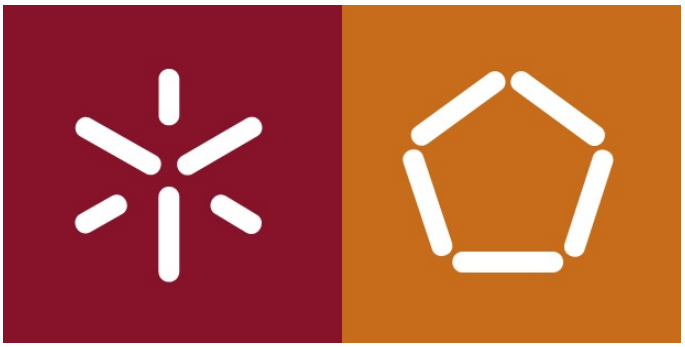


Universidade do Minho - Escola de Engenharia

Relatório do trabalho prático de Análise de Dados

Northwind Data Warehouse

Autores :
Carlos Campos (A74745)

Diana Costa (A78985)

José Oliveira (A78806)

Vitor Castro (A77870)

Versão 1.0
21 de Janeiro de 2019

Resumo

Neste relatório será feita uma abordagem ao projeto de Análise de Dados, ao qual está associado a análise, planeamento, arquitetura e implementação de um SBDMD. O objeto de estudo será a tão conhecida base de dados NorthWind, que armazena dados relativos ao comércio de produtos alimentares de uma empresa fictícia. Será também analisado, através do Microsoft PowerBI Desktop, o Data Warehouse implementado, de forma a extrair dos dados a informação pertinente requerida.

Conteúdo

1 Introdução

Este projeto foi elaborado no âmbito da Unidade Curricular de Análise de Dados, do 4º ano do Mestrado Integrado em Engenharia Informática, com vista ao planeamento e execução de projetos de SBDMD e construção de plataformas para suporte à visualização de dados. Pretendia-se, mais concretamente, que fosse analisada a base de dados da Microsoft "Northwind", e a partir daqui, planeado, arquitetado e implementado um SBDMD. Através dos dados relativos ao comércio de produtos alimentares de uma empresa fictícia, armazenados na BD "Northwind", seria necessário elaborar um sistema de povoamento inicial, contemplando as estruturas analíticas necessárias ao seu refrescamento incremental e/ou diferencial. De forma a, finalmente, analisar os dados do data warehouse, teriam de ser definidos indicadores de Business Intelligence em dashboards.

Assim, neste relatório, pode-se constatar o trabalho desenvolvido pelo grupo referente a quatro fases principais. Numa primeira fase, analisou-se a fundo a "Northwind", e tomaram-se decisões acerca de que informação seria analisada, já que esta base de dados é extensa e possui várias vertentes de análise. Numa outra fase, foi desenhado o modelo dimensional e definidas as perguntas a serem respondidas pelo DW. Na terceira fase, o grupo descreveu todas as etapas, com a ajuda do Pentaho, desde a extração dos dados da fonte de informação, até à sua inserção no DW. Ainda neste passo, contemplaram-se a implementação das estruturas que acomodariam os dados transformados e/ou a transformar, assim como as necessárias ao refrescamento de novos dados. Por fim, para concretizar o objetivo real deste data warehouse, o grupo procedeu à análise dos dados, através do Power BI.

Desta forma, o relatório começa por introduzir o problema e analisar a base de dados escolhida pelos docentes. De seguida, apresentam-se todas as etapas de construção do DW, desde o modelo dimensional, até à implementação do sistema físico. Depois, é feita uma análise dos dados, quer através do Microsoft Power BI, quer por uma abordagem mais teórica, em que se respondem a todas as perguntas que o DW tinha de solucionar. O relatório termina com as conclusões, em que o grupo faz uma análise do trabalho desenvolvido ao longo do semestre, tendo em conta as dificuldades obtidas.

2 Análise da Base de Dados

Nesta secção relata-se todo o estudo do grupo, considerando cada passo de análise, desde a simples averiguação das entidades da BD Northwind, até à vertente de análise escolhida para ser, posteriormente, contemplada no data warehouse.

Antes de mais, é necessária uma compreensão e pesquisa acerca da utilidade da base de dados a utilizar para a construção do SBDMD. Assim, é importante esclarecer que a Northwind é uma base de dados de "amostra", usada pela Microsoft para demonstrar as features de alguns dos seus produtos, incluindo o SQL Server e o Microsoft Access. Esta BD contém os dados de vendas da Northwind Traders, uma empresa fictícia de transporte e exportação de produtos alimentares. Para tal, e como se esquematiza na imagem abaixo, a Northwind possui clientes, que compram determinados produtos, provenientes de fornecedores. A empresa armazena os produtos e mantém um stock, sendo gerida/mantida por vários funcionários. De forma a que as encomendas dos produtos cheguem aos clientes específicos, surgem os expedidores, responsáveis pelo transporte dos alimentos até cada cliente. Tudo isto gera encomendas e recibos, tanto da compra de alimentos pela NorthWind Traders, como pela venda dos mesmos a clientes.

Figura 1: Esquema abrangente de todas as funcionalidades da base de dados Northwind

Posto isto, prossegue-se para uma análise mais detalhada de todos os elementos da base de dados.

2.1 Entidades e atributos

A Northwind apresenta muitas entidades, com vários atributos. Para conseguir um maior alcance de análise, e de forma a que não se tornasse demasiado confuso, o grupo achou por bem organizar as entidades e atributos da seguinte forma:

- **Fornecedores/Vendedores** - quem abastece a companhia. Possuem atributos como identificação, empresa, contactos e localização;
- **Clientes** - quem compra através da Northwind. É armazenada informação como identificação, empresa, contactos e localização;
- **Detalhes de funcionários** - quem trabalha para a empresa. Possuem atributos como identificação, contactos, estatuto e localização;
- **Informação de produto** - produtos que a Northwind traders troca. Têm detalhes como nome, descrição, diferentes custos, categoria e níveis;
- **Detalhes de inventário** - inventário que a empresa possui. Dispõe de atributos, nomeadamente tipo, datas, produto a que se refere e quantidades;
- **Expedidores** - detalhes dos expedidores que exportam os produtos desde a companhia até ao cliente final. Possuem atributos como identificação, empresa, contactos e localização;
- **Transações de pedidos de compra** - detalhes das transações que acontecem entre os fornecedores e a companhia. Têm atributos como a identificação do fornecedor, o produto em questão, e quantidades, datas e custos relativas à encomenda;
- **Transação de ordem de venda** - detalhes das transações que acontecem entre o cliente e a companhia. Dispõe de atributos, nomeadamente a identificação do funcionário, expedidor, produto e cliente, e quantidades, datas e custos referentes à encomenda;
- **Transação de inventário** - detalhes das transações que acontecem ao nível do inventário. É armazenada informação como tipo, produto a que se refere, quantidade de produto e encomendas relacionadas;
- **Faturas** - detalhes da faturação relativa às encomendas/vendas. Possuem, acima de tudo, atributos referentes a datas e custos monetários de encomendas de ordem de venda.

2.2 Relacionamentos

O relacionamentos entre as entidades são muitos e variados. Como tal, apresentam-se os relacionamentos entre entidades mais pertinentes e que merecem análise. A imagem seguinte pode ser considerada como um pré modelo lógico da BD Northwind:

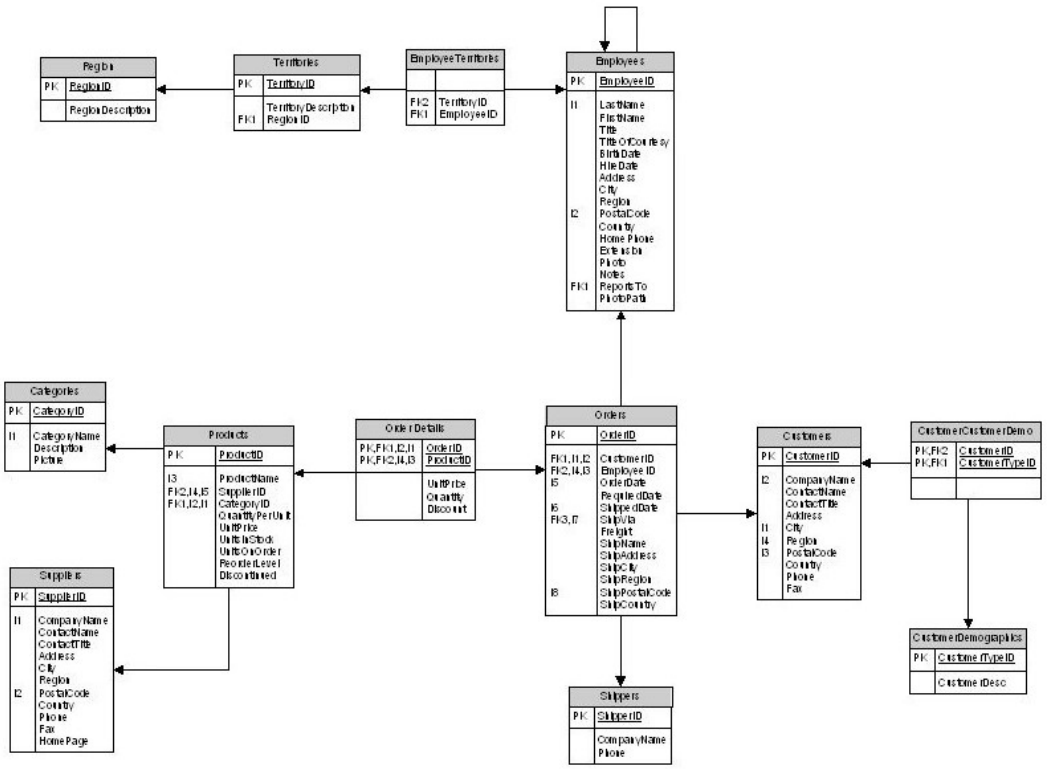


Figura 2: Pré modelo lógico - highlitght nos relacionamentos

Assim, o foco principal está nas Encomendas, que se relacionam com os Clientes, Expedidores e Funcionários. Esta entidade não se relaciona com os Produtos diretamente, mas sim através da entidade Detalhes de Encomenda. Estes produtos estão relacionados com os respectivos Fornecedores.

2.3 Modelo Lógico

Apresenta-se, finalmente, o modelo lógico da Northwind, fornecido pela equipa docente, com todas as entidades, atributos e relacionamentos representados.

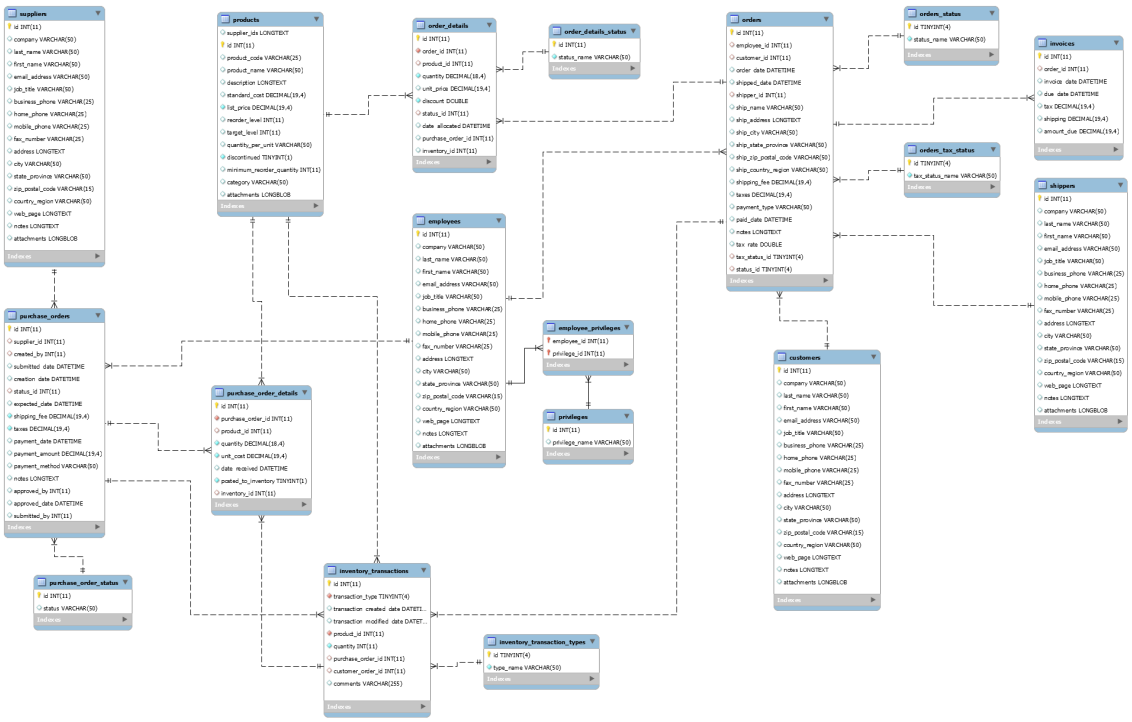
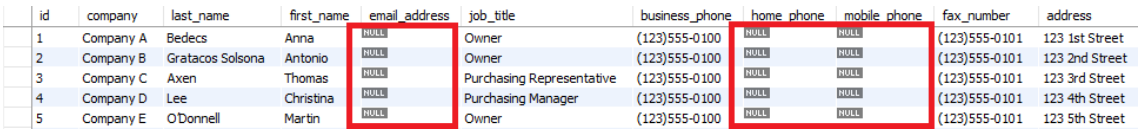


Figura 3: Modelo lógico da base de dados Northwind

2.4 Povoamento

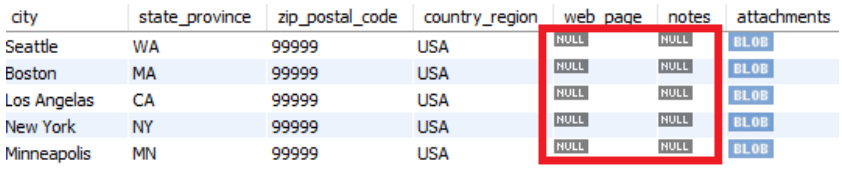
Para análise do povoamento, e uma vez que a equipa docente conhece e forneceu o mesmo ao grupo, optou-se por mostrar apenas alguns detalhes, de forma a ver que atributos é que têm tendência a não serem introduzidos (nulos), tal como a forma como os restantes estão representados. Assim, é possível conhecer os aspectos que o grupo teve em conta ao analisar o povoamento.

- **Clientes**: Dados organizados e legíveis. Algumas colunas têm valores nulos predominantes.



id	company	last_name	first_name	email_address	job_title	business_phone	home_phone	mobile_phone	fax_number	address
1	Company A	Bedecs	Anna	NULL	Owner	(123)555-0100	NULL	NULL	(123)555-0101	123 1st Street
2	Company B	Gratacos Solsona	Antonio	NULL	Owner	(123)555-0100	NULL	NULL	(123)555-0101	123 2nd Street
3	Company C	Axen	Thomas	NULL	Purchasing Representative	(123)555-0100	NULL	NULL	(123)555-0101	123 3rd Street
4	Company D	Lee	Christina	NULL	Purchasing Manager	(123)555-0100	NULL	NULL	(123)555-0101	123 4th Street
5	Company E	O'Donnell	Martin	NULL	Owner	(123)555-0100	NULL	NULL	(123)555-0101	123 5th Street

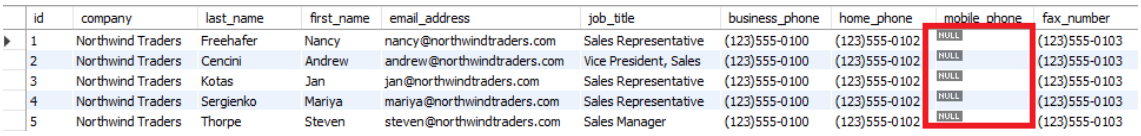
Figura 4: Alguns dados relativos a clientes da Northwind



city	state_province	zip_postal_code	country_region	web_page	notes	attachments
Seattle	WA	99999	USA	NULL	NULL	BLOB
Boston	MA	99999	USA	NULL	NULL	BLOB
Los Angelas	CA	99999	USA	NULL	NULL	BLOB
New York	NY	99999	USA	NULL	NULL	BLOB
Minneapolis	MN	99999	USA	NULL	NULL	BLOB

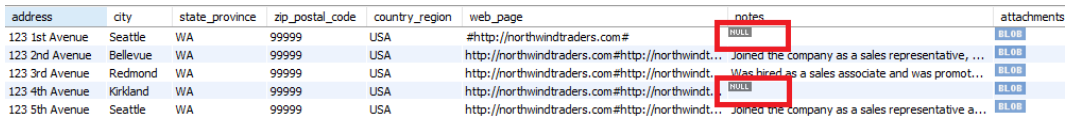
Figura 5: Alguns dados relativos a clientes da Northwind - continuação

- **Funcionários**: Dados organizados e legíveis. Algumas colunas têm valores nulos predominantes, noutras encontram-se em minoria.



id	company	last_name	first_name	email_address	job_title	business_phone	home_phone	mobile_phone	fax_number
1	Northwind Traders	Freehafer	Nancy	nancy@northwindtraders.com	Sales Representative	(123)555-0100	(123)555-0102	NULL	(123)555-0103
2	Northwind Traders	Cencini	Andrew	andrew@northwindtraders.com	Vice President, Sales	(123)555-0100	(123)555-0102	NULL	(123)555-0103
3	Northwind Traders	Kotas	Jan	jan@northwindtraders.com	Sales Representative	(123)555-0100	(123)555-0102	NULL	(123)555-0103
4	Northwind Traders	Sergienko	Mariya	mariya@northwindtraders.com	Sales Representative	(123)555-0100	(123)555-0102	NULL	(123)555-0103
5	Northwind Traders	Thorpe	Steven	steven@northwindtraders.com	Sales Manager	(123)555-0100	(123)555-0102	NULL	(123)555-0103

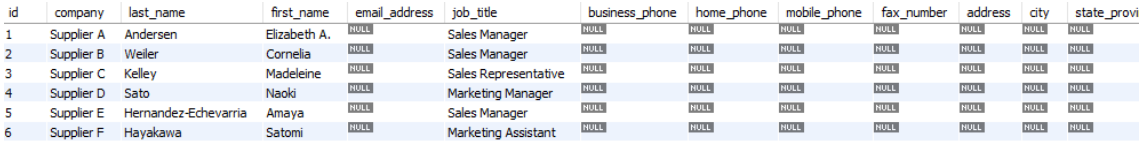
Figura 6: Alguns dados relativos a funcionários da Northwind



address	city	state_province	zip_postal_code	country_region	web_page	notes	attachments
123 1st Avenue	Seattle	WA	99999	USA	#http://northwindtraders.com#	NULL	BLOB
123 2nd Avenue	Bellevue	WA	99999	USA	http://northwindtraders.com#http://northwindt...	Joined the company as a sales representative, ...	BLOB
123 3rd Avenue	Redmond	WA	99999	USA	http://northwindtraders.com#http://northwindt...	Was hired as a sales associate and was promot...	BLOB
123 4th Avenue	Kirkland	WA	99999	USA	http://northwindtraders.com#http://northwindt...	NULL	BLOB
123 5th Avenue	Seattle	WA	99999	USA	http://northwindtraders.com#http://northwindt...	Joined the company as a sales representative a...	BLOB

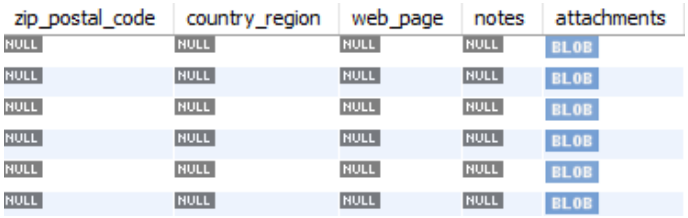
Figura 7: Alguns dados relativos a funcionários da Northwind - continuação

- **Fornecedores**: Dados organizados e legíveis. Empresas fornecedoras apenas identificadas por letras. A maioria das colunas têm valores nulos predominantes.



id	company	last_name	first_name	email_address	job_title	business_phone	home_phone	mobile_phone	fax_number	address	city	state_provi
1	Supplier A	Andersen	Elizabeth A.	NULL	Sales Manager	NULL	NULL	NULL	NULL	NULL	NULL	NULL
2	Supplier B	Weiler	Cornelia	NULL	Sales Manager	NULL	NULL	NULL	NULL	NULL	NULL	NULL
3	Supplier C	Kelley	Madeleine	NULL	Sales Representative	NULL	NULL	NULL	NULL	NULL	NULL	NULL
4	Supplier D	Sato	Naoki	NULL	Marketing Manager	NULL	NULL	NULL	NULL	NULL	NULL	NULL
5	Supplier E	Hernandez-Echevarria	Amaya	NULL	Sales Manager	NULL	NULL	NULL	NULL	NULL	NULL	NULL
6	Supplier F	Hayakawa	Satomi	NULL	Marketing Assistant	NULL	NULL	NULL	NULL	NULL	NULL	NULL

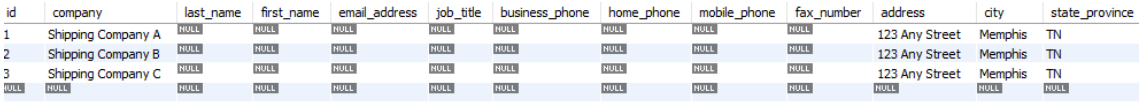
Figura 8: Alguns dados relativos a fornecedores da Northwind



zip_postal_code	country_region	web_page	notes	attachments
NULL	NULL	NULL	NULL	BLOB
NULL	NULL	NULL	NULL	BLOB
NULL	NULL	NULL	NULL	BLOB
NULL	NULL	NULL	NULL	BLOB
NULL	NULL	NULL	NULL	BLOB
NULL	NULL	NULL	NULL	BLOB

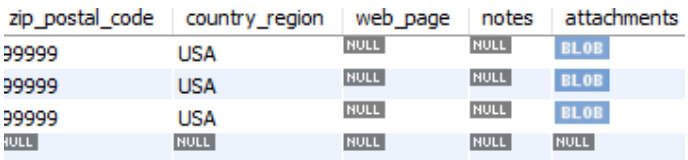
Figura 9: Alguns dados relativos a fornecedores da Northwind - continuação

- **Expedidores**: Dados organizados e legíveis. Empresas expedidoras apenas identificadas por letras. A maioria das colunas têm valores nulos predominantes.



id	company	last_name	first_name	email_address	job_title	business_phone	home_phone	mobile_phone	fax_number	address	city	state_province
1	Shipping Company A	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	123 Any Street	Memphis	TN
2	Shipping Company B	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	123 Any Street	Memphis	TN
3	Shipping Company C	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	123 Any Street	Memphis	TN
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL

Figura 10: Alguns dados relativos a expedidores da Northwind



zip_postal_code	country_region	web_page	notes	attachments
99999	USA	NULL	NULL	BLOB
99999	USA	NULL	NULL	BLOB
99999	USA	NULL	NULL	BLOB
NULL	NULL	NULL	NULL	NULL

Figura 11: Alguns dados relativos a expedidores da Northwind - continuação

- **Produtos**: Dados organizados e legíveis. Identificadores dos fornecedores dos produtos em forma de lista. Algumas colunas têm valores nulos predominantes.

supplier_ids	id	product_code	product_name	description	standard_cost	list_price	reorder_level	target_level	quantity_per_unit	discontinued
4	1	NWTB-1	Northwind Traders Chai	NULL	13.5000	18.0000	10	40	10 boxes x 20 bags	0
10	3	NWTCO-3	Northwind Traders Syrup	NULL	7.5000	10.0000	25	100	12 - 550 ml bottles	0
10	4	NWTCO-4	Northwind Traders Cajun Seasoning	NULL	16.5000	22.0000	10	40	48 - 6 oz jars	0
10	5	NWTO-5	Northwind Traders Olive Oil	NULL	16.0125	21.3500	10	40	36 boxes	0
2;6	6	NWTJP-6	Northwind Traders Boysenberry Spread	NULL	18.7500	25.0000	25	100	12 - 8 oz jars	0

Figura 12: Alguns dados relativos a produtos da Northwind

minimum_reorder_quantity	category	attachments
10	Beverages	BLOB
25	Condiments	BLOB
10	Condiments	BLOB
10	Oil	BLOB
25	Jams, Preserves	BLOB

Figura 13: Alguns dados relativos a produtos da Northwind - continuação

- **Encomendas**: Dados organizados e legíveis. Algumas colunas têm valores nulos predominantes, noutras encontram-se em minoria.

id	employee_id	customer_id	order_date	shipped_date	shipper_id	ship_name	ship_address	ship_city	ship_state_province	ship_zip_postal
40	4	10	2006-03-24 00:00:00	2006-03-24 00:00:00	2	Roland Wacker	123 10th Street	Chicago	IL	99999
41	1	7	2006-03-24 00:00:00	NULL	NULL	Ming-Yang Xie	123 7th Street	Boise	ID	99999
42	1	10	2006-03-24 00:00:00	2006-04-07 00:00:00	1	Roland Wacker	123 10th Street	Chicago	IL	99999
43	1	11	2006-03-24 00:00:00	NULL	3	Peter Krschne	123 11th Street	Miami	FL	99999
44	1	1	2006-03-24 00:00:00	NULL	NULL	Anna Bedecs	123 1st Street	Seattle	WA	99999

Figura 14: Alguns dados relativos a encomendas da Northwind

ship_country_region	shipping_fee	taxes	payment_type	paid_date	notes	tax_rate	tax_status_id	status_id
USA	9.0000	0.0000	Credit Card	2006-03-24 00:00:00	NULL	0	NULL	3
USA	0.0000	0.0000	NULL	NULL	NULL	0	NULL	0
USA	0.0000	0.0000	NULL	NULL	NULL	0	NULL	2
USA	0.0000	0.0000	NULL	NULL	NULL	0	NULL	0
USA	0.0000	0.0000	NULL	NULL	NULL	0	NULL	0

Figura 15: Alguns dados relativos a encomendas da Northwind - continuação

- **Detalhes de Encomenda**: Ponto de ligação entre encomendas e produtos encomendados. Algumas colunas têm valores nulos predominantes.

id	order_id	product_id	quantity	unit_price	discount	status_id	date_allocated	purchase_order_id	inventory_id
27	30	34	100.0000	14.0000	0	2	NULL	96	83
28	30	80	30.0000	3.5000	0	2	NULL	NULL	63
29	31	7	10.0000	30.0000	0	2	NULL	NULL	64
30	31	51	10.0000	53.0000	0	2	NULL	NULL	65
31	31	80	10.0000	3.5000	0	2	NULL	NULL	66

Figura 16: Alguns dados relativos a detalhes de encomendas da Northwind

- **Transações de pedidos de compra**: Dados organizados e legíveis. Algumas colunas com valores predominantemente nulos.

id	supplier_id	created_by	submitted_date	creation_date	status_id	expected_date	shipping_fee	taxes	payment_date	payment_amount	payment_method
90	1	2	2006-01-14 00:00:00	2006-01-22 00:00:00	2	NULL	0.0000	0.0000	NULL	0.0000	NULL
91	3	2	2006-01-14 00:00:00	2006-01-22 00:00:00	2	NULL	0.0000	0.0000	NULL	0.0000	NULL
92	2	2	2006-01-14 00:00:00	2006-01-22 00:00:00	2	NULL	0.0000	0.0000	NULL	0.0000	NULL
93	5	2	2006-01-14 00:00:00	2006-01-22 00:00:00	2	NULL	0.0000	0.0000	NULL	0.0000	NULL
94	6	2	2006-01-14 00:00:00	2006-01-22 00:00:00	2	NULL	0.0000	0.0000	NULL	0.0000	NULL
95	4	2	2006-01-14 00:00:00	2006-01-22 00:00:00	2	NULL	0.0000	0.0000	NULL	0.0000	NULL
96	1	5	2006-01-14 00:00:00	2006-01-22 00:00:00	2	NULL	0.0000	0.0000	NULL	0.0000	NULL

Figura 17: Alguns dados relativos a pedidos de compra (a fornecedor) da Northwind

notes	approved_by	approved_date	submitted_by
NULL	2	2006-01-22 00:00:00	2
NULL	2	2006-01-22 00:00:00	2
NULL	2	2006-01-22 00:00:00	2
NULL	2	2006-01-22 00:00:00	2
NULL	2	2006-01-22 00:00:00	2
NULL	2	2006-01-22 00:00:00	2
Purchase generated based on Order #30	2	2006-01-22 00:00:00	5

Figura 18: Alguns dados relativos a pedidos de compra (a fornecedor) da Northwind - continuação

2.5 Vertente de análise escolhida

De forma a resumir e sumarizar tudo o que foi descrito nas secções anteriores, apresentam-se algumas vertentes de análise possíveis para a NorthWind:

- Relatórios de vendas: Este relatório pode ser usado para detetar vendas por cliente, funcionário, produto e fornecedor;
- Relatório de preenchimento de solicitações: Relatório para deteção de tempos desde que um produto é pedido até que é entregue a um cliente. Útil para conseguir melhorias de tempos de entrega, e descobrir onde se situa o ponto de falha/atraso;
- Relatório ao nível do funcionário: Este relatório pode ser usado para analisar o desempenho dos funcionários, e ver como podem melhorar e progredir, seja oferecendo prémios para os melhores desempenhos, ou oferecendo treino aos piores desempenhos, ou ambos;
- Distribuição de encomendas e análise de inventário de produtos: Relatório útil para encontrar encomendas a nível global, localizar o inventário, nível de pedido, nível de reordenamento da empresa para o aprimoramento, entre outros.

Estas são algumas das potenciais áreas de negócios que podem ser melhoradas, de forma a lidar com negócios com uma gestão eficiente no que toca a vendas, tempo e melhoria de lucro.

Para o tempo de elaboração do projeto disponível, **o grupo decidiu optar por uma vertente focada nos relatórios de vendas (da Northwind a clientes), assim como analisar algumas relações da empresa com o exterior**. Mais concretamente, o grupo achou por bem analisar as seguintes vertentes:

- Relatórios de vendas/encomendas: é possível efetuar o tracking do tempo que cada encomenda demora entre as várias fases, tal como o custo destes estágios. Será possível analisar a quantidade encomendada ou o preço típico por encomenda;
- Relatórios simples ao nível do funcionário, cliente, fornecedor e expedidor: é assim possível saber quem são os funcionários com melhor desempenho de vendas, fornecedores mais populares na empresa, clientes que mais compram e expedidores que mais transportam;
- Relatórios simples de produtos: permite o conhecimento dos produtos mais afluentes ou escassos nas vendas.

3 Data Warehouse

3.1 Funcionamento do Sistema

O ETL é fundamental para qualquer iniciativa de DW. A imagem abaixo descreve qualquer processo geral de ETL e a sua aplicabilidade. No presente trabalho prático, foram extraídos dados de **apenas uma** fonte de informação, como exemplificado na figura.

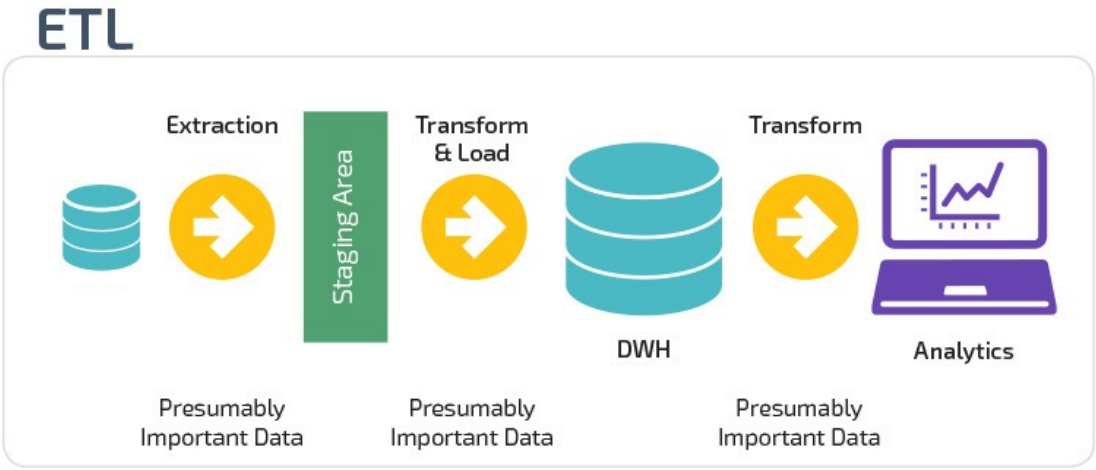


Figura 19: Processo geral de ETL

Assim, devemos tratar o ETL como sendo o "cordão umbilical" que une e possibilita a condução dos dados ao DW. Este processo inicia-se com a extração dos dados da fonte de informação para uma área de retenção. Nesta primeira etapa não é extraída a totalidade dos dados das bases de dados, mas apenas os dados pertinentes e que merecem análise. A fase seguinte inclui a limpeza, consolidação e conformidade dos dados que se encontram na AR. Aqui transformam-se os dados e adequam-se os mesmos para o DW final. Por fim ocorre a inserção dos dados preparados no data warehouse, que fica com informação tratada, com qualidade e valor para posterior para análise e apoio de decisões.

3.2 Seleção de Dados

De maneira a recolher informação útil e relevante, de acordo com a vertente de análise que o grupo escolheu, é necessário definir quais os dados que devem ser analisados, bem como as informações que se pretende extrair.

Analisando a Northwind, e de acordo com o que o grupo definiu, verifica-se a existência de alguns elementos fundamentais para a compreensão e análise das encomendas e tudo à volta destas:

- Encomendas
  - quantidade encomendada
  - preço total da encomenda
  - preço da unidade
  - preço de envio
  - desconto

Tendo isto em conta, elaborou-se um conjunto de questões possíveis de serem respondidas para obter os resultados pretendidos. é de notar que, como o grupo tem 5 medidas, o conjunto de perguntas a serem respondidas era vasto. Assim, selecionaram-se, para cada medida, apenas as questões mais pertinentes.

- Quantidade
    1. Qual a quantidade típica de um produto encomendada por mês e dia da semana (épocas mais suscetíveis a compra)?
    2. Qual a quantidade de um produto encomendada por categoria de produto (**produtos mais comprados/populares**)?
    3. Qual a quantidade de um produto encomendada por empresa de cliente?
    4. Qual a quantidade de um produto encomendada por cidade de cliente?
    5. Qual a quantidade de um produto encomendada por estado de cliente?
    6. Qual quantidade de um produto encomendada típica de ser transportada por cada expedidor (**expedidores com maior capacidade de transporte**)?
    7. Qual a quantidade de um produto encomendada típica vendida por um funcionário (**funcionários que mais quantidade vendem**)?
- Preço Total
    8. Qual o preço total típico de uma encomenda por mês e dia da semana (épocas mais suscetíveis a gastos)?
    9. Qual o preço total de encomenda por categoria de produto (produtos mais caros ou mais comprados)?
    10. Qual o preço total de encomenda por empresa de cliente (empresa que mais gasta na Northwind)?
    11. Qual o preço total de encomenda por cidade de cliente (**cidades com maior poder monetário**)?
    12. Qual o preço total de encomenda por estado de cliente?
    13. Qual o preço total de encomenda típico de ser vendida por um funcionário (**funcionários que mais lucram para a Northwind**)?
    14. Qual o preço total de encomenda de acordo com o fornecedor do produto (**fornecedores que fornecem produtos que mais rendem à Northwind**)?
- Preço da unidade
    15. Qual o preço por unidade de encomenda por categoria de produto (**produtos de luxo/bens essenciais**)?
    16. Qual o preço por unidade de encomenda típico de ser transportada por cada expedidor (expedidores que transportam produtos de luxo/bens essenciais)?
    17. Qual o preço por unidade de encomenda típico de ser vendida por um funcionário (funcionários que vendem produtos de luxo/bens essenciais)?
    18. Qual o preço por unidade de encomenda de acordo com o fornecedor do produto (fornecedores de bens de luxo/bens essenciais)?
- Preço de envio
    19. Qual o preço de envio de encomenda por categoria de produto (**produtos com envio mais caro**)?
    20. Qual o preço de envio de encomenda por cidade de cliente (locais mais caros para enviar encomendas)?

21. Qual o preço de envio de encomenda típico de cada expedidor (**expedidores mais caros**)?
22. Qual o preço de envio de encomenda de acordo com o fornecedor do produto (fornecedores que poderão de ter que diminuir/aumentar o preço de venda de acordo com o preço de envio, para maior lucro)?

- Desconto

23. Qual o desconto de encomenda por categoria de produto (**produtos mais sujeitos a desconto**)?
24. Qual o desconto de encomenda típico por funcionário (funcionários que mais vendem produtos promocionais)?

- Outras

25. Qual é o produto mais encomendado (**favoritismo de clientes em relação a produtos**)?
26. Qual é o número de encomendas por cada empresa de fornecedores?
    - Qual a empresa de fornecedores que mais vende (favoritismo de clientes em relação a fornecedores)?
27. Qual é o número de encomendas por empresa de expedidor?
    - Qual o expedidor que mais transporta (favoritismo da empresa em relação a fornecedores)?
28. Qual é o número de encomendas por funcionário?
    - Qual o funcionário que mais vende (**prémios de desempenho**)?

3.3 DataMarts

De acordo com a seleção dos dados apresentada no subtópico anterior, verifica-se a existência de um datamart fundamental, apresentado de seguida.

3.3.1 Dimensões e Factos

Com base nos exemplos de questões anteriores, é necessária a sua interpretação de forma a definir correctamente as dimensões e factos do modelo. Assim, sabendo que as dimensões são utilizadas com o objetivo de filtrar e categorizar os factos (medidas), sugerem-se as seguintes dimensões:

- Data - informações sobre a data, dia da semana, dia, mês, ano, trimestre e semestre;
- Produto - informações sobre o nome, quantidade que a unidade taz e categoria;
- Expedidor - informações sobre nome da empresa;
- Funcionário - informações sobre nome, cargo e cidade;
- Cliente - informações sobre nome, empresa, cidade e estado;
- Fornecedor - informações sobre nome da empresa.

Da mesma forma, pela interpretação das questões, identificam-se os factos que representam aquilo que se pretende analisar:

- Encomendas - quantidade, preço total, preço da unidade, preço de envio e desconto.

3.3.2 Esquema

Identificadas as perguntas a serem respondidas, as dimensões e os factos necessários à elaboração do sistema, surge o modelo dimensional em **estrela**. O grupo decidiu optar por este tipo de esquema devido à sua simplicidade de leitura, escrita e simplificação e otimização de queries (as queries são escritas com simples inner joins entre os factos e um pequeno número de dimensões; condições where servem apenas para filtrar os atributos desejados; agregações rápidas).

É de notar que a data é uma ***role-playing dimension***, uma vez que é referida múltiplas vezes na tabela de factos, com diferentes propósitos/papéis.

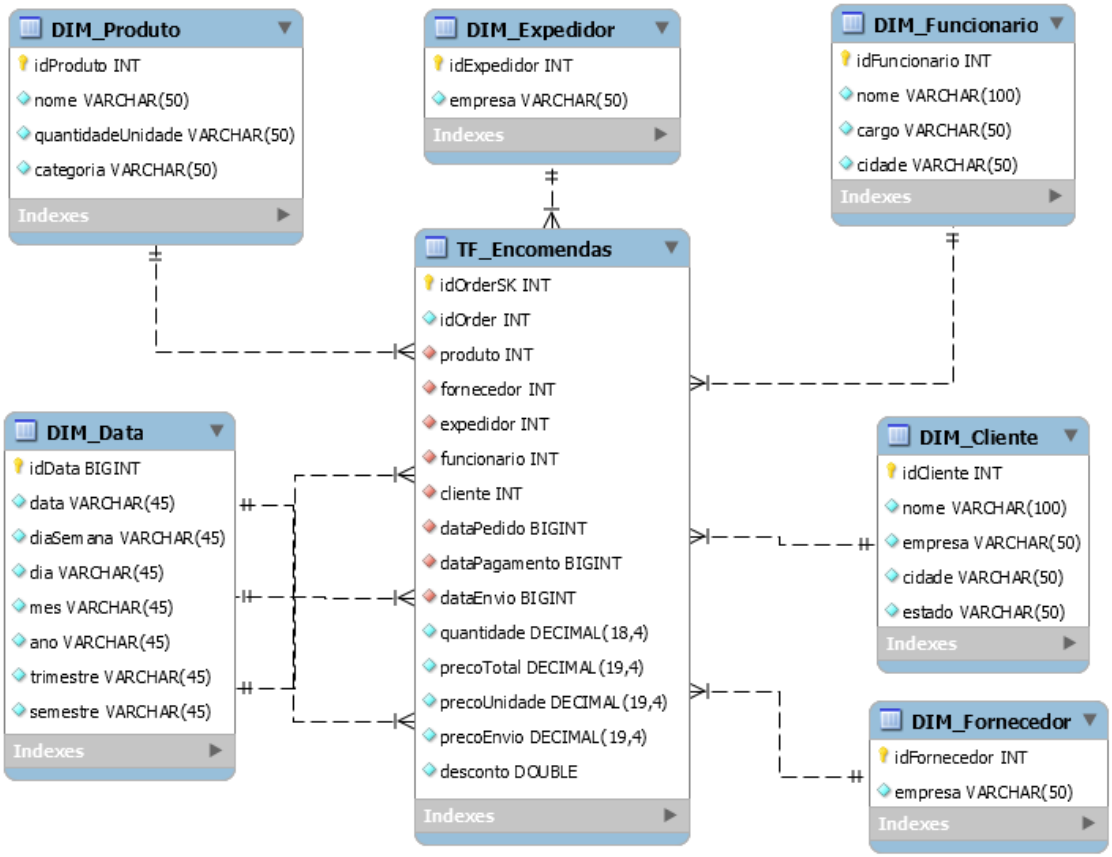


Figura 20: Modelo Dimensional

3.4 Mapa lógico de dados

Para melhor clarificar o processo de ETL, mapeando a origem e o destino dos dados, o grupo elaborou um mapa lógico de dados, demonstrado de seguida. Este demonstrou-se bastante útil, uma vez que definiu logo à partida as transformações que seriam necessárias entre os dados brutos da Northwind até aos tratados do *data warehouse*

Destino					Origem				Transformação
Tabela	Coluna	Tipo Dados	Tipo Tabela	Tipo SCD	BD	Tabela	Coluna	Tipo Dados	
DIM_Produto	idProduto	INT	Dimensão	1	Northwind	products	id	INT(11)	Direto
DIM_Produto	nome	VARCHAR(50)	Dimensão	1	Northwind	products	product_name	VARCHAR(50)	Direto
DIM_Produto	quantidade	VARCHAR(50)	Dimensão	1	Northwind	products	quantity_per_unit	VARCHAR(50)	Direto
DIM_Produto	categoria	VARCHAR(50)	Dimensão	1	Northwind	products	category	VARCHAR(50)	Direto
DIM_Expedidor	idExpedidor	INT	Dimensão	1	Northwind	shippers	id	INT(11)	Direto
DIM_Expedidor	empresa	VARCHAR(50)	Dimensão	1	Northwind	shippers	company	VARCHAR(50)	Direto
DIM_Funcionario	idFuncionario	INT	Dimensão	1	Northwind	employees	id	INT(11)	Direto
DIM_Funcionario	nome	VARCHAR(50)	Dimensão	1	Northwind	employees	first_name, last_name	VARCHAR(50)	CONCAT(first_name, " ", last_name)
DIM_Funcionario	cargo	VARCHAR(50)	Dimensão	1	Northwind	employees	job_title	VARCHAR(50)	Direto
DIM_Funcionario	cidade	VARCHAR(50)	Dimensão	1	Northwind	employees	city	VARCHAR(50)	Direto
DIM_Cliente	idCliente	INT	Dimensão	1	Northwind	costumers	id	INT(11)	Direto
DIM_Cliente	nome	VARCHAR(100)	Dimensão	1	Northwind	costumers	first_name, last_name	VARCHAR(50)	CONCAT(first_name, " ", last_name)
DIM_Cliente	empresa	VARCHAR(50)	Dimensão	1	Northwind	costumers	company	VARCHAR(50)	Direto
DIM_Cliente	cidade	VARCHAR(50)	Dimensão	1	Northwind	costumers	city	VARCHAR(50)	Direto
DIM_Cliente	estado	VARCHAR(50)	Dimensão	1	Northwind	costumers	state_province	VARCHAR(50)	Direto
DIM_Fornecedor	idFornecedor	INT	Dimensão	1	Northwind	suppliers	id	INT(11)	Direto
DIM_Fornecedor	empresa	VARCHAR(50)	Dimensão	1	Northwind	suppliers	company	VARCHAR(50)	Direto
DIM_Data	idData	BIGINT	Dimensão	1	Northwind	orders	order_date, paid_date, shipped_date	DATETIME	CAST(CONCAT(YEAR(dates),0, MONTH(dates),0,DAY(dates)) AS UNSIGNED) *a palavra "dates" indica a union das operações para as três datas
DIM_Data	data	VARCHAR(45)	Dimensão	1	Northwind	orders	order_date, paid_date, shipped_date	DATETIME	CONCAT(YEAR(dates), "-", MONTH(dates), "-", DAY(dates))
DIM_Data	diaSemana	VARCHAR(45)	Dimensão	1	Northwind	orders	order_date, paid_date, shipped_date	DATETIME	WEEKDAY(dates)
DIM_Data	dia	VARCHAR(45)	Dimensão	1	Northwind	orders	order_date, paid_date, shipped_date	DATETIME	DAY(dates)
DIM_Data	mês	VARCHAR(45)	Dimensão	1	Northwind	orders	order_date, paid_date, shipped_date	DATETIME	MONTH(dates)
DIM_Data	ano	VARCHAR(45)	Dimensão	1	Northwind	orders	order_date, paid_date, shipped_date	DATETIME	YEAR(dates)
DIM_Data	trimestre	VARCHAR(45)	Dimensão	1	Northwind	orders	order_date, paid_date, shipped_date	DATETIME	QUARTER(dates)
DIM_Data	semestre	VARCHAR(45)	Dimensão	1	Northwind	orders	order_date, paid_date, shipped_date	DATETIME	FLOOR((QUARTER(dates)/3)+1)
TF_Encomendas	idOrderSK	INT	Factos	N/A				INT	Surrogate key

Figura 21: Mapa lógico de dados

TF_Encomendas	idOrder	INT	Factos	N/A	Northwind	orders	id	INT(11)	Direto	WHERE order_details.order_id = orders.id
TF_Encomendas	produto	INT	Factos	N/A	Northwind	order_details	product_id	INT(11)		
TF_Encomendas	fornecedor	INT	Factos	N/A	Northwind	products	supplier_ids	LONGTEXT	SELECT CAST(SUBSTRING_INDEX (p.supplier_ids,';',1) AS UNSIGNED) FROM products p, order_details od, orders d WHERE od.order_id = o.id AND od.product_id = p.id UNION SELECT CAST(SUBSTRING_INDEX (p.supplier_ids,';',-1) AS UNSIGNED) FROM products p, order_details od, orders d WHERE od.order_id = o.id AND od.product_id = p.id	
TF_Encomendas	expedidor	INT	Factos	N/A	Northwind	orders	shipper_id	INT(11)	Direto	
TF_Encomendas	funcionário	INT	Factos	N/A	Northwind	orders	employee_id	INT(11)	Direto	
TF_Encomendas	cliente	INT	Factos	N/A	Northwind	orders	costumer_id	INT(11)	Direto	
TF_Encomendas	dataPedido	BIGINT	Factos	N/A	Northwind	orders	order_Date	DATETIME	CAST(CONCAT(YEAR(orders.order_date),0, MONTH(orders.order_date),0, DAY(orders.order_date)) AS UNSIGNED)	
TF_Encomendas	dataPagamento	BIGINT	Factos	N/A	Northwind	orders	paid_Date	DATETIME	CAST(CONCAT(YEAR(orders.paid_date),0, MONTH(orders.paid_date),0, DAY(orders.paid_date)) AS UNSIGNED	
TF_Encomendas	dataEnvio	BIGINT	Factos	N/A	Northwind	orders	shipped_Date	DATETIME	CAST(CONCAT(YEAR(orders.shipped_date),0, MONTH(orders.shipped_date),0, DAY(orders.shipped_date)) AS UNSIGNED	
TF_Encomendas	quantidade	DECIMAL(18,4)	Factos	N/A	Northwind	order_details	quantity	DECIMAL(18,4)	WHERE order_details.order_id = orders.id	
TF_Encomendas	precoTotal	DECIMAL(19,4)	Factos	N/A	Northwind	order_Details	unit_price, quantity, discount	DECIMAL(18,4), DECIMAL(19,4), DOUBLE	SELECT (od.unit_price*od.quantity) - (od.unit_price*od.quantity*od.discount) FROM orders o, order_details od WHERE od.order_id = o.order_id	
TF_Encomendas	precoUnidade	DECIMAL(19,4)	Factos	N/A	Northwind	order_Details	unit_price	DECIMAL(19,4)	WHERE order_details.order_id = orders.id	
TF_Encomendas	precoEnvio	DECIMAL(19,4)	Factos	N/A	Northwind	orders	shipping_fee	DECIMAL(19,4)	Direto	
TF_Encomendas	desconto	DOUBLE	Factos	N/A	Northwind	order_Details	discount	DOUBLE	WHERE order_details.order_id = orders.id	

Figura 22: Mapa lógico de dados - continuação

3.5 Preenchimento

Uma vez definido e implementado o *data warehouse*, foi necessário efetuar o povoamento inicial, com todos os dados (passados) que se encontravam na Northwind. Para tal, o grupo usou o *software* Pentaho Data Integration (Kettle), responsável por tratar de todos os passos do ETL, desde a extração da Northwind até ao carregamento nos *datamarts* definidos anteriormente.

Para suportar as fases do ETL, o grupo criou uma área de retenção (demonstrada abaixo), que se trata da área de transição dos dados entre a BD inicial e o DW final. Assim, o processo de preenchimento resume-se a: (1) Extrair dados da Northwind para a Área de Retenção, (2) Transformar os dados na Área de Retenção, (3) conciliar e extrair os dados da Área de Retenção para o DW final. Esta *staging area* tem atributos e tipos iguais aos finais do DW.

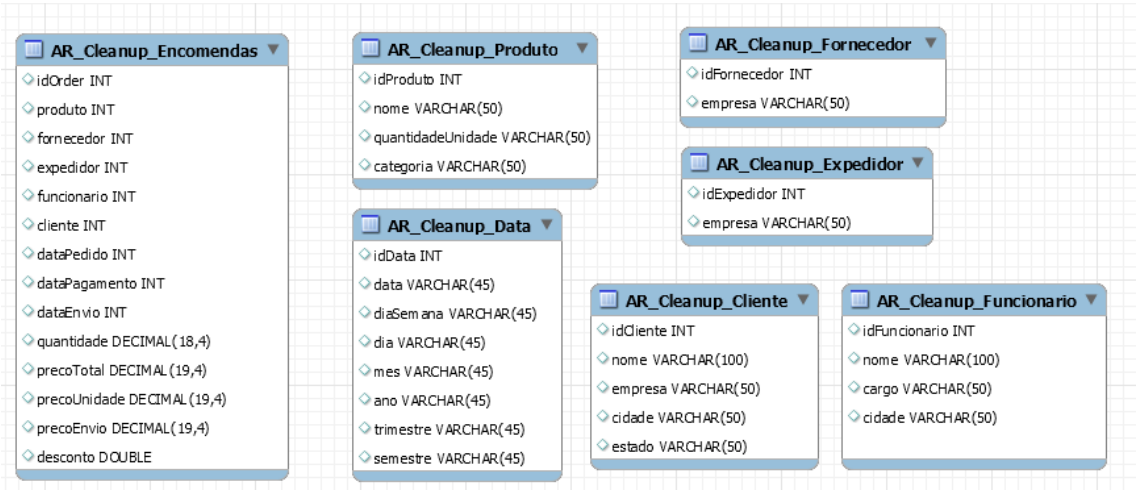


Figura 23: Esquema da área de retenção

3.5.1 Extração e Transformação

Mostram-se, de seguida, as transformações necessárias à extração dos dados da Northwind para a área de retenção. Nesta fase já se efetua parte da transformação dos dados, correspondende à **limpeza** e operações de agregação e filtragem.

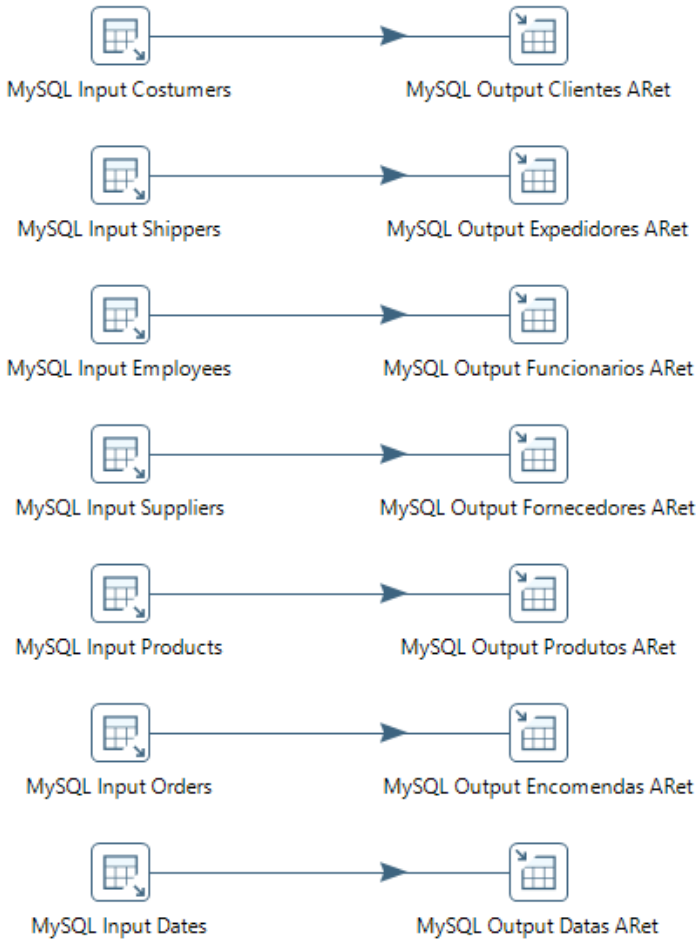


Figura 24: Transformação correspondente à extração

Assim, começa-se por extrair os dados das tabelas originais da Northwind "Costumers", "Shippers", "Employees", "Suppliers", "Products"e "Orders". Esta fase baseia-se em simples *selects*, *joins* das tabelas originais e operações de agregação e concatenação dos dados, de forma a compatibilizar com os que estão na área de retenção, que já correspondem aos tipos finais. Também as "Dates"a extrair são todas as correspondentes às presentes nas encomendas da Northwind. Nesta etapa, transforma-se já o id da data num número inteiro do tipo 20060105.

Depois de tudo isto, os dados são armazenados na área de retenção nas tabelas correspondentes às que vão ser as dimensões e factos. Para o **tratamento dos nulos** foi elaborado um *trigger before insert*. Este trigger foi necessário já que existiam três tipos de dados desconhecidos na Northwind que afetavam o trabalho do grupo:

- Nulos no atributo 'Quantidade' dos Produtos;
- Nulos nas várias datas das 'Encomendas';
- Nulos nos expedidores.

Desta forma, qualquer quantidade desconhecida nos produtos foi substituída pela string 'desconhecido'.

```
DROP TRIGGER IF EXISTS nullTreatmentProd;
DELIMITER //

CREATE TRIGGER nullTreatmentProd
BEFORE INSERT ON arearet.ar_cleanup_produto FOR EACH ROW
BEGIN

    IF(NEW.quantidadeUnidade IS NULL)
    THEN SET NEW.quantidadeUnidade = 'desconhecido';
    END IF;

END;
//

DELIMITER ;
```

Figura 25: 1º trigger para tratamento dos nulos

As datas das encomendas têm todas um id inteiro no formato 20060105. No que toca às datas desconhecidas, o grupo optou por lhes atribuir o id 0. Também nos expedidores em falta, foi decidido que o inteiro 0 identificaria o desconhecido, já que os ids começam sempre no número 1. Assim, o 0 ficou reservado para estes casos, atuando como um 'código' de erro.

```
-- Número 0 fica reservado para código de erro
-- Não há datas nem expedidores que possam tomar o valor 0
-- (Expedidor chaves começam a 1 e datas são inteiros no formato 20060101)
DROP TRIGGER IF EXISTS nullTreatmentOrders;
DELIMITER //

CREATE TRIGGER nullTreatmentOrders
BEFORE INSERT ON arearet.ar_cleanup_encomendas FOR EACH ROW
BEGIN

    IF(NEW.dataPagamento IS NULL)
    THEN SET NEW.dataPagamento = 0;
    END IF;

    IF(NEW.dataEnvio IS NULL)
    THEN SET NEW.dataEnvio = 0;
    END IF;

    IF(NEW.expedidor IS NULL)
    THEN SET NEW.expedidor = 0;
    END IF;

END;
//
```

Figura 26: 2º trigger para tratamento dos nulos

Por fim, para estes ids serem todos reconhecidos no data warehouse e ligados às respectivas dimensões, tanto no povoamento inicial como em casos futuros, foram inseridos na área de retenção uma data com id=0 e atributos todos a 'desconhecido', e um expedidor com id=0 e atributos a 'desconhecido'.

```
INSERT INTO ar_cleanup_data
VALUES (0, 'desconhecido', 'desconhecido', 'desconhecido', 'desconhecido',
        'desconhecido', 'desconhecido', 'desconhecido');


INSERT INTO ar_cleanup_expedidor VALUES (0,'desconhecido');
```

Figura 27: Inserts necessários para os relacionamentos posteriores corretos no DW

3.5.2 Carregamento

Uma vez tratados os dados, o carregamento é direto das tabelas da área de retenção para as dimensões e factos do DW. De forma a identificar unicamente cada registo de encomenda na tabela de factos, foi efetuada a **conciliação** dos dados, que ocorre ao mesmo tempo do *load*. Assim, ao serem inseridos os dados na TF, cada registo fica com uma *surrogate key* a ele associada.

É importante notar que, antes de se efetuar o povoamento da tabela de factos, é necessário que todas as dimensões estejam preenchidas. Justifica-se, assim, o hop desabilitado na última transformação da imagem seguinte, que foi ativado posteriormente.

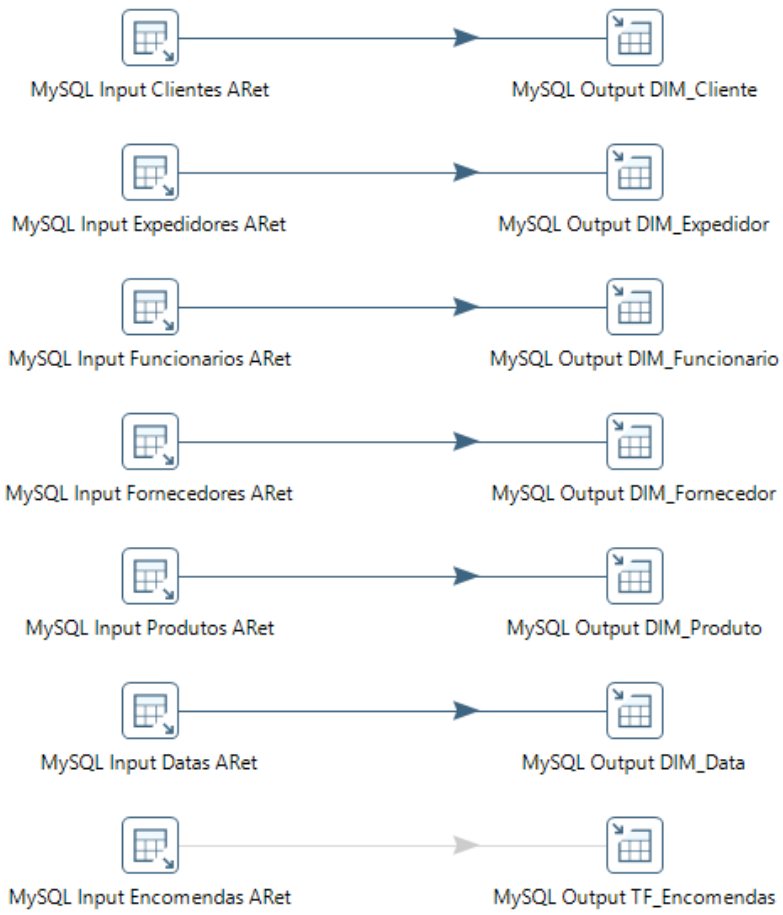


Figura 28: Transformação correspondente ao carregamento

3.6 Refrescamento dos dados

Um DW deve ser atualizado regularmente, de modo a garantir que as informações derivadas deste sejam atualizadas. Assim, o grupo decidiu implementar um sistema de refrescamento **diferencial e incremental**. Desta forma, os dados são carregados na totalidade da Northwind para a área de retenção, mas apenas são carregados para o DW os dados que foram inseridos de novo ou atualizados. Como não é considerado histórico de dados, qualquer dado modificado na BD original é também modificado no DW. Por exemplo, se um cliente mudar a sua localização, o DW não manterá um histórico de moradas, e apenas considera os dados atuais.

Assim sendo, o grupo criou uma transformação no Pentaho para limpar a área de retenção, demonstrada de seguida.

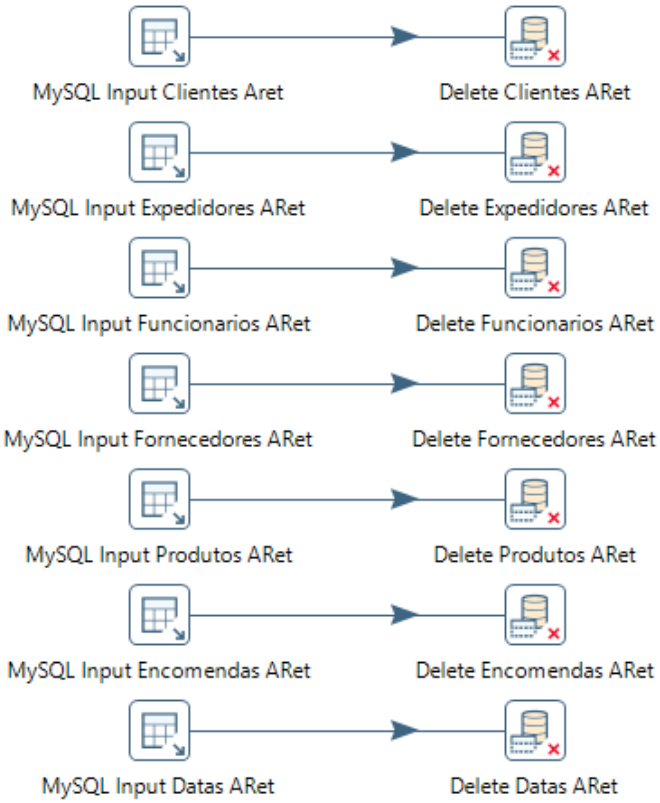


Figura 29: Limpeza dos dados da área de retenção

De seguida, é efetuada novamente a extração dos dados da Northwind para a área de retenção. Esta extração é, naturalmente, efetuada da mesma maneira do povoamento inicial, uma vez que o grupo optou pelo refrescamento diferencial e incremental, onde os dados movem-se na totalidade para a AR.

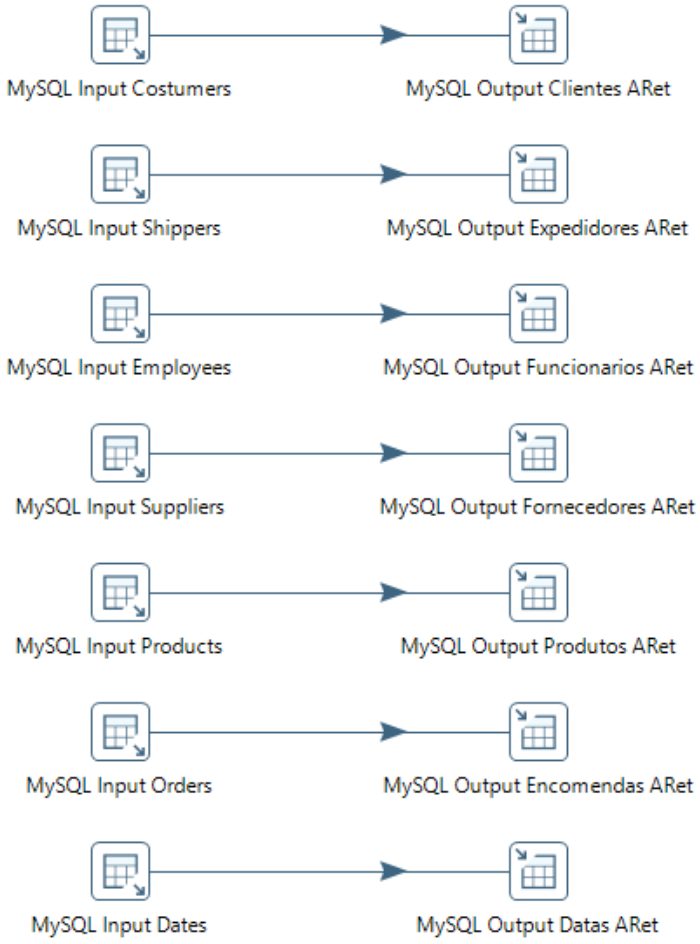


Figura 30: Transformação correspondente à extração

A transformação decorre da mesma maneira do povoamento inicial, onde os nulos são tratados e é garantida a conformidade dos dados.

A próxima fase é o carregamento para o DW. Este carregamento baseia-se em transformações de *insert/update*, para corresponder ao tipo de refrescamento optado pelo grupo.

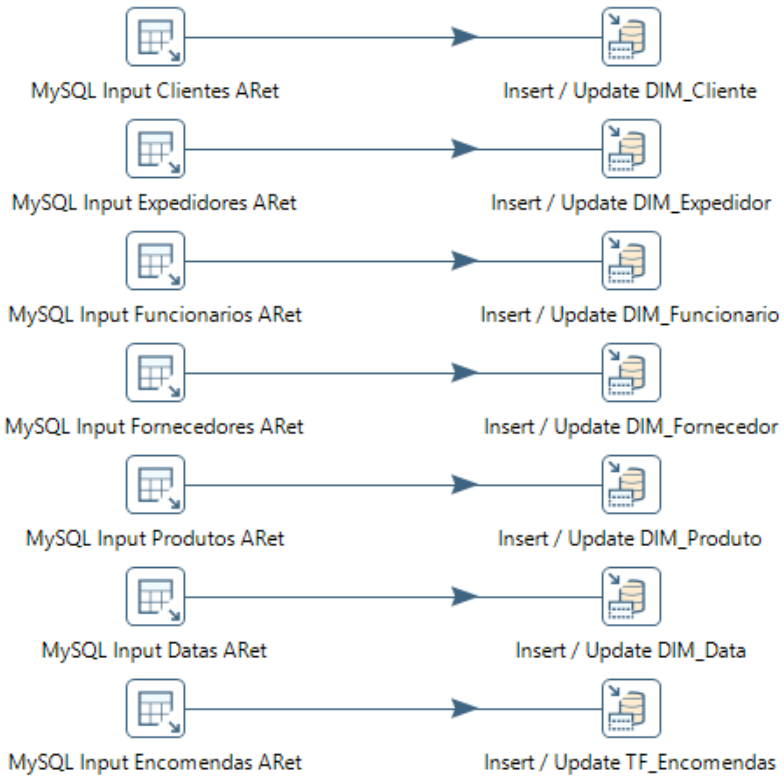


Figura 31: Transformação correspondente ao carregamento no refrescamento

Desta maneira, um dado atualizado é updated também no DW, e um dado inserido, é inserted como um novo registo. Os dados da base de dados original que já se encontravam no data warehouse e que não sofreram mudanças, não são alterados.

Nota: poderia ser utilizada esta transformação, quer para o povoamento inicial, quer para o refrescamento. Poderia também ser definida uma transformação de insert/update para as dimensões, e outra de insert na tabela de factos. Desta forma, na TF só ocorrem inserts, tal como na solução corrente, e nas DIM ocorre insert no povoamento inicial e insert/update nos refrescamentos.

3.6.1 Testes

Para testar o refrescamento diferencial e incremental, o grupo criou os dois seguintes testes:

- Atualizar a morada de um cliente;
- Adicionar uma nova encomenda.

Para o teste do update, atualizou-se a morada de um cliente, originalmente de Minneapolis (MN), para New York (NY).

```
-- TEST #1: O cliente com o id = 5 morava originalmente em Minneapolis, MN.
--          Mudou-se para cidade = 'New York' e estado = 'NY'
UPDATE customers
SET city = 'New York', state_province = 'NY'
WHERE id = 5;


-- REVERT #1
UPDATE customers
SET city = 'Minneapolis', state_province = 'MN'
WHERE id = 5;

select * from customers;
```

Figura 32: Atualização da morada de um cliente (cidade + estado)

Para testar a inserção de apenas uma nova encomenda, adicionou-se à BD Northwind um novo registo da tabela 'Orders'. Como o produto é obtido através da tabela 'Order_details', foi também preciso adicionar um registo à mesma.

```
-- TEST #2: Nova encomenda: - O id é AI (vai ser 82 porque o último id na BD é 81)
--                          - employee_id: 4
--                          - costumer_id: 3
--                          - Order_date: '2018-04-25 17:05:10'
--                          - Shipped_date: '2018-10-01 10:33:54'
--                          - Expedidor q realizou o transporte: id = 2
--                          - Ship_name: 'Soo Jung Lee'
--                          - Ship_address: '789 29th Street'
--                          - Ship_city: 'Denver'
--                          - Ship_state_province: 'CO'
--                          - Ship_zip_postal_code: '99999'
--                          - Ship_country_region: 'USA'
--                          - Shipping_fee: 7.0000
--                          - Taxes: 0.0000
--                          - Payment_type: 'Credit Card'
--                          - Paid_date: '2018-04-25 18:10:01'
--                          - Notes: null
--                          - Tax_rate: 0
--                          - Tax_status_id: null
--                          - Status_id: 3
INSERT INTO orders (employee_id, customer_id, order_date, shipped_date, shipper_id,
                    ship_name, ship_address, ship_city, ship_state_province, ship_zip_postal_code,
                    ship_country_region, shipping_fee, taxes, payment_type, paid_date, notes,
                    tax_rate, tax_status_id, status_id)
VALUES (4, 3, '2018-04-25 17:05:10', '2018-10-01 10:33:54', 2, 'Soo Jung Lee', '789 29th Street',
       'Denver', 'CO', '99999', 'USA', 7.0000, 0.0000, 'Credit Card', '2018-04-25 18:10:01', null,
       0, null, 3);
```

Figura 33: Inserção de uma nova encomenda

```
-- REVERT #2
DELETE FROM orders WHERE id = 82;
ALTER TABLE orders AUTO_INCREMENT = 81;


select * from orders;


-- é preciso preencher o order_Details também porque é onde está o id produto das orders
-- processo de extração retira dados da tabela orders e order_details
-- id é AI: vai ser 92 porque o último na tabela é 91
INSERT INTO order_details (order_id, product_id, quantity, unit_price, discount, status_id,
                           date_allocated, purchase_order_id, inventory_id)
VALUES (82, 34, 100.0000, 14.0000, 0, 2, null, null, null);


-- REVERT
DELETE FROM order_details WHERE id = 92;
ALTER TABLE order_details AUTO_INCREMENT = 91;


select * from order_details;
```

Figura 34: Inserção de uma nova encomenda - continuação

Executando os passos indicados na secção anterior, obtém-se o seguinte resultado de execução no Pentaho, que indica o sucesso dos testes. Observa-se a inserção (sublinhado a amarelo na coluna 'Saída') de uma nova encomenda, com duas novas datas. Foram apenas duas novas datas inseridas uma vez que a data de pedido coincide com a data de pagamento, inserindo apenas uma destas juntamente com a data de envio. É visível também (sublinhado a amarelo na coluna 'Atualizados') apenas o update de um registo, que corresponde à nova morada do cliente.

Execution Results

Execution History | Logging | Step Metrics | Performance Graph | Metrics | Preview data

#	Nome do step	Copia nr	Lidos	escritos	Entrada	Sa◆da	Atualizados	Rejected	Erros	Ativo	Tempo	Velocidade (r/s)
8	Insert / Update DIM_Expedidor	0	4	4	4	0	0	0	0	Finished	0.1s	42
9	Insert / Update DIM_Funcionario	0	9	9	9	0	0	0	0	Finished	0.1s	96
10	Insert / Update DIM_Fornecedor	0	10	10	10	0	0	0	0	Finished	0.1s	103
11	Insert / Update DIM_Produto	0	45	45	45	0	0	0	0	Finished	0.1s	369
12	Insert / Update DIM_Data	0	35	35	35	2	0	0	0	Finished	0.2s	183
13	Insert / Update TF_Encomendas	0	68	68	68	1	0	0	0	Finished	0.4s	153
14	Insert / Update DIM_Cliente	0	29	29	29	0	1	0	0	Finished	0.3s	92

Figura 35: Resultado da nova inserção + novo update

Para terminar, mostram-se os dados concretos, no DW, correspondentes à inserção da nova encomenda e update da morada do cliente. Ainda mais, compara-se o antes e depois do refrescamento de forma a melhorar o teste.

Como se vê, o cliente com o id = 5 possuía os seguintes atributos:

idCliente	nome	empresa	cidade	estado
1	Anna Bedecs	Company A	Seattle	WA
2	Antonio Gratacos Solsona	Company B	Boston	MA
3	Thomas Axen	Company C	Los Angelas	CA
4	Christina Lee	Company D	New York	NY
5	Martin O'Donnell	Company E	Minneapolis	MN

Figura 36: Morada do cliente antes do refrescamento

Com o refrescamento, os valores a sublinhado foram atualizados.

idCliente	nome	empresa	cidade	estado
1	Anna Bedecs	Company A	Seattle	WA
2	Antonio Gratacos Solsona	Company B	Boston	MA
3	Thomas Axen	Company C	Los Angelas	CA
4	Christina Lee	Company D	New York	NY
5	Martin O'Donnell	Company E	New York	NY

Figura 37: Morada do cliente depois do refrescamento

Quanto aos dados das encomendas, provenientes do povoamento inicial, o DW tinha o seguinte aspeto:

idOrderSK	idOrder	produto	fornecedor	expedidor
48	70	8	8	3
49	69	80	2	1
50	67	74	2	2
51	60	72	5	3
52	63	3	10	2
53	63	8	8	2
54	58	20	2	1
55	58	52	1	1
56	80	56	1	0
57	81	81	3	0
58	81	56	1	0
59	32	43	4	2
60	38	43	4	3
61	41	43	4	0
62	42	6	6	1
63	44	43	4	0
64	77	6	6	3
65	72	43	4	3
66	67	74	6	2
67	58	20	6	1

(...)

Figura 38: Conteúdo da tabela de factos antes do refrescamento

Com a inserção da nova encomenda, com o id = 82, foi adicionado mais um registo.

idOrderSK	idOrder	produto	fornecedor	expedidor
49	69	80	2	1
50	67	74	2	2
51	60	72	5	3
52	63	3	10	2
53	63	8	8	2
54	58	20	2	1
55	58	52	1	1
56	80	56	1	0
57	81	81	3	0
58	81	56	1	0
59	32	43	4	2
60	38	43	4	3
61	41	43	4	0
62	42	6	6	1
63	44	43	4	0
64	77	6	6	3
65	72	43	4	3
66	67	74	6	2
67	58	20	6	1
68	82	34	4	2

(...)

Figura 39: Conteúdo da tabela de factos depois do refrescamento

4 Business Intelligence

Uma vez implementado e povoado o *data warehouse*, o grupo usou o *Microsoft Power BI* de modo a visualizar os dados e tirar alguma informação útil dos mesmos. Esta ferramenta de análise e visualização de dados permite avaliar os dados de uma forma interativa e visual, com a vantagem da partilha de dashboards entre todos os elementos do grupo.

Desta forma, com o objetivo de responder às perguntas apresentadas na secção de seleção de dados (sec. 3.2), o grupo criou um relatório referente ao *datamart* existente, em que cada página do mesmo, é referente ao estudo de uma medida da tabela de factos.

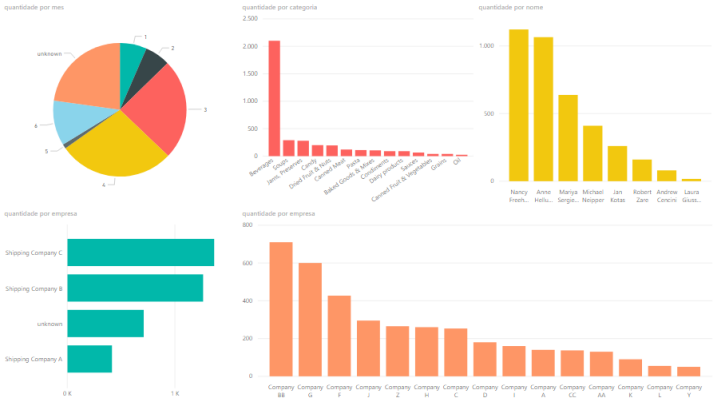


Figura 40: Análise de encomendas por quantidade

Figura 41: Análise de encomendas por preço total

Figura 42: Análise de encomendas por preço de unidade

Figura 43: Análise de encomendas por preço de envio

Figura 44: Análise de encomendas

5 Análise de Resultados

Para a análise de resultados, respondem-se, nesta secção, a todas as perguntas da secção 3.2 com o gráfico obtido, assim como uma interpretação textual do mesmo gráfico.

Começa-se, então, pela medida "quantidade".

1. Qual a quantidade típica de um produto encomendada por mês e dia da semana (épocas mais suscetíveis a compra)?

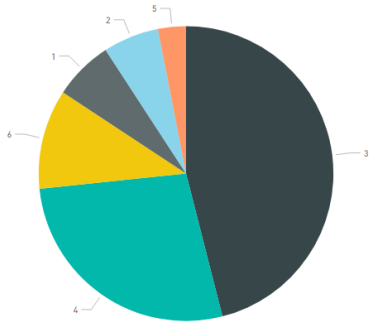


Figura 45: Soma da quantidade de produto por mês

- Como se pode ver através do gráfico, da soma total das quantidades de produtos transacionados, cerca de 46,03% foram pedidas no mês de Março (3), e 27,32% foram pedidas no mês de Abril (4). O que indica que nestes dois meses, o fluxo de produtos pretendidos é muito maior no que nos restantes juntos. - É também de salientar que não existem quaisquer dados referentes a encomendas processadas no segundo semestre do ano, na base de dados original.

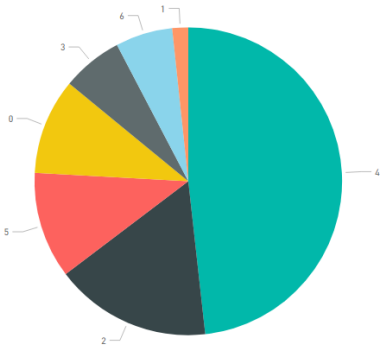


Figura 46: Soma da quantidade de produto por dia da semana

- Com este gráfico concluímos que cerca de 48,24% da soma das quantidades de produtos, é requerido à uma quarta-feira (4), que é quase o mesmo que todos os outros dias da semana juntos.

2. Qual a quantidade de um produto encomendada por categoria de produto (**produtos mais comprados/populares**)?

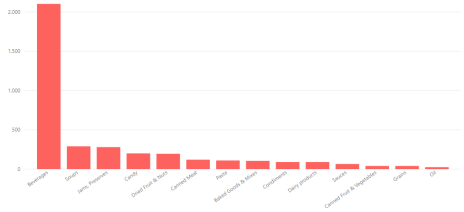


Figura 47: Soma da quantidade de produto por categoria

- O tipo de produtos que é mais pedido em maiores quantidades, é o das bebidas, com uma soma total na casa dos 2 milhares. Sendo que este apresenta uma soma de quantidades muito superior, em relação às outras categorias.

3. Qual a quantidade de um produto encomendada por empresa de cliente?

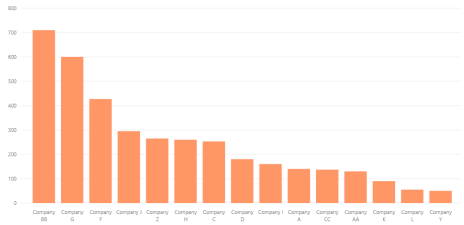


Figura 48: Soma da quantidade de produto por empresa cliente

- A empresa que mais produtos pediu para encomendar em relação à sua quantidade, é a *Company BB*, com uma soma total de 710 unidades. Esta mesma é seguida pela *Company G* com 600 unidades, e a *Company F* com 427 unidades.

4. Qual a quantidade de um produto encomendada por cidade de cliente?

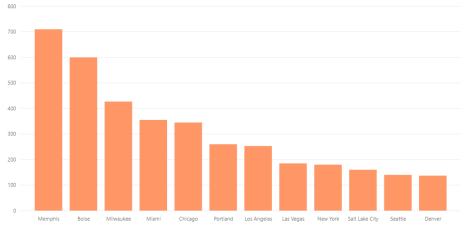


Figura 49: Soma da quantidade de produto por cidade da empresa cliente

- A cidade para a qual mais produtos foram pretendidos em relação à sua quantidade, é a *Memphis*, com uma soma total de 710 unidades (cidade onde é sediada a *Company BB*). Esta mesma é seguida por *Boise* com 600 unidades (cidade onde é sediada a *Company G*), e por *Milwaukee* com 427 unidades (cidade onde é sediada a *Company F*). - É de reparar que o número de cidades não indicam o mesmo número de empresas clientes, o que sugere que existem empresas sediadas na mesma cidade.

5. Qual a quantidade de encomenda por estado de cliente?

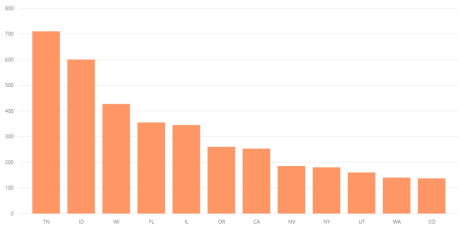


Figura 50: Soma da quantidade de produto por estado da empresa cliente

- O estado para o qual mais produtos foram pretendidos em relação à sua quantidade, é o estado de *TN*, com uma soma total de 710 unidades (cidade onde é sediada a *Company BB*). Esta mesma é seguida por *ID* com 600 unidades (cidade onde é sediada a *Company G*), e por *WI* com 427 unidades (cidade onde é sediada a *Company F*). - Neste gráfico dá para evidenciar que o número de cidades é igual ao número de estados, o que indica que neste sistema não existem duas cidades do mesmo estado.

6. Qual quantidade de um produto encomendada típica de ser transportada por cada expedidor (**expedidores com maior capacidade de transporte**)?

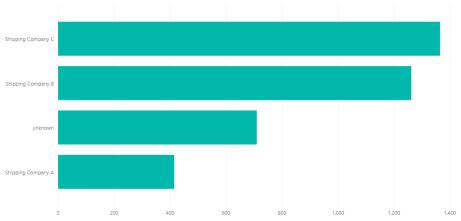


Figura 51: Soma da quantidade de produto por expedidor

- Neste ponto dá para ver que a *Shipping Company C* é responsável por transportar cerca de 1365 unidades, e com este valor é a empresa que mais transportou produtos em relação ao número de unidades. - Cerca de 710 unidades, foram transportadas sem que seja conhecido a companhia que fez esse efeito.

7. Qual a quantidade de um produto encomendada típica vendida por um funcionário (**funcionários que mais quantidade vendem**)?

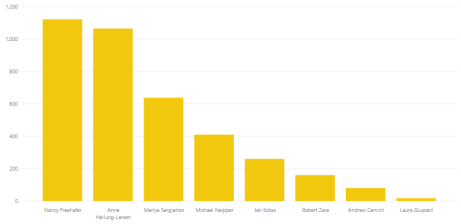


Figura 52: Soma da quantidade de produto por funcionário

- A nível de vendas de produtos em quantidade, os funcionários que se destacam são a *Nancy Freehafer* com 1122 unidades, a *Anne Hellung-Larsen* com 1065 unidades e a *Mariya Sergienko* com 638 unidades.

Enquandram-se, agora, os indicadores relativos à medida "preço total".

8. Qual o preço total típico de uma encomenda por mês, dia da semana (épocas mais suscetíveis a gastos)?

Figura 53: Média do preço total de encomenda por mês

- Como se pode ver através do gráfico, o mês de Março (3), foi o mês onde foram encomendados os produtos cujo o preço total gera mais receitas. O que indica que nesse mês, o fluxo monetário é quase superior ao dos restantes meses juntos. - É também de salientar que não existem quaisquer dados referentes a encomendas processadas no segundo semestre do ano, na base de dados original.

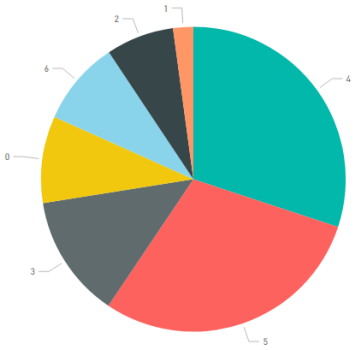


Figura 54: Média do preço total de encomenda por dia da semana

- Com este gráfico concluímos que cerca de 30,07% da receita gerada por produto, foi a uma quarta-feira (4). O segundo dia é a quinta feira com 29,36% da receita gerada por produto.

9. Qual o preço total de encomenda por categoria de produto (produtos mais caros ou mais comprados)?

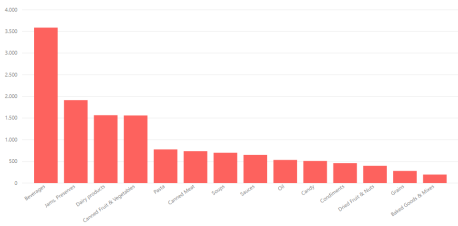


Figura 55: Média do preço total de encomenda por categoria de produto

- O tipo de produtos que gera mais receita por encomenda é o das bebidas, com uma média de 3587,38 unidades monetárias. Sendo que esse mesmo mês apresenta quase o dobro das receitas por produto em relação ao segundo classificado.

10. Qual o preço total de encomenda por empresa de cliente (empresa que mais gasta na Northwind)?

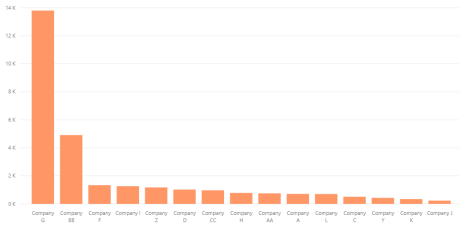


Figura 56: Média do preço total de encomenda por empresa cliente

- A empresa cliente que mais dinheiro gastou em média por produto, é a *Company G*, com um preço total por produto de 13800 unidades monetárias. Esta mesma é seguida pela *Company BB* com 4910,42 unidades monetárias por produto, e a *Company F* com 1334,58 unidades monetárias por produto.

11. Qual o preço total de encomenda por cidade de cliente (**cidades com maior poder monetário**)?

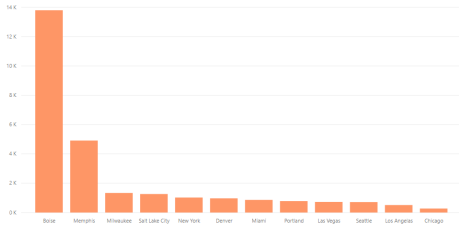


Figura 57: Média do preço total de encomenda por cidade da empresa cliente

- A cidade que mais dinheiro gastou em média por produto, é a *Boise* com um preço total por produto de 13800 unidades monetárias (cidade onde é sediada a *Company G*). Esta mesma é seguida por *Memphis*, com 4910,42 unidades monetárias por produto (cidade onde é sediada a *Company BB*), e por *Milwaukee* com 1334,58 unidades monetárias por produto (cidade onde é sediada a *Company F*). - É de reparar que o número de cidades não indicam o mesmo número de empresas clientes, o que sugere que existem empresas sediadas na mesma cidade.

12. Qual o preço total de encomenda por estado de cliente?

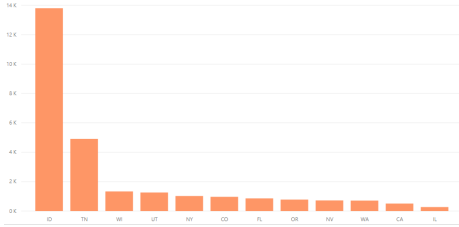


Figura 58: Média do preço total de encomenda por estado da empresa cliente

- O estado que mais dinheiro gastou em média por produto, é o *ID* com um preço total por produto de 13800 unidades monetárias (cidade onde é sediada a *Company G*). Esta mesma é seguida por *TN*, com 4910,42 unidades monetárias por produto (cidade onde é sediada a *Company BB*), e por *WI* com 1334,58 unidades monetárias por produto (cidade onde é sediada a *Company F*). - Neste gráfico dá para evidenciar que o número de cidades é igual ao número de estados, o que indica que neste sistema não existem duas cidades do mesmo estado.

13. Qual o preço total de encomenda típico de ser vendida por um funcionário (**funcionários que mais lucram para a Northwind**)?

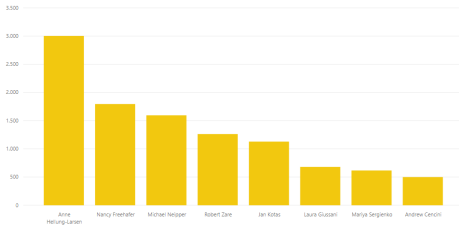


Figura 59: Média do preço total de encomenda por funcionário

- A nível de receita média gerada por produto, os funcionários que se destacam são a *Anne Hellung-Larsen* com 3002,02 unidades monetárias por produto, a *Nancy Freehafer* com 1794,54 unidades monetárias por produto, e o *Michael Neipper* com 1594,50 unidades monetárias por produto.

14. Qual o preço total de encomenda de acordo com o fornecedor do produto (**fornecedores que fornecem produtos que mais rendem à Northwind**)?

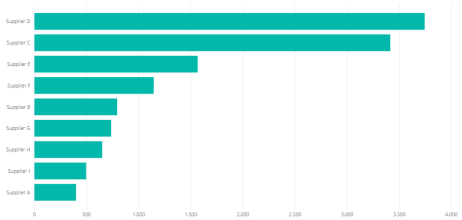


Figura 60: Média do preço total de encomenda por fornecedor

- No que toca a preço total por produto, os fornecedores que ocupam os dois primeiros lugares são a *Supplier D* com 3743,80 unidades monetárias por produto, e a *Supplier C* com 3413,58 unidades monetárias por produto. - Estes dois apresentam uma barra mais do que duas vezes maior em relação aos restantes fornecedores.

Apresentam-se, de seguida, os indicadores referentes ao "preço por unidade".

15. Qual o preço por unidade de encomenda por categoria de produto (**produtos de luxo/bens essenciais**)?

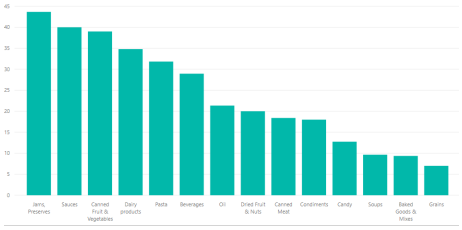


Figura 61: Média do preço da unidade por categoria do produto

- Com base nos resultados obtidos pelo gráfico, os produtos que podem ser considerados como bens mais luxuosos, são as compotas e as conservas, que possuem um preço por unidade em média de 43,67 unidades monetárias. - Por outro lado os produtos que podem ser considerados como bens de mais necessidade, são os cereais com um preço por unidade em média de 7 unidades monetárias.

16. Qual o preço por unidade de encomenda típico de ser transportada por cada expedidor (expedidores que transportam produtos de luxo/bens essenciais)?

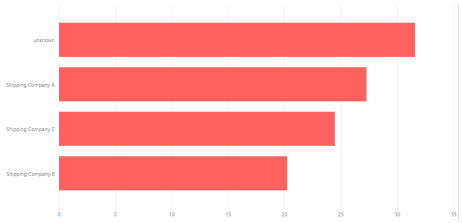


Figura 62: Média do preço da unidade por expedidor

- A *Shipping Company A* é a transportadora que transporta por norma os bens mais caros, com uma média de 27,26 unidades monetárias por produto. - Por outro lados, a *Shipping Company B* é que por norma transporta os bens mais baratos, com uma média de 20,23 unidade monetárias por produto. - É também de referir que os produtos mais caros não têm transportadora referida nas fontes de dados.

17. Qual o preço por unidade de encomenda típico de ser vendida por um funcionário (funcionários que vendem produtos de luxo/bens essenciais)?

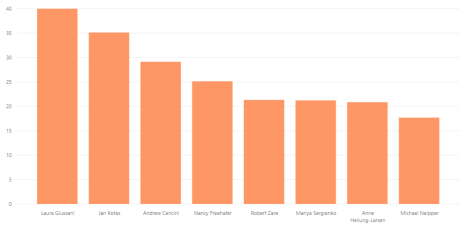


Figura 63: Média do preço da unidade por funcionário

- O funcionário que vende os produtos mais caros é a *Laura Giussani*, com uma média de 40 unidades monetárias por produto. - Por outro lado o funcionário que em média vende os

produtos mais baratos é o *Michael Neipper*, com uma média de 17,69 unidades monetárias por produto.

18. Qual o preço por unidade de encomenda de acordo com o fornecedor do produto (fornecedores de bens de luxo/bens essenciais)?

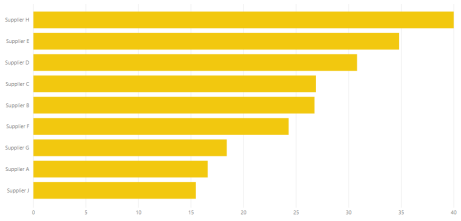


Figura 64: Média do preço da unidade por fornecedor

- O fornecedor dos produtos mais caros é a *Supplier H*, com uma média 40 unidades monetárias por produto. - O fornecedor dos produtos mais baratos é a *Supplier J*, com uma média 15,46 unidades monetárias por produto.

Mostram-se, agora, todos os indicadores referentes ao "preço de envio".

19. Qual o preço de envio de encomenda por categoria de produto (**produtos com envio mais caro**)?

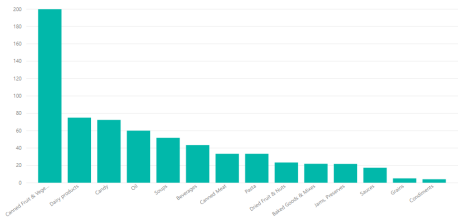


Figura 65: Média do preço de envio por categoria de produto

- Com base nos resultados obtidos pelo gráfico, os produtos que possuem o transporte mais caros são as frutas e os vegetais enlatados, com uma média de 200 unidades monetárias. - Por outro lado os produtos que que possuem o transporte mais barato são os temperos, com uma média de 4 unidades monetárias.

20. Qual o preço de envio de encomenda por cidade de cliente (locais mais caros para enviar encomendas)?

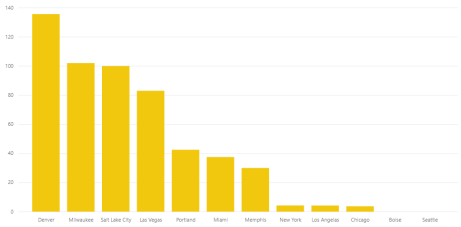


Figura 66: Média do preço de envio por cidade de empresa cliente

- A cidade que provoca maior despesa em enviar produtos é *Denver*, com uma média 135,67 unidades monetárias de envio do produto. - As cidades que provocam menor despesa em

enviar produtos são *Seatle* e *Boise*, em que não apresentam sequer despesa de envio do produto.

21. Qual o preço de envio de encomenda típico de cada expedidor (**expedidores mais caros**)?

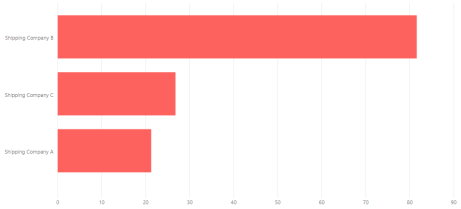


Figura 67: Média do preço de envio por expedidor

- O expedidor que provoca maior despesa em transportar produtos é a *Shipping Company B*, com uma média 81,60 unidades monetárias de envio do produto. - O expedidor que provoca menor despesa em transportar produtos é a *Shipping Company A*, com uma média 21,25 unidades monetárias de envio do produto.

22. Qual o preço de envio de encomenda de acordo com o fornecedor do produto (fornecedores que poderão de ter que diminuir/aumentar o preço de venda de acordo com o preço de envio, para maior lucro)?

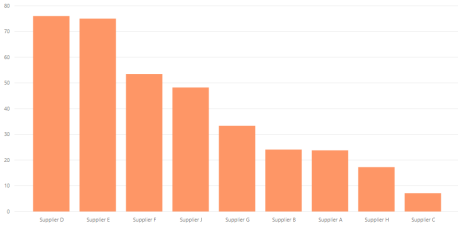


Figura 68: Média do preço de envio por fornecedor

- O fornecedor que disponibiliza com os produtos que têm o preço de envio mais caro é a *Supplier D*, com uma média 76 unidades monetárias de envio do produto. - O fornecedor que disponibiliza com os produtos que têm o preço de envio mais barato é a *Supplier C*, com uma média 7,11 unidades monetárias de envio do produto.

Demonstram-se, de seguida, os indicadores referentes ao "desconto".

23. Qual o desconto de encomenda por categoria de produto (**produtos mais sujeitos a desconto**)?

    - Não foi possível responder a esta resposta, porque não existiam descontos nas fontes de dados (todos os valores a zero). Ainda assim, esta medida é importante num contexto futuro. Isto também é um indicador de que a empresa precisa de implementar descontos, uma vez que pode ser benéfico para o lucro da mesma.

24. Qual o desconto de encomenda típico por funcionário (funcionários que mais vendem produtos promocionais)?

    - Não foi possível responder a esta resposta, porque não existiam descontos nas fontes de dados (todos os valores a zero). Ainda assim, esta medida é importante num contexto futuro. Isto também é um indicador de que a empresa precisa de implementar descontos, uma vez que pode ser benéfico para o lucro da mesma.

Por fim, demonstram-se alguns indicadores mais gerais, não dependentes de nenhuma medida, mas que o grupo considerou relevantes para análise.

25. Qual é o produto mais encomendado (**favoritismo de clientes em relação a produtos**)?

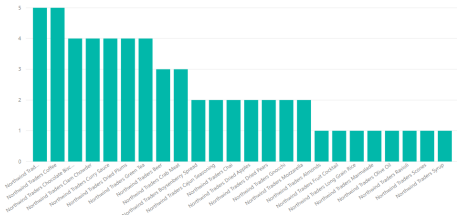


Figura 69: Produtos mais encomendados

- Os produtos mais encomendados são o *Northwind Traders Chocoate* e o *Northwind Traders Coffe*, com 5 ocorrências cada um.

26. Qual é o número de encomendas por cada empresa de fornecedores?

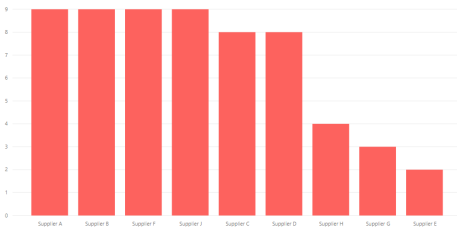


Figura 70: Número de encomendas por fornecedor

- Qual a empresa de fornecedores que mais vende (favoritismo de clientes em relação a fornecedores)?
  - As empresas de fornecedores que mais vendem são a *Supplier A*, a *Supplier B*, a *Supplier F* e a *Supplier J*, com 9 ocorrências cada uma.

27. Qual é o número de encomendas por empresa de expedidor?

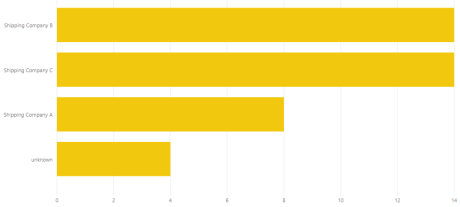


Figura 71: Número de encomendas por expedidor

- Qual o expedidor que mais transporta (favoritismo da empresa em relação a fornecedores)?
  - Os expedidores que mais transportam são a *Shipping Company B* e a *Shipping Company C*, com 14 ocorrências cada uma.

28. Qual é o número de encomendas por funcionário?

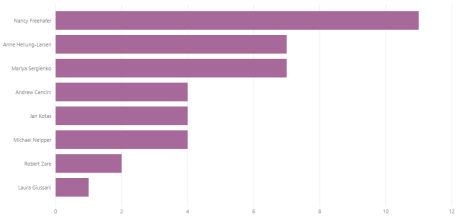


Figura 72: Número de encomendas por funcionários

- Qual o funcionário que mais vende (**prémios de desempenho**)?
  - O funcionário que mais vende é a *Nancy Freehafer* com um total de 11 vendas. Esta funcionária merece ser valorizada, em relação aos restantes, com prémios de desempenho.

6 Conclusões e Sugestões

A resolução do trabalho prático foi bastante importante e enriquecedora, pois permitiu aos membros do grupo perceber e interiorizar melhor os conceitos abordados nas aulas da Unidade Curricular de Análise de Dados.

Deste modo, foram aprofundados conhecimentos relativos ao processo de ETL, e à necessidade de uma boa modelação do problema e implementação. O grupo teve alguma dificuldade inicial, pois foi preciso escolher uma vertente de análise de entre várias possíveis na vasta BD Northwind. O desafio estava em escolher uma vertente e perguntas a serem respondidas que levassem a um modelo dimensional simples e de fácil leitura. Após algumas reuniões, o grupo ultrapassou este problema facilmente.

No que diz respeito à implementação concreta do processo de ETL, a dificuldade foi perceber quais as estruturas necessárias que tivessem arcabouço para lidar com a transformação dos dados, e como as dividir. Para tal, a área de retenção revelou-se de bastante utilidade, pois permitiu armazenar os dados enquanto eram "cozinhados"para o DW final.

Depois do processo de povoamento inicial, surgiu outro contratempo, já que era necessário escolher uma vertente de refrescamento: incremental e/ou diferencial. O grupo acabou por escolher uma mistura entre as duas, o que lhe pareceu a abordagem mais simples e eficiente para o tempo disponível. Depois disto, foi apenas mandatório que se realizassem alguns testes, de modo a concluir que o refrescamento estava, de facto, a ser feito corretamente.

Quanto à utilização do PowerBI como forma de análise dos dados do DW, o grupo considera que adquiriu competências numa ferramenta poderosa, intuitiva e relevante para o futuro dos mesmos.

Em suma, é feita uma apreciação positiva relativamente ao trabalho realizado, visto que a implementação de todas as funcionalidades propostas foram conseguidas com sucesso. O grupo conseguiu tirar partido dos conhecimentos adquiridos neste projeto, sentindo-se capaz de, num contexto futuro, aplicar os conceitos subjacentes de forma eficaz. É evidente que, num outro contexto (como um projeto de grandes dimensões), seria benéfico que fossem implementadas um maior conjunto de funcionalidades adicionais para uma melhor gestão do sistema.