PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA
UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

MESTRADO

MÉTODO PARA ESTRUTURAR DADOS BRUTOS DE ANÁLISE DE CROMATOGRAFIA

Luiz Eduardo Davila de Paiva

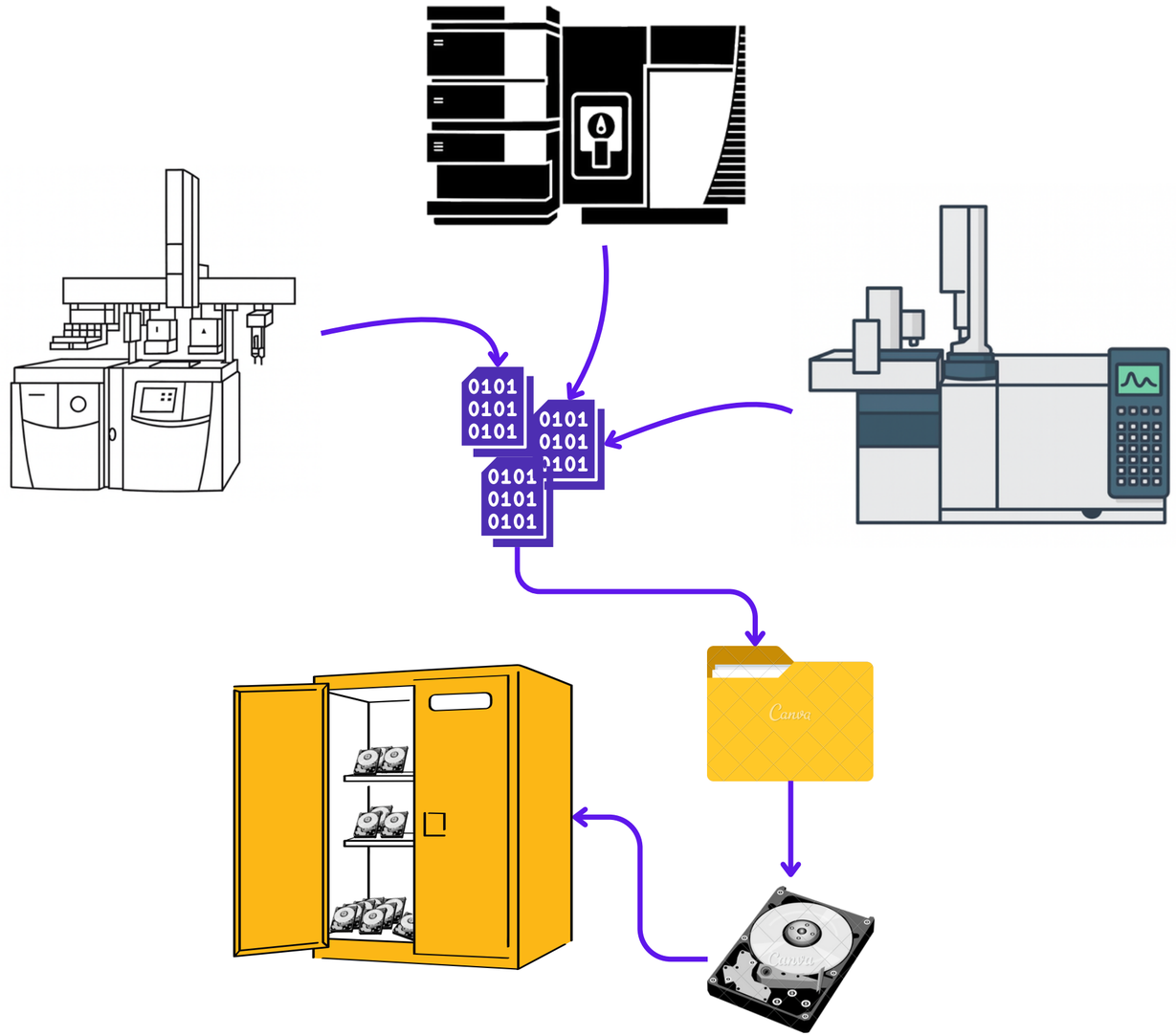
Prof. Sergio Manuel Serra da Cruz
Profa. Giseli Rabello Lopes

INTRODUÇÃO

As técnicas de cromatografia geram enormes volumes de dados analíticos, compostos por milhares de pontos que relacionam tempo de retenção, intensidade e massa. A dependência de arquivos brutos e softwares proprietários dificulta a reprodutibilidade e a reanálise histórica.

[LAI et al., 2024]

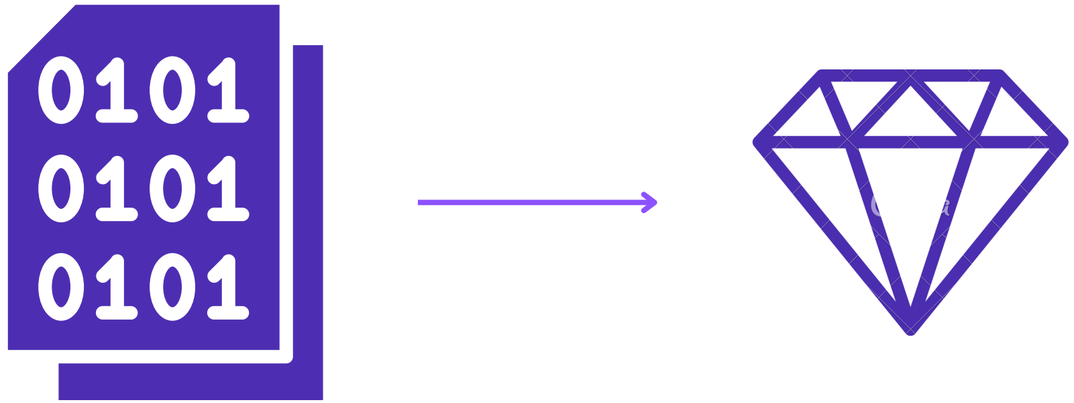
INTRODUÇÃO



Em muitos laboratórios, os resultados permanecem espalhados em arquivos brutos e planilhas, sem padronização ou integração com sistemas laboratoriais. Isso prejudica a rastreabilidade, dificulta consultas históricas e limita o uso de métodos de ciência de dados.

[MARDAL et al., 2023]

INTRODUÇÃO

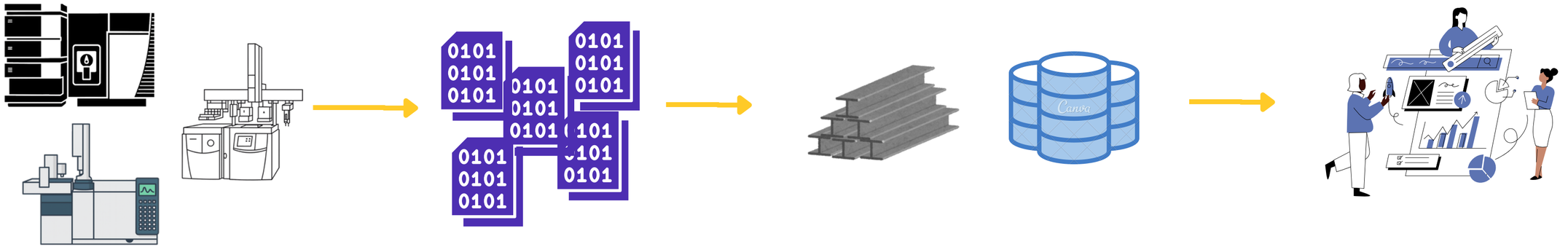


A Engenharia de Dados oferece processos ETL capazes de extrair, transformar e padronizar dados cromatográficos de diferentes instrumentos. Essa integração reduz redundâncias, garante integridade e prepara os dados para análises avançadas em bancos relacionais e plataformas analíticas.

KIMBALL & ROSS, 2013; SANTOS, 2019.

OBJETIVO

Equipamentos | **Dados** | **Engenharia de dados** | **Ciência de Dados**

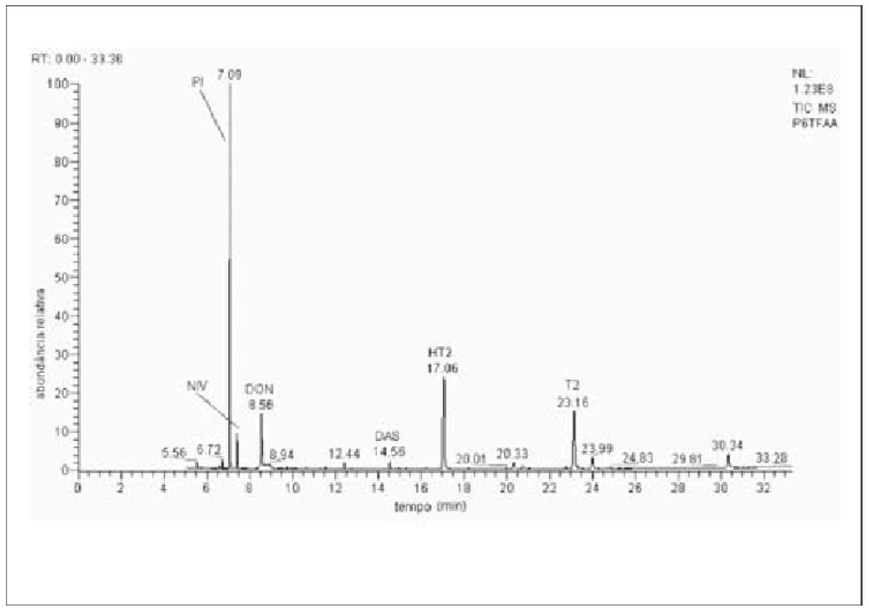


Um Data Lake científico centraliza dados um repositório único e escalável. Com governança adequada, evita-se a formação de um Data Swamp e torna-se possível aplicar ferramentas avançadas, como comparação automatizada de cromatogramas e técnicas de similaridade como DTW.

INMON, 2016; NEVALA, 2018; LAWAL, 2024.

REFERENCIAL TEÓRICO

Cromatografia e Espectrometria de Massas

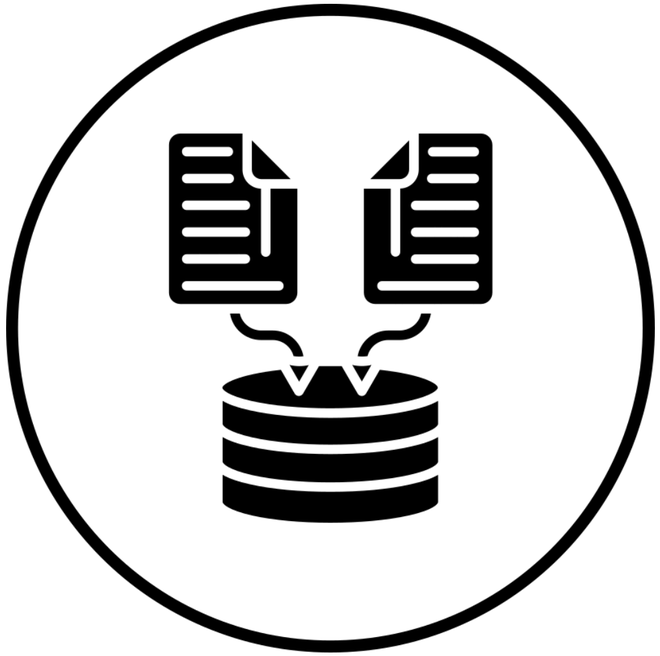


A cromatografia acoplada à espectrometria de massas (LC-HRMS e GC-MS) é uma abordagem essencial para a identificação e quantificação de compostos em matrizes complexas. Cada corrida gera arquivos volumosos contendo milhões de pontos de intensidade relacionados ao tempo de retenção e à razão m/z, geralmente em formatos proprietários como .raw, .d. A ausência de padronização entre fabricantes dificulta a interoperabilidade, limita o reaproveitamento dos dados e restringe sua integração com sistemas laboratoriais. Formatos abertos como mzML surgem como alternativas.

MARDAL et al., 2023; QIN et al., 2024; JARMUSCH et al., 2021; ROSENTHAL et al., 2024; RENNER & REUSCHENBACH, 2023.

REFERENCIAL TEÓRICO

Dados Brutos, Padronização e Bancos de Dados Analíticos



Os dados brutos produzidos por LC-HRMS e GC-MS representam um dos maiores desafios da ciência analítica devido à heterogeneidade de formatos e à dependência de softwares proprietários, como Xcalibur e MassHunter. Essa fragmentação dificulta comparabilidade, rastreabilidade e reuso científico. O uso de bancos de dados analíticos permite estruturar, armazenar e relacionar dados cromatográficos e metadados, possibilitando consultas rápidas e integração entre diferentes fontes de informação. Repositórios como ScreenDB demonstram o valor da padronização e do acesso aberto para estudos retrospectivos e mineração de dados.

QIN et al., 2024; ROSENTHAL et al., 2024; JARMUSCH et al., 2021; RENNER & REUSCHENBACH, 2023; MARDAL et al., 2023; INMON, 2016.

REFERENCIAL TEÓRICO

Engenharia de Dados, ETL e Princípios FAIR

A Engenharia de Dados fornece metodologias para coletar, transformar e integrar grandes volumes de dados em estruturas organizadas e rastreáveis. O processo ETL (extrair, transformar e carregar) permite converter arquivos proprietários ou abertos em formatos padronizados compatíveis com bancos de dados analíticos. Além de organizar e padronizar, a engenharia de dados também sustenta a governança da informação, garantindo versionamento, transparência e segurança. Quando alinhados aos princípios FAIR, os pipelines tornam os dados encontráveis, acessíveis e reutilizáveis, ampliando o valor científico e permitindo reinterpretações e estudos retrospectivos.

KIMBALL & ROSS, 2013; SANTOS et al., 2019; QIN et al., 2024; ROSENTHAL et al., 2024; LAWAL, 2024; WILKINSON et al., 2016; INMON, 2016; NEVALA, 2018.

TRABALHOS RELACIONADOS

CARACTERÍSTICAS	MetHouse (Gaida & Neumann, 2007)	Metabolite Atlases (Yao et al., 2015)	Scalable Analysis (Mårdal et al., 2023)
Técnicas Analíticas	GC-MS e LC-MS	MS-MS e LC-MS	LC-HRMS
Técnicas Computacionais	Data Warehouse (OLAP), XCMS (para processamento de picos), BioMart, JPOX/JDO2	SciDB (banco de dados paralelo para HPC), MongoDB (para metadados), IPython/Jupyter notebooks (interface de análise).	ScreenDB (estrutura de banco de dados SQL), R/Bioconductor (XCMS), Linguagens de script como Python e R.
Dataset	Dados de metabolômica	Dados de metabolômica	Dados de triagem forense de drogas
Tipo de Dados Analisados	Dados brutos de full scan, Picos processados (peak-picking), Metadados experimentais (modelo ArMet).	Dados brutos de full scan, Picos processados (peak-picking), Metadados experimentais (modelo ArMet).	Dados brutos de LC-HRMS, informações de íons fragmentados e íons precursores, Dados de QC (controle de qualidade).
Volume de Dados	41 experimentos	Nâo informado	40.000 amostras
Eficiência e Desempenho	O data warehouse é otimizado para recuperação e análise. A importação e recuperação de dados passaram por benchmarking.	Permite consultas rápidas aos dados brutos através do SciDB. Facilita a visualização e refino de anotações de características.	Permite análise retrospectiva e re-análise rápida de alvos, como detecção de novos perfis em segundos. Robusto e escalável para grandes biomonitoramentos.
Limitações Destacadas	A principal limitação é a dependência de processos de pré-processamento demorados realizados pelo pacote XCMS, que é o fator limitante em termos de velocidade	A alta dependência nas condições experimentais causa inconsistências em experimentos de larga escala	Limitada pela sua dependência de funcionalidades do SQL para indexação e consultas, o que significa que a arquitetura atual não é ideal para atingir todas as metas de pesquisa.

METODOLOGIA

Abordagem Metodológica

A pesquisa adotará uma abordagem qualitativa e experimental, baseada na comparação entre resultados antes e depois da aplicação do método proposto. O desenvolvimento integra três dimensões complementares: Engenharia de Dados, responsável pela estruturação dos dados cromatográficos; Engenharia de Software, voltada à construção da aplicação web de visualização. Essa integração caracteriza uma abordagem interdisciplinar típica de pesquisas aplicadas em sistemas inteligentes.

Yin (2018); Gregor & Hevner (2013).

MATERIAIS E MÉTODOS

Cenários

LBCD

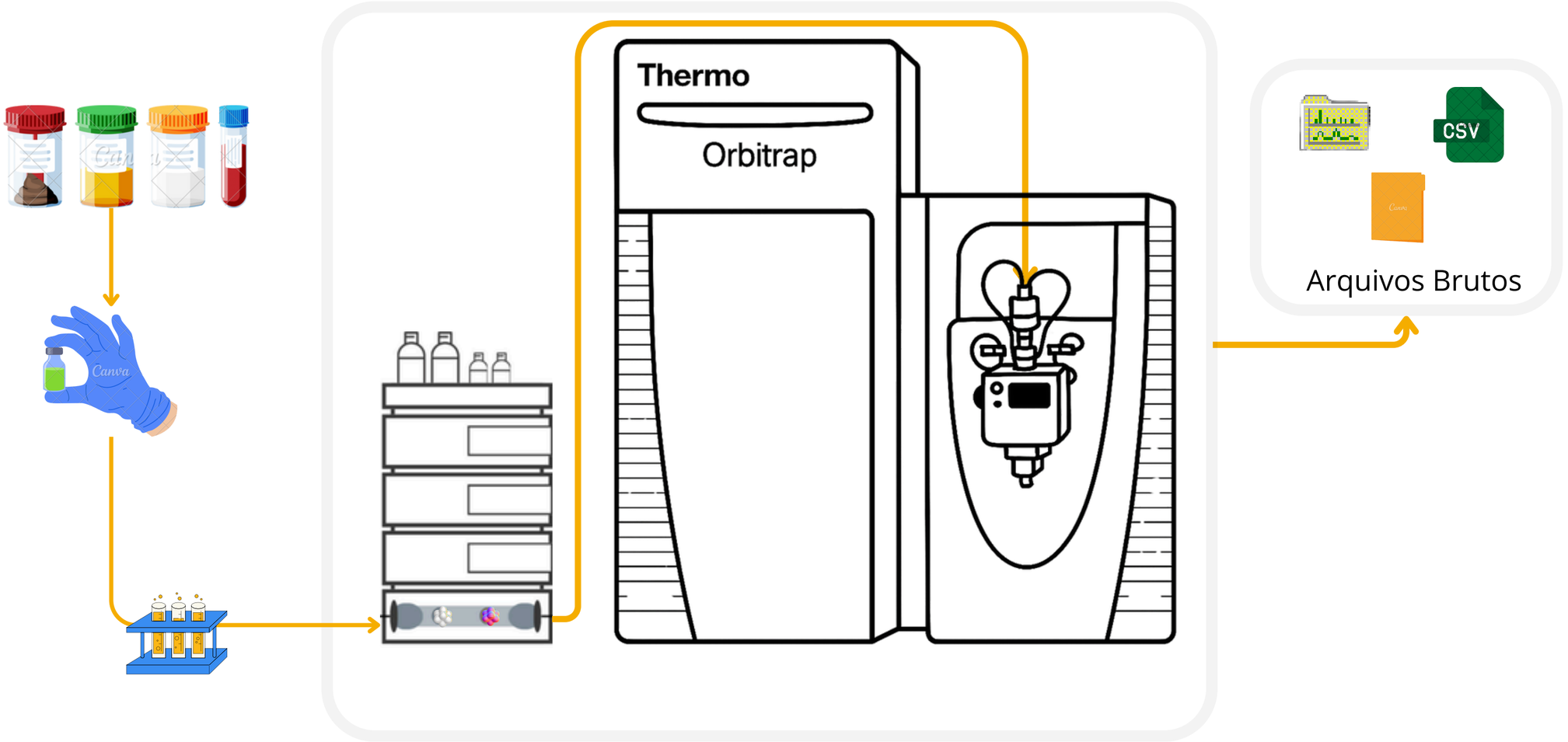
NAF

IBAMA
MMA

O método será aplicado a dados provenientes do IBAMA, do Núcleo de Análises Forenses e do Laboratório Brasileiro de Controle de Dopagem, permitindo avaliar seu desempenho em diferentes contextos analíticos e demonstrar sua aplicabilidade prática.

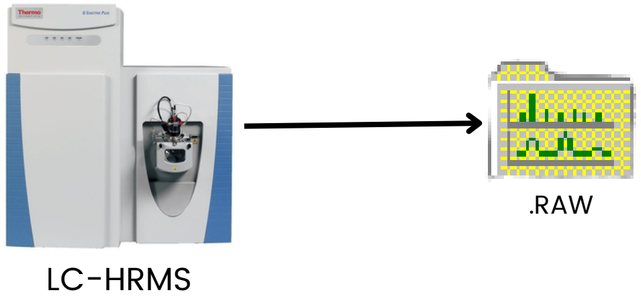
MATERIAIS E MÉTODOS

Processo de Injeção



MATERIAIS E MÉTODOS

Desenvolvimento: Etapa 1 - Análise e Caracterização dos Arquivos

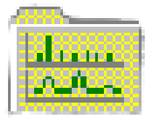


Nesta etapa, são avaliadas as características dos arquivos brutos produzidos por instrumentos LC-HRMS e GC-MS, incluindo formatos proprietários (.raw e .d). A análise estrutural permite identificar metadados, atributos relevantes e o escopo necessário para extração. Essa etapa fornece o entendimento técnico essencial para o processo de ETL e para a definição da arquitetura do banco de dados analítico.

Rosenthal et al. (2024); Qin et al. (2024).

MATERIAIS E MÉTODOS

Desenvolvimento: Etapa 2 - Definição e Aplicação do Pipeline ETL



.RAW

.D

Transformar

O pipeline ETL será implementado para converter dados brutos em estruturas padronizadas. A etapa de extração envolve leitura automatizada dos arquivos cromatográficos e a inserção dos dados tratados em um banco relacional analítico. O pipeline é fundamentado nos princípios de modelagem e integração de dados utilizados em engenharia de dados.

Kimball & Ross (2013); Santos et al. (2019).

MATERIAIS E MÉTODOS

Desenvolvimento: Etapa 3 - Modelagem e Estruturação do DB

ClickHouse

Será utiliuzado um banco de dados analítico, o ClickHouseDB, escolhido por sua arquitetura em colunas e alto desempenho em consultas de grandes volumes de dados. A modelagem relacional e analítica garantirá integridade, rastreabilidade e consistência entre amostra, substância, lote e pontos cromatográficos. Cada ponto será tratado como uma observação temporal associada a metadados experimentais. O uso de compressão, particionamento e índices permitirá consultas rápidas e escalabilidade, essenciais para grandes volumes de dados laboratoriais.

Inmon (2016); Mardal et al. (2023).

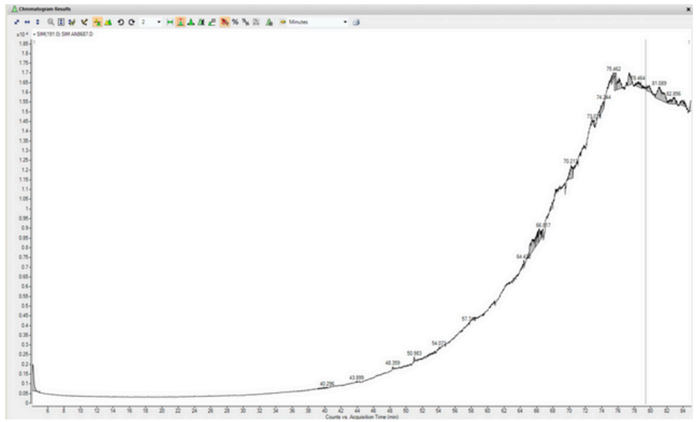
MATERIAIS E MÉTODOS

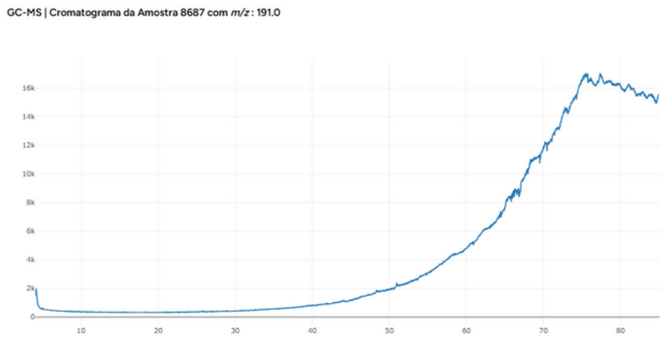
Desenvolvimento: Etapa 4 - Desenvolvimento da Aplicação Web

Será construída uma aplicação web em Python (Flask), integrando visualizações interativas com Plotly para exibição de cromatogramas. A interface permitirá navegação, consulta estruturada e seleção de amostras, oferecendo ao usuário uma forma intuitiva e eficiente de explorar os dados armazenados no banco analítico.

MATERIAIS E MÉTODOS

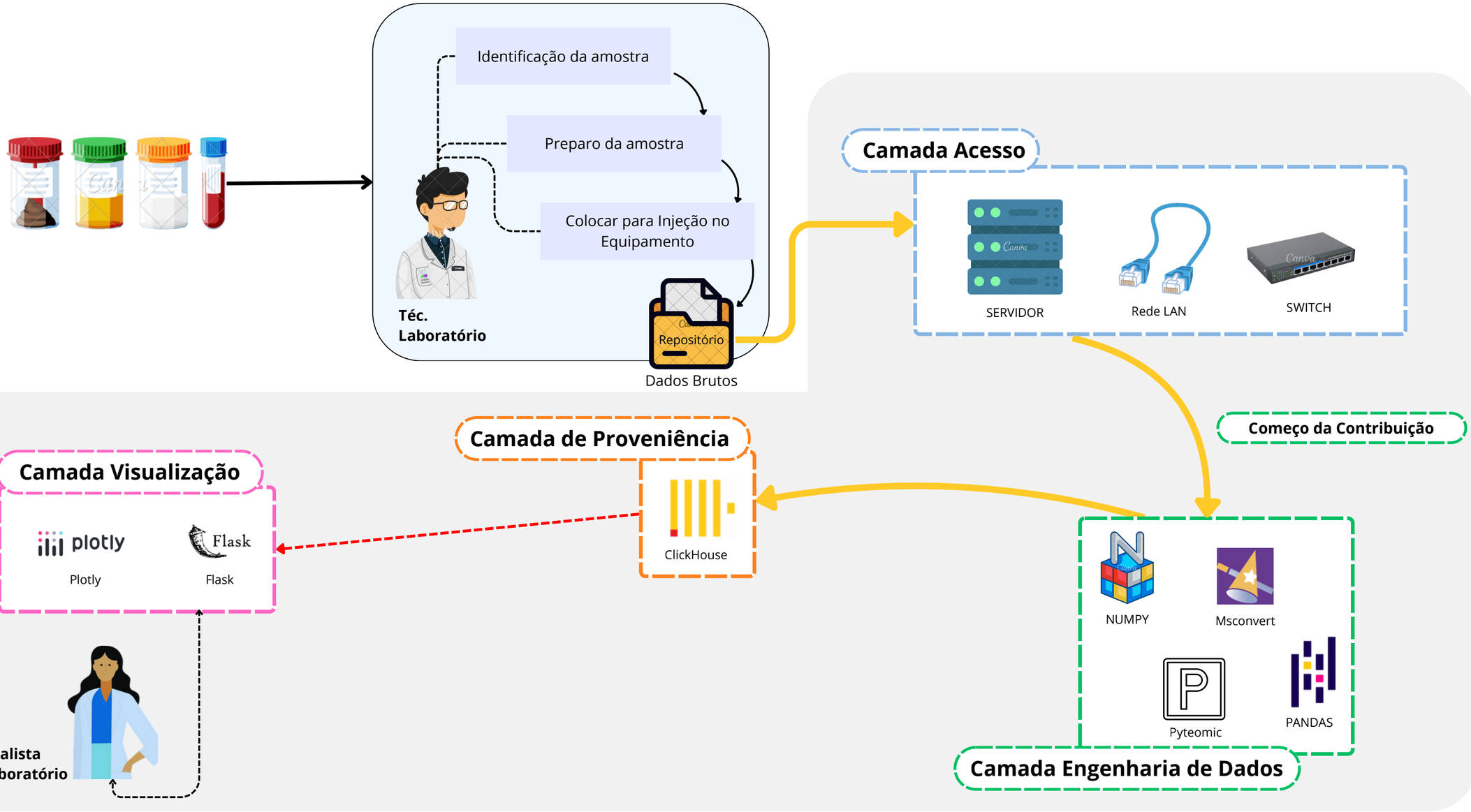
Desenvolvimento: Etapa 5 - Validação Funcional e Científica





A validação do método foi realizada de forma visual, comparando os cromatogramas e espectros reconstruídos a partir dos dados armazenados no repositório analítico com aqueles gerados diretamente a partir dos arquivos brutos nos softwares proprietários dos fabricantes. Essa abordagem permitiu verificar a preservação dos perfis cromatográficos, intensidades relativas e tempos de retenção ao longo de todo o workflow, assegurando a integridade, a reprodutibilidade e a utilidade científica dos dados armazenados, bem como a adequação do sistema aos requisitos operacionais da análise laboratorial.

ARQUITETURA



CRONOGRAMA

ETAPAS	1ºTRI 2025	2ºTRI 2025	3ºTRI 2025	4ºTRI 2025	1ºTRI 2026	2ºTRI 2026	3ºTRI 2026	4ºTRI 2026	1ºTRI 2027
Disciplinas									
Revisão sobre metodos de recuperação de dados									
Investigação do problema para recuperar os dados									
Pesquisa sobre acesso a arquivos brutos e suas bibliotecas									
Aplicação inicial convertendo o arquivo bruto em mzML e aplicando ETL									
Implementação de uma tela de visualização dos dados armazenados									
Aplicar formulas para realizar o processamento dos dados no DataLake									
Execução final utilizando ETL e Ciencia de Dados para enriquecer os dados									
Análise de resultados e interpretação final									
Redação da dissertação									
Apresentação do Projeto									
Revisões e ajustes de colaboração									
Revisão crítica da dissertação por pares									
Finalização de apêndices e material suplementar									
Formatação final segundo normas institucionais									
Simulação da apresentação para banca									
Submissão da versão final para banca									
DEFESA									
Ajustes pós-defesa e depósito final da dissertação									

CONCLUIDO	
EM ANDAMENTO	
PROXIMAS ETAPAS	

REFERÊNCIAS

JARMUSCH, A. K.; CLEMENTS, K. D.; HERRERA, A.; QUINN, R. A. Reproducible mass spectrometry data processing using open formats and standardized workflows. Analytical Chemistry, 2021.

LAI, A. et al. Challenges in mass spectrometry data management and reanalysis: toward open and interoperable solutions. Journal of Proteome Research, 2024.

MARDAL, M. et al. ScreenDB: A mass spectrometry database for toxicology data management and retrospective analysis. Analytical Chemistry, 2023.

QIN, X. et al. Advances in high-resolution mass spectrometry workflows and data standardization. Analytica Chimica Acta, 2024.

RENNEER, M.; REUSCHENBACH, B. Interoperability challenges in chromatographic data management systems. TrAC Trends in Analytical Chemistry, 2023.

ROSENTHAL, M. et al. Standardization of mass spectrometry data formats and implications for analytical reproducibility. Journal of Chromatography A, 2024.

INMON, W. H. Data Lake Architecture: Designing the Data Lake and Avoiding the Data Swamp. Amsterdam: Elsevier, 2016.

KIMBALL, R.; ROSS, M. The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling. 3. ed. New York: Wiley, 2013.

LAWAL, M. Data governance models for scientific data lifecycle management. Data & Knowledge Engineering, 2024.

NEVALA, T. Data Lake governance: preventing the evolution into Data Swamps. Information Systems Journal, 2018.

SANTOS, L. et al. ETL processes and data integration frameworks: concepts and applications. Journal of Information Systems Engineering, 2019

WILKINSON, M. et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 2016.

GIL, A. C. Métodos e técnicas de pesquisa social. 7. ed. São Paulo: Atlas, 2019.

GREGOR, S.; HEVNER, A. Positioning and presenting design science research for maximum impact. MIS Quarterly, 2013.

HEVNER, A. R. et al. Design Science in Information Systems Research. MIS Quarterly, 2004.

YIN, R. K. Case Study Research and Applications: Design and Methods. 6. ed. Thousand Oaks: Sage, 2018.

KEOGH, E.; RATH, C. On the importance of choosing the right distance measure for time series. Proceedings of the 5th ACM SIGKDD, 2001.

SENIN, P. Dynamic Time Warping algorithm review. Information Systems, 2008


PPGI
PROGRAMA
DE PÓS-GRADUAÇÃO
EM INFORMÁTICA
UNIVERSIDADE FEDERAL DO RIO DE JANEIRO


LBCD


NAF


IBAMA
MMA

OBRIGADO