Documentação de Datasets: Caso de Uso UberEats

Introdução

Este documento fornece uma visão geral das entidades de negócio e suas respectivas fontes de dados, que serão utilizadas nos laboratórios práticos para desenvolver pipelines de dados utilizando Apache Spark e suas APIs.

Visão Geral da Arquitetura de Dados

Os dados estão distribuídos em múltiplos sistemas, simulando uma arquitetura típica de microserviços:

- **PostgreSQL**: Armazena dados relacionados a motoristas e inventário
- **MySQL**: Mantém informações sobre restaurantes, produtos, avaliações e menu
- **MongoDB**: Contém dados de usuários, itens, recomendações e tickets de suporte
- **Apache Kafka**: Gerencia streams de eventos como pedidos, pagamentos, status, GPS e rotas

Entidades de Negócio

1. Usuários

Representam os clientes que fazem pedidos na plataforma.

Fontes:

- `mongodb/users`: Dados principais dos usuários
- `mssql/users`: Informações complementares e profissionais

Atributos principais:

- `user_id`: Identificador único do usuário
- `cpf`: Documento de identificação brasileiro
- `email`: Endereço de email para contato
- `delivery_address`: Endereço de entrega
- `phone_number`: Número de telefone
- `country`: País (predominantemente Brasil)
- `city`: Cidade

2. Restaurantes

Estabelecimentos cadastrados que oferecem produtos para delivery.

Fonte:

- `mysql/restaurants`

Atributos principais:

- `restaurant_id`: Identificador único do restaurante
- `name`: Nome do estabelecimento
- `cnpj`: Documento de identificação empresarial brasileiro
- `address`: Endereço físico
- `cuisine_type`: Tipo de culinária (ex: Italiana, Japonesa)
- `opening_time`/`closing_time`: Horários de funcionamento
- `average_rating`: Avaliação média
- `num_reviews`: Número total de avaliações

3. Motoristas/Entregadores

Parceiros responsáveis pela entrega dos pedidos.

Fonte:

- `postgres/drivers`

Atributos principais:

- `driver_id`: Identificador único do motorista
- `first_name`/`last_name`: Nome do motorista
- `license_number`: Número da habilitação
- `vehicle_type`: Tipo de veículo (Carro, Moto, Bicicleta, etc.)
- `vehicle_make`/`vehicle_model`: Fabricante e modelo do veículo
- `vehicle_year`: Ano do veículo

4. Produtos

Itens disponíveis para pedido nos restaurantes.

Fontes:

- `mysql/products`: Catálogo geral de produtos
- `mongodb/items`: Instâncias de produtos em pedidos

Atributos principais:

- `product_id`: Identificador único do produto
- `name`: Nome do produto
- `price`: Preço de venda
- `unit_cost`: Custo unitário
- `restaurant_id`: Restaurante ao qual o produto pertence
- `cuisine_type`: Tipo de culinária
- `calories`: Informação nutricional
- `is_vegetarian`/`is_gluten_free`: Características dietéticas

5. Pedidos

Solicitações de entrega feitas pelos usuários.

Fonte:

- `kafka/orders`

Atributos principais:

- `order_id`: Identificador único do pedido
- `user_key`: Referência ao usuário
- `restaurant_key`: Referência ao restaurante
- `driver_key`: Referência ao motorista
- `order_date`: Data e hora do pedido
- `total_amount`: Valor total do pedido
- `payment_key`: Referência ao pagamento

6. Status de Pedidos

Acompanhamento do ciclo de vida de cada pedido.

Fonte:

- `kafka/status`

Valores possíveis:

- Order Placed (Pedido Realizado)
- In Analysis (Em Análise)
- Accepted (Aceito)
- Preparing (Em Preparação)
- Ready for Pickup (Pronto para Retirada)
- Picked Up (Retirado)
- Out for Delivery (Em Rota de Entrega)
- Delivered (Entregue)
- Completed (Finalizado)

7. Pagamentos

Transações financeiras associadas aos pedidos.

Fonte:

- `kafka/payments`

Atributos principais:

- `payment_id`: Identificador único do pagamento
- `order_key`: Referência ao pedido
- `amount`: Valor total cobrado
- `method`: Método de pagamento (Cartão, Boleto, Carteira digital)
- `status`: Status da transação (succeeded, failed, pending, refunded)
- `currency`: Moeda da transação
- `card_brand`: Bandeira do cartão
- `provider`: Provedor de pagamento (PayPal, Adyen, etc.)

8. Avaliações

Feedback dos clientes sobre os restaurantes.

Fonte:

- `mysql/ratings`

Atributos principais:

- `rating_id`: Identificador único da avaliação
- `restaurant_identifier`: Referência ao restaurante
- `rating`: Pontuação (escala de 1-5)
- `timestamp`: Data e hora da avaliação

9. Menu

Organização dos produtos em seções nos restaurantes.

Fonte:

- `mysql/menu`

Atributos principais:

- `menu_section_id`: Identificador único da seção
- `restaurant_id`: Restaurante ao qual a seção pertence
- `name`: Nome da seção (ex: Entradas, Pratos Principais)
- `description`: Descrição da seção
- `active`: Indica se a seção está ativa

10. Inventário

Controle de estoque dos produtos nos restaurantes.

Fonte:

- `postgres/inventory`

Atributos principais:

- `stock_id`: Identificador único do registro de estoque
- `restaurant_id`: Referência ao restaurante
- `product_id`: Referência ao produto
- `quantity_available`: Quantidade disponível em estoque
- `last_updated`: Data e hora da última atualização

11. Rotas

Trajetos percorridos pelos entregadores.

Fonte:

- `kafka/route`

Atributos principais:

- `route_id`: Identificador único da rota
- `order_id`: Referência ao pedido
- `driver_id`: Referência ao motorista
- `start_time`/`end_time`: Horários de início e fim
- `start_lat`/`start_lon`: Coordenadas de origem
- `end_lat`/`end_lon`: Coordenadas de destino
- `distance_km`: Distância percorrida
- `estimated_duration_min`: Tempo estimado

12. Dados GPS

Pontos de localização em tempo real durante as entregas.

Fonte:

- `kafka/gps`

Atributos principais:

- `gps_id`: Identificador único do ponto GPS
- `order_id`: Referência ao pedido
- `timestamp`: Data e hora da captura
- `lat`/`lon`: Coordenadas geográficas
- `altitude`: Altitude em metros
- `speed_kph`: Velocidade em km/h
- `direction_deg`: Direção em graus
- `accuracy_m`: Precisão em metros

13. Recibos

Comprovantes fiscais dos pedidos.

Fonte:

- `kafka/receipts`

Atributos principais:

- `receipt_id`: Identificador único do recibo
- `order_id`: Referência ao pedido
- `payment_id`: Referência ao pagamento
- `total_amount`: Valor total
- `item_count`: Quantidade de itens
- `receipt_generated_at`: Data e hora de geração

14. Eventos de Pagamento

Registro de ocorrências relacionadas às transações.

Fonte:

- `kafka/events`

Atributos principais:

- `event_id`: Identificador único do evento
- `payment_id`: Referência ao pagamento
- `event`: Objeto com informações do evento
    - `event_name`: Tipo do evento (succeeded, settled, closed, etc.)
    - `timestamp`: Data e hora do evento

15. Suporte ao Cliente

Tickets de atendimento e solução de problemas.

Fonte:

- `mongodb/support`

Atributos principais:

- `ticket_id`: Identificador único do ticket
- `user_id`: Referência ao usuário
- `order_id`: Referência ao pedido
- `category`: Categoria do problema (Late Delivery, Payment Issue, etc.)
- `description`: Descrição do problema
- `status`: Estado atual (Open, In Progress, Resolved, Closed)
- `opened_at`: Data e hora de abertura

16. Buscas

Consultas realizadas pelos usuários na plataforma.

Fonte:

- `kafka/search`

Atributos principais:

- `search_id`: Identificador único da busca
- `user_id`: Referência ao usuário
- `query_text`: Texto pesquisado
- `filters`: Filtros aplicados
- `result_count`: Quantidade de resultados
- `clicked_product_id`: Produto selecionado após a busca
- `timestamp`: Data e hora da pesquisa

17. Recomendações

Sistema de sugestões personalizadas para os usuários.

Fonte:

- `mongodb/recommendations`

Atributos principais:

- `event_id`: Identificador único do evento de recomendação
- `user_id`: Referência ao usuário
- `event_type`: Tipo do evento (view, click, add_to_cart, recommendation_served)
- `product_id`: Produto recomendado ou interagido
- `timestamp`: Data e hora do evento

18. Turnos de Motoristas

Períodos de trabalho dos entregadores.

Fonte:

- `kafka/shift`

Atributos principais:

- `shift_id`: Identificador único do turno
- `driver_id`: Referência ao motorista
- `start_time`/`end_time`: Início e fim do turno
- `shift_type`: Período (Morning, Afternoon, Evening, Overnight)
- `shift_duration_min`: Duração em minutos
- `earnings_brl`: Ganhos durante o turno
- `num_orders`: Quantidade de pedidos atendidos
- `distance_covered_km`: Distância total percorrida
- `shift_rating`: Avaliação do desempenho

Relações entre Entidades

As principais relações entre as entidades são:

1. **Usuários → Pedidos**: Usuários realizam pedidos
2. **Restaurantes → Produtos → Menu**: Restaurantes oferecem produtos organizados em seções de menu
3. **Pedidos → Produtos (Itens)**: Pedidos contêm vários itens
4. **Pedidos → Pagamentos → Recibos**: Pedidos geram pagamentos que produzem recibos
5. **Pedidos → Status**: Pedidos possuem status que mudam ao longo do ciclo de vida
6. **Motoristas → Pedidos → Rotas → GPS**: Motoristas entregam pedidos seguindo rotas com pontos GPS
7. **Usuários → Avaliações → Restaurantes**: Usuários avaliam restaurantes após pedidos
8. **Usuários → Suporte**: Usuários abrem tickets de suporte relacionados a pedidos
9. **Usuários → Buscas → Recomendações**: Interações dos usuários geram recomendações

Desafios de Integração

Ao trabalhar com esses datasets, você enfrentará desafios típicos de ambientes de dados distribuídos:

1. **Dados Fragmentados**: Informações relacionadas estão em diferentes sistemas
2. **Formatos Heterogêneos**: Cada fonte possui sua própria estrutura e formato
3. **Chaves de Referência Inconsistentes**: Diferentes nomes de chaves entre sistemas
4. **Dados em Tempo Real vs. Batch**: Combinação de processamento em lote e streaming
5. **Qualidade de Dados**: Valores ausentes, duplicados ou inconsistentes