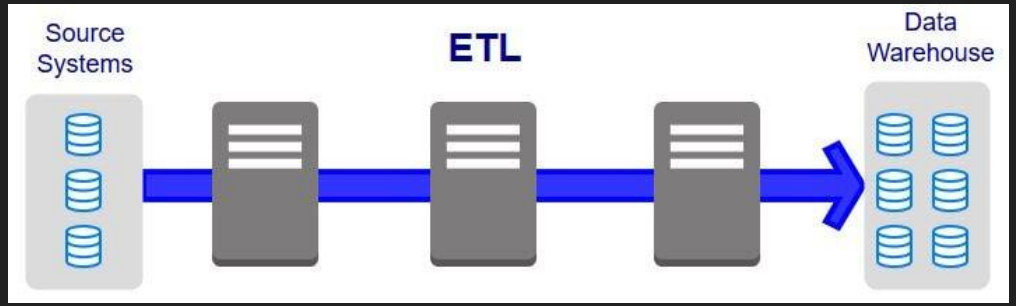
ETL - Com informações financeiras

Lucas de Castro Oliveira

O que significa ETL

- ETL vem do inglês Extract, Transform, Load. São operações envolvendo grande volumes de dados compostas em três fases: extração, transformação e carregamento (ou armazenamento).

- Muito comum quando se quer transferir um ou vários conjuntos de dados de uma fonte para a outra



Exemplo utilizado - Informações de fundos da CVM

- A comissão de valores mobiliários é uma entidade brasileira que homologa e regulamenta a venda de papéis financeiros tais como fundos de investimento, ações na bolsa da valores, e outros ativos do mercado financeiro.

- Possui um portal de dados abertos em: http://dados.cvm.gov.br/ onde é possível fazer o download de diversas planilhas tais como movimentações em fundos de investimento, informações de cadastro, balancetes e entre outros.

Exemplo utilizado - Informações de fundos da CVM

- Construí um exemplo de ETL com uma planilha de dados chamada inf_diario_fi_201911.csv, que nada mais é um arquivo CSV onde cada linha representa uma informação diária em relação ao mês de novembro de um fundo de investimento(diferenciada por um CNPJ)

- Esta planilha contém em torno de 200.000 linhas em 23MB de informações textuais.

- Objetivo: extrair as informações da planilha, fazer uma pequena transformação e carregar em um banco de dados.

Tecnologias Utilizadas

- A seguinte stack foi utilizada para resolver este problema:
  a. Linguagem de programação Java8.
  b. Framework Spring juntamente com os módulos:
     - Spring Batch (útil para tarefas de processamento em lote, perfeito para ETL).
     - Spring JPA (para fazer mapeamento das classes para tabelas em bancos de dados e cuidar das queries SQL).
     - Spring Web, para expor as informações armazenadas no banco de dados em uma API REST.
  c. Banco de dados MySQL 8 como destino final dos dados.

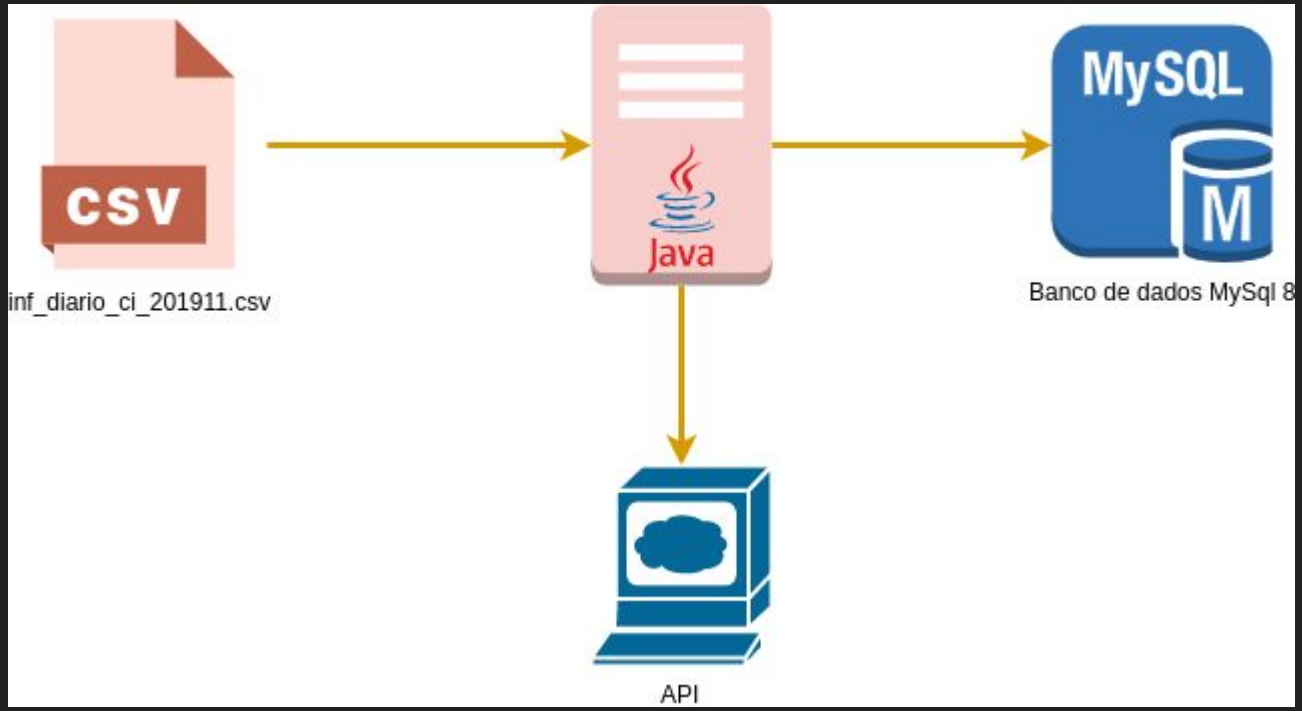


Diagrama do esquema criado

Estudos futuros

- Melhorar a apresentação dos dados. Utilizar planilha de informações cadastrais da CVM.
- Criar um cliente JavaScript para renderizar estas informações em gráficos.
- Aumentar a performance da operação com tunning da paralelização.
- Fazer testes com outros bancos de dados.
- Utilizar Eleasticsearch para indexar informações de modo a acelerar buscar.
- Fazer WebScrapping para baixar planilha da CVM todo o dia….

Obrigado

https://github.com/lcastrooliveira/funds_daily_report