Tabela de conteúdos

Metodologia

Nesta secção serão descritas todas as opções tomadas pelo grupo, seguidas de explicações, gráficos e imagens que os descrevam.
O software *BI server* e *Data Integration* foi instalado pelos membros do grupo em sistemas de base UNIX, nomeadamente Macintosh e Linux, como tal o processo de instalação poderá diferir relativamente aos *howtos* fornecidos pelo docente.

Modelo multidimensional - *Steelwheels*

Inicialmente foi desenvolvido um modelo multidimensional, baseado numa estrutura em estrela, resultado de uma simplificação da amostra *Steelwheels* explorada previamente na componente prática. Para tal, foram analisados os factos e dimensões da amostra, obtendo três dimensões distintas, unidas através de uma tabela de factos. Consoante o enunciado, foi selecionada uma dimensão relativa à localização geográfica. O modelo desenvolvido está presente na figura 1 que se segue.

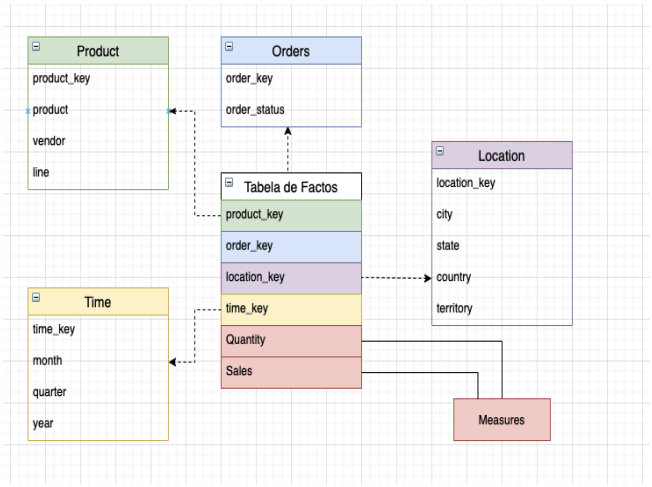


Illustration 1: Modelo Simplificado Steelwheels

Esquemas relacionais

Seguidamente, foram construídos 6 modelos de esquema relacional, resultantes de uma simplificação do modelo desenvolvido na secção anterior. Para tal foram exportadas quatro estruturas de dados do software Pentaho Data Integration, nomeadamente: (i) product; (ii) time; (iii) customer; (iv) orders. Adicionalmente foi introduzida uma tabela de factos, no centro do modelo, constituído pelas chaves estrangeiras das tabelas (i), (ii), (iii) e (iv), tal como por métricas de avaliação destas dimensões.

De acordo com o enunciado, foram retiradas algumas referências geográficas destas tabelas, tal como o estado e país, pois cada modelo representa os dados de um dado país. Adicionalmente foram realizadas alterações nos nomes de atributos (idiomas diferentes), valores monetários (EURO / DOLAR..), separação de atributos, entre outros. Os países utilizados para a criação destes modelos foram: (i) EUA; (ii) Espanha; (iii) França; (iv) Austrália; (v) Nova Zelândia; (vi) Reino Unido.
Segue-se a figura 2 que representa a versão simplificada do ER da BD original e as variações criadas para cada um dos restantes países.

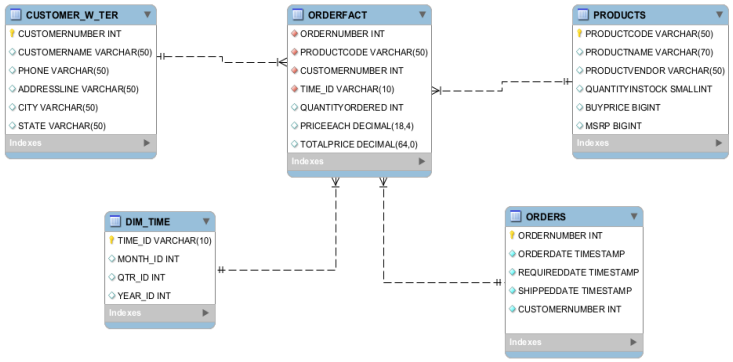


Illustration 2: ER - Tabela base

As seguintes estruturas demonstram os restantes países, seguido de uma breve explicação das alterações realizadas.

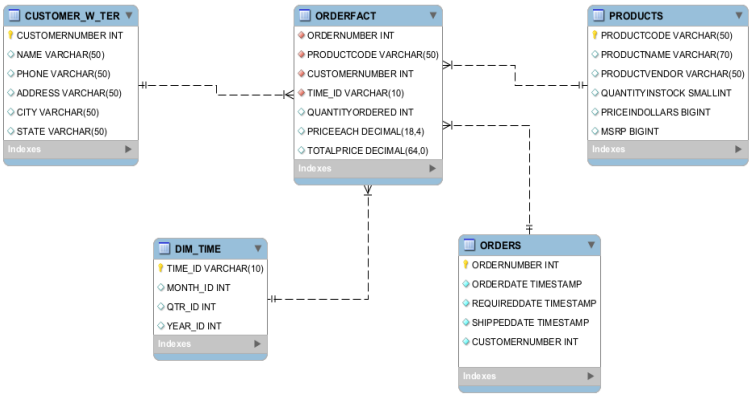


Illustration 3: ER – EUA

O esquema da figura 3 representa uma estrutura do país (i), onde os atributos ADDRESSLINE e CUSTOMERNAME foram modificados para ADDRESS e NAME, respetivamente. Adicionalmente a moeda utilizada foi definida para o DÓLAR.

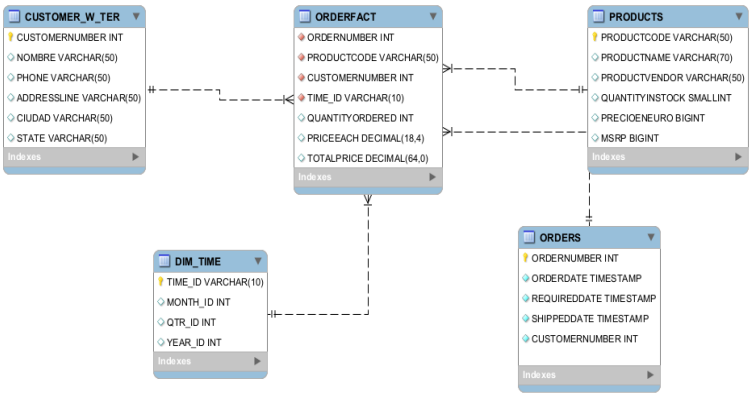


Illustration 4: ER - Espanha

O esquema da figura 4 representa uma estrutura do país (ii), onde os atributos CUSTOMERNAME, CITY e BUYPRICE foram traduzidos para espanhol. Adicionalmente a moeda utilizada foi definida para o EURO.

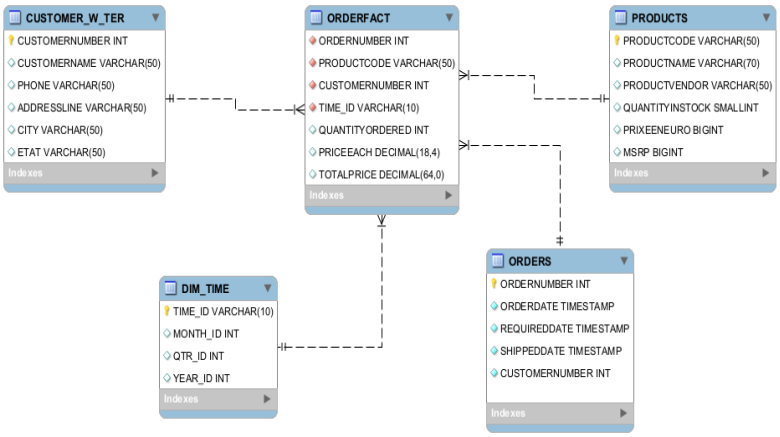


Illustration 5: ER - França

O esquema da figura 5 representa uma estrutura do país (iii), onde os atributos STATE e BUYPRICE foram traduzidos para francês. Adicionalmente a moeda utilizada foi definida para o EURO.

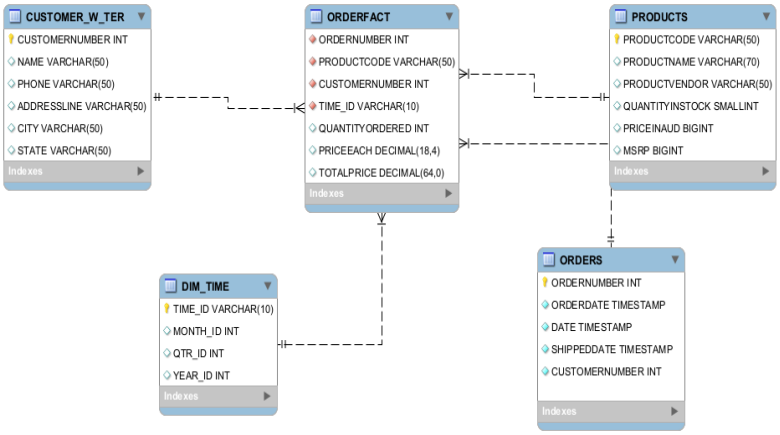


Illustration 6: ER - Austrália

O esquema da figura 6 representa uma estrutura do país (iv), onde os atributos CUSTOMERNAME e REQUIREDDATE foram modificados para NAME e DATE, respetivamente. Adicionalmente a moeda utilizada foi definida para o dólar australiano.

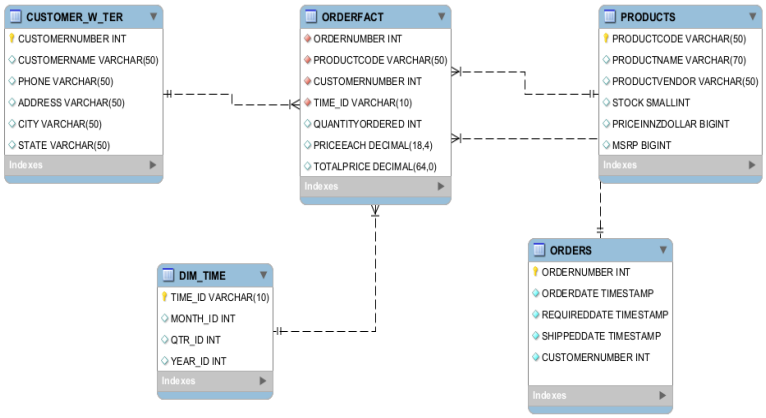


Illustration 7: ER - Nova Zelândia

O esquema da figura 7 representa uma estrutura do país (v), onde os atributos foram alterados de modo a respeitarem o contexto geográfico, ao alterar a moeda para dólares da Nova Zelândia. Adicionalmente o campo ADDRESSLINE da tabela CUSTOMER_W_TER foi alterado para ADDRESS e alterado o campo QUANTITYINSTOCK da tabela PRODUCTS para STOCK.

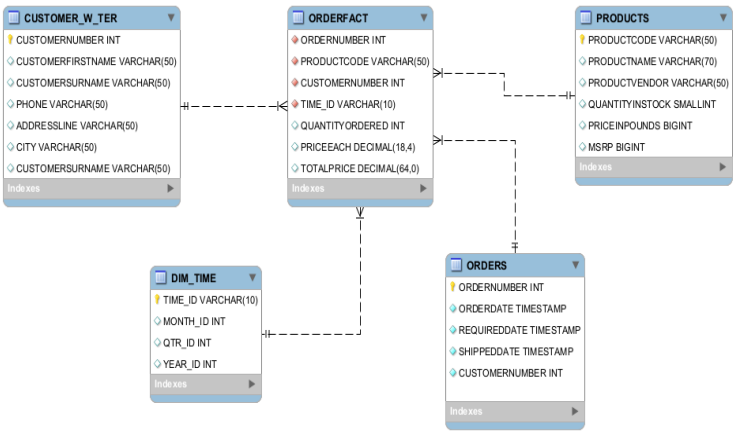


Illustration 8: ER - Reino Unido

O esquema da figura 8 representa uma estrutura do país (vi), onde os atributos foram alterados de modo a respeitarem o contexto geográfico, ao dividir o campo *CUSTOMERNAME* em dois, nomeadamente *CUSTOMERFIRSTNAME* e *CUSTOMERSURNAME*. Adicionalmente foi realizada uma alteração na moeda utilizada para *pound*.

Exportação das fontes de dados

Após uma breve discussão, o grupo decidiu criar as 6 fontes para cada um dos países nos seguintes formatos de exportação:

- USA – exportação de dados no formato .csv;
- Espanha – exportação de dados no formato .mdb;
- França – exportação de dados no formato .xls;
- Austrália – exportação de dados em MySQL;
- Nova Zelândia – exportação de dados em Sqlite;
- Reino Unido – exportação de dados em json;

Excel

A figura 9 representa a criação da fonte de dados referente à França. Utilizando o Kettle, criou-se esta configuração de forma a exportar as tabelas da base de dados (simplificada) SteelWheels para um único ficheiro Excel. Cada tabela da base de dados irá corresponder a uma sheet deste ficheiro. Foi também configurada através de JavaScript a conversão de dólares americanos para euros, de forma a demonstrar a funcionalidade de Data Transformation.

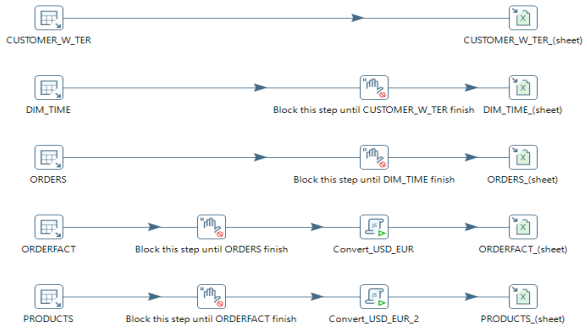


Illustration 9: Criação de fonte para exportação de dados em Excel.

MySQL remoto

A figura 10 representa a criação da fonte de dados referente à Austrália. Utilizando o Kettle, criou-se esta configuração de forma a exportar as tabelas da base de dados (simplificada) SteelWheels em SQL para formato de base de dados MySQL. Para tal foi necessário criar uma base de dados, neste caso "australia_db" através do XAMPP. Posteriormente, realizou-se uma conexão com a mesma no Kettle de forma a se poder exportar as tabelas. Foi também configurada através de JavaScript a conversão de dólares americanos para dólares australianos, de forma a demonstrar a funcionalidade de Data Transformation.

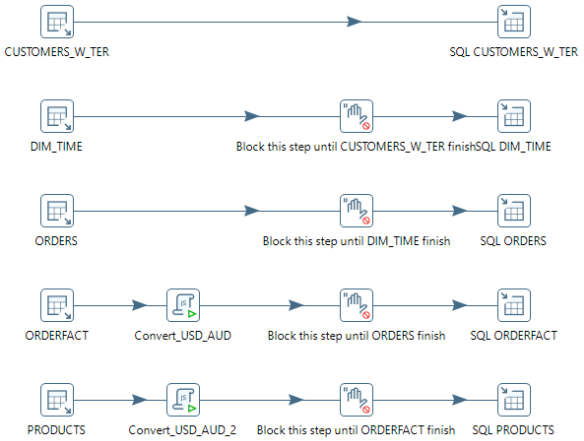


Illustration 10: Criação de fonte para exportação de dados em MySQL.

CSV

A exportação da fonte de dados csv foi realizada para o país EUA, como é possível verificar na figura 11. De forma semelhante aos casos anteriores, foram realizadas *querys* para obter os dados referentes de cada tabela deste país. Seguidamente foi feita uma verificação por tuplos nulos e renomeados os campos de modo a corresponderem ao ER da figura 3. Para finalizar os dados foram exportados em formato CSV para uma pasta EUA, onde cada ficheiro representa uma tabela distinta. Neste caso não foi realizada uma transformação monetária, pelo facto dos valores já estarem em dólar americano.

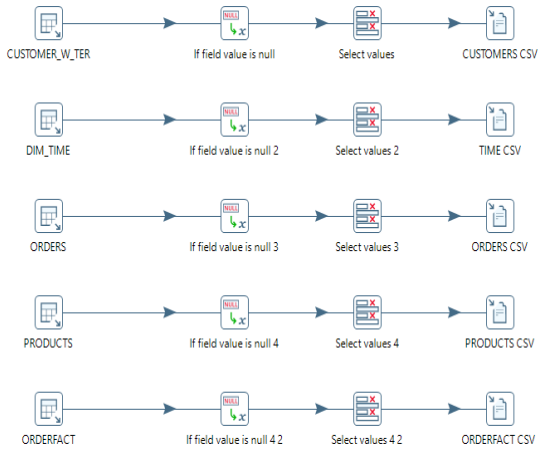


Illustration 11: Criação de fonte para exportação de dados em CSV.

Microsoft Access

Para a exportação de dados para MDB utilizou-se os dados referentes ao país Espanha, como realizado anteriormente realizou-se as devidas queries para cada tabela pretendida, sempre mantendo a ligação apenas com o país em questão, para que todos os resultados obtidos pertencem ao mesmo. Na tabela de ORDERFACT foi necessário realizar mais uma vez a conversão da moeda de USD para EURO, outra observação a ter em consideração nesta mesma tabela, foi necessário alterar o tipo das colunas PRICEEACH e TOTALPRICE de *number* para *string* pois não estava a permitir introduzir os valores no ficheiro MDB devido a um erro de precisão. No final deste processo obteve-se um único ficheiro MDB contendo este todas as tabelas referente ao país Espanha.

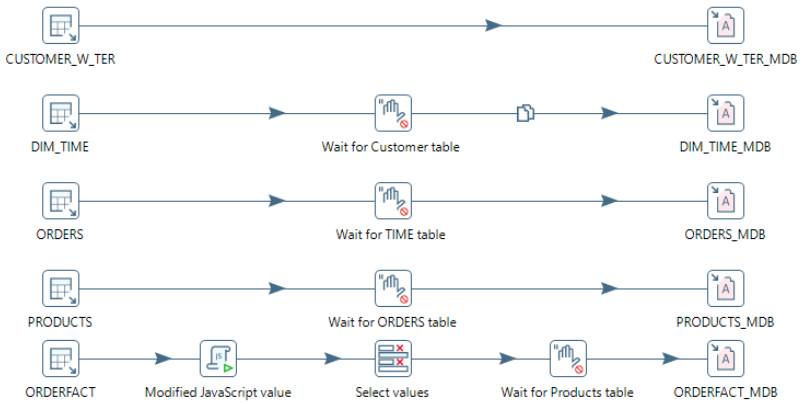


Illustration 12: Criação de fonte para exportação de dados em MDB

Sqlite

Para a exportação de dados Sqlite, foi primeiro realizada uma exportação da estrutura de dados no formato mysql, seguido de uma transformação para formato sqlite. Este passo permite obter um ficheiro sqlite com a estrutura de dados necessária para integrar os dados adquiridos pelo Kettle. O processo de carregar os dados e modificá-los está visível na figura 13. A exportação foi realizada para a Nova Zelândia.

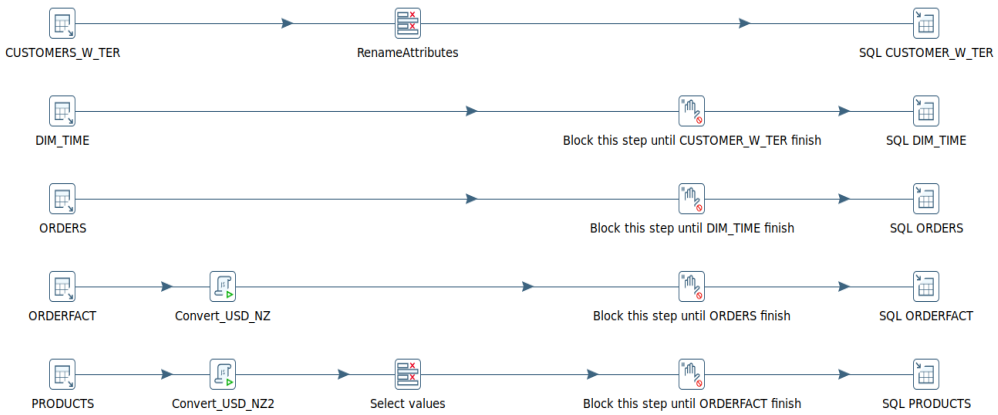


Illustration 13: Criação de fonte para exportação de dados em SQlite.

JSON

Para a exportação de dados JSON, foram adquiridos os dados através de *tables inputs*, este foram depois formatados e convertidos para os formatos adequados. Para finalizar foi realizado uma exportação em modo "*write to file*", visível na figura 14 referente ao Reino Unido.

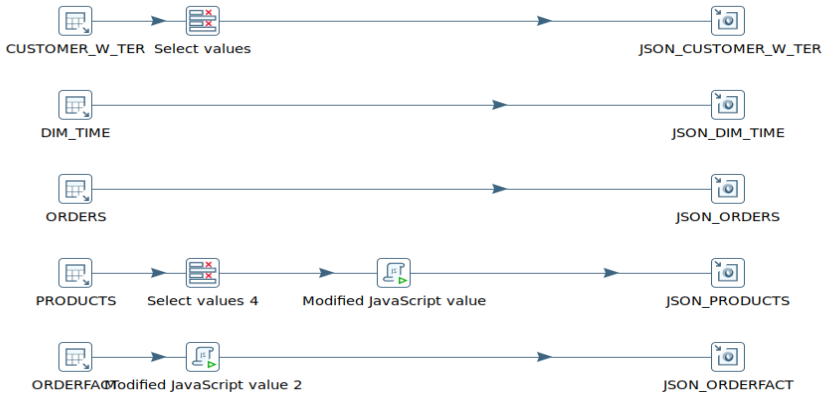


Illustration 14: Criação de fonte para exportação de dados em JSON.

Recolha/escolha de valores da amostra SteelWheels

Após a criação das seis fontes de dados referidas anteriormente, foram selecionadas três para efetuar adulterações, nomeadamente ao tornar alguns dados incompletos, de modo a demonstrar funcionalidade de Data Cleaning (preenchimento de valores ausentes com o valor mais provável). Em todas as fontes, excluindo EUA, os valores monetários foram transformados para a moeda em questão no contexto geográfico. As três fontes selecionadas foram **Austrália**, **Espanha** e **EUA**.

Na primeira o CUSTOMERNUMBER da tabela ORDERS foi colocado a nulo para os ids 10120 e 12270. Adicionalmente na tabela PRODUCTS foram removidos o MSRP, PRICEINAUD e QUANTITYINSTOCK para os registos S18_4600, S18_3482 e S24_1785, respetivamente.

Para a fonte de Espanha o PRECIOENEURO da tabela PRODUCTS foi colocado a nulo para os ids S12_1099 e S18_2325, tal como o MSRP para o id S12_4473. Adicionalmente na tabela CUSTOMER_W_TER foi removido a CIUDAD para o registo 458 e na tabela DIM_TIME removido o QTR_ID para o id 2003-0627.

Finalmente na última fonte o PRINCEINDOLLARS da tabela PRODUCTS foi colocado a nulo para os ids S10_1678 e S18_3685, tal como o MSRP para o id S18_2625. Foram ainda removidos o YEAR_ID e QTR_ID dos registos 29-04-2003 e 31-01-2003, na tabela DIM_TIME.

Howtos

De forma a sistematizar os problemas encontrados e facilitar a execução de trabalho futuro usando as ferramentas do Pentaho, esta secção contém uma listagem de itens que, no momento da realização do trabalho, não constavam nos howtos online disponibilizados, de modo a que outros grupos possam usufruir destas novas soluções.

JAVA_HOME em sistemas UNIX.

1. Instalar JAVA 8 - algumas funcionalidades não funcionam em outras versões após consulta de documentação
    1. ***sudo apt install openjdk-8-jdk***
2. Descobrir caminho para a pasta JAVA
    1. ***dirname $(dirname $(readlink -f $(which javac)))***
3. Adicionar o caminho anterior ao ambiente de desenvolvimento
    1. Editar o ficheiro */etc/profile*
        1. Adicionar export ***JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64***
    2. Reiniciar sessão do utilizador
4. Verificar sucesso da operação
    1. ***echo $JAVA_HOME*** - deverá retornar o caminho detetado anteriormente
    2. ***java -version*** - deverá retornar a versão atualmente instalada do java

Considerações adicionais

1. Verificar a versão do java necessária para o software Pentaho aquando da instalação, pois a versão 8 poderá não ser a utilizada atualmente.
2. Como opção poderá definir o caminho da variável num ficheiro *.bashrc* ou *.zprofile*, dependo do interpretador de comandos UNIX que esteja a utilizar.
3. Poderá ser necessário exportar a variável PATH
    1. ***export PATH=${PATH}:${JAVA_HOME}/bin***

Software Data Integration em sistemas UNIX.

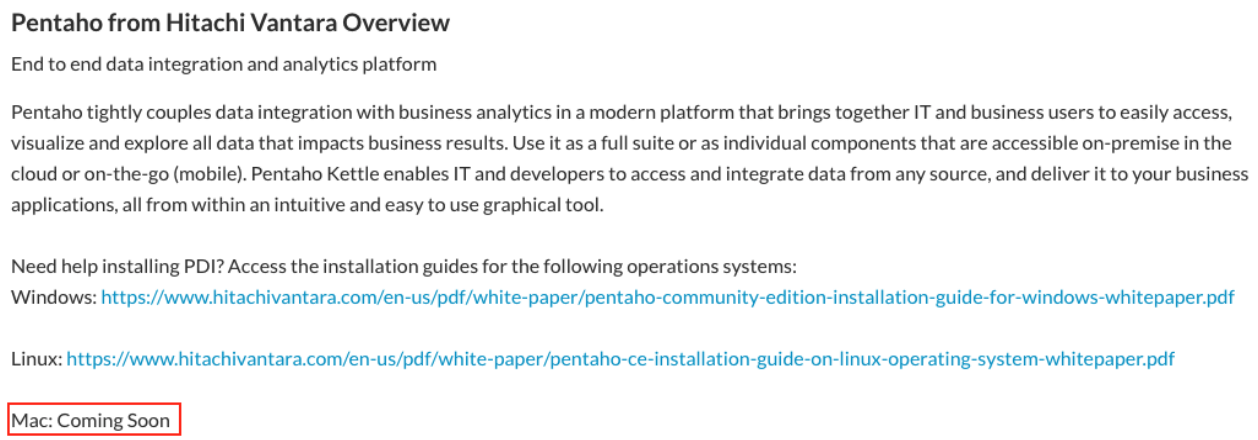
1. Requisitos
    1. Java
    2. Variáveis de ambiente configuradas
2. Download do ficheiro zip
    1. ***https://sourceforge.net/projects/pentaho/***
3. Extrair ficheiros para uma pasta à sua escolha
4. Executar o ficheiro spoon.sh que se encontra dentro da pasta data-integration
    1. *./spoon.sh*

Considerações adicionais

1. Como opção poderá seguir o tutorial fornecido na documentação da plataforma Pentaho. O sistema pedirá que efetue o registo, mas na realidade pode fornecer dados aleatórios pois não é realizada uma verificação do email.
    1. ***https://www.hitachivantara.com/en-us/pdf/white-paper/pentaho-ce-installation-guide-on-linux-operating-system-whitepaper.pdf***

Considerações MacOS:

Ao tentar instalar o pentaho community edition, deparamo-nos com a informação que este ainda não estava disponível para MacOS, para tal teve-se de recorrer à versão enterprise.
Nota: A versão enterprise quando se desinstala e se volta a instalar efetua uma nova renovação de licença.


Pentaho from Hitachi Vantara Overview

End to end data integration and analytics platform

Pentaho tightly couples data integration with business analytics in a modern platform that brings together IT and business users to easily access, visualize and explore all data that impacts business results. Use it as a full suite or as individual components that are accessible on-premise in the cloud or on-the-go (mobile). Pentaho Kettle enables IT and developers to access and integrate data from any source, and deliver it to your business applications, all from within an intuitive and easy to use graphical tool.

Need help installing PDI? Access the installation guides for the following operations systems:
Windows: https://www.hitachivantara.com/en-us/pdf/white-paper/pentaho-community-edition-installation-guide-for-windows-whitepaper.pdf

Linux: https://www.hitachivantara.com/en-us/pdf/white-paper/pentaho-ce-installation-guide-on-linux-operating-system-whitepaper.pdf

Mac: Coming Soon

Figure 1: Mensagem suporte do Pentaho com a informação da ausência da versão community para MacOS.

Instalação plugin Saiku:

Nos sistemas MacOS tendo a versão enterprise instalada existe uma alteração na diretoria.
Sendo assim deve-se introduzir a pasta que está dentro do ficheiro zip (disponível aqui), na seguinte diretoria: **/Applications/Pentaho/server/pentaho-server/pentaho-solutions/system**, ficando assim **/Applications/Pentaho/server/pentaho-server/pentaho-solutions/system/saiku**. Os restantes passos podem ser seguidos no Howto I.
Para iniciar o servidor local é executar o ficheiro **start.command**, disponível na diretoria **/Applications/Pentaho**, para desligar e salvar as alterações do servidor local é só executar o ficheiro **stop.command** disponível na mesma diretoria.

Conexão à base de dados no data-integration

1. *Ao configurar uma nova ligação à base de dados que contém a amostra Steelwheels* a password da conexão poderá ser não só "pentaho_user", mas também "**password**".

Registo de Trabalho

		Global	Individual	Individual	Individual	
	Total ->	94	34	34	34	
Entrega	**Segmento**	**# Participantes**	**Leonardo Abreu**	**José Freitas**	**João Franco**	**Descrição**
RT04	150	1	X	O	O	Verificação e correção em erros na Exportação de dados das 6 fontes pedidas
RT04	151	1	X	O	O	Verificação e correção em erros na Exportação de dados das 6 fontes pedidas
RT04	152	1	X	O	O	Verificação e correção em erros na Exportação de dados das 6 fontes pedidas
RT04	153	1	X	O	O	Verificação e correção em erros na Exportação de dados das 6 fontes pedidas
RT04	154	1	X	O	O	Verificação e correção em erros na Exportação de dados das 6 fontes pedidas
RT04	155	1	X	O	O	Criação do Repositório Consolidado
RT04	156	1	X	O	O	Criação do Repositório Consolidado
RT04	157	1	X	O	O	Criação do Repositório Consolidado
RT04	158	1	X	O	O	Criação de uma tranformation para introduzir os dados das 6 fontes de dados
RT04	159	1	X	O	O	Criação de uma tranformation para introduzir os dados das 6 fontes de dados
RT04	160	1	X	O	O	Criação de uma tranformation para introduzir os dados das 6 fontes de dados
RT04	161	1	X	O	O	Criação de uma tranformation para introduzir os dados das 6 fontes de dados
RT04	162	1	X	O	O	Criação de uma tranformation para introduzir os dados das 6 fontes de dados
RT04	163	1	O	X	O	Criação da tranformação para extração de dados para um ficheiro sqlite do país: Nova Zelândia
RT04	164	1	O	X	O	Criação da tranformação para extração de dados para um ficheiro sqlite do país: Nova Zelândia
RT04	165	1	O	X	O	Criação da tranformação para extração de dados para um ficheiro sqlite do país: Nova Zelândia
RT04	166	1	O	X	O	Correção em erros na Exportação de dados em CSV: EUA
RT04	167	1	O	X	O	Tornar alguns dados incompletos
RT04	168	1	O	X	O	Verificação e correção em erros na Exportação de dados das 6 fontes pedidas
RT04	169	1	O	X	O	Verificação e correção em erros na Exportação de dados das 6 fontes pedidas
RT04	170	1	O	X	O	Criação do Repositório Consolidado

RT04	171	1	O	X	O	Criação do Repositório Consolidado
RT04	172	1	O	X	O	Escrita do documento RT04
RT04	173	1	O	X	O	Escrita do documento RT04
RT04	174	1	O	X	O	Escrita do documento RT04
RT04	175	1	O	X	O	Registo de horas
RT04	176	1	O	O	X	Verificação e correção da transformação para extração de dados em Excel (renomeação das tabelas)
RT04	177	1	O	O	X	Verificação e correção da transformação para extração de dados em Excel (fazer a renomeação das tabelas conforme os diagramas E-R)
RT04	178	1	O	O	X	Verificação e correção da transformação para extração de dados em Excel (correção do script de conversão de USD para EUR para incluír o MSRP)
RT04	179	1	O	O	X	Verificação e correção da transformação para extração de dados em Excel (Correção de algumas queries para evitar linhas repetidas nas tabelas)
RT04	180	1	O	O	X	Correção de algumas queries para evitar linhas repetidas nas tabelas e script de conversão de USD para AUD para incluir o MSRP
RT04	181	1	O	O	X	Renomeação das tabelas conforme os diagramas E-R e apagar alguns dados exportados para posteriomente fazermos operações de data cleaning
RT04	182	1	O	O	X	Criação do Repositório Consolidado
RT04	183	1	O	O	X	introdução da tabela Product no repositório)
RT04	184	1	O	O	X	introdução da tabela Product no repositório, verificação dos vários exports.)
RT04	185	1	O	O	X	introdução da tabela Product adição de modulos "Block this step until finish" pois o export dava conflito com alguns formatos e apagava dados
RT04	186	1	O	O	X	Ajudar o Leonardo na transformação para introdução dos dados das 6 fontes (introdução da tabela DIM_TIME no repositório)
RT04	187	1	O	O	X	Ajudar o Leonardo na transformação para introdução dos dados das 6 fontes (introdução da tabela Orders no repositório)
RT04	188	1	O	O	X	Ajudar o Leonardo na transformação para introdução dos dados das 6 fontes (introdução da tabela Orders no repositório)