Eventos de fluxo

Documentação da arquitetura de dados

Este documento apresenta uma visão geral sobre a arquitetura de dados projetada para armazenar eventos de fluxo registrados em estabelecimentos comerciais do setor alimentício. Este projeto foi estabelecido como uma forma de avaliação em um processo seletivo para Engenheiro de Dados na empresa Geofusion.

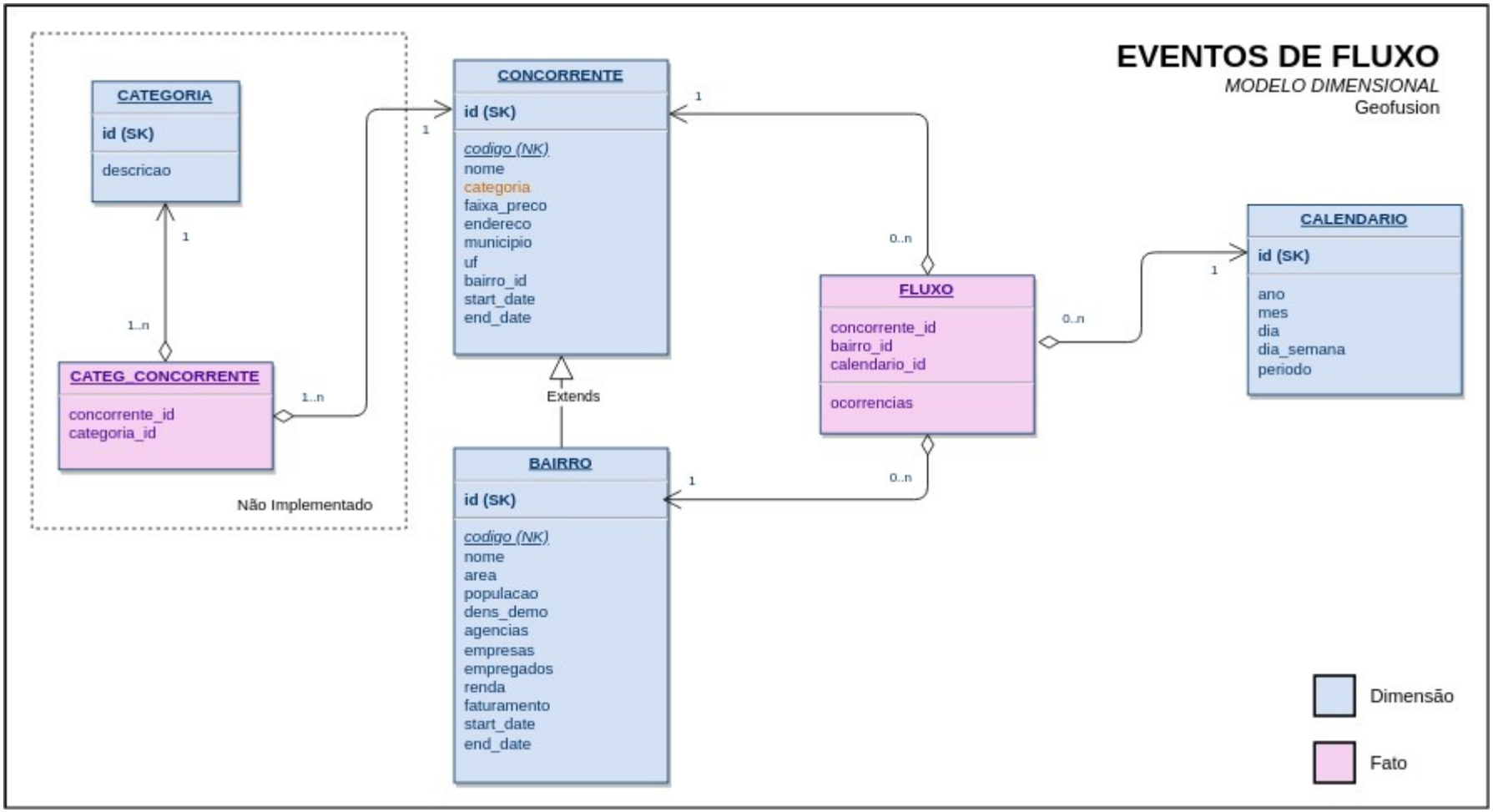
Autor:
André Costa
(andre@costa.eti.br)
https://www.linkedin.com/in/a-l-costa

11 de março de 2019

Requisitos e fontes de dados originais

- Tanto dos dados originais, quanto os requisitos, estão diretamente associados à análise de movimentação comercial no setor de alimentos, em nível de bairros
- As fontes de dados foram disponibilizadas em arquivos com quantidade considerável de dados à serem inseridos do data warehouse, propício à utilização de um pipeline ETL baseado em processo de batch, ao invés de streaming
- Há interesse explicito em o sistema de dados suportar armazenamento e acesso escaláveis, e preferencialmente robusto à falhas
- Foram disponibilizados 5 arquivos de dados que, essencialmente, descrevem os estabelecimentos comerciais e as características demográficas de suas localizações, juntamente com registros de visitação de clientes feitos a partir de seus celulares, com dia e hora de registro das ocorrências
- Tempo é um fator importante neste projeto, e há interesse específico em se fazer a análise dos dados com base nos períodos do dia (manhã, tarde e noite), e nos dias da semana

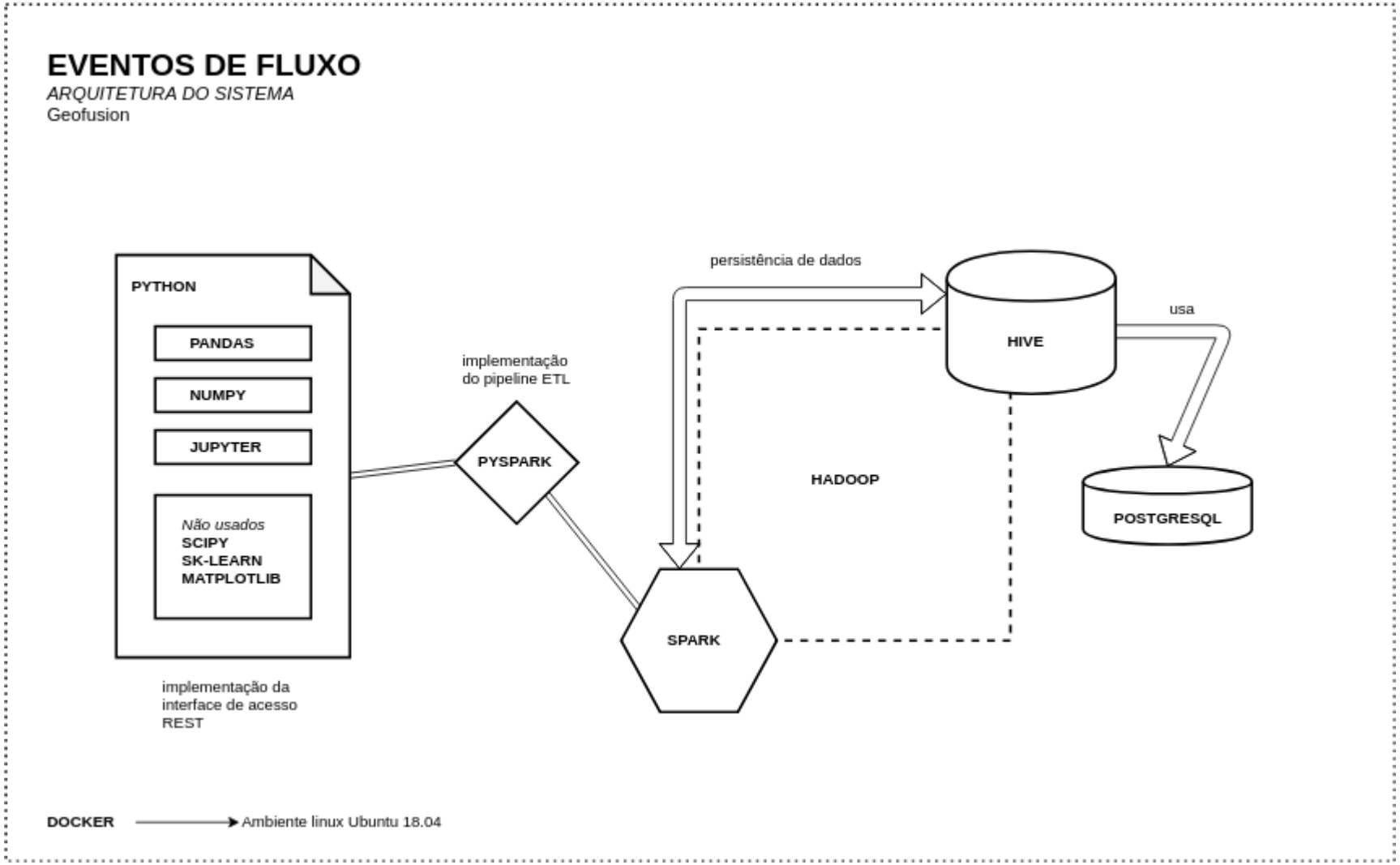
Modelo de dados



Considerações sobre o modelo de dados

- Após analisarmos as características do problema, concluímos que apenas duas dimensões são estritamente necessárias: a dimensão CONCORRENTE, e a dimensão CALENDARIO
- Porém, considerando a necessidade de utilização de atributos de *tipo 2* em diversos itens da dimensão CONCORRENTE, decidimos por utilizar a mini-dimensão BAIRRO para armazenar os atributos com maior probabilidade de serem atualizados
- Adicionalmente, houve a possibilidade de diversos atributos de grande volume, como endereço e lista de categorias, ficarem na dimensão principal, otimizando assim o armazenamento do data warehouse
- Algo que consideramos, por também otimizar o armazenamento, e por facilitar análises envolvendo subcategorias, foi criar uma dimensão CATEGORIA, e uma tabela adicional de fato para conectar à dimensão CONCORRENTE. Contudo, consideramos que o custo de implementação neste momento seria inviável, pesando também o fato de não ser um requisito explicito

Tecnologias utilizadas



Descrição das tecnologias

- Com o objetivo de ser um sistema escalável, escolhemos utilizar o Hadoop como plataforma, por ele oferecer um sistema de arquivos distribuído, e estar associado à muitas tecnologias específicas para a engenharia de dados
- Os principais software utilizados foram o Spark, para processamento intermediário do ETL, e do Hive, para armazenamento do data warehouse de forma gerenciada dentro do HDFS do Hadoop. O Hive também foi consideravelmente utilizado na etapa de transformação dos dados
- Escolhemos a linguagem de programação Python para implementar o pipeline ETL, e o acesso REST aos dados. Um grande peso foi familiaridade com a linguagem, mas além disso, o Python ofereçe uma suite de bibliotecas de machine learning e processamento de dados que podem ser extremamente úteis para a análise dos dados
- Decidimos por utilizar o jupyter notebook como interface de programação porque ele facilita o processo como um todo, principalmente de documentação