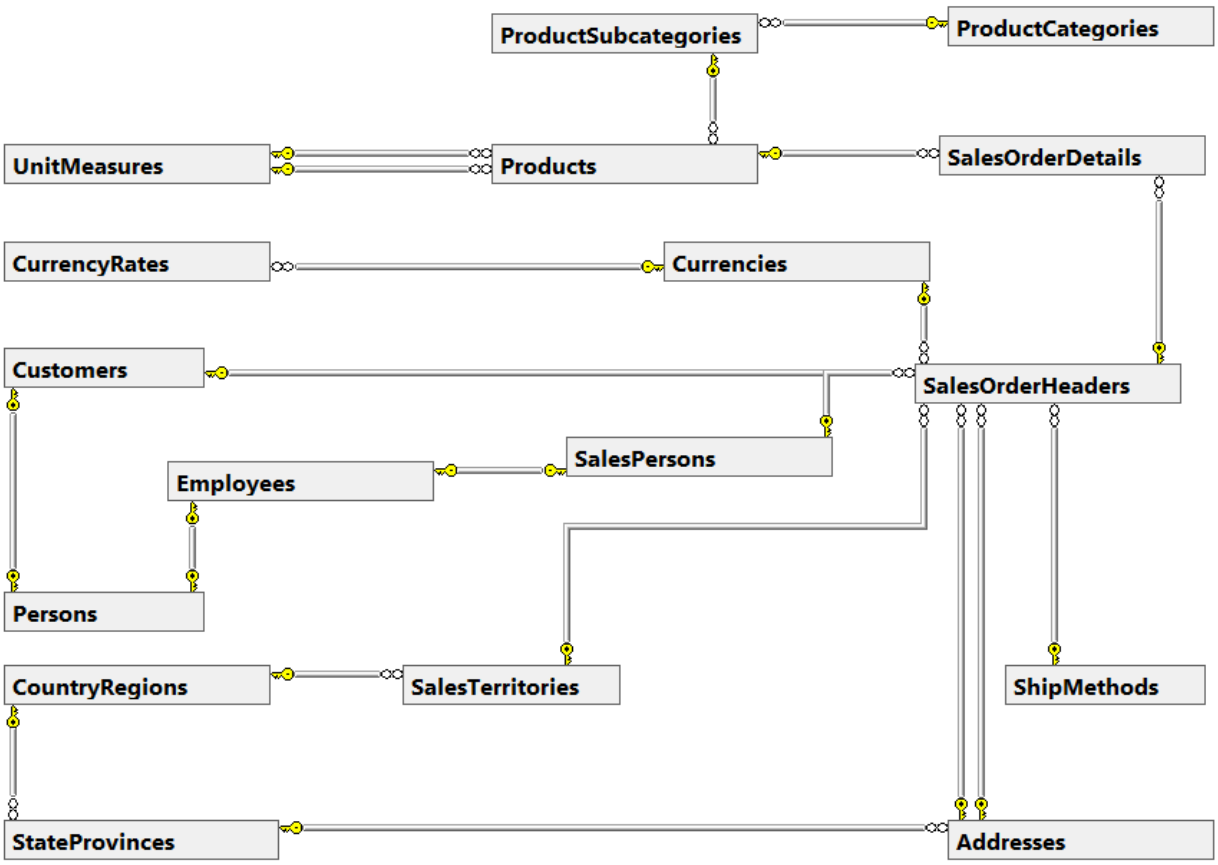
Descrição do problema

O presente trabalho tem por objetivo o desenvolvimento de um armazém de dados, com base nos dados de vendas a clientes, efetuadas por uma empresa nacional denominada *Bikes & Bikes*. O ramo de atividade desta empresa consiste na produção e venda de bicicletas e acessórios. Os clientes da empresa são originários dos mais diversos países, efetuando as suas encomendas através da *web*. Entre outras áreas, o sistema operacional da empresa suporta a área da gestão de vendas a clientes. Na figura seguinte apresenta-se o modelo de dados da parte do sistema operacional da empresa *Bikes & Bikes* referente a essa área.



O sistema operacional possui capacidades muito limitadas a nível da realização de análises de dados às vendas efetuadas aos clientes. Na sequência desta limitação, e também para permitir análises de cariz histórico, pretende-se que **desenvolva um *armazém de dados*** que permita a realização de **análises variadas às vendas efetuadas aos clientes**.

No desenvolvimento do armazém de dados há alguns aspetos a ter em consideração, relacionados com os dados que se encontram na base de dados do sistema operacional:

- Os dados referentes aos produtos encontram-se armazenados na tabela *Products*. O mesmo produto pode ter sido introduzido erradamente em duplicado, estando representado com identificadores diferentes (i.e., *ProductNumber*) na base de dados. Para além disso, o mesmo produto pode não ter sido introduzido de forma rigorosamente igual.
- Os valores monetários associados às vendas encontram-se registados na moeda do país do cliente. Esta informação encontra-se no atributo *CurrencyCode* da tabela *SalesOrderHeaders*. No caso de vendas efetuadas a clientes da zona Euro, é sempre usado o Euro (*EUR*).
- Os custos de expedição/frete (atributo *Freight*) e o imposto pago (atributo *TaxAmount*) encontram-se ao nível da venda (global) e não ao nível de cada linha individual (referente ao produto) que faz parte de cada venda.
- O sistema operacional regista sempre a data em que um registo é inserido numa tabela. Igualmente, sempre que ocorre uma alteração a um atributo de um registo, a respetiva data em que tal ocorreu também é registada.

Requisitos do trabalho

Nos pontos seguintes enumeram-se os requisitos que devem ser respeitados no trabalho a efetuar:

- Desenvolver um processo de análise dimensional, a fim de definir e criar o esquema conceptual para o armazém de dados, com base na informação anteriormente disponibilizada e seguindo a metodologia de *Kimball*. Todas as opções tomadas na criação do modelo dimensional devem ser devidamente justificadas no relatório a elaborar. O armazém de dados a desenvolver deverá ser concebido de modo a permitir a realização de consultas/análises **aos dados das vendas no nível de granularidade mais elementar**, de modo a possibilitar a maior flexibilidade possível para a realização futura de análises de dados.
- Proceder à extração, transformação, limpeza, integração e carregamento dos dados no armazém de dados, por intermédio de um *Integration Services Project do Visual Studio* e dos componentes mais adequados que este disponibiliza para as referidas tarefas. Mais especificamente, pretende-se que:

- Todo o processo de extração, transformação, limpeza, integração e carregamento dos dados deve ser o mais eficiente possível. Por questões de confidencialidade relacionadas com o negócio, foi disponibilizada para a construção do armazém de dados uma pequena amostra, já antiga, dos dados que se encontram no sistema operacional. No entanto, o volume de dados é muito elevado.
- Os dados podem estar afetados por problemas de qualidade (por exemplo: valores em falta em atributos de preenchimento obrigatório; valores que violam o domínio do atributo; valores que constituem violações a restrições de integridade/regras de negócio). Para a identificação dos problemas que possam existir sugere-se a realização de um processo de *Data Profiling* prévio sobre os dados.
- O processo de transformação e limpeza deve filtrar todos os registos que apresentem problemas de qualidade de dados. Assim, qualquer registo afetado por problemas de qualidade, cuja correção não seja possível de efetuar de forma imediata e automática, não deve ser carregado no armazém de dados, sendo armazenado em tabelas especialmente criadas para o efeito na *Staging Area*. A correção destes problemas e posterior carregamento desses registos para o armazém de dados fica excluída do âmbito deste trabalho.
- A fase de limpeza de dados deve ser efetuada de modo a eliminar todas as redundâncias (registos duplicados iguais ou aproximadamente iguais) que possam existir nos dados, oriundos do sistema operacional.
- O carregamento dos dados no armazém de dados (tabelas de dimensões e de factos) deve ser concebido de modo a poder ser executado de forma incremental, isto é, em cada execução só serem carregados os dados novos e atualizados os que já existem, caso estes tenham sofrido alterações.
- As análises de dados às vendas são apenas realizadas na unidade monetária local em que ocorreram (i.e., análises envolvendo uma só moeda) ou na unidade monetária de uniformização (i.e., análises envolvendo várias moedas). Sendo uma empresa nacional, a unidade monetária de uniformização é o Euro (Eur).
- O *Integration Services Project* deve ser elaborado de modo a não existirem caminhos (*paths*) absolutos nos diversos componentes. Todos estes caminhos têm de ser configuráveis mediante a sua especificação num ficheiro de parametrizações, assim como: nome do servidor de base de dados da empresa *Bikes & Bikes*; nome da base de dados da empresa *Bikes & Bikes*; nome do servidor de base de dados da *staging*

area e do armazém de dados (mesmo servidor para ambas); nome da base de dados da *staging area*; e, nome da base de dados do armazém de dados.

- Proceder à elaboração de 20 análises multidimensionais de dados sobre o armazém de dados criado. Estas análises dimensionais devem ser efetuadas tendo por base um cubo de dados resultantes de um *Analysis Services Project* do *Visual Studio*. Das 20 análises de dados, 10 são realizadas "à escolha", mas devem possuir graus de dificuldade diferentes e têm de incidir sobre as hierarquias que existam nos dados. As restantes 10 análises de dados encontram-se apresentadas de seguida, representando análises típicas que a gestão da empresa pretende efetuar às vendas a clientes. Note-se que estas 10 análises são meramente indicativas, pelo que o *data mart* não pode ser criado unicamente para lhes dar resposta.

    1. Valores totais (incluindo frete e imposto) em dólares australianos (AUD) referentes às vendas efetuadas nesta moeda, no primeiro semestre de 2012, detalhados por vendedor e por categoria do produto.
    2. Valores totais dos fretes suportados no transporte dos produtos vendidos durante o mês de dezembro de 2012, detalhados por método de envio e por subcategoria de produto.
    3. Valores totais dos impostos referentes às vendas realizadas durante o ano de 2014, com possibilidade de análise detalhada (i.e., *drill down*) ao nível do semestre, trimestre e mês, detalhados por tipo de cliente (atributo *PersonType*).
    4. Valores totais (incluindo frete e impostos) e respetivas quantidades das vendas efetuadas a clientes, por cada mês do ano de 2013, com possibilidade de análise agregada (i.e., *roll up*) ao nível do trimestre, semestre ou ano.
    5. Valores totais das vendas a clientes (sem incluir frete e imposto) efetuadas no primeiro trimestre de 2013, detalhados por vendedor e pela categoria do produto, com possibilidade de análise detalhada (i.e., *drill down*) ao nível da subcategoria e do produto.
    6. Valores totais dos impostos em libras (GBP) referentes às vendas realizadas durante a primavera e verão de 2013 efetuadas nesta moeda, detalhados pela subcategoria de produto, com possibilidade de análise agregada (i.e., *roll up*) ao nível da categoria.
    7. Valores totais referentes aos descontos praticados sobre o preço unitário de venda durante o ano de 2013, com possibilidade de análise detalhada (i.e., *drill down*) ao

nível do semestre, trimestre e mês, detalhados pelos territórios de venda, com possibilidade de análise agregada (i.e., *roll up*) ao nível do país/região.

8. Quantidades vendidas a clientes por vendedor e por cidade da morada de expedição, com possibilidade de análise agregada (i.e., *roll up*) ao nível do estado/província, para as vendas expedidas no último dia de cada mês do ano de 2013.
9. Valores totais (incluindo fretes e impostos) das vendas a clientes por cidade da morada de faturação e por categoria de produto, com possibilidade de análise detalhada (i.e., *drill down*) ao nível da subcategoria de produto, para o ano de 2013.
10. Valores totais das vendas (sem incluir frete e imposto) por moeda e por território, com possibilidade de análise agregada (i.e., *roll up*) ao nível do país/região do cliente, no terceiro quadrimestre de 2013.

A realização do trabalho é feita em **grupos de dois alunos** e envolve duas **partes complementares**.

Na **1ª parte** pretende-se que seja elaborado: arquitetura do armazém de dados para a situação descrita; modelo dimensional subjacente (com a apresentação dos atributos e respetivos tipos de dados para cada dimensão e tabela de factos); estruturas de dados (i.e., tabelas; ficheiros) a criar na *staging area*; mapeamento de dados entre os sistemas fonte, a *staging area* e o armazém de dados (o que inclui que para os atributos das dimensões seja apresentada a estratégia de *Slowly Changing Dimension* (SCD); e, eventuais operações de transformação e limpeza de dados a realizar nos diversos atributos). Estes artefactos devem resultar na elaboração de um relatório.

Na 2ª parte do trabalho pretende-se a implementação do armazém de dados, com a correspondente implementação dos processos de extração, transformação, limpeza, integração e carregamento dos dados. Esta parte deve ser documentada através da realização de um relatório final no qual constem todos os elementos relevantes para a avaliação final do trabalho, como: arquitetura do armazém de dados (final); modelo dimensional (final); estruturas de dados criadas na *staging area* (final); mapeamento de dados entre os sistemas fonte, a *staging area* e o armazém de dados (final); processos de extração, transformação, limpeza, integração e carregamento de dados efetuados; *scripts* SQL criados; análises dimensionais efetuadas mediante a apresentação das respetivas consultas; justificação das opções tomadas; melhoramentos possíveis; etc.

Prazo e Forma de Entrega da 1ª parte do Trabalho:

- O **relatório** tem de ser **submetido no Moodle** até às **23h59** do dia **3 de dezembro de 2023**.
- O trabalho deve ser submetido sob a forma de um **único ficheiro PDF**, com o seguinte nome: **XXXXXXX_YYYYYYY.pdf**, em que XXXXXXX e YYYYYYY representam os **números dos alunos** que constituem o grupo.

Prazo e Forma de Entrega da 2ª parte do Trabalho:

- **O Trabalho** (relatório final + *integration services project* + *analysis services project*) tem de ser **submetido no Moodle** até às **23h59** do dia **7 de janeiro de 2024**.
- O trabalho deve ser submetido sob a forma de um **único ficheiro ZIP**, com o seguinte nome: **XXXXXXX_YYYYYYY.zip**, em que XXXXXXX e YYYYYYY representam os **números dos alunos** que constituem o grupo.
- A discussão do trabalho será efetuada com **ambos os elementos do grupo obrigatoriamente presentes**, na semana de 8 a 12 de janeiro de 2024, em dia e hora a combinar oportunamente com cada grupo.

O **incumprimento do prazo de entrega** implica uma **penalização na nota**, dessa parte do Trabalho, **de 20% por cada dia de atraso.**