens (PACS); e bases colaborativas disponibilizadas pela Internet (Dados abertos). Na pipeline da Figura 1 anterior ao *data pool*, a ingestão dos dados ocorre em 5 camadas, de cima para baixo: (1) fontes de dados clínicos intra e inter-hospitalares, que são extraídos de cada fonte por protocolos específicos (*e.g.* consultas SQL, requisições HTTP e web *scrapping*), (2) os dados são transformados em estruturas de dados básicas (vetores, matrizes, árvores, grafos, etc.), (3) processos de tratamento e anonimização, (4) formatação dos dados binários, (5) consolidação de dados e carga no SGBD do *data pool*.

3.2. Modelagem relacional com metadados

O modelo do *data pool* é baseado em uma implementação típica de bancos de dados estruturados com objetos complexos: os dados complexos são "blobs" (*binary large objects* – arquivos DICOM, HDF, Parquet, etc.) armazenados em um sistema de arquivos indexado pelo caminho em uma árvore-B; e um SGBD relacional armazena os metadados de cada blob, permitindo consultar rapidamente quais blobs se aplicam a um estudo particular. A primeira versão segue a estrutura básica dos objetos no padrão DICOM [Mildenberger et al. 2002], cujo esquema está na Figura 2.

Pacientes vão ao hospital fazer exames, que são nomeados no DICOM como *estudos*. Cada estudo pode conter múltiplas sessões de captura de dados, que são armazenados como *séries* de *imagens*. Cada estudo é avaliado pelo técnico ou especialista, sendo marcados com *laudos* diagnósticos correspondentes aos achados identificados, de uma lista definida pelos médicos, e preenchida na tabela *diag*. Este esquema foi projetado de

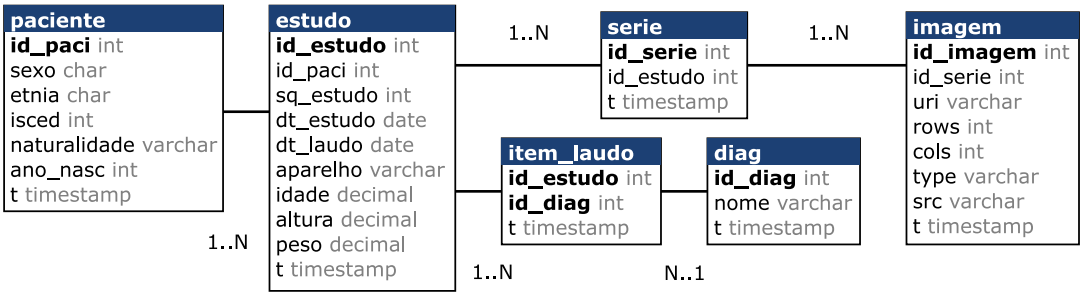

Figura 2. Modelo físico: um paciente tem estudos, compostos por séries de imagens (blobs) e que são opcionalmente vinculados a laudos diagnósticos.

acordo com as demandas dos projetos da instituição, voltados principalmente à análise de imagens médicas, contudo pode ser inadequado para outras instituições e portanto deve ser estendido com entidades e campos necessários a cada projeto. Após a carga, o sistema do data pool provê uma interface para consulta de imagens, medidas e diagnósticos, que são equivalentes às variáveis (X/Y) de uma análise estatística. Utilizando uma biblioteca padronizada de pré-processamento (PrepLib na Figura 1) é possível garantir a repetibilidade do preparo de dados para os vários projetos de pesquisa em múltiplas instituições.

3.3. Preparo e exportação de dados

Ao desenhar um estudo clínico, um pesquisador delimita quais coortes (grupos de pacientes) e tipos de exames, imagens e equipamentos serão inclusos no estudo. Este projeto implicitamente define predicados de consultas SQL sobre os metadados do SGBD. Ferramentas do OMOP [Voss et al. 2015] também seguem esta linha, apresentando uma interface gráfica para definição das consultas, que são então compiladas para o dialeto nativo do SGBD (*e.g.* MSSQL, Oracle ou Postgres). O SGBD do *data pool* executa a seleção, cujos metadados retornados por uma consulta de coorte contém as informações da estrutura do dado (tipo, formato, codificação, tamanho), variáveis demográficas dos pacientes (idade, sexo, etnia, etc.), medidas laboratoriais ou laudos diagnósticos (variáveis explicadas), e o caminho do exame (imagem.uri) para acesso ao blob via diretório de rede compartilhado. A consulta pode ser acessada por ferramentas como pandas [RRID:SCR_018214] e dplyr [RRID:SCR_016708] via SQL e exportadas como XML/JSON/CSV por uma API.

3.4. Configuração do sistema

O SGBD, aplicações web e processos de extração foram executados em uma máquina virtual com 2 núcleos de um Xeon X5650, 2GB de RAM e 100GB de disco. Os processos de treinamento e execução de modelos estatísticos foram executados em um HPC Foxconn M100-NIH (2x Xeon SKL Gold, 1TB de RAM e 8x V100-SXM2-16GB divididos em 14 nós), cujo armazenamento é alocado em um grupo RAID único de 2PB e os nós de trabalho são conectados por um barramento InfiniBand. Durante o preprocessamento, os dados são organizados em arquivos binários (*e.g.* HDF) e carregados por páginas de 1 a 8GB na memória RAM do nó que executa o treinamento dos modelos.

4. Resultados

4.1. Aplicações em casos de uso reais

Utilizou-se o *data pool* na extração e preprocessamento de imagens anonimizadas de um PACS, organizando-as em formato tensorial para processamento em frameworks de

aprendizagem profunda (caixa em destaque na Figura 1). Em um dos casos de teste, o *data pool* deu suporte à organização, consulta e pré-processamento de aproximadamente 200 mil imagens de eletrocardiografia (ECG) de 100 mil pacientes. Após a primeira rodada de ajuste e validação, o *data pool* recebeu mais imagens e devolveu ao modelo classificador as imagens correspondentes aos erros tipo II (falsos negativos). Estas imagens com predições erradas foram revisadas por cinco médicos em uma interface web (Figura 3). Neste artigo não é avaliado o resultado da aprendizagem, mas sim o objetivo do datapool, que é executar seleções exatas de imagens e metadados de acordo com os predicados SQL, e diminuiu o tempo de preparo dos dados de 30 para 4h por iteração do caso de uso "ECG" – nas primeiras iterações do projeto ECG (4 meses), as imagens eram preparadas sob demanda diretamente da fonte DICOM. Após o *data pool*, as imagens foram preparadas em lotes e armazenadas em formatos eficientes para utilização durante os 20+ meses subsequentes.

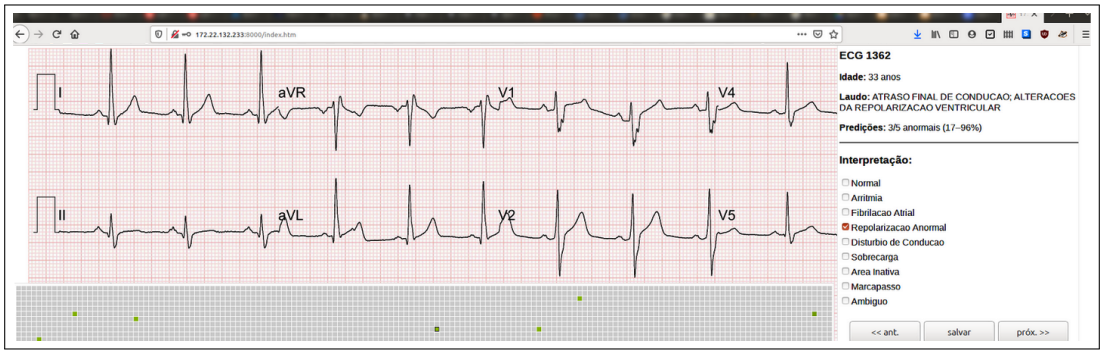

Figura 3. Tela de revisão de laudos em imagens anonimizadas. O médico olha o traço do sinal, revisa a previsão da rede e marca a interpretação correta.

5. Discussão e Conclusões

Neste trabalho foi apresentado o modelo de *data pool* – um sistema simples de organização de bases de dados voltado à pesquisa em saúde, cuja implementação foi utilizada para suportar aplicações de aprendizagem profunda voltadas à uma equipe médica de eletrocardiografia, completando todas as etapas de mineração de dados [de Amo 2004] em um contexto médico. Este modelo foi desenvolvido usando modelagem relacional e linguagem SQL–pois ambas têm amplo suporte de ferramentas e fornecedores há décadas– e pode ser facilmente estendido para mais aplicações e tipos de dados complexos. Uma grande vantagem de organizar os dados internamente em *data pools* é que o compartilhamento e adaptação a padrões de outras instituições é extremamente facilitado, e todo o arcabouço de indexação e mineração de dados complexos existentes nos SGBDs é automaticamente disponibilizado para a etapa de preparação de dados, como índices colunares [Larson et al. 2011], consultas por similaridade [Traina Jr et al. 2019], detecção de anomalias/eventos raros [Cabral and Cordeiro 2020] e tratamento de valores faltantes [Rodrigues et al. 2020].

Como trabalhos futuros, a implementação de *data pool* apresentada poderá ser estendida para abranger os demais projetos de pesquisa da instituição e readequada para fácil conexão com padrões internacionais, *e.g.* adicionando campos do OMOP CDM/Conditions à tabela *diag* e vinculando os diagnósticos da instituição a um vocabulário padrão CID-10 ou SNOMED. A adoção de modelos semelhantes por múltiplas instituições permite dividir o trabalho de estruturação de dados entre as instituições, facilitando a interoperabilidade de dados e a execução de estudos multicêntricos, *e.g.* como a N3C.

Referências

Benson, T. (2012). *Principles of health interoperability HL7 and SNOMED*. Springer Science & Business Media.

Cabral, E. F. and Cordeiro, R. L. (2020). Fast and scalable outlier detection with sorted hypercubes. In *Proc. 29th ACM CIKM*, pages 95–104.

Canêo, P. K. and Rondina, J. M. (2014). Prontuário eletrônico do paciente: conhecendo as experiências de sua implantação. *JHI*, 6(2).

de Amo, S. (2004). Técnicas de mineração de dados. *JAI*.

de Azevedo-Marques, P. M. and Salomão, S. C. (2009). Pacs: sistemas de arquivamento e distribuição de imagens. *Rev. bras. fis. med.*, 3(1):131–139.

DiCenso, A., Bayley, L., and Haynes, R. B. (2009). Accessing pre-appraised evidence: fine-tuning the 5s model into a 6s model. *Evidence-Based Nursing*, 12(4):99–101.

FAPESP (2020). FAPESP COVID-19 Data Sharing/BR. `https://repositoriodatasharingfapesp.uspdigital.usp.br/`.

Furuie, S. S., Gutierrez, M. A., Figueiredo, J., Tachinardi, U., Rebelo, M., Bertozzo, N., Moreno, R., Motta, G., Nardon, F., and Oliveira, P. (2003). Prontuário eletrônico de pacientes: integrando informações clínicas e imagens médicas. *Rev. bras. eng. biomed*, pages 125–137.

Kang, B., Yoon, J., Kim, H. Y., Jo, S. J., Lee, Y., and Kam, H. J. (2021). Deep-learning-based automated terminology mapping in omop-cdm. *JAMIA*. [ocab030].

Larson, P.-Å., Clinciu, C., Hanson, E. N., Oks, A., Price, S. L., Rangarajan, S., Surna, A., and Zhou, Q. (2011). Sql server column store indexes. In *Proc. ACM SIGMOD Conf. MOD*, pages 1177–1184.

Mildenberger, P., Eichelberg, M., and Martin, E. (2002). Introduction to the dicom standard. *European radiology*, 12(4):920–927.

Miller, R. J. (2018). Open data integration. *Proc. VLDB Endow.*, 11(12):21302139.

Rodrigues, L. S., Cazzolato, M. T., Traina, A. J. M., and Traina, C. (2020). Taking advantage of highly-correlated attributes in similarity queries with missing values. In *Lecture Notes in Computer Science*, volume 12440, pages 168–176. Springer.

Segaran, T. and Hammerbacher, J. (2009). *Beautiful data: the stories behind elegant data solutions*. O'Reilly Media, Inc.

Tito, L., Motinha, C., Santiago, F., Ocaña, K., Bedo, M., and de Oliveira, D. (2020). Xi-dl: um sistema de gerência de data lake para monitoramento de dados da saúde. In *Anais do XXXV SBBD*, pages 151–156, Porto Alegre, RS, Brasil. SBC.

Traina Jr, C., Moriyama, A., Rocha, G., Cordeiro, R., Ciferri, C. D., and Traina, A. (2019). The similarql framework: similarity queries in plain sql. In *Proc. 34th ACM/SIGAPP SAC*, pages 468–471.

Voss, E. A., Makadia, R., Matcho, A., Ma, Q., Knoll, C., Schuemie, M., DeFalco, F. J., Londhe, A., Zhu, V., and Ryan, P. B. (2015). Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. *JAMIA*, 22(3):553–564.