# UM ESTUDO SOBRE INTEGRAÇÃO DE FONTES DE DADOS HETEROGÊNEAS UTILIZANDO AS FERRAMENTAS ONTOP E DREMIO

Gabriel de Melo Osório (PIC/UEM), Raqueline Ritter de Moura Penteado (Orientadora). E-mail: rrmpenteado@uem.br.

Universidade Estadual de Maringá, Departamento de Informática, Maringá, PR.

**Área e subárea do conhecimento: Ciência da Computação/Sistemas de Informação.**

**Palavras-chave**: integração de dados, *dremio*, *ontop*.

## RESUMO

Dados são os principais recursos consumidos por analistas de negócios. Com a combinação deles, analistas geram informações que auxiliam tomadas de decisões em ambientes empresariais. Com o advento da Indústria 4.0 e da Internet das Coisas, tais informações podem ser geradas a partir da integração semântica entre diversas fontes de dados heterogêneas, autônomas e distribuídas. Porém, a heterogeneidade representa um desafio para a integração de dados. Sistemas OBDA (*Ontology-based Data Access*) exploram o conceito de ontologias para integração semântica de dados. O *Ontop* é um sistema OBDA que faz a integração semântica entre fontes de dados relacionais heterogêneas. Entretanto, nos últimos anos, o *Ontop* passou a dar suporte ao *Dremio* para viabilizar também a integração de fontes de dados que adotam modelos lógicos distintos. O projeto em questão realizou um estudo sobre a integração de dados viabilizada pelo *Dremio* ao *Ontop,* implementando e analisando um estudo de caso, a fim de gerar de diminuir a curva de aprendizado da utilização das ferramentas em trabalhos futuros.

## INTRODUÇÃO

A integração de dados provenientes de fontes diversas desempenha um papel crucial na geração de informações relevantes em uma variedade de setores, como indústria, saúde e educação. Essas informações têm o potencial de orientar decisões tanto gerenciais quanto operacionais. No entanto, a heterogeneidade intrínseca a essas fontes de dados representa um desafio para a integração, devido à necessidade de estabelecer a interoperabilidade entre sistemas distintos. Essa heterogeneidade pode ser classificada em quatro categorias, sendo elas: sistema, envolvendo diferença de hardware ou sistema operacional; sintaxe, envolvendo diferenças de linguagens e representações de dados; estrutura, em que o modelo lógico dos dados são diferentes; e, por fim, semântica, quando fontes de dados usam diferentes termos para os mesmos conceitos (Cui, Z.; O'Brien, P., 2000).

O *Ontop* (Calvanese et al., 2017) é um sistema OBDA que propõe uma alternativa para superar o desafio de integração semântica. Ele cria grafos RDF virtuais para a integração de fontes autônomas e distribuídas e simplifica o acesso aos dados por meio da representação semântica unificada fornecida por uma ontologia, que oferece um vocabulário para expressar e comunicar o conhecimento de um domínio, permitindo uma visão coesa das diversas fontes de dados (Cui, Z.; O'Brien, P., 2000). Uma consulta SPARQL, baseada na ontologia, é traduzida para SQL (*Structured Query Language*), por meio de mapeamentos pré-definidos no *Ontop*, e enviada para um SGBD, que retorna o resultado ao *Ontop*.

No âmbito da pesquisa realizada neste projeto, destaca-se o papel do *Dremio*, uma ferramenta que estende a capacidade de integração de dados do *Ontop* ao lidar com a heterogeneidade entre bases de dados que adotam diferentes modelos lógicos, incluindo o relacional e o baseado em documentos. O *Dremio* é um *Data Lakehouse* responsável por integrar bases de dados com modelos lógicos distintos por meio da virtualização das mesmas no formato relacional (Dremio, 2021).

O presente projeto se propôs a compreender a operacionalização do *Dremio* e explorar a integração de fontes de dados com modelos lógicos distintos. Para isso foi realizado um estudo de caso que integrou fontes de dados relacional e baseada em documentos, empregando o *MySQL* e o *MongoDB*, respectivamente.

## MATERIAIS E MÉTODOS

No estudo de caso, uma corporação se deparou com dois problemas ao tentar realizar a integração entre as bases de dados de duas lojas distintas, sendo eles: *i)* a heterogeneidade semântica entre as bases, que tratam de dados do mesmo contexto porém com semânticas distintas; e, *ii)* a heterogeneidade estrutural dos dados. Uma base utilizava o modelo relacional para armazenamento de dados e a outra utilizava o modelo orientado a documentos.

Para explorar o problema de heterogeneidade estrutural entre as bases do estudo utilizou-se o *Dremio* (versão 4.9.1) e para tratar a semântica utilizou-se o *Ontop* (versão 4.2.0) (como um *plugin* no *Protégé* (versão 5.5.0), responsável por dar suporte a construção e manipulação de ontologias.). O estudo de caso também envolveu o *MySQL* (versão 8.0.26) e o *MongoDB* (versão 1.30.1). O *MySQL* adota o modelo de dados relacional e o *MongoDB*, o modelo de dados baseado em documentos. Além dos manuais das ferramentas, utilizou-se de fóruns referentes ao *Dremio*, uma vez que o material disponível foi insuficiente para o seu entendimento.

## RESULTADOS E DISCUSSÃO

A Figura 1 mostra uma consulta SPARQL requisitada ao *Ontop* (a), a consulta SQL gerada pelo *Ontop* e enviada ao *Dremio* (b) e o resultado retornado pelo *Dremio* ao *Ontop* (c). A consulta recupera o CPF, o sexo e a data de nascimento dos clientes. A

consulta envolveu as duas bases uma vez que o atributo sexo está na base orientada a documentos e a data de nascimento na base relacional.

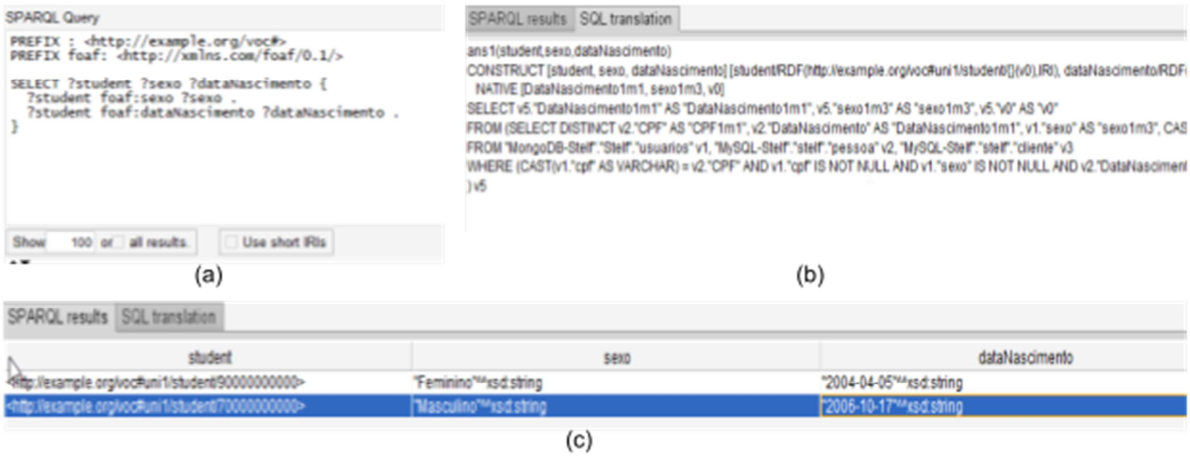
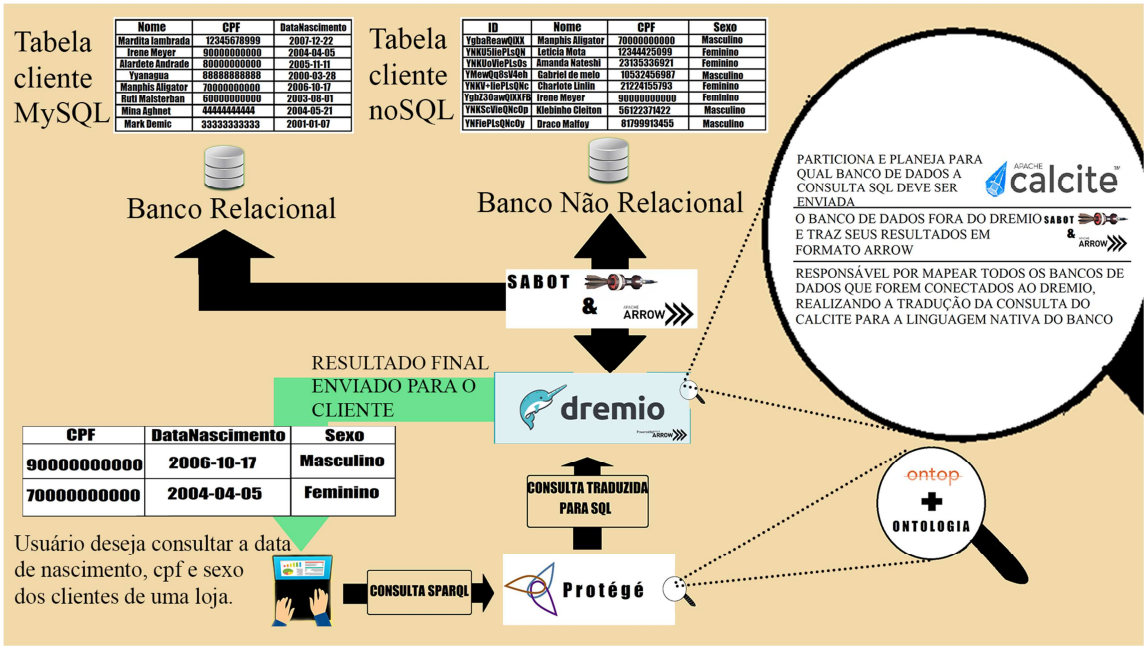



Figura 1 - Consulta SPARQL (a), consulta SQL (b) e resultado da consulta (c).




| CPF | DataNascimento | Sexo |
|---|---|---|
| 90000000000 | 2006-10-17 | Masculino |
| 70000000000 | 2004-04-05 | Feminino |

Figura 2 - Processo envolvido na integração de dados envolvendo as ferramentas *Protégé*, *Ontop*, *Dremio*, *MySQL* e *MongoDB*.

A execução da consulta envolveu diversas etapas. A Figura 2 mostra as etapas envolvidas no processamento da consulta. O usuário requisitou a consulta SPARQL ao *Ontop* por meio do *Protégé*. Explorando mapeamentos pré-definidos, o *Ontop* gera a consulta SQL que é enviada ao *Dremio*, que utiliza diversas tecnologias para retornar a resposta da consulta ao *Ontop* e por fim ao usuário. A primeira tecnologia é o *Apache Calcite*, responsável por analisar e otimizar a consulta, buscando por meio de diagramas relacionais as bases que serão necessárias para respondê-la.

Com as bases elencadas, entra em ação o *Apache Arrow*, que é utilizado pela a *engine* de execução do *Dremio*, denominada de *Sabot*, traduzindo a requisição para a base consultada dependendo do modelo lógico da base. O formato de dados utilizado pelo *Sabot* é interoperável entre todos os bancos sem a necessidade de serialização e deserialização dos dados em uma potencial tradução, o qual é armazenado em memória principal, no formato colunar, considerada a mais eficiente para consultas em diversos estudos. Por fim, após a tradução para este formato, ela é enviada às bases pelo protocolo *Apache Arrow Flight*, um protocolo de comunicação RPC otimizado para dados colunares, que prontamente retorna o resultado da consulta no formato relacional (Dremio, 2019).

## CONCLUSÕES

A partir do estudo foi possível compreender os elementos envolvidos na integração semântica e estrutural de bases heterogêneas. O maior desafio encontrado no estudo foi a falta de material sobre o *Dremio*, visto que o fabricante da ferramenta fornece uma visão superficial da tecnologia, dificultando a extração de conhecimento sobre as tecnologias e processos envolvidos na ferramenta.

O uso da ferramenta *Dremio* com o *Ontop* viabiliza a integração estrutural de dados além da integração semântica de fontes de dados relacionais distintas por meio de ontologias. Sendo assim, uma consulta SPARQL, que utiliza o vocabulário de uma ontologia, recupera e combina dados alocados em fontes distintas heterogêneas estrutural e semanticamente. Esse tipo de integração pode viabilizar a geração de informações para tomadas de decisão por gestores de instituições.

## REFERÊNCIAS

CALVANESE, D.; BENJAMIN, C.; KOMLA-EBRI, S.; KONTCHAKOV, R.; LANTI, D.; REZK, M.; RODRIGUEZ-MURO, M.; XIAO, G.. Ontop: Answering SPARQL queries over relational databases. Semantic Web, vol. 8., 2016.

CUI, Z.; O'BRIEN, P.. Domain Ontology Management Environment. Proceedings of the 33rd Annual Hawaii International Conference on System Sciences, Maui, HI, USA, 2000.

DREMIO. Using arrow, calcite and parquet to build a relational cache. Disponível em https://www.dremio.com/resources/webinars/apache-arrow-calcite-parquet-relational-cache/, 2019, Acesso em 18/08/2021.

DREMIO. Dremio architecture guide architecting. Disponível em https://www.dremio.com/downloads/DremioArchitectureGuide.pdf, 2021, Accesso em 18/08/2021.