Tabela de conteúdos

PBI - Fase 1

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.

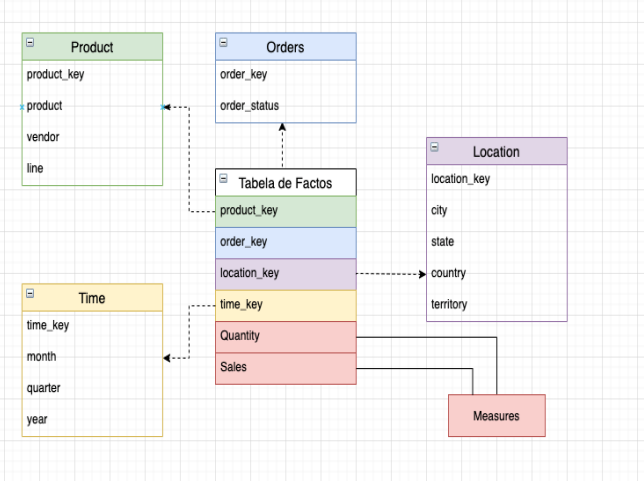


Illustration 1: Modelo Simplificado Steelwheels

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.

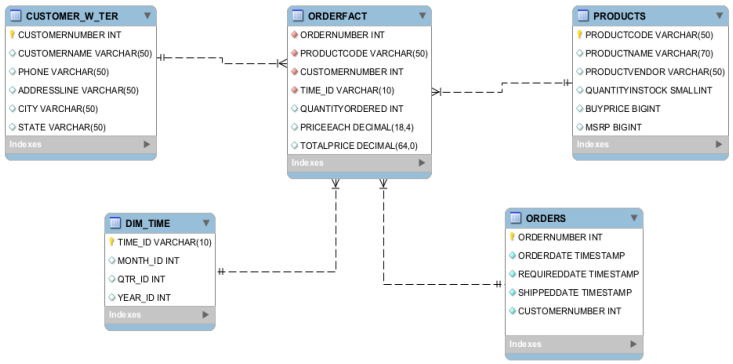


Illustration 2: ER - Tabela base

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

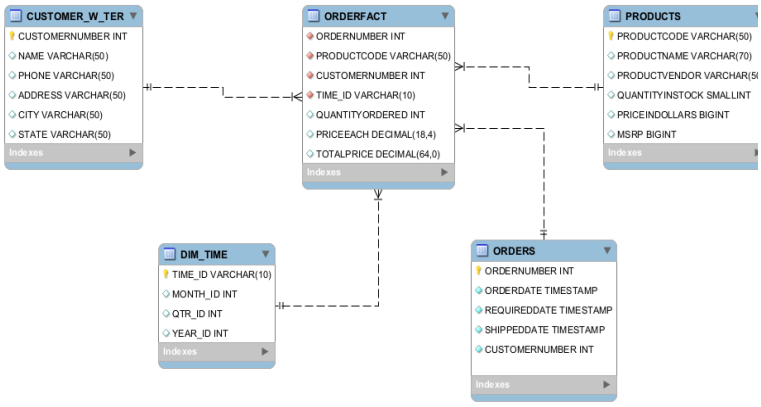


Illustration 3: ER – EUA

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.

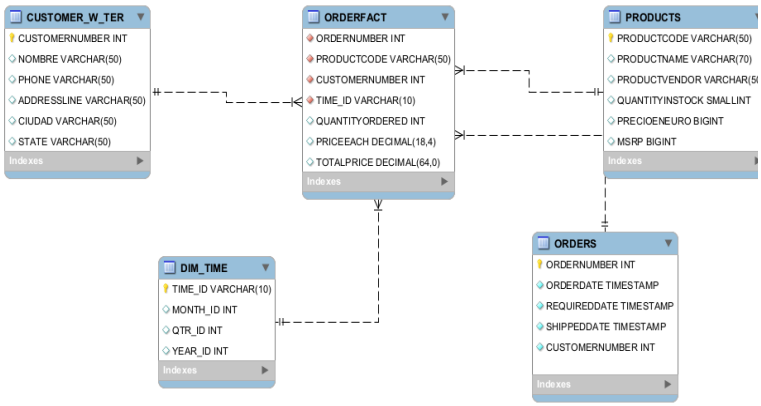


Illustration 4: ER - Espanha

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.

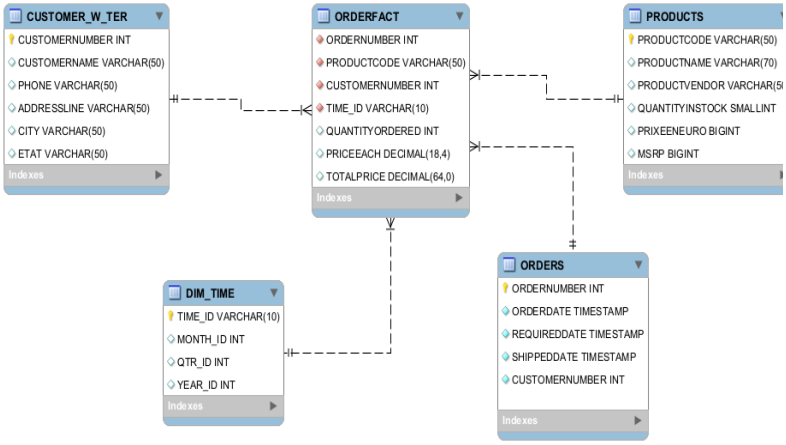


Illustration 5: ER - França

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.

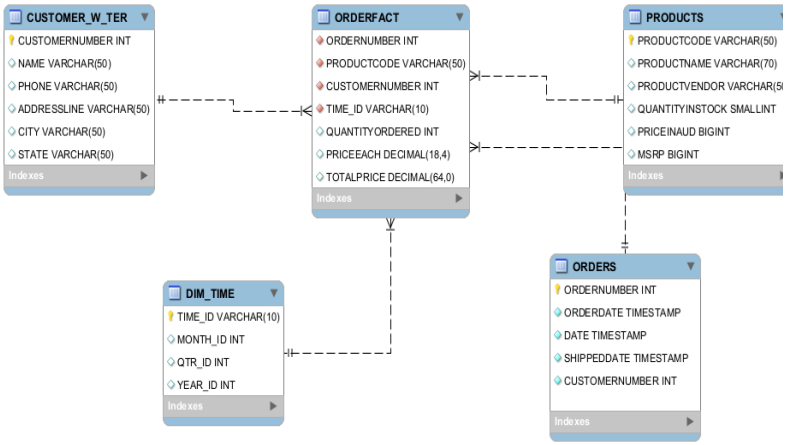


Illustration 6: ER - Austrália

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.

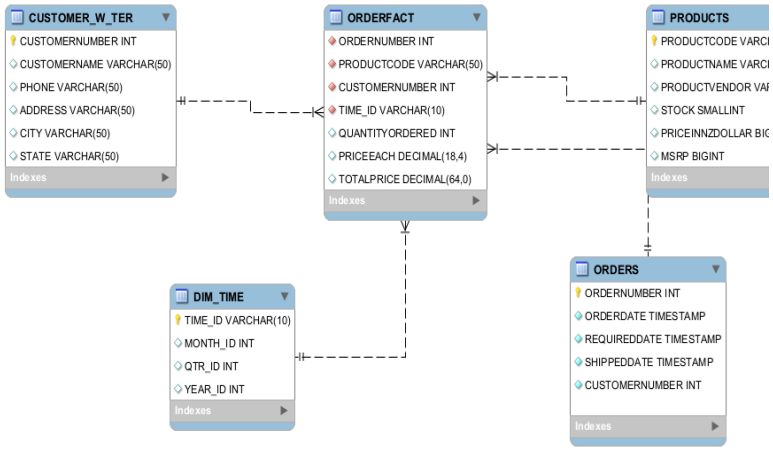


Illustration 7: ER - Nova Zelândia

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.

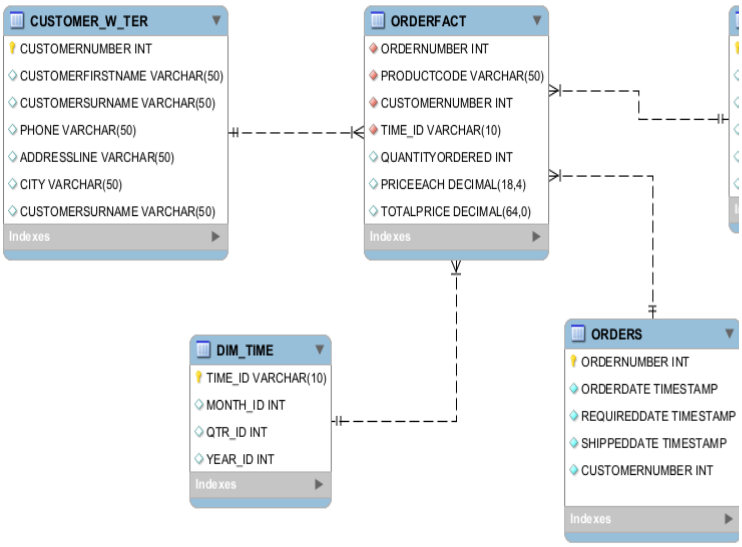


Illustration 8: ER - Reino Unido

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.

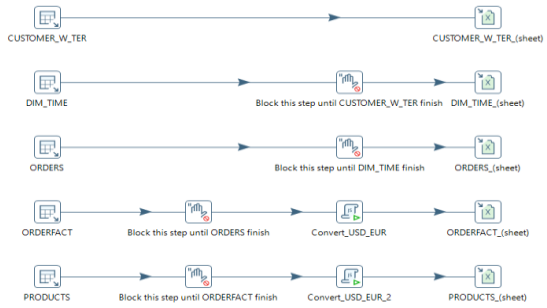


Illustration 9: Criação de fonte para exportação de dados em Excel.

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.

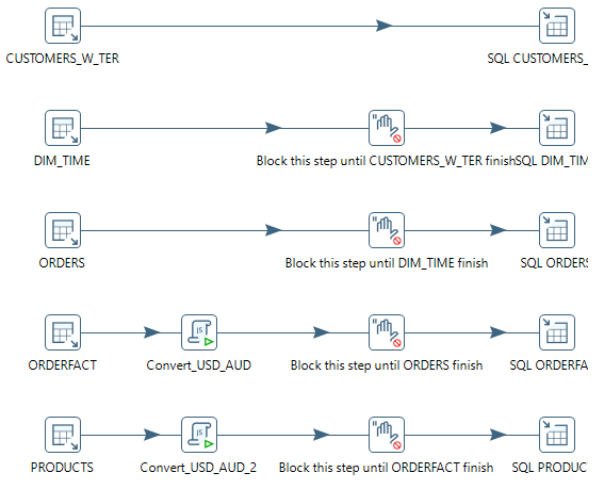


Illustration 10: Criação de fonte para exportação de dados em MySQL.

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.

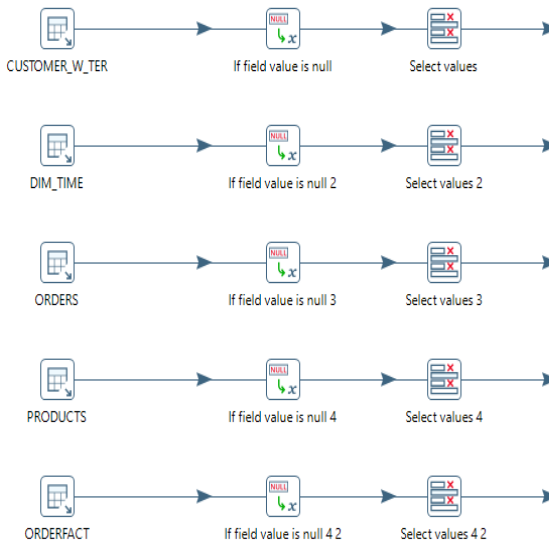


Illustration 11: Criação de fonte para exportação de dados em CSV.

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.

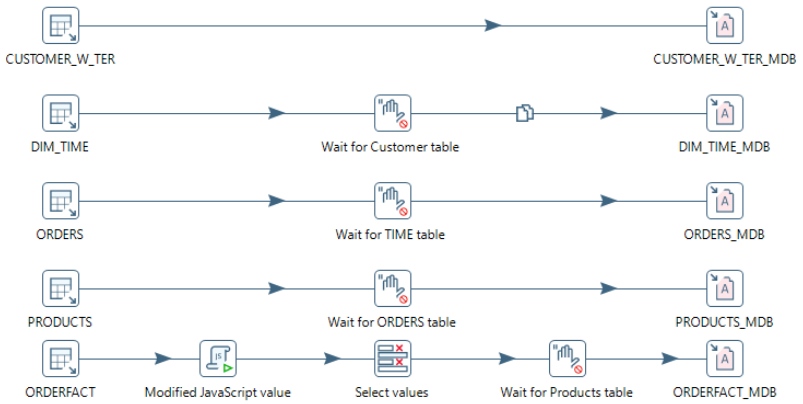


Illustration 12: Criação de fonte para exportação de dados em MDB

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.

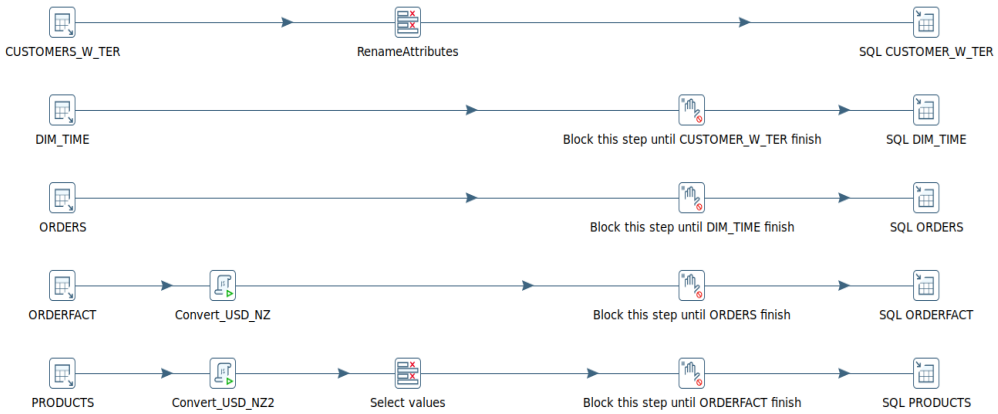


Illustration 13: Criação de fonte para exportação de dados em SQlite.

MySQL Remoto

Para a exportação de dados em MySQL remoto, 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 "*table output*", visível na figura 14 referente ao Reino Unido.

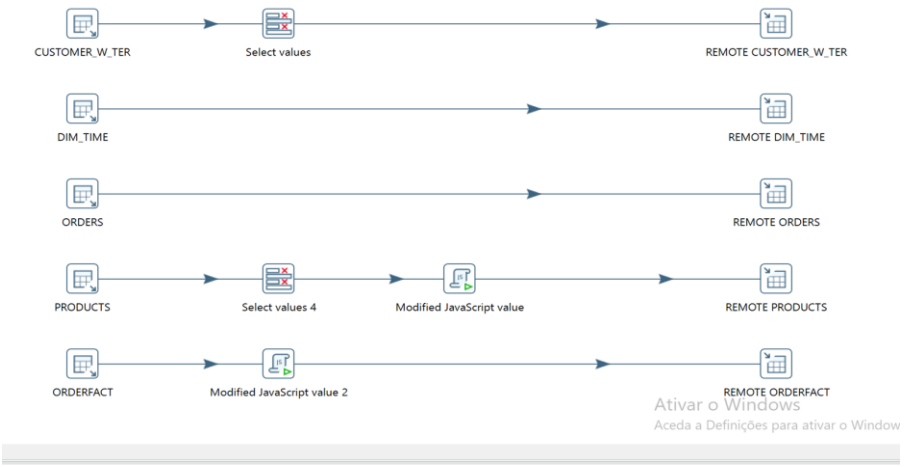


Illustration 14: Criação de fonte para exportação de dados em JSON.

Adulterações de valores da amostra

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**, **Reino Unido** e **EUA**.

Alterações

Na primeira o CUSTOMERNUMBER da tabela ORDERS foi colocado a nulo para os ids 10120 e 10125. Adicionalmente na tabela PRODUCTS foram removidos o PRICEINAUD para os registos S10_1678, S10_1949, S10_2016, S10_4698 e S10_4962

Para a fonte do Reino Unido foram removidos o YEAR_ID do registo 2003-03-18 e MONTH_ID do registo 2004-06-01, na tabela DIM_TIME.

Finalmente na última fonte o PRICEINDOLLARS da tabela PRODUCTS foi colocado a nulo para os ids S12_3148, S18_1367 e S18_4721.

União das fontes de dados

O processo de juntar novamente as fontes de dados, numa só plataforma consolidade consiste em importar os dados das diversas fontes para o kettle, renomear e transformar os atributos de modo a respeitarem a estrutura do modelo base na figura 2 e adicioná-los numa base de dados MySQL.

Carregamento de dados

De modo a satisfazer estes critérios a primeira fontes de dados de cada tabela foi adicionada em modo *truncate*, com o objetivo de limpar dados anteriores no sistema e estabelecer a estrutura sql da mesma. Para esta tarefa foi selecionada a base de dados MySQL referente à Austrália, pois já estava no formato desejado do repositório consolidado. As tabelas que se seguem realizam um insert nas

respetivas tabelas, após esperarem que a Austrália finalize a importação de dados.

Foram re-adicionadas as referências geográficas retiradas anteriormente.

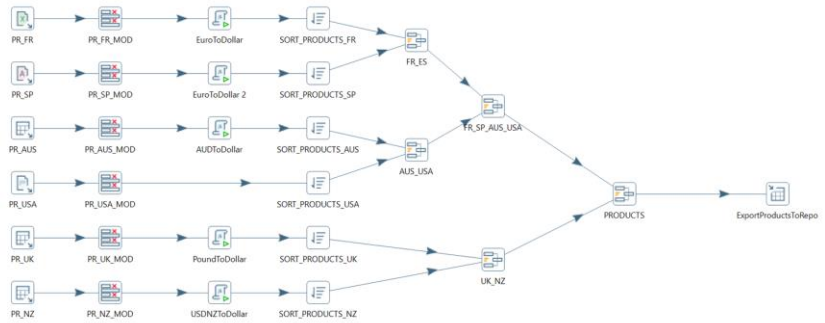


Illustration 15: DI - Products

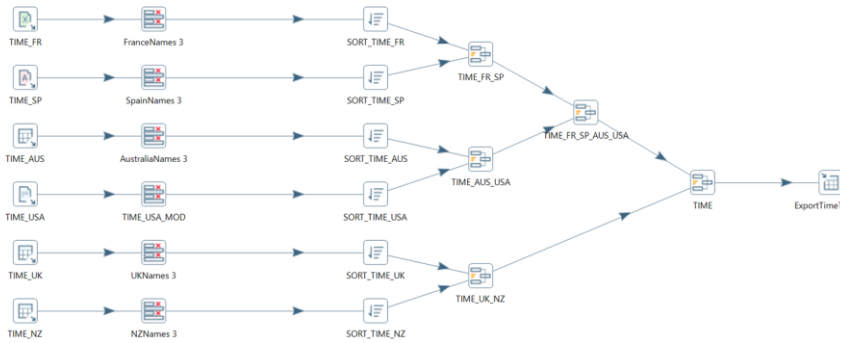


Illustration 16: DI – DIM_TIME

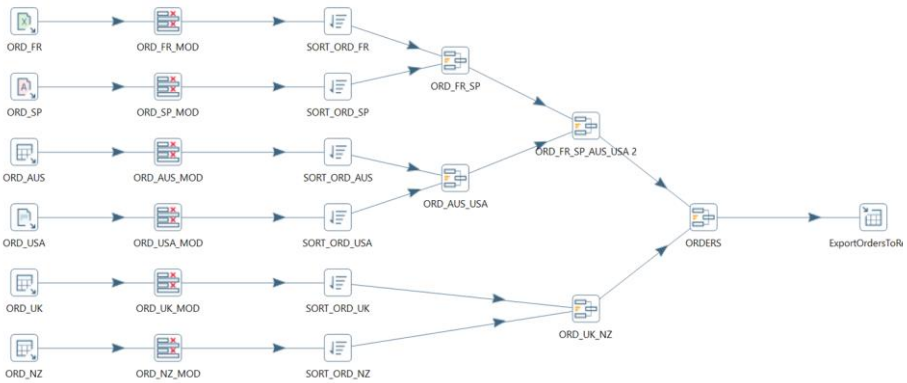


Illustration 17: DI – ORDERS

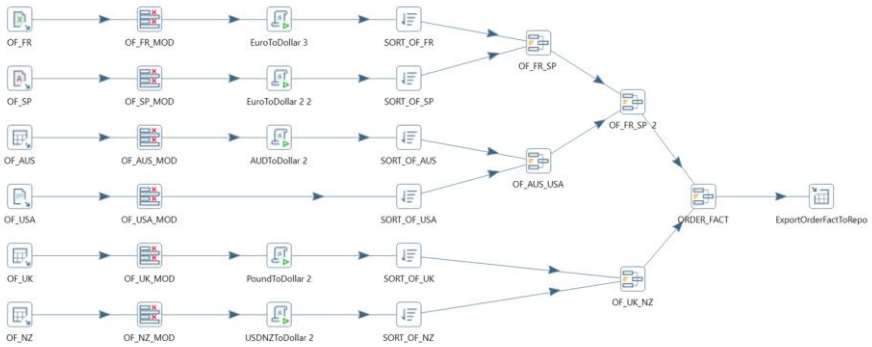


Illustration 18: DI – ORDER_FACT

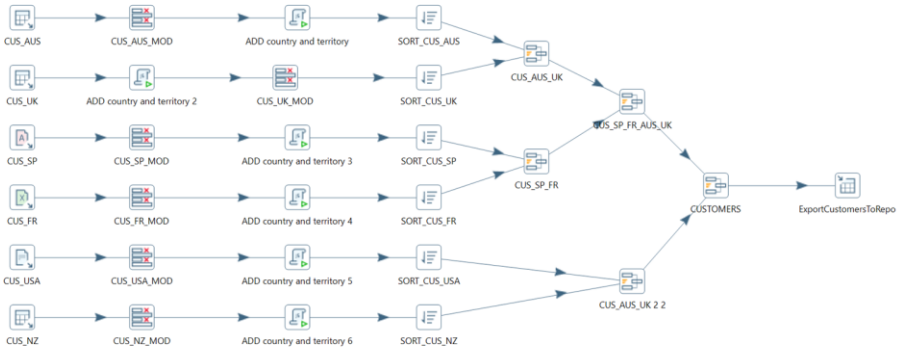


Illustration 19: DI – CUSTOMERS

Pré-processamento de dados

As diversas fontes de dados estavam estruturadas em diversos formatos e possuíam unidades, formatos e valores nulos. Como tal, foram realizadas tarefas para transformação de valores monetários, remoção de valores duplicados e cálculo de valores em falta.

Os campos nulos nos atributos PRICEINDOLLARS e PRICEINAUD foram resolvidos através do cálculo da média de valores dessas mesmas colunas a partir dos tuplos não nulos. A média foi calculada através da função SQL avg(), que substituiu os campos nulos através de um script javascript. Seguidamente os dados foram atualizados no repositório consolidado, como é possível verificar na figura 20.

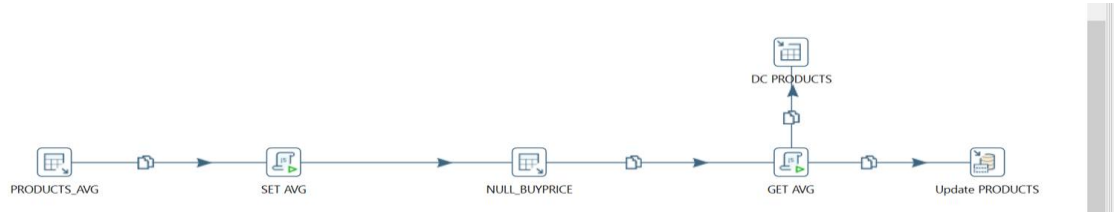


Illustration 20: Remoção de valores nulos

Relativamente ao CUSTOMERNUMBER, este valor foi obtido através da moda estatística, ou seja os valores nulos foram substituídos pelo CUSTOMERNUMBER mais frequente nos dados. O restante processo foi realizado de forma semelhante às médias, na figura 21.

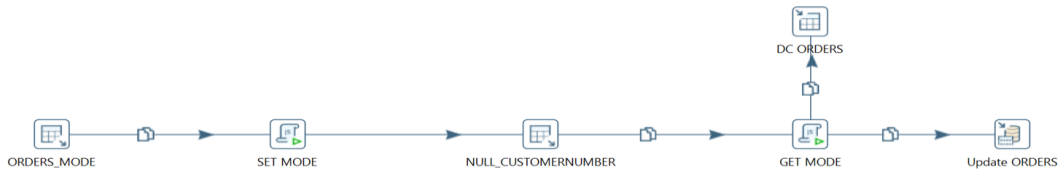


Illustration 21: Remoção de ids nulos

Finalmente as datas foram calculadas de acordo com o TIME_ID presente na tabela DIM_TIME em formato dd/mm/aaaa. Ou seja, caso o tuplo tenha qualquer um dos restantes níveis em falta (trimestre, mês e/ou ano), o valor pode ser calculado através de um *parse* no campo TIME_ID.

Illustration 22: Tratamento de datas nulas

A remoção de valores duplicados foi estabelecida através de um novo carregamento do repositório consolidado para o kettle onde as operações de seleção possuíam o atributo *distinct* na chave primária, pois trata-se de uma cláusula para eliminar repetições em consultas. Adicionalmente estas chaves primária foram ainda ordenadas e agrupadas de modo a facilitar a visualização dos dados.

Este processo está presente na figura 23.

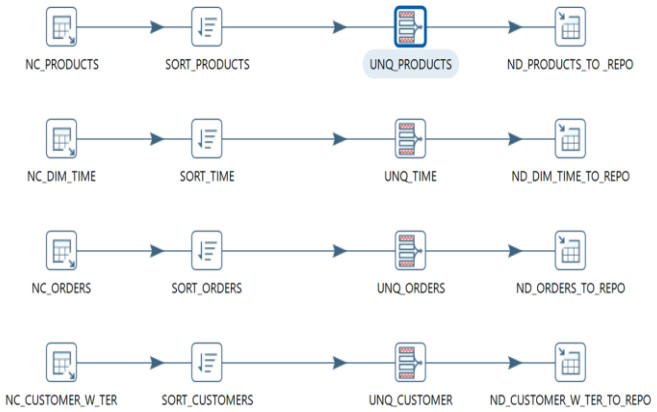


Illustration 23: Remoção de duplicados

Cubo de dados

Foi utilizado o software do schema workbench para desenhar o cubo no Pentaho que implemente o modelo multi-dimensional acima definido. Os passos que seguem indicam a ordem das tarefas realizadas de modo a obter a estrutura desejada.

1. Estabelecer uma ligação ao repositório consolidado
2. Criar um cubo de dados
3. Definir a tabela de factos: ORDERFACT
4. Criar as dimensões do cubo: PRODUCT, CUSTOMER, TIME
    1. Definir uma hierarquia de conceitos padrão com a chave estrangeira
    2. Definir tabela para cada dimensão
    3. Definir os níveis para cada dimensão e a suas chaves estrangeiras
5. Estabelecer as métricas de avaliação: QUANTITY, PRICEACH, TOTALPRICE
6. Publicar o cubo

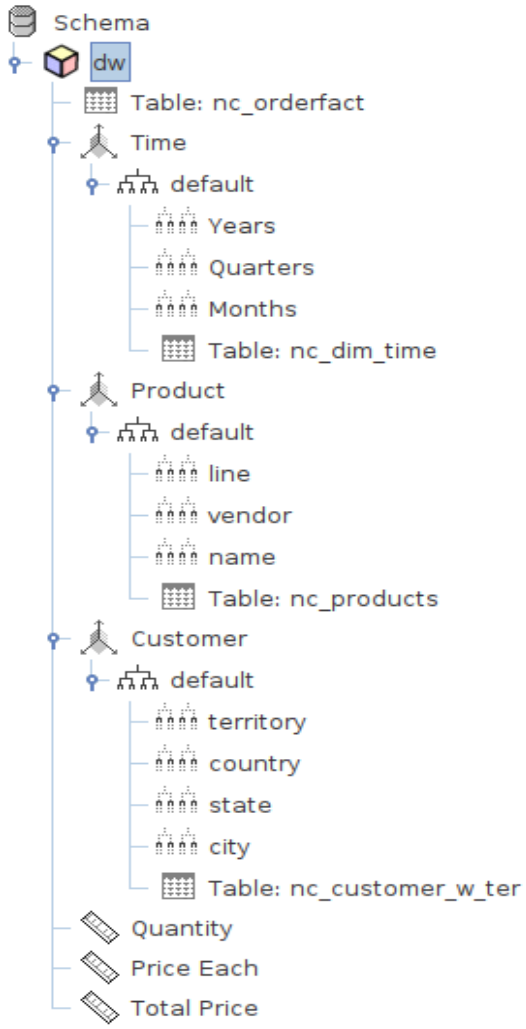


Illustration 24: Schema Workbench - Cubo de dados

Componente OLAP – Saiku

De modo a demonstrar as operações OLAP, foram implementadas dois roll-ups, dois drill-down e dois slice and dice.

Roll-up

- **"Product" para "Line"**

Um processo semelhante poderá ser realizado de forma a identificar a eficácia de um produto consoante a sua linha, num dado país. Isto poderá ser útil de modo a analisar se um produto possui um determinado número de vendas devido à qualidade do mesmo ou se já existe um contexto cultural benéfico / prejudicial que afete o seu número de vendas. Neste exemplo em concreto, a empresa poderá verificar que tipologias de produtos mais vende em cada um dos territórios.

country	Australia	New Zealand	Spain	UK	USA
name	Quantity	Quantity	Quantity	Quantity	Quantity
1968 Ford Mustang	60	190	40	48	222
1958 Chevy Corvette Limited Edition	-	-	574	74	456
1966 Shelby Cobra 427 S/C	168	64	244	98	347
1982 Camaro Z28	92	42	300	84	282
1949 Jaguar XK 120	120	118	316	262	392
1952 Alpine Renault 1300	218	-	176	-	402
1956 Porsche 356A Coupe	84	-	322	78	267
1957 Corvette Convertible	142	238	148	-	254
1961 Chevrolet Impala	48	216	138	-	286
1965 Aston Martin DB5	94	-	202	122	368
1952 Citroen-15CV	178	126	270	324	267

Illustration 25: Roll-up antes Product -> Line

country	Australia	New Zealand	Spain	UK	USA
line	Quantity	Quantity	Quantity	Quantity	Quantity
Classic Cars	3636	3052	8760	3014	11625
Motorcycles	1752	1952	1560	742	5080
Planes	1626	1034	2202	958	3476
Ships	112	744	2776	1662	2395
Trains	66	212	1018	336	912
Trucks and Buses	1410	404	3418	582	3932
Vintage Cars	3890	3394	5124	2732	8239

Illustration 26: Roll-up depois Product -> Line

- **"Country" para "Territory"**

A utilização desta operação neste caso pode ser utilizada pela organização de forma a analisar quais os territórios mais suscetíveis a uma maior quantidade de vendas. Esta informação permite aos gestores identificar padrões culturais que possam afetar as vendas dos seus produtos e consequentemente realizar adaptações da estratégia em territórios desfavoráveis.

country	Australia	New Zealand	Spain	UK	USA
line	Quantity	Quantity	Quantity	Quantity	Quantity
Classic Cars	3636	3052	8760	3014	11625
Motorcycles	1752	1952	1560	742	5080
Planes	1626	1034	2202	958	3476
Ships	112	744	2776	1662	2395
Trains	66	212	1018	336	912
Trucks and Buses	1410	404	3418	582	3932
Vintage Cars	3890	3394	5124	2732	8239

Illustration 28: Roll-up antes Country -> Territory

territory	APAC	EMEA	NA
line	Quantity	Quantity	Quantity
Classic Cars	6688	11774	11625
Motorcycles	3704	2302	5080
Planes	2660	3160	3476
Ships	856	4438	2395
Trains	278	1354	912
Trucks and Buses	1814	4000	3932
Vintage Cars	7284	7856	8239

Illustration 27: Roll-up depois Country -> Territory

Drill-down

- **"Year" para "Quarters"**

A organização pode verificar em que trimestre tende a vender determinado tipo de produtos. Desta forma garantindo a existência de stock, em trimestres de maior procura, evitando que a uma venda não ocorra por falta de produtos em stock local.

line	Classic Cars		Motorcycles		Planes		Ships	
Years	Price Each	Quantity	Price Each	Quantity	Price Each	Quantity	Price Each	Quantity
2003	10222.46 $	3996	3946.15 $	1711	2466.01 $	1020	2792.61 $	1191
2004	14460.76 $	5896	4346.22 $	1893	4522.97 $	1906	4077.53 $	1733
2005	5988.45 $	2879	3839.59 $	1803	2401.35 $	1120	1243.83 $	489

Illustration 30: Drill-down antes Year -> Quarters

line	Classic Cars		Motorcycles		Planes		Ships	
Quarters	Price Each	Quantity	Price Each	Quantity	Price Each	Quantity	Price Each	Quantity
1	1252.78 $	521	-	-	-	-	-	-
2	2351.45 $	918	1150.55 $	486	1102.51 $	458	683.83 $	293
3	2902.50 $	1026	633.09 $	227	-	-	500.36 $	234
4	3715.73 $	1531	2162.51 $	998	1363.50 $	562	1608.42 $	664
1	3122.21 $	1295	1707.43 $	780	759.60 $	318	1158.22 $	527
2	3842.50 $	1452	339.33 $	187	704.90 $	311	264.18 $	109
3	961.41 $	429	820.32 $	323	765.45 $	321	-	-
4	6534.64 $	2720	1479.14 $	603	2293.02 $	956	2655.13 $	1097
1	2983.93 $	1268	1197.87 $	477	2275.92 $	1057	1147.88 $	457
2	3004.52 $	1611	2641.72 $	1326	125.43 $	63	95.95 $	32

Illustration 29: Drill-down depois Year -> Quarters

- **"Country" para "City"**

De forma contrária ao segundo exemplo do roll-up, ao realizar um drill-down nos países é possível especificar um foco mais eficaz na gestão de vendas. Esta informação poderá ser extremamente útil para a organização, salientando cidades de um dado país cujos investimentos poderão ser mais benéficos, ou seja, com a realização desta operação, será possível a organização perceber quais as cidade mais lucrativas de cada país. Adicionalmente, estes dados poderão ser filtrados e entregues aos representantes locais, o que os ajudará na tomada de decisões.

country	Australia		New Zealand		Spain		UK		USA	
vendor	Price Each	Quantity	Price Each	Quantity	Price Each	Quantity	Price Each	Quantity	Price Each	Quantity
Autoart Studio Design	200.00 $	60	386.56 $	190	200.00 $	40	200.00 $	48	471.47 $	222
Carousel DieCast Legends	410.44 $	260	320.74 $	106	1863.56 $	1118	528.88 $	256	1671.31 $	1085
Classic Metal Creations	2136.36 $	706	1536.78 $	572	2915.30 $	1302	1129.74 $	462	5057.11 $	1969
Exoto Designs	933.32 $	356	967.58 $	354	2285.96 $	948	1049.11 $	564	2580.52 $	912
Gearbox Collectibles	1330.90 $	552	1139.62 $	332	3199.36 $	1234	642.22 $	244	4320.18 $	1663
Highway 66 Mini Classics	530.62 $	176	-	-	1047.22 $	514	200.00 $	52	1465.16 $	694
Min Lin Diecast	1048.68 $	368	663.64 $	330	1372.08 $	468	400.00 $	112	2423.81 $	952
Motor City Art Classics	-	-	200.00 $	96	350.94 $	142	600.00 $	184	948.05 $	345
Red Start Diecast	200.00 $	54	-	-	600.00 $	224	200.00 $	74	864.90 $	287

Illustration 31: Drill-down antes Country -> City

city	Chatswood		North Sydney		South Brisbane		Glen Waverly		Melbourne		Auckland	
vendor	Price Each	Quantity	Price Each	Quantity	Price Each	Quantity	Price Each	Quantity	Price Each	Quantity	Price Each	Quantity
Autoart Studio Design	-	-	200.00 $	60	-	-	-	-	-	-	386.56 $	190
Carousel DieCast Legends	296.74 $	166	113.70 $	94	-	-	-	-	-	-	320.74 $	106
Classic Metal Creations	400.00 $	120	936.36 $	298	200.00 $	68	-	-	600.00 $	220	1162.16 $	496
Exoto Designs	386.08 $	180	347.24 $	130	200.00 $	46	-	-	-	-	967.58 $	354
Gearbox Collectibles	200.00 $	52	600.00 $	182	-	-	200.00 $	98	330.90 $	220	939.62 $	278
Highway 66 Mini Classics	-	-	200.00 $	46	130.62 $	50	-	-	200.00 $	80	-	-
Min Lin Diecast	-	-	341.52 $	114	200.00 $	70	173.68 $	90	333.48 $	94	510.86 $	298
Motor City Art Classics	-	-	-	-	-	-	-	-	-	-	-	-

Illustration 32: Drill-down depois Country -> City

Slice and Dice

- **"Line" para "classic car" e "motocycles"**

Esta operação permite projetar e selecionar a dimensão produtos no nível "line", podendo obter registos detalhados das suas subcategorias.

	2003				2004				2005	
line	1	2	3	4	1	2	3	4	1	2
Classic Cars	521	918	1026	1531	1295	1452	429	2720	1268	1611
Motorcycles	-	486	227	998	780	187	323	603	477	1326
Planes	-	458	-	562	318	311	321	956	1057	63
Ships	-	293	234	664	527	109	-	1097	457	32
Trains	81	41	93	77	126	77	-	360	183	-
Trucks and Buses	138	454	-	918	43	184	162	751	658	666
Vintage Cars	273	659	577	1340	914	946	581	2569	903	763

Illustration 33: Slice and Dice antes Line em Classic Cars e Motorcyle

line	2003	2004	2005
Classic Cars	3996	5896	2879
Motorcycles	1711	1893	1803

Illustration 34: Slice and Dice depois Line em Classic Cars e Motorcyle

- **"Quarters" para primeiro trimestre de cada ano**

Nesta operação é possível comparar a quantidade de vendas do primeiro trimeste de cada ano, definindo métricas de comparação anuais entre vendas.

	2003	2004	2005
line	1	1	1
Classic Cars	521	1295	1268
Motorcycles	-	780	477
Planes	-	318	1057
Ships	-	527	457
Trains	81	126	183
Trucks and Buses	138	43	658
Vintage Cars	273	914	903

Illustration 35: Slice and Dice antes Time no primeiro trimestre

line	2003	2004	2005
Classic Cars	3996	5896	2879
Motorcycles	1711	1893	1803
Planes	1020	1906	1120
Ships	1191	1733	489
Trains	292	563	183
Trucks and Buses	1510	1140	1324
Vintage Cars	2849	5010	1666

Illustration 36: Slice and Dice depois Time no primeiro trimestre

PBI - Fase 2

Exportação conjuntos de dados

Foi utilizado o software kettle referido na primeira fase, para gerar seis coleções de dados distintas com o objetivo de identificar regras de associação presentes nos dados. A imagem 37 demonstra a estrutura no kettle que realiza a query à base de dados Steelwheels e exporta para um ficheiro CSV o resultado.

Illustration 37: Estrutura Kettle para criação dos conjuntos de dados

Inicialmente foram elaboradas seis perguntas de negócio relevantes, uma para cada conjuntos de dados. Estas poderão ser alteradas futuramente, consoante os resultados obtidos nos próximos passos.

Segue-se as perguntas definidas tais como os campos selecionados para fazer parte do ficheiro csv resultante:

1. Vendedor de produtos que tem sucesso num dado país, consegue transmitir esse sucesso para os mesmos países desse território no ano 2005;

PRODUCTVENDOR	COUNTRY	TERRITORY	YEAR_ID

2. Cliente que compra uma linha de produtos também compra outra linha de produtos nos EUA;

PRODUCTLINE	CUSTOMERNUMBER	COUNTRY

3. Clientes com crédito limite elevado realizam compras com valores mais elevados na América do Norte;

BUYPRICE	CUSTOMERNUMBER	CREDITLIMIT	TERRITORY

4. Relação entre quantidade em stock e soma do número de artigos em França para 2003;

PRODUCTCODE	QUANTITYINSTOCK	QUANTITYORDERED

5. Clientes que compram “Classic Cars” com limite de crédito superior a 100.000 também compram outro tipo de produtos;

PRODUCTLINE	CUSTOMERNUMBER	CREDITLIMIT

6. Cliente que compra grandes quantidades de “Classic Cars” compra menos quantidades “Trains”.

PRODUCTVENDOR	COUNTRY	TERRITORY	YEAR_ID

O grupo tentou criar uma amostra de perguntas de negócio diversificadas que demonstrem a flexibilidade e capacidades das ferramentas utilizadas. Futuramente estas serão testadas de forma a verificar se geraram regras de associação de sucesso.

Transformação do CSV para ARFF

Foi desenvolvido um script em Java que realiza uma pesquisa na diretoria que contém os ficheiros CSV, armazena o nome dos ficheiros num array e seguidamente converte-os para formato ARFF, com auxílio da Função ArffSaver().

Segue-se os blocos de código referentes ao main, listagem de ficheiros csv e conversao para ARRF.

```
public static void main(String[] args) throws Exception {

      String folderPath = "C:\\Users\\tadeu\\Desktop\\SAD2021\\SADG05\\PBI\\PBI2\\Current\\";

      ArrayList<String> csvFilesList = GetCSVFilesList(folderPath);
      for (int i = 0; i < csvFilesList.size(); i++){
          String fileName = csvFilesList.get(i);
          fileName = fileName.substring(0,fileName.length()-4);

          String filePath = folderPath + fileName;

          ConvertCSVtoARFF(filePath, fileName);
      }
   }
```

```
private static ArrayList<String> GetCSVFilesList(String path) {
      File folder = new File(path);
      File[] listOfFiles = folder.listFiles();

      ArrayList<String> result = new ArrayList<String>();

      for(int i = 0; i < listOfFiles.length; i++){
          String fileName = listOfFiles[i].getName();
          String fileExtension = fileName.substring(fileName.length()-3);

          if( fileExtension.equals("csv")){
              result.add(fileName);
          }
      }
      return result;
   }
```

```
private static void ConvertCSVtoARFF(String filePath, String fileName){
      try{
         //Load CSV
         CSVLoader loader = new CSVLoader();
         loader.setSource(new File(filePath+".csv"));
         Instances data = loader.getDataSet();

         System.out.println(data.size());

         // Save ARFF
         ArffSaver saver = new ArffSaver();
         saver.setInstances(data);

         // Save as ARFF
         saver.setFile(new File("${path}\\"+fileName+".arff"));
         saver.writeBatch();
         System.out.println("SUCCESS: File Created");
      }
      catch(IOException E){
         System.out.println("ERROR: The File "+ fileName +" cannot be converted");
         System.out.println(E);
      }
   }
```

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

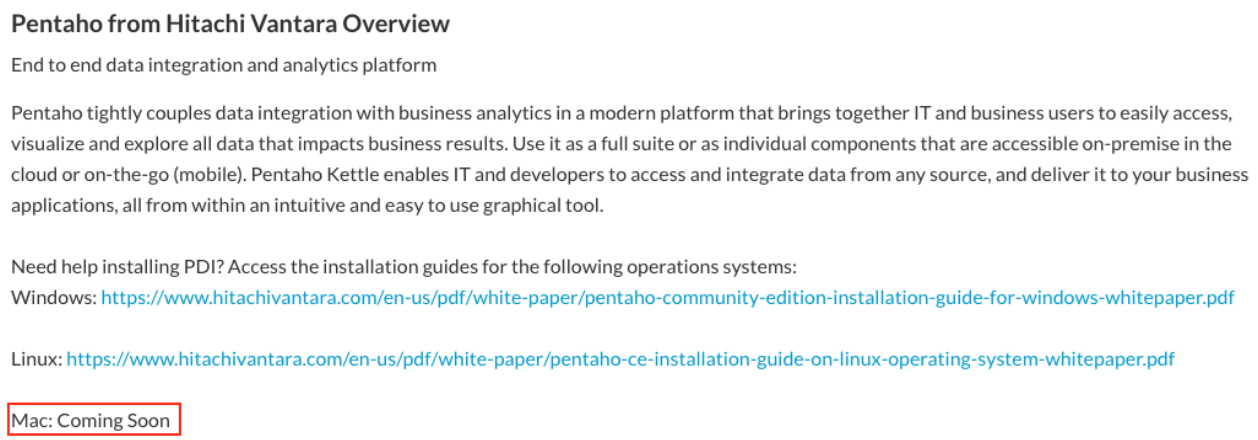
1. *./spoon.sh*

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**	
198,5		98		98	98	
# Participantes		**Leonardo Abreu**		**José Freitas**	**João Franco**	**Descrição**
1	O	O	X			Correcao de entrega anterior - repositorio consolidado
1	O	O	X			Conjunto de dados
1	O	O	X			Conjunto de dados
1	O	O	X			Instalação Eclipse IDE e configuração da API Weka
1	O	O	X			Instalação Netbeans (devido a problemas com o Eclipse) e configuração da API Weka
1	O	O	X			Criação do programa em Java para conversão das tasks csv para arff
1	O	O	X			Estudando a documentação da API Weka
1	O	O	X			Criação do programa em Java para conversão das tasks csv para arff
1	O	O	X			Criação do programa em Java para conversão das tasks csv para arff
1	O	O	X			Criação do programa em Java para conversão das tasks csv para arff
1	O	O	X			Implementação do algoritmo apriori para extração das regras de associação
1	O	O	X			Escrita relatório
1	O	O	X			Escrita relatório
1	O	X	O			Correcao de entrega anterior - repositorio consolidado
1	O	X	O			Correcao de entrega anterior - repositorio consolidado
1	O	X	O			Conjunto de dados
1	O	X	O			Conjunto de dados
1	O	X	O			Inicializacao de java API para transformacoes
1	O	X	O			Inicializacao de java API para transformacoes
1	O	X	O			Inicializacao de java API para transformacoes
1	O	X	O			Inicializacao de java API para transformacoes
1	O	X	O			Inicializacao de java API para transformacoes
1	O	X	O			Inicializacao de java API para transformacoes
1	O	X	O			Inicializacao de java API para transformacoes
1	O	X	O			Inicializacao de java API para transformacoes
1	O	X	O			Escrita relatório
1	X	O	O			Correcao de entrega anterior - repositorio consolidado
1	X	O	O			Correcao de entrega anterior - repositorio consolidado
1	X	O	O			Conjunto de dados

1	X	O	O	Conjunto de dados
1	X	O	O	Correcao de entrega anterior - repositorio consolidado
1	X	O	O	Instalação do NetBeans e Eclipse
1	X	O	O	Criação da classe de java para conversão .CSV para .ARFF
1	X	O	O	Criação da classe de java para conversão .CSV para .ARFF
1	X	O	O	Criação da classe de java para conversão .CSV para .ARFF
1	X	O	O	Implementação do algoritmo apriori para extração das regras de associação
1	X	O	O	Implementação do algoritmo apriori para extração das regras de associação
1	X	O	O	Implementação do algoritmo apriori para extração das regras de associação
1	X	O	O	Escrita relatório