Business Intelligence – MEI – University of Coimbra

Análise dos voos nos EUA

<Meta 2>

Equipa "<nome da equipa>":
André Miguel Pereira Pinho, <apinho@student.dei.uc.pt>
Rafael Filipe Pereira Pinho, <rfpinho@student.dei.uc.pt>

Parte 4

Seleção de software

O projeto

- Este projeto tem como objetivo fornecer dados estatísticos e ratings operacionais dos voos domésticos nos EUA:
  - Importantes para a gestão global dos transportes pelo DOT
  - Para as companhias aéreas compararem o desempenho do seu serviço com os seus concorrentes

Requisitos de software

- Armazenamento e acesso a dados
  - Bom desempenho para processamento analítico
  - Integração com ferramentas externas para carregamento e análise de dados
- ETL
  - Modelação visual, automatização do processo ETL
  - Integração com ficheiros .csv
- OLAP
  - Fazer slice e dice, drill-down e roll-up
  - Integração fácil e eficiente com bases de dados

Abordagem

- Para selecionar as ferramentas, comparámos algumas ferramentas diferentes com as ferramentas exploradas durante as aulas, onde analisámos as suas vantagens e desvantagens

Critérios de seleção da base de dados

- 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
- Interface da base de dados simples e intuitiva

Comparação das bases de dados

	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

Critérios de seleção da ferramenta ETL

- 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

	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

Critérios de seleção da ferramenta OLAP

- 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

	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

Lista de software selecionado

- Base de dados: PostgreSQL
- ETL: Pentaho data integration (kettle)
- OLAP: nenhuma, fizemos queries ad hoc

Parte 5

ETL

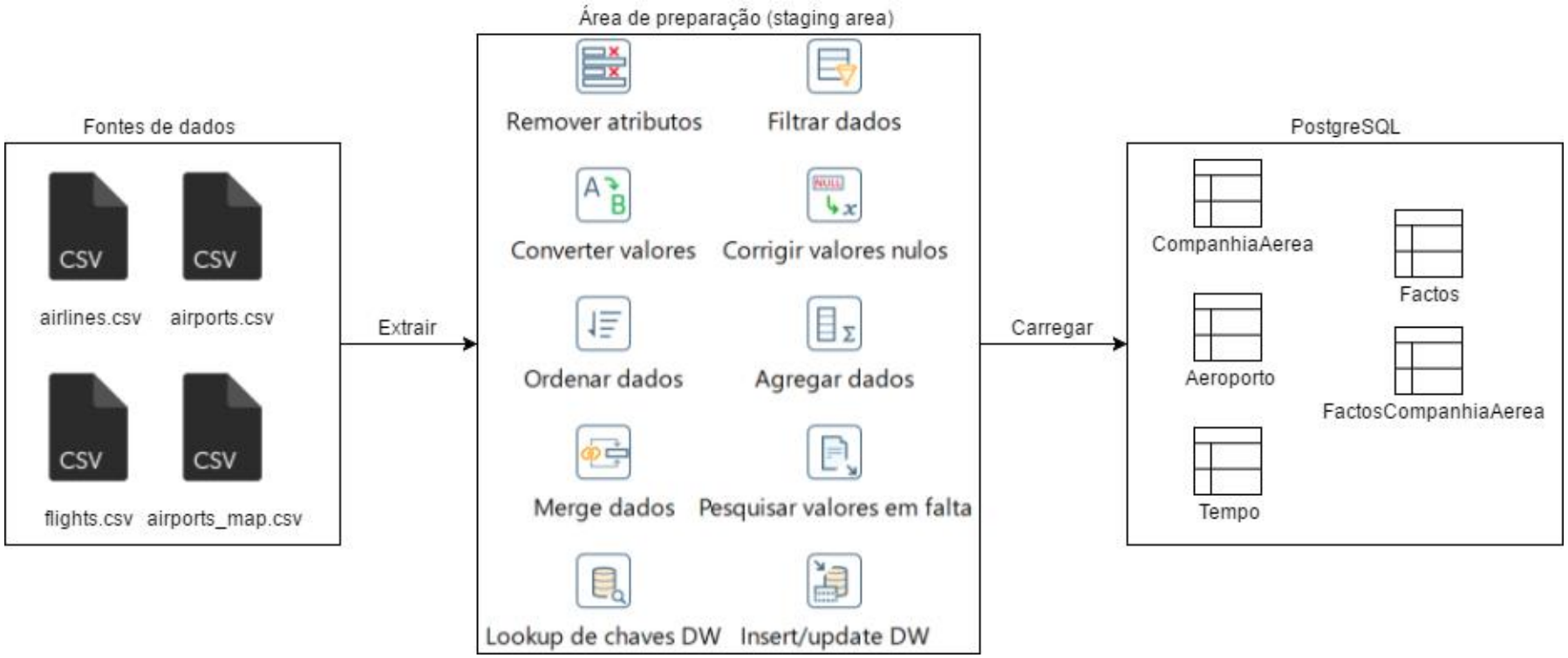
Modelo de dados

Fontes de dados

- Kaggle [1], forneceu informação
  - Voos domésticos em 2015
  - Companhias aéreas e dos aeroportos
- Governo dos EUA [2], forneceu informação
  - Lucros das companhias aéreas
  - Ficheiro para fazer o mapeamento dos aeroportos nos voos do mês de outubro

[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

Plano ETL

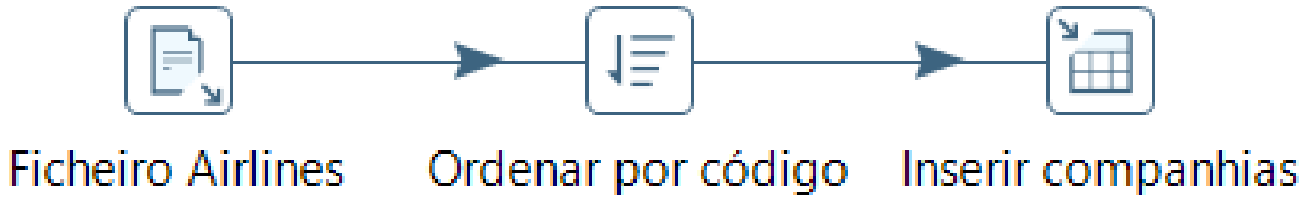

Área de preparação (staging area)
Fontes de dados
CSV
airlines.csv
CSV
airports.csv
CSV
flights.csv
CSV
airports_map.csv
Extrair
Remover atributos
Filtrar dados
Converter valores
Corrigir valores nulos
Ordenar dados
Agregar dados
Merge dados
Pesquisar valores em falta
Lookup de chaves DW
Insert/update DW
Carregar
PostgreSQL
CompanhiaAerea
Factos
Aeroporto
FactosCompanhiaAerea
Tempo

Área de preparação (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

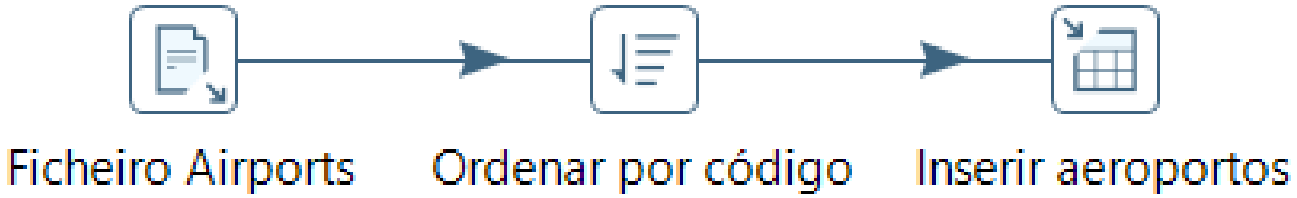
Atividade #01

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



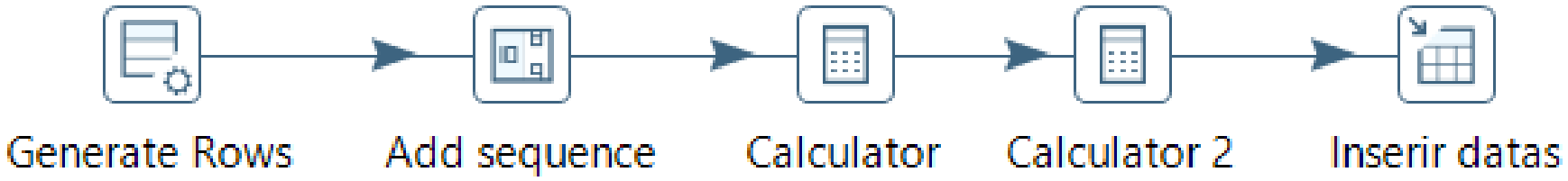
Atividade #02

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



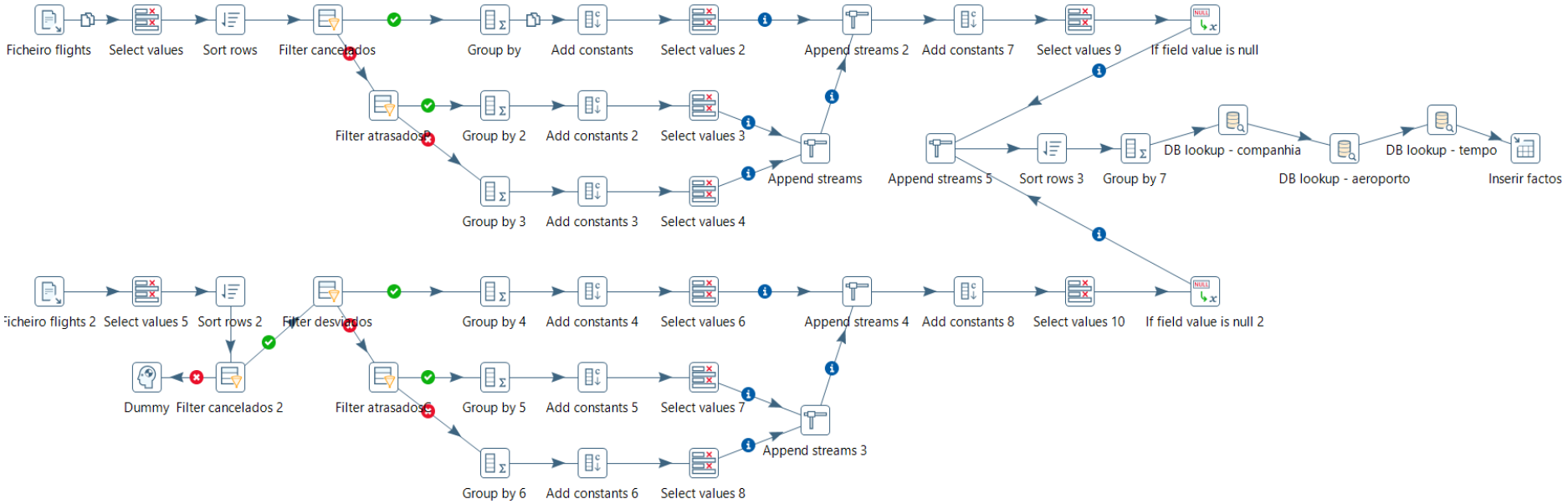
Atividade #03

- Dimensão Tempo
  - Para carregar a dimensão tempo utilizámos um gerador de linhas com o formato date, e com uma sequência incrementamos a data dia a dia e inserimos no data warehouse



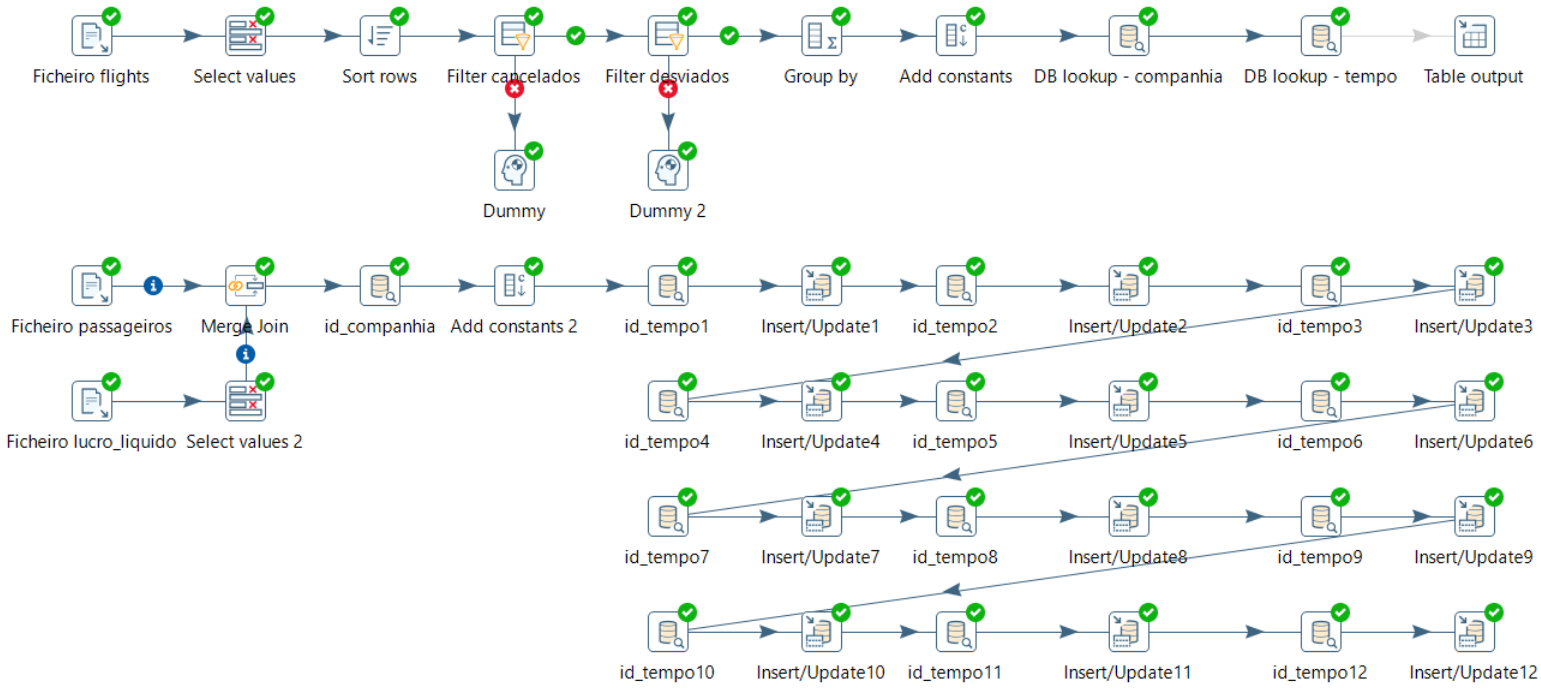
Atividade #04

- Tabela de Factos



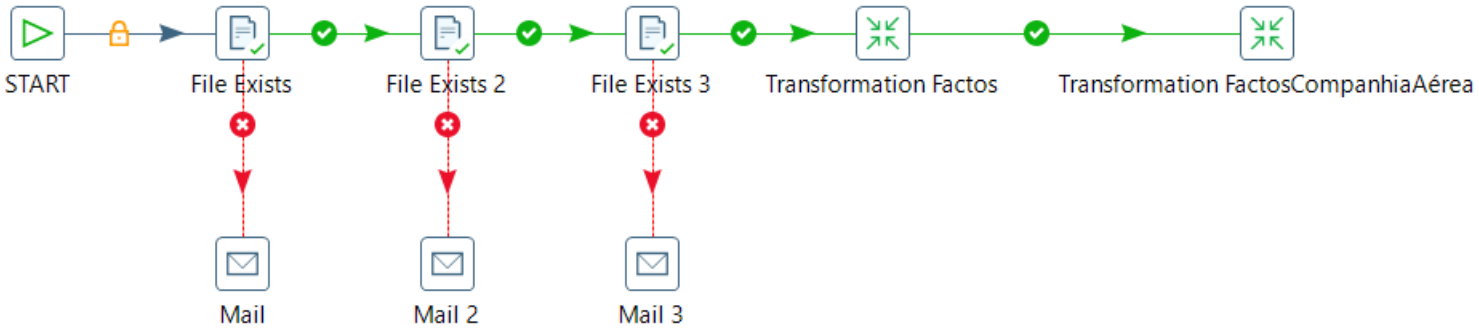
Atividade #05

- Tabela de FactosCompanhiaAerea



Atualização automática do DW

- Criámos um cron scheduler (kettle job) que dispara no início de cada ano para fazer a atualização das tabelas de factos



ETL - Métricas

- Tamanho dos dados de origem: 580 Mb divididos em 3 ficheiros .csv
- Tamanho dos dados após o primeiro carregamento: 74 Mb
  - Factos com 72 Mb (389.023 registos)
  - FactosCompanhiaAerea: 1 Mb (4.932 registos)
- Tempo decorrido na primeira carga: 1 hora e 2 minutos
- Tamanho dos dados nas cargas subsequentes: será semelhante à primeira carga

ETL - Fora do scope

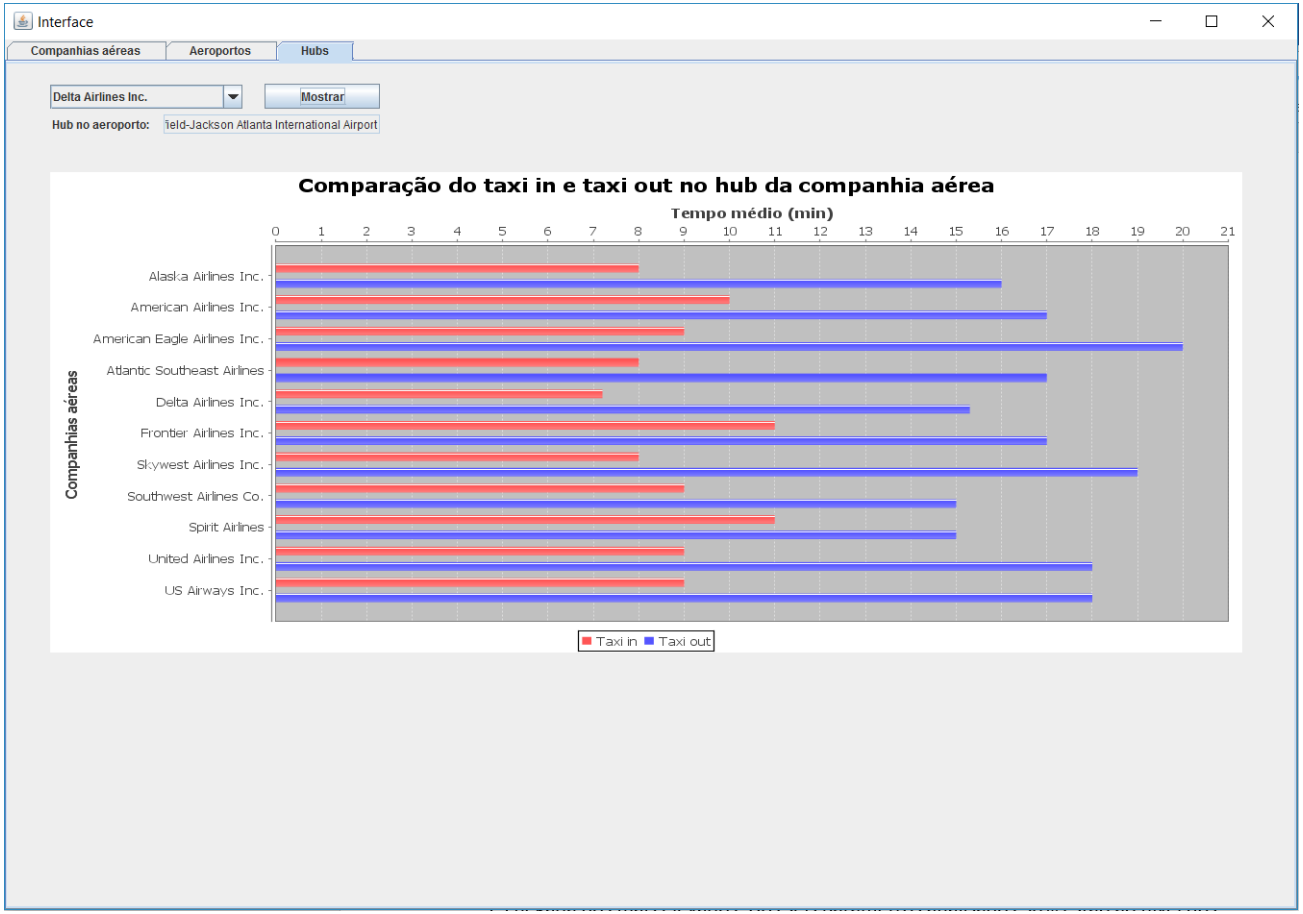
- Todos os problemas identificados foram resolvidos com sucesso

OLAP

Usabilidade e acesso a dados

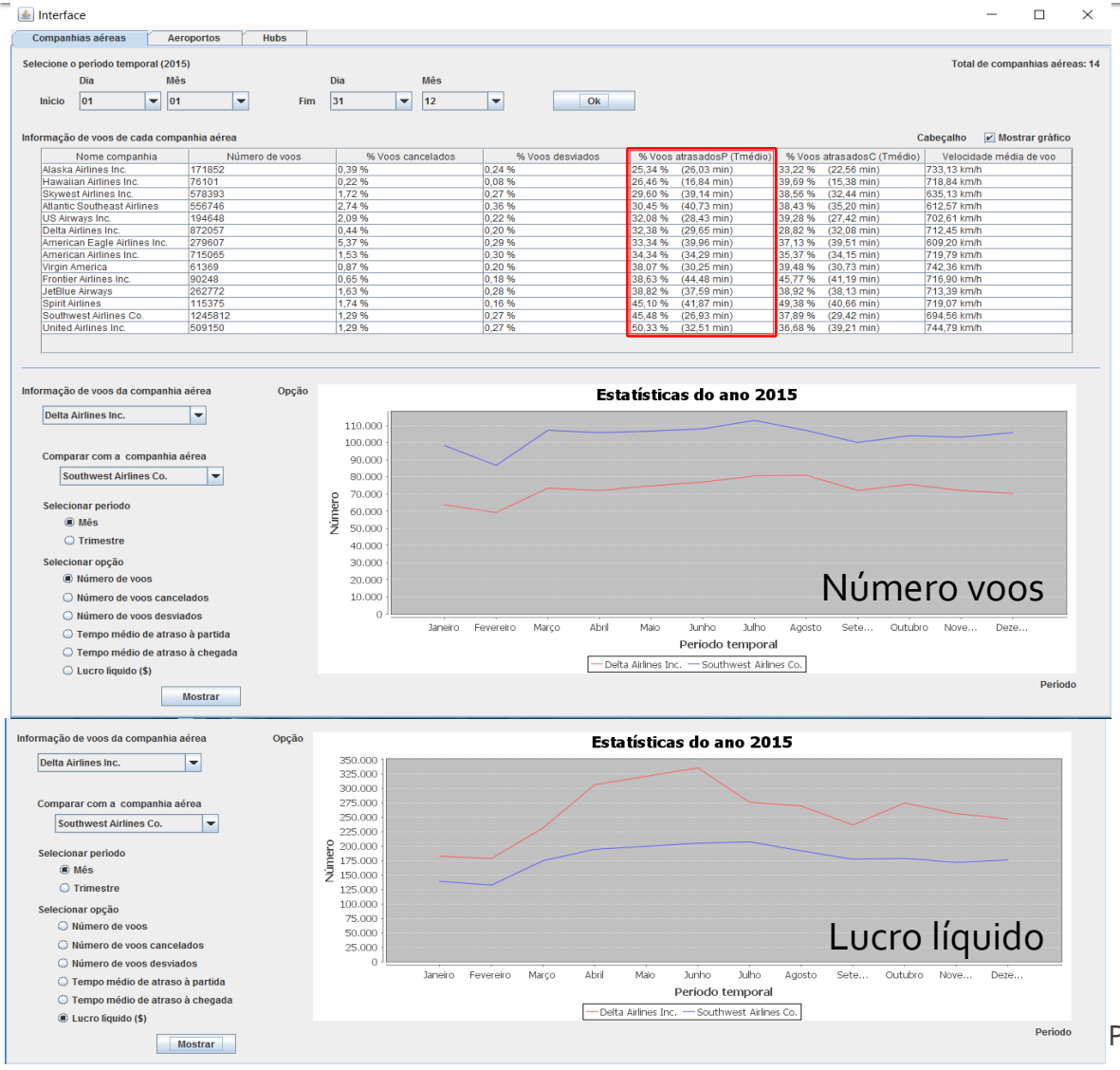
- A informação é apresentada aos utilizadores através de uma dashboard
  - Contém tabelas e gráficos comparativos
  - Várias opções de drill-down e roll-up

Análise #01 - Hubs das companhias aéreas



Neste tipo de análise ficou demostrado que o investimento (aluguer de espaço) em hubs não trás vantagens significativas em termos de taxi_in e taxi_out

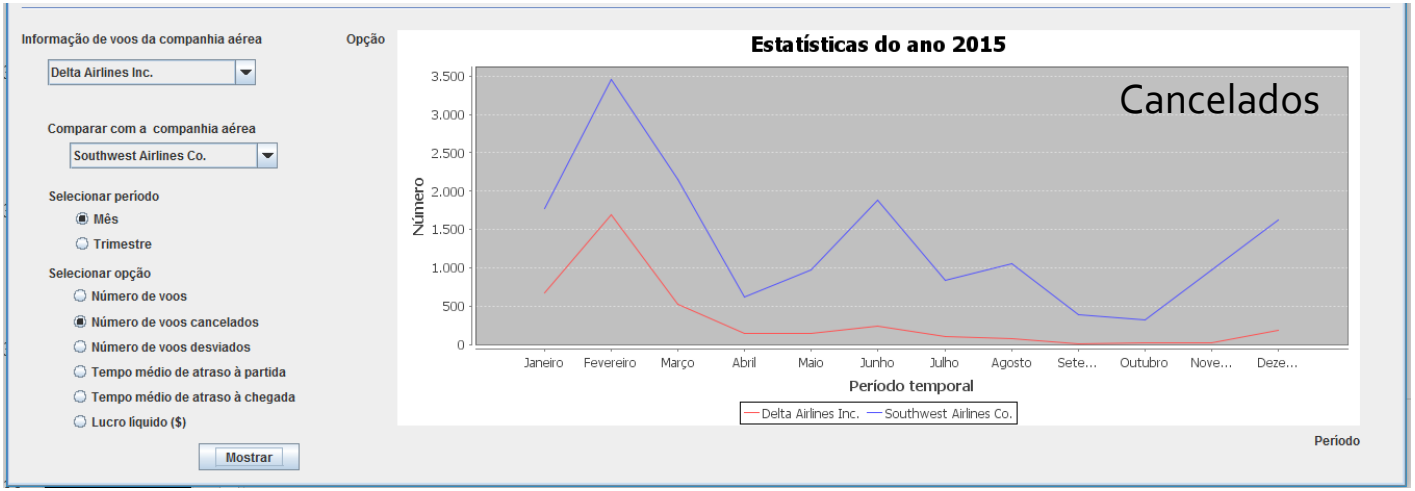
Análise #02 - Benchmarking entre as duas maiores companhias aéreas

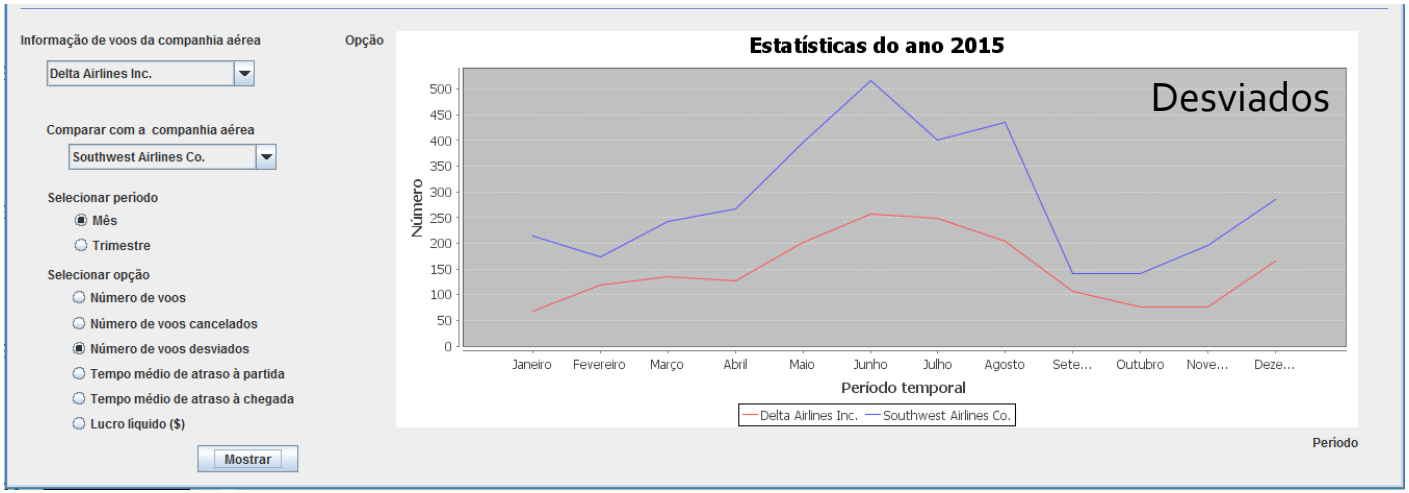


A tabela permite às companhias compararem rácios operacionais com os seus concorrentes assim como analisar os seus pontos fortes e fracos

A companhia com o maior número de voos 'Southwest Airlines Co.', têm um lucro líquido inferior à 'Delta Airlines Inc.'

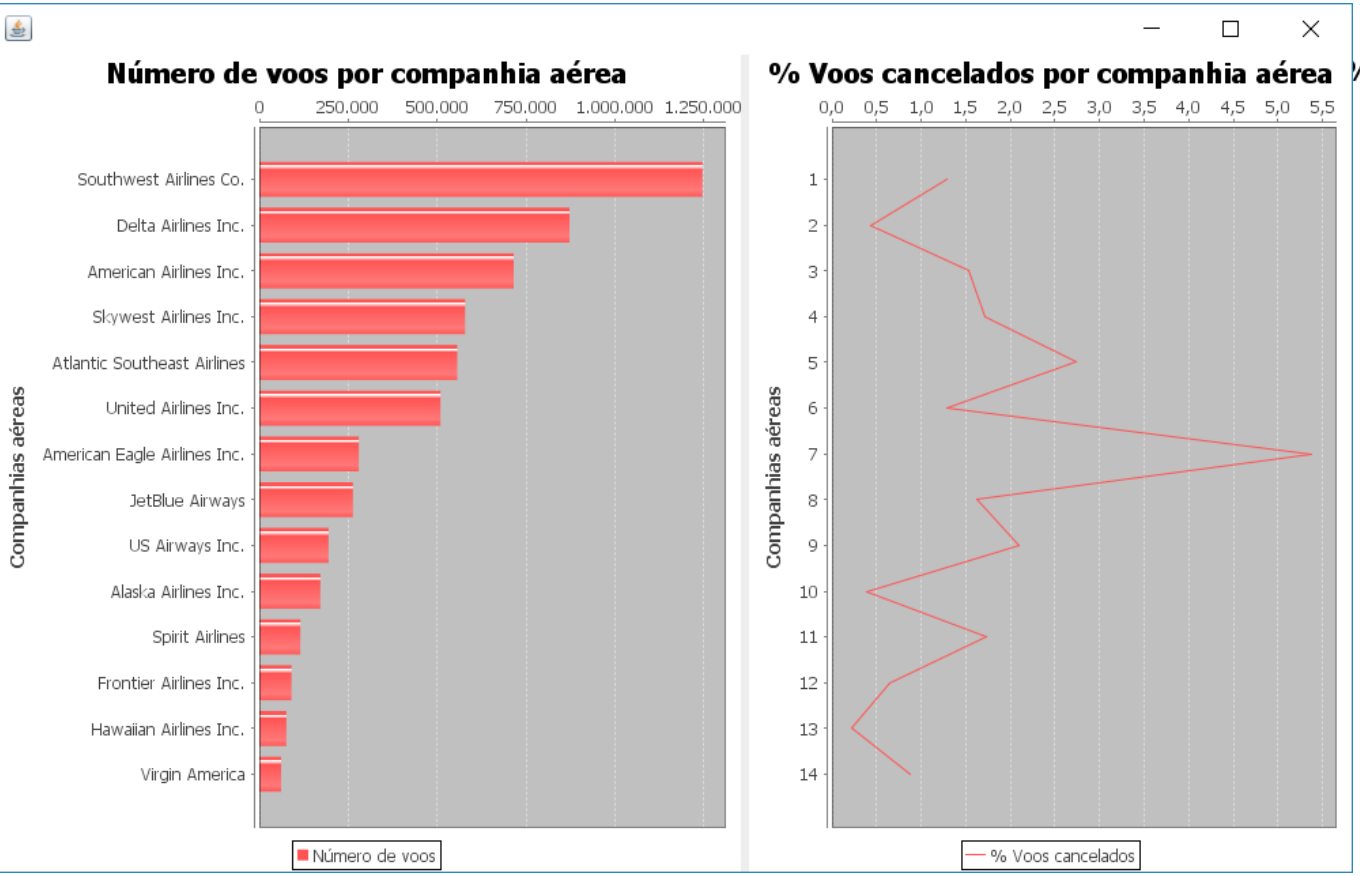
Análise #03 - Benchmarking entre as duas maiores companhias [cont.]

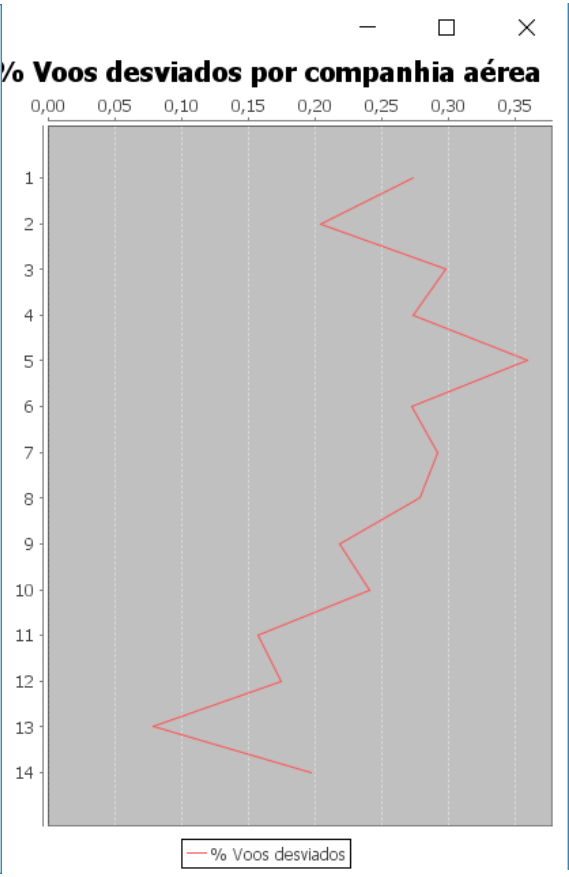




Um dos motivos desta diferença, deve-se ao maior número de voos cancelados e desviados

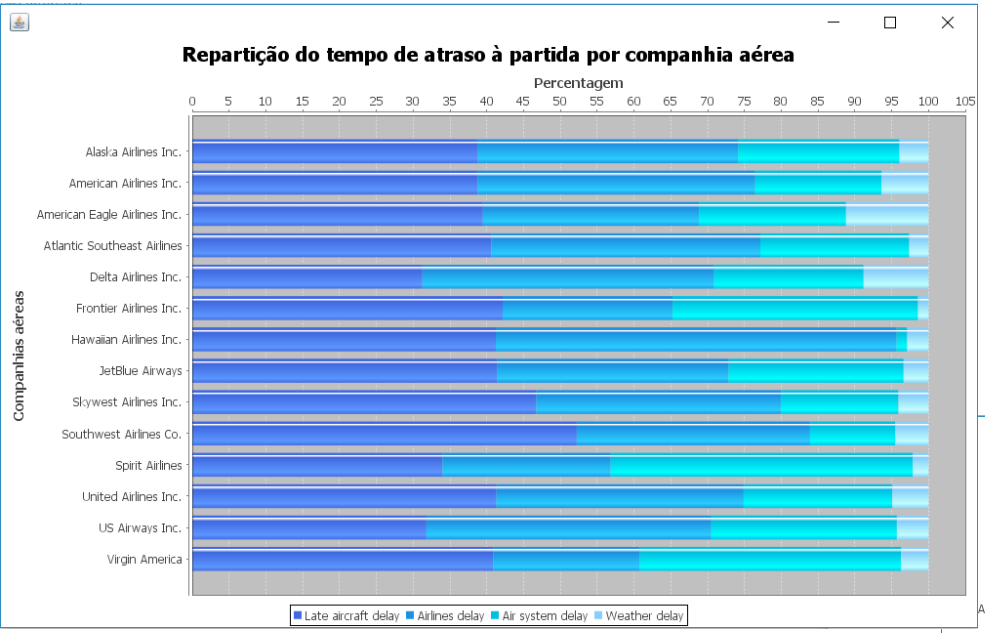
Análise #04 – Comparação da percentagem de voos cancelados e desviados por companhia aérea



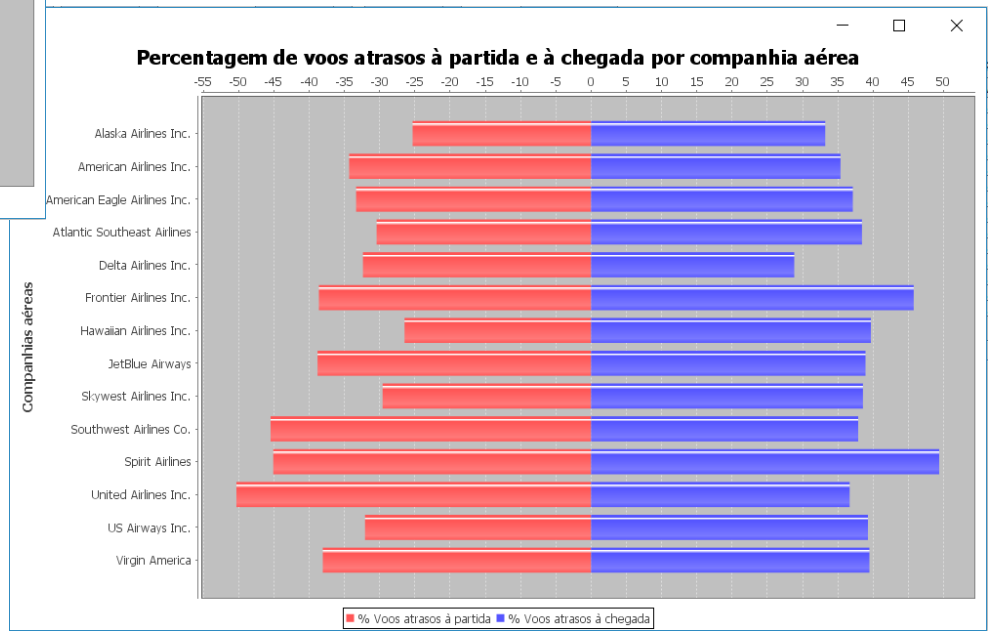


A companhia com pior desempenho operacional é a 'American Eagle Airlines Inc.', devido à maior percentagem de voos cancelados, e também por ter uma das maiores percentagens de voos desviados

Análise #05 - Comparação dos atrasos detalhados por companhia aérea

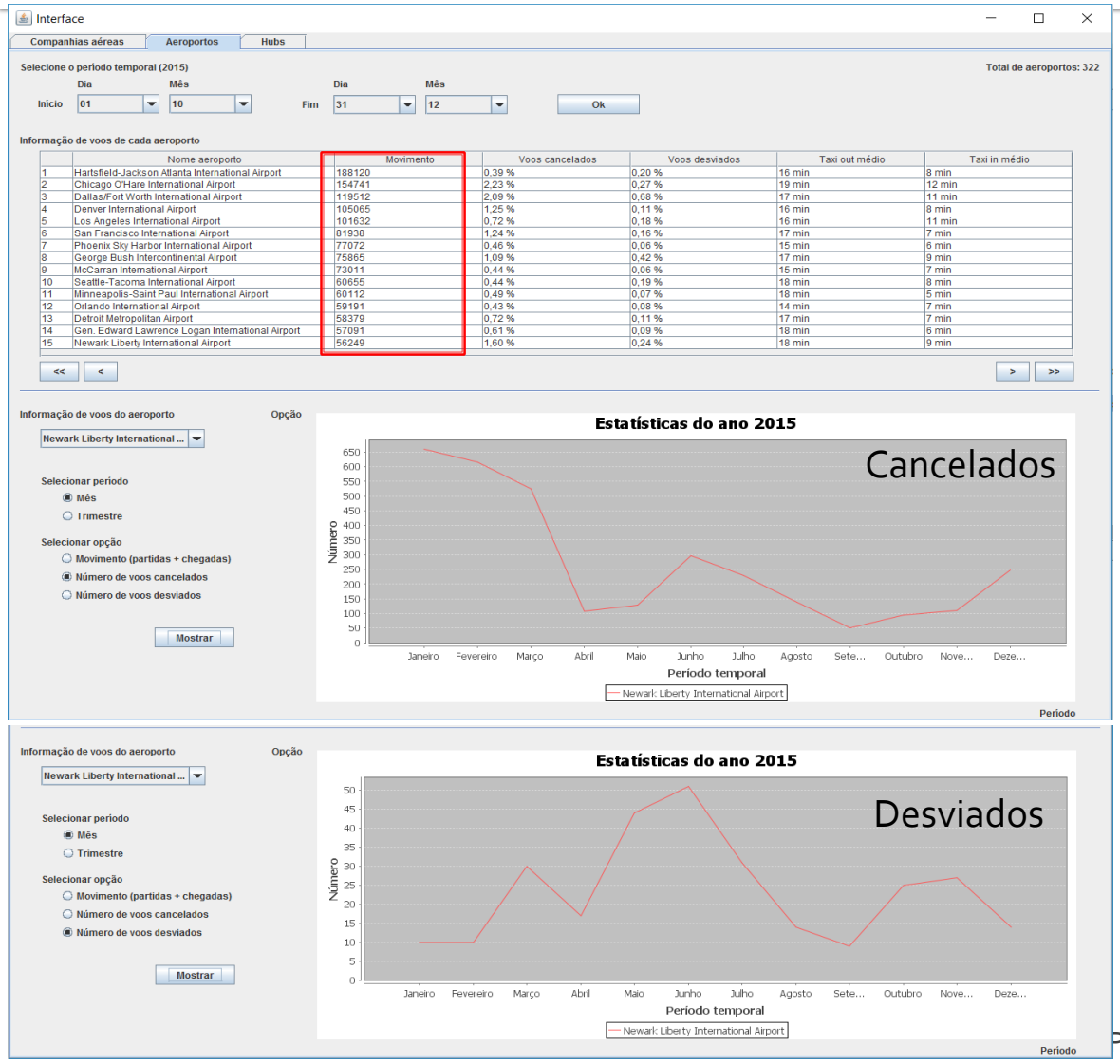


O gráfico da esquerda permite visualizar a repartição dos tempos de atraso à partida num dado período de tempo definido
- Sistema aéreo
- Verificações técnicas
- Condições meteorológicas



O gráfico da direita permite comparar o tempo médio de atraso à partida e à chegada num dado período de tempo definido

Análise #06 - Perspetiva da gestão aeroportuária pela DOT



A tabela permite ao DOT obter um panorama geral da gestão aeroportuária do país, como por exemplo os aeroportos mais movimentados, com mais voos cancelados, desviados, ...

Constatamos que existem:
- Mais voos cancelamentos no inverno
- Mais voos desviados no verão

Objetivos de negócio e DSS

- Fornecer **dados estatísticos e ratings operacionais** dos voos domésticos, importantes para a gestão global dos transportes pelo DOT
- Permitir ás companhias aéreas **compararem o desempenho do seu serviço** com os seus concorrentes
- Permitir ás companhias aéreas analisar os seus **pontos fortes e fracos** a fim de reduzir as suas **ineficiências operacionais**
- Pela análise dos resultados ficou demostrado que o investimento em hubs não trás vantagens significativas

Otimização de desempenho do DW

- Indexação das chaves estrangeiras das duas tabelas de factos para acelerar as junções
- Duas vistas materializadas para armazenar resultados pré calculados e acelerar a resposta às queries

Questões