Estudo estatístico dos Voos Brasileiros

Projeto da Disciplina de BI

Prof. Anderson Nascimento
prof.anderson@ica.ele.puc-rio.br

Componentes do Projeto:

Ana Lucia Rucos – anarucos@gmail.com
Daniel Galdino – daniel.galdinex@gmail.com
Leandro Saldini – lsaldini@gmail.com
Leonardo Lins – leonardolins@yahoo.com
Marcus Vidon – marcusvidon@gmail.com

Histórico de Versões

Data	Versão	Descrição	Autor	Aprovado por
01/10/2021	1	Esboço do projeto	Ana Rucos	Daniel Galdino
24/10/2021	2	Revisão	Daniel Galdino	Ana Rucos
01/11/2021	3	Introdução, Dashboard, Conclusão	Marcus Vidon	Ana Rucos
05/11/2021	4	Revisão Final	Daniel Galdino	Todos

Sumário

1. Introdução

Este documento tem como objetivo descrever todas as etapas envolvidas no processo de criação de *dashboard* para análise das estatísticas de voos, nacionais e internacionais fornecidos pela Agência Nacional de Aviação Civil.

Nos tópicos a seguir serão detalhadas as componentes dos dados de origem, o processo de internalização dessas informações em um *data warehouse* e, finalmente, as visões disponibilizadas em *dashboard*.

Os *softwares* utilizados nesse processo foram:

- Modelagem do *Datawarehouse*: *Power Architect (V1.0.9)*;
- Fluxo *ETL*: *Pentaho (v9.2.0.0)*;
- Banco de Dados: *Postgres (v5.7)*;
- *Dashboard*: *Power BI (versão 2.97.861)*.

2. Estudo de Caso

2.1. Descrição do Estudo de Caso

Com o intuito de ampliar o conhecimento da sociedade brasileira e de subsidiar a realização de pesquisas, estudos e análises mais abrangentes sobre o setor, a ANAC tem disponibilizado na seção "Dados e Estatísticas" do seu portal na internet, relatórios, estudos e informações sobre as condições de mercado.

Nesse sentido, a Agência agrega, para a livre consulta de qualquer interessado, a série histórica dos dados estatísticos do transporte aéreo do Brasil com elevado grau de detalhamento.

Estão contempladas informações sobre a quantidade de passageiros, carga e mala postal transportados, distância voada, combustível consumido, entre outras, por etapa de voo e por empresa aérea.

Essas informações podem ser utilizadas para apurar importantes indicadores do setor, como demanda (RPK, RTK), oferta (ASK, ATK), participação de mercado, taxa de ocupação das aeronaves (Load Factor), entre outros.

Os dados estatísticos do transporte aéreo do Brasil encontram-se regulamentados pela Resolução ANAC no 191/2011 e pelas Portarias ANAC no 1.189 e 1.190/SRE/2011.

De acordo com a mencionada regulamentação, os dados são mensalmente fornecidos à ANAC, até o dia 10 do mês subsequente ao de referência pelas empresas brasileiras e estrangeiras que exploram os serviços de transporte aéreo público regular e não regular no Brasil.

Com base nas informações disponíveis foi elaborado um estudo para identificar a evolução do setor aéreo e o impacto da Pandemia nos voos nacionais e internacionais no Brasil.

3. Descrição do Modelo Transacional

O modelo transacional foi extraído do site da ANAC através de upload de uma tabela única de dados em Excel. Além disso, algumas informações foram enriquecidas com fontes externas, como por exemplo, a localização geográfica (latitude e longitude) dos aeroportos.

3.1. Fonte 1 - Tabela de Dados Estatísticos - ANAC

A fonte principal de dados utilizada no projeto foi a tabela de Dados Estatísticos disponibilizada no site da ANAC na página: https://www.anac.gov.br/acesso-a-informacao/dados-abertos/areas-de-atuacao/voos-e-operacoes-aereas/dados-estatisticos-do-transporte-aereo

Os dados são disponibilizados mensalmente e são acumulativos, ou seja, mensalmente é disponibilizado no servidor da ANAC uma tabela com o histórico de dados acrescido das informações atualizadas do último mês.

Segue abaixo detalhamento dos campos informados na tabela Excel:

Campo	Descrição
Empresa Aérea	Empresa Aérea responsável por operar as etapas.
Natureza do Voo	Refere-se à natureza das etapas e possui o valor "Doméstico" caso as etapas tenham o pouso e a decolagem realizadas no Brasil e sejam operadas por Empresas brasileiras ou possuem o valor "Internacional" caso contrário.
Tipo de Voo	Faz referência ao tipo de operação das etapas: *Improdutivas* (Non-revenue flights): etapas que não geraram receita à empresa aérea (como realização de treinamentos, voo para manutenção de aeronaves); *Regulares* (Scheduled revenue flights): etapas remuneradas que são realizadas sob uma numeração de Horário de Transporte (HOTRAN). Recebem esse nome, pois possuem a característica de serem realizadas regularmente; e *Não Regulares* (Non-scheduled revenue flights): etapas remuneradas que não são realizadas sob uma numeração de Horário de Transporte (HOTRAN). Recebem esse nome, pois possuem a característica de serem realizadas de forma não continuada. Aqui estão os voos Charters, Fretamentos.
ASK (Available seat kilometer)	Refere-se ao volume de Assentos Quilômetros Oferecidos, ou seja, a soma do produto entre o número de assentos oferecido e a distância das etapas.
RPK (Revenue seat kilometer)	Refere-se ao volume de Passageiros Quilômetros Transportados, ou seja, a soma do produto entre o número de passageiros pagos e a distâncias das etapas.
ATK (Available tonne kilometer)	Refere-se ao volume de Tonelada Quilômetro Oferecida, ou seja, a soma do produto entre o payload, que é a capacidade total de peso disponível na aeronave, expressa em quilogramas, disponível para efetuar o transporte de passageiros, carga e correio, e a distância das etapas, dividido por 1.000.

RTK (Revenue tonne kilometer)	Refere-se ao volume de Toneladas Quilômetros Transportadas, ou seja, a soma do produto entre os quilogramas carregados pagos, onde cada passageiro possui o peso estimado de 75 Kg, e a distância das etapas, dividido por 1.000.
Combustível	Refere-se à quantidade, em litros, de combustível consumida pela aeronave na execução da referida etapa. Informação disponível apenas para empresas brasileiras.
Distância	Refere-se à distância, expressa em quilômetros, entre os aeródromos de origem e destino da etapa, considerando a curvatura do planeta Terra.
Horas Voadas	Refere-se ao número de horas de voo entre os aeródromos de origem e destino da etapa.
Decolagens	Refere-se ao número de decolagens que ocorreram entre os aeródromos de origem e destino da etapa.
Carga Paga Km	Refere-se ao volume de Carga Paga (kg) em cada quilômetro, ou seja, a soma do produto entre a quantia (kg) de carga paga e a distâncias das etapas.
Carga Grátis Km	Refere-se ao volume de Carga Grátis (kg) em cada quilômetro, ou seja, a soma do produto entre a quantia (kg) de carga grátis e a distâncias das etapas.
Correio Km	Refere-se ao volume de Correio (kg) em cada quilômetro, ou seja, a soma do produto entre a quantia (kg) de correio e a distâncias das etapas.
Assentos	É o número de assentos disponíveis em cada etapa de voo de acordo com a configuração da aeronave na execução da etapa; e
Payload (Kg) (Payload capacity)	é a capacidade total de peso na aeronave, expressa em quilogramas, disponível para efetuar o transporte de passageiros, carga e correio.
Etapa Combinada (On flight origin and destination - OFOD)	As etapas combinadas identificam os pares de aeródromos de origem, onde houve o embarque do objeto de transporte, e destino, onde houve o desembarque do objeto de transporte, independentemente da existência de aeródromos intermediários, atendidos por determinado voo. É a etapa de voo vista com foco no objeto de transporte (pessoas e/ou cargas), com base no embarque e desembarque nos aeródromos relacionados. Os dados estatísticos da etapa combinada informam a origem e destino no voo, dos passageiros e cargas transportadas, independente das suas escalas.
Empresa Aérea	Empresa Aérea responsável por operar as etapas.
Natureza do Voo	Refere-se à natureza das etapas, e possui o valor "Doméstico" caso as etapas tenham o pouso e a decolagem realizadas no Brasil e sejam operadas por Empresas brasileiras ou possuem o valor "Internacional" caso contrário.
Tipo de Voo	Faz referência ao tipo de operação das etapas: *Improdutivas* (Non-revenue flights): etapas que não geraram receita à empresa aérea (como realização de treinamentos, voo para manutenção de aeronaves); *Regulares* (Scheduled revenue flights): etapas remuneradas que são realizadas sob uma numeração de Horário de Transporte (HOTRAN). Recebem esse nome, pois possuem a característica de serem realizadas regularmente; e

	Não Regulares (Non-scheduled revenue flights): etapas remuneradas que não são realizadas sob uma numeração de Horário de Transporte (HOTRAN). Recebem esse nome, pois possuem a característica de serem realizadas de forma não continuada. Aqui estão os voos Charters, Fretamentos.
Passageiros Pagos	Refere-se aos passageiros que ocupam assentos comercializados ao público e que geram receita, com a compra de assentos, para a empresa de transporte aéreo. Incluem-se nesta definição as pessoas que viajam em virtude de ofertas promocionais, as que se valem dos programas de fidelização de clientes, as que se valem dos descontos concedidos pelas empresas, as que viajam com tarifas preferenciais, as pessoas que compram passagem no balcão ou através do site de empresa de transporte aéreo e as pessoas que compram passagem em agências de viagem.
Passageiros Grátis	Refere-se aos passageiros que ocupam assentos comercializados ao público, mas que não geram receita, com a compra de assentos, para a empresa de transporte aéreo. Incluem-se nesta definição as pessoas que viajam gratuitamente, as que se valem dos descontos de funcionários das empresas aéreas e seus agentes, os funcionários de empresas aéreas que viajam a negócios pela própria empresa e os tripulantes ou quem estiver ocupando assento destinado a estes.
Carga Paga	Refere-se à quantidade total, expressa em quilogramas, de todos os bens que tenham sido transportados na aeronave, exceto correio e bagagem, e tenham gerado receitas direta ou indireta para a empresa aérea.
Carga Grátis	Refere-se à quantidade total, expressa em quilogramas, de todos os bens que tenham sido transportados na aeronave, exceto correio e bagagem, e não tenha gerado receitas diretas ou indiretas para a empresa aérea.
Correio	Refere-se à quantidade de objetos transportados de rede postal em cada trecho de voo realizado, expresso em quilogramas.
Bagagem	Refere-se à quantidade total de bagagem despachada, expressa em quilogramas.

Obs.: Os Dados Estatísticos possuem grande semelhança em relação ao programa estatístico da *International Civil Aviation Organization* (ICAO), em especial no significado das variáveis. Assim, para facilitar a comparação, os nomes de algumas variáveis são apresentados também em inglês.
Apesar de estarem dispostas como informações de etapas básicas, as variáveis RPK, RTK, Carga Paga Km, Carga Grátis Km e Correio Km das empresas estrangeiras são computadas por meio das informações congêneres advindas das etapas combinadas. Assim, ao se desagregar essas variáveis de demanda das empresas estrangeiras por aeroporto, ou por rota, e as comparar com as informações de oferta (ASK e ATK, originadas das etapas básicas), é possível que os valores das taxas de aproveitamento estejam subdimensionados ou superdimensionados, e, em alguns casos, estarem superiores a 100%. Por exemplo, o Load Factor (RPK/ASK) de empresas estrangeiras, desagregado por aeroporto, pode ser superior a 100%, devido à formatação dos dados.

3.2. Fonte 2 - Tabela de Aeroporto - fonte Externa

Como um dos objetivos finais é plotar um mapa com o posicionamento dos aeroportos, a dimensão aeroporto precisou ser enriquecida com uma fonte externa com informações de latitude e longitude deles.
Tais informações foram extraídas da internet e exportadas em Excel para posterior enriquecimento.

Fonte de dados originais (complementares):
https://datahub.io/core/airport-codes#data
https://ourairports.com/world.html

4. Proposta de Processo de BI

Esta seção apresenta o processo de BI fim a fim proposto para o projeto.

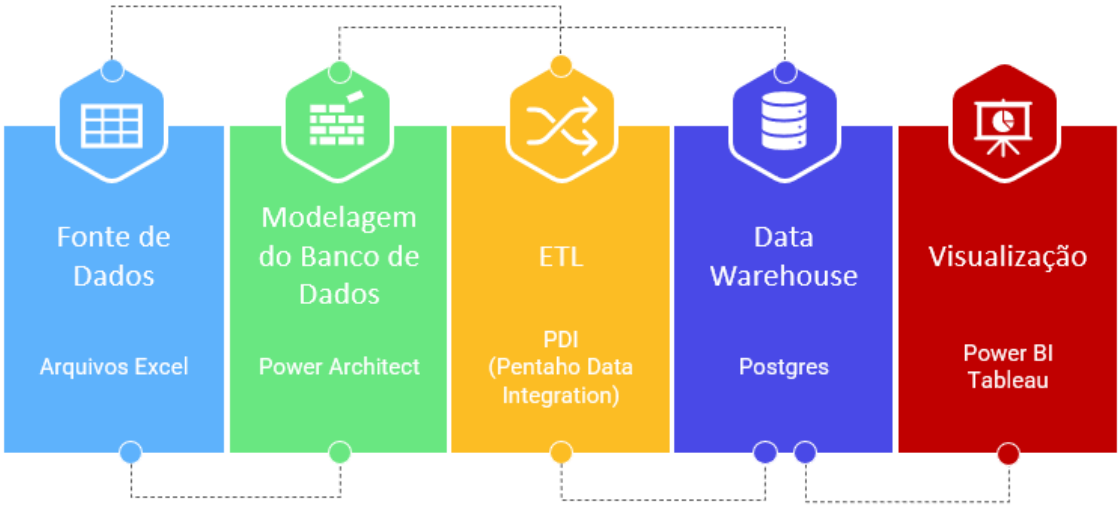

Figura 1 – Processo de BI

As fontes de dados utilizadas no projeto são em arquivos Excel conforme detalhado nos itens 3.1 e 3.2 citados anteriormente.

A modelagem do Datawarehouse foi feita no Power Architect *(V1.0.9)*: ferramenta livre e de código aberto.

A ferramenta de ETL utilizada na extração, transformação e carregamento dos dados foi o PDI (Pentaho Data Integration - *v9.2.0.0*), software de código aberto desenvolvido em Java.

Os dados foram armazenados no Postgres (*v5.7)*: sistema gerenciador de banco de dados de objeto relacional também de código aberto.

A visualização dos dados foi através do Power BI (*versão 2.97.861)* que é um software da Microsoft desenvolvido para fornecer análises interativas de BI (*Business Inteligence*).

5. Modelo Multidimensional

A tabela de dados disponibilizada pela ANAC fornece informações estatísticas sobre os voos como as listadas acima no item 3.1. Além dos dados estatísticos, foram identificadas informações dimensionais como dados temporais, cadastros de aeroportos e empresas.

Estas 3 classes formaram as 3 dimensões do modelo multidimensional relacionadas em diagrama estrela com a tabela fato (ft_dadosanac) conforme figura abaixo:

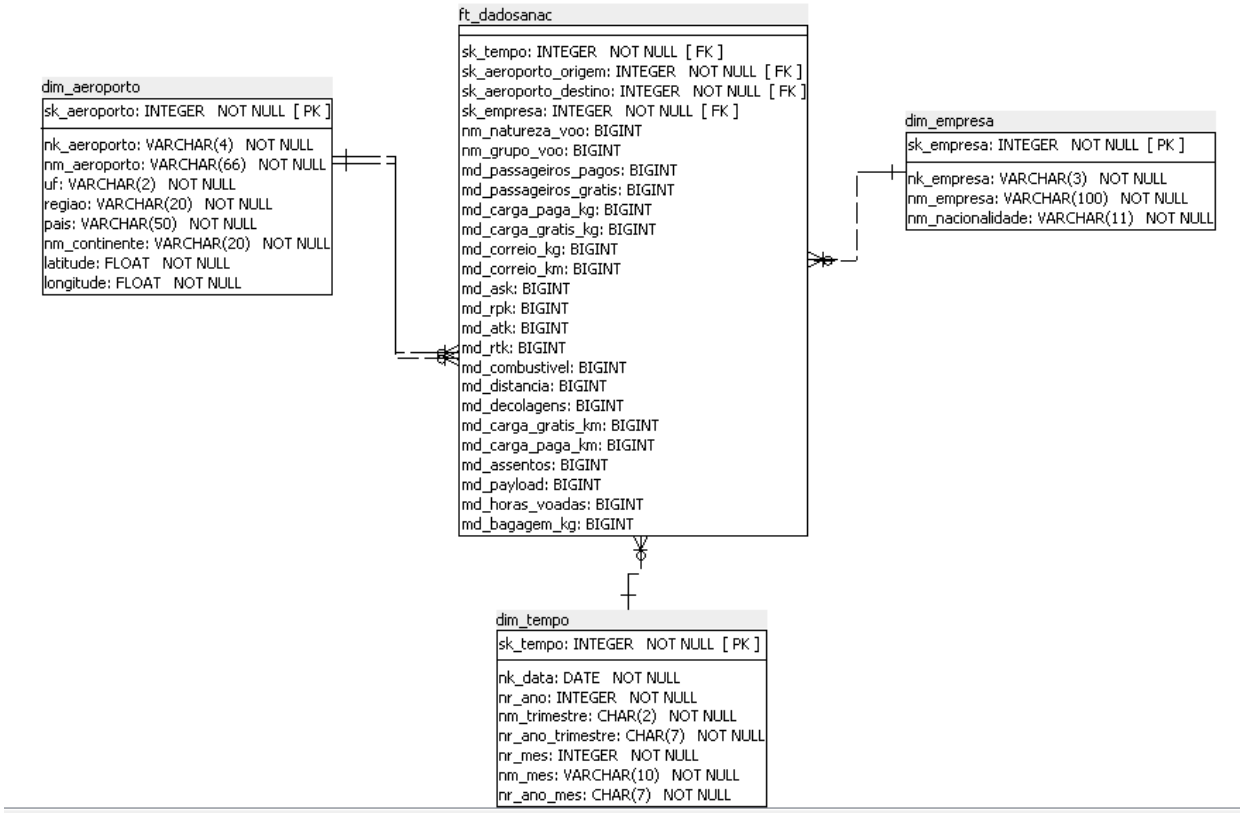

Figura 2 – Diagrama multidimensional

6. Elaboração do Data Warehouse

6.1. Definição do DW

O Data Warehouse será a fonte integradora de informações. A tecnologia será utilizada com o intuito de servir de base para a camada de aplicação que será responsável por fornecer dados para a tomada de decisão. Essa é a estrutura propriamente dita de armazenamento das informações decisivas. Apenas os dados com valor para a gestão corporativa estão reunidos no DW.

O DW com dados estatísticos da ANAC apresenta 3 tabelas dimensionais. São elas:

- Dimensão de Tempo (dim_tempo):
  Esta dimensão contém as referências de tempo utilizadas para cruzamento com a tabela fato assim como as hierarquias temporais como ano, trimestre e mês. A menor granularidade desta dimensão é mês, já que os dados disponibilizados pela ANAC são mensais;

- Dimensão Aeroporto (dim_aeroporto):
  Nesta dimensão é feito o cadastro de todos os aeroportos nacionais e internacionais que apresentam algum relacionamento com os aeroportos brasileiros. A dimensão aeroporto se relaciona com a fato duas vezes, já que os voos apresentam sempre como característica um aeroporto de origem e outro de destino;

- Dimensão Empresa (dim_empresa):
  Contém o cadastro das empresas brasileiras ou estrangeiras que tiveram operações vinculadas a aeroportos brasileiros.

O DW apresenta ainda duas dimensões degeneradas (que não apresentam sua própria dimensão) e estão na tabela fato. São elas: **natureza do voo** e **grupo do voo**. Os detalhes das mesmas são apresentados no item 6.2.4.

6.1.1. Arquitetura

A arquitetura utilizada no projeto é de Data Marts integrados. Apesar de os DM serem implementados separadamente por grupos de trabalho ou departamentos, eles são integrados ou interconectados, provendo uma visão corporativa maior dos dados e informações.

Este tipo de implementação se fez útil pela riqueza das informações que podem ser de grande valia para os usuários de diversos departamentos.

A qualidade e variedade do cadastro das empresas e dos aeroportos nas suas respectivas dimensões assim como as estatísticas apresentadas podem ser acessados e utilizados por DM de outros departamentos (compartilhamento).

6.1.2. Abordagem de Construção

A abordagem de construção utilizada nesse projeto foi a *Bottom Up* já que apresenta uma rápida implementação e é focada no problema.

Esse tipo de construção se baseia no “dividir para conquistar”. Como não tínhamos uma visão geral da complexidade do nosso banco, se tornou importante a construção incremental dos Data Marts até se chegar ao DW final. Isso capacita o projeto a apresentar um desenvolvimento evolutivo e ter menores riscos.

Outra vantagem é que esse tipo de construção nos deu um retorno mais rápido, já que a estrutura inicial (Data Marts) já poderia ser consultada antes mesmo de se completar a construção do DW final.

6.1.3. Arquitetura Física

A arquitetura física utilizada nesse projeto foi a *On-Premises*. O volume de dados não era tão grande e foi utilizada uma máquina virtual com todo dado carregado para tratamento, armazenamento e visualização das informações.

As ferramentas utilizadas foram, em sua maioria, de código aberto e de simples instalação, justificando assim o uso de arquitetura *on-premises*.

6.2. Arquitetura final do DW

Abaixo apresentamos a arquitetura final do DW com descrição de todos os campos envolvidos.

6.2.1. Tabela Dimensional: dim_tempo

A dim_tempo contém informações de tempo e se relaciona com a fato pela chave sk_tempo.

sk_tempo	Inteiro / Chave Primária
nk_data	Tipo: Date Valor completo da data. Para esta tabela, como a periodicidade é mensal, refere-se ao primeiro dia do mês.
nr_ano	Tipo: INTEGER Valor numérico do ano (YYYY)
nm_trimestre	Tipo: CHAR Sigla do trimestre (T1, T2 ou T3)
nr_ano_trismestre	Tipo: CHAR

	Junção do campo nm_trimestre com o nr_ano (Ex.: T1/2021, T2/2021)
nr_mes	Tipo: INTEGER Valor numérico do mês [1-12]
nm_mes	Tipo: VARCHAR Texto com o nome do mês (Ex.: Janeiro)
nr_ano_mes	Tipo: CHAR Valor de referência do ano/mês (Formato: YYYY/MM)

6.2.2. Tabela Dimensional: dim_aeroporto

A dim_aeroporto contém informações de cadastro dos aeroportos mundiais que tem alguma relação com os aeroportos brasileiros. Esta dimensão se relaciona com a tabela fato tanto pela ft_dadosanac.sk_aeroporto_destino quanto pela ft_dadosanac.sk_aeroporto_origem.

sk_aeroporto	Inteiro / Chave Primária
nk_aeroporto	Tipo: VARCHAR Sigla internacional dos aeroportos
nm_aeroporto	Tipo: VARCHAR Nome do aeroporto e/ou nome da cidade ou pais que está localizado.
uf	Tipo: VARCHAR Sigla da Unidade Federativa Brasileira. Campo preenchido apenas para os aeroportos brasileiros
regiao	Tipo: VARCHAR Região brasileira onde o aeroporto está localizado. Campo preenchido apenas para aeroportos brasileiros
pais	Tipo: VARCHAR Indica nome do país onde está localizado o aeroporto
nm_continente	Tipo: VARCHAR Indica nome do continente onde está localizado o aeroporto
latitude	Tipo: REAL Indica o valor na posição geográfica latitude do aeroporto
longitude	Tipo: REAL Indica o valor na posição geográfica longitude do aeroporto

6.2.3. Tabela Dimensional: dim_empresa

A dim_empresa contém informações de cadastro das empresas que operam nos aeroportos brasileiros e se relaciona com a fato pela chave sk_empresa.

sk_empresa	Inteiro / Chave Primária
nk_empresa	Tipo: VARCHAR Sigla da empresa de transporte aéreo.
nm_empresa	Tipo: VARCHAR Nome (razão social) da empresa de transporte aéreo.
nm_nacionalidade	Tipo: VARCHAR Nacionalidade da empresa aérea (ESTRANGEIRA / BRASILEIRA)

6.2.4. Tabela Fato: ft_ dadosanac

A ft_dadosanac é a tabela fato com as informações estatísticas de voos disponibilizadas pela ANAC mensalmente.

sk_tempo	Inteiro / Chave Primária Relação com a dim_tempo
sk_aeroporto_origem	Inteiro / Chave Primária Relação com a dim_aeroporto e indica o aeroporto de origem do voo
sk_aeroporto_destino	Inteiro / Chave Primária Relação com a dim_aeroporto e indica o aeroporto de destino do voo
sk_empresa	Inteiro / Chave Primária Relação com a dim_empresa
nm_natureza_voo	Tipo: INTEIRO Informa a natureza das etapas de voo: 0- DOMÉSTICA 1- ITERNACIONAL
nm_grupo_voo	Tipo: INTEIRO Informa o tipo de voo: 0- IMPRODUTIVO (não geram receita a empresa) 1- REGULAR (que geram receita e são realizados sob numeração HOTRAN) 2- NÃO REGULAR (que geram receita e NÃO são realizados sob numeração HOTRAN)
md_passageiros_pagos	Tipo: INTEIRO Informa número de passageiros que ocupam assentos comercializados e que geram receitas às empresas aéreas.
md_passageiros_pagos	Tipo: INTEIRO Informa número de passageiros que ocupam assentos comercializados e que não geram receitas às empresas aéreas.
md_carga_paga_kg	Tipo: INTEIRO Refere-se à quantidade total, expressa em quilogramas, de todos os bens que tenham sido transportados na aeronave, exceto correio e bagagem, e tenham gerado receitas direta ou indireta para a empresa aérea.
md_carga_gratis_kg	Tipo: INTEIRO Refere-se à quantidade total, expressa em quilogramas, de todos os bens que tenham sido transportados na aeronave, exceto correio e bagagem, e não tenham gerado receitas direta ou indireta para a empresa aérea.
md_ correio_kg,	Tipo: INTEIRO Refere-se à quantidade de objetos transportados de rede postal em cada trecho de voo realizado, expresso em quilogramas.
md_correio_km	Tipo: INTEIRO Refere-se ao volume de Correio (kg) em cada quilômetro, ou seja, a soma do produto entre a quantia (kg) de correio e a distâncias das etapas.
md_ask	Tipo: INTEIRO Refere-se ao volume de Assentos Quilômetros Oferecidos, ou seja, a soma do produto entre o número de assentos oferecido e a distância das etapas.
md_ rpk	Tipo: INTEIRO

	Refere-se ao volume de Passageiros Quilômetros Transportados, ou seja, a soma do produto entre o número de passageiros pagos e a distâncias das etapas.
md_atk	Tipo: INTEIRO Refere-se ao volume de Tonelada Quilômetro Oferecida, ou seja, a soma do produto entre o payload, que é a capacidade total de peso disponível na aeronave, expressa em quilogramas, disponível para efetuar o transporte de passageiros, carga e correio, e a distância das etapas, dividido por 1.000.
md_rtk	Tipo: INTEIRO Refere-se ao volume de Toneladas Quilômetros Transportadas, ou seja, a soma do produto entre os quilogramas carregados pagos, onde cada passageiro possui o peso estimado de 75 Kg, e a distância das etapas, dividido por 1.000.
md_combustivel	Tipo: INTEIRO Refere-se à quantidade, em litros, de combustível consumida pela aeronave na execução da referida etapa. Informação disponível apenas para empresas brasileiras.
md_distancia	Tipo: INTEIRO Refere-se à distância, expressa em quilômetros, entre os aeródromos de origem e destino da etapa, considerando a curvatura do planeta Terra.
md_decolagens	Tipo: INTEIRO Refere-se ao número de decolagens que ocorreram entre os aeródromos de origem e destino da etapa.
md_ carga_gratis_km	Tipo: INTEIRO Refere-se ao volume de Carga Grátis (kg) em cada quilômetro, ou seja, a soma do produto entre a quantia (kg) de carga grátis e a distâncias das etapas.
md_ carga_ paga_km	Tipo: INTEIRO Refere-se ao volume de Carga Paga (kg) em cada quilômetro, ou seja, a soma do produto entre a quantia (kg) de carga paga e a distâncias das etapas.
md_assentos	Tipo: INTEIRO É o número de assentos disponíveis em cada etapa de voo de acordo com a configuração da aeronave na execução da etapa; e
md_payload	Tipo: INTEIRO Informa a capacidade total de peso na aeronave, expressa em quilogramas, disponível para efetuar o transporte de passageiros, carga e correio.
md_horas_voadas	Tipo: INTEIRO Refere-se ao número de horas de voo entre os aeródromos de origem e destino da etapa.
md_bagagem_kg	Tipo: INTEIRO Refere-se à quantidade total de bagagem despachada, expressa em quilogramas.

7. Projeto de ETL

7.1. Descrição do Projeto de ETL

O processo de ETL foi desenvolvido em 4 etapas distintas:

- Popula dimensão tempo (carga única);
- Popula dimensão aeroporto (carga mensal);
- Popula dimensão empresa (carga mensal);
- Popula a fato (carga mensal);

Como a fonte de dados principal apresenta dados acumulativos e pelo fato de não existir, em nenhuma outra fonte, um cadastro prévio de empresa e aeroporto, estas duas tabelas dimensionais são populadas através dos dados transacionais disponíveis. Desta forma, antes de carregar a fato com os dados mais recentes, é necessário fazer o insert na dimensão dim_empresa e dim_aeroporto com os novos registros disponíveis no relatório (se existentes).

Já a dimensão tempo foi carregada uma única vez, com histórico de 2010 até 2032.

7.1.1. ETL Popula dimensão Empresa (popula_dim_empresa)

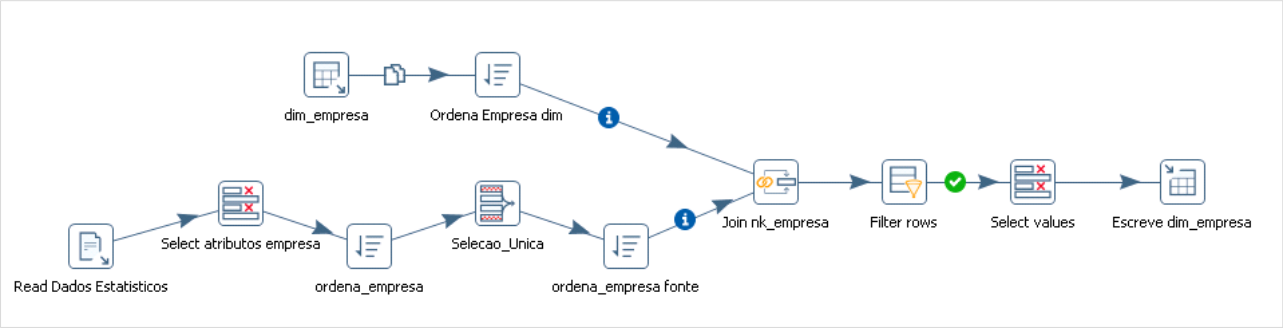

Figura 3 – Fluxo de carga inicial – dim_empresa

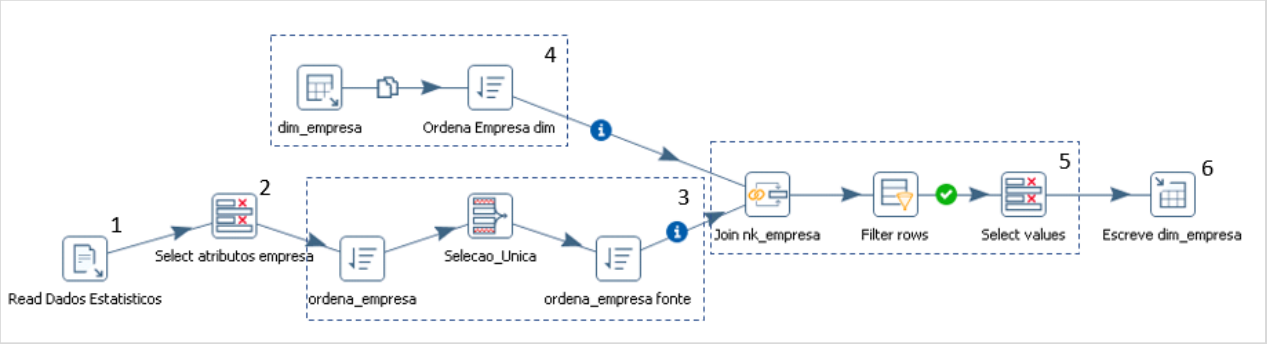

Figura 4 – Fluxo de carga incremental – dim_empresa

O fluxo de carga dos dados na dim_empresa segue as seguintes etapas:

1. Leitura da tabela de Dados Estatísticos disponíveis pela ANAC;
2. Seleção apenas do atributo empresa;
3. Identificação dos registros únicos de empresa;
4. Leitura e ordenação dos registros já presentes na dim_empresa;
5. Identificação dos registros faltantes na dim_empresa;
6. *Insert* das informações das empresas faltantes na dim_empresa.

7.1.2. ETL Popula dimensão Aeroporto (popula_dim_aeroporto)

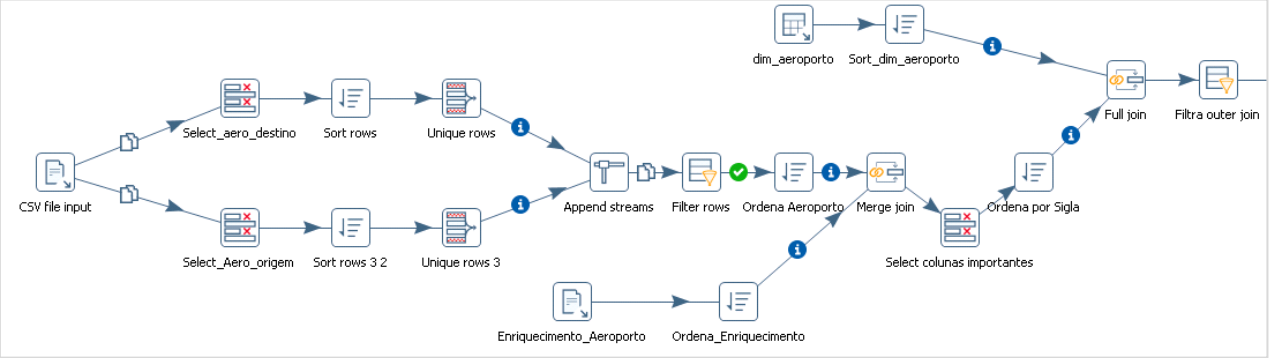

Figura 5 – Fluxo de carga inicial – dim_aeroporto

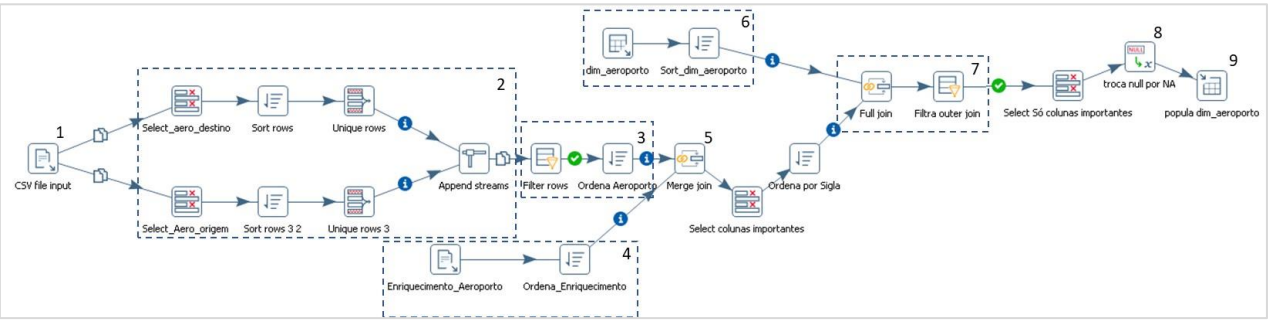

Figura 6 – Fluxo de carga incremental – dim_aeroporto

O fluxo de carga dos dados na dim_aeroporto segue as seguintes etapas:

1. Leitura da tabela de Dados Estatísticos disponíveis pela ANAC;
2. Seleção apenas dos atributos aeroporto origem e destino e união dos 2 atributos distintos em uma lista única;
3. Identificação dos registros únicos de aeroporto (independentemente de ser origem ou destino);
4. Leitura da fonte de enriquecimento das informações de aeroporto que contém as coordenadas geográficas;
5. Merge da informação de aeroporto com o enriquecimento de latitude e longitude;
6. Leitura e ordenação dos registros já presentes na dim_aeroporto;
7. Identificação dos registros faltantes na dim_aeroporto;

8. Tratamento dos campos nulos de registro de UF e Município trocando nulo por NA. (os registros de aeroporto internacionais não apresentam estes campos preenchidos na tabela original);
9. *Insert* dos cadastros de aeroportos faltantes na dim_aeroporto.

7.1.3. ETL Popula dimensão fato (popula_fato)

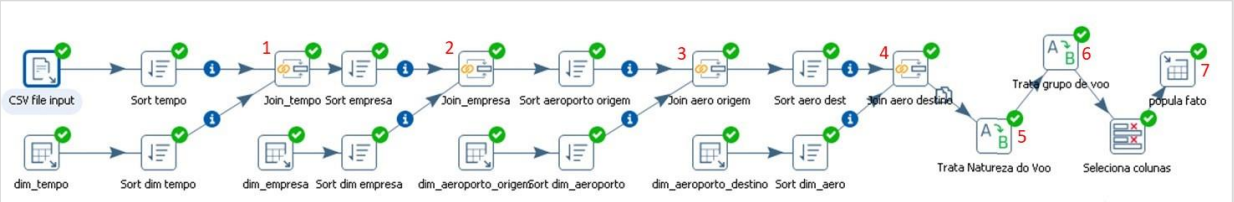

Figura 7 – Fluxo de carga – ft_dadosanac

Após executar a carga dos novos dados nas dimensões dim_empresa e dim_aeroporto, é necessário popular a tabela fato. Seguem as etapas:

1. Leitura da tabela de estatística e join com a dim_tempo para identificar a sk_tempo (o join é feito pela chave Ano/Mês);
2. Join com a dimensão dim_empresa para identificar a sk_empresa (join feito pela sigla da Empresa);
3. Join com a dimensão dim_aeroporto para identificar a sk_aeroporto_origem (join feito pela sigla do Aeroporto);
4. Join com a dimensão dim_aeroporto para identificar a sk_aeroporto_destino (join feito pela sigla do Aeroporto);
5. Tratamento da dimensão degenerada Natureza de Voo;
6. Tratamento da dimensão degenerada Grupo de Voo;
7. Carga na tabela fato das medidas e suas respectivas chaves.

8. Dashboard

8.1. Descrição da Elaboração

Após toda etapa de ETL foi possível carregar o Data Warehouse para a ferramenta de confecção do Dashboard, que para esse trabalho foi o Power BI.

Como o objeto da análise são os voos, desde o início do projeto tinha-se o desejo de utilizar mapas para descrever as rotas envolvidas. A tabela original continha as coordenadas geográficas dos aeroportos de origem e destino, porém no processo de elaboração do *DW* percebeu-se que não havia necessidade de se criar duas tabelas com os dados dos aeroportos – origem e destino, pois geraria uma redundância de informações desnecessária. Contudo, no *Power BI*, o desenho das rotas demandava a diferenciação das localizações de origem e

destino, por esse motivo a dimensão aeroporto teve de ser trazida duas vezes, possibilitando a diferenciação entre os pontos de partida e chegada.

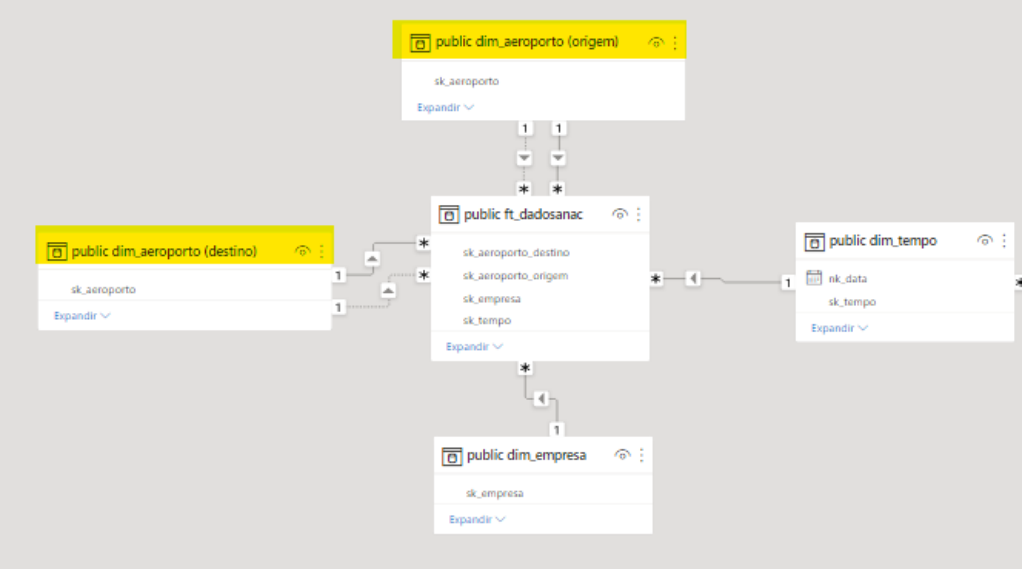

Figura 8 – Relacionamento das tabelas no Power BI

8.2. Telas do Dashboard

As visões elaboradas dividem-se em dois grupos:

- **Mapas**: Composto de três telas que permitem ao usuário a visualização das informações das rotas agregadas por cidade ou país;

- **Estatísticas**: São três telas que apresentam diversas estatísticas, em formato gráfico, como a quantidade de aeroportos envolvidos, evolução de passageiros, cargas – permite inclusive avaliação do comportamento pré e pós início da pandemia de COVID-19.

- **Página 1 – Passageiros por Cidade**:
  - **Descrição**: Permite ao cliente visualizar no mapa as cidades de origens e destino dos voos realizados nos intervalos de tempo desejados;
  - **Principais *Features*:** Visualização, na rota do mapa, da sumarização dos passageiros pagos, filtragem por tipo de voo, filtro das rotas por empresa, avaliação da participação de cada empresa no total de passageiros.

Figura 9 – Dashboard 1

- **Página 2 – Decolagem por País**:
  - **Descrição**: Visualização, em mapa, das estatísticas de decolagem na rota entre o país de origem e destino;
  - **Principais *Features*:** Visualização das rotas (bem definidas) no mapa, da sumarização da quantidade de decolagem e filtro das rotas por empresa.

Figura 10 – Dashboard 2

- **Página 3 – Decolagens por Cidade**:

    - **Descrição**: Permite ao cliente visualizar no mapa as cidades de origens e destino dos voos realizados nos intervalos de tempo desejados;

    - **Principais *Features*:** Visualização, na rota do mapa, da sumarização da quantidade de decolagem e filtro das rotas por empresa.

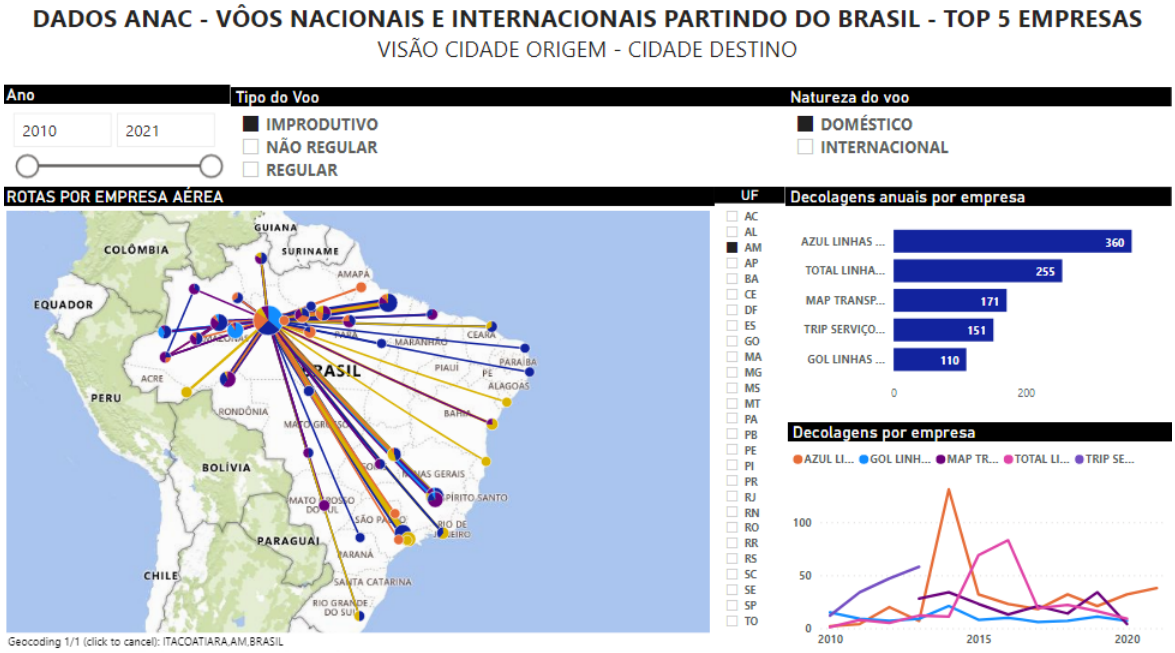

Figura 11 – Dashboard 3

- **Página 4 – Aeroportos**:

    - **Descrição**: Apresenta estatísticas das quantidades de aeroportos, de cada país envolvidos nas rotas analisadas;

    - **Principais *Features*:** Gráfico de blocos com as quantidades de aeroportos em cada país, filtragem por continente e quantitativo de decolagens por aeroporto. O eixo RJ-SP se destaca do restante do país em quantidade de decolagens e que 86% da nossa base são relacionadas a voos domésticos.

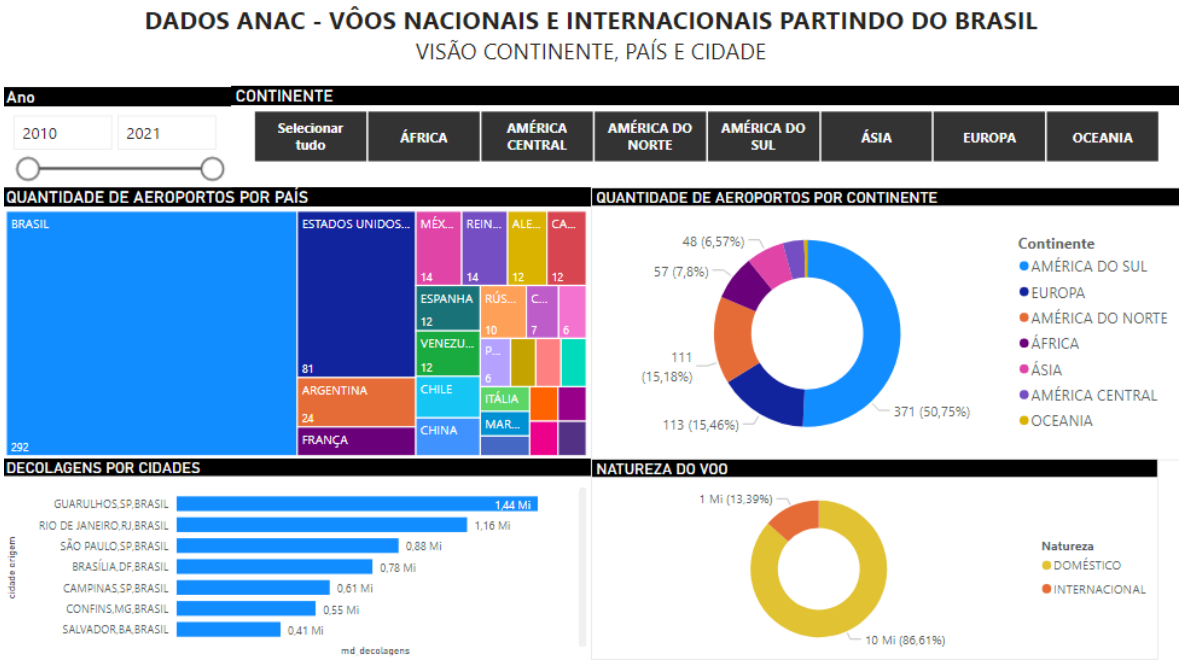

Figura 12 – Dashboard 4

- **Página 5 – Evolução Temporal das Medidas**:

  - **Descrição**: Possibilita a visualização temporal de diversas medidas possibilitando verificar o grande impacto gerado pela pandemia de COVID 19 no setor aéreo. É possível identificar claramente as duas ondas do Covid-19, sendo a primeira delas a mais impactante.

  - **Principais *Features*:** Filtragem por período de tempo e por classificação pré e pós pandemia.

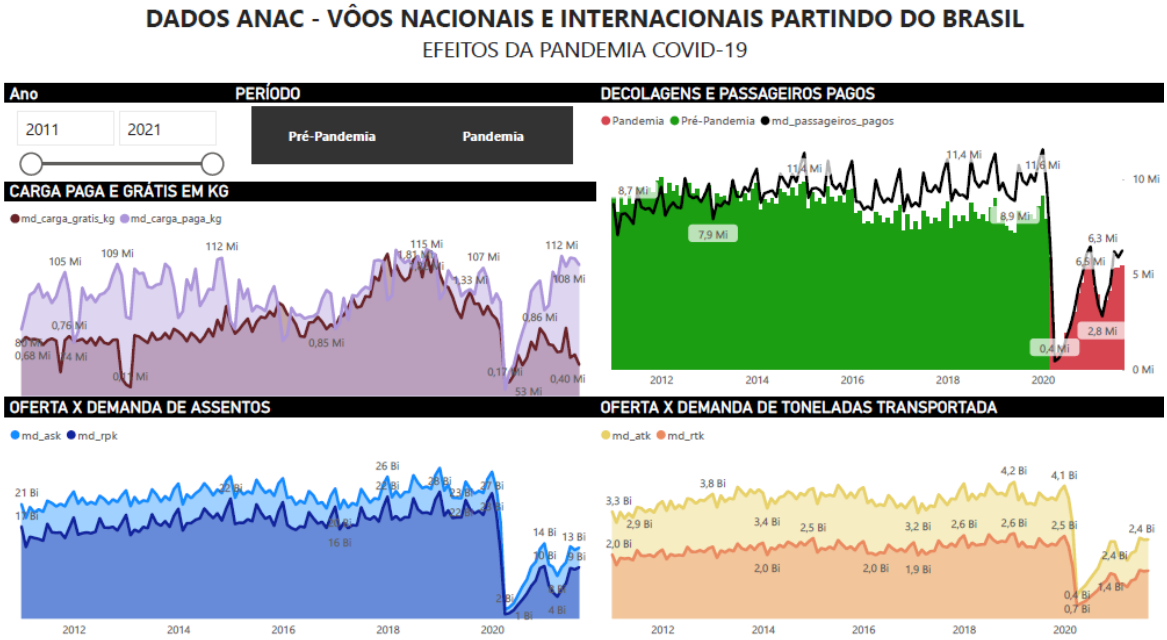

Figura 13 – Dashboard 5

- **Página 6 – Impacto COVID-19 nas estatísticas das Empresas**:
    - **Descrição**: Apresenta resumos de medições sumarizados pelas empresas envolvidas. A empresa Azul se destaca por ser a mais próxima em alcançar em 2021 o patamar de decolagens de 2019. Ela também é a que mais possui horas de voos improdutivos, como por exemplo, treinamentos e manutenções.

    - **Principais *Features*:** Filtragem por período de tempo, empresa e período pré ou pós pandemia;

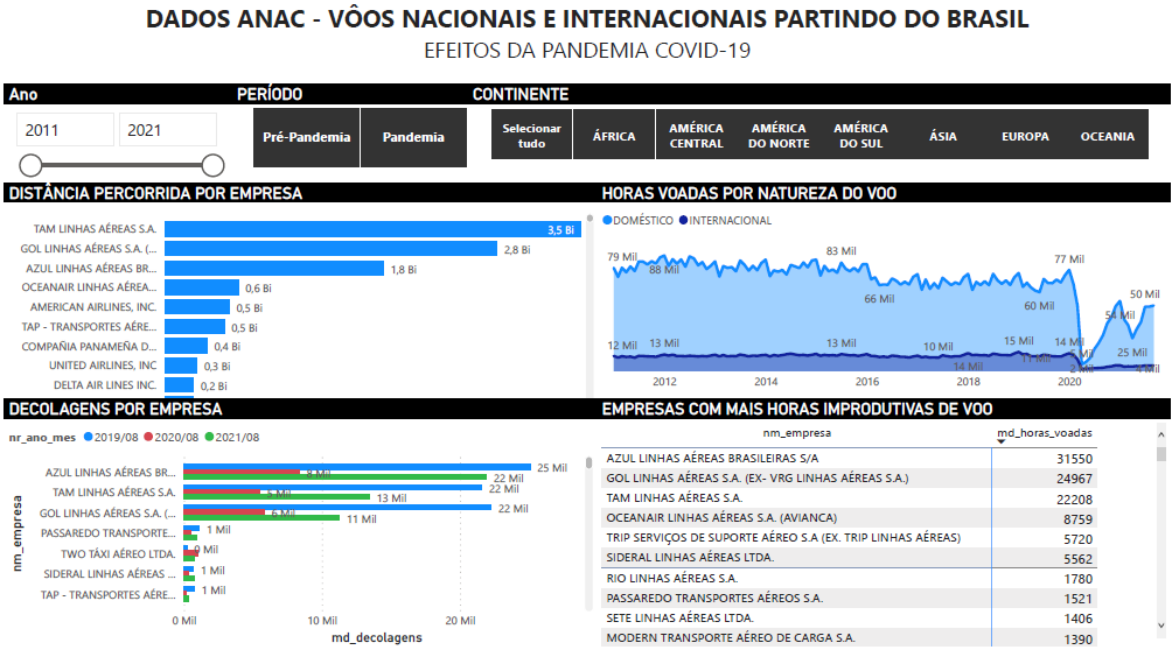

Figura 14 – Dashboard 6

9. Conclusão

O projeto alcançou o resultado desejado, ou seja, foi possível elaborar um fluxo ETL de alimentação de um D*ata Warehouse* que por sua vez é base para atualização de *Dashboards* para análise dos dados originalmente fornecidos em formato E*xcel* pela Anac.

Os relatórios produzidos em formato de mapas permitiram uma visão simples, porém, super detalhada da evolução temporal das rotas de voos. Já os gráficos baseados em séries temporais demonstraram, a partir de março de 2020, o forte impacto da pandemia de COVID-19 nas medidas disponibilizadas e, consequentemente, as enormes perdas sofridas pelo setor aéreo. Foi possível identificar também uma leve recuperação do setor no segundo e terceiro trimestre de 2021, mas ainda assim abaixo da movimentação em 2019.

10. Anexos

Todos os anexos serão disponibilizados via GitHub no caminho:

https://github.com/anarucos/bi_master

11. Arquivos

Os arquivos referentes ao projeto estão na pasta no GitHub:

- Documento Final:
  https://github.com/anarucos/bi_master
    - Projeto_Anac.pdf
- Código:
  https://github.com/anarucos/bi_master/tree/main/codigo
    - criação do dw.sql
    - popula_dim_aeroporto.ktr
    - popula_dim_empresa.ktr
    - popula_dim_tempo.txt
    - popula_fato.ktr
    - PowerArchitect - Projeto Anac_v1.architect
- Fonte de dados:
  https://github.com/anarucos/bi_master/tree/main/fonte_de_dados
    - Dados Estatísticos.zip
    - Enriquecimento_Aeroporto.csv
- Visualização:
  https://github.com/anarucos/bi_master/tree/main/visualizacao
    - Projeto Anac.pbix