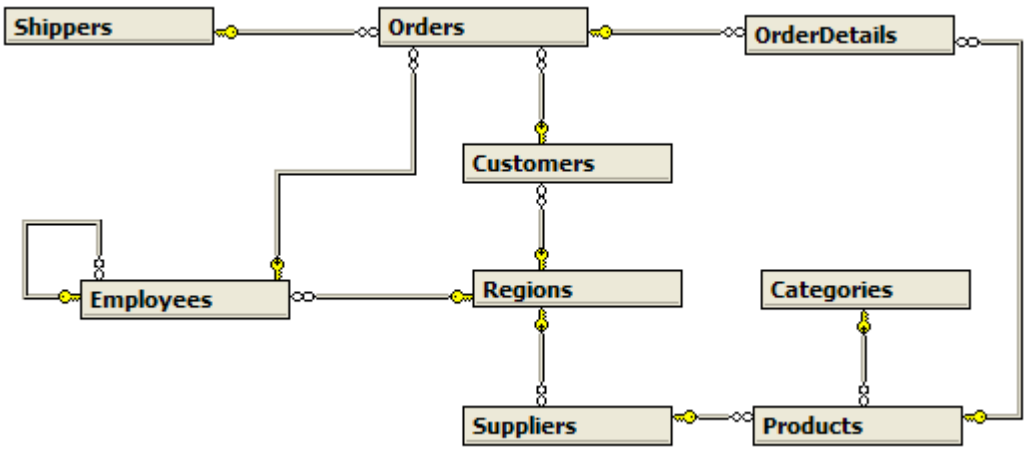
Descrição do Problema

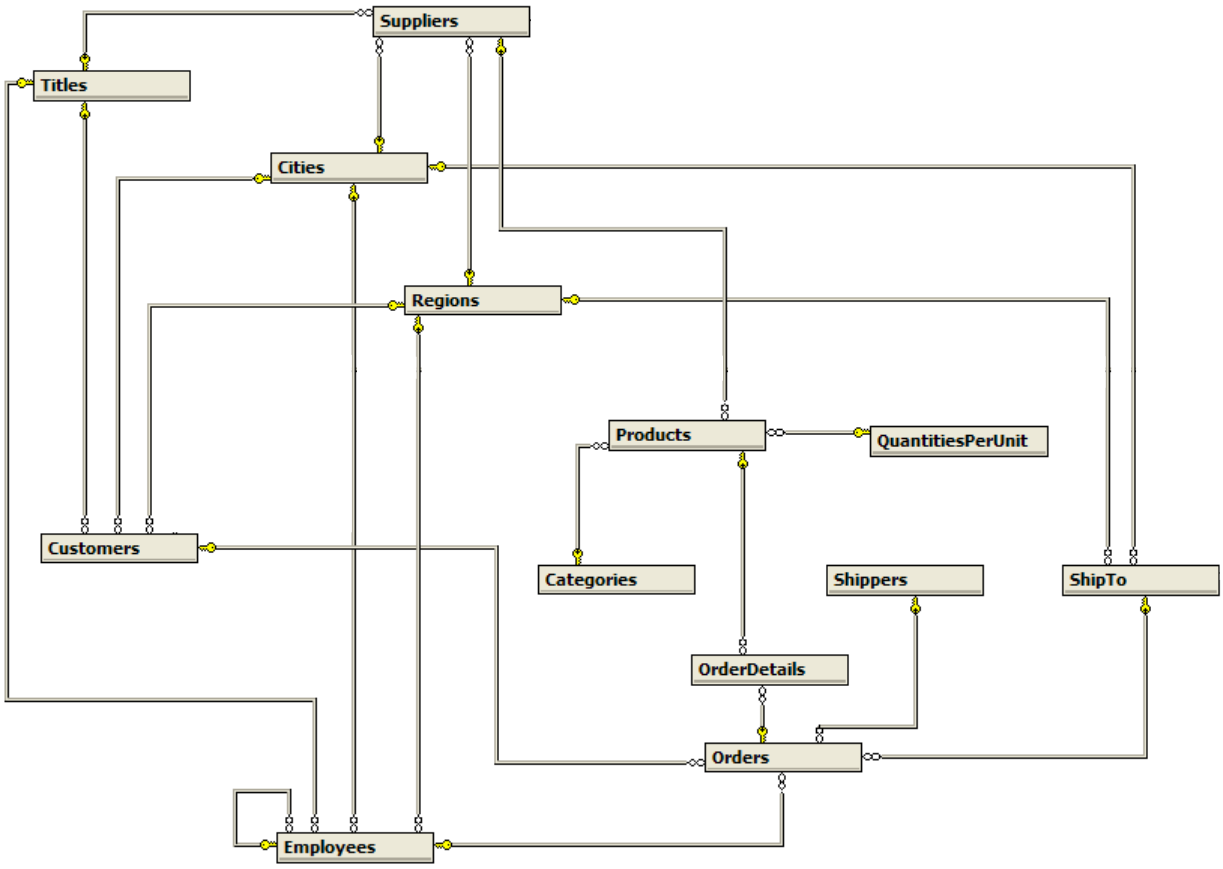
O presente trabalho tem por objetivo o desenvolvimento de um armazém de dados, a partir dos dados de encomendas de produtos, efetuadas pelos clientes de uma empresa de retalho que se dedica à comercialização de bens alimentares. A empresa em questão encontra-se sediada nos Estados Unidos da América (local de fundação), possuindo uma delegação em Inglaterra (aberta posteriormente). No momento da abertura da sede foi desenvolvida uma aplicação informática com o propósito de suportar a gestão das encomendas dos clientes. O modelo de dados inerente a essa aplicação informática encontra-se apresentado na figura seguinte.



Uma vez que a aplicação informática respetiva apresenta algumas deficiências e que o suporte fornecido pela empresa que a desenvolveu nem sempre é o melhor, aquando da abertura da delegação em Inglaterra, foi decidido encomendar o desenvolvimento de uma outra aplicação informática. Como o objetivo é o mesmo, isto é, fornecer um suporte automatizado às encomendas dos clientes, os dados manipulados nesta nova aplicação são, sensivelmente, os mesmos. No entanto, estes mesmos dados encontram-se organizados de forma diferente. Na figura da página seguinte apresenta-se o modelo de dados inerente à nova aplicação informática desenvolvida.

Face a esta realidade, os dois sistemas operacionais funcionam isoladamente, o que não permite a realização de análises conjuntas aos dados existentes em ambos. Na sequência desta impossibilidade, a gestão da empresa pretende que se **desenvolva um armazém de dados** que

permita a realização de diversas **análises variadas às encomendas efetuadas** (sejam estas à sede, à delegação, ou a ambas).



No desenvolvimento do armazém de dados há alguns aspetos a ter em consideração, relacionados com os dados que se encontram em cada uma das fontes de dados (ficheiros ou base de dados):

- Os dados do sistema operacional da sede encontram-se armazenados sobre a forma de ficheiros binários indexados. Uma vez que não há forma de aceder diretamente a estes dados (e.g., via OLEDB ou ADO.NET), a empresa que desenvolveu a aplicação concebeu uma funcionalidade que coloca todos esses dados sob a forma de ficheiros de texto (no caso, ficheiros *csv*).
- Os dados que se encontram nas tabelas/ficheiros *Categories*, *Region* e *Shippers* são exatamente os mesmos. Isto acontece em virtude de, no momento de entrada em funcionamento do sistema operacional da delegação, terem sido colocados nestas tabelas os mesmos dados que existiam nos ficheiros do sistema operacional da sede. Por questões de

simplificação, assume-se que estes dados se manterão sempre sincronizados entre os dois sistemas operacionais.

- Há exceção do preço unitário do produto, que no sistema operacional da sede se encontra em dólares e no sistema operacional da delegação se encontra em libras, e das unidades existentes em stock, todos os restantes dados existentes sobre os Produtos são exatamente iguais. O motivo pelo qual isto acontece é o mesmo do que o referido no ponto anterior. Igualmente, por questões de simplificação, assume-se que os dados iguais se manterão sempre sincronizados automaticamente entre ambos os sistemas operacionais.
- O conjunto de clientes que se encontram na tabela *Customers* e no ficheiro *Customers* são diferentes. No entanto, um determinado cliente pode estar presente em ambas as fontes de dados, o que significa que efetuou, ao longo do tempo, encomendas à sede e à delegação. Não é garantido que a forma como o mesmo cliente se encontra representado em cada uma das fontes seja rigorosamente igual.
- Existem diferenças ao nível do formato de representação usado em certos atributos entre tabelas/ficheiros análogos. No sistema operacional da sede, as datas encontram-se representadas sob o formato *mm-dd-aaaa*, enquanto no sistema operacional da delegação, as datas encontram-se representadas sob a forma *aaaa-mm-dd*.
- Existem diferenças de representação a nível do número de atributos usado para armazenar os mesmos dados. Na tabela *Employees* do sistema operacional da delegação são utilizados os atributos *TitleOfCourtesy*, *FirstName* e *LastName*, enquanto no ficheiro *Employees* todos estes dados se encontram representados num único atributo: *Name*.
- Os dados referentes às encomendas são totalmente diferentes. No entanto, o número identificador único de cada encomenda segue a mesma convenção de representação em ambos as fontes de dados, ou seja, numeração automática iniciada em 1 e com incremento de 1.
- Nas tabelas *Customers*, *Suppliers* e *ShipTo* do sistema operacional da delegação, o país encontra-se representado pelo respetivo código (e.g., PT para Portugal), ao invés do que acontece no sistema operacional da sede, em que o país surge por extenso.
- Existem diferenças ao nível das unidades monetárias utilizadas. No sistema operacional da sede, os valores monetários encontram-se em dólares, enquanto no sistema operacional da delegação, os valores monetários encontram-se em libras.

- O sistema operacional da delegação regista sempre a data em que um registo foi inserido numa tabela. Igualmente, sempre que ocorre uma alteração a um atributo, a respetiva data é também registada. No caso do sistema operacional da sede, nenhuma das duas datas é registada.

Requisitos do trabalho

Nos pontos seguintes enumeram-se os requisitos que devem ser respeitados pelo 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 atrás disponibilizada e seguindo a metodologia de *Kimball*. Todas as opções tomadas na criação do modelo dimensional devem ser devidamente justificadas. O armazém de dados a desenvolver deverá ser concebido de modo que permita a realização de consultas/análises **aos dados das encomendas no nível de granularidade mais detalhado/elementar**, de modo a possibilitar a maior flexibilidade futura possível.

- 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 extracção, integração, transformação/limpeza e carregamento dos dados deve ser o mais eficiente possível. Por questões de confidencialidade relacionadas com o negócio, apenas foi disponibilizada para a construção do armazém de dados uma pequena amostra dos dados que se encontram em ambos os sistemas operacionais. No entanto, o volume de dados em ambos os sistemas é muito elevado.
    - Os dados podem estar afetados por problemas de qualidade (por exemplo: valores em falta em atributos que era suposto de preenchimento obrigatório; valores que violam o domínio do atributo). 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 integração de dados deve ser efetuada de modo a eliminar todas as redundâncias (registos duplicados iguais e/ou aproximadamente iguais) que possam existir nos dados, oriundos de cada sistema operacional.
- O carregamento dos dados no armazém de dados (tabelas de dimensões e tabela 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 tenham sofrido alterações.
- Sendo uma empresa norte-americana, a unidade monetária a considerar para realização de análises globais às encomendas (i.e., envolvendo simultaneamente encomendas à sede e à delegação) é o dólar.
- O *Integration Services Project* deve ser elaborado de modo que não existam 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 sucursal; nome da base de dados da sucursal; nome do servidor de base de dados do armazém de dados e da *staging area* (mesmo servidor para ambas); nome da base de dados da *staging área*; 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 algumas têm de incidir sobre as hierarquias que existem nos dados. As restantes 10 análises de dados encontram-se apresentadas de seguida, representando análises típicas que a gestão da empresa pretende efetuar às encomendas de clientes. Note-se que estas 10 análises são meramente indicativas, pelo que o armazém de dados não pode ser criado unicamente para lhes dar resposta.

1. Valores totais com desconto das encomendas efetuadas no primeiro semestre de 2021, com possibilidade de análise detalhada (i.e., *drill down*) ao nível do trimestre e do mês, detalhados por país do cliente e por categoria do produto.
2. Valores totais dos fretes (em libras) despendidos no transporte dos produtos durante o primeiro trimestre de 2022, detalhados por transportador e por país de expedição, mas apenas para encomendas efetuadas à delegação.
3. Quantidades de unidades encomendadas no ano de 2020, com possibilidade de análise detalhada (i.e., *drill down*) ao nível do semestre e do trimestre, detalhadas pelos funcionários que as processaram e por cidade de expedição, mas apenas para os funcionários da sede.
4. Valores totais com e sem desconto das encomendas efetuadas, durante o ano de 2021, com possibilidade de análise detalhada (i.e., *drill down*) ao nível do semestre, trimestre e mês, detalhados por fornecedor do produto e pela respetiva categoria.
5. Valores totais dos descontos efetuados no último dia de cada mês do ano de 2021, detalhados por categoria de produto, com possibilidade de análise detalhada (i.e., *drill down*) ao nível do produto, mas apenas para encomendadas efetuadas à sede.
6. Valores totais com desconto (em libras) no mês de abril de 2022, referentes às encomendas efetuadas à delegação, detalhadas por cidade do cliente, com possibilidade de análise detalhada (i.e., *drill down*) ao nível do cliente.
7. Valores totais sem desconto e respetivos valores dos descontos referentes às encomendas ocorridas no 2º quadrimestre de 2021, detalhadas por país de expedição e por produto, com possibilidade de análise agregada (i.e., *roll up*) ao nível do fornecedor do produto.
8. Valores totais com desconto e respetivas quantidades encomendadas durante a primavera e verão de 2021, detalhadas pelos funcionários que as processaram e pelos transportadores que as levaram até aos clientes.
9. Valores totais dos fretes por cliente, com possibilidade de análise agregada (i.e., *roll up*) ao nível da região do cliente, e por mês do ano de 2022, com possibilidade de análise agregada (i.e., *roll up*) ao nível do trimestre e do semestre, apenas dos produtos que pertencem às categorias "Beverages", "Confections", "Grains/Cereals" e "Produce", encomendados à sede.

10. Valores totais sem desconto (em dólares) das encomendas efetuadas na 1ª semana de cada mês de 2021 à delegação, dos produtos que pertencem às categorias "Condiments", "Dairy Products", "Meat/Poultry" e "Seafood", detalhadas por fornecedor, com possibilidade de análise detalhada (i.e., *drill down*) ao nível do produto, e por cidade do cliente, com possibilidade de análise detalhada (i.e., *drill down*) ao nível do cliente.

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. **Do relatório, também, deve constar o plano de trabalho, detalhado com as diversas atividades/tarefas desenvolvidas e respetivo(s) elemento(s) do grupo encarregue(s) da sua realização**.

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. **Do relatório, também, deve constar o plano de trabalho, detalhado com as diversas atividades/tarefas desenvolvidas e respetivo(s) elemento(s) do grupo encarregue(s) da sua realização**.

Prazo e Forma de Entrega (1ª parte do Trabalho)

- O **relatório** tem de ser **submetido no Moodle** até às **23h59** do dia **24 de novembro de 2024**.
- 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 (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 **5 de janeiro de 2025**.
- 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 **os elementos do grupo obrigatoriamente presentes**, em dia e hora a combinar oportunamente com cada grupo.

Observações Finais

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

- Em caso de deteção de compartilhamento de trabalho entre grupos, seja total ou parcialmente, todas as partes serão penalizadas com a classificação de zero.

- Casos de apropriação ilícita de trabalho, seja total ou parcialmente, serão penalizadas com a classificação de zero.