UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE CIÊNCIAS MATEMÁTICAS E DE COMPUTAÇÃO

SCC0245 - PROCESSAMENTO ANALÍTICO DE DADOS (2025)

PROFª DR. CRISTINA AGUIAR

01 DE OUTUBRO DE 2025

Data Warehousing para Análise de CRM

Nome	**NUSP**
Gabriel Ribeiro Fonseca de Freitas	12542651
Henrique Souza Marques	11815722
Laura Fernandes Camargos	13692334
Leticia Barbanera Menezes	14588642

Conteúdo

1 Introdução

Nos últimos anos, a quantidade de informações geradas e armazenadas tem experimentado um crescimento significativo. A expansão das soluções tecnológicas para diversas áreas, aliada à crescente adoção dessas ferramentas, tem proporcionado uma abundância de fontes para a obtenção de dados. Exemplos disso incluem dispositivos como GPSs, computadores, celulares, alarmes e sensores, que, atualmente, funcionam essencialmente como ferramentas de captação de informações. Conforme divulgado na plataforma Statista, de acordo com o relatório da Comissão Europeia, todos os dias são criados cerca de 328,77 milhões de terabytes, ou 0,33 zettabytes, de dados. Isto equivale a cerca de 2,31 zettabytes por semana e 120 zettabytes por ano, o que ilustra a imensa escala da produção de dados. A tendência de crescimento foi tão marcante que surgiu a necessidade de cunhar um novo conceito: o termo ”Big Data”.

O conceito de Big Data engloba a crescente quantidade massiva de informações disponíveis, que incluem dados estruturados, semi-estruturados e não estruturados, gerados rapidamente — muitas vezes de forma instantânea. Contudo, Big Data não se resume apenas ao grande volume de informações. Esses dados apresentam características particulares que os tornam de grande interesse comercial. Além de seu grande volume, eles são gerados em alta velocidade, provenientes de uma vasta gama de fontes e em uma grande diversidade de formatos. Esse modelo é frequentemente descrito pelos ”3 Vs”: volume, velocidade e variedade.

Entretanto, à medida que essa quantidade de dados cresce, surgem desafios significativos. O armazenamento e a gestão dessa escala crescente de dados não podem ser realizados pelas tecnologias tradicionais de banco de dados. Ficou claro, portanto, que novos métodos e tecnologias precisariam ser desenvolvidos para possibilitar a extração de valor a partir desse imenso volume de informações.

Paralelamente a esse movimento, e influenciando-se mutuamente, consolidou-se a área de Ciência de Dados, que se especializa em gerar valor a partir dos dados. Essa tarefa reveste-se de grande relevância por diversos motivos, como a necessidade de tomar decisões mais assertivas e baseadas em evidências e a possibilidade de otimizar processos e reduzir custos em organizações, além de permitir a personalização de produtos e serviços para atender melhor os consumidores. Contudo, para que esse valor seja efetivamente extraído, é essencial que os dados estejam organizados de forma a viabilizar sua utilização adequada.

Como mencionado anteriormente, as fontes de informações são variadas. Entretanto, via de regra, o armazenamento dos dados nos chamados ambientes operacionais é realizado de maneira a otimizar tarefas e processos internos, os quais têm um foco distinto do objetivo analítico da Ciência de Dados. Nesse contexto, torna-se necessária a criação de um ambiente específico para o gerenciamento desses dados, com o propósito de garantir a eficiência de tarefas analíticas voltadas para a tomada de decisão.

Diante das urgências mencionadas, foi desenvolvida a tecnologia de Data Warehousing (DWing), associada ao sistema de armazenamento denominado Data Warehouse (DW). Uma questão que pode surgir é: por que não utilizar apenas as tecnologias de bancos de dados operacionais, mas em larga escala? De fato, é possível implementar um DW utilizando as tecnologias de bancos de dados tradicionais — o que é denominado implementação ou máquina ROLAP (Relational OLAP). No entanto, é importante ressaltar que as características de uso de ambos os sistemas são significativamente distintas, o que justifica o desenvolvimento de tecnologias específicas para DW, como os servidores MOLAP (Multidimensional OLAP).

Como mencionado, enquanto os bancos de dados operacionais são voltados para o processamento de transações OLTP (Online Transaction Processing), os Data Warehouses têm como foco o processamento de consultas OLAP (Online Analytical Processing). Isso implica que uma tecnologia operacional deve ser otimizada para realizar com eficiência operações frequentes de inserção, remoção e atualização. Em contrapartida, os ambientes informacionais são projetados para maximizar a eficiência das operações de leitura.

Devido a essas diferenças, os primeiros são caracterizados por um modelo normalizado, que garante propriedades como atomicidade, isolamento e, em última instância, consistência. Já os ambientes de

Data Warehouse são orientados aos tópicos do negócio que abordam, construindo uma perspectiva multidimensional dos dados. Além disso, enquanto as transações em um banco de dados operacional são, geralmente, pequenas e simples, acessando um número reduzido de registros, as consultas em um Data Warehouse são mais longas e complexas, pois exigem a correlação de diversas tabelas e o acesso a um grande número de registros. Outra distinção importante é que, enquanto os bancos de dados operacionais lidam com dados recentes e dinâmicos, os ambientes informacionais, como os Data Warehouses, lidam com dados históricos e estáveis, frequentemente associados ao próprio conceito de Big Data.

O Data Warehouse pode ser compreendido como uma central de dados – um repositório de grande porte, modelado e organizado de maneira a proporcionar fácil acesso e suporte eficaz a tarefas analíticas, o que, dada a quantidade de dados envolvidos, é de extrema importância. Por sua vez, o Data Warehousing engloba todo o sistema que alimenta o DW, incluindo a infraestrutura necessária para sua existência.

Diversas fontes de dados, provenientes de servidores distintos, alimentam o DW. Para que esses dados possam ser incorporados ao DW de maneira eficaz, é fundamental que estejam preparados para uso. Isso exige que os dados sejam integrados, homogeneizados, limpos e verificados adequadamente. O processo que abrange a extração dos dados de suas fontes originais, sua transformação para se adequar à estrutura do DW e garantir a correção das informações, e, por fim, o carregamento desses dados no ambiente informacional é denominado processo ETL (Extract, Transform, Load).

O presente trabalho tem como objetivo descrever o plano de implementação de um DW e de execução de DWing. Utilizando técnicas de modelagem multidimensional, apresentamos uma visão detalhada de como os dados serão relacionados e armazenados. Além disso, com base no conhecimento prévio dos perfis dos dados, descrevemos as tarefas a serem executadas, assim como as tecnologias que serão empregadas em cada etapa do processo, desde a extração das bases de dados heterogêneas nos servidores até o carregamento do DW.

2 Descrição do Problema

A análise de CRM (Customer Relationship Management) envolve uma série de técnicas e metodologias voltadas para a extração de insights valiosos a partir dos dados dos clientes de uma empresa. Ao examinar as tendências de mercado, o cenário competitivo e o feedback dos clientes, as organizações são capazes de identificar oportunidades de crescimento, mitigar riscos e alocar recursos de maneira estratégica. Adicionalmente, ao analisar as métricas de desempenho das campanhas e as taxas de resposta dos clientes, os profissionais de marketing podem aprimorar a segmentação, ajustar as mensagens e otimizar a alocação de canais, com o objetivo de maximizar o retorno sobre os investimentos. No mesmo contexto, ao observar o comportamento do cliente e as métricas de engajamento, as organizações têm a capacidade de identificar clientes em risco e implementar estratégias específicas para a retenção desses clientes. Além disso, ao analisar dados históricos de vendas e interações com clientes potenciais, as equipes de vendas podem identificar padrões, prever o desempenho futuro das vendas e alocar recursos de maneira mais eficiente.

A análise de CRM abrange desde a coleta de dados relacionados a qualquer interação entre o cliente e a marca, os quais são provenientes de diversas fontes, como dados demográficos, histórico de compras, comportamento de compra, rastreamento de uso do site e outras mídias de comunicação da marca, entre outros. Esses dados passam por um processo de integração, seguido de sua visualização e análise, com o objetivo de fornecer uma compreensão holística do funcionamento do negócio e embasar ações estratégicas. Essas ações podem incluir o refinamento de campanhas de marketing, a personalização das comunicações com clientes, a otimização de processos de vendas ou a melhoria no atendimento ao cliente.

Nesse contexto, e considerando a quantidade significativa de trabalho envolvida no processo ETL, especialmente em grandes corporações, é evidente que essas organizações podem se beneficiar

substancialmente da criação de um DW voltado para a análise de seus clientes e de suas vendas. Isso se deve ao fato de que, ao contar com dados já organizados e prontos para uso, os cientistas de dados podem direcionar seus esforços para o que realmente importa: extrair valor dos dados.

O presente trabalho, portanto, busca simular o desenvolvimento de um Data Warehouse (ou Data Mart, dependendo do escopo considerado) com o objetivo de apoiar análises de CRM.

3 Metodologia

3.1 Do DW

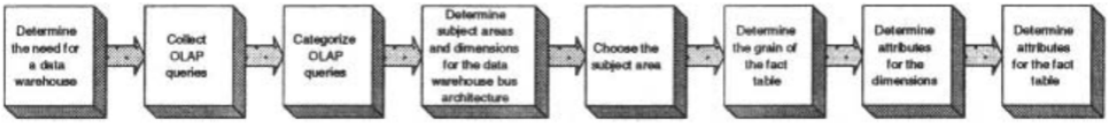

Figura 1: Diagrama representativo das etapas da metodologia adotada para a modelagem do DW

A metodologia de desenvolvimento do DW adotada neste trabalho foi baseada na estratégia proposta por Kimball. Inicialmente, foi realizada uma coleta de análises possíveis e perguntas que poderiam ser respondidas. Essas questões foram então categorizadas em classes, conforme o tópico abordado. Embora todas as questões estivessem, essencialmente, voltadas para o CRM, a análise poderia ser abordada sob diferentes perspectivas. Foram formuladas consultas relacionadas a vendas, ao mercado, às características do e-commerce, entre outras.

Em seguida, foi necessário avaliar os dados disponíveis e determinar o que poderia ser feito com eles. Como dependíamos de dados públicos sobre usuários da plataforma, os quais são, por questões legais relacionadas à Lei Geral de Proteção de Dados (LGPD), geralmente mantidos de forma privada, desconfiávamos que nem todas as análises planejadas seriam viáveis. Dessa forma, optamos por direcionar o foco do nosso DW para as vendas e o mercado da empresa. PAs consultas ”objetivo”, utilizadas neste trabalho, estão apresentadas na seção 6.

Com esse foco nas consultas principais, e com base no modelo relacional disponível na descrição da base de dados que utilizamos, foi possível determinar as dimensões e os fatos para a nossa modelagem. Em seguida, definimos a granularidade de cada fato, ou seja, o nível máximo de detalhe que seria armazenado, bem como os atributos das dimensões. Para isso, avaliamos as formas como os fatos poderiam ser agrupados e analisados. Inicialmente, focamos apenas em alocar os atributos disponíveis nas respectivas dimensões planejadas.

Em um segundo momento, consideramos a expansão desses atributos. Como os DW sacrificam eficiência de armazenamento em prol da eficiência nas consultas, a redundância se torna uma característica presente e até desejável nesses sistemas. Assim, adicionamos redundância onde julgamos necessário, sempre tendo em mente as consultas motivadoras definidas. Por exemplo, um atributo relacionado ao ”dia”de uma data poderia ser enriquecido com informações adicionais, como o dia do ano correspondente, o dia da semana e a identificação de se trata de um feriado ou data comemorativa.

Além disso, desenvolvemos atributos denominados ”derivados”, aqueles que poderiam ser calculados a partir dos dados sempre que uma consulta fosse realizada, mas que, devido à alta frequência de consultas — como os atributos relacionados aos Key Performance Indicators (KPIs) —, era desejável armazená-los de forma antecipada, prontos para uso. Um exemplo disso é o campo de ”atraso”. Essa informação poderia ser obtida facilmente comparando a data de entrega de uma compra com a data limite de entrega. No entanto, essa comparação seria mais custosa em termos de processamento do que simplesmente verificar se uma flag está ativa (1) ou inativa (0). Em grandes escalas, essa pequena diferença de processamento reflete em horas adicionais de carregamento nas consultas.

Por fim, a partir dos atributos das dimensões e das conexões entre os fatos e as dimensões, definimos os atributos das tabelas-fato, ou seja, os atributos centrais das análises. Novamente, esses atributos foram definidos tanto com base nos dados disponíveis quanto com base nos dados que poderiam ser calculados a partir deles.

3.2 Do DWing

O processo de DWing engloba o sistema completo de infraestrutura e procedimentos que viabilizam o fluxo de dados dos sistemas operacionais para o DW, garantindo que as informações estejam prontas e organizadas para a análise de CRM. O pilar desse sistema, como explicado anteriormente, é o processo ETL, que será executado em três fases distintas: Extração, Transformação e Carga.

3.2.1 Extração (Extract)

A fase de extração consiste na coleta dos dados a partir das bases de dados heterogêneas da Olist (os repositórios originais). Dada a natureza dos dados, que são fornecidos em arquivos estáticos do tipo CSV, o método de extração inicial será a leitura direta desses arquivos, simulando a extração dos sistemas operacionais.

- **Fontes:** As dez tabelas de dados brutos (Olist Customer, Order, Order Item, Order Payment, Order Review, Product, Seller, Geolocation, Closed Deals e Marketing Qualified Deals) serão utilizadas como a fonte primária de dados.
- **Frequência:** Será executada uma Carga Inicial Completa para popular o DW com todo o histórico de 2016 a 2018. Para fins de manutenção, o sistema será projetado para suportar cargas incrementais periódicas (por exemplo, diárias), focando apenas nos novos registros gerados após a última carga.

3.2.2 Transformação (Transform)

Esta é a fase mais crítica, onde os dados brutos são adaptados ao modelo multidimensional, conforme definido na seção 5.

Limpeza e Qualidade de Dados

Serão implementadas rotinas para:

- Tratar valores nulos e inconsistências, como preencher categorias de produtos faltantes ou uniformizar formatos de dados.
- Executar deduplicação para garantir a integridade das chaves naturais (*nk_costumer*, *nk_seller*, etc.).

Integração e Geração de Atributos

- **Cálculo de Distância:** A distância entre o cliente e o vendedor (*Dseller_customer_distance*) será calculada na fase de transformação, utilizando as coordenadas de latitude/longitude da *Dim_Geolocation* para a *Facts_Order_Itens*.
- **Geração de Chaves Substitutas (*Surrogate Keys*):** Serão criados identificadores únicos, inteiros e sequenciais (as chaves primárias *id* de cada dimensão), para substituir as chaves naturais (*nk_*) e otimizar as operações de *join* no DW.

- **Enriquecimento de Dimensões:** Serão calculados e inseridos atributos derivados em dimensões-chave para análises de CRM, como:
  - **Dim_Costumers:** Geração dos campos Recência (*recency*), Frequência (*frequency*) e a *flag* de Recorrência (*flag_recurrent*), obtidos através do histórico de compras do cliente.
  - **Dim_Orders:** Cálculo da *flag* Entrega Atrasada (`flag_latedelivery`) e derivação das medidas totais (como o valor total dos itens, do frete e do pedido) para otimizar as consultas na *Facts_Order_Itens*.

3.2.3 Carga (Load) e Tecnologias

Nesta etapa, os dados transformados são carregados nas tabelas-alvo do Data Warehouse.

Ordem de Carga e SCD O carregamento seguirá a ordem hierárquica do modelo, começando pelas tabelas de Dimensão para garantir que todas as chaves substitutas estejam disponíveis antes de carregar as tabelas Fato. Para a maioria das dimensões, será adotada a estratégia Tipo 1 para as Slowly Changing Dimensions (SCD), onde as alterações nos atributos sobrescrevem o valor anterior, pois a análise de CRM foca no estado atual do cliente.

Tecnologias Envolvidas O sistema de DWing será construído sobre uma arquitetura baseada em cloud para garantir escalabilidade, desempenho superior em consultas OLAP e alinhamento com as práticas atuais de mercado.

- **Plataforma de Data Warehouse (DW): Google BigQuery** O Google BigQuery será utilizado como o DW central. Escolhido por ser uma solução serverless e colunar, ele é otimizado para o processamento massivo e rápido de consultas analíticas, sendo ideal para suportar as análises de CRM complexas no volume de dados necessário. A arquitetura colunar maximiza a eficiência de leitura, que é o principal foco de um DW.
- **Processo de Transformação e Carga (ELT)**: Python/Pandas e SQL O processo de gestão dos dados seguirá o modelo ETL, aproveitando o poder de processamento do BigQuery para a etapa de transformação:
  - **Extração e Carga:** O Python, utilizando a biblioteca Pandas e conectores apropriados, será responsável por extrair os dados das fontes CSV e carregá-los diretamente no BigQuery, em uma área de staging (carregamento bruto).
  - **Transformação:** A maior parte da lógica de transformação será executada utilizando SQL, diretamente no BigQuery. Isso inclui: a geração das chaves substitutas (ids), a implementação da lógica SCD Tipo 1 e o cálculo dos atributos derivados complexos (`recency`, `frequency`, distância euclidiana). A utilização do SQL no BigQuery garante que o poder de processamento da plataforma seja aproveitado, otimizando o tempo de execução.
- **Visualização e Análise (Business Intelligence): Google Looker Studio** O Google Looker Studio será a ferramenta de Business Intelligence (BI) primária. Conectado diretamente ao BigQuery, ele permitirá a construção de dashboards e relatórios para as análises de CRM, facilitando o acesso rápido e intuitivo aos dados estruturados no esquema estrela.

4 Bases de Dados

A Olist é uma plataforma brasileira de marketplace que opera no segmento de e-commerce. A empresa atua como uma provedora de soluções tecnológicas, oferecendo sua plataforma como Software

como Serviço (SaaS). A Olist conecta pequenos e médios vendedores a consumidores em diversos segmentos de produtos, facilitando o processo de compra e venda por meio de sua plataforma online. A empresa disponibilizou, através da plataforma Kaggle, uma série de repositórios de dados próprios, que contêm mais de 100.000 registros referentes ao período de 2016 a 2018. Esses dados foram utilizados como fontes de informação para o desenvolvimento do DW deste trabalho.

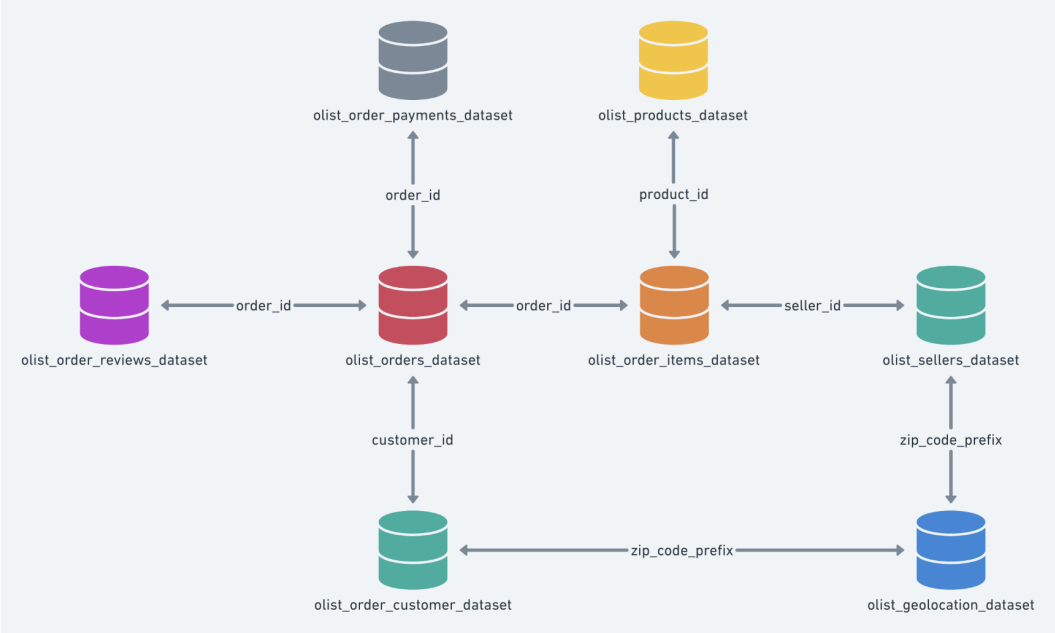

Figura 2: Esquema dos relacionamentos entre os repositórios fontes do DW

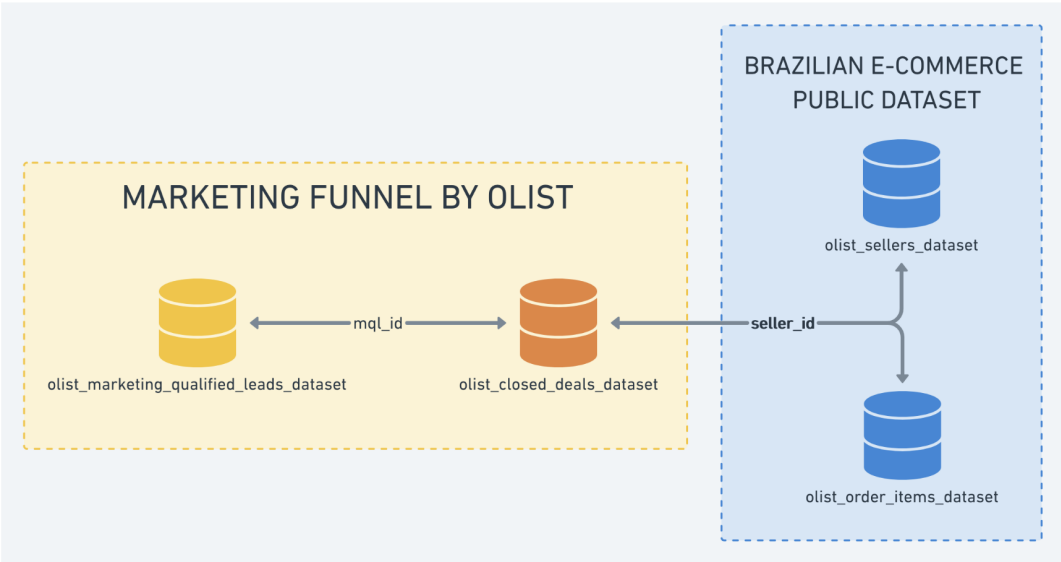

Figura 3: Outro esquema dos relacionamentos entre os repositórios fontes do DW

Foram modelados 3 processos principais do negócio: o cadastro de vendedores na plataforma, a venda de itens e a avaliação dos itens. A escolha desses processos foi fundamentada, além das restrições impostas pela LGPD, pelo seu papel essencial no contexto de CRM. O cadastro de vendedores cria uma base de dados estratégica, o monitoramento da performance dos vendedores e, assim, o prestígio da empresa dentre os clientes. O processo de venda de itens, por sua vez, possibilita o rastreamento do

comportamento de compra dos clientes, viabilizando a personalização das estratégias de vendas e a fidelização. Já a avaliação dos itens é crucial para entender a satisfação do cliente, identificar áreas de melhoria contínua nos produtos e serviços, e fortalecer o relacionamento com os consumidores. Os detalhes referentes à modelagem - como escolha de fatos e dimensões - será exibida na seção.

5 Modelagem

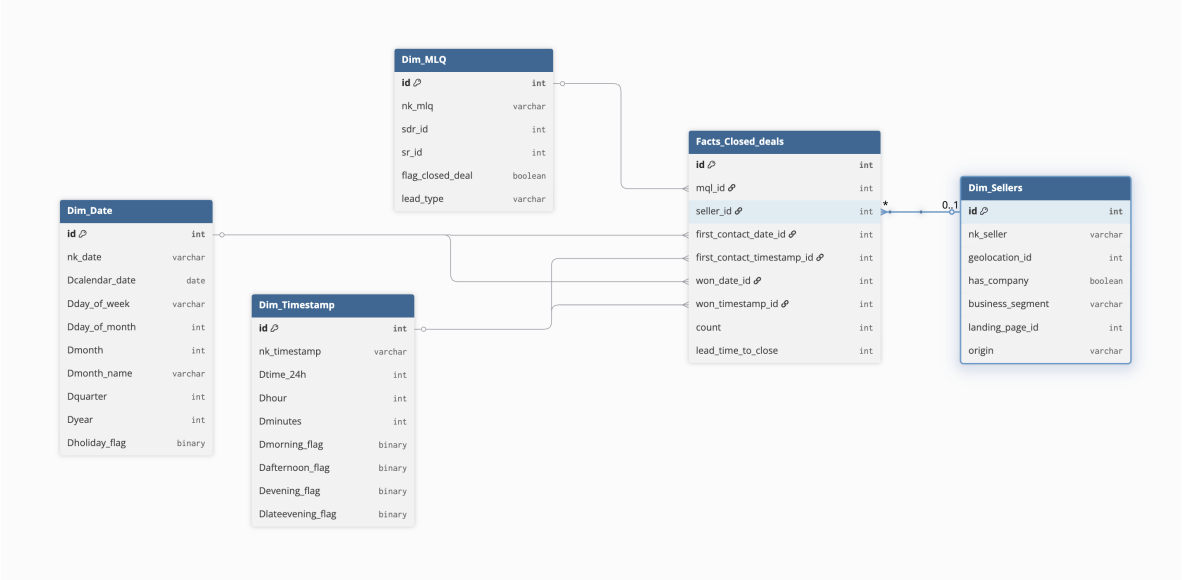

Figura 4: Esquema I

Em relação ao fato de vendedores (”Fact_Closed_Deals”), o grão adotado foi a unidade de lead que ingressava no funil da Olist, ou seja, cada tentativa de cadastro de uma entidade como vendedora fidelizada na plataforma. Este fato abrange as dimensões ”Dim_MLQ”, que distingue entre os leads convertidos e os que não foram fechados; ”Dim_Sellers”, uma vez que, a partir do momento em que um contrato é fechado, a entidade se torna um vendedor autorizado na plataforma; ”Dim_Date”e ”Dim_Time”, visto que várias datas são armazenadas e, consequentemente, fragmentadas em partes atômicas. Suas medidas numéricas são: contagem (1 unidade para toda tupla) e tempo transcorrido entre primeiro contato e fechamento do contrato, em dias. Trata-se de um fato não aditivo, pois, embora o primeiro atributo possa ser corretamente somado por meio das dimensões, o segundo geraria informações incorretas ou não significativas se manipulado dessa forma. Nesse caso, acreditamos que a agregação através da média aritmética, ou mediana, seja mais informativa.

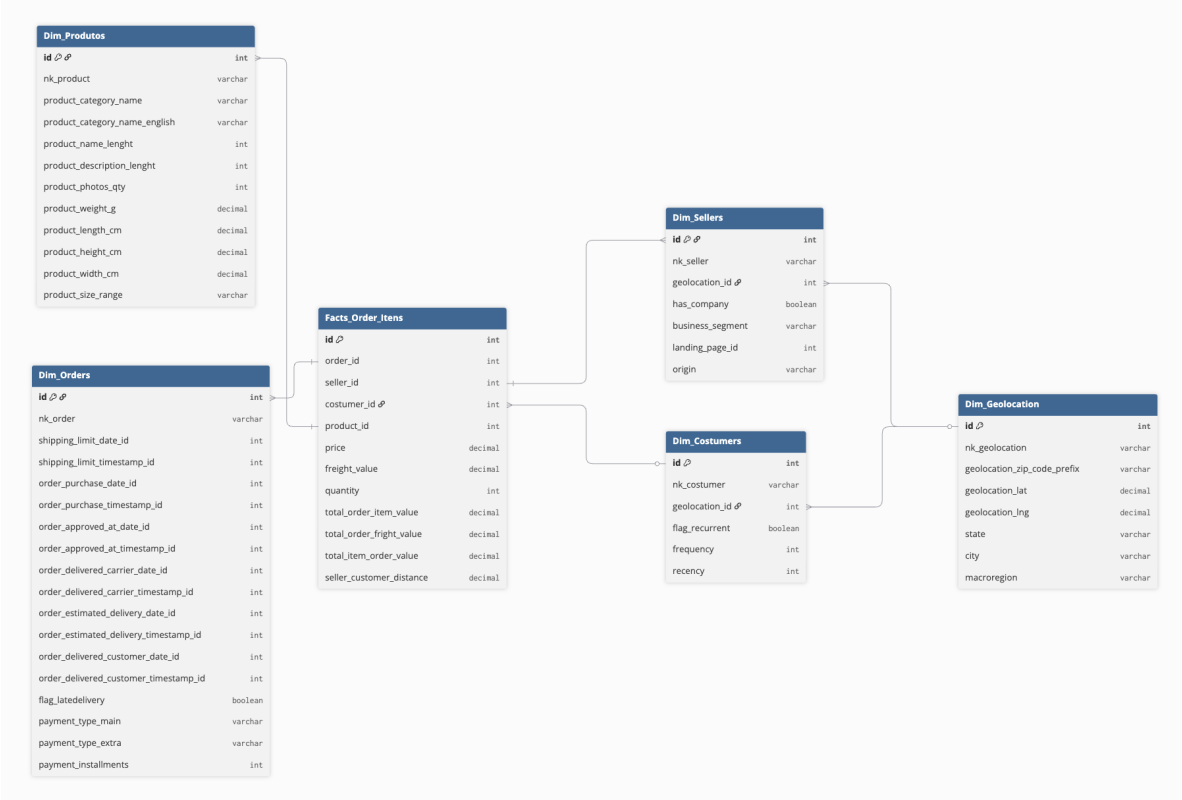

Figura 5: Esquema II

Em relação ao fato de itens vendidos (”Fact_Order_Items”), o grão adotado foi a unidade/objeto que compunha um pedido de um cliente. Este fato se relaciona com as dimensões ”Dim_Orders”, uma vez que cada item está vinculado a um pedido; ”Dim_Products”, já que cada item é categorizado dentro de uma hierarquia de classes de produtos na plataforma; ”Dim_Seller”e ”Dim_Customer”, pois um pedido é feito por um cliente a um vendedor (vale ressaltar que a Olist não permitia a associação de múltiplos vendedores em um único pedido; portanto, se um cliente desejasse comprar produtos de vendedores distintos, era necessário realizar compras separadas). Ele apresenta como medidas numéricas: preço por unidade, valor do frete por unidade, quantidade daquele produto incluída no pedido, valor total dos itens, valor total do frete, valor total do pedido com esses itens, e distância entre cliente e vendedor. Os últimos quatro atributos são derivados. O valor total dos itens é obtido pelo produto entre quantidade e preço unitário. O valor total do frete é obtido pelo produto entre a quantidade e valor do frete por item. O valor total do pedido com esse item é dado pela soma dos dois anteriores. Por fim, a distância é obtida como a distância euclidiana entre os pontos de latitude e longitude dos indivíduos. Os atributos desse fato são aditivos para todas as dimensões, uma vez que a soma simples de cada um deles resulta na agregação correta dos valores em níveis de granularidade progressivamente maiores. É importante destacar que, embora sejam referidos como ”totais”, esses valores não significam, necessariamente, o valor total final pago pelo usuário. Isso porque esses valores referem-se a somente um item (o especificado no id) mas um mesmo pedido pode incluir mais de um tipo de produto, desde que sejam do mesmo vendedor.

Figura 6: Esquema III

Por fim, em relação ao fato de avaliações de produtos (”Fact_Reviews”), o grão estabelecido foi uma avaliação deixada por um cliente em relação a um tipo de item adquirido em um momento específico. Por exemplo, caso um cliente realize uma compra de 5 unidades do produto A e 3 unidades do produto B, ele terá o direito de deixar uma avaliação para o produto A e uma avaliação para o produto B. Caso o cliente retorne no futuro e compre novamente os mesmos produtos, poderá avaliá-los novamente. Esse fato se relaciona com ”Dim_Seller”, ”Dim_Customer”e ”Dim_Orders”, pois, para que uma avaliação ocorra, é necessário que um cliente tenha feito um pedido a um vendedor. Além disso, está relacionado com ”Dim_Products”, pois a avaliação é baseada no tipo de produto, atributo dessa dimensão. Também se vincula a ”Dim_Date”e ”Dim_Time”, uma vez que múltiplas datas são armazenadas e exploradas por meio dessas dimensões. Este fato é não aditivo, pois o atributo numérico, referente à nota atribuída pelo cliente ao produto, de 0 a 5 estrelas, não é representativo caso seja somado diretamente. A agregação mais apropriada, nesse caso, é realizada por meio da média das notas.

Passa-se, então, para a descrição das dimensões. A **Dim_MLQ** possui o atributo *id*, que é assinalado de maneira automática e única para cada tupla. Esse identificador é específico para o DW em questão. A chave primária da tabela de dados original foi renomeada para uma chave natural *nk_mlq*. O *srd_id* representa o representante de desenvolvimento de vendas responsável pelo atendimento ao lead, e o *sr_id* representa o vendedor responsável pela tarefa. O *lead_type* é o tipo de produto a ser vendido ou a subárea de negócio, conforme autoinformado pelo lead. O *flag_closed_deal* diferencia entre os leads que se tornaram vendedores registrados na plataforma e aqueles que não fecharam contrato.

A dimensão **Dim_Sellers** contém atributos que descrevem os vendedores na plataforma. O *id* é o identificador único do vendedor, enquanto *nk_seller* representa a chave primária associada à tupla na base de dados original. O *geolocation_id* vincula o vendedor a uma localização específica, e o *has_company* indica se o vendedor é pessoa física ou jurídica. O *business_segment* classifica o vendedor segundo seu segmento de mercado, e o *landing_page_id* conecta o vendedor à sua página de vendas. Por fim, o *origin* descreve a origem do vendedor na plataforma.

A dimensão **Dim_Customers** descreve os atributos dos clientes registrados na plataforma. O *id* é o identificador único do cliente no DW, enquanto *nk_costumer* é sua chave primária na tabela fonte e *nk_costumer_unique* seu identificador único. A diferença é significativa: na base de dados original,

havia *costumer_id* e *costumer_unique_id*. O primeiro era único para cada tupla. O segundo era único para o cliente, de forma que poderia haver repetição caso o cliente realizasse mais de uma compra na plataforma. A mesma lógica foi importada para o DW. O *geolocation_id* vincula o cliente a uma localização específica, para a qual são enviadas suas compras. O atributo *flag_recurrent* indica se o cliente é recorrente, ou seja, se realizou ao menos uma compra nos últimos 3 meses. *Frequency* mensura a frequência exata das compras do cliente nesse espaço de tempo, enquanto *recency* avalia o tempo desde a última compra realizada. Esses atributos ajudam a caracterizar o comportamento do cliente, possibilitando a segmentação e análise de suas características. Os últimos três atributos serão obtidos a partir da análise do histórico de compras do cliente. Caso um cliente tenha realizado mais de uma compra, diferentes *ids* (e *nk_costumer*) estarão associados ao mesmo *nk_costumer_unique*. Para determinar a *flag*, basta verificar se há recompra. Para determinar a frequência, é preciso verificar as datas das compras e considerar apenas aquelas feitas nos últimos 90 dias. Para povoar o atributo de *recency*, obtém-se a última compra associada ao cliente e subtrai-se ela da data atual.

A dimensão **Dim_Geolocation** descreve as informações geográficas associadas aos vendedores e clientes. Ela permite associar um CEP a um ponto (latitude, longitude) de forma a mensurar a distância entre o vendedor e o cliente. O *id* é o identificador único da geolocalização no DW, enquanto *nk_geolocation* é sua chave natural. O *geolocation_zip_code_prefix* refere-se ao prefixo do código postal, e *Geolocation_lat* e *geolocation_lng* representam, respectivamente, as coordenadas de latitude e longitude desse ponto. Os atributos *state*, *city* e *macoregion* indicam, respectivamente, o estado, a cidade e a macrorregião em que a geolocalização está situada. O estado e macrorregião serão obtidos a partir de uma base externa que realize a associação correta entre a cidade e esses atributos.

A dimensão **Dim_Produtos** descreve os atributos dos produtos disponíveis na plataforma. O *id* é o identificador único do produto no DW, enquanto *nk_product* é a chave natural. O *product_category_name* e *product_category_name_english* indicam o nome da categoria do produto em português e em inglês, respectivamente. Os atributos *product_name_lenght* e *product_description_lenght* fornecem a quantidade de caracteres do nome e da descrição do produto. O *product_photos_qty* indica a quantidade de fotos associadas ao produto. As dimensões *product_weight_g*, *product_length_cm*, *product_height_cm* e *product_width_cm* fornecem informações sobre o peso e as dimensões físicas. Por fim, o *product_size_range* descreve a faixa de tamanhos disponível para o produto, entre pequeno, médio e grande, de forma a possibilitar uma consulta associada mais rápida. Ele será obtido através da multiplicação da largura, comprimento e altura do produto, e posterior categorização do resultado. Pretendemos utilizar a seguinte distinção: Pequeno (P) quando $V \leq 5{,}000$ cm^3, Médio (M) quando $5{,}000 < V \leq 20{,}000$ cm^3 e Grande (G) quando $V > 20{,}000$ cm^3.

A dimensão **Dim_Orders** descreve os atributos relacionados aos pedidos. O *id* é o identificador único do pedido, enquanto *nk_order* é sua chave natural. Os atributos relacionados às datas e horários dos pedidos incluem *shipping_limit_date_id*, *shipping_limit_timestamp_id*, *order_purchase_date_id*, *order_purchase_timestamp_id*, *order_approved_at_date_id*, *order_approved_at_timestamp_id*, *order_delivered_carrier_date_id*, *order_delivered_carrier_timestamp_id*, *order_estimated_delivery_date_id*, *order_estimated_delivery_timestamp_id*, *order_delivered_customer_date_id* e *order_delivered_customer_timestamp_id* , que registram as datas e horários em que o pedido foi feito, aprovado, enviado, estimado para entrega e realmente entregue ao cliente. O atributo *flag_latedelivery* indica se houve atraso na entrega. Os atributos *payment_type_main*, *payment_type_extra* e *payment_installments* fornecem informações sobre o método de pagamento, o tipo extra de pagamento (se houver) e o número de parcelas do pagamento. Aqui, o tipo de pagamento estava armazenado em apenas uma coluna na base. Como o cliente tinha a opção de escolher até duas formas de pagamento, optou-se por associar cada um a um atributo.

As dimensões **Dim_Date** e **Dim_Timestamp** armazenam informações temporais sobre o momento no ano e no dia em que ocorrem os eventos. A dimensão **Dim_Date** inclui atributos como o *id* (identificador único da data) e o *nk_date* (chave natural). O *Calendar_date* contém a data em formato padrão, enquanto *Day_of_week* fornece o nome do dia da semana e *Day_of_month* o número do dia do mês. Já o *Month* e *Month_name* representam, respectivamente, o número e o nome do mês, e o *Quarter*

indica o quadrimestre do ano. O *Year* armazena o ano e o *Holiday_flag* uma flag binária que identifica se a data corresponde a um dia próximo de feriado - o range definido foi de uma semana. Todos os atributos dessa dimensão serão extraídos da data em formato padrão disponível na base de dados de origem. Por sua vez, a dimensão **Dim_Timestamp** contém atributos como o *id* (identificador único do timestamp) e o *nk_timestamp* (chave natural). O *Time_24h* representa a hora no formato de 24 horas, enquanto *Hour* e *Minutes* armazenam, respectivamente, a hora e o minuto. As flags *Morning_flag*, *Afternoon_flag*, *Evening_flag* e *Lateevening_flag* indicam se o timestamp ocorre durante a manhã, tarde, noite ou madrugada, respectivamente. Todos os atributos dessa dimensão serão extraídos da hora em formato padrão disponível na base de dados de origem.

As comunicações entre as tabelas-fato e suas dimensões no DW são realizadas por meio das suas respectivas chaves primárias, que possuem o nome uniformizado como *id*. Foi adotada a estratégia de associar um valor único adicional a cada tupla, ao invés de "reutilizar" sua chave na base de que é proveniente, para assegurar sua unicidade no DW. No entanto, a chave primária original, chamada de "chave natural"por Kimball, não pode ser removida, pois ela é fundamental para identificar unicamente as tuplas durante operações como junções. As chaves primárias do DW serão geradas com um gerador de números sequenciais.

Na dimensão **Date**, os atributos seguem a hierarquia $Dyear > Dquarter > Dmonth > Dday\_of\_month$. Na dimensão **Timestamp**, temos a hierarquia $Dhour > Dminutes$, enquanto os atributos $morning\_flag, afternoon\_flag, evening\_flag$ e $lateevening\_flag$ derivam de $Dtime\_24h$. Na dimensão **Orders**, os atributos de data seguem a hierarquia *order_purchase_date_id* > *order_approved_at_date_id* > *order_delivered_carrier_date_id* > *order_delivered_customer_date_id*, pois esperamos que uma compra siga essa ordenação cronológica.

Para visualização da constelação de fatos, com todos os esquemas-estrela concatenados, acessar **modelagem**.

6 Consultas

Esta seção detalha as consultas que serão suportadas pela aplicação, alinhando as operações clássicas de Data Warehouse com as necessidades analíticas do e-commerce. As consultas incluem:

1. **Roll-Up/Drill-Down**, para análise hierárquica e agregações em diferentes níveis de granularidade;
2. **Slice and Dice**, para filtragens específicas sobre subconjuntos dos dados;
3. **Pivot**, que permite a reorganização de medidas e dimensões para facilitar comparações;
4. **Drill-Across**, a qual integra múltiplas tabelas fato para examinar correlações entre diferentes tipos de eventos, como as recomendações cruzadas entre assuntos;

As principais questões a serem investigadas com os dados disponíveis dizem respeito a vendas e receita.

6.1 Roll-up / Drill-down

- "Qual o percentual de pedidos entregues atrasados por bimestre em 2016? Como é o comportamento semanal no bimestre de maior percentual?"
  Esse conjunto de consultas realiza as operações de *slice*, ao restringir o atributo dos anos a um valor particular (2016), e de *drill-down*, pois primeiro investiga o objetivo em um nível de granularidade maior - bimestres - e depois enfoca em um deles, destrinchando-o em um nível de granularidade menor - semanas.
- "Como o churn rate (clientes que não voltaram a comprar após 3 meses) evoluiu (por bimestre) em 2018, por categoria de produto e segmento de cliente? Esse comportamento é semelhante ao

observado nos últimos 2 anos da empresa (por semestre)?"
Na respectiva consulta, utiliza-se *dice* em dois momentos: ao selecionar apenas os clientes que não voltaram a recomprar nos últimos meses e ao estabelecer um intervalo de anos de interesse (na segunda questão - "últimos 2 anos"). Além, realiza-se *slice* na primeira pergunta, pois estabelece o valor fixo para o ano (2018). O conjunto de consultas configura um *roll-up*, já que parte-se da granularidade menor, bimestres, na primeira questão, para uma maior, semestres, na segunda pergunta.

6.2 Slice and Dice

- "Qual foi o crescimento relativo de vendas em datas comemorativas, por comemoração (ex.: Black Friday, Natal) ao longo dos anos, por categoria de produto?"
  Essa consulta realiza operação *slice*, restringe-se às tuplas cuja flag de feriados está ativada (valor fixo = 1).

- "Qual o percentual de pedidos entregues atrasados por bimestre em 2016? Como é o comportamento semanal no bimestre de maior percentual?"
  Esse conjunto de consultas realiza as operações de *slice*, ao restringir o atributo dos anos a um valor particular (2016), e de *drill-down*, pois primeiro investiga o objetivo em um nível de granularidade maior - bimestres - e depois enfoca em um deles, destrinchando-o em um nível de granularidade menor - semanas.

- "Qual o tempo médio, em dias, entre a primeira compra e a segunda, de acordo com a origem do cliente?"
  Nesta, realiza-se *dice* implicitamente, uma vez que o tempo médio mencionado só pode ser calculado para aqueles clientes que realizaram pelo menos duas compras, ou seja, também restringe-se apenas aos clientes que realizaram mais que 1 compra na plataforma.

- "Como o churn rate (clientes que não voltaram a comprar após 3 meses) evoluiu (por bimestre) em 2018, por categoria de produto e segmento de cliente? Esse comportamento é semelhante ao observado nos últimos 2 anos da empresa (por semestre)?"
  Na respectiva consulta, utiliza-se *dice* em dois momentos: ao selecionar apenas os clientes que não voltaram a recomprar nos últimos meses e ao estabelecer um intervalo de anos de interesse (na segunda questão - "últimos 2 anos"). Além, realiza-se *slice* na primeira pergunta, pois estabelece o valor fixo para o ano (2018). O conjunto de consultas configura um *roll-up*, já que parte-se da granularidade menor, bimestres, na primeira questão, para uma maior, semestres, na segunda pergunta.

6.3 Pivot

- "Qual a performance de qualidade, ou seja, a satisfação do cliente de cada segmento de vendedor em relação a cada categoria de produto que eles oferecem?"
  Essa consulta realiza a operação *pivot*, pois a métrica de *review_score* é analisada simultaneamente em duas dimensões: segmento de vendedor e categoria de produto. O resultado assume forma matricial, em que os segmentos aparecem como linhas e as categorias de produto como colunas, preenchidas com a média das notas de satisfação. Essa consulta permite identificar os pontos positivos e negativos na percepção do cliente em cada nicho de atuação.

- "Quais produtos são frequentemente comprados em conjunto?"
  Esta consulta realiza a operação *pivot*, pois reorganiza os dados da *Fact_Order_Items* em formato matricial. O pivot ocorre ao cruzar a dimensão de produtos com ela mesma, permitindo visualizar combinações recorrentes. Esse resultado é útil para estratégias de *cross-selling* e gestão de

estoque. Além disso, o resultado pode orientar a criação de combos promocionais ou sugestões automáticas na plataforma de e-commerce.

6.4 Drill-Across

- Queremos saber se o preço do item influencia, de alguma forma, a quantidade de itens comprados (por exemplo, se colocar o preço de um item mais barato estimula mais o consumo). Para isso, queremos avaliar as relações entre o preço unitário do produto e o valor total de itens da compra, por categoria de produto. Nesta análise, realiza-se *drill-across*, uma vez que é explorada a relação entre o preço unitário do produto (que pertence à dimensão de itens) e o valor total dos itens da compra (que está na dimensão de pedidos).

- ”Qual a relação entre tempo de entrega e satisfação geral do cliente com a compra?”
  Esta consulta realiza a operação *drill-across*, pois cruza as métricas de *Fact_Orders*, que contém informações sobre prazos de entrega e custos logísticos e de *Fact_Review*, que armazena as avaliações dos clientes (*review_score* e comentários). Através dessa consulta é possivel ver verificar se tempos de entrega elevados resultam em notas menores de satisfação.

7 Referências

Material didático disponibilizado na plataforma da disciplina.

CUNNINGHAM, Colleen; SONG, Il-Yeol; CHEN, Peter P. Data warehouse design to support customer relationship management analyses. *Journal of Database Management*, v. 17, n. 2, p. 61–83, abr./jun. 2006.

SONG, Il-Yeol; LEVAN-SHULTZ, Kelly. Data warehouse design for e-commerce environments. Philadelphia: Drexel University, College of Information Science and Technology, [s.d.].