

Análise dos voos nos EUA

Relatório final



Inteligência no Negócio
2016/2017

André Pinho, <apinho@student.dei.uc.pt>
Rafael Pinho, <rfpinho@student.dei.uc.pt>

Tabela de conteúdos

1. Identificação do projeto	2
1.1. Contexto	2
1.2. Fontes de dados	2
1.3. Objetivos da solução de BI	2
2. Protótipos	3
3. Modelo de dados para o data warehouse	4
4. Seleção de software	5
4.1. Bases de dados	5
4.2. ETL	6
4.3. OLAP	6
5. ETL	7
5.1. Plano ETL	7
5.2. Área temporária (staging area)	7
5.3. Ações importantes no plano ETL	7
5.4. Atualização automática do data warehouse	9
5.5. Métricas	9
6. OLAP	10
6.1. Na perspetiva das companhias aéreas	10
6.2. Na perspetiva da gestão dos transportes aéreos pelo DOT	14
6.3. Otimização do data warehouse	15
7. Data mining	16
7.1. Objetivos	16
7.2. Fontes de dados	16
7.3. Atributos selecionados para cada análise	16
7.4. Preparação dos dados	17
7.5. Previsão do número de voos e de passageiros	17
7.6. Classificação do desempenho operacional	18
8. Conclusão	20

Glossário

DOT - departamento de transportes dos EUA

IATA - código que identifica a companhia aérea/aeroporto

Taxi out - intervalo de tempo em minutos desde a hora de partida (da porta de embarque) até que as rodas deixam o solo

Taxi in - intervalo de tempo em minutos desde que as rodas tocam no solo até à hora de chegada (à porta de desembarque)

Hub - é um espaço alugado num dado aeroporto, que as companhias aéreas usam como a base das suas operações

1. Identificação do projeto

1.1. Contexto

O nosso projeto teve como principal objetivo a análise dos voos domésticos nos EUA e tem dois destinatários, o Departamento de transportes dos EUA (DOT) e as companhias aéreas.

O DOT, que tem a responsabilidade de coordenar, garantir e supervisionar os transportes aéreos dos EUA. Desta responsabilidade resulta a necessidade e a obrigação de garantir um serviço eficiente, rápido, seguro, acessível, garantindo a segurança dos EUA e a satisfação dos utilizadores.

As companhias aéreas deste mercado, que estão em permanente avaliação por parte do DOT e necessariamente dos utilizadores. Têm de estar, num processo de melhoria contínua, otimizando os seus rácios inerentes à operação. Por outro lado, interessam-lhes comparar o desempenho do seu serviço com as suas concorrentes, e com isso conseguir analisar os seus pontos fortes e fracos.

1.2. Fontes de dados

Os nossos dados provêm de duas fontes distintas, de um desafio publicado no kaggle [1] e do site do governo dos EUA [2].

O Kaggle forneceu informação dos voos domésticos em 2015, das companhias aéreas e dos aeroportos. São cerca de 580 Mb de dados divididos em 3 ficheiros .csv: Airlines.csv com 1 Kb (14 registos), Airports.csv com 24 Kb (322 registos) e Flights.csv com 579 Mb (5.819.079 registos).

O site do governo dos EUA forneceu informação dos lucros das companhias aéreas, do número de passageiros transportados pelas companhias aéreas, um ficheiro para fazer o mapeamento dos aeroportos que são identificados por um ID em vez do seu código IATA e informação dos voos domésticos referentes aos anos de 2014, 2016 e 2017 (apenas os primeiros 3 meses). Os dados para cada um destes anos estão divididos em 12 ficheiros .csv, cada um com cerca de 16 Mb (aproximadamente 500.000 registos por mês).

1.3. Objetivos da solução de BI

O objetivo da nossa solução de BI é fornecer dados estatísticos, ratings operacionais e previsões que podem funcionar como ferramentas importantes na gestão global dos transportes pelo DOT. Fornece também dados importantes quer individuais, quer comparativos que podem ajudar a melhorar alguns rácios operacionais e de gestão das companhias.

Irá permitir ainda a realização de operações de "benchmarking" a dois níveis:

1. Pelo DOT entre aeroportos para identificar, por um lado, as melhores práticas em alguns, por outro lado, as ineficiências de outros.
2. Às companhias aéreas para comparar com as melhores práticas dos concorrentes a fim de reduzir as suas ineficiências operacionais.

De seguida apresentamos um conjunto de questões às quais tentámos dar resposta na perspetiva das companhias aéreas e dos aeroportos.

1. Companhias aéreas

- Quais as companhias aéreas com maior/menor número de voos realizados?
Com mais/menos voos cancelados?

Com mais/menos voos desviados?

Com mais/menos voos pontuais e atrasados?

Com maior/menor velocidade de voo?

- Quais os tipos de atraso mais frequentes?
- Qual a relação entre os lucros líquidos das companhias aéreas.
- O investimento em hubs compensa? Por exemplo as companhias aéreas que possuem um hub num dado aeroporto são privilegiadas em relação às outras em termos de taxi in e taxi out?
- Qual a companhia aérea que devemos voar para evitar atrasos significativos?
- Qual o número de voos e passageiros previstos para os próximos três anos para cada companhia aérea em cada ano?
- Quais as companhias com melhor desempenho operacional até ao momento em 2017?

2. Aeroportos

- Quais os aeroportos com mais/menos movimento?

Com mais/menos voos cancelados?

Com mais/menos voos desviados?

Com mais/menos tempo despendido no taxi out e no taxi in?

- Qual é a altura do ano com mais/menos voos cancelados e desviados?
- Qual o número de partidas e passageiros previstos para os próximos três anos para cada aeroporto em cada ano?

2. Protótipos

Nas figuras abaixo apresentamos os mockups para as interfaces criadas nos estágios iniciais do projeto.

Interface 1

Companhias aéreas | Aeroportos

Total de companhias aéreas: 14

Selecione o período temporal

Início: // / / Fim: // / / Ok

Informação de voos de cada companhia aérea

Nome companhia	Número de voos	% Voos cancelados	% Voos desviados	Tempo médio de atraso	Velocidade média de voo
Delta Air Lines	50000	0.6%	0.5%	2 min	480 mph
United Air Lines	45000	0.3%	0.2%	1 min	500 mph
American Airlines	42000	0.9%	0.8%	3 min	460 mph
Alaska Airlines	40000	15%	12%	5 min	420 mph
JetBlue Airways	36000	12%	10%	4 min	440 mph

Informação de voos da companhia aérea

Selecione período

☐ Mês

☒ Trimestre

Selecione opção

☒ Número de voos

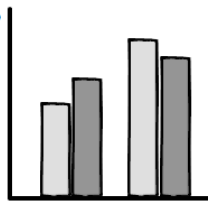
☐ Número de voos cancelados

☐ Número de voos desviados

☐ Tempo médio de atraso

Mostrar

Opção



Período

Esta interface permite às companhias aéreas comparar os rácios operacionais com os seus concorrentes e analisar os seus próprios rácios operacionais ao longo do tempo.

Interface 2

Companhias aéreas
Aeroportos

Total de aeroportos: 322

Selecione o período temporal

Início

/ /

Fim

/ /

Ok

Informação de voos de cada aeroporto

	Nome aeroporto	Movimento	Voos cancelados	Voos desviados	Taxi out médio	Taxi in médio
1	John F. Kennedy International Airport	250000	200	150	22	16
2	Newark Liberty International Airport	200000	250	100	24	18
3	San Francisco International Airport	180000	150	200	20	14
4	Denver International Airport	175000	50	300	16	10
5	Los Angeles International Airport	150000	100	250	18	12

Informação de voos do aeroporto

Selecione período

Mês
Trimestre

Selecione opção

Movimento (partidas + chegadas)
Número de voos cancelados
Número de voos desviados

Mostrar

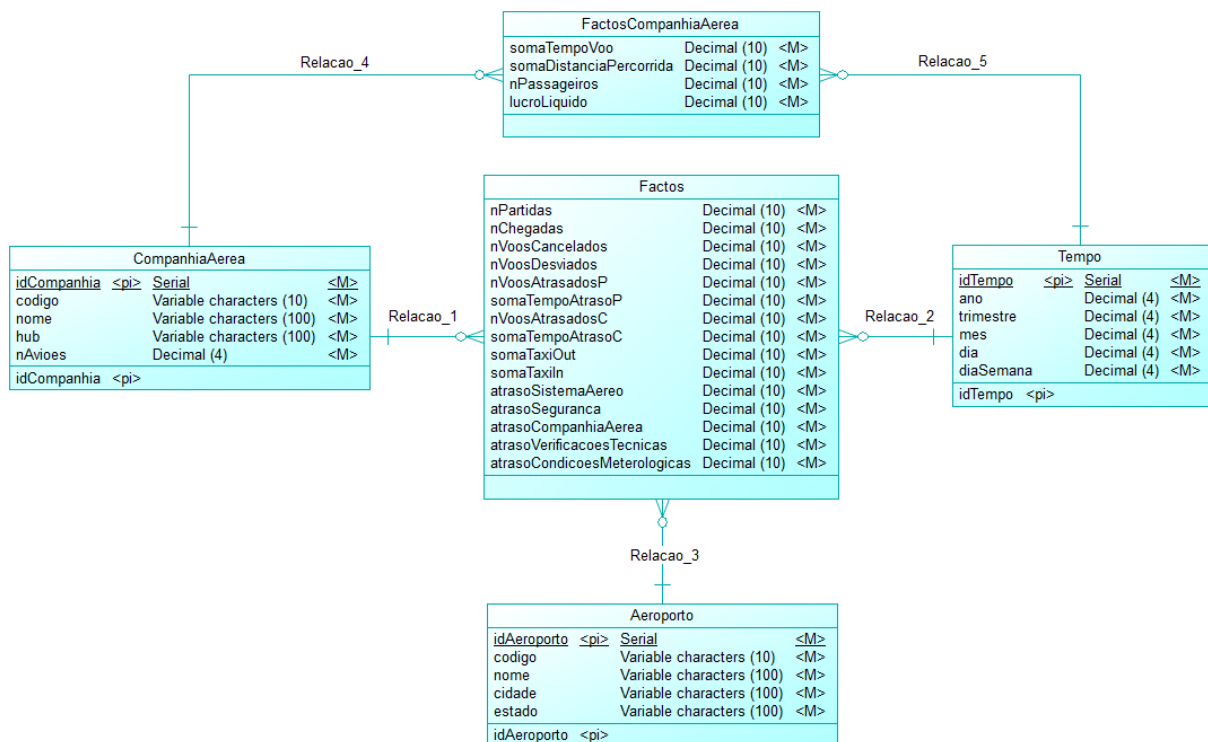
Opção

Período

Esta interface permite ao DOT comparar os rácios operacionais entre aeroportos e analisar os rácios operacionais de um aeroporto ao longo do tempo.

3. Modelo de dados para o data warehouse

Na figura abaixo apresentamos o diagrama ER do modelo de dados multidimensional para o data warehouse, que reúne todos os dados necessários para responder às perguntas identificadas anteriormente.



O nosso modelo de dados é constituído por duas estrelas (Factos e FactosCompanhia Aerea).

A tabela de Factos da primeira estrela está relacionada com as dimensões Tempo, CompanhiaAerea e Aeroporto, e é constituída pelos seguintes atributos: nº de partidas, nº de chegadas, nº de voos cancelados, nº de voos desviados, nº de voos atrasados à partida, soma do tempo de atraso à partida, nº de voos atrasados à chegada, soma do tempo de atraso à chegada, soma do taxi out, soma do taxi in, soma do atraso relativo ao sistema aéreo, soma do atraso relativo à segurança, soma do atraso relativo à companhia aérea, soma do atraso relativo às verificações técnicas e soma do atraso relativo às condições meteorológicas.

A tabela de FactosCompanhiaAerea da segunda estrela está relacionada com as dimensões Tempo e CompanhiaAerea, e é constituída pelos seguintes atributos: soma do tempo de voo, soma da distância percorrida, nº de passageiros e lucro líquido.

Na dimensão Tempo escolhemos uma granularidade fina, ao nível do dia porque se adequa às necessidades da análise.

4. Seleção de software

Para selecionar as ferramentas que iremos usar para o armazenamento dos dados, para realizar o processo ETL e para fazer o OLAP começámos por definir os seus requisitos obrigatórios:

- Para o armazenamento e acesso a dados pretendemos uma base de dados com bom desempenho para processamento analítico e que permita a integração de ferramentas externas para carregamento e análise de dados.
- Para o ETL pretendemos um software de modelação visual, que automatize o processo ETL e que integre com ficheiros .csv.
- Para o OLAP pretendemos algo que permita fazer slice e dice, drill-down e roll-up, e de integração fácil e eficiente com bases de dados.

A abordagem para selecionar as ferramentas passou por comparar algumas ferramentas diferentes com as ferramentas exploradas durante as aulas, e analisar as suas vantagens e desvantagens.

4.1. Bases de dados

Para a seleção da base de dados foram usados os seguintes critérios:

- Relacional para suportar um modelo multidimensional.
- Uso gratuito.
- Desempenho para o processamento analítico e queries complexas.
- Conectividade para a integração com ferramentas externas de carregamento e análise de dados (data mining, OLAP, reporting).
- Interface da base de dados simples e intuitiva.

Comparação das bases de dados usando os critérios definidos:

	PostgreSQL	Oracle	MySQL
Relacional	Sim	Sim	Sim
Gratuita	Sim	+/- (até 10 Gb)	Sim
Desempenho	Sim	Sim	Não
Conectividade	Sim	Sim	Sim
Interface	Sim	Sim	Sim

4.2. ETL

Para a seleção da ferramenta ETL foram usados os seguintes critérios:

- Gratuito até ao final do projeto.
- Boa documentação.
- Modelação visual com a definição de fluxo do processo.
- Automatização do processo ETL.
- Integração de dados a partir de ficheiros .csv.

Comparação das ferramentas ETL usando os critérios definidos:

	Pentaho (kettle)	Talend Open Studio	CloverETL
Gratuita	Sim	Sim	Sim
Boa documentação	+/-	Não	+/-
Modelação visual	Sim	Sim	Sim
Automatização do ETL	Sim	Sim	Sim
Integração de dados a partir de ficheiros .csv	Sim	Sim	Sim

4.3. OLAP

Para a seleção da ferramenta OLAP foram usados os seguintes critérios:

- Gratuito até ao final do projeto.
- Boa documentação.
- Criar e modular cubos OLAP.
- Integração eficiente com bases de dados para produzir boas análises com queries simples.
- Integração com Java.

Comparação das ferramentas OLAP usando os critérios definidos:

	Queries ad hoc	Mondrian	Tableau
Gratuito	Sim	Sim	Sim
Boa documentação	Sim	+/-	Sim
Criar e modular cubos OLAP	Não	Sim	Sim
Integração com bases de dados	Sim	Sim	Sim
Integração com Java	Sim	Sim	Não

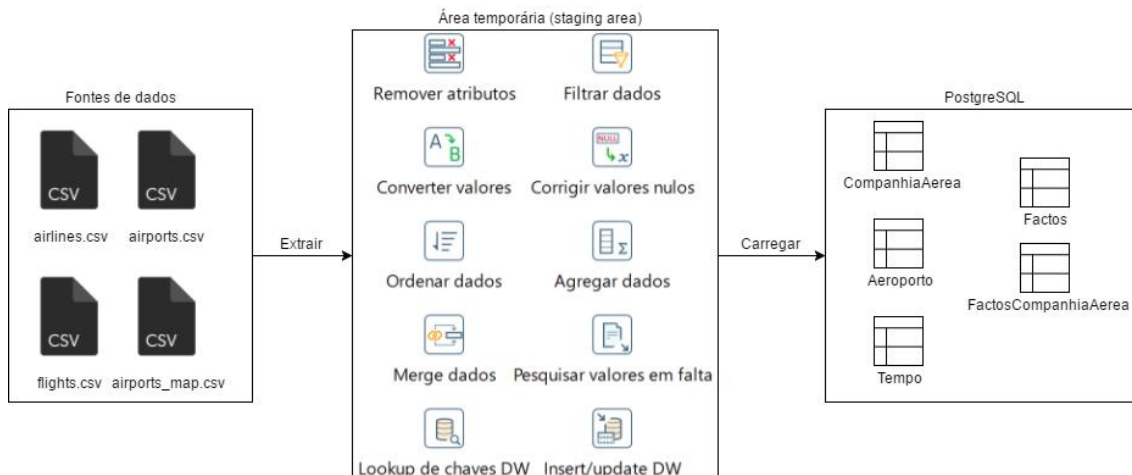
No nosso trabalho utilizámos o PostgreSQL como motor de base de dados, o Pentaho data integration (kettle) para o ETL e queries ad hoc para o OLAP.

5. ETL

Nesta seção, iremos concentrar-nos no planeamento do processo ETL, na execução do ETL e na criação das ferramentas para as atividades OLAP do utilizador.

5.1. Plano ETL

Na figura abaixo apresentamos um esquema sumário do plano ETL realizado.



- Na extração dos dados fizemos o download dos ficheiros .csv para um diretório dentro da área temporária.
- Na transformação de dados removemos atributos para selecionar os pertinentes, filtrámos dados para preparar as agregações, convertemos valores e corrigimos valores nulos, fizemos ordenamentos, agregações e funções de grupo (somatórios, médias e contagens), fizemos o merge dos fluxos provenientes de diferentes ficheiros, procurámos as chaves estrangeiras das dimensões e inserimos/atualizámos os registos.
- No carregamento de dados começámos por remover os índices, depois carregámos as tabelas das dimensões, depois carregámos os dados nas tabelas de factos e finalmente inserimos os índices novamente.

5.2. Área temporária (staging area)

- Temos todos os ficheiros .csv dentro de um diretório.
- Corremos um script previamente criado para a fazer a correção dos aeroportos nos voos realizados em Outubro (eram identificados por um id em vez do seu código IATA).
- Usámos o Kettle para transformar e carregar os dados.
- O ETL é executado de forma automática.
- Os dados dos voos serão atualizados uma vez por ano, através de um cron scheduler criado para o efeito.

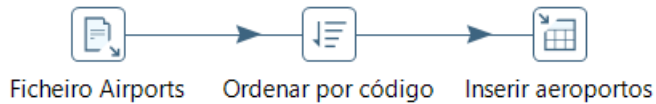
5.3. Ações importantes no plano ETL

Atividade #1 - Carregar a dimensão CompanhiaAerea



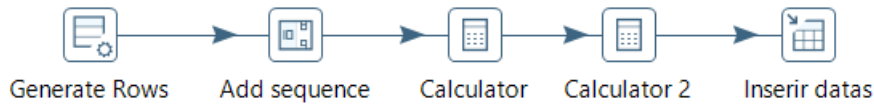
Para carregar a dimensão companhia aérea lemos o ficheiro airlines.csv, ordenámos os registos pelo código IATA e inserimos no data warehouse.

Atividade #2 - Carregar a dimensão Aeroporto



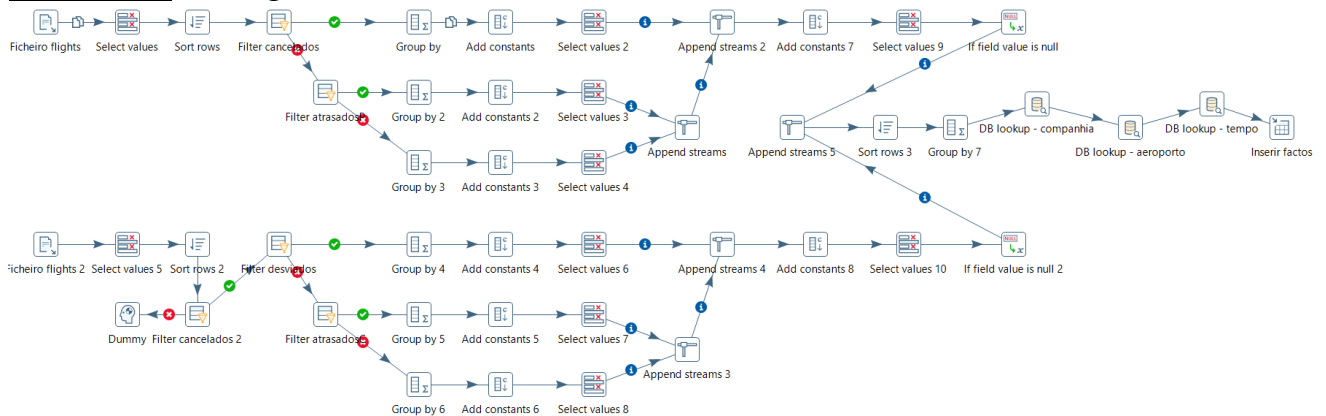
Para carregar a dimensão aeroporto lemos o ficheiro airports.csv, ordenámos os registos pelo código IATA e inserimos no data warehouse.

Atividade #3 - Carregar a dimensão Tempo



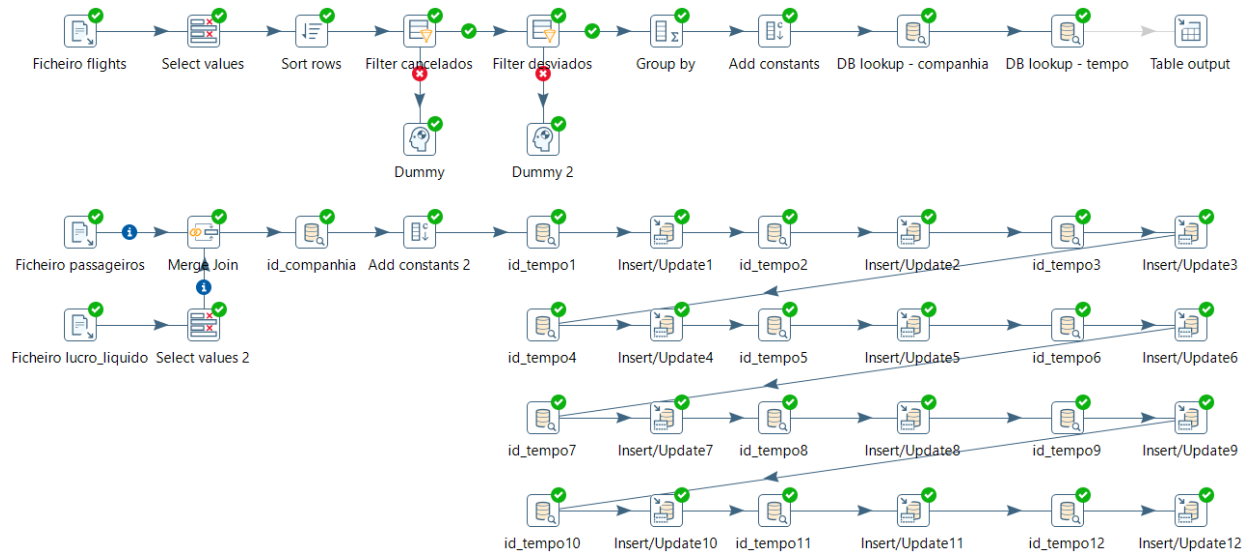
Para carregar a dimensão tempo utilizámos um gerador de linhas com o formato date, e com uma sequência incrementámos a data dia a dia e inserimos no data warehouse.

Atividade #4 - Carregar a tabela de Factos



Para carregar a tabela de Factos utilizámos o ficheiro flights.csv e dividimos o fluxo em duas partes para fazer as agregações corretamente. Na parte de cima agregámos por tempo (ano, mês, dia), companhia aérea e aeroporto de origem porque o número de partidas, o taxi out, o número de voos cancelados e os atrasos á partida têm de ficar associados ao aeroporto de origem. Na parte de baixo agregámos por tempo (ano, mês, dia), companhia aérea e aeroporto de destino porque o número de chegadas, o taxi in, o número de voos desviados e os atrasos á chegada têm de ficar associados ao aeroporto de destino. Depois fizemos o merge dos vários sub fluxos, ordenámos, agregámos, fizemos o lookup das chaves estrangeiras e inserimos no data warehouse.

Atividade #5 - Carregar a tabela de FactosCompanhiaAerea

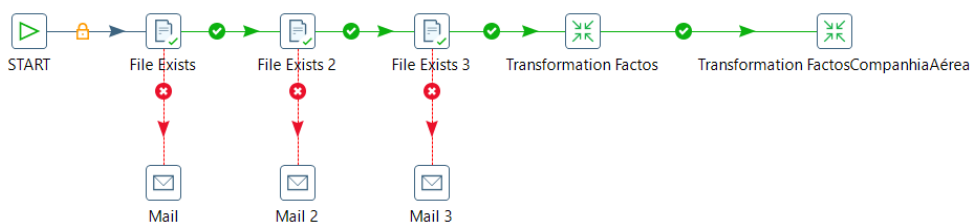


Para carregar a tabela FactosCompanhiaAerea utilizámos o ficheiro flights.csv, ordenámos os voos por tempo (ano, mês, dia) e aeroporto, descartámos os voos cancelados e desviados para fazer as agregações do tempo de voo e da distância percorrida. De seguida fizemos o lookup das chaves estrangeiras e inserimos no data warehouse.

O fluxo de baixo insere o lucro líquido e o número de passageiros mensal na tabela FactosCompanhiaAerea.

5.4. Atualização automática do data warehouse

Para a atualização automática do data warehouse criámos um cron scheduler (kettle job) que dispara no início de cada ano para fazer a atualização das tabelas de factos. Para isso também é necessário que os ficheiros sejam colocados numa diretoria definida antes do dia e da hora de disparo.



O cron scheduler quando é disparado começa por verificar se os três ficheiros estão na diretoria especificada. Se os ficheiros estiverem na diretoria especificada o ETL da tabela de Factos e da tabela FactosCompanhiaAerea é executado, caso contrário um email é enviado ao utilizador.

5.5. Métricas

Tamanho dos dados de origem (ano de 2015): 580 Mb divididos em 3 ficheiros .csv

Tamanho dos dados após o primeiro carregamento: 74 Mb

- Dimensão CompanhiaAerea com 88 Kb (14 registos)
- Dimensão Aeroporto com 40 Kb (322 registos)
- Dimensão Tempo com 80 Kb (365 registos)

- Factos com 72 Mb (389.023 registos)
- FactosCompanhiaAerea: 600 Kb (4.932 registos)

Tempo decorrido na primeira carga: 1 hora e 2 minutos

- Dimensão CompanhiaAerea: 1s
- Dimensão Aeroporto: 2s
- Dimensão Tempo: 1s
- Factos: 3335s
- FactosCompanhiaAerea: 307s

Nas cargas subsequentes o tamanho dos dados e o tempo utilizado para cada atualização será semelhante ao tempo decorrido na primeira carga, uma vez que o ficheiro dos registos de voos tem aproximadamente o mesmo número de registos.

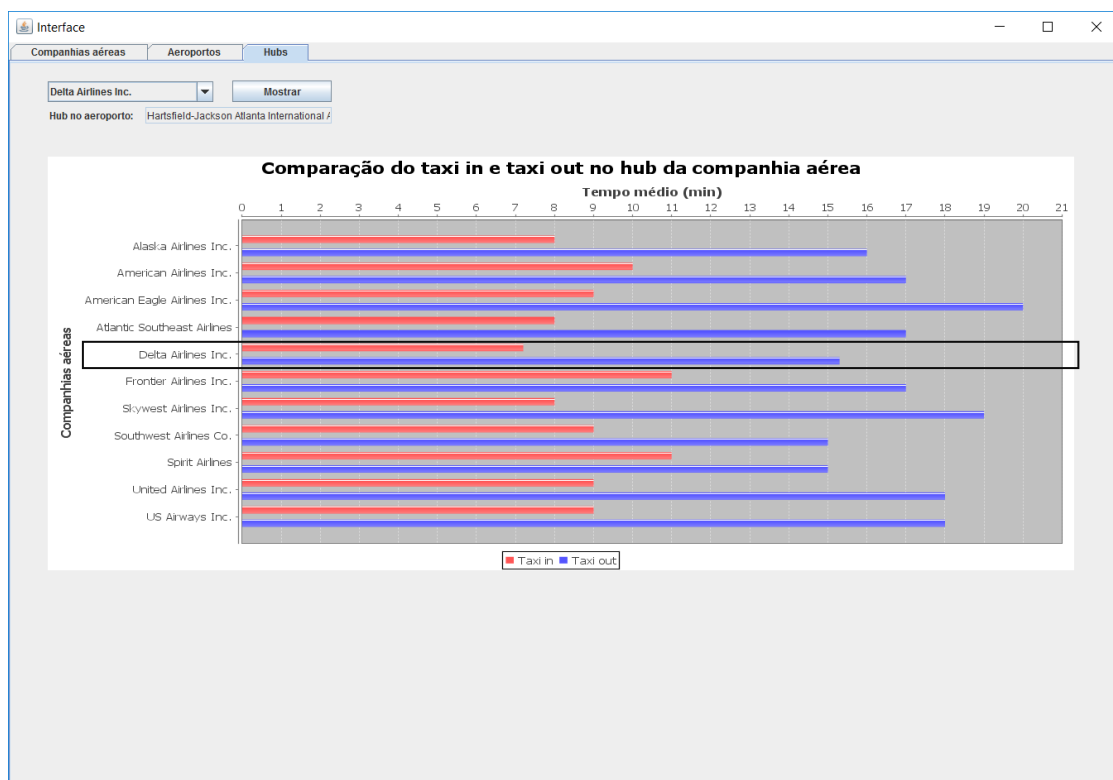
6. OLAP

Para as atividades de OLAP desenvolvemos uma dashboard na swing do Java, que contém tabelas e gráficos comparativos com várias opções de drill-down e roll-up. Os resultados a seguir apresentados têm como objetivo dar resposta aos objetivos que definimos na primeira meta do projeto. De uma forma geral, os resultados abrangem os parâmetros fundamentais da análise das companhias aéreas e dos aeroportos, para permitir aos decisores fazerem múltiplas análises de acordo com os seus objetivos de gestão. Das análises possíveis destacamos as seguintes:

6.1. Na perspetiva das companhias aéreas

Análise #1 - Hubs das companhias aéreas

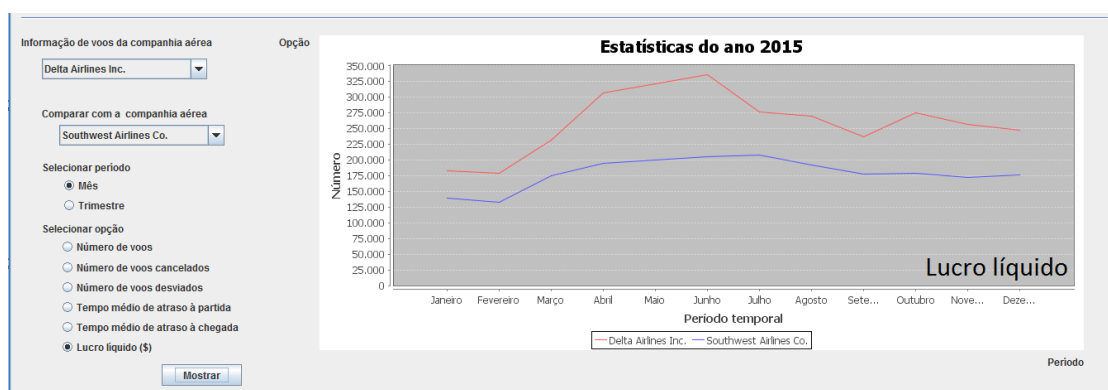
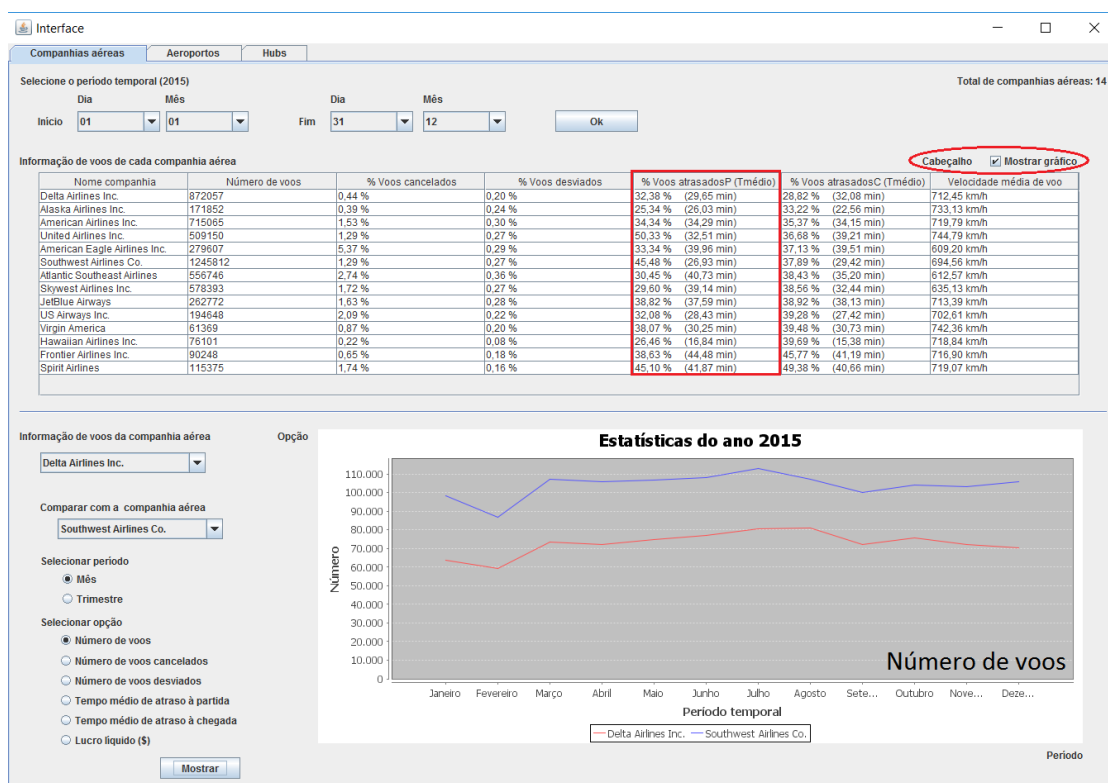
A interface abaixo permite às companhias aéreas analisar se o investimento em hubs compensa. Por exemplo se as companhias aéreas que possuem um hub num dado aeroporto são privilegiadas em relação às outras em termos de taxi in e taxi out.



Pela análise dos resultados ficou demonstrado que o investimento (aluguer de espaço) em hubs trás uma ligeira vantagem em termos de taxi in e taxi out.

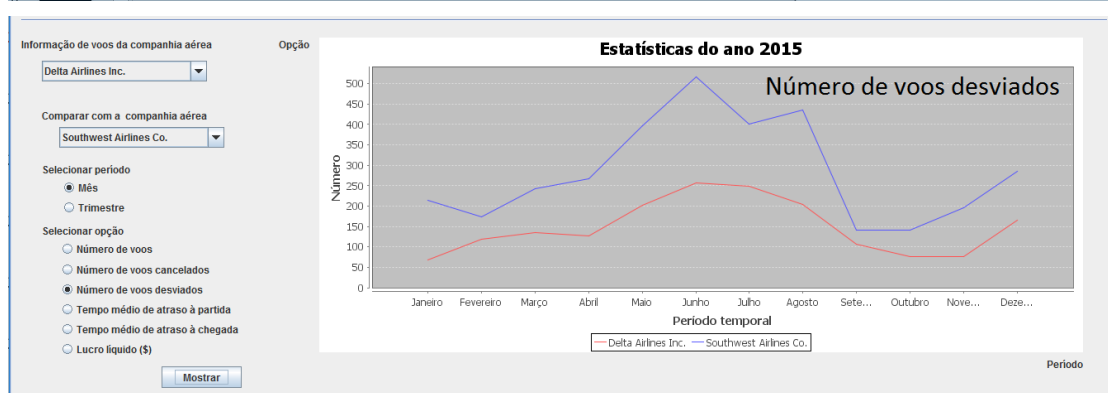
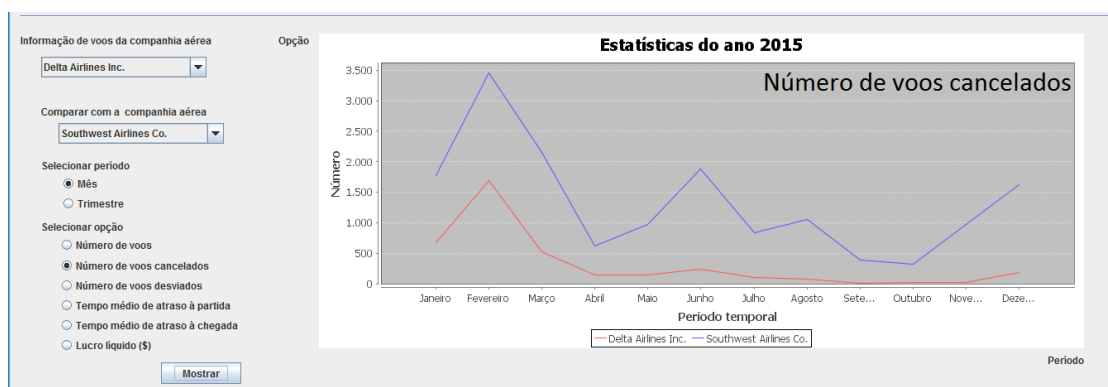
Análise #02 - Benchmarking entre as duas maiores companhias aéreas

A interface abaixo permite às companhias aéreas comparem o desempenho do seu serviço com as suas concorrentes, e com isso conseguir analisar os seus pontos fortes e fracos. Permite fazer vários tipos de análise ao nível dos rácios operacionais num intervalo de tempo definido, tanto de forma tabular como comparação direta de forma gráfica. Um clique no cabeçalho de uma coluna da tabela permite ordenar essa coluna (quadrado vermelho), e caso o pisto do cabeçalho esteja selecionado (círculo vermelho) permite obter uma visualização gráfica mais detalhada do rácio que veremos mais adiante (análises #3, #4 e #5).



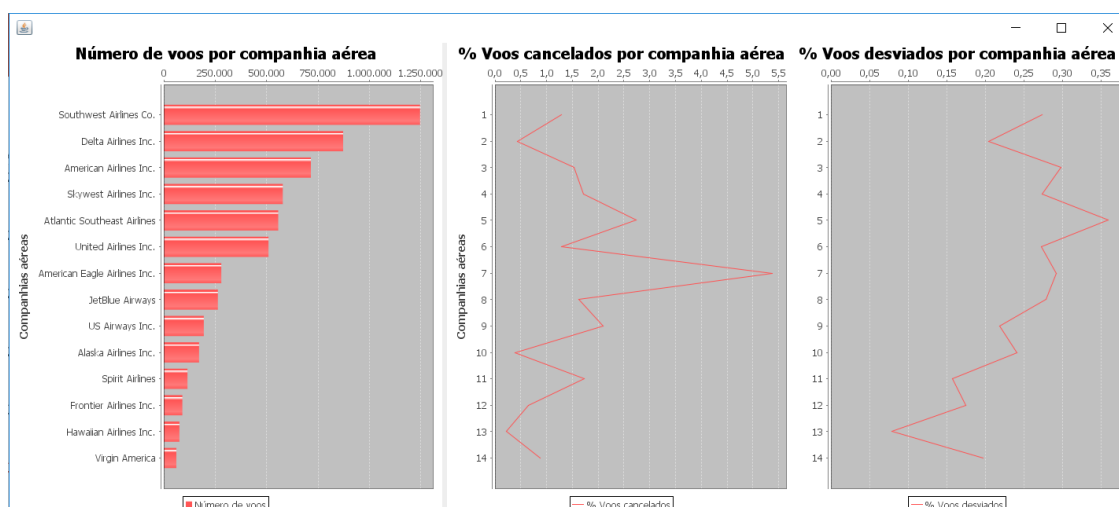
Nesta análise fizemos o benchmarking dos lucros líquidos entre as duas maiores companhias aéreas a 'Delta Airlines Inc.' e a 'Southwest Airlines Co.' e constatámos que a companhia com o maior número de voos, a 'Southwest Airlines Co.' tem um

lucro líquido inferior à 'Delta Airlines Inc.'. Pela análise da tabela acima ou pelos gráficos seguintes é possível verificar que a 'Southwest Airlines Co.' tem um maior número de voos cancelados e desviados, o que ajuda a justificar esta diferença de lucros líquidos.



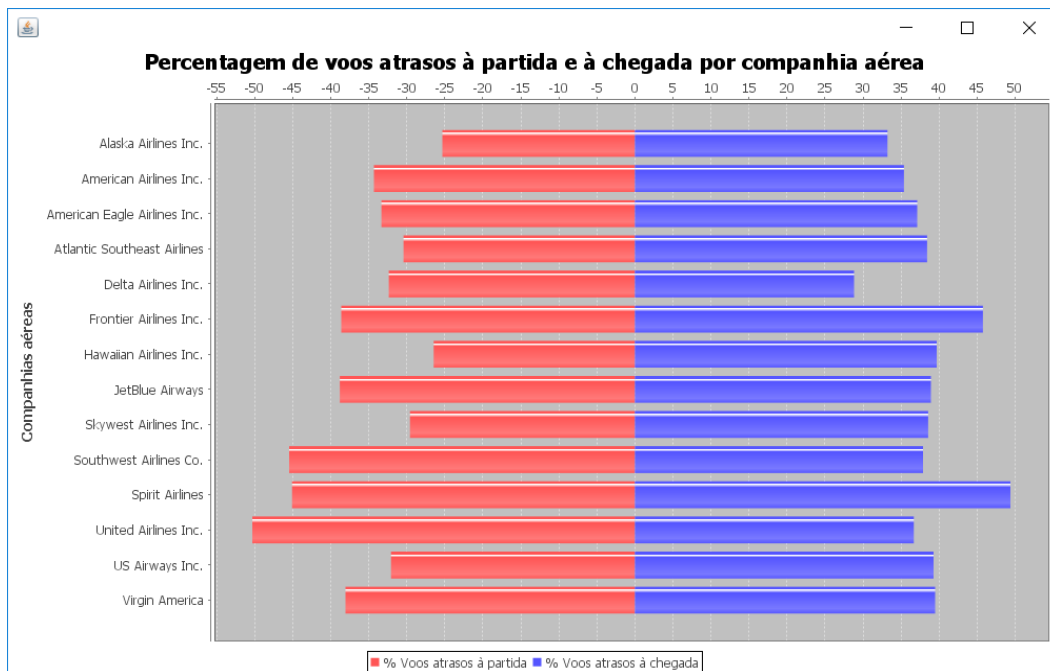
Análise #03 - Comparação da percentagem de voos cancelados e desviados por companhia aérea

A interface abaixo permite às companhias aéreas comparar a percentagem de voos cancelados e desviados com os seus concorrentes num dado período de tempo.



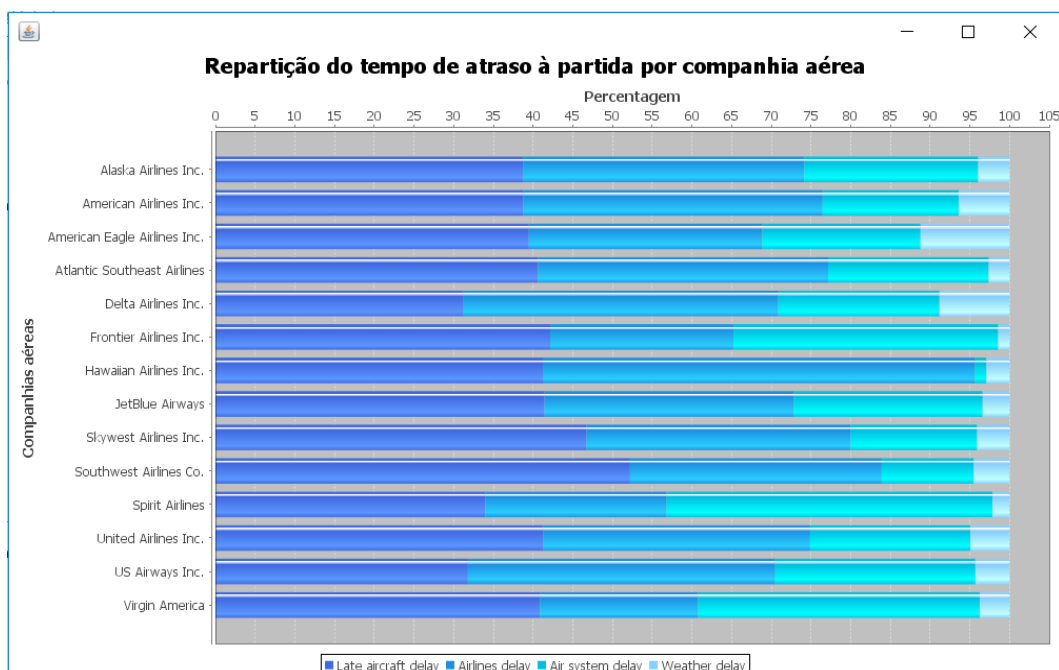
A companhia que apresenta pior desempenho operacional é a 'American Eagle Airlines Inc.', devido à maior percentagem de voos cancelados e também por ter uma das maiores percentagens de voos desviados. No sentido oposto, a 'Delta Airlines Inc.' e a 'Hawaiian Airlines Inc.' são as que apresentam melhor desempenho operacional.

Análise #04 - Comparação dos voos atrasados à partida e à chegada por companhia aérea
A interface abaixo permite às companhias aéreas comparar a percentagem de voos atrasados à partida e à chegada com os seus concorrentes num dado período de tempo. As causas dos atrasos analisaremos na análise #5.



Quando voamos e pretendemos evitar atrasos devemos optar pela ‘Delta Airlines Inc.’, que é a companhia aérea mais pontual á chegada ao destino.

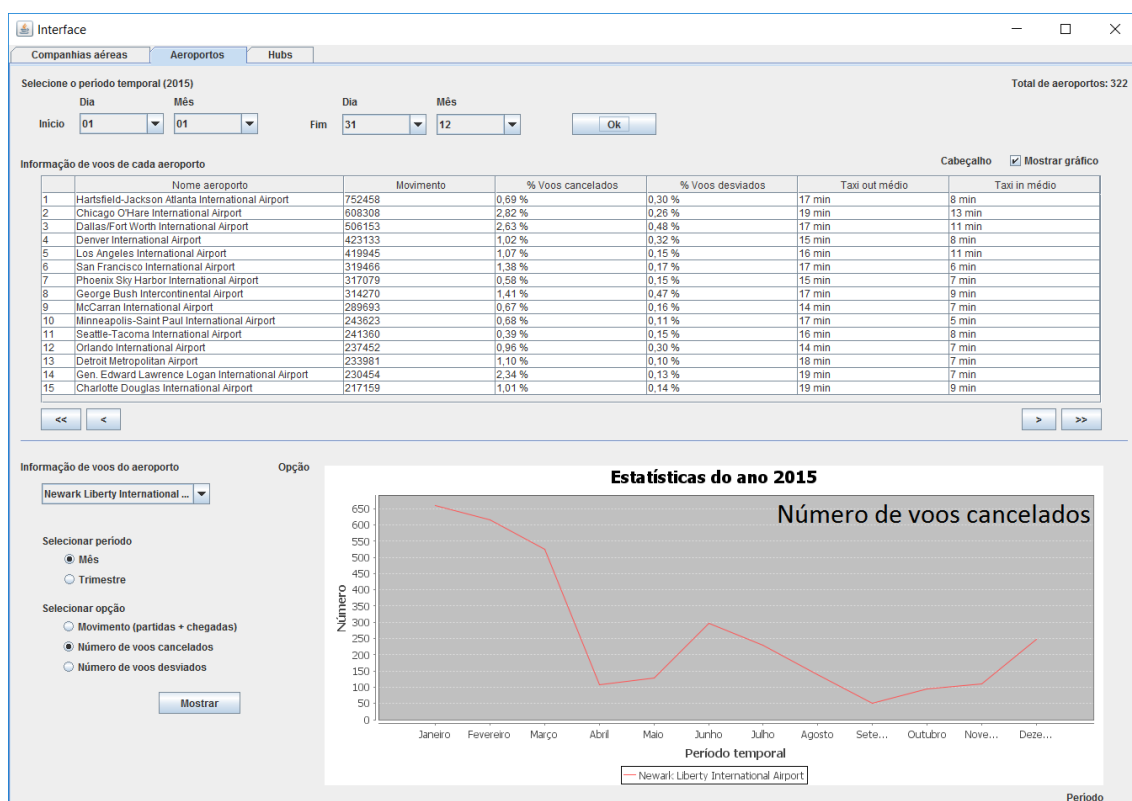
Análise #05 - Comparação dos atrasos detalhados por companhia aérea
A interface abaixo permite às companhias aéreas comparar e visualizar a repartição dos tempos de atraso à partida em detalhe (sistema aéreo, verificações técnicas, condições meteorológicas) num dado período de tempo.



6.2. Na perspectiva da gestão dos transportes aéreos pelo DOT

Análise #01 - Qual é a altura do ano com mais/menos voos cancelados e desviados?

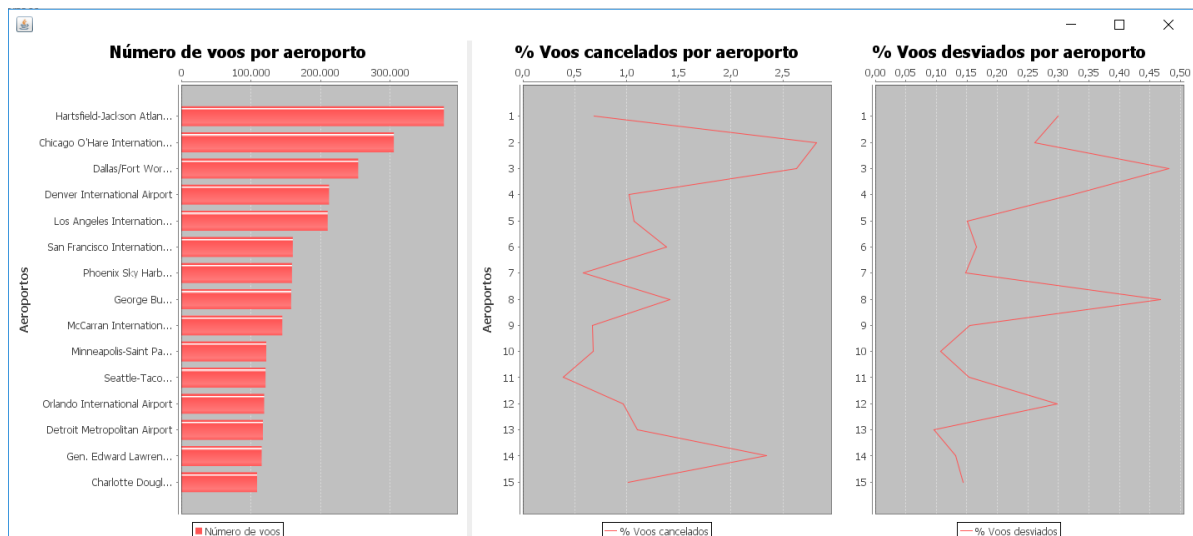
A interface abaixo permite ao DOT comparar o desempenho operacional dos vários aeroportos no país, tanto de forma tabular como comparação direta de forma gráfica num intervalo de tempo definido. Um clique no cabeçalho de uma coluna da tabela permite ordenar essa coluna, e caso o pisto do cabeçalho esteja selecionado permite obter uma visualização gráfica mais detalhada do rácio que veremos mais adiante (análise #2).



A tabela acima permite ao DOT obter um panorama geral da gestão aeroportuária do país, como por exemplo os aeroportos mais movimentados, com mais voos cancelados, desviados, ... Pelas análises gráficas realizadas a alguns aeroportos foi possível constatar que existem mais voos cancelados no inverno e mais voos desviados no verão. O motivo dos voos cancelados no Inverno são as condições meteorológicas. O motivo dos voos desviados no Verão é o aumento significativo no tráfego nos aeroportos e as férias dos funcionários.

Análise #02 - Quais os aeroportos mais movimentados e comparação da percentagem de voos cancelados e desviados por aeroporto

A interface abaixo permite ao DOT comparar a percentagem de voos cancelados e desviados num dado período de tempo nos aeroportos mais movimentados.



O aeroporto mais movimentado nos EUA é o 'Hartsfield-Jackson Atlanta International Airport' com um número médio de partidas e chegadas diárias de 2062 voos em 2015, seguido pelos aeroportos de 'Chicago O'Hare International Airport' e de 'Dallas/Fort Worth International Airport'.

6.3. Otimização do data warehouse

Para otimizar o desempenho do data warehouse utilizámos algumas técnicas tais como índices e vistas materializadas.

Indexámos as chaves estrangeiras da tabela Factos (idCompanhia, idTempo, idAeroporto) e da tabela FactosCompanhiaAerea (idCompanhia, idTempo) para acelerar as junções com as dimensões. Optámos por índices B-Tree.

```
1  /* Criar os índices */
2  CREATE INDEX index_1 ON factos (idCompanhia);
3  CREATE INDEX index_2 ON factos (idAeroporto);
4  CREATE INDEX index_3 ON factos (idTempo);
5  CREATE INDEX index_4 ON factosCompanhiaAerea (idCompanhia);
6  CREATE INDEX index_5 ON factosCompanhiaAerea (idTempo);
```

Criámos também duas vistas materializadas (uma para cada tabela de factos) para armazenar resultados pré calculados e acelerar a resposta às queries. O atributo idTempo de ambas as vistas também foi indexada para acelerar as pesquisas.

```
8  /* Criar vistas materializadas */
9  CREATE MATERIALIZED VIEW tabela1 AS
10 select t.idTempo as idTempo, ca.nome as nome, sum(f.nPartidas) as nPartidas, sum(f.nChegadas) as nChegadas,
11        sum(f.nVoosCancelados) as nVoosCancelados, sum(f.nVoosDesviados) as nVoosDesviados, sum(f.nVoosAtrasadosP) as nVoosAtrasadosP,
12        sum(f.somaTempoAtrasoP) as somaTempoAtrasoP, sum(f.nVoosAtrasadosC) as nVoosAtrasadosC, sum(f.somaTempoAtrasoC) as somaTempoAtrasoC,
13        sum(atrasoSistemaAereo) as atrasoSistemaAereo, sum(atrasoCompanhiaAerea) as atrasoCompanhiaAerea, sum(atrasoVerificacoesTecnicas) as
14        atrasoVerificacoesTecnicas, sum(atrasoCondicoesMeterologicas) as atrasoCondicoesMeterologicas
15 from factos f, companhiaAerea ca, tempo t
16 where f.idCompanhia = ca.idCompanhia and f.idTempo = t.idTempo
17 group by t.idTempo, ca.nome;
18
19 CREATE MATERIALIZED VIEW tabela2 AS
20 select t.idTempo as idTempo, a.nome as nome, sum(f.nPartidas) as nPartidas, sum(f.nChegadas) as nChegadas, sum(f.nVoosCancelados)
21 as nVoosCancelados, sum(f.nVoosDesviados) as nVoosDesviados, sum(f.somaTaxiOut) as somaTaxiOut, sum(f.somaTaxiIn) as somaTaxiIn
22 from factos f, aeroporto a, tempo t
23 where f.idAeroporto = a.idAeroporto and f.idTempo = t.idTempo
24 group by t.idTempo, a.nome;
25
26 /* Indexar as vistas materializadas */
27 CREATE INDEX index_6 ON tabela1 (idTempo);
28 CREATE INDEX index_7 ON tabela2 (idTempo);
```


A tabela1 é uma vista materializada da tabela de Factos, ficou com 712 Kb (4.926 registos). A tabela2 é uma vista materializada da tabela de FactosCompanhiaAerea, ficou com 13 Mb (113.350 registos). O tempo de resposta baixou cerca de 30 vezes (de 3 segundos para cerca de uma décima de segundo).

7. Data mining

Nesta seção, iremos concentrar-nos nas ferramentas de data mining para complementar a análise de dados OLAP com capacidades descritivas e preditivas. Em termos de software utilizámos o Pentaho data integration (kettle) e o Weka. O Kettle para inserir novos dados no data warehouse, para extrair valores do data warehouse e para criar os ficheiros .arff. O Weka para fazer as previsões e para classificar as companhias aéreas.

7.1. Objetivos

Para a análise de data mining tivemos dois objetivos. O primeiro foi prever o número de voos e o número de passageiros para os próximos três anos, tanto para as companhias aéreas como para os aeroportos. O segundo objetivo foi classificar o desempenho operacional das companhias aéreas até ao momento no ano de 2017.

7.2. Fontes de dados

Nesta fase do trabalho utilizámos a informação atual do nosso data warehouse referente aos voos domésticos em 2015, às companhias aéreas e aos aeroportos inseridos na meta anterior, e novos dados disponibilizados no site do governo dos EUA. O site do governo forneceu informação dos voos domésticos referentes aos anos de 2014, 2016 e 2017 (apenas os primeiros 3 meses) usados para a classificação [2]. Os dados para cada um destes anos estão divididos em 12 ficheiros .csv, cada um com cerca de 16 Mb (aproximadamente 500.000 registos por mês). Também forneceu dados históricos relativos ao número de voos e passageiros para cada companhia aérea e aeroporto nos últimos 10 anos usados para a previsão [3].

7.3. Atributos selecionados para cada análise

- | | |
|--|---------------------------------|
| - Previsão companhias aéreas | - Previsão aeroportos |
| @attribute companhiaAerea string | @attribute aeroporto string |
| @attribute ano numeric | @attribute ano numeric |
| @attribute nVoos numeric | @attribute nPartidas numeric |
| @attribute nPassageiros numeric | @attribute nPassageiros numeric |
| - Classificação do desempenho operacional | |
| @attribute percentagemCancelados numeric | |
| @attribute percentagemDesviados numeric | |
| @attribute percentagemVoosAtrasadosP numeric | |
| @attribute percentagemVoosAtrasadosC numeric | |
| @attribute tempoMedioAtrasoP numeric | |
| @attribute tempoMedioAtrasoC numeric | |

7.4. Preparação dos dados

Começámos por executar o processo ETL desenvolvido na meta anterior para inserir e limpar os novos dados referentes aos anos de 2014, 2016 e 2017 (primeiros 3 meses). Após a execução do processo ETL usámos ao fluxo kettle abaixo para gerar os ficheiros .arff com a informação necessária tanto para as previsões como para as classificações.



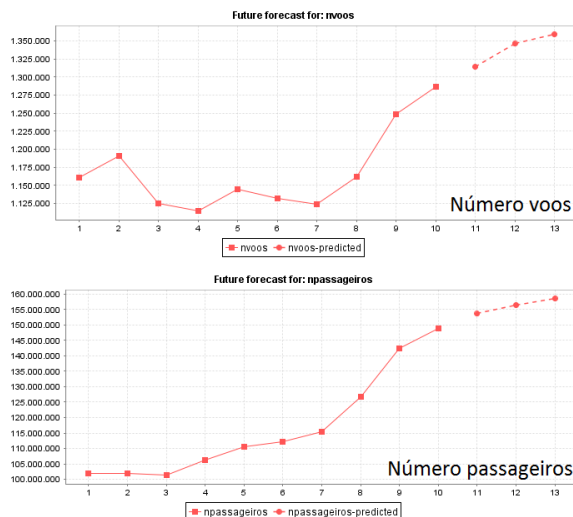
Este fluxo começa por fazer uma query à base de dados com uma query previamente definida e depois escreve o output num ficheiro em formato .arff.

7.5. Previsão do número de voos e de passageiros

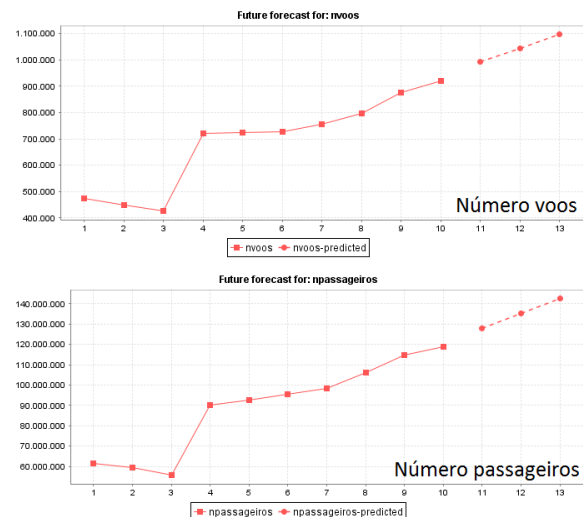
Para a previsão usámos o algoritmo 'MultilayerPerceptron' com uma learningRate de 0.01, porque após várias experiências foi aquele que deu os melhores resultados.

Resultados da previsão para algumas das companhias aéreas

Southwest Airlines Co.



Delta Airlines Inc.



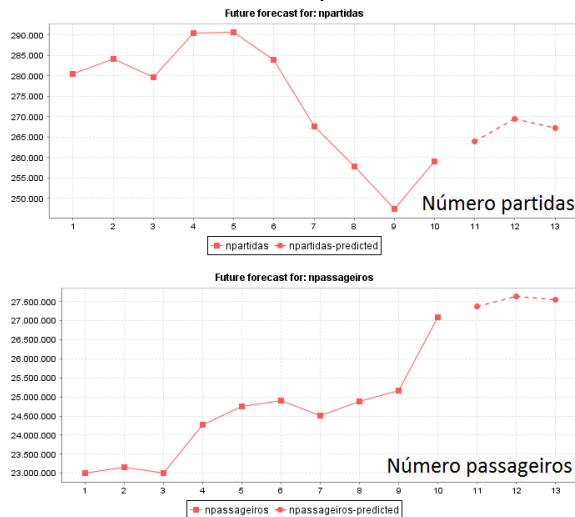
Nota: os pontos de 1 a 10 no eixo do X correspondem aos anos de 2007 a 2016, enquanto que os pontos 11, 12 e 13 referem-se à previsão para os anos de 2017, 2018 e 2019.

A partir dos gráficos de previsão acima podemos ver uma tendência crescente no número de voos e passageiros para as duas maiores companhias aéreas de voos domésticos nos EUA, a 'Southwest Airlines Co.' e a 'Delta Airlines Inc.'.

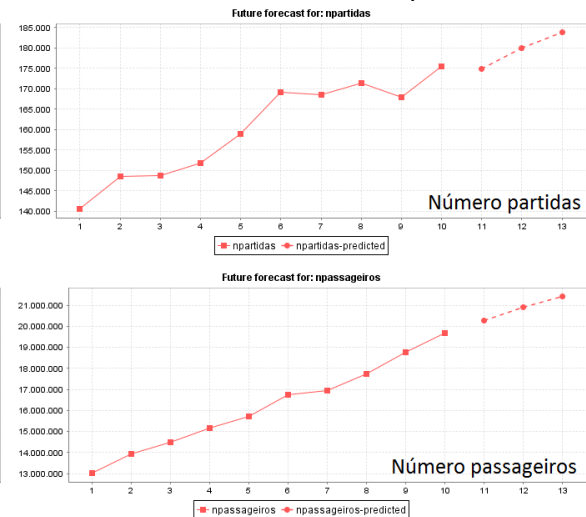
Para um histórico de 10 anos, o nosso modelo de previsão consegue prever o número de voos com uma taxa de erro média inferior a ~15.000 voos em 1.200.000 de voos (~1.25%), e o número de passageiros com uma taxa de erro média inferior a ~2.750.000 passageiros em 150 milhões de passageiros (~1.83%).

Resultados da previsão para alguns dos aeroportos

Denver International Airport



San Francisco International Airport



Nota: os pontos de 1 a 10 no eixo do X correspondem aos anos de 2007 a 2016, enquanto que os pontos 11, 12 e 13 referem-se à previsão para os anos de 2017, 2018 e 2019.

No aeroporto de San Francisco o número de voos tem tendência a aumentar nos próximos 3 anos de forma gradual. No aeroporto de Denver a tendência é que aumente em 2017 e 2018 e decresça em 2019.

No aeroporto de Denver para um histórico de 10 anos, o erro de previsão no 3º ano foi de 3.163 para o número de partidas (1.18 %) e de 309.560 para o número de passageiros (1.12 %).

No aeroporto de San Francisco para um histórico de 10 anos, o erro de previsão no 3º ano foi de 2.327 para o número de partidas (1.27 %) e de 140.967 para o número de passageiros (0.66 %).

7.6. Classificação do desempenho operacional

Para a classificação optámos por usar o algoritmo 'RandomForest', porque após comparar com outros algoritmos (tabela abaixo) foi aquele que deu a melhor percentagem de instâncias corretamente classificadas.

Algoritmo de classificação	Cross-validation (folds = 5)	
	Corretamente classificadas	Incorretamente classificadas
OneR	51.05 %	48.95 %
SMO	64.77 %	35.23 %
BayesNet	73.63 %	26.37 %
J48	91.56 %	8.44 %
<u>RandomForest</u>	100.00 %	0.00 %

Na figura seguinte apresentamos o sumário e a matriz de confusão para o algoritmo RandomForest:

```

=== Summary ===
Correctly Classified Instances      474      100 %
Incorrectly Classified Instances    0         0 %
Kappa statistic                    1
Mean absolute error                0.0307
Root mean squared error            0.0731
Relative absolute error            18.4389 %
Root relative squared error        25.3709 %
Total Number of Instances         474

=== Confusion Matrix ===
      a  b  c  d  e  f  g  h  i  j  <-- classified as
a = Um
b = Dois
c = Tres
d = Quatro
e = Cinco
f = Seis
g = Sete
h = Oito
i = Nove
j = Dez
0  0  0  0  0  0  0  0  0  0 |
0  8  0  0  0  0  0  0  0  0 |
0  0  49  0  0  0  0  0  0  0 |
0  0  0  86  0  0  0  0  0  0 |
0  0  0  0  113  0  0  0  0  0 |
0  0  0  0  0  95  0  0  0  0 |
0  0  0  0  0  0  71  0  0  0 |
0  0  0  0  0  0  0  36  0  0 |
0  0  0  0  0  0  0  0  15  0 |
0  0  0  0  0  0  0  0  0  1 |

```

Para treinar o modelo de classificação utilizámos os dados referentes aos anos de 2014, 2015 e 2016 mês a mês para cada companhia aérea, num total de 474 instâncias de treino que foram classificadas da seguinte forma:

- Dividimos a gama de valores de cada um dos atributos em 10 partes iguais, em que os valores extremos são o menor e o maior valor da gama de valores.
- Para cada registo classificámos cada um dos atributos de 1 a 10 de acordo com o seu valor.
- Por fim a classificação do desempenho operacional é igual ao valor médio de classificação dos atributos, ou seja, um valor de 1 a 10.

Na tabela abaixo encontra-se a classificação do desempenho operacional atribuída pelo classificador a cada companhia aérea no ano de 2017.

Companhia aérea	Classificação de 2016 (primeiros 3 meses)	Classificação de 2017 (primeiros 3 meses)	
Alaska Airlines Inc.	7	5	↓
American Airlines Inc.	6	7	↑
Atlantic Southeast Airlines	6	5	↓
Delta Airlines Inc.	7	7	—
Frontier Airlines Inc.	7	6	↓
Hawaiian Airlines Inc.	9	8	↓
JetBlue Airways	5	4	↓
Skywest Airlines Inc.	5	4	↓
Southwest Airlines Co.	7	6	↓
Spirit Airlines	5	6	↑
United Airlines Inc.	6	6	—
Virgin America	5	3	↓

A partir da tabela acima podemos constatar que a qualidade do desempenho operacional, regra geral diminuiu ligeiramente de 2016 para 2017, em igual período. As exceções são a 'Delta Airlines Inc.' e a 'United Airlines Inc.' que conseguiram manter o nível do ano transato, e a 'American Airlines Inc.' e a 'Spirit Airlines' que até conseguiram melhorar ligeiramente.

8. Conclusão

Com este trabalho projetamos e implementámos uma solução de BI completa destinada ao DOT e às companhias aéreas, com o propósito de satisfazer os seguintes objetivos de negócio:

- Fornecer dados estatísticos e ratings operacionais dos voos domésticos, importantes para a gestão global dos transportes pelo DOT.
- Permitir às companhias aéreas compararem o desempenho do seu serviço com os seus concorrentes.
- Permitir às companhias aéreas analisar os seus pontos fortes e fracos a fim de reduzir as suas ineficiências operacionais.
- Permitir saber se o investimento em hubs é vantajoso em termos de taxi in e taxi out.
- Ajudar as companhias aéreas a otimizar a sua frota de aviões e o número de funcionários a médio prazo através da previsão.

Referências

[1] <https://www.kaggle.com/usdot/flight-delays>

[2] https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time

[3] https://www.transtats.bts.gov/Data_Elements.aspx?Data=1