Engenharia de Dados

Tecnologias de Big Data - Projeto

Arquitetura de Big Data no Processamento de Dados Epidemiológicos

Dados do SUS: Leishmaniose Visceral Humana

LABDATA | FIA FUNDAÇÃO INSTITUTO DE ADMINISTRAÇÃO BUSINESS SCHOOL

15/08/2023

Agenda

1. Definição do problema
2. Contextualização
3. Definição da arquitetura
   - Ingestão, armazenamento, processamento e exploração
4. Desenvolvimento
5. Disponibilização
6. Conclusões

1. Definição do Problema

O objetivo do trabalho é construir uma arquitetura e um pipeline de processamento de dados epidemiológicos e complementares para que a análise como um todo desses dados seja o mais automatizada possível.

FIA BUSINESS SCHOOL lab.data

2. Contextualização do Problema

A compreensão de como doenças avançam no espaço e tempo são de grande relevância para toda a sociedade, principalmente para as Vigilâncias Epidemiológicas que podem construir planos de ação com base em análises dessa natureza.

Portanto, o monitoramento em tempo real, ou o mais próximo disso, é crucial para que novos padrões espaciais e/ou temporais sejam detectados a tempo e de forma eficiente. Ademais, a visualização destas análises também é muito importante para os tomadores de decisão.

Logo, um pipeline automatizado de aquisição, processamento, mineração e disponibilização dessas informações pode aumentar a eficiência de profissionais que lidam com essas informações e também diminuir chances de erros humanos no processo como um todo.

2. Contextualização do Problema

- Doença analisada: Leishmaniose Visceral Humana
  Doença parasitária, vetorial, crônica, sistêmica, que afeta o sistema imunológico de animais e seres humanos.

- Fonte: DataSus: Sinan
  O Sistema de Informação de Agravos de Notificação - **Sinan** é alimentado, principalmente, pela notificação e investigação de casos de doenças e agravos
  Veja em: https://datasus.saude.gov.br/transferencia-de-arquivos/

- Período analisado
  Casos de todo o Brasil notificados entre 2007 e 2020

Os dados de agravos no portal do Sinan são fornecidos em um formato desenvolvido pelo Ministério da Saúde (dbc), para que seja trabalhado em outras ferramentas além do Tabnet (fornecido pelo ministério) é necessário convertê-los. O readdbc é uma biblioteca em R que foi desenvolvida para esta finalidade.
Veja em: https://github.com/danicat/read.dbc

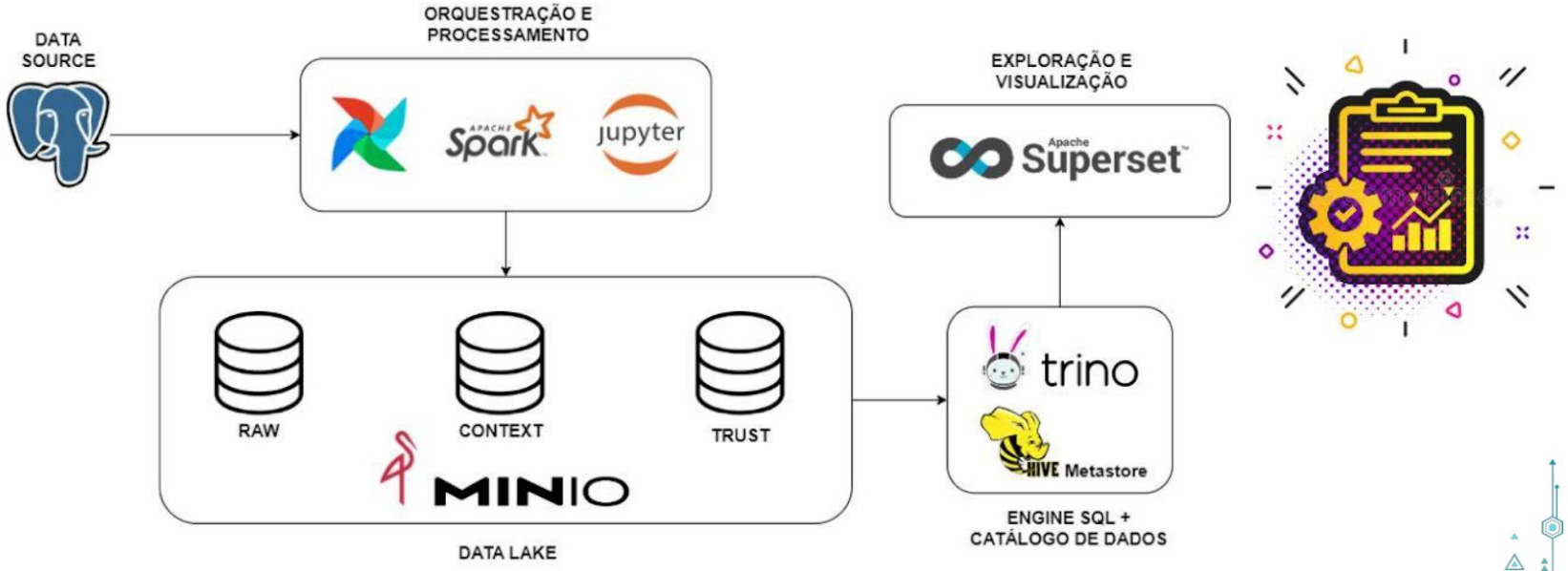
A seguir será explicado o fluxo de conversão e preparação necessários para trabalhar com estes dados.

FIA BUSINESS SCHOOL lab.data

3. Arquitetura Sugerida

Projeto

2. CONCEITO

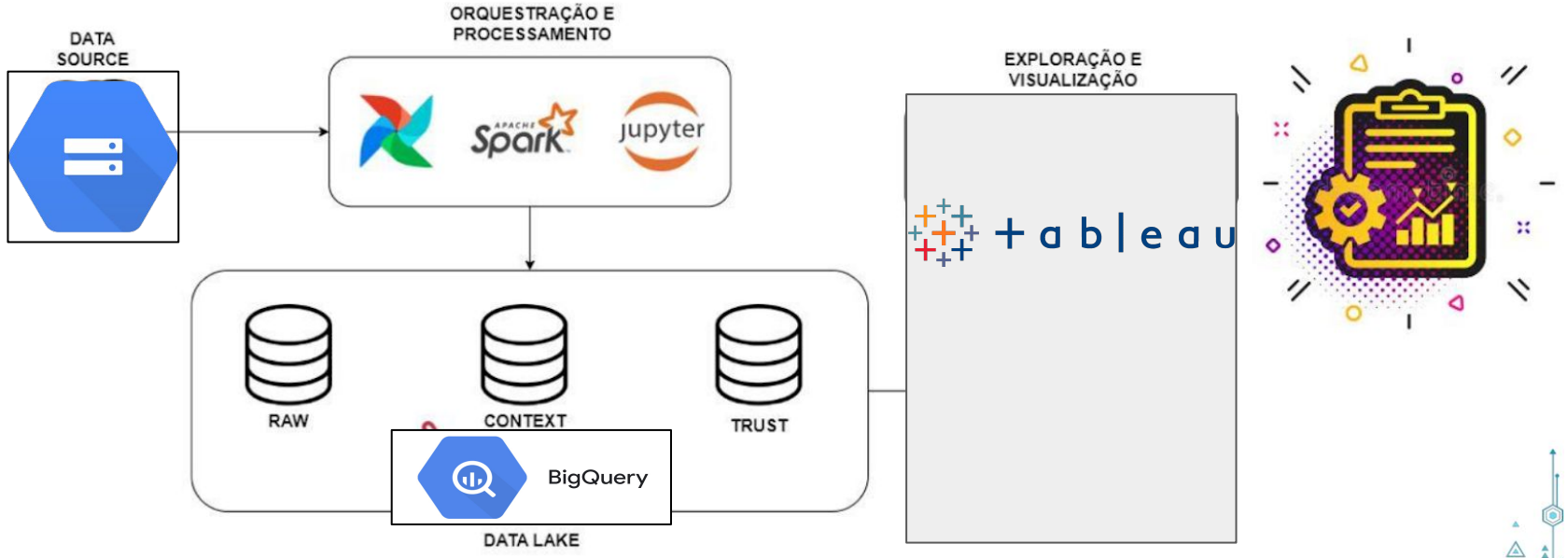


Os dados brutos são processados, transferidos e carregados em um data lake através de Apache Airflow (orquestrador), Spark, e Jupyter Notebook (exploração). O data lake é composto por camadas raw, context e trust no MINIo, cuja principal característica é a boa integração com o S3 da Amazon.
O trino e hive metastore, como catálogo de dados operam entre data lake e ferramenta de visualização e exploração, neste caso o Superset. O Superset é o responsável pela disponibilização dos dados.

3. Arquitetura Adotada

Projeto

2. CONCEITO



Os dados brutos, no formato dbc, estão no bucket do GCP.
Estes são orquestrados pelo Airflow, o pipeline consiste em conversões de tipo muito básicas (preservando o dado bruto) e em seguida o fluxo envia os dados para a camada raw do data lake.

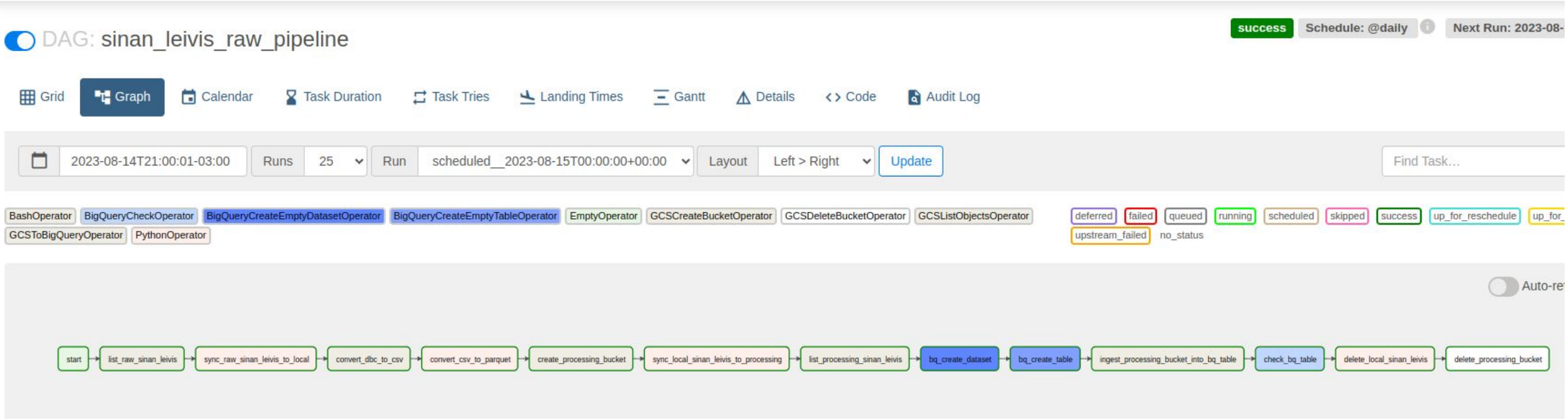
Posteriormente, os dados na camada raw são processados e armazenados em context. Por fim, a camada trust armazena os dados transformados, enriquecidos (pós processo de mineração) e prontos para visualização.

FIA BUSINESS SCHOOL lab.data

4. Desenvolvimento

Para cada base de dados há três DAGs para orquestrar o processamento: raw, context e trus.
Neste caso, como só uma base está sendo trabalhada, são três ao todo.

github : https://github.com/FlaviaLopes/spatio-temporal-analysis-techniques



5. Disponibilização

As análises serão disponibilizadas de forma visual pelo Tableau no perfil abaixo.

Tableau : https://public.tableau.com/app/profile/flavia.lopes/viz

6. Conclusões

Próximos passos:

- alguns algoritmos de clusterização serão experimentados na etapa de mineração no Spark, devido a alta carga de processamento necessária.