



Universidade do Minho
Licenciatura em Engenharia Informática

Perfil de Especialização em Sistemas de Armazéns de Dados

Ano Letivo de 2024/2025

Data Warehouse para Análise e Mitigação de Atrasos em Voos Comerciais

Alexandre Eduardo Viera Martins, pg53604

Bruno Afonso, pg55925

Inês Meneses de Castro, pg57550

Norberto Miguel Luzes Pais Pinto, pg55907

Ricardo Miguel Queirós de Jesus, pg57898

Abril 2025

Data de Recepção	
Responsável	
Avaliação	
Observações	

Data Warehouse para Análise e Mitigação de Atrasos em Voos Comerciais

Alexandre Eduardo Viera Martins, pg53604

Bruno Afonso, pg55925

Inês Meneses de Castro, pg57550

Norberto Miguel Luzes Pais Pinto, pg55907

Ricardo Miguel Queirós de Jesus, pg57898

Abril 2025

<< Dedicatória opcional>>

Resumo

O presente trabalho aborda o desenvolvimento de um sistema de suporte à tomada de decisão para análise e previsão de atrasos em voos, com o objetivo de melhorar a eficiência operacional no setor da aviação. Diante dos desafios inerentes à gestão aeroportuária e à operação das companhias aéreas, especialmente devido à variabilidade das condições meteorológicas, volume de voos e fatores operacionais, foi proposto um sistema integrado que permite identificar padrões históricos, prever atrasos e recomendar ações para mitigar os seus impactos.

Para tal, foi adotada uma abordagem baseada num Data Warehouse estruturado segundo um modelo dimensional, onde foram integrados dados operacionais, meteorológicos e informações detalhadas sobre aeroportos, companhias aéreas, aeronaves e rotas.

A análise e segmentação dos aeroportos permitiu identificar diferentes perfis operacionais, possibilitando uma melhor compreensão dos fatores que influenciam os atrasos. Com base nessa segmentação, foi desenvolvido um sistema OLAP capaz de fornecer recomendações úteis para a gestão aeroportuária, visando otimizar a operação e reduzir os atrasos.

O sistema proposto integra dados diversos e utiliza modelos avançados para apoiar decisões estratégicas e operacionais. Os resultados obtidos demonstram o potencial da solução para contribuir na melhoria da eficiência do setor em estudo, oferecendo informações valiosas que podem ser aplicadas na antecipação e mitigação de atrasos, beneficiando tanto as entidades empresariais como os passageiros.

Área de Aplicação: Sistemas de Apoio à Tomada Decisão e Data Warehousing para Análise Operacional no Setor Aeroportuário

Palavras-Chave: <<Conjunto de palavras-chave que permitirão referenciar domínios de conhecimento, tecnologias, estratégias, etc., directa ou indirectamente referidos no relatório. Por exemplo: Bases de Dados Relacionais, Gestão de Índices, JAVA, Protocolos de Comunicação.>>

Índice Geral

1. Definição do Sistema	1
1.1 Contexto da Aplicação	1
1.2 Motivação e Objetivos do Trabalho	2
1.3 Análise da Viabilidade do Processo	3
1.4 Recursos e Equipa de Trabalho	4
1.5 Plano de Execução do Projeto	4
2. Levantamento e Análise de Requisitos	5
2.1 Método Adotado	5
2.2 Organização dos Requisitos Levantados	5
2.2.1. Classificação dos Requisitos	5
2.2.2.1 Requisitos de Descrição	5
2.2.2.2 Requisitos de Exploração	7
2.2.2.3 Requisitos de Controlo	9
2.3 Análise e Validação Geral dos Requisitos	10
3. Modelação Dimensional de Dados	12
3.1 Apresentação da Abordagem Realizada	12
3.2 A Matriz de Decisão	13
3.3 Definição e Caracterização de Dimensões	14
3.4 Definição e Caracterização da Tabela de Factos e Respetivo Grão	16
3.4.1 Tabela de Factos	16
3.4.2 Grão	16
3.5 Configuração dos Esquemas	17
3.6 As Vistas dos Agentes de Decisão	18
4. Arquitetura Geral do Sistema	19
4.1 Apresentação Geral	19
4.2 Fontes de Dados	19
4.2.1 Dataset de Companhias Aéreas	20
4.2.2 Dataset com Rotas de Voo	20
4.2.3 Dataset de Aeronaves	22
4.2.4 Dataset com Condições Climatéricas	24
4.2.5 Dataset de Aeroportos	24
4.3 Área de Preparação de dados	25
4.3.1 Companhia Aérea	25
4.3.2 Aeronaves	26
4.3.3 Aeroporto	27
4.3.4 Voo	28
4.3.4 Condições Climatéricas	31
4.4 O Armazém de Dados	32
4.5 Exploração e Visualização de Dados	34
4.6 Aquisição de Conhecimento	34
5. O Povoamento de Dados	35
5.1 Apresentação da Abordagem Realizada	35

5.2 Mapeamento de Dados	36
5.2.1 Companhia Aerea	37
5.2.2 Aeronave	37
5.2.3 Aeroporto	37
5.2.4 Voo	38
5.2.5 Clima	39
5.3 Modelação do Sistema de Povoamento	39
5.4 Implementação do Sistema de Povoamento	39
5.5 Validação e Testes	40
6. Exploração e Análise de Dados	41
6.1 Organização geral do sistema de dashboarding	41
6.2 Serviços de exploração e análise implementados	42
7. Caracterização de Perfis de Clientes	43
7.1 Definição do problema e compreensão dos elementos de análise envolvidos	43
7.2 Seleção e preparação dos dados	43
7.3 Identificação e fundamentação da técnica de análise	43
7.4 Construção do modelo do modelo de análise	43
7.5 Validação do desempenho do modelo	43
7.6 Avaliação dos resultados	43
8. Personalização de Ofertas de Produtos e Serviços	44
8.1 Definição do problema e compreensão dos elementos de análise envolvidos	44
8.2 Seleção e preparação dos dados	44
8.3 Identificação e fundamentação da técnica de análise	44
8.4 Construção do modelo do modelo de análise	44
8.5 Validação do desempenho do modelo	44
8.6 Avaliação dos resultados	44
9. Conclusões e Trabalho Futuro	45
9.1 Conclusões	45
9.2 Trabalho Futuro	45
9. Bibliografia	46
Referências	47
Anexos	49

Índice de Figuras

Figura 1 - Modelo Dimensional.

3

Índice de Tabelas

Tabela 1 - Ilustração de inserção de uma tabela e sua legenda.

3

1. Definição do Sistema

1.1 Contexto da Aplicação

Atualmente, o setor de aviação enfrenta diversos desafios relacionados com a gestão eficiente das operações, sendo os **atrasos em voos** um dos fatores mais críticos. A variabilidade das condições climáticas, a gestão da frota, a infraestrutura aeroportuária e a coordenação entre diferentes entidades operacionais afetam diariamente a pontualidade dos voos e, consequentemente, a eficiência do setor. Neste contexto, a informação recolhida ao longo do tempo acerca dos voos, rotas e condições meteorológicas representa uma fonte extremamente valiosa para suportar tomadas de decisão informadas.

O aumento da concorrência entre companhias aéreas e a exigência por parte dos passageiros de um serviço cada vez mais fiável e previsível torna essencial o desenvolvimento de sistemas inteligentes de apoio à decisão. Neste contexto da aviação comercial, temos como foco a **análise e previsão de atrasos em voos**, com base em dados históricos e fatores externos como clima e desempenho aeroportuário.

Deste modo a Organização da Aviação Civil Internacional, cujo objetivo é favorecer a segurança, a economia, a eficiência e o desenvolvimento dos serviços aéreos quis implementar um sistema de datawarehousing para armazenar dados sobre os voos e os seus atrasos de modo a tomar decisões informadas que favoreçam não só as companhias aéreas mas também os aeroportos e consequentemente a experiência dos seus clientes.

A nossa proposta de sistema visa integrar diferentes fontes de dados operacionais com o objetivo de **compreender de forma aprofundada os padrões de atraso** por companhias aéreas, rotas, aeroportos e outras variáveis relevantes. A aplicação prática deste projeto poderá servir de base para a integração em plataformas internas das companhias aéreas (ou serviços associados), permitindo **atuar proativamente na mitigação dos atrasos** e no reforço da eficiência operacional. O sistema poderá, por exemplo, **prever atrasos em determinadas rotas** e recomendar ações específicas para minimizar o seu impacto, contribuindo para a melhoria do desempenho global da operação aérea.

1.2 Motivação e Objetivos do Trabalho

A motivação para o desenvolvimento deste sistema assenta na crescente necessidade de **aumentar a eficiência operacional no setor da aviação**. A aviação comercial enfrenta desafios constantes relacionados com a **otimização de recursos, satisfação dos passageiros, conformidade regulatória e sustentabilidade ambiental**. As companhias aéreas operam num mercado altamente competitivo, onde a **pontualidade** é um fator decisivo tanto em termos de desempenho como de reputação.

Os atrasos têm impactos significativos que se refletem em várias frentes:

- **Custos operacionais elevados**, incluindo tempo de tripulação, uso de pista e consumo adicional de combustível;
- **Perda de receitas indiretas**, como conexões perdidas e compensações a passageiros;
- **Impacto reputacional**, que pode afetar a preferência dos clientes por determinadas companhias;
- **Desorganização logística**, com efeitos em cascata ao longo do dia em operações interligadas.

Estes atrasos resultam frequentemente de uma combinação complexa de fatores – desde condições meteorológicas adversas até limitações nos aeroportos, passando por questões de tráfego aéreo, manutenção ou gestão interna das companhias. A antecipação e mitigação destes atrasos, portanto, exige **abordagens inteligentes baseadas em dados**, com capacidade de análise histórica, previsão de cenários e apoio à decisão em tempo real.

A proposta deste trabalho é desenvolver um sistema que use **a análise de dados** para identificar padrões de atraso, prever ocorrências futuras e fornecer recomendações operacionais concretas. Com isso, pretende-se **transformar dados operacionais em conhecimento acionável** para melhorar a pontualidade, a eficiência e a experiência do passageiro.

Através deste sistema, poderemos:

- **Compreender os padrões de atraso** com base em rotas, horários, companhias e condições operacionais;
- **Prever atrasos** com base em dados históricos e variáveis externas, como condições meteorológicas e eficiência aeroportuária;
- **Apoiar decisões operacionais e estratégicas**, como a reconfiguração de horários, alocação de frota ou ajustes na escala de voos;
- **Mitigar o impacto dos atrasos**, através de recomendações fundamentadas e adaptadas ao contexto.

Objetivos específicos:

- Recolher e integrar dados relevantes (voos, clima, companhias, horários, rotas);
- Identificar e analisar padrões históricos de atrasos;
- Desenvolver modelos preditivos para estimar a probabilidade e duração de atrasos;
- Fornecer recomendações operacionais baseadas em evidência;
- Criar dashboards de **visualização** para apoio à decisão.

1.3 Análise da Viabilidade do Processo

A aplicação prática deste caso de estudo revela-se particularmente relevante, pois permite promover a eficiência operacional das companhias aéreas, através da **antecipação e mitigação de atrasos de voos**. O sistema proposto contribui para uma **gestão mais precisa e informada**, permitindo atuar de forma proativa perante eventos que afetam a pontualidade dos voos.

Este caso de estudo é viável porque possibilita identificar padrões históricos de atraso, compreender as suas causas e recomendar medidas corretivas, como a reconfiguração de horários, ajustes na escala de voos ou alterações na alocação de aeronaves. Estas ações podem traduzir-se numa operação mais eficiente, menor acumulação de atrasos em cadeia e melhor utilização dos recursos.

A viabilidade do projeto assenta, portanto, em vários fatores, tais como:

- Disponibilidade de dados: existem conjuntos de dados públicos e abertos (como o OpenFlight e o Kaggle) que contêm registo detalhados de voos, atrasos, rotas e condições meteorológicas. Estes dados permitem uma base sólida para análises e previsões.
- Tecnologias disponíveis: a arquitetura do sistema pode ser construída com ferramentas amplamente utilizadas e acessíveis, como Python para análise e modelação de dados, SQL para gestão de armazéns de dados, e Apache NiFi para ingestão e processamento de dados..
- Relevância e aplicabilidade do domínio: o problema dos atrasos é recorrente e transversal a todas as companhias aéreas. Um sistema com capacidade de prever e mitigar esses efeitos têm elevada aplicabilidade prática e grande potencial de integração com sistemas internos de operação e planeamento.
- Escalabilidade: o sistema poderá ser evoluído de forma progressiva, permitindo a incorporação de múltiplas companhias aéreas, rotas e aeroportos.

Exemplo concreto de viabilidade:

Suponhamos que, através do sistema, uma companhia aérea como a **TAP** identifica que uma determinada rota (**Lisboa–Caraíbas**) apresenta **atrasos regulares superiores a 20 minutos**, especialmente em dias com **baixa visibilidade e vento no aeroporto de origem**. Com base nas análises realizadas, o sistema pode **recomendar a reconfiguração do horário de partida, o reforço dos tempos**

de escala ou até o redirecionamento pontual de recursos. Estas medidas permitem reduzir os impactos negativos, melhorar a fiabilidade da operação e otimizar o desempenho da companhia, identificando também quais são as rotas e companhias com maior propensão para atrasos.

1.4 Recursos e Equipa de Trabalho

No que toca aos Recursos Humanos constituintes deste trabalho, a equipa desenvolvedora do projeto consiste em 5 estudantes do perfil de Sistemas de Data Warehouse apresentados no início do relatório.

Para o levantamento, análise e modelação dos processos, utilizamos diversas tecnologias especializadas que facilitaram cada etapa do projeto: **Indyco** e **Draw.io** foi utilizado para a modelação dimensional e lógica dos dados, **Bizagi Modeler** permitiu o mapeamento estruturado dos fluxos dos processos através de diagramas BPMN. O **Apache NiFi** e o **MySQL** foi responsável pelo povoamento do data warehouse e pelo tratamento inicial através da implementação dos fluxos de extração, transformação e carga (ETL) dos dados. Finalmente, para a análise e aquisição de conhecimento, recorremos ao **Power BI** para o desenvolvimento de dashboards interativas e ao **Jupyter Notebooks** para a execução de análises estatísticas e aquisição de conhecimento.

Para a construção do data warehouse propriamente dito, utilizámos uma base de dados relacional SQL, onde foram implementados os esquemas dimensionais e tabelas de fatos definidos na modelação dimensional.

1.5 Plano de Execução do Projeto

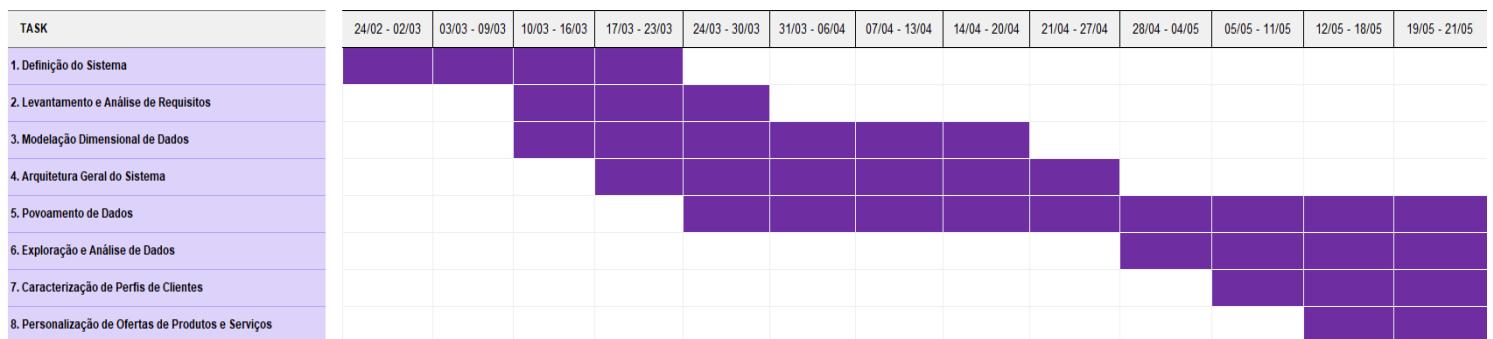


Figura 1 - Diagrama de Gantt

2. Levantamento e Análise de Requisitos

2.1 Método Adotado

Durante a fase de levantamento de requisitos para a criação do sistema de Data Warehousing, a equipa de trabalho recorreu a diferentes métodos para garantir que o projeto respondesse às expectativas dos clientes. Entre as abordagens utilizadas, destacam-se **entrevistas com os stakeholders**, que permitiram recolher dados sobre as companhias aéreas. A equipa também organizou sessões de **brainstorming** para explorar potenciais informações e métodos mais apropriados para este tipo de problema, apoiado pela experiência e conhecimento da equipa responsável. **Questionários e Pesquisas** foram outro método utilizado para saber o nível de satisfação dos clientes das companhias aéreas e as suas reclamações. A **análise de documentos e relatórios existentes** foi um dos principais métodos adotados pela equipa, com o intuito de verificar que tipo de dados existentes por parte dos clientes são essenciais e aqueles que foram precisos recolher posteriormente.

2.2 Organização dos Requisitos Levantados

2.2.1. Classificação dos Requisitos

Os requisitos levantados foram organizados nas seguintes categorias:

- **Requisitos de Descrição;**
- **Requisitos de Exploração** - refere-se à utilização prática e à viabilidade dos requisitos para orientar o desenvolvimento do sistema;
- **Requisitos de Controlo** - mecanismos para garantir a gestão adequada dos requisitos durante todo o ciclo de vida do projeto.

2.2.2.1 Requisitos de Descrição

Requisito #: 1 Tipo: Descrição

Descrição: **Registo de horários programados e reais de partida/chegada.**

Rationale: Calcular atrasos.

Origem: Entrevistas com analistas

Prioridade: **Must**

Data: 08/04/2025

Requisito #: 2 Tipo: Descrição

Descrição: **Informações da companhia aérea.**

Rationale: Obter a identificação detalhada da companhia aérea.

Origem: Análise de Documentos

Prioridade: **Must**

Data: 08/04/2025

Requisito #: 3 Tipo: Descrição

Descrição: **Informações das rotas.**

Rationale: Obter a informação detalhada sobre as rotas de cada avião (origem, destino, distância,etc).

Origem: Análise de Documentos

Prioridade: **Must**

Data: 08/04/2025

Requisito #: 4 Tipo: Descrição

Descrição: **Condições climatéricas.**

Rationale: Obter dados meteorológicos no momento do voo.

Origem: Análise de Documentos

Prioridade: **Must**

Data: 08/04/2025

Requisito #: 5 Tipo: Descrição

Descrição: **Identificação do aeroporto.**

Rationale: Obter dados sobre os aeroportos que embarcam e desembarcam.

Origem: Análise de Documentos

Prioridade: **Must**

Data: 08/04/2025

Requisito #: 6 Tipo: Descrição

Descrição: **Feedback dos passageiros.**

Rationale: O nível de satisfação, reclamações e experiências dos passageiros com as companhias aéreas.

Origem: Questionários e Pesquisas

Prioridade: **Must**

Data: 08/04/2025

Requisito #: 9	Tipo: Descrição
Descrição: Incluir a causa oficial do atraso.	
<i>Rationale:</i> Permite distinguir entre atrasos meteorológicos, técnicos, operacionais...	
Origem: Relatórios de operações aéreas	
Prioridade: <i>Should</i>	
Data: 14/04/2025	
<hr/>	
Requisito #: 10	Tipo: Descrição
Descrição: Dados sobre tripulação (opcional, anonimizado).	
<i>Rationale:</i> Pode ajudar a identificar padrões de atraso relacionados com recursos humanos.	
Origem:	
Prioridade: <i>Could</i>	
Data: 14/04/2025	
<hr/>	
<h3>2.2.2.2 Requisitos de Exploração</h3>	
Requisito #: 11	Tipo: Exploração
Descrição: Gerar relatório de atrasos por companhias aéreas.	
<i>Rationale:</i> Avaliar a pontualidade de cada companhia	
Origem: Entrevistas com analistas	
Prioridade: <i>Must</i>	
Data: 08/04/2025	
<hr/>	
Requisito #: 12	Tipo: Exploração
Descrição: Analisar Aeroportos com mais atrasos.	
<i>Rationale:</i> Identificar aeroportos problemáticos e oportunidades de melhoria.	
Origem: Brainstorming	
Prioridade: <i>Must</i>	
Data: 08/04/2025	
<hr/>	
Requisito #: 13	Tipo: Exploração
Descrição: Avaliar desempenho por modelo de avião.	
<i>Rationale:</i> Detectar padrões de falha ou ineficiência técnica	
Origem: Entrevistas com manutenção	
Prioridade: <i>Should</i>	
Data: 08/04/2025	

Requisito #: 14 Tipo: **Exploração**

Descrição: **Analisar eficiência de aeroportos.**

Rationale: Identificar gargalos operacionais externos à companhia

Origem: Relatórios existentes

Prioridade: **Must**

Data: 08/04/2025

Requisito #: 15 Tipo: **Exploração**

Descrição: **Visualizar picos de atrasos ao longo do ano.**

Rationale: Otimizar escalas e logística em períodos críticos

Origem: Questionários e análise de dados sazonais

Prioridade: **Must**

Data: 08/04/2025

Requisito #: 16 Tipo: **Exploração**

Descrição: **Filtrar análises por horário do dia (manhã, tarde, noite, madrugada)**

Rationale: Atrasos podem variar ao longo do dia.

Origem: Entrevistas com analistas operacionais

Prioridade: **Should**

Data: 08/04/2025

Requisito #: 17 Tipo: **Exploração**

Descrição: **Gerar alertas automáticos para rotas com alto índice de atraso.**

Rationale: Permite ação proativa da equipa ou ajustar horários de partidas/chegadas.

Origem: Reuniões com operações

Prioridade: **Should**

Data: 08/04/2025

Requisito #: 18 Tipo: **Exploração**

Descrição: **Gerar painel comparativo entre duas ou mais companhias aéreas**

Rationale: Comparação direta de desempenho.

Origem: Brainstorming

Prioridade: **Must**

Data: 08/04/2025

Requisito #: 19 Tipo: **Exploração**

Descrição: **Analisar atrasos por tipo de atraso.**

Rationale: Aprofundar a análise causal e otimizar processos específicos.

Origem: Análise de indicadores de qualidade aeroportuária

Prioridade: **Should**

Data: 14/04/2025

2.2.2.3 Requisitos de Controlo

Requisito #: 20 Tipo: Controlo

Descrição: **Atualização diária dos dados**

Rationale: Manter os relatórios atualizados

Origem:

Prioridade: **Must**

Data: 08/04/2025

Requisito #: 21 Tipo: Controlo

Descrição: **Disponibilidade**

Rationale: DW deve estar acessível 24/7 para consultas e relatórios

Origem:

Prioridade: **Must**

Data: 08/04/2025

Requisito #: 22 Tipo: Controlo

Descrição: **Resposta rápida em análises complexas**

Rationale: Tempo de resposta inferior a 5s em consultas padrão

Origem:

Prioridade: **Must**

Data: 08/04/2025

Requisito #: 23 Tipo: Controlo

Descrição: **Segurança de dados pessoais**

Rationale: Garantir a segurança dos dados de passageiros

Origem:

Prioridade: **Must**

Data: 08/04/2025

Requisito #: 24 Tipo: Controlo

Descrição: **Escalabilidade**

Rationale: Capacidade de crescer conforme aumenta o volume de dados históricos

Origem:

Prioridade: **Should**

Data: 08/04/2025

Requisito #: 25

Tipo: Controlo

Descrição: **Integração com sistemas externos**

Rationale: Garantir atualizações em tempo real e dados oficiais

Origem:

Prioridade: *Could*

Data: 14/04/2025

Requisito #: 14

Tipo: **Exploração**

Descrição: **Prever atrasos com base no aeroporto, data e clima.**

Rationale: Permitir que o cliente planeje melhor a viagem

Origem: Brainstorming

Prioridade: *Must*

Data: 14/04/2025

2.3 Análise e Validação Geral dos Requisitos

Após o levantamento dos requisitos, a equipa procedeu à análise e validação para garantir que temos os dados necessários documentados para apresentar aos agentes de decisão. Esta etapa envolveu, nomeadamente, revisões internas e sessões de validação com representantes das companhias aéreas e da equipa de trabalho.

A equipa considera que após esta revisão, os requisitos contemplam todas as áreas críticas do sistema, incluindo dados operacionais, desempenho por companhias aéreas, fatores climáticos, rotas, aeroportos, e satisfação dos clientes. Foi também verificado que não existem conflitos entre os requisitos funcionais e não funcionais e que utilizam vocabulário padronizado. Concluímos que com base na análise da equipa técnica, todos os requisitos são tecnicamente viáveis com os recursos e tecnologias disponíveis.

A equipa promoveu sessões de validação com os interessados e representantes no desenvolvimento deste projeto das seguintes áreas:

- **Operações de voo de diferentes companhias aéreas**
- **Gestão de aeroportos**
- **Serviço ao cliente**
- **Tecnologia da Informação (TI)**
- **Executivos das companhias aéreas**

Durante essas sessões, os requisitos foram apresentados e discutidos, e ajustes foram feitos com base nas sugestões recebidas. Os requisitos foram considerados **adequados e suficientes** para suportar os objetivos do sistema de Data Warehousing proposto.

3. Modelação Dimensional de Dados

3.1 Apresentação da Abordagem Realizada

Para a construção do nosso sistema de apoio à decisão baseado em dados de voos e companhias aéreas, optámos pela modelação dimensional, seguindo a abordagem clássica proposta por **Ralph Kimball**, focada na organização dos dados em torno de fatos e dimensões. Este método foi escolhido pela sua simplicidade, flexibilidade analítica e boa performance em consultas analíticas.

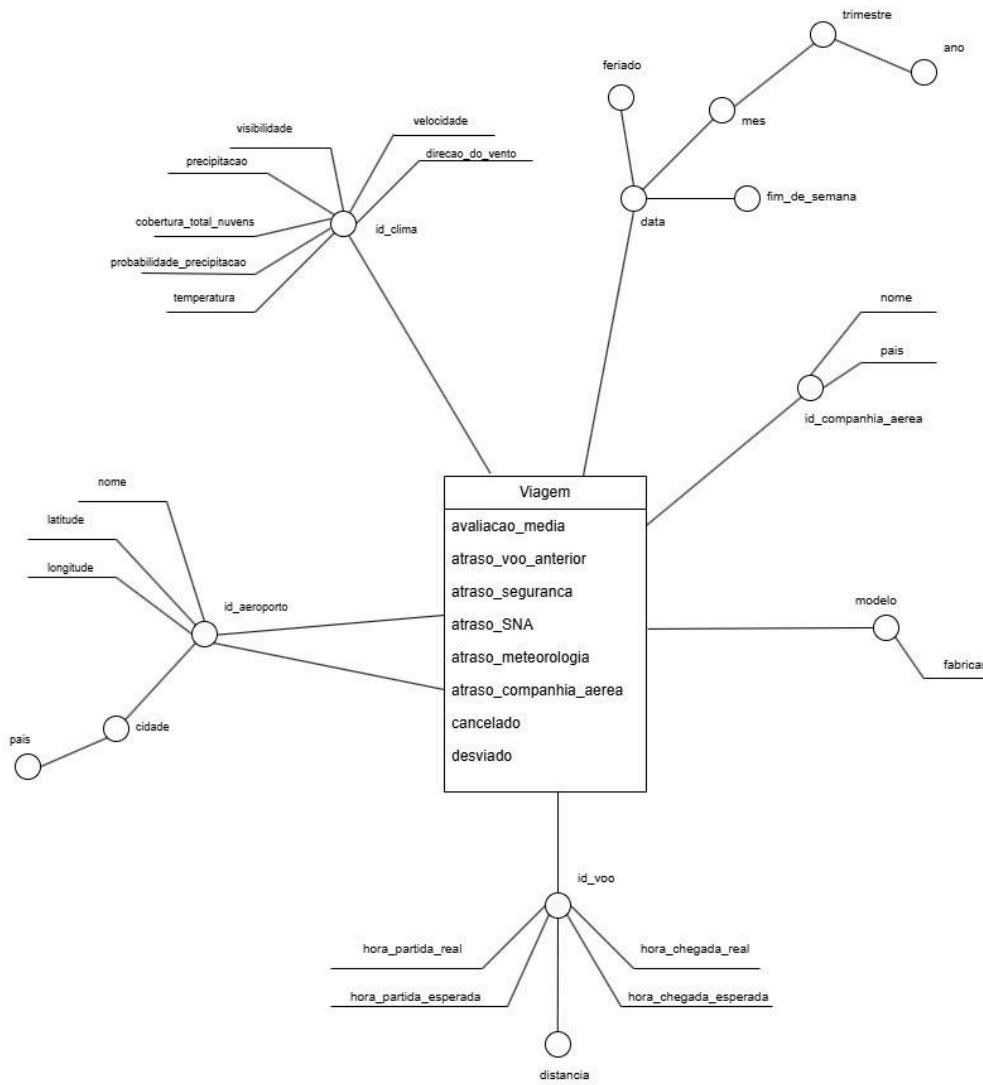


Figura 1 - Modelo Dimensional.

Ferramentas de Modelação e Implementação

Ao longo do processo, foram utilizadas diversas ferramentas para cobrir cada etapa da solução:

- **MySQL** como sistema de base de dados para implementar o armazém de dados;
- **Apache NiFi** para orquestração da ingestão e fluxo de dados desde os ficheiros CSV até às tabelas de staging e produção;
- **Draw.io** e/ou **Lucidchart** para criação dos diagramas de modelação dimensional;
- **Python** e scripts SQL para validação e transformação de dados durante a fase de limpeza e carga;
- Ferramentas de visualização como **Power BI** ou **Jupyter Notebooks** para testes de exploração dos dados.

3.2 A Matriz de Decisão

Caracterização do Data Mart	
Identificação:	
Descrição Geral: Informação para suporte à tomada de decisão na área de análise de performance de voos, fornecendo dados selecionados acerca dos mesmos em diversos aeroportos, de forma a poder ser feita uma análise dos voos realizados e dos atrasos inerentes aos voos para fazer uma avaliação do desempenho dos voos, aviões, companhias aéreas, aeroportos, etc...	
Estrutura Base	
Tabelas de Factos >>	TF-Viagem
<< Dimensões	
Voo	✓
Clima	✓
Aeroporto	✓
Avião	✓
Calendário	✓
Companhia Aérea	✓
Número Dimensões	6
Tipo	Snapshot

Periodicidade	Semanal
Descrição	Voos de passageiros efetuados, com foco em métricas de atraso
Utilidade estratégica	Análise da performance dos voos. Identificação de padrões de atraso. Apoio à tomada de decisão na gestão de frota e escolha de rotas. Apoio na mitigação de fatores recorrentes de atrasos.
Utilizadores	Diretores de aeroportos e gestores das companhias aéreas.
Observações	
Nada a assinalar.	

3.3 Definição e Caracterização de Dimensões

No âmbito do modelo dimensional para o nosso Data Warehouse, foram definidas e implementadas várias dimensões que permitem contextualizar e descrever de forma detalhada os factos registados na tabela de factos (Viagem). Estas dimensões foram selecionadas com base nos requisitos de análise definidos para o sistema e visam suportar a exploração temporal, geográfica e operacional dos dados relacionados com voos comerciais.

Dimensão Tempo (`dim_data`)

- **Descrição:** Permite analisar os voos ao longo do tempo, facilitando comparações sazonais, mensais ou diárias.
- **Atributos:** data, dia, mês, ano, trimestre, feriado, fim_de_semana.
- **Hierarquia:** Dia → Mês → Trimestre → Ano.
- **Tipo:** Dimensão de tempo.
- **Justificação:** Essencial para análises temporais.

Dimensão Companhia Aérea (`dim_companhia`)

- **Descrição:** Armazena os dados relativos às companhias aéreas.
- **Atributos:** id_companhia, nome, país.
- **Tipo:** Dimensão estática (não muda frequentemente).
- **Justificação:** Necessária para análises por companhia aérea, comparação de performance entre operadores, etc.

Dimensão Aeroporto (`dim_aeroporto`)

- **Descrição:** Contém a informação sobre os aeroportos de origem e destino dos voos.
- **Atributos:** id_aeroporto, nome, cidade , país, latitude, longitude.
- **Hierarquia:** Aeroporto → Cidade → País
- **Tipo:** Dimensão estática (não muda frequentemente).
- **Justificação:** Crucial para análise geográfica de rotas e avaliação de atrasos por origem/destino.

Dimensão Avião (dim_aeronave)

- **Descrição:** Representa os modelos de aviões utilizados nos voos.
- **Atributos:** modelo, fabricante.
- **Tipo:** Dimensão estática (não muda frequentemente)..
- **Justificação:** Permite estudar o desempenho por tipo de aeronave e eventuais correlações com atrasos ou avaliações.

Dimensão Meteorologia Estimada (dim_clima)

- **Descrição:** Resulta do acesso a uma API e no caso de os registos serem null de uma inferência indireta com base nos atrasos reportados para cada tipo de causa (ex: segurança, meteorologia).
- **Atributos:** id_clima, temperatura, precipitacao, cobertura, direcao, velocidade, visibilidade.
- **Tipo:** Dimensão estática / Dimensão derivada (construída com base em regras de inferência).
- **Justificação:** Importante para análises de impacto do clima no desempenho dos voos, mesmo quando os dados meteorológicos reais não estão disponíveis.

Dimensão Voo (dim_Voo)

- **Descrição:** Resulta do acesso a uma API e no caso de os registos serem null de uma inferência indireta com base nos atrasos reportados para cada tipo de causa (ex: segurança, meteorologia).
- **Atributos:** id_clima, temperatura, precipitacao, cobertura, direcao, velocidade, visibilidade.
- **Tipo:** Dimensão estática / Dimensão derivada (construída com base em regras de inferência).
- **Justificação:** Importante para análises de impacto do clima no desempenho dos voos, mesmo quando os dados meteorológicos reais não estão disponíveis.

3.4 Definição e Caracterização da Tabela de Factos e Respetivo Grão

3.4.1 Tabela de Factos

A tabela de factos do nosso Data Warehouse constitui o núcleo central do modelo dimensional, armazenando os dados quantitativos e mensuráveis dos voos.

Nome da Tabela: Viagem

Chaves Estrangeiras:

- id_voo → dim_voo
- id_companhia_aerea → dim_companhia
- id_aeroporto_origem → dim_aeroporto
- id_aeroporto_chegada → dim_aeroporto
- id_aeronave → dim_aeronave
- id_clima → dim_clima

Medidas (Factos):

- avaliacao_media
- atraso_voo_anterior
- atraso_seguranca
- atraso_SNA
- atraso_meteorologia
- atraso_companhia_area
- cancelado
- desviado

3.4.2 Grão

“Uma linha na tabela de factos Viagem representa um voo específico, realizado numa determinada data, por uma companhia aérea, com um avião, de um determinado aeroporto de origem, para um aeroporto de destino associado a condições climáticas específicas.”

3.5 Configuração dos Esquemas

A configuração do armazém de dados foi estruturada com base em um **esquema global centralizado**, adotando uma **arquitetura em floco de neve**, que permite garantir maior normalização das dimensões e, por consequência, maior integridade dos dados. O modelo está organizado em torno de uma **tabela de factos principal – Viagem**.

Organização Geral

O esquema global foi dividido em três áreas:

- **Área de staging (RAW e AUD)**: onde os dados são inicialmente armazenados após a ingestão via NiFi.
- **Área de preparação e tratamento**: onde ocorrem validações, correções e tradução de campos (como de inglês para português), e criação de históricos.
- **Área de produção (Data Warehouse)**: composta pela tabela de factos e suas dimensões associadas.

Tabela de Factos

A tabela **Viagem** armazena todos os eventos relativos aos voos, identificando cada operação por data, companhia aérea, origem, destino e modelo de avião. Esta tabela também centraliza os valores dos diferentes tipos de atraso, que posteriormente servem de base para análises operacionais e inferência de fatores climáticos.

Dimensões e Subdimensões

A estrutura de floco de neve normaliza as dimensões, dividindo-as em subdimensões para melhor organização. Exemplos:

- **Dim_Tempo** → ligada a subdimensões Mês, Trimestre, Ano, Fim de semana, Feriado.
- **Dim_Aeroporto** → ligada a subdimensões Cidade e País.
- **Dim_Aeronave**
- **Dim_CompanhiaAerea**
- **Dim_Clima** → ligada a subdimensões temperatura, precipitação, probabilidade_precipitação, direção_vento, velocidade_vento, visibilidade, cobertura_nuvens.

Data Mart

O Data Mart construído no âmbito deste projeto tem como principal finalidade **suportar a tomada de decisão operacional e estratégica** no setor da aviação, com foco na **análise de**

desempenho de voos e identificação de padrões de atraso. Trata-se de um repositório temático que consolida dados relevantes sobre voos comerciais, integrando informações sobre rotas, aeroportos, companhias aéreas, aeronaves e condições climatéricas.

Através deste Data Mart, é possível:

- **Avaliar o desempenho de companhias aéreas e aeroportos**, com base em indicadores como atraso médio, percentagem de voos cancelados ou desviados e avaliações médias dos voos;
- **Identificar padrões recorrentes de atraso**, considerando fatores operacionais e externos, como condições meteorológicas ou falhas anteriores;
- **Apoiar a reconfiguração de horários de voo**, alocação da frota e ajustes na escala operacional com base em evidência histórica;
- **Comparar a performance entre operadores e infraestruturas**, promovendo uma cultura de melhoria contínua e eficiência;
- **Facilitar a exploração interativa dos dados**, através da integração com dashboards em Power BI, que permitem análises por tempo, localização, tipo de atraso e operador.

Este Data Mart, modelado em esquema de floco de neve, permite realizar análises aprofundadas a partir de uma **tabela de factos central (Viagem)** e **seis dimensões** (Tempo, Companhia Aérea, Aeroporto, Avião, Clima e Voo), estando orientado para **consultas analíticas com granularidade ao nível do voo individual**.

A sua construção visa, portanto, **transformar dados operacionais complexos em conhecimento acionável**, promovendo decisões mais rápidas, fundamentadas e eficazes.

3.6 As Vistas dos Agentes de Decisão

O sistema foi concebido para servir um conjunto diversificado de agentes de decisão no setor da aviação, com **vistas analíticas partilhadas**, mas ajustadas às **necessidades de análise e tomada de decisão específicas** de cada perfil. Todas as vistas são suportadas por um modelo

dimensional no Data Warehouse e exploradas através de dashboards interativas desenvolvidas em Power BI, como será referido no ponto 5.

Importa referir que, para os perfis de **agente de viagens, gestor do aeroporto e gestor da companhia aérea**, a **vista apresentada é comum**, dado que todos estes utilizadores partilham necessidades operacionais semelhantes, baseadas na análise de atrasos, desempenho por rotas e eficiência de operação.

Agente de Viagens / Gestor do Aeroporto / Gestor da Companhia Aérea

Estes perfis têm como objetivo garantir a eficiência operacional, minimizar os atrasos e melhorar a experiência do passageiro. As suas principais necessidades incluem:

- Análise detalhada de atrasos por aeroporto de origem e destino.
- Monitorização da média de atrasos por modelo de avião e por companhia aérea.
- Avaliação das causas de atraso: meteorologia, segurança, gestão do voo anterior, entre outras.
- Identificação de companhias e rotas com maior taxa de cancelamento ou desvio.
- Análise por períodos críticos (horário do dia, dia da semana, sazonalidade).
- Comparação de desempenho entre diferentes operadores e infraestruturas.

A partilha da mesma vista analítica garante consistência de dados e facilita a comunicação entre diferentes áreas operacionais.

Serviço ao Cliente

O foco deste perfil está na **satisfação e retenção dos passageiros**, sendo essencial o acesso a dados relacionados com a qualidade do serviço prestado. As principais necessidades são:

- Acesso a indicadores de atrasos significativos que impactam a experiência do cliente.
- Percentagem de voos cancelados ou desviados por companhia aérea.
- Cruzamento de atrasos com períodos de maior tráfego para reforço de comunicação proativa.
- Apoio ao tratamento de reclamações com base em dados históricos concretos.
- Identificação de padrões que possam justificar medidas compensatórias.

Tecnologia da Informação (TI)

A equipa de TI é responsável pela **infraestrutura e disponibilidade do sistema**, tendo necessidades mais técnicas relacionadas com a manutenção e evolução da solução:

- Garantia de disponibilidade do sistema 24/7 para os utilizadores finais.
- Validação do tempo de resposta nas consultas e dashboards.
- Monitorização da atualização diária dos dados no Data Warehouse.
- Integração com fontes externas e APIs (meteorologia, dados operacionais).
- Escalabilidade da solução com base no crescimento de dados históricos.

Executivos das Companhias Aéreas (Decisão Estratégica)

Este perfil tem uma perspetiva mais agregada e focada na **avaliação de desempenho global e definição de estratégias de longo prazo**. As suas necessidades incluem:

- Indicadores comparativos entre companhias e aeroportos.
- Análise da fiabilidade operacional com base na média de atrasos e % de atrasos >30 min.
- Visão integrada do impacto dos cancelamentos e desvios.
- Apoio à definição de políticas operacionais e investimentos estratégicos.
- Avaliação da reputação e posicionamento da companhia com base em dados reais.

Todas estas vistas são alimentadas por dados consolidados no Data Warehouse, permitindo análises multidimensionais com filtros por tempo, localização, tipo de atraso e fatores externos. A partilha de uma base comum de dados garante **coerência entre as áreas operacionais, técnicas e estratégicas**, promovendo uma **tomada de decisão informada e colaborativa**.

4. Arquitetura Geral do Sistema

4.1 Apresentação Geral

Os dados são recolhidos a partir de diversas fontes heterogéneas, passando por uma camada de preparação que assegura a limpeza e transformação dos mesmos. Em seguida, os dados são integrados num armazém de dados estruturado segundo um modelo dimensional apresentado anteriormente. Este armazém serve de base para as análises realizadas por dashboards interativos e para os processos analíticos mais avançados de descoberta de conhecimento.

Para tal, o grupo recorreu ao uso de base de dados MySQL e ferramenta ETL/ELT Apache NiFi.

4.2 Fontes de Dados

Para posteriormente realizarmos o povoamento dos dados e a sua analise, tivemos de recolher datasets que permitissem responder aos fundamentos e aos objetivos deste projeto. Como tal, o grupo definiu que os principais domínios de dados necessários seriam:

- **Companhias Aéreas**
- **Voo**
- **Aeronaves**
- **Condições Climáticas**
- **Aeroportos**

Fonte de Dados	Tipo	Formato	Atualização	Dados Principais
Kaggle	Pública	CSV	Indefinido	Companhias Aéreas, Voo, Aeronaves
Open-meteo	Externa/API	CSV/JSON	Hora a Hora	Condições Climáticas
OpenFlights	Pública	CSV	Esporádica	Aeroportos

Preferencialmente, o grupo procurou usar sempre que possível datasets existentes ao invés da criação de uma dataset, de modo aos resultados da analises, mesmo que não tão bons e incompletos serem o mais parecido com a realidade. A periodicidade da atualização

depende do domínio em questão, para os Aeroportos, Aeronaves e Companhias_Aereas, a sua atualização é inconstante uma vez que se trata de informação que não varia com grande frequência. No caso do Voo, este é atualizado semanalmente.

4.2.1 Dataset de Companhias Aéreas

Este dataset retirado do site “[Kaggle - Airline Database](#)” com informação de mais de 5000 companhias aéreas. Em seguida apresentamos as 8 colunas que constituem cada entrada.

Airline ID	INT	Identificador único da companhia aérea no OpenFlights
Name	VARCHAR(75)	Nome da companhia aérea
Alias	INT	Apelido da companhia aérea.
IATA	VARCHAR(2)	Código IATA de 2 letras
ICAO	VARCHAR(3)	Código ICAO de 3 letras
Callsign	VARCHAR(45)	Indicativo de chamada da companhia aérea
Country	VARCHAR(45)	País ou território onde a companhia aérea está registrada.
Active	VARCHAR(1)	"Y" se a companhia aérea está ou esteve recentemente em operação, "N" se está extinta.

4.2.2 Dataset com Rotas de Voo

(falar dos dados criados adicionados aeronave...)

Este dataset foi retirado do “[Kaggle - Flight Delay and Cancellation Dataset \(2019-2023\)](#)” que contém vários dados sobre viagens. Em seguida, a tabela apresenta os dados neste dataset.

Posteriormente, o grupo decidiu atribuir um ID_aeronave para posteriormente relacionar as viagens e rotas com as aeronaves utilizadas.

FL_DATE	INT	Data do Voo (yyyymmdd)
----------------	-----	------------------------

AIRLINE_CODE	VARCHAR(2)	Código Único da Companhia Aérea
DOT_CODE	INT	Número de identificação atribuído pelo US DOT para identificar uma companhia aérea única.
FL_NUMBER	INT	Número do Voo
ORIGIN	VARCHAR(45)	Aeroporto de Origem
ORIGIN_CITY	VARCHAR(45)	Aeroporto de Origem, Nome da Cidade
DEST	VARCHAR(45)	Aeroporto de Destino
DEST_CITY	VARCHAR(45)	Aeroporto de Destino, Nome da Cidade
CRS_DEP_TIME	INT	Hora de Partida CRS (Sistema de Reservas Computadorizado) - (hhmm)
DEP_TIME	FLOAT	Hora Real de Partida
DEP_DELAY	FLOAT	Diferença entre a Hora Real de Partida e a Hora de Partida Prevista. (Horas de Partida adiantadas são representadas por números negativos)
TAXI_OUT	FLOAT	Tempo que um avião leva, após sair do portão, até efetuar a descolagem (em minutos).
WHEELS_OFF	FLOAT	Hora de Descolagem (hora local: hhmm)
WHEELS_ON	FLOAT	Hora de Aterragem (hora local: hhmm)
TAXI_IN	FLOAT	o tempo que o avião leva desde o momento em que toca na pista até chegar ao portão no aeroporto de destino (em minutos)..
CRS_ARR_TIME	INT	Hora de Chegada CRS
ARR_TIME	FLOAT	Hora Real de Chegada
ARR_DELAY	FLOAT	Diferença entre a Hora Real de Chegada e a Hora de Chegada Prevista. (Horas de

		Chegada adiantadas são representadas por números negativos)
CANCELLED	FLOAT	Indicador de Voo Cancelado (1=Sim)
CANCELLATION_CODE	VARCHAR(45)	Especifica a razão pela qual o voo foi cancelado
DIVERTED	FLOAT	Indicador de Voo Desviado (1=Sim)
CRS_ELAPSED_TIME	FLOAT	Tempo Planeado de Voo pelo Sistema de Reservas (em minutos)
ELAPSED_TIME	FLOAT	Tempo de Voo (em minutos)
AIR_TIME	FLOAT	Tempo de Voo no Ar (em minutos)
DISTANCE	FLOAT	Distância entre aeroportos
DELAY_DUE_CARRIER	FLOAT	Atraso da Companhia Aérea (em minutos)
DELAY_DUE_WEATHER	FLOAT	Atraso devido às condições climatéricas (em minutos)
DELAY_DUE_NAS	FLOAT	Atraso no Sistema Nacional de Aviação (em minutos)
DELAY_DUE_SECURITY	FLOAT	Atraso de Procedimentos Segurança (em minutos)
DELAY_DUE_LATE_AIRCRAFT	FLOAT	Tempo de atraso causado por um voo anterior (em minutos)

4.2.3 Dataset de Aeronaves

(Temos que mudar)

Por sua vez, o dataset dos aviões foi retirado do "[Kaggle](#) - Aircraft Performance (Aircraft Bluebook)" que contém vários dados sobre 861 aviões e as suas características. Em seguida, a tabela apresenta os dados neste dataset.

Model	VARCHAR(120)	Nome do avião
Company	VARCHAR(120)	Nome da empresa (fabricante ou operadora)

Engine Type	VARCHAR(45)	Tipo de motor usado no avião
HP or lbs thr ea engine	INT	Potência no eixo (HP) ou empuxo (lbf) em condições padrão da atmosfera (ISA). Unidades: HP ou lbf.
Max speed Knots	INT	Velocidade máxima do avião.
Rcmnd cruise Knots	INT	Velocidade de cruzeiro ideal do avião
Stall Knots dirty	INT	Velocidade de estol do avião em configuração “suja” (flaps estendidos, trem de pouso abaixado, etc.)
Fuel gal/lbs	INT	Capacidade de combustível do avião.
All eng service ceiling	INT	Altitude máxima de densidade com todos os motores funcionando
Eng out service ceiling	INT	Altitude máxima de densidade com apenas um motor funcionando
All eng rate of climb	INT	Taxa de subida do avião com todos os motores funcionando.
Eng out rate of climb	INT	Taxa de subida do avião com apenas um motor funcionando.
Takeoff over 50ft	INT	Velocidade de subida do avião durante a decolagem normal para ultrapassar obstáculo de 50 pés.
Takeoff ground run	INT	Distância percorrida no solo para decolar.
Takeoff over 50ft	INT	Velocidade de descida durante pouso normal para ultrapassar obstáculo de 50 pés.
Landing ground run	INT	Distância percorrida no solo durante o pouso.
Gross weight lbs	INT	Peso bruto do avião
Empty weight lbs	INT	Peso vazio do avião

Length ft/in	VARCHAR(15)	Comprimento total do avião
Height ft/in	VARCHAR(15)	Altura total do avião
Wingspan ft/in	VARCHAR(15)	Envergadura (largura total das asas).
Range N.M.	INT	Alcance do avião.

4.2.4 Dataset com Condições Climatéricas

Para este dataset o grupo utilizou uma API do site [Open-meteo](#). Introduzindo a Latitude, Longitude, e o intervalo de tempo pretendido podemos obter informações como a Temperatura, Probabilidade de Precipitação, Precipitação (chuva, neve, etc), Cobertura total de nuvens, Visibilidade, Velocidade e Direção do Vento.

4.2.5 Dataset de Aeroportos

Este dataset foi retirado apartir de uma *site online* chamado “[Open Flights](#)” que contém mais de 10 000 aeroportos, estações de comboios e terminais de embarcações por todo o mundo. Em seguida está representada na tabela abaixo com as informações sobre o dataset, os seus campos, o tipo das variáveis e a descrição.

ID_aeroporto	INT	ID do aeroporto
Nome	VARCHAR(120)	Nome completo do aeroporto
Cidade	VARCHAR(45)	Cidade onde está localizado o aeroporto
País	VARCHAR(45)	País do aeroporto
Código_IATA	VARCHAR(3)	Código IATA, usado comercialmente
Código_ICAO	VARCHAR(4)	Código ICAO, usado para tráfego aéreo
Latitude	FLOAT	Coordenada de localização geográfica
Longitude	FLOAT	Coordenada de localização geográfica
Altitude	INT	Altitude em pés

Fuso_Horário	INT	Diferença em horas em relação ao UTC
Código_DST	VARCHAR(1)	Código de horário de verão (U = desconhecido, N = none, E = Europe, etc.)
Zona_Horária	VARCHAR(45)	Zona horária local
Tipo	VARCHAR(45)	Tipo de localização
Fonte	VARCHAR(45)	Fonte dos dados

4.3 Área de Preparação de dados

De forma a garantir a qualidade, consistência e integridade dos dados, foi criada uma área de preparação baseada em ficheiros intermediários e scripts de transformação. Esta área desempenha um papel fundamental antes da integração no armazém de dados.

A preparação inclui:

- **Limpeza de dados** (ex: eliminação de nulos, alteração de nomes de atributos)
- **Enriquecimento** com dados de voos, com informação sobre as companhia aéreas, aeronaves, aeroportos e condições climatéricas.
- **Cálculo de indicadores derivados** (cálculo da média dos valores sobre o clima).

Embora não se tenha adotado um Data Lake, foi estruturada uma **zona de staging** que funciona como ponto de transição entre os dados brutos e os dados preparados.

4.3.1 Companhia Aérea

O tratamento de dados da Companhia Aérea passa por pegar no csv e introduzi-lo através do uso do NiFi numa tabela Companhia_AereaFONTE criada no MySQL. Com a criação de triggers, adicionamos os campos Operação, Etiqueta e Utilizador aos dados anteriores na tabela Companhia_AereaAUD.

Posteriormente, a tabela Companhia_AereaAUD é copiada para a Companhia_AereaRAW para garantir que temos uma “reserva” dos dados no caso de haver algum problema.

Passamos agora à limpeza dos dados inseridos em Companhia_AereaRAW, onde fazemos a verificação de nulos, caso haja o registo é inserido na tabela Companhia_AereaERRO, após a criação dos campos descrição e tipo, onde posteriormente podem ser vistos e se possível corrigidos para inserir outra vez no fluxo. Depois de garantir que todos os registos vindos da Companhia_AereaRAW são válidos, alteramos os nomes dos campos para português. Tendo este tratamento feito os registos são agora válidos para povoar a tabela Companhia_Aerea apenas com os campos pretendidos. Adicionalmente, é feito o povoamento da tabela Companhia_AereaHST com os campos DataHoraModificação e Modificação.

Ficamos portanto com uma tabela com os seguintes campos:

id_companhia_aerea	INT PRIMARY KEY	Identificador único da companhia aérea no OpenFlights
nome	VARCHAR(150)	Nome da companhia aérea
país	VARCHAR(50)	País ou território onde a companhia aérea está registrada.

4.3.2 Aeronaves

O tratamento dos dados das Aeronaves inicia-se com a ingestão do ficheiro CSV por meio do Apache NiFi, sendo os dados inicialmente carregados na tabela AeronaveFONTE, existente na base de dados MySQL.

A seguir, são aplicadas triggers nesta tabela para enriquecer os registos com três campos adicionais: Operacao, Etiqueta e Utilizador. Os dados resultantes desse enriquecimento são armazenados na tabela AeronaveAUD.

Todos os dados da AeronaveAUD são replicados na tabela Companhia_AereaRAW, que funciona como uma cópia de salvaguarda.

Na etapa seguinte, é realizado o processo de validação e tratamento de erros. Os registos que apresentem valores nulos ou inconsistências são automaticamente redirecionados para a tabela AeronaveERRO.

Depois de validados os dados da AeronaveRAW, procede-se à normalização dos nomes dos campos, traduzindo-os para português conforme a convenção adotada. Os registos

já limpos e com formato padronizado são finalmente inseridos na tabela definitiva Aeronave, contendo apenas os campos essenciais ao projeto.

Além disso, é mantida a tabela Companhia_AereaHST para assegurar rastreabilidade e controlo de versões.

Modelo	VARCHAR(100) PRIMARY KEY	Nome do Motor como identificador único
Fabricante	VARCHAR(100)	Nome do Fabricante

4.3.3 Aeroporto

O processo de tratamento dos dados relativos ao Aeroporto inicia-se com a importação do ficheiro CSV através da ferramenta Apache NiFi, que insere os registo na tabela Aeroporto_FONTE, previamente criada numa base de dados MySQL.

A partir desta tabela, são definidos triggers que permitem adicionar automaticamente três campos adicionais a cada registo: Operacao, Etiqueta e Utilizador. Estes dados enriquecidos são então armazenados na tabela AeroportoAUD, que funciona como registo de auditoria das operações realizadas.

Para garantir a segurança e recuperação dos dados, os registo da tabela de auditoria (AeroportoAUD) são copiados integralmente para a tabela AeroportoRAW, que atua como reserva de segurança (backup), assegurando a preservação da informação em caso de falhas ou inconsistências no fluxo.

Segue-se a fase de validação e limpeza dos dados. Nesta etapa, são verificados os campos nulos ou inválidos. Os registo que apresentem problemas são redirecionados para a tabela AeroportoERRO. Esta tabela contém ainda os campos adicionais Descricao e Tipo, que especificam o motivo do erro.

Os registo limpos e validados são então utilizados para povoar a tabela final Aeroporto, Cidade e Pais.

Por fim, é ainda realizada a atualização da tabela AeroportoHST, que guarda um histórico de alterações com os campos DataHoraModificacao e Modificacao, permitindo a rastreabilidade de mudanças ao longo do tempo.

id_cidade	INT PRIMARY KEY	Identificador único da cidade
cidade	VARCHAR(100)	Nome da Cidade

país	VARCHAR(100) PRIMARY KEY	Nome do país como identificador único
-------------	--------------------------	---------------------------------------

ID_aeroporto	INT PRIMARY KEY	ID do aeroporto
Nome	VARCHAR(100)	Nome completo do aeroporto
Cidade	VARCHAR(100)	Cidade onde está localizado o aeroporto
País	VARCHAR(100)	País do aeroporto
Código_IATA	CHAR(3) UNIQUE	Código IATA, usado comercialmente
Latitude	DECIMAL(15,10)	Coordenada de localização geográfica
Longitude	DECIMAL(15,10)	Coordenada de localização geográfica

4.3.4 Voo

Anteriormente, ao tratamento dos dados do dataset de Voos, devido a falta de dados que associasse o voo às aeronaves e avaliação dos clientes foi realizado um código python de modo a integrar o modelo de aeronave associado a cada voo e a média de avaliação de satisfação, dependendo da companhia aérea.

Para o dataset de Voos começa com a importação dos dados em formato CSV para uma base de dados onde os dados são carregados na tabela VooFONTE.

Foram definidas triggers que, ao detetar inserções, modificações e remoções, acrescentam novos campos. Estes dados são registados na tabela VooAUD, que atua como uma camada de auditoria.

Os dados da VooAUD são copiados para a tabela VooRAW. Segue-se a etapa de validação e limpeza de dados, onde são verificados os campos nulos ou inválidos. Em caso de falha é enviado para a tabela VooQUA.

Após a validação, os dados corretos e completos são renomeados para seguir a nomenclatura em português, e são então carregados na tabela final distancia, Voo e VooRDY.

id_distancia	INT AUTO_INCREMENT PRIMARY KEY	ID da distância
valor	DECIMAL(5,1)	valor da distância

id_voo	INT AUTO_INCREMENT PRIMARY KEY	Identificador único do voo
nr_voo	INT	Número do voo
hora_partida_esperada	DECIMAL(5,1)	Hora de Partida CRS (Sistema de Reservas Computadorizado) - (hhmm)
hora_partida_real	DECIMAL(5,1)	Hora Real de Partida
hora_chegada_esperada	DECIMAL(5,1)	Hora de Chegada CRS
hora_chegada_real	DECIMAL(5,1)	Hora Real de Chegada

data_partida	DATE	Data do Voo (yyyymmdd)
id_voo	INT AUTO_INCREMENT PRIMARY KEY	Identificador único do voo
nr_voo	INT	Número do voo
nome_companhia_aerea	VARCHAR(100)	Número da companhia aérea único
aeroporto_origem	CHAR(3)	Aeroporto de Origem
aeroporto_destino	CHAR(3)	Aeroporto de Destino
CRS_DEP_TIME	INT	Hora de Partida CRS (Sistema de Reservas Computadorizado) - (hhmm)

hora_partida_real	DECIMAL(5,1)	Hora Real de Partida
hora_partida_esperada	DECIMAL(5,1)	Diferença entre a Hora Real de Partida e a Hora de Partida Prevista. (Horas de Partida adiantadas são representadas por números negativos)
hora_chegada_esperada	DECIMAL(5,1)	Hora de Chegada CRS
hora_chegada_real	DECIMAL(5,1)	Hora Real de Chegada
cancelado	DECIMAL(6,1)	Indicador de Voo Cancelado (1=Sim)
desviado	DECIMAL(6,1)	Indicador de Voo Desviado (1=Sim)
distancia	DECIMAL(6,1)	Distância entre aeroportos
atraso_companhia_aerea	DECIMAL(5,1)	Atraso da Companhia Aérea (em minutos)
atraso_meteorologia	DECIMAL(5,1)	Atraso devido às condições climatéricas (em minutos)
atraso_SNA	DECIMAL(5,1)	Atraso no Sistema Nacional de Aviação (em minutos)
atraso_segurança	DECIMAL(5,1)	Atraso de Procedimentos Segurança (em minutos)
atraso_voo_anterior	DECIMAL(5,1)	Tempo de atraso causado por um voo anterior (em minutos)
Modelo_Avião	VARCHAR(100)	Modelo do Avião
avaliacao	DECIMAL(5,1)	Média da avaliação de satisfação do voo

4.3.4 Condições Climatéricas

Por sua vez, para conseguir obter informação sobre as condições climatéricas através do dia_partida, longitude e latitude do aeroporto de partida, recorremos a uma API - OpenMeteo. Esta API oferecia, entre outros, os campos indicados abaixo com um valor de hora em hora. De modo a realizar o povoamento na tabela do clima, utilizamos um script para calcular a média de cada campo. É de realçar que muitos dos registos vindos desta API, sendo ela gratuita, encontram-se a null.

id_clima	INT PRIMARY KEY	ID do clima
temperatura	DECIMAL(15,5)	Temperatura média
precipitacao	DECIMAL(15,5)	Precipitação Total
probabilidade	DECIMAL(15,5)	Probabilidade de Precipitação
visibilidade	DECIMAL(15,5)	Visibilidade
cobertura	DECIMAL(15,5)	Cobertura de nuvens
velocidade	DECIMAL(15,5)	Velocidade do vento
direcao	DECIMAL(15,5)	Direção do vento

4.4 O Armazém de Dados

O armazém de dados foi estruturado com base num **modelo dimensional**, facilitando a análise eficiente por parte dos utilizadores finais. Este modelo é composto por:

- **Tabelas de Dimensão:** companhias aéreas, aeroportos, aeronaves e condições climatéricas.

```
CREATE TABLE Companhia_Aerea (
    id_companhia_aerea INT PRIMARY KEY,
    nome VARCHAR(150) UNIQUE NOT NULL,
    país VARCHAR(50) NOT NULL
);
```

```
CREATE TABLE Aeronave (
    Modelo VARCHAR(100) PRIMARY KEY,
    Fabricante VARCHAR(100)
);
```

```
CREATE TABLE Aeroporto (
```

```

        ID_aeroporto INT PRIMARY KEY,
        Codigo_IATA CHAR(3) UNIQUE,
        Nome VARCHAR(100),
        Cidade VARCHAR(100),
        Pais VARCHAR(100),
        Latitude DECIMAL(15,10),
        Longitude DECIMAL(15,10)
    );

```

```

CREATE TABLE clima (
    id_clima INT PRIMARY KEY,
    temperatura DECIMAL(5,5),
    precipitacao DECIMAL(5,5),
    probabilidade DECIMAL(5,5),
    visibilidade DECIMAL(5,5),
    cobertura DECIMAL(5,5),
    velocidade DECIMAL(5,5),
    direcao DECIMAL(5,5)
);

```

```

CREATE TABLE Voo (
    id_voo INT AUTO_INCREMENT PRIMARY KEY,
    nr_voo INT,
    id_distancia INT,
    hora_partida_real DECIMAL(5,1),
    hora_partida_esperada DECIMAL(5,1),
    hora_chegada_real DECIMAL(5,1),
    hora_chegada_esperada DECIMAL(5,1),
    FOREIGN KEY (id_distancia) REFERENCES RDY.distancia(id_distancia)
    ON DELETE CASCADE
);

```

- **Tabela de Factos:** registos de voos com medidas associadas, como atrasos, distância, voo cancelado, voo desviado, etc.

```
CREATE TABLE Viagem (
```

```

        id_viajem INT AUTO_INCREMENT PRIMARY KEY,
        avaliacao DECIMAL(5,1),
        cancelado DECIMAL(6,1),
        desviado DECIMAL(6,1),
        atraso_companhia_aerea DECIMAL(5,1),
        atraso_metereologia DECIMAL(5,1),
        atraso_SNA DECIMAL(5,1),
        atraso_seguranca DECIMAL(5,1),
        atraso_voo_anterior DECIMAL(5,1),
        id_voo INT,
        nome_companhia_aerea VARCHAR(150),
        aeroporto_origem CHAR(3),
        aeroporto_destino CHAR(3),
        Modelo_Avião VARCHAR(100),
        id_clima INT,
        FOREIGN KEY (id_voo) REFERENCES RDY.Voo(id_voo) ON DELETE CASCADE,
            FOREIGN KEY (nome_companhia_aerea) REFERENCES
RDY.Companhia_Aerea(nome) ON DELETE CASCADE,
            FOREIGN KEY (aeroporto_origem) REFERENCES
RDY.Aeroporto(Codigo_IATA) ON DELETE CASCADE,
            FOREIGN KEY (aeroporto_destino) REFERENCES
RDY.Aeroporto(Codigo_IATA) ON DELETE CASCADE,
            FOREIGN KEY (Modelo_Avião) REFERENCES RDY.Aeronave(Modelo) ON
DELETE CASCADE,
            FOREIGN KEY (id_clima) REFERENCES RDY.clima(id_clima) ON DELETE
CASCADE
);

```

4.5 Exploração e Visualização de Dados

A camada de exploração e visualização de dados tem como principal objetivo facilitar o acesso e a análise da informação por diferentes tipos de utilizadores, disponibilizando dashboards intuitivos, interativos e adaptados às suas necessidades específicas. Esta camada

é fundamental para a monitorização de operações, apoio à decisão estratégica e identificação de anomalias ou oportunidades.

A exploração dos dados será feita através de ferramentas de **Business Intelligence (BI)** modernas, como:

- **Power BI**

A primeira etapa da exploração dos dados foi realizada por meio da análise estatística descritiva das variáveis numéricas, como ilustrado na figura abaixo. Essa análise apresenta informações importantes, como:

- O número de valores não nulos (count),
- A média (mean),
- O desvio padrão (std),
- O valor mínimo (min) e máximo (max),
- Os percentis (25%, 50% e 75%).

Estas métricas fornecem-nos uma visão geral da distribuição dos dados e ajudam a identificar possíveis outliers, valores em falta e padrões relevantes. Com base nesta análise inicial, é possível tomar decisões mais informadas relativamente ao pré-processamento dos dados, ao tratamento de valores inconsistentes ou nulos, bem como orientar a criação de novas variáveis que possam melhorar o desempenho dos modelos futuros.

A plataforma será usada para diferentes perfis, com permissões, filtros e visualizações adequadas:

- **Executivos;**
- **Analistas de Dados;**

```
df_flights.info( )

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3000000 entries, 0 to 2999999
Data columns (total 38 columns):
 #   Column           Dtype  
--- 
 0   nr_voo            int64  
 1   data_partida      object  
 2   nome_companhia_aerea  object  
 3   id_companhia_aerea  object  
 4   pais_companhia_aerea  object  
 5   aeroporto_origem    object  
 6   nome_aeroporto_origem  object  
 7   pais_aeroporto_origem  object  
 8   cidade_aeroporto_origem  object  
 9   latitude_aeroporto_origem  float64 
 10  longitude_aeroporto_origem  float64 
 11  aeroporto_destino    object  
 12  nome_aeroporto_destino  object  
 13  pais_aeroporto_destino  object  
 14  cidade_aeroporto_destino  object  
 15  latitude_aeroporto_destino  float64 
 16  longitude_aeroporto_destino  float64 
 17  hora_partida_real    float64 
 18  hora_partida_esperada  int64  
 19  hora_chegada_real    float64 
 20  hora_chegada_esperada  int64  
 21  cancelado           float64 
 22  desviado             float64 
 23  distancia            float64 
 24  atraso_companhia_aerea  float64 
 25  atraso_metereologia   float64 
 26  atraso_SNA            float64 
 27  atraso_seguranca     float64 
 28  atraso_voo_anterior   float64 
 29  Modelo_Avião         object  
 30  Fabricante           object  
 31  avaliacao            float64 
 32  temperatura          float64 
 33  precipitacao          float64 
 34  visibilidade          float64 
 35  cobertura_nuvens     float64 
 36  velocidade            float64 
 37  direcao               float64 
dtypes: float64(21), int64(3), object(14)
memory usage: 869.8+ MB
```

```
# Summary statistics for numerical variables
df_flights.describe().T
```

	count	mean	std	min	25%	50%	75%	max
nr_voo	3000000.0	2511.535519	1747.258040	1.000000e+00	1051.000000	2152.000000	3797.000000	9562.000000
latitude_aeroporto_origem	2999265.0	36.688177	6.011262	-1.433100e+01	32.898602	37.362598	40.777199	71.285402
longitude_aeroporto_origem	2999265.0	-94.806949	18.432813	-1.766460e+02	-110.941002	-87.904800	-80.943100	145.729004
latitude_aeroporto_destino	2999238.0	36.684660	6.013082	-1.433100e+01	32.898602	37.362598	40.777199	71.285402
longitude_aeroporto_destino	2999238.0	-94.809625	18.448716	-1.766460e+02	-110.941002	-87.904800	-80.943100	145.729004
hora_partida_real	2922385.0	1329.775913	499.310052	1.000000e+00	916.000000	1323.000000	1739.000000	2400.000000
hora_partida Esperada	3000000.0	1327.061984	485.878854	1.000000e+00	915.000000	1320.000000	1730.000000	2359.000000
hora_chegada_real	2920058.0	1466.511162	531.883849	1.000000e+00	1053.000000	1505.000000	1913.000000	2400.000000
hora_chegada Esperada	3000000.0	1490.560665	511.547566	1.000000e+00	1107.000000	1516.000000	1919.000000	2400.000000
cancelado	3000000.0	0.026380	0.160263	0.000000e+00	0.000000	0.000000	0.000000	1.000000
desviado	3000000.0	0.002352	0.048440	0.000000e+00	0.000000	0.000000	0.000000	1.000000
distancia	3000000.0	809.361552	587.893938	2.900000e+01	377.000000	651.000000	1046.000000	5812.000000
atraso_companhia_aerea	533863.0	24.759086	71.771845	0.000000e+00	0.000000	4.000000	23.000000	2934.000000
atraso_meteorologia	533863.0	3.985260	32.410796	0.000000e+00	0.000000	0.000000	0.000000	1653.000000
atraso_SNA	533863.0	13.164728	33.161122	0.000000e+00	0.000000	0.000000	17.000000	1741.000000
atraso_seguranca	533863.0	0.145931	3.582053	0.000000e+00	0.000000	0.000000	0.000000	1185.000000
atraso_voo_anterior	533863.0	25.471282	55.766892	0.000000e+00	0.000000	0.000000	30.000000	2557.000000
avaliacao	3000000.0	3.381326	1.785002	0.000000e+00	2.000000	3.300000	4.600000	10.000000
temperatura	3000000.0	20.309715	3.299005	1.500000e+01	17.654811	20.207847	22.756925	45.729279
precipitacao	3000000.0	3.981724	5.134459	1.189148e-06	1.409097	2.815961	4.221969	77.854046
visibilidade	3000000.0	11.592150	2.082938	5.044410e-01	9.823237	11.588020	13.354516	17.244127
cobertura_nuvens	3000000.0	25.875419	15.811842	7.554426e-07	12.753382	25.485473	38.204644	114.877667
velocidade	3000000.0	15.047048	9.772495	5.000005e+00	9.153204	13.271599	17.393906	68.917697
direcao	3000000.0	181.520829	104.996867	1.031431e-04	90.695192	181.461644	272.065112	413.557120

```
# Summary statistics for categorical variables
df_flights.describe(include='object').T
```

	count	unique		top	freq
data_partida	3000000	1704		2019-07-25	2379
nome_companhia_aerea	3000000	18		Southwest Airlines	576470
id_companhia_aerea	3000000	18		WN	576470
pais_companhia_aerea	3000000	2		United States	2899533
aeroporto_origem	2999265	378		ATL	153556
nome_aeroporto_origem	2999265	378	Hartsfield Jackson Atlanta International Airport	153556	
pais_aeroporto_origem	2999265	6		United States	2981743
cidade_aeroporto_origem	2999265	362		Chicago	157368
aeroporto_destino	2999238	378		ATL	153569
nome_aeroporto_destino	2999238	378	Hartsfield Jackson Atlanta International Airport	153569	
pais_aeroporto_destino	2999238	6		United States	2981627
cidade_aeroporto_destino	2999238	362		Chicago	158087
Modelo_Avião	3000000	105		Boeing 737-300	39478
Fabricante	3000000	6		Boeing	1836802

4.6 Aquisição de Conhecimento

A aquisição de conhecimento consiste na extração de informação relevante a partir dos dados tratados, permitindo ao sistema gerar insights e apoiar a tomada de decisões. Esta etapa foi essencial para transformar os dados brutos em atributos informativos, enriquecendo a base analítica do projeto.

Foram criadas diversas **novas variáveis (features)**, com o objectivo de captar comportamentos complexos relacionados com o desempenho dos aeroportos, padrões de atraso e condições meteorológicas. Entre os atributos derivados, destacam-se:

- **Informações operacionais dos aeroportos:**
 - dias_desde_ultimo_atraso, nr_total_voos, nr_rotas_distintas, nr_companhias_distintas, Unique_Flights.
- **Métricas de desempenho e qualidade:**
 - atraso_medio, avaliacao_media, distancia_media.
- **Médias de variáveis meteorológicas:**
 - precipitacao_media, visibilidade_media, cobertura_nuvens_media, velocidade_media, direcao_media.
- **Padrões temporais identificados nos aeroportos de origem e destino:**
 - Dias médios entre voos (*_avg_days_between), horários e dias de maior movimento (*_peak_day, *_peak_hour, *_peak_month).

Estas variáveis foram construídas com base na análise exploratória dos dados e em conhecimento do domínio, com o intuito de melhorar o desempenho dos modelos preditivos e enriquecer as análises subsequentes. A criação destas features representa um passo fundamental na geração de conhecimento, permitindo análises mais profundas e decisões mais informadas.

5. O Povoamento de Dados

5.1 Apresentação da Abordagem Realizada

O processo de povoamento de dados foi estabelecido com base numa abordagem ETL (Extract, Transform, Load) clássica para garantir a integridade, qualidade e rastreabilidade dos dados. Utilizámos o Apache NiFi como ferramenta de orquestração dos fluxos de dados, devido à sua interface gráfica intuitiva, controlo de fluxo e integração nativa com sistemas de bases de dados como o MySQL, onde reside o nosso Data Warehouse.

O fluxo ETL inicia-se com a extração dos ficheiros CSV. Estes dados são inseridos em tabelas de staging (*FONTE*), passando posteriormente por transformações em tabelas *RAW* e de *ERROS*, até alcançarem as tabelas de produção (*RDY*). Foram ainda utilizadas triggers SQL para rastrear modificações nas tabelas, e garantir a manutenção de históricos (*AUD* e *HST*).

A tabela de Voos, ao contrário da companhia Aerea, Aeronaves e Aeroportos, não tem uma tabela AUD ou HST, uma vez que uma vez realizado o voo, ele é registado e não pode ser alterado. Caso haja algum erro no registo de voos, este é revisto manualmente e reinserido no fluxo após o seu tratamento.

5.2 Mapeamento de Dados

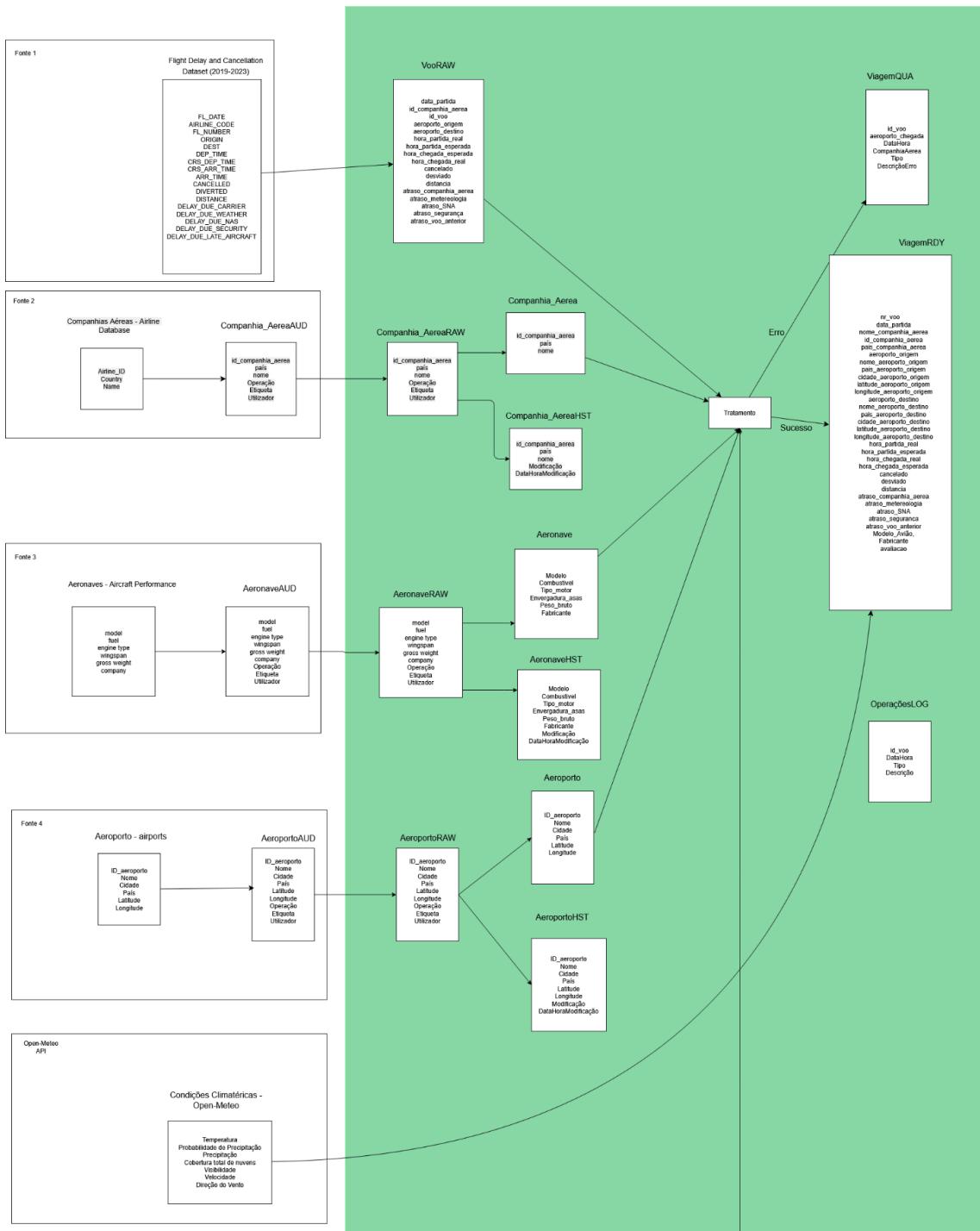


Figura 3 - Source-to-target datamap.

O mapeamento source-to-target foi definido de forma a alinhar os dados brutos com a estrutura do armazém de dados. O seguinte exemplo resume os principais mapeamentos efetuados:

5.2.1 Companhia Aerea

Campo Fonte CSV	Campo Destino CSV	Transformações Aplicadas
Airline_ID	id_companhia_aerea	Verificação de nulos, Alteração do nome
Country	país	Verificação de nulos, Alteração do nome
Name	nome	Verificação de nulos, Alteração do nome

5.2.2 Aeronave

Campo Fonte CSV	Campo Destino CSV	Transformações Aplicadas
model	id_companhia_aerea	Verificação de nulos, Alteração do nome
company	país	Verificação de nulos, Alteração do nome

5.2.3 Aeroporto

Campo Fonte CSV	Campo Destino CSV	Transformações Aplicadas
ID_aeroporto	ID_aeroporto	Verificação de nulos
Nome	Nome	Verificação de nulos
Cidade	Cidade	Verificação de nulos
País	País	Verificação de nulos
Latitude	Latitude	Verificação de nulos
Longitude	Longitude	Verificação de nulos

5.2.4 Voo

Campo Fonte CSV	Campo Destino CSV	Transformações Aplicadas
FL_DATE	data_partida	Verificação de nulos, Alteração do nome

AIRLINE_CODE	nome_companhia_aerea	Verificação de nulos, Alteração do nome
FL_NUMBER	id_voo	Verificação de nulos, Alteração do nome
ORIGIN	aeroporto_origem	Verificação de nulos, Alteração do nome
DEST	aeroporto_destino	Verificação de nulos, Alteração do nome
DEP_TIME	hora_partida_real	Verificação de nulos, Alteração do nome
CRS_DEP_TIME	hora_partida_esperada	Verificação de nulos, Alteração do nome
ARR_TIME	hora_chegada_real	Verificação de nulos, Alteração do nome
CRS_ARR_TIME	hora_chegada_esperada	Verificação de nulos, Alteração do nome
CANCELLED	cancelado	Verificação de nulos, Alteração do nome
DIVERTED	desviado	Verificação de nulos, Alteração do nome
DISTANCE	distancia	Verificação de nulos, Alteração do nome
DELAY_DUE_WEATHER	atraso_meteorologia	Verificação de nulos, Alteração do nome
DELAY_DUE_CARRIER	atraso_companhia_aerea	Verificação de nulos, Alteração do nome
DELAY_DUE_NAS	atraso_SNA	Verificação de nulos, Alteração do nome
DELAY_DUE_SECURITY	atraso_seguranca	Verificação de nulos, Alteração do nome
DELAY_DUE_LATE_AIRCR AFT	atraso_voo_anterior	Verificação de nulos, Alteração do nome

5.2.5 Clima

Campo Fonte CSV	Campo Destino CSV	Transformações Aplicadas
temperature_2m	temperatura	Alteração do nome, cálculo da média

precipitation_probability	probabilidade	Alteração do nome, cálculo da média
precipitation	precipitacao	Alteração do nome, cálculo da média
cloud_cover	cobertura	Alteração do nome, cálculo da média
visibility	visibilidade	Alteração do nome, cálculo da média
wind_speed_180m	velocidade	Alteração do nome, cálculo da média
wind_direction_180m	direcao	Alteração do nome, cálculo da média

5.3 Modelação do Sistema de Povoamento

A modelação do sistema de povoamento foi realizada com base na notação BPMN (Business Process Model and Notation), facilitando a compreensão dos diversos fluxos de dados.

O modelo é composto por 5 tarefas principais:

1. **Ingestão:** Leitura dos ficheiros CSV (NiFi).
2. **Carga para FONTE:** Inserção dos dados brutos em tabelas *FONTE* no MySQL.
3. **Transformação:** Validação de campos, tratamento de nulos, normalização de nomes e cálculos da média de certos campos. Aqui são criadas as tabelas *RAW*, *ERRO*, e posteriormente os dados válidos são encaminhados para *RDY*.
4. **Criação de Histórico:** Registo das alterações com triggers, nas tabelas *AUD* e *HST*.
5. **Carga Final (DW):** Inserção nas tabelas de dimensão e povoamento da **tabela de factos**, com os campos integrados e derivados.

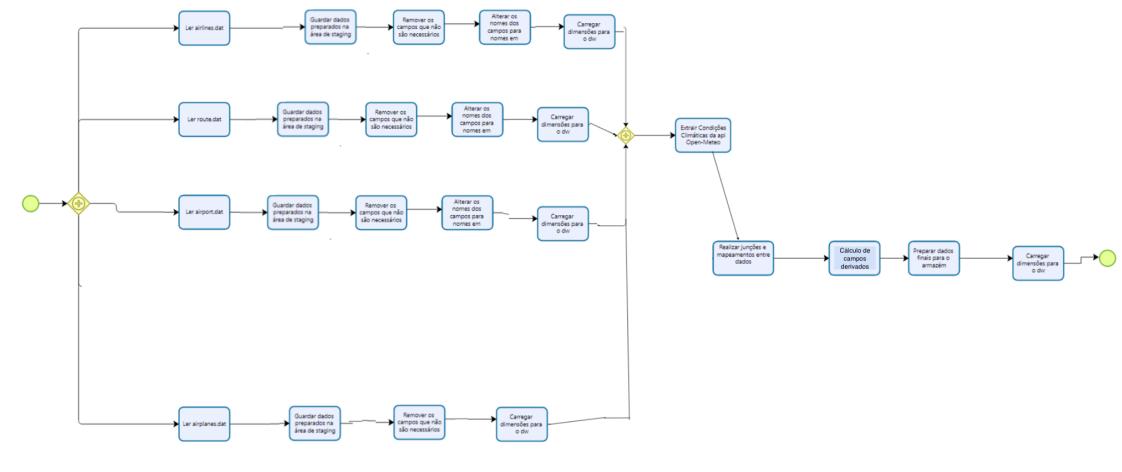
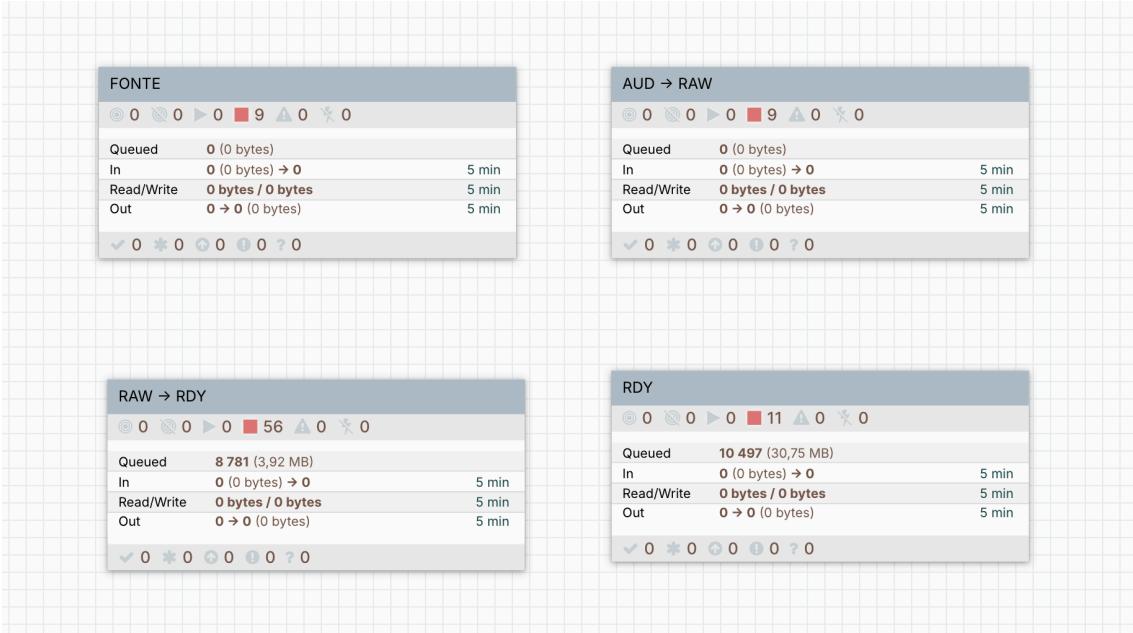


Figura 4 - BPMN para o mapeamento do fluxo de dados.



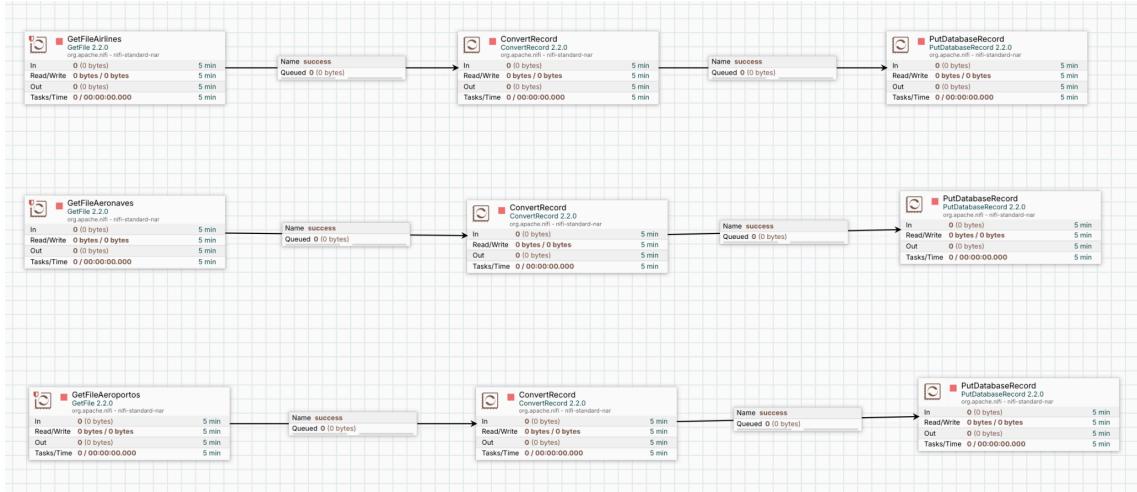
5.4 Implementação do Sistema de Povoamento

Após termos o esquema definido e os datasets escolhidos, procedeu-se à implementação do **Sistema de Povoamento**, composto por várias etapas articuladas que garantem a transformação progressiva dos dados desde os ficheiros de origem até à inserção final em tabelas analíticas e de histórico. Este sistema foi concebido com foco na rastreabilidade, controlo de qualidade e fiabilidade dos dados. Segue-se a descrição pormenorizada de cada fase:

1. Importação Inicial – De CSV para Tabela FONTE

A primeira etapa consiste na leitura dos ficheiros CSV contendo os dados brutos. Estes são importados diretamente para a **tabela FONTE**, que funciona como ponto de entrada do

sistema. Esta tabela reflete a estrutura original dos ficheiros, sem transformação ou validação, servindo como base para rastreamento e reprocessamento, caso necessário.

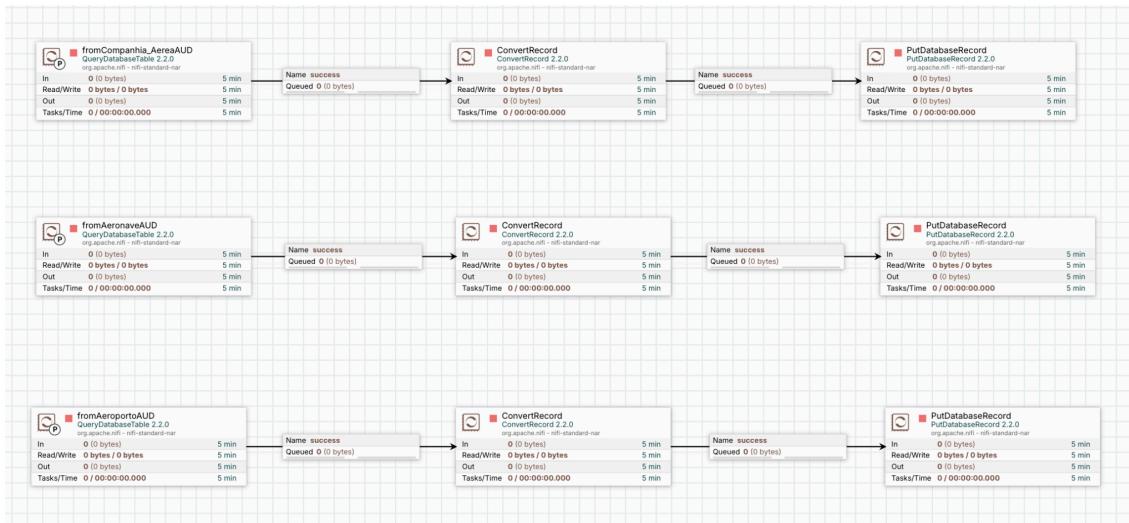


2. Registo de Operações com Triggers – Tabela AUD

Para garantir a **auditabilidade** das operações sobre a tabela FONTE, foram definidos **triggers** que registam automaticamente as ações de inserção, remoção e atualização de dados. Estes registo são armazenados na **tabela AUD**, permitindo manter um histórico completo de todas as alterações efetuadas na FONTE.

3. Transferência para RAW – Resiliência em Caso de Falha

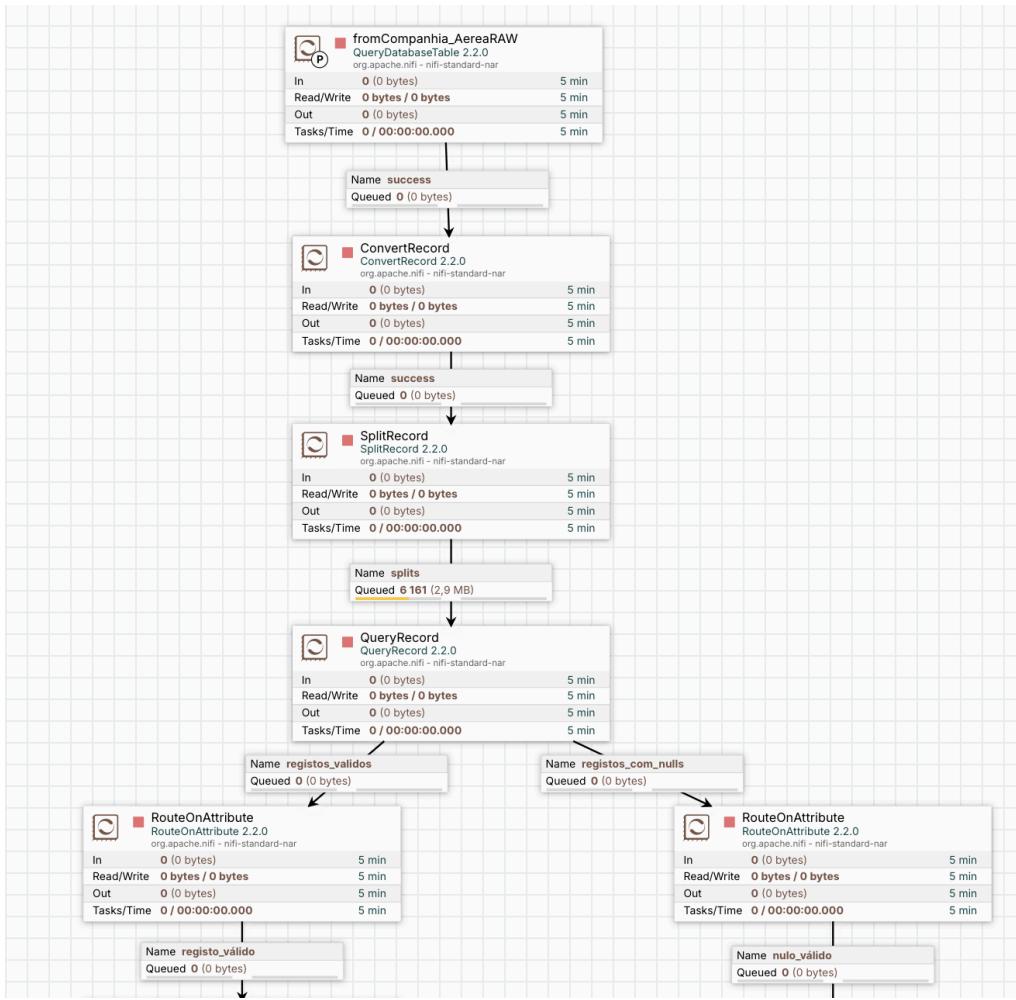
A partir dos registo da tabela AUD, procede-se à passagem dos dados para a **tabela RAW**, que constitui a primeira camada de staging no sistema de povoamento. Esta tabela armazena os dados com **mínima transformação**, funcionando como zona de buffer. A existência da RAW permite retomar o processo em caso de falha sem recorrer à reimportação dos ficheiros originais.

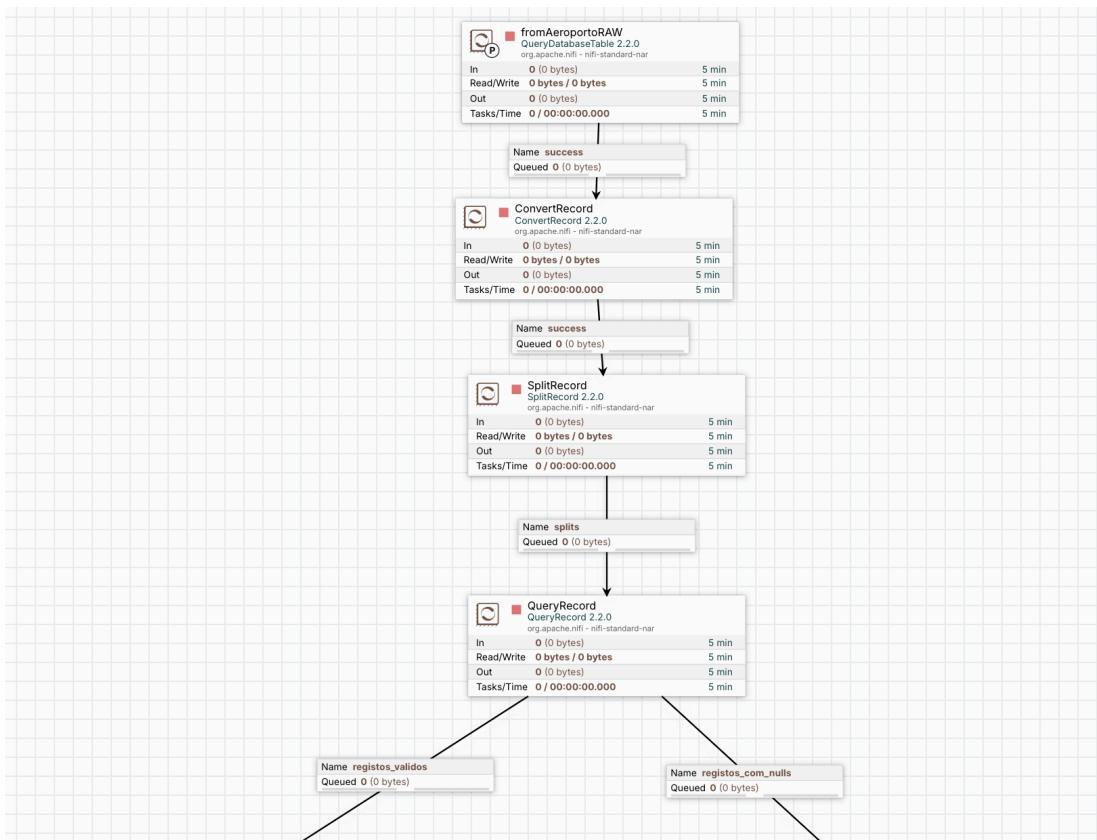
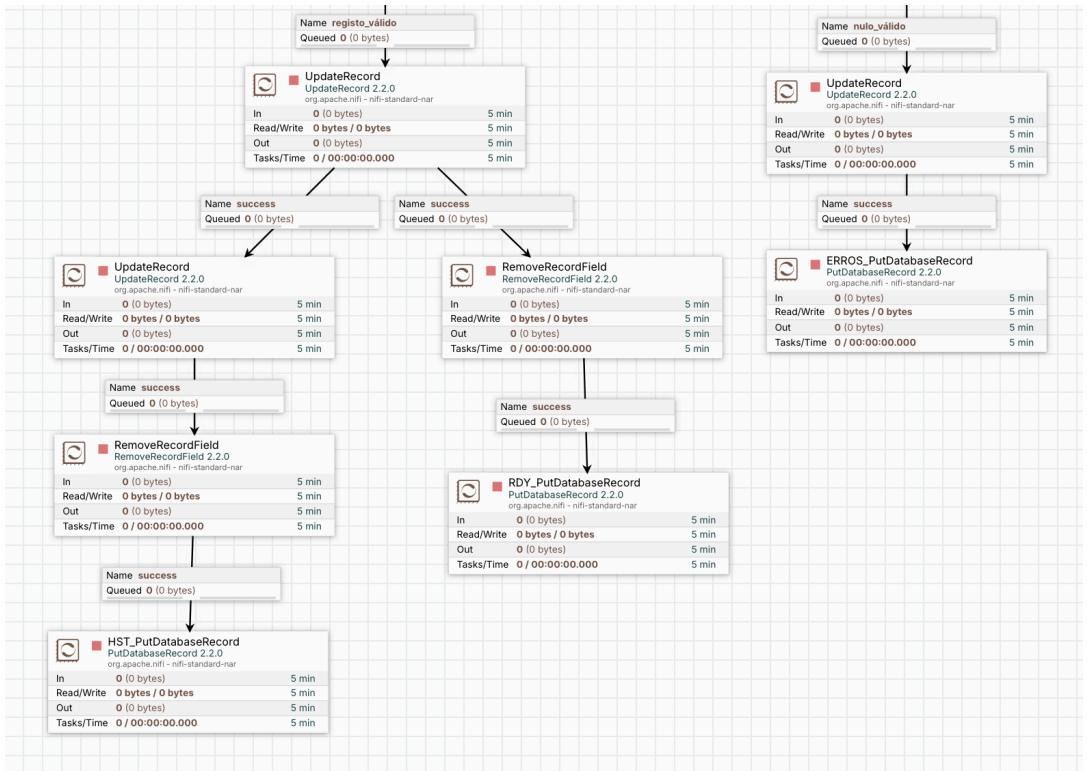


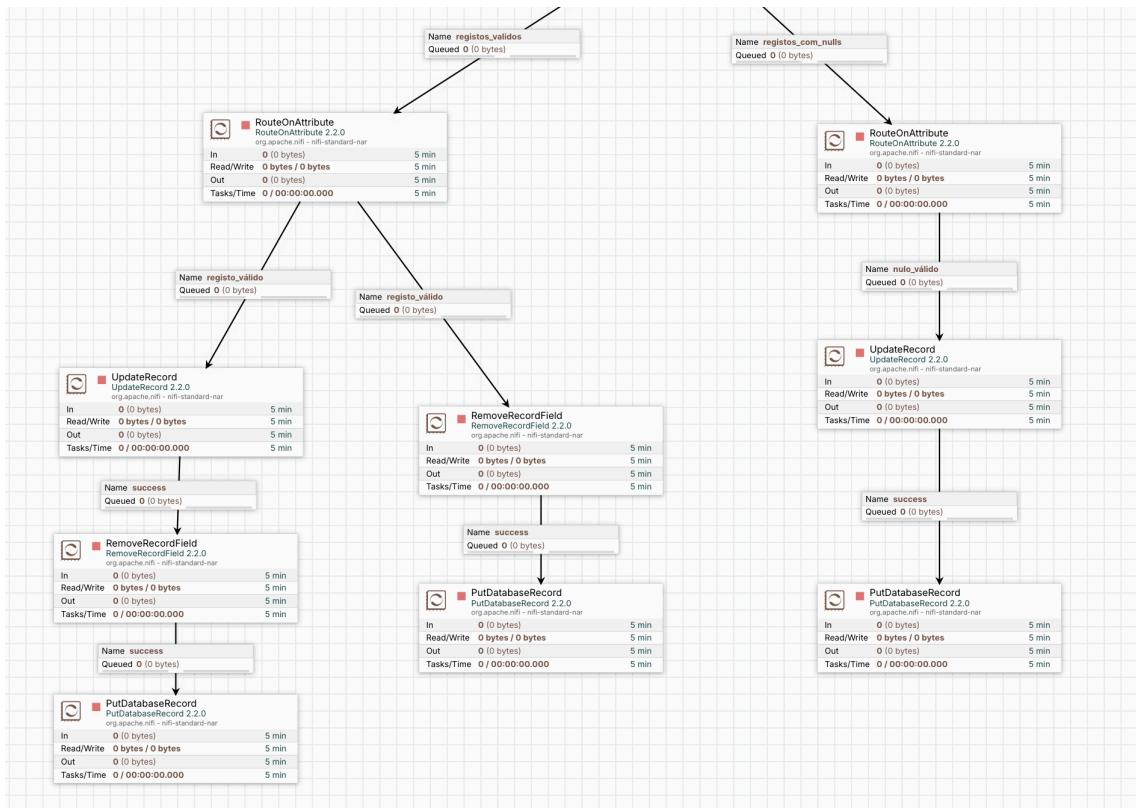
4. Tratamento de Companhia Aérea e Aeronaves

Os dados referentes a **companhias aéreas** e **aeronaves** passam por uma fase de limpeza e preparação:

- São realizadas **verificações de campos nulos** e **alterações de nomes dos campos para português**, por uma questão de normalização e coerência com o resto do sistema.
- Os registos **válidos** são inseridos na **tabela RDY**, que representa a camada pronta para análise.
- Simultaneamente, é mantido o histórico desses registos na **tabela HST**.
- Os registos com campos obrigatórios nulos ou inválidos são direcionados para a **tabela ERRO**, onde ficam disponíveis para revisão e eventual correção manual.



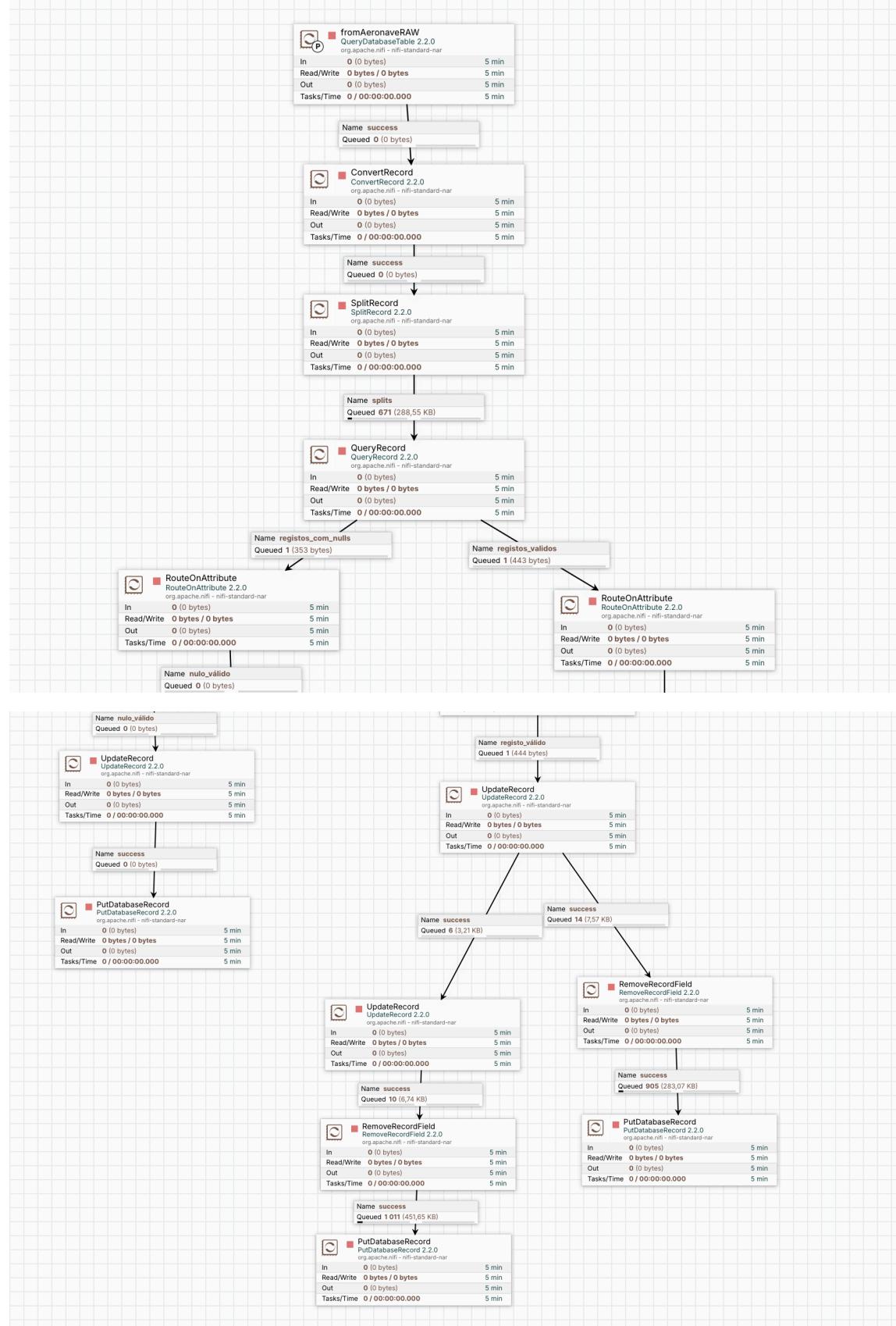




5. Validação de Dados das Aeronaves

As **aeronaves** seguem um processo semelhante ao das companhias aéreas:

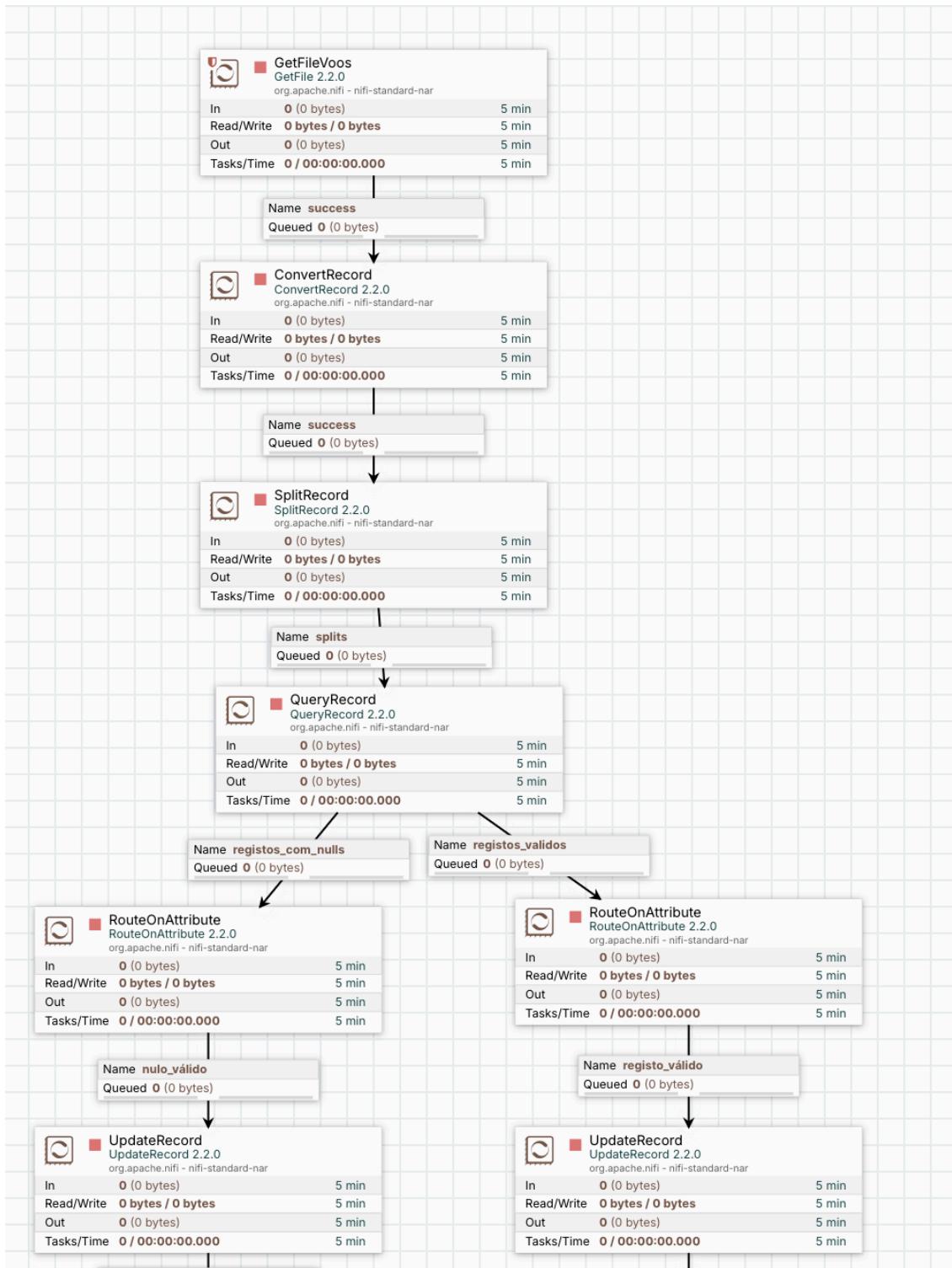
- Os dados são verificados quanto à presença de nulos nos campos críticos.
- Registos válidos são encaminhados para as tabelas **RDY** e **HST**.
- Registos com problemas são inseridos na tabela **ERRO**, assegurando que dados incorretos não prosseguem no pipeline.

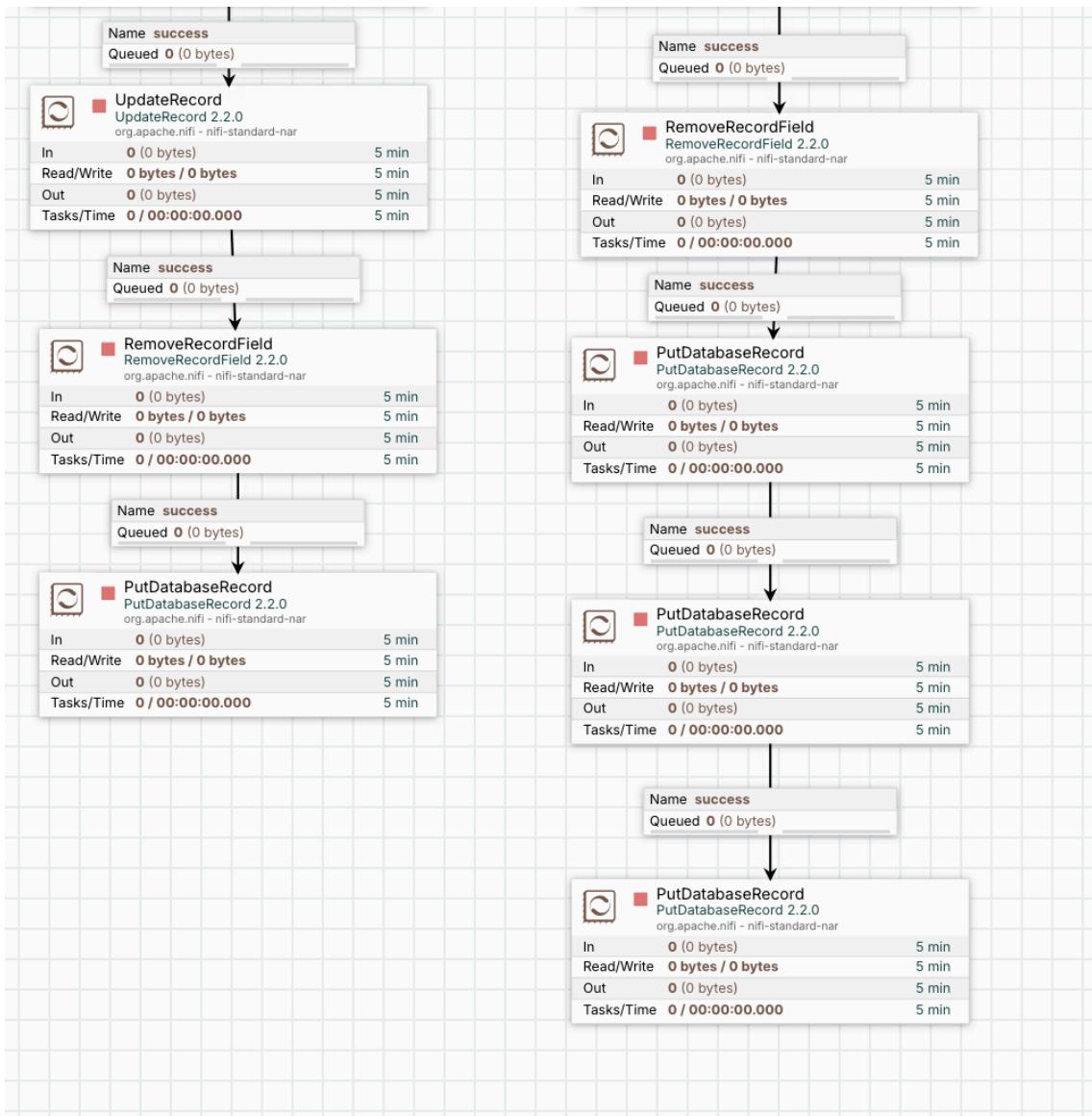


6. Validação e Tratamento dos Voos

No caso dos **voos**, o processo de validação também inclui:

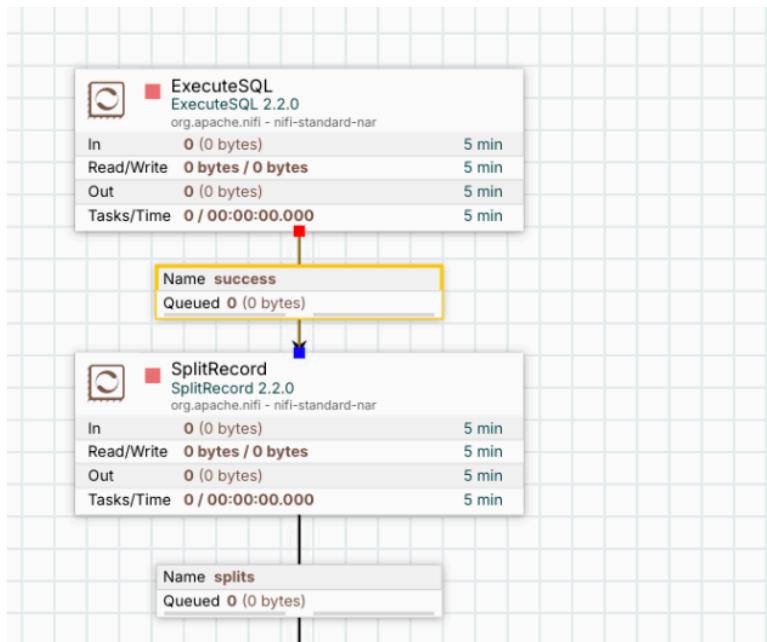
- Verificação de **valores nulos** e **alteração de nomes de campos** para manter consistência semântica no sistema.
- Após esta preparação, os dados são povoados nas respectivas tabelas de acordo com o seu estado (válido, histórico, erro).





7. Integração das Entidades – Enriquecimento por Join

Com todas as entidades principais devidamente preparadas e validadas nas tabelas **RDY**, realiza-se um **processo de join** entre os voos e as entidades relacionadas (companhias aéreas, aeronaves, etc.). Este passo tem como objetivo enriquecer os dados dos voos com toda a informação adicional disponível.



SELECT

```

v.id_voo,
v.data_partida,
v.nome_companhia_aerea,
v.avaliacao,
v.Modelo_Avião,
c.país,
v.aeroporto_origem,
v.aeroporto_destino,
v.hora_partida_real,
v.hora_partida_esperada,
v.hora_chegada_real,
v.hora_chegada_esperada,
v.cancelado,
v.desviado,
v.distancia,
```

```

v.atraso_companhia_aerea,
v.atraso_meteorologia,
v.atraso_SNA,
v.atraso_segurança,
v.atraso_voo_anterior,
a.Modelo,
a.Fabricante,
ap1.Nome AS nome_aeroporto_origem,
ap1.Cidade AS cidade_aeroporto_origem,
ap1.Pais AS pais_aeroporto_origem,
ap1.Latitude,
ap1.Longitude

FROM VooRDY v
JOIN Companhia_Aerea c ON v.nome_companhia_aerea = c.nome
JOIN Aeronave a ON v.Modelo_Avião = a.Modelo
JOIN Aeroporto ap1 ON v.aeroporto_origem = ap1.Código_IATA;


---

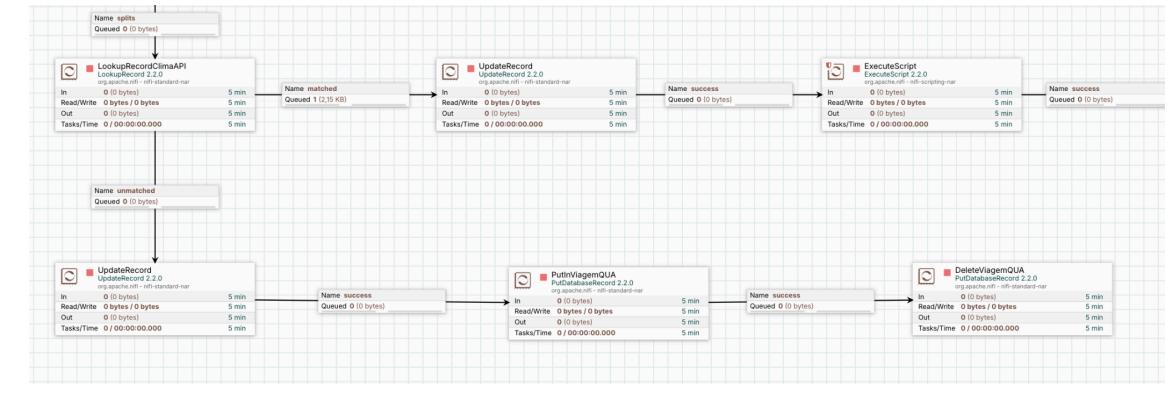

// _____ // _____

```

8. Enriquecimento com Dados Climáticos

Uma das etapas mais relevantes do sistema consiste na **integração dos dados climáticos**, associando:

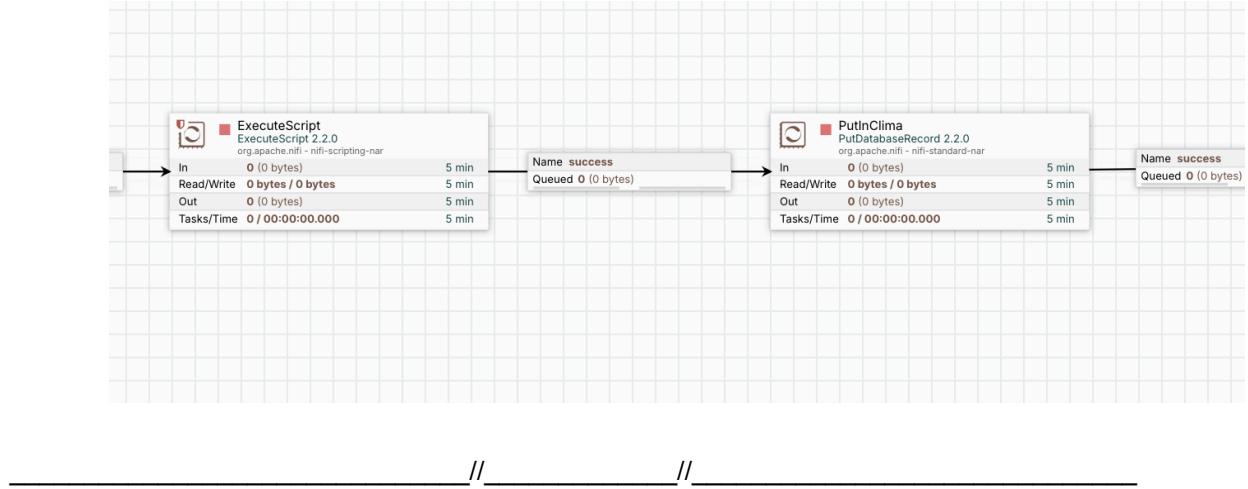
- As **coordenadas geográficas** dos aeroportos com a **data do voo**, de modo a obter as condições meteorológicas correspondentes.
- Nos casos em que **não há correspondência direta** com os dados meteorológicos, o registo é inserido na tabela **VooQUA**, permitindo a sua monitorização e eventual posterior tratamento.



9. Criação de Campos Derivados

Para enriquecer a análise, são executados **scripts de transformação** que criam novos campos, tais como:

- **Médias** de temperatura, precipitação, visibilidade e outros indicadores climáticos, ajudando a caracterizar o ambiente operacional de cada voo.



```

import org.apache.commons.io.IOUtils
import java.nio.charset.StandardCharsets
import groovy.json.JsonSlurper
import groovy.json.JsonOutput

def media(lista) {

    if (!(lista instanceof List)) return null

    def validos = lista.findAll { it instanceof Number }

    return validos ? validos.sum() / validos.size() : null
}
  
```

```

flowFile = session.get()

if (flowFile != null) {

    try {

        def inputStream = session.read(flowFile)

        def conteudo = IOUtils.toString(inputStream,
StandardCharsets.UTF_8)

        inputStream.close()

        def json = new JsonSlurper().parseText(conteudo)

        // Se json for lista com um elemento, pega o elemento para
simplificar

        if (json instanceof List && json.size() == 1 && json[0]
instanceof Map) {

            json = json[0]

        }

        // Substitui os arrays pelas médias mantendo os demais campos
intactos

        if (json.temperatura != null) json.temperatura =
media(json.temperatura)

        if (json.precipitacao != null) json.precipitacao =
media(json.precipitacao)

        if (json.probabilidade != null) json.probabilidade =
media(json.probabilidade)

        if (json.visibilidade != null) json.visibilidade =
media(json.visibilidade)

        if (json.cobertura != null) json.cobertura =
media(json.cobertura)

        if (json.velocidade != null) json.velocidade =
media(json.velocidade)

        if (json.direcao != null) json.direcao = media(json.direcao)

        def novoConteudo = JsonOutput.toJson(json)

        flowFile = session.write(flowFile, { outputStream ->

            outputStream.write(novoConteudo.getBytes(StandardCharsets.UTF_8))
        })
    }
}

```

```

        } as OutputStreamCallback)

    session.transfer(flowFile, REL_SUCCESS)

} catch (e) {

    log.error("Erro ao processar JSON: " + e.message, e)

    session.transfer(flowFile, REL_FAILURE)

}

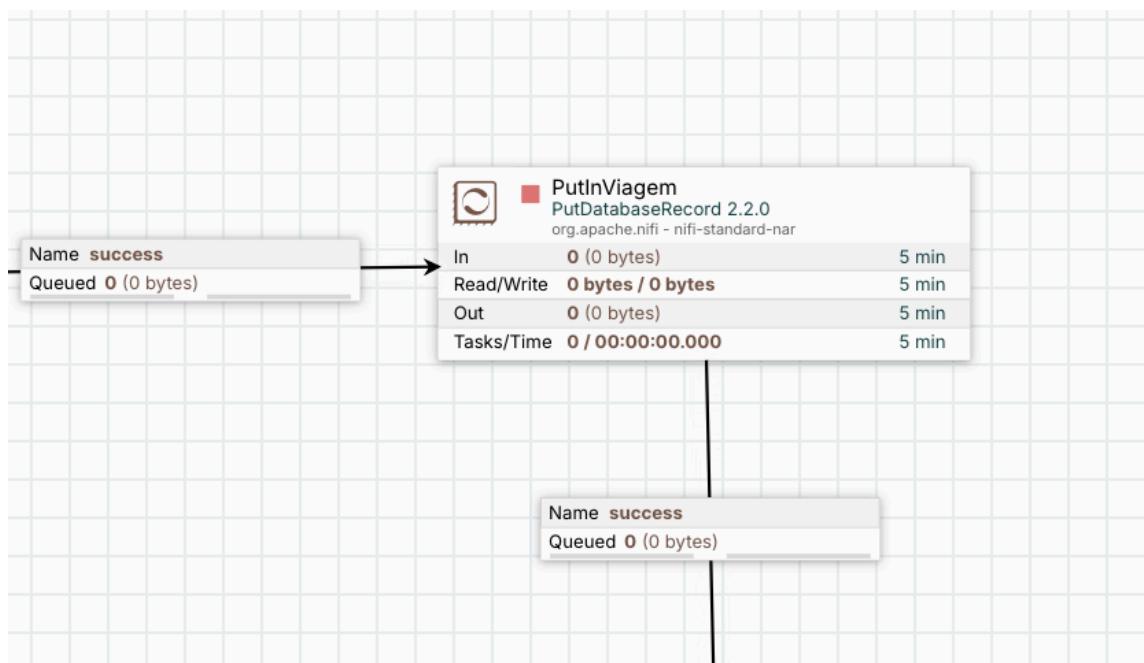
}

```

10. Povoamento das Tabela Viagem

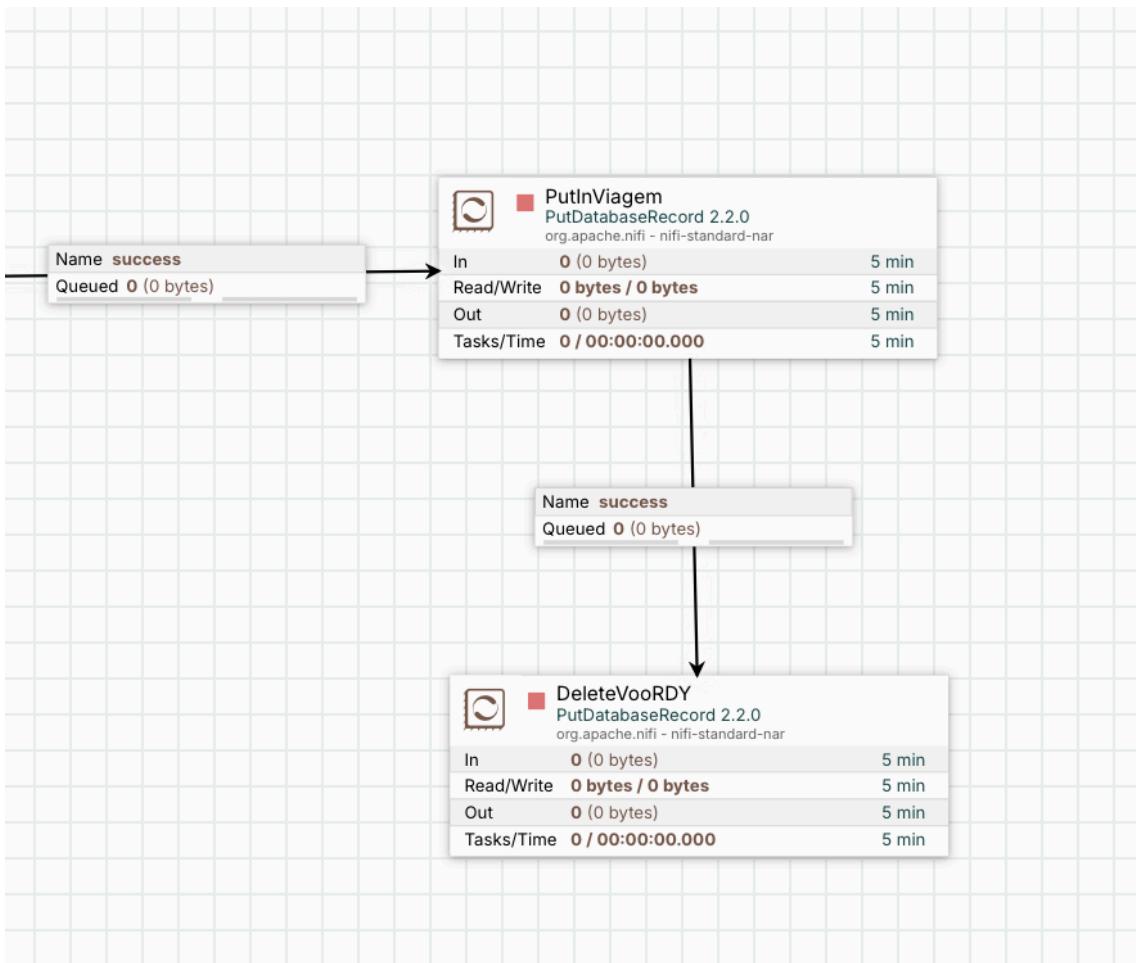
Após a consolidação de todas as dimensões envolvidas, procede-se ao **povoamento da tabela de factos e dimensão:**

- **Viagem**, que agrupa as informações de operação, clima e entidades envolvidas, numa estrutura otimizada para análise multidimensional.



11. Limpeza Pós-Povoamento

Por fim, é efetuada a **remoção do registo de voo na tabela VooRDY**, após a sua inserção com sucesso nas tabelas **Viagem** e **Voo**. Este passo evita duplicações e mantém a integridade do sistema, garantindo que apenas registos não processados continuam visíveis na tabela RDY.



5.5 Validação e Testes

Após a preparação de dados com a ferramenta Apache NiFi deparamo-nos com algumas complicações. Os nossos datasets com mais de 1000 registos, por exemplo o dataset dos aeroportos ou o dos Vooos que contém 3 milhões de registo, não estava a ser guardado na sua totalidade na base de dados. Mesmo alterando o número máximo de lote de inserção do NiFi nas tabelas e de “transporte” entre processos o NiFi apenas guardava 999 registos na base de dados, o que posteriormente levou o resultado da junção de tabelas ser pouco eficiente, uma vez que as tabelas no MySQL não tinham a informação necessária para fazer o JOIN das tabelas.

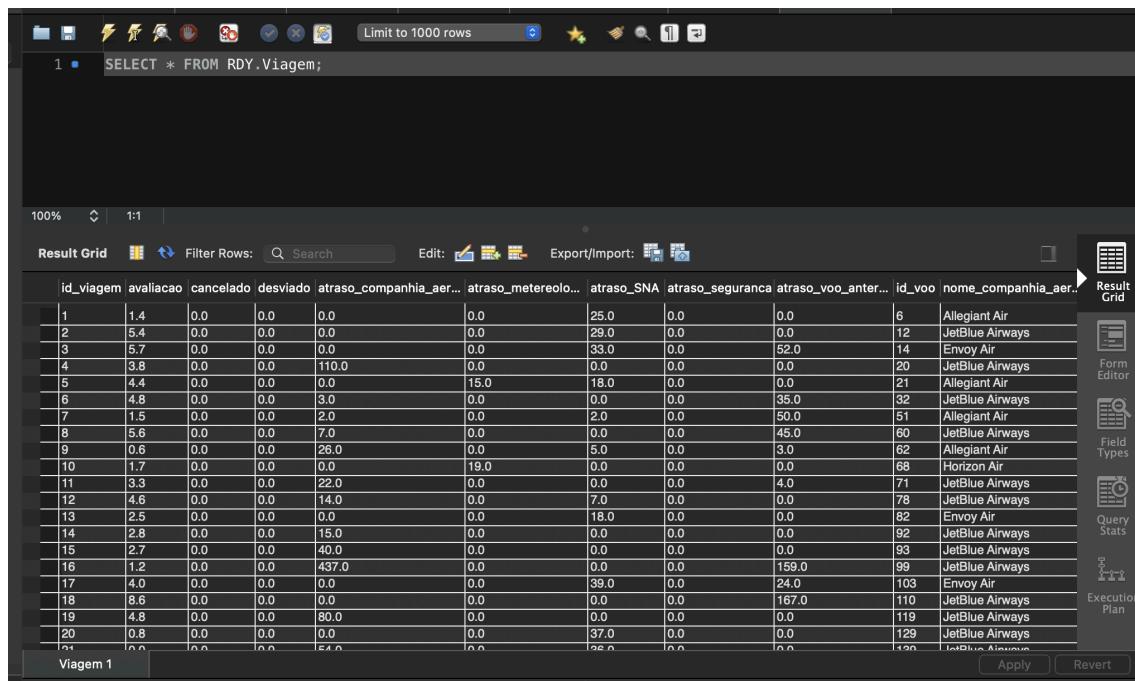
O povoamento da tabela de factos revelou-se um dos maiores desafios no desenvolvimento do sistema, principalmente devido a limitações operacionais do Apache NiFi. Esta ferramenta, embora poderosa na integração e transformação de dados, não oferece suporte direto para operações complexas de junção (JOIN). Esta limitação impediu a execução direta de JOINs entre as tabelas dimensionais já construídas e os dados temporários da pipeline em NiFi, o que seria necessário para preencher algumas colunas da tabela de factos.

caso estas não fossem atributos descritivos. Tivemos então neste ponto que simplificar o esquema lógico e dimensional.

Uma outra solução considerada pelo grupo foi a criação de tabelas intermediárias na base de dados. Essas tabelas atuariam como repositórios temporários, permitindo que os dados fossem primeiro carregados individualmente. No entanto, optámos por não implementar esta solução, pois considerámos que introduziria redundância de dados e aumentaria a complexidade do sistema. Além disso, a criação e manutenção destas tabelas intermediárias, num cenário real, exigiria mais recursos e dificultaria o controlo da integridade e consistência dos dados ao longo do tempo.

Tivemos também problemas com datasets que tinham informação duplicada quando procuramos os aeroportos pelo nome mas que no dataset tinham código_IATA diferentes.

No entanto e apesar de todos os problemas indicados acima foi possível fazer o povoamento da tabela de factos.



The screenshot shows a database interface with a query editor at the top containing the SQL command: `SELECT * FROM RDY.Viagem;`. Below the editor is a result grid titled "Result Grid". The grid displays data from the "Viagem" table with the following columns: id_viagem, avaliacao, cancelado, desviado, atraso_companhia_aer, atraso_meteorolo..., atraso_SNA, atraso_seguranca, atraso_voo_anter..., id_voo, nome_companhia_aer. The data consists of approximately 20 rows of flight-related statistics and airline names. To the right of the grid is a vertical toolbar with icons for Result Grid, Form Editor, Field Types, Query Stats, and Execution Plan. At the bottom right of the grid are buttons for "Apply" and "Revert".

	id_viagem	avaliacao	cancelado	desviado	atraso_companhia_aer...	atraso_meteorolo...	atraso_SNA	atraso_seguranca	atraso_voo_anter...	id_voo	nome_companhia_aer
1	1.4	0.0	0.0	0.0	0.0	25.0	0.0	0.0	0.0	6	Allegiant Air
2	5.4	0.0	0.0	0.0	0.0	29.0	0.0	0.0	0.0	12	JetBlue Airways
3	5.7	0.0	0.0	0.0	0.0	33.0	0.0	52.0	0.0	14	Envoy Air
4	3.8	0.0	0.0	110.0	0.0	0.0	0.0	0.0	0.0	20	JetBlue Airways
5	4.4	0.0	0.0	0.0	15.0	18.0	0.0	0.0	0.0	21	Allegiant Air
6	4.8	0.0	0.0	3.0	0.0	0.0	0.0	35.0	0.0	32	JetBlue Airways
7	1.5	0.0	0.0	2.0	0.0	2.0	0.0	50.0	0.0	51	Allegiant Air
8	5.6	0.0	0.0	7.0	0.0	0.0	0.0	0.0	45.0	60	JetBlue Airways
9	0.6	0.0	0.0	26.0	0.0	5.0	0.0	0.0	3.0	62	Allegiant Air
10	1.7	0.0	0.0	0.0	19.0	0.0	0.0	0.0	0.0	68	Horizon Air
11	3.3	0.0	0.0	22.0	0.0	0.0	0.0	0.0	4.0	71	JetBlue Airways
12	4.6	0.0	0.0	14.0	0.0	7.0	0.0	0.0	0.0	78	JetBlue Airways
13	2.5	0.0	0.0	0.0	0.0	18.0	0.0	0.0	0.0	82	Envoy Air
14	2.8	0.0	0.0	15.0	0.0	0.0	0.0	0.0	0.0	92	JetBlue Airways
15	2.7	0.0	0.0	40.0	0.0	0.0	0.0	0.0	0.0	93	JetBlue Airways
16	1.2	0.0	0.0	437.0	0.0	0.0	0.0	159.0	0.0	99	JetBlue Airways
17	4.0	0.0	0.0	0.0	0.0	39.0	0.0	24.0	0.0	103	Envoy Air
18	8.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	167.0	110	JetBlue Airways
19	4.8	0.0	0.0	80.0	0.0	0.0	0.0	0.0	0.0	119	JetBlue Airways
20	0.8	0.0	0.0	0.0	0.0	37.0	0.0	0.0	0.0	129	JetBlue Airways
21	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	130	JetBlue Airways

Concluímos portanto, num caso real, os datasets usados encontravam-se incompletos, não contendo a informação necessária e suficiente para o povoamento completo da tabela de factos.

Tendo isto em consideração, nas fases futuras optamos por utilizar diretamente os csvs, após um breve tratamento, para podermos usar um maior conjunto de dados, permitindo então uma análise mais significativa, e recomendações melhores.

6. Exploração e Análise de Dados

6.1 Organização geral do sistema de dashboarding

O sistema de visualização foi desenvolvido integralmente no Power BI, com base no ficheiro de dados consolidado **voos_completos_clima.csv**. A informação foi previamente tratada no Power Query para garantir a tipagem correta das colunas e a integridade das métricas calculadas, nomeadamente os campos de atraso, cancelamento e desvio.

A estrutura dos dashboards, do ponto de vista da análise por aeroporto, foi organizada segundo duas perspectivas principais:

- **Aeroportos de Origem**
- **Aeroportos de Destino**

Cada painel analítico explora diferentes aspectos da operação aeroportuária, com foco nos atrasos, cancelamentos e voos desviados. Os visuais incluem gráficos de linha, barras empilhadas, colunas e área, para permitir análise comparativa e temporal. Os valores apresentados são agregados com base em funções **SUM()** e **COUNT()** aplicadas diretamente no Power BI, sem necessidade de cálculos externos.

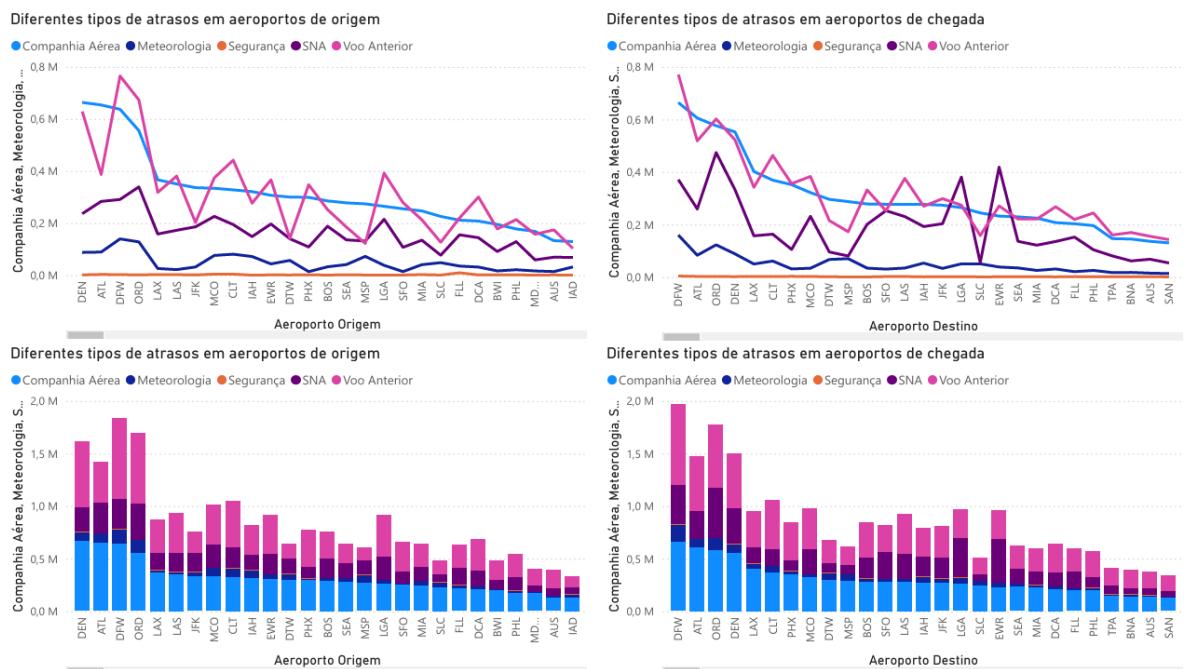


Figura 5 - Análise dos diferentes tipos de atrasos em aeroportos de origem e destino.

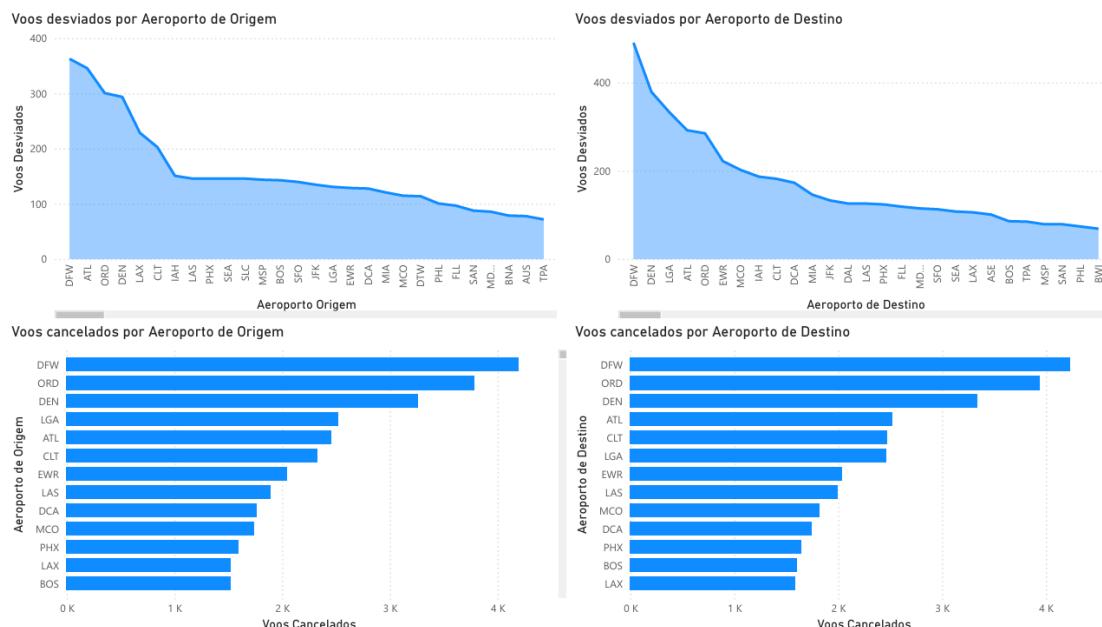


Figura 6 - Distribuição dos voos desviados e cancelados por aeroportos de origem e destino.

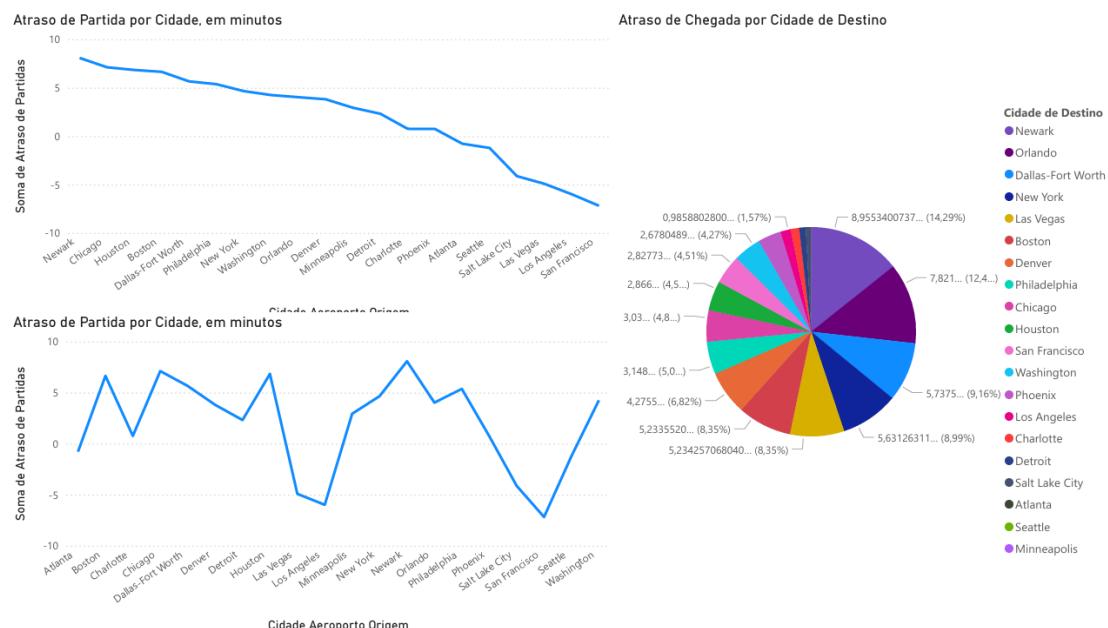


Figura 7 - Atrasos de voos de partida e de chegada por cidade.

Complementando a perspetiva anterior, foi desenvolvida uma análise detalhada focada nas companhias aéreas, com o objetivo de identificar padrões de desempenho e comportamentos operacionais. Os painéis analíticos desenvolvidos incluem:

- Gráficos de barras com a média e soma dos atrasos por companhia aérea;
- Gráfico de barras com a percentagem de voos com atrasos significativos (mais do que 30 minutos), por companhia aérea;
- Gráfico de barras vertical com a percentagem de voos cancelados por companhia aérea;
- Tabela resumo com a soma total e média de atrasos, percentagem de voos com atraso significativo, percentagem de voos cancelados e um total de voos registados por companhia aérea.
- Gráfico de barras horizontal com os atrasos médios por modelo de avião, segmentando por companhia aérea, com um filtro lateral para seleção da companhia.

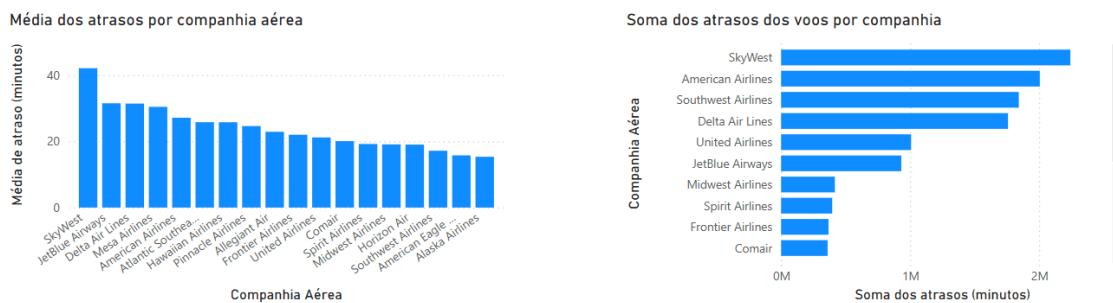


Figura 8 - Média e soma dos atrasos por companhia aérea

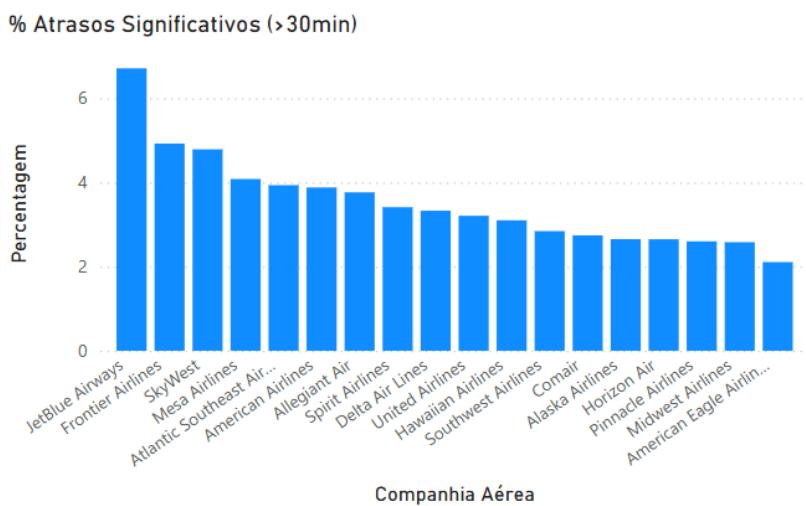


Figura 9 - Percentagem de atrasos significativos(>30 minutos) por companhia aérea

Percentagem de voos cancelados por Companhia Aérea

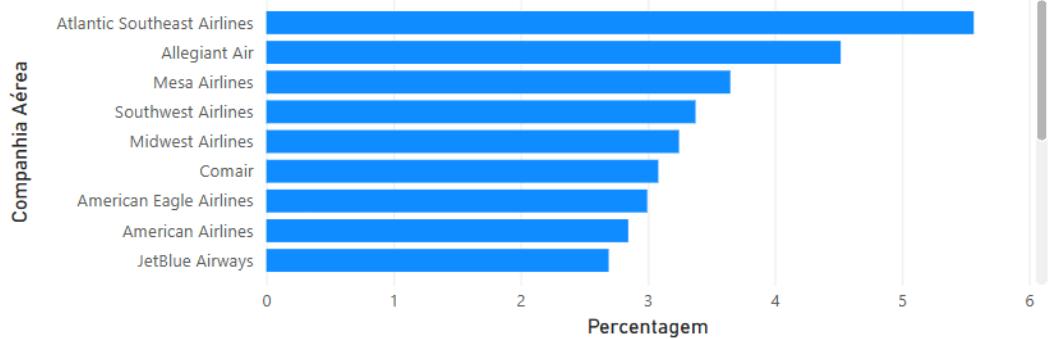


Figura 10 - Percentagem de voos cancelados por Companhia Aérea

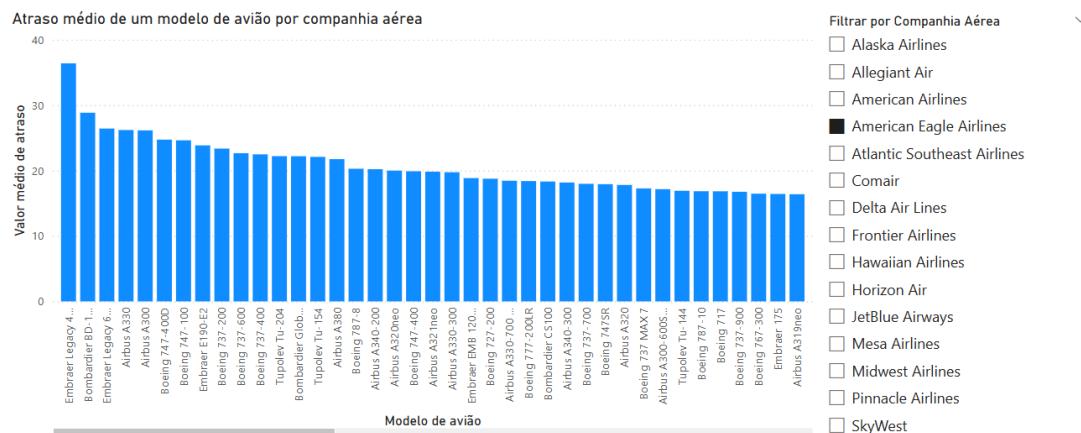


Figura 11 - Gráfico de barras com a percentagem de voos cancelados por Companhia Aérea

Dados gerais das companhias aéreas

Companhia Aérea	Soma dos atrasos (minutos)	Média dos atrasos (minutos)	% de atrasos significativos (>30%)	% voos cancelados	Total de voos
JetBlue Airways	931.321,00	31,51	6,71	2,69	112844
Frontier Airlines	367.576,00	21,99	4,93	2,58	64466
SkyWest	2.241.650,00	42,12	4,79	2,25	343737
Mesa Airlines	354.100,00	30,43	4,08	3,65	65012
Atlantic Southeast Airlines	99.441,00	25,79	3,94	5,57	19082
American Airlines	2.004.811,00	27,14	3,88	2,85	383106
Allegiant Air	299.695,00	22,89	3,77	4,52	52738
Spirit Airlines	396.139,00	19,22	3,41	2,42	95711
Delta Air Lines	1.758.471,00	31,42	3,33	1,51	395239
United Airlines	1.007.314,00	21,18	3,21	2,18	254504
Hawaiian Airlines	133.772,00	25,77	3,10	1,21	32114
Southwest Airlines	1.841.977,00	17,17	2,84	3,38	576470
Comair	360.426,00	20,07	2,74	3,08	107050
Alaska Airlines	274.241,00	15,35	2,65	1,93	100467
Horizon Air	55.874,00	19,02	2,65	1,81	20634
Pinnacle Airlines	352.619,00	24,62	2,60	2,13	112463
Midwest Airlines	416.615,00	19,08	2,58	3,25	143107
American Eagle Airlines	321.918,00	15,78	2,11	3,00	121256

Figura 12 - Tabela com os dados gerais das companhias aéreas sobre atrasos e voos cancelados

6.2 Serviços de exploração e análise implementados

Foram desenvolvidos vários serviços de exploração e análise de dados com o objetivo de fornecer suporte às decisões operacionais e estratégicas dos diferentes agentes envolvidos. Estes serviços foram implementados na plataforma **Power BI**, com ligação direta ao Data Warehouse, permitindo uma visualização interativa e dinâmica dos indicadores.

A exploração dos dados focou-se nas seguintes áreas principais:

1. Análise de Atrasos

- Atraso médio por **modelo de aeronave**, com possibilidade de filtragem por companhia aérea.
- Atrasos médios por **companhia aérea**, com comparação entre operadoras.
- Atraso médio por **aeroporto de origem e de destino**, analisado por cidade e terminal.
- Identificação das **principais causas de atraso**: meteorologia, segurança, voo anterior, gestão aeroportuária.

2. Cancelamentos e Voos Desviados

- Percentagem de voos cancelados e desviados por companhia aérea.
Volume absoluto de cancelamentos e desvios por aeroporto de origem e destino.

3. Atrasos Significativos

- Percentagem de voos com atrasos superiores a 30 minutos, por companhia aérea.
- Identificação de operadoras com maior impacto negativo na experiência do passageiro.

4. Comparação de Desempenho

- Tabelas resumo com indicadores agregados por companhia aérea:
 - Soma total dos atrasos
 - Média de atraso
 - % de atrasos >30 min
 - % de cancelamentos
 - Total de voos

7. Caracterização de Perfis de Clientes

Devido aos problemas de inserção dos datasets com grandes quantidades de registo com o Apache NiFi, o grupo decidiu criar um ficheiro python que faz o mesmo que faz a junção da informação necessária à análise com os datasets reais encontrados.

Como o acesso à API OpenMeteo apenas é realizado no NiFi, tivemos de criar código que através de uma lógica relacionada com o tipo de atraso, atribuí-se valores aos campos do clima.

7.1 Definição do problema e compreensão dos elementos de análise envolvidos

O principal objetivo desta análise é identificar diferentes perfis de aeroportos com base no seu atrasos e outras condicionais. Através da segmentação de voos, pretende-se possibilitar uma personalização mais eficaz de organização e uma melhoria das estratégias de marketing.

Serão considerados elementos como companhias aéreas, aeronaves, aeroportos e condições climatéricas. A resolução do problema passa pela aplicação de técnicas de clustering que permitam agrupar voos com características similares, revelando assim padrões de comportamento relevantes.

7.2 Seleção e preparação dos dados

A preparação dos dados para o profiling e sistema de recomendação incluiu diversas etapas essenciais para garantir a qualidade e a coerência dos dados para a análise e modelagem. Inicialmente, foram tratados os voos cancelados e desviados assim como os valores nulos encontrados no conjunto de dados. Verificamos que tínhamos um dataset com 2.64% de voos que foram cancelados e 0.24% de voos que tiveram de ser desviados. Em relação aos dados nulos, uma vez que os datasets não foram criados por nós verificamos que grande parte dos dados tinham valores a null, nomeadamente os campos que indicam o tipo de atraso, que em 3 milhões de registo do dataset, 82.02% encontravam-se a null. Os nulls em questão não tiveram impacto pois decidimos descartar as colunas destes dados pois apesar de fornecerem um contexto que poderia ser importante, são informações que podemos calcular de forma menos detalhada. Serve relembrar que não teríamos estes nulls caso estivéssemos a usar a informação diretamente do datawarehouse.

Em seguida decidimos focar na análise de dados do ponto de vista de um aeroporto, ou seja, escolhemos fazer o clustering e os sistemas de recomendação com base nos

aeroportos. Para isso foi realizado um processo de feature engineering para a criação das variáveis relacionadas com os aeroportos que seriam usadas posteriormente para a análise, profiling e criação dos clusters, guardadas no dataframe **airport_data**. Estas features vão ser formadas a partir das viagens em que o aeroporto é origem ou destino, e para isso fizemos o somatório ou a média de diversos dados.

Ainda sobre Feature Engineering, foram utilizadas RFM Features, onde o objetivo encontrava-se em:

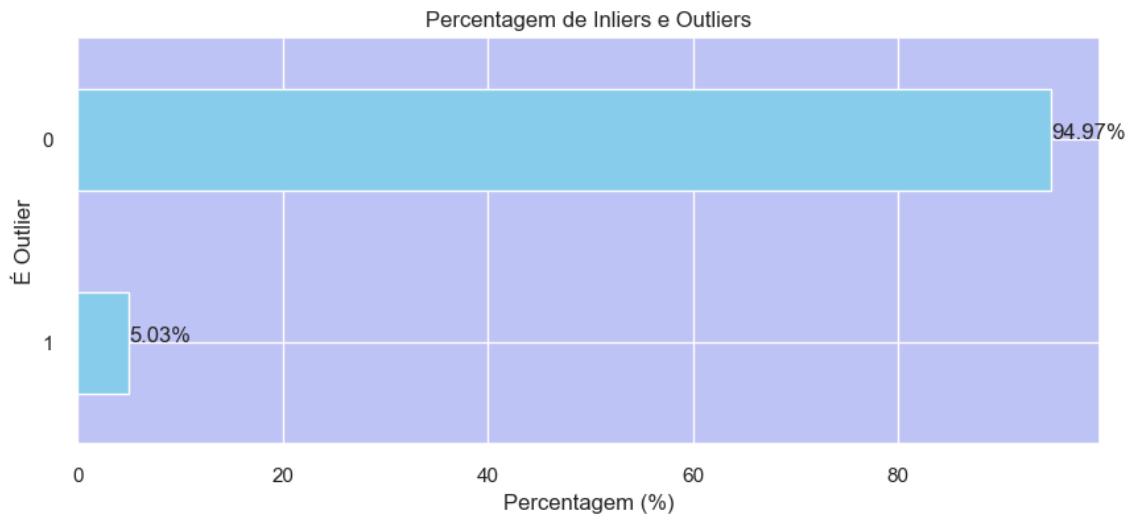
- **Recency** - usar a **recência** para analisar quão recente foi o último atraso.
- **Frequency** - A frequência permite-nos descobrir se os atrasos são acontecimentos recorrentes.
- **Monetary** - Qual é a quantidade total de atrasos

Foi também realizado o cálculo dos unique flights, número de rotas distintas, distância média de voo, número de companhias aéreas distintas, avaliação média das viagens, condições climatéricas médias e features de análise de comportamento, como por exemplo o dia da semana com mais atrasos, mês e o número de dias médio entre atrasos.

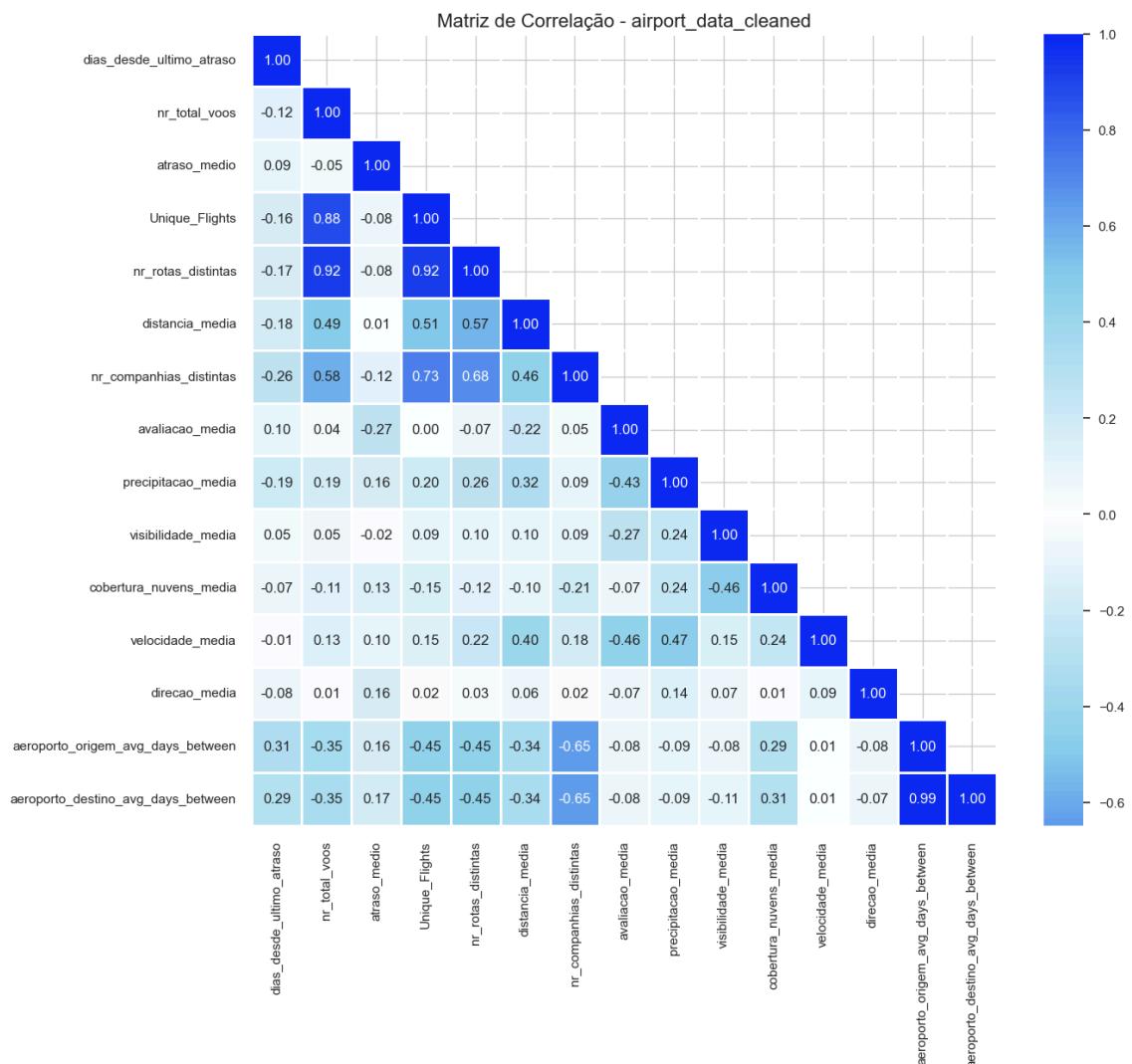
Essas etapas de tratamento e engenharia de features foram fundamentais para garantir que os dados refletissem corretamente os padrões operacionais, proporcionando uma base sólida para o desenvolvimento e validação dos sistemas de recomendação. Adicionalmente, foram incluídas variáveis categóricas representativas dos clusters operacionais, que facilitam a segmentação e filtragem durante a recomendação.

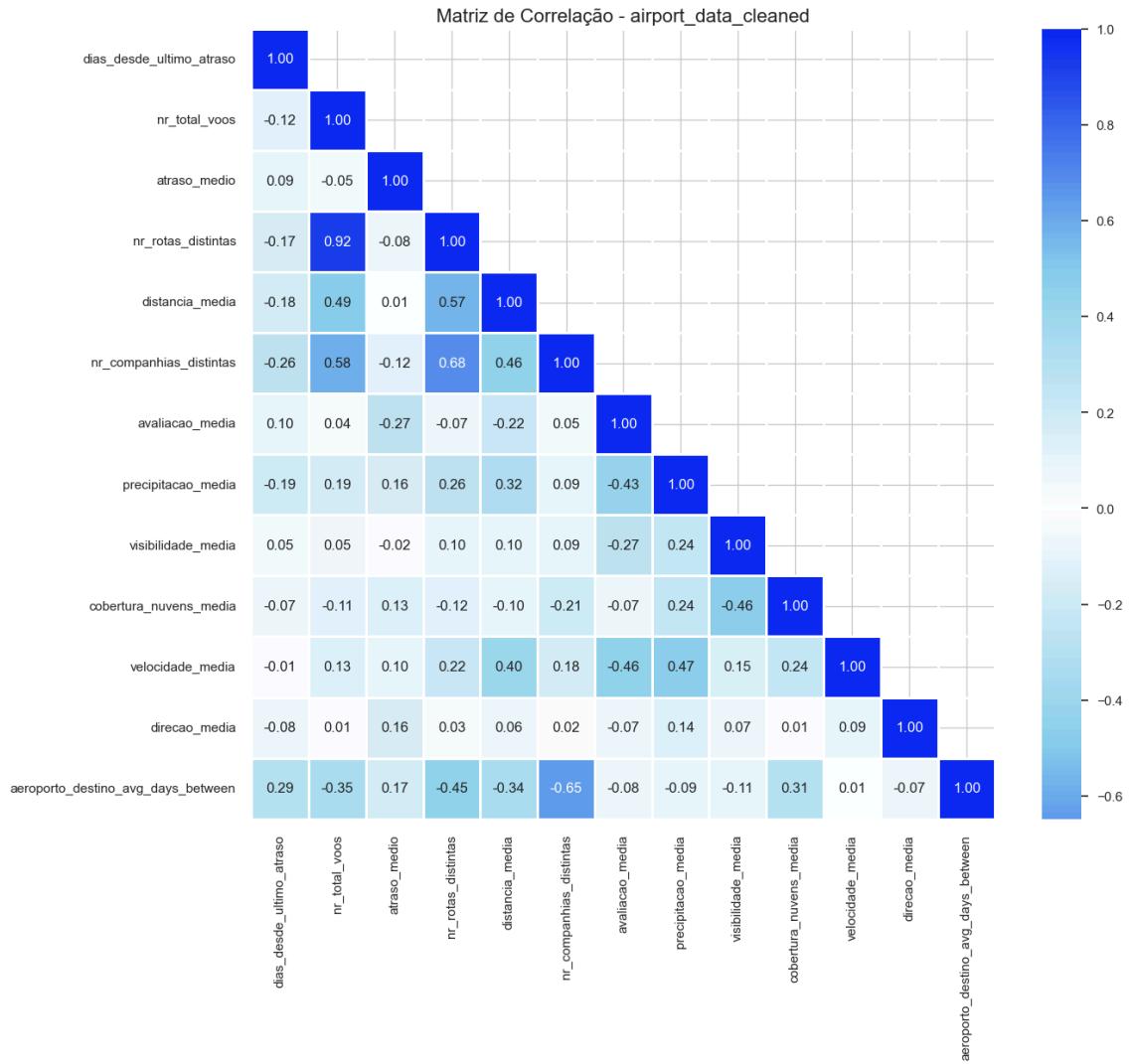
7.3 Identificação e fundamentação da técnica de análise

Para a criação dos clusters, foi realizada a **deteção e tratamento de outliers** para remover valores atípicos que poderiam distorcer os resultados dos clusters.



Verificamos que não temos muitos outliers no dataset depois de limpo. Além disso, procedeu-se a uma **análise de correlação** entre as variáveis para identificar relações relevantes e evitar redundâncias no conjunto de dados.





Após a realização da primeira matriz de correlação rapidamente descobrimos que havia 5 features com correlação demasiado alta entre elas, e as mesmas diminuem a qualidade do modelo a ser desenvolvido, por isso decidimos apagar duas delas. A primeira era a feature aeroporto_origem_avg_days_between que tinha grande correlação com aeroporto_destino_avg_days_between e isto indica que a frequência entre atrasos de voos chegados é semelhante aos atrasos de voos a partir. A outra feature descartada foi o unique_flights que apresenta grande correlação com o número de rotas distintas e com o número total de voos.

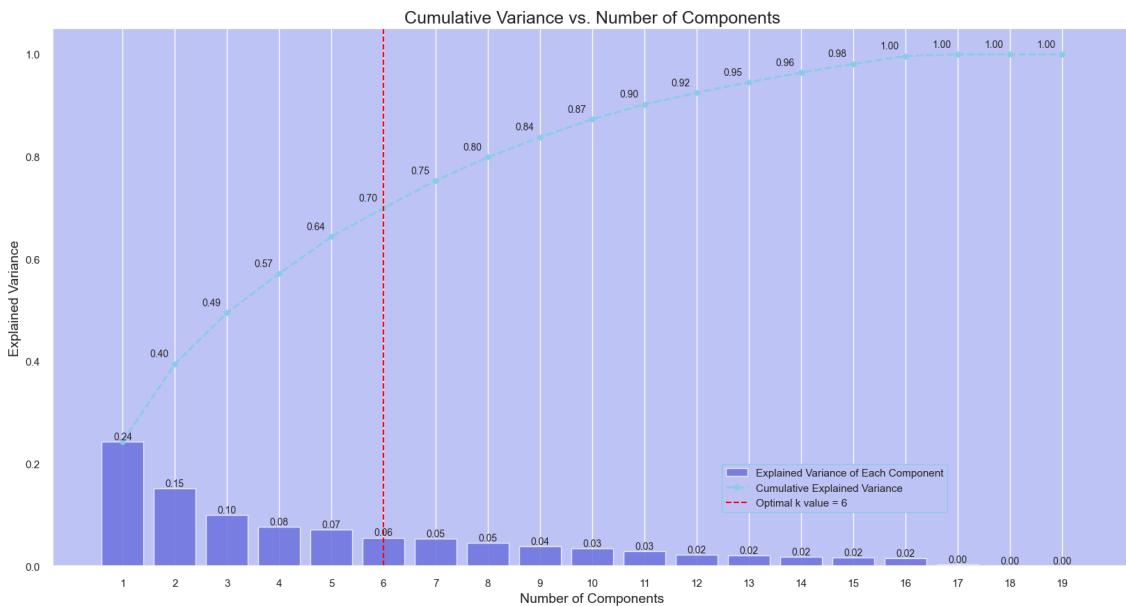
A técnica selecionada foi o **K-Means Clustering**, por se adequar bem à segmentação de grandes volumes de dados e à identificação de agrupamentos naturais entre clientes. Esta técnica permite agrupar clientes em segmentos com base em distâncias euclidianas entre variáveis num espaço multidimensional.

A escolha baseou-se na sua simplicidade, eficácia e capacidade de adaptação a diferentes perfis de dados.

7.4 Construção do modelo de análise

Na etapa de preparação dos dados, as variáveis numéricas foram normalizadas utilizando o **StandardScaler** e **Cyclic Scaling**, garantindo que todas as features tivessem escala comparável. Para as variáveis cíclicas, como hora do dia ou dia do ano, foram definidos períodos específicos para preservar sua natureza periódica durante a análise.

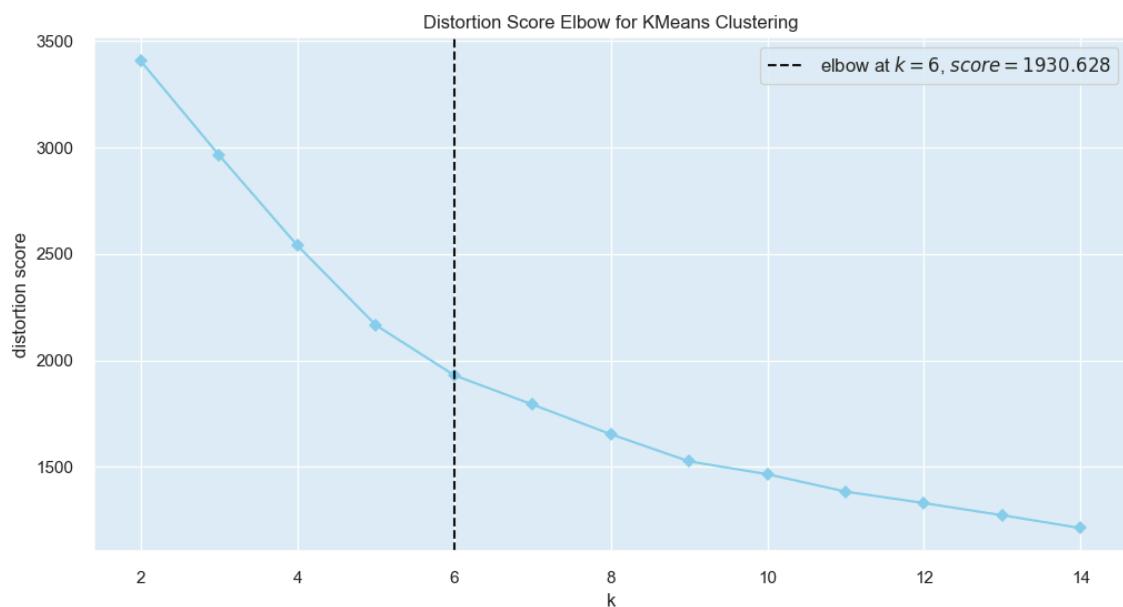
Com os dados tratados, aplicou-se a técnica de **Análise de Componentes Principais (PCA)** para redução de dimensionalidade, facilitando a visualização e eliminando a multicolinearidade entre variáveis, além de melhorar a eficiência computacional dos algoritmos subsequentes.

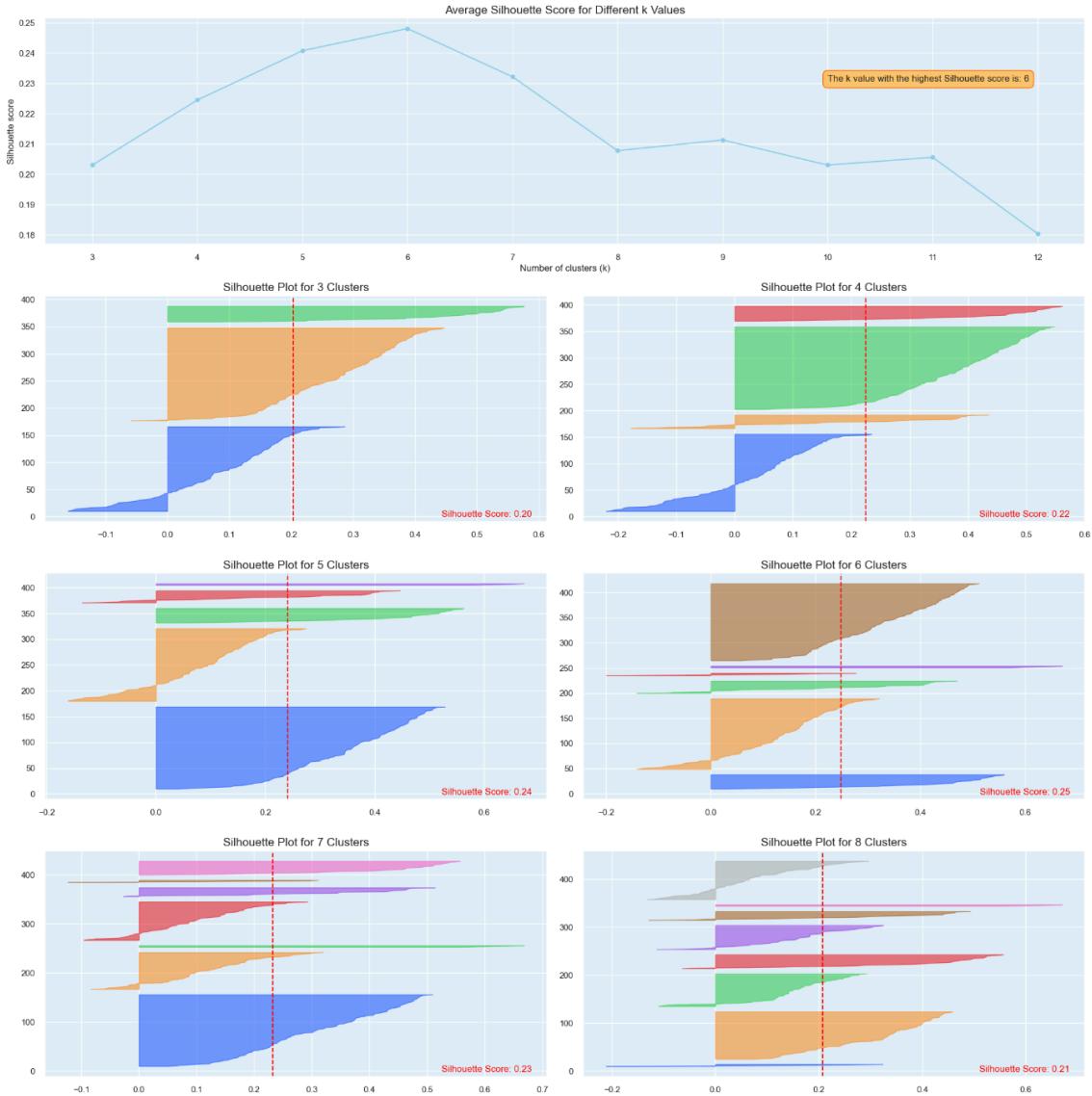


- Escolha do número de componentes:** A linha vertical vermelha sugere que 6 componentes são suficientes para representar adequadamente a estrutura dos dados, explicando 70% da variância total. Esta escolha representa um equilíbrio entre a redução de dimensionalidade e a preservação da informação relevante.
- Distribuição da variância:** Os primeiros componentes capturam proporcionalmente mais variância, demonstrando que existem estruturas dominantes nos dados. A curva de variância cumulativa mostra um padrão típico de "cotovelo", onde os ganhos marginais de adicionar componentes diminuem após certo ponto.
- Complexidade dos dados:** O fato de precisarmos de 6 componentes para explicar 70% da variância e de 14 componentes para chegar a 98% sugere que os dados têm uma estrutura relativamente complexa, com informações distribuídas em várias dimensões.

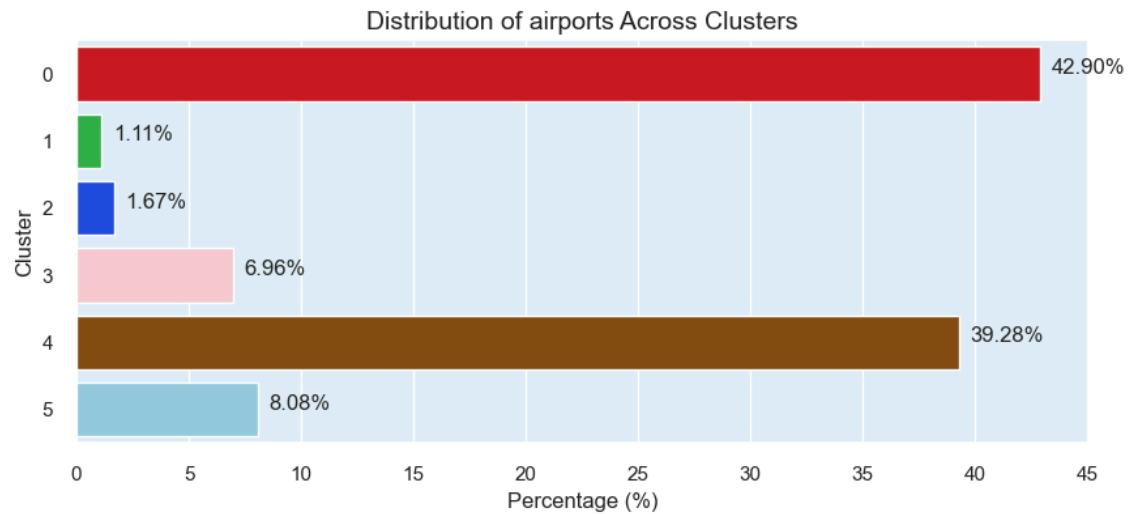
4. **Eficiência da redução:** Usando apenas 6 componentes (redução significativa em relação ao número total de 19 dimensões originais), conseguimos preservar 70% da informação contida nos dados originais, o que é uma compressão eficiente.

Para determinar o número ideal de clusters, foram utilizados dois métodos complementares: o **elbow method**, que analisa a soma dos quadrados das distâncias intra-cluster, e o **silhouette method**, que avalia a separação e coesão dos grupos formados. A partir das análises, foi definido um valor de **k** para o algoritmo.



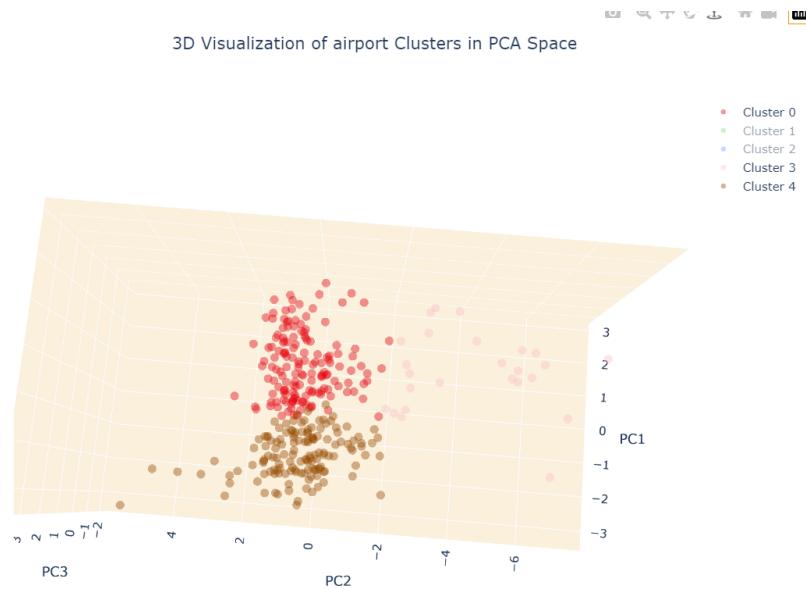


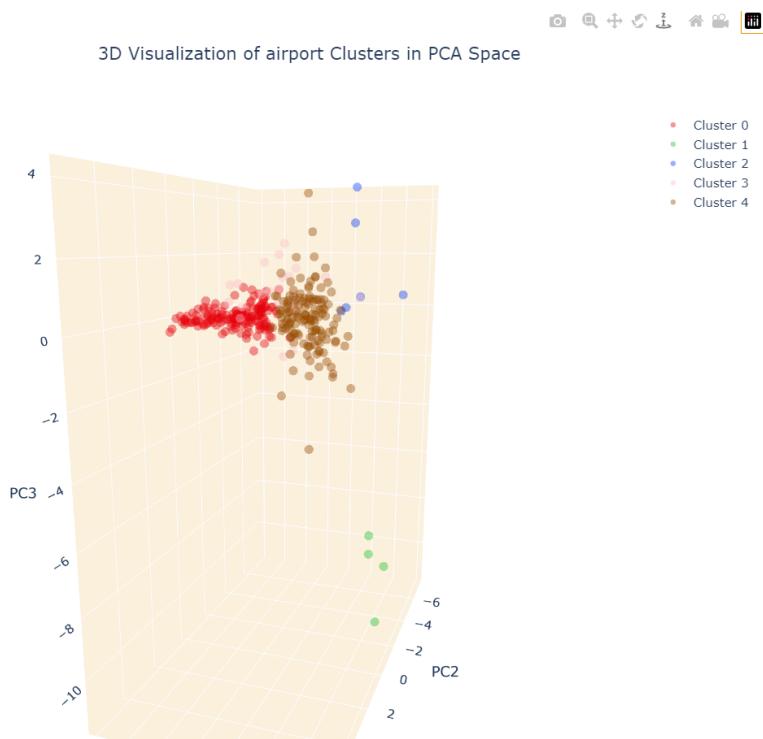
Por fim, aplicou-se o algoritmo **k-means**, com o número de clusters definido pelos métodos anteriores, para agrupar os aeroportos em perfis operacionais homogêneos. Esta segmentação permitiu identificar padrões distintos no desempenho e nas condições dos aeroportos, fundamentando a etapa de filtragem do sistema de recomendação.



Apesar de a distribuição não ser totalmente equilibrada, verifica-se uma forte semelhança entre os elementos presentes nos clusters de menor dimensão. Por esse motivo, considerou-se adequada a decisão de manter os 6 clusters.

7.5 Validação do desempenho do modelo





Nas duas figuras acima verifica-se a separação dos clusters de maneira relativamente clara apesar de existir um pequeno overlap entre o cluster 0 e o 4. Também conseguimos verificar que os clusters apresentam-se de forma mais isolada em relação aos outros, esta diferença é mais pronunciada no cluster 1 que encontra-se muito isolado dos outros.

7.6 Avaliação dos resultados

A aplicação do modelo resultou na identificação de seis perfis principais de clientes:

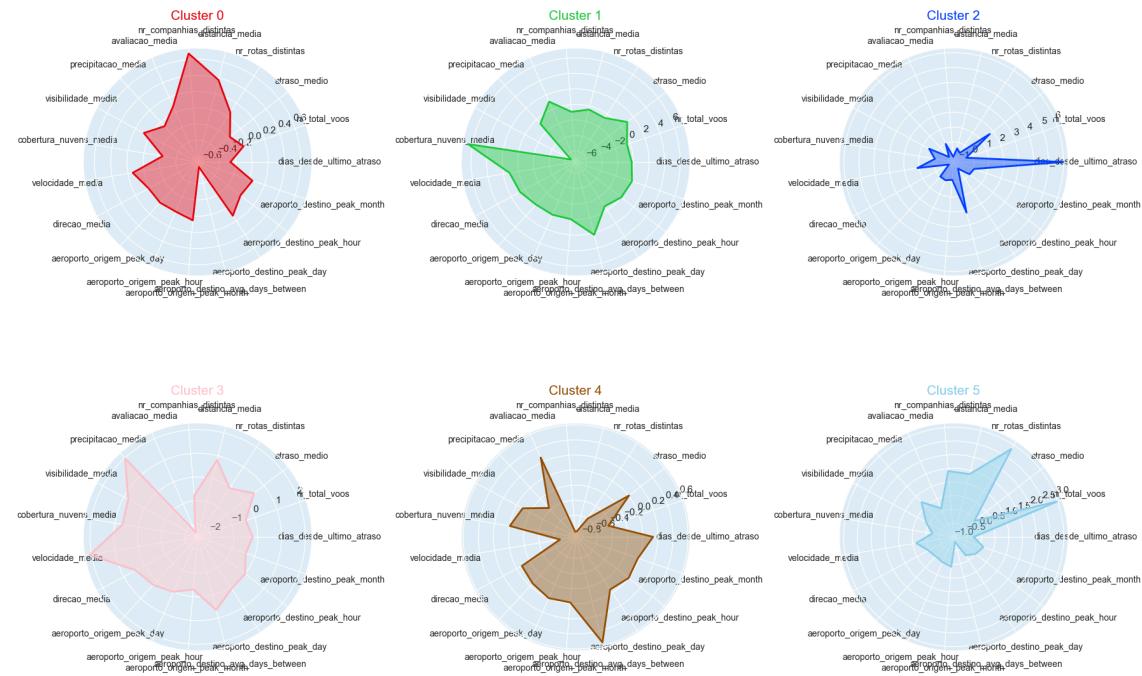


Figura 13 - Distribuição dos clusters resultantes do modelo aplicado.

• **Cluster 0 - Destinos Turisticos**

O cluster de aeroportos em análise apresenta características típicas de infraestruturas fortemente orientadas para o turismo. Entre os aeroportos incluídos neste grupo encontra-se o HSV (Huntsville International Airport), localizado no Alabama. De forma geral, os aeroportos deste cluster registam um elevado volume de tráfego aéreo, com um grande número total de voos e uma significativa diversidade de companhias aéreas a operar. A pontualidade destaca-se como um fator comum, com um atraso médio negativo nas chegadas, o que indica que os voos tendem a aterrhar antes da hora prevista. O número de dias desde o último atraso é muito reduzido, refletindo a intensa atividade e a gestão eficiente das operações. Este cluster apresenta também a maior distância média por viagem, sugerindo a prevalência de voos internacionais ou de longo curso. Do ponto de vista sazonal, os meses de verão — especialmente junho, julho e agosto — concentram o maior volume de tráfego, coincidindo com o período de férias. Em termos semanais, os dias de partida mais movimentados ocorrem à segunda-feira, quinta-feira e sexta-feira, enquanto as chegadas se intensificam ao fim de semana e à segunda-feira, um padrão típico de destinos turísticos de curta ou média duração.

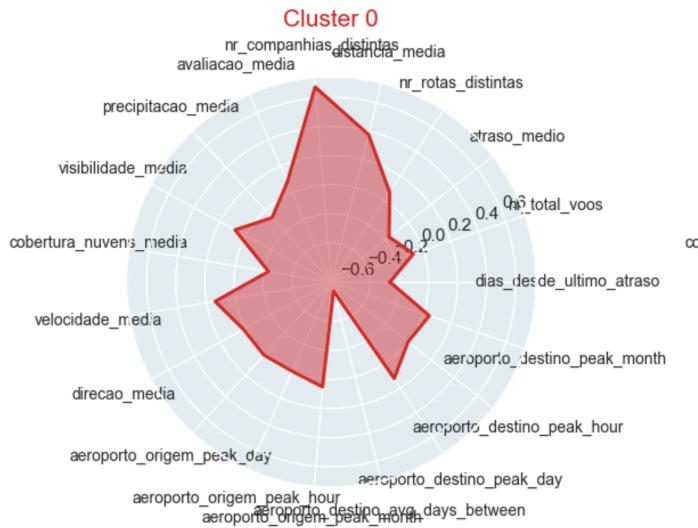


Figura 14 - Cluster 0 - Destinos turísticos.

A partir do gráfico e dos histogramas, destacam-se os seguintes pontos:

- **nr_companhias_distintas, nr_rotas_distintas, e distancia_media:** Clientes que usam várias companhias e rotas diferentes, com distâncias médias elevadas.
- **total_voos:** Volume de voos moderado a elevado.
- **velocidade_media, cobertura_nuvens_media, e precipitacao_media:** Voo em condições meteorológicas geralmente mais favoráveis.
- **dias_desde_ultimo_atraso e atraso_medio:** Indicadores de atraso negativos — ou seja, atrasos muito frequentes.
- **aeroponto_origem/destino_peak_***: Participação baixa em períodos de maior movimento (hora, dia, mês).
- **Cluster 1 - Aeroportos no Alasca**

O cluster de aeroportos em análise encontra-se inteiramente localizado no Alasca, incluindo o aeroporto DLG (Dillingham). Estes aeroportos partilham condições meteorológicas adversas típicas da região, que influenciam significativamente a sua operação. As chegadas concentram-se maioritariamente entre domingo e terça-feira, enquanto os voos de origem mais populares apresentam uma distribuição mais equilibrada ao longo da semana. O número de companhias aéreas distintas que operam nestes aeroportos é reduzido, refletindo um mercado mais restrito e especializado. Apesar da existência de atrasos regulares, estes mantêm-se relativamente baixos, o que sugere que o planeamento e a gestão operacional para fazer face às condições meteorológicas adversas são eficazes e bem estruturados.

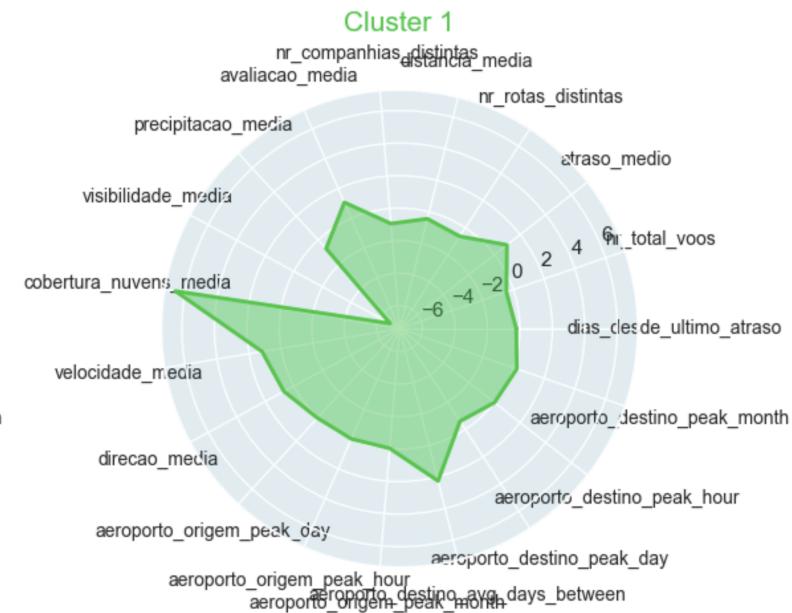


Figura 15 - Cluster 1 - Aeroportos no Alasca

Cluster 2 - Cluster de Aeroportos Locais de Pequena Dimensão com Operação Restrita

Este cluster aparenta representar aeroportos locais de pequena dimensão, exemplificado pelo aeroporto MMH (Mammoth Yosemite Airport), na Califórnia. Caracteriza-se por apresentar atrasos médios elevados, apesar de dispor de um número reduzido de rotas e uma distância média de voo relativamente curta. O número de companhias aéreas que operam nestes aeroportos é bastante limitado, refletindo um mercado mais restrito. As condições meteorológicas são, em geral, favoráveis, o que indica que os atrasos registados poderão estar associados a fatores operacionais ou de capacidade, em vez de condições ambientais adversas.

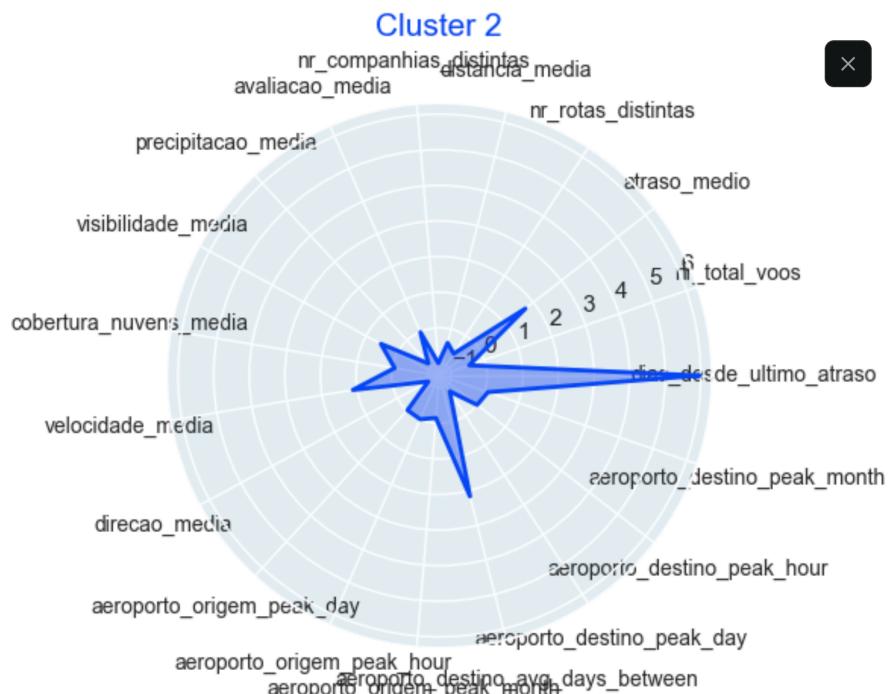


Figura 16 - Cluster 2 - Cluster de Aeroportos Locais de Pequena Dimensão com Operação Restrita

Cluster 3- Cluster de Aeroportos em Zonas Desérticas com Alta Movimentação e Condições Meteorológicas Desafiantes

Este cluster agrupa aeroportos localizados em regiões desérticas, exemplificados pelo aeroporto AZA (Phoenix-Mesa Gateway Airport). Caracterizam-se por uma baixa frequência de atrasos, apesar de apresentarem um elevado número total de voos. A média dos atrasos situa-se próxima de zero, indicando boa pontualidade geral. Alguns aeroportos deste cluster possuem uma grande diversidade de rotas, enquanto outros mantêm um número moderado, relativamente elevado para aeroportos de menor dimensão. O número de companhias aéreas distintas que operam nestes aeroportos é moderado. As condições ambientais típicas incluem ventos fortes, que podem impactar a operação aérea. Em termos de movimentação semanal, terças e quartas-feiras são os dias menos movimentados, com maior afluência nos restantes dias da semana. Do ponto de vista sazonal, os meses de maior atividade coincidem com a primavera e o final do verão.

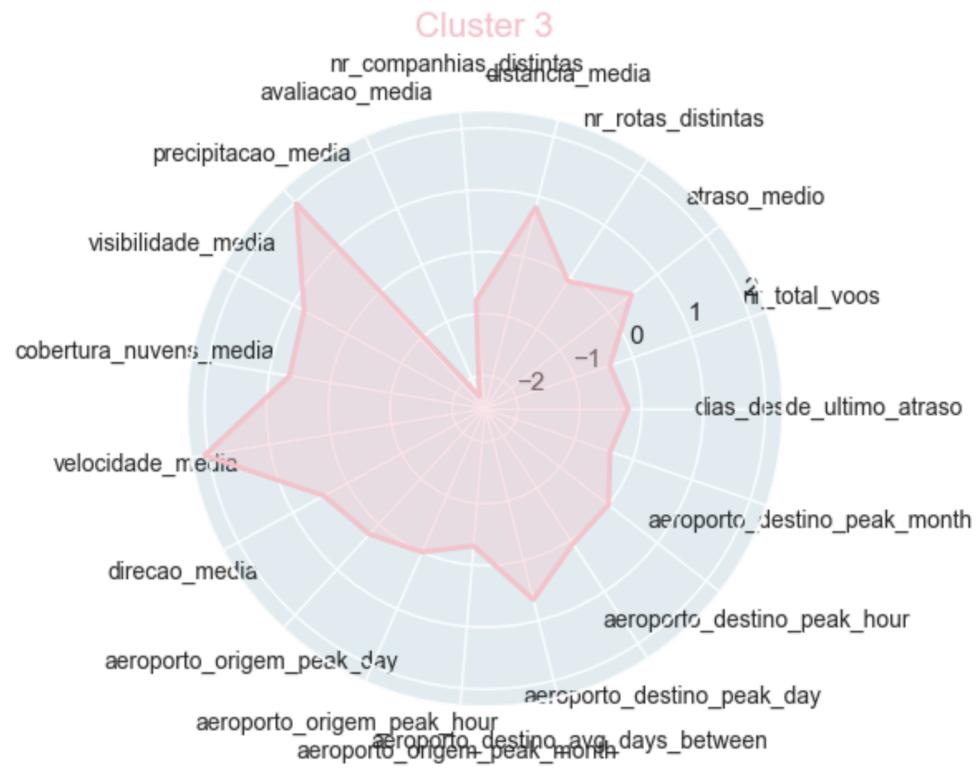


Figura 17 - Cluster 3 - Cluster de Aeroportos em Zonas Desérticas com Alta Movimentação e Condições Meteorológicas Desafiantes

Cluster 4 - Cluster de Aeroportos Regionais com Perfil Empresarial e Conectividade Local

Este cluster reúne aeroportos regionais de dimensão média a pequena, exemplificados pelo ABI (Abilene Regional Airport) no Texas. Estes aeroportos funcionam sobretudo como destinos para viagens regionais e de negócios, caracterizando-se por um volume moderado de voos e uma rede reduzida de rotas, operadas por um número limitado de companhias aéreas. A pontualidade é, em geral, elevada, com atrasos médios baixos, refletindo uma operação eficiente e ajustada às necessidades locais. A atividade nestes aeroportos não apresenta fortes flutuações sazonais associadas ao turismo de lazer, indicando que são utilizados maioritariamente para deslocações profissionais, eventos regionais ou viagens de curta duração. A sua localização em cidades de dimensão média ou pequena torna-os fundamentais para a conectividade local e para o suporte à economia regional, servindo como importantes hubs para o transporte aéreo dentro das suas respetivas áreas de influência.

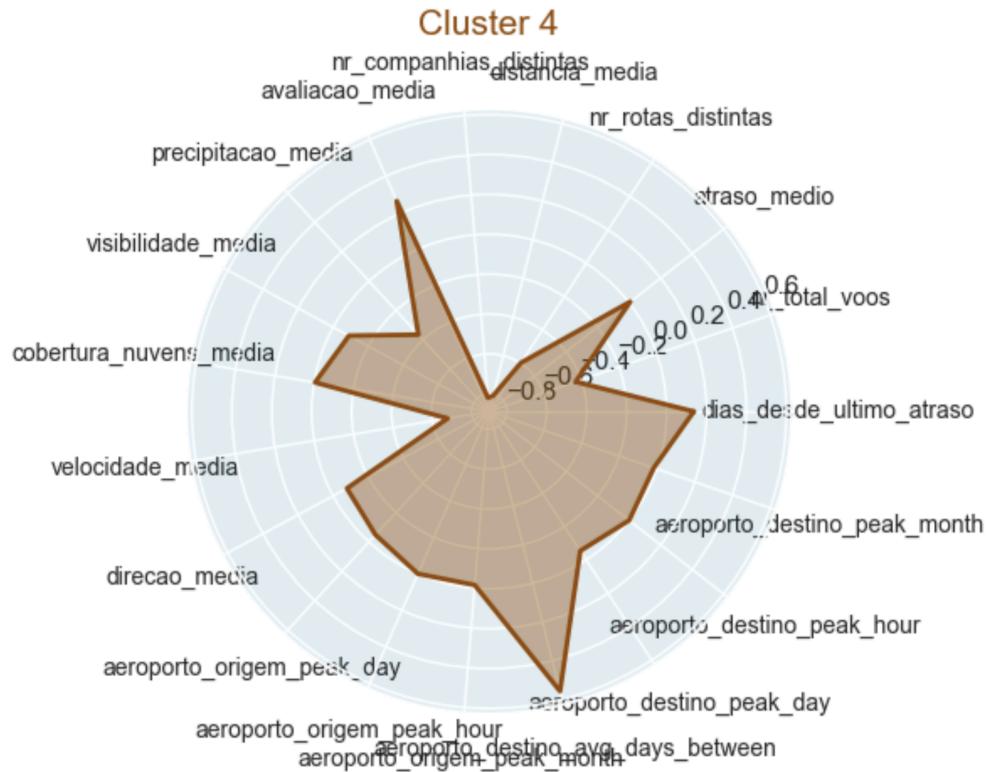


Figura 18 - Cluster 4 -Cluster de Aeroportos Regionais com Perfil Empresarial e Conectividade Local

A partir do Cluster 4, podemos deduzir as seguintes características:

- **atraso_medio**: negativo — reforça a boa performance operacional atual.
- **aeroporto_destino_peak_month/hour/day** e **aeroporto_origem_peak_month**: Positivos — estas rotas coincidem com os períodos de maior movimento nos aeroportos, o que sugere importância estratégica ou sazonalidade turística.
- **nr_total_voos**: Ligeiramente acima da média — estas rotas têm presença regular, mas não excessiva, na malha aérea.
- **nr_companhias_distintas**: Valor negativo — poucas companhias operam nestas rotas, reforçando um caráter mais especializado ou regional.
- **nr_rotas_distintas**: Negativo moderado — baixa diversidade de rotas, com foco em conexões específicas e recorrentes.
- **distancia_media**: Negativa — rotas curtas, possivelmente regionais ou intermunicipais.
- ex aeroporto abi abilene texas

Cluster 5 - Cluster de Aeroportos de Grande Dimensão com Elevada Diversidade de Rotas e Operação Eficiente

Este cluster caracteriza-se por aeroportos que registam um imenso número de voos diáários, mantendo um atraso médio baixo, o que demonstra uma operação eficiente apesar da elevada intensidade de tráfego. Apresentam a maior diversidade e quantidade de rotas em comparação com os outros clusters, refletindo a sua relevância como pontos estratégicos para ligações de curta, média e longa distância. Além disso, contam com um grande número de companhias aéreas distintas, o que contribui para uma oferta variada de serviços e destinos. As condições meteorológicas são geralmente favoráveis, permitindo uma operação estável ao longo do ano. Em termos sazonais, os meses mais movimentados são agosto e, com menor intensidade, julho e março, indicando picos específicos de procura ao longo do verão e na primavera.

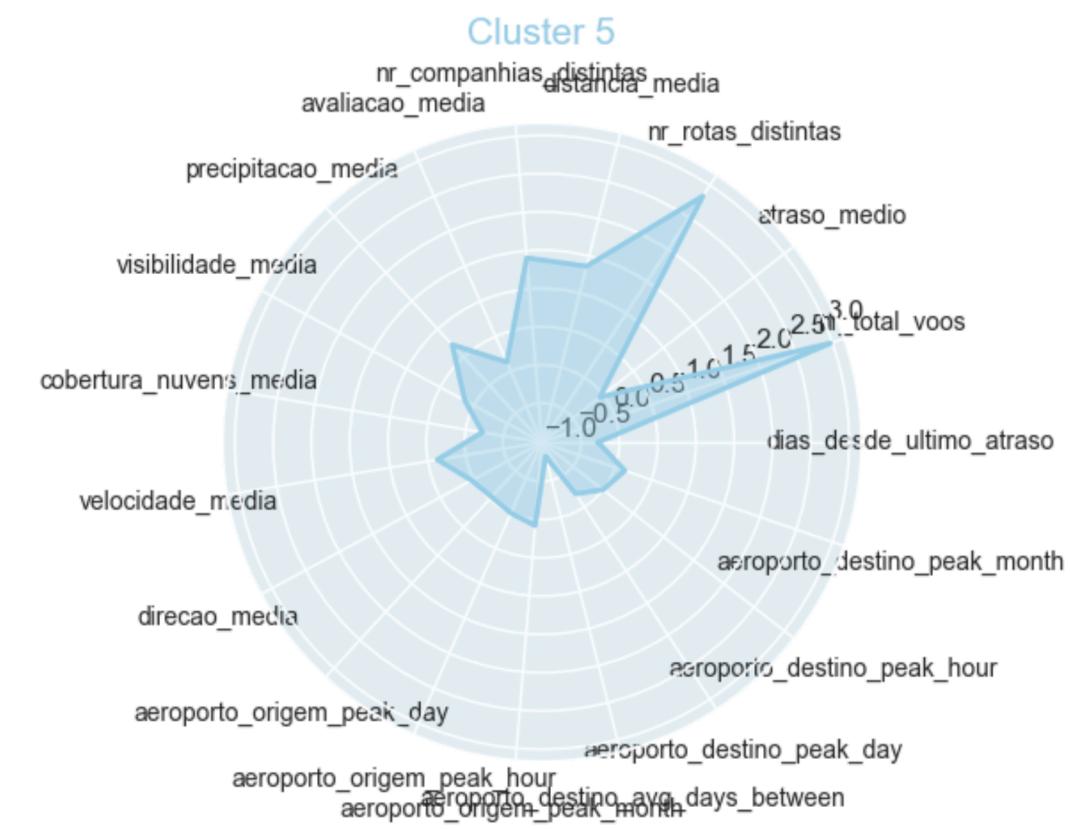


Figura 19 - Cluster 5 - .

De acordo com a informação disposta no cluster 5, concluímos as seguintes características:

- **atraso_medio:** Extremamente negativo — rotas com pontualidade exemplar, uma das marcas mais fortes deste cluster .
- **nr_total_voos:** Muito elevado — este cluster reúne rotas com altíssimo volume de operações.
- **nr_rotas_distintas:** Positivo expressivo — grande diversidade de destinos e percursos.

- `nr_companhias_distintas`: Muito positivo — alta diversidade de operadores nesses aeroportos.
- `precipitacao_media` e `cobertura_nuvens_media`: Ligeiramente negativos — condições climáticas geralmente favoráveis, possivelmente contribuindo para a alta pontualidade.
- Indicadores de pico (peak hour/day/month): Negativos ou neutros — estas rotas não operam majoritariamente em períodos de pico, o que diminui a pressão sobre a infraestrutura e pode favorecer a pontualidade.

8. Personalização de Ofertas de Produtos e Serviços

8.1 Definição do problema e compreensão dos elementos de análise envolvidos

O setor da aviação enfrenta desafios contínuos relacionados à gestão eficiente das operações aeroportuárias, sendo os atrasos nos voos um dos problemas mais críticos que afetam a satisfação dos passageiros, a eficiência operacional das companhias aéreas e a gestão dos aeroportos. Estes atrasos resultam de uma combinação complexa de fatores, incluindo condições meteorológicas adversas, limitações na infraestrutura aeroportuária, volume elevado de voos, problemas operacionais e fatores externos como questões de segurança.

Diante deste cenário, o objetivo principal deste trabalho é desenvolver um sistema de recomendação capaz de apoiar gestores aeroportuários e operadores a identificarem aeroportos com maior potencial para minimizar atrasos. A recomendação visa otimizar a escolha de aeroportos para determinadas operações, rotas e períodos, com base em dados históricos de performance, características climáticas e operacionais. Como isto não é sempre possível, temos também o objetivo de informar os consumidores de possíveis atrasos que possam ocorrer.

Para resolver este problema, foram adotadas diversas estratégias como uma abordagem híbrida que combina técnicas de análise de clusters que utiliza a segmentação dos aeroportos em perfis operacionais homogêneos, filtragem baseada em conteúdo para seleção dos aeroportos mais adequados, e modelos preditivos de regressão para estimar o risco e o tempo esperado de atraso. Esta combinação permite uma recomendação mais precisa e contextualizada, ajustada às condições específicas de cada aeroporto e suas variáveis ambientais.

8.2 Seleção e preparação dos dados

Para o desenvolvimento dos sistemas de recomendação, foram utilizados os dados previamente agrupados nos clusters de aeroportos. Essa segmentação facilitou a formulação dos modelos, permitindo trabalhar com perfis operacionais homogêneos que refletem

características similares dos aeroportos. A partir desses grupos, selecionaram-se as variáveis relevantes para a recomendação, mantendo a coerência e representatividade dos dados dentro de cada cluster.

Assim, os sistemas de recomendação basearam-se em informações consolidadas e organizadas pelos clusters, simplificando a análise e tornando os modelos mais eficientes e direcionados às especificidades de cada grupo de aeroportos.

8.3 Identificação e fundamentação da técnica de análise

No desenvolvimento do sistema de recomendação, optou-se por uma abordagem que combina filtragem baseada em conteúdo e um modelo híbrido em cascata, fundamentada nas especificidades do problema e nas características dos dados disponíveis.

A **filtragem baseada em conteúdo** foi escolhida devido à sua capacidade de recomendar companhias aéreas e rotas com base nas características concretas dos aeroportos, como o perfil operacional definido pelos clusters, métricas de atraso e condições meteorológicas. Esta técnica é adequada quando se dispõe de informações detalhadas sobre os itens a recomendar (neste caso, aeroportos), permitindo gerar recomendações interpretáveis e diretamente relacionadas ao contexto específico de cada aeroporto.

Por outro lado, o modelo **híbrido em cascata** foi implementado para aumentar a precisão e robustez das recomendações. Esta abordagem permite reduzir progressivamente o espaço de busca e refinar as recomendações com base em previsões mais detalhadas, adaptando-se às diversas variáveis operacionais e meteorológicas que influenciam os atrasos.

Quanto à **filtragem colaborativa**, esta técnica foi considerada inadequada para o contexto do presente trabalho, uma vez que depende fortemente do histórico de interações ou preferências dos usuários (passageiros) para gerar recomendações. No entanto, o conjunto de dados disponível não contém informações suficientes sobre as preferências individuais dos passageiros ou padrões de comportamento que permitissem a aplicação eficaz deste método. Além disso, o foco do sistema está em apoiar decisões operacionais e estratégicas baseadas em dados agregados e características dos aeroportos, o que é melhor atendido por abordagens baseadas em conteúdo e modelagem preditiva.

8.4 Construção do modelo de análise

Filtragem baseada em conteúdo

Estes modelos simples utilizam informações diretamente associadas ao aeroporto selecionado para recomendar rotas e companhias aéreas. A partir do cluster ao qual o aeroporto pertence, o sistema identifica aeroportos similares e seleciona as companhias aéreas que operam nesses locais, ordenando-as segundo métricas de desempenho, como o atraso médio observado. Esta abordagem permite gerar recomendações rápidas e interpretáveis, baseadas no contexto operacional e histórico, mas não incorpora previsões avançadas ou múltiplas camadas de decisão.

Modelo híbrido em cascata (Cascade Hybrid)

Para aumentar a precisão e robustez das recomendações, foi desenvolvido um modelo híbrido em cascata que combina múltiplas etapas de filtragem e predição, funcionando da seguinte forma:

1. Filtragem inicial por clusters:

O sistema seleciona todos os aeroportos pertencentes aos clusters operacionais pré-definidos, que agrupam aeroportos com perfis e características semelhantes. Esta etapa amplia o conjunto inicial de candidatos com base em agrupamentos previamente identificados, garantindo que a análise subsequente se concentre em aeroportos que compartilham padrões operacionais e ambientais relevantes. Desta forma permite selecionar para quais perfis de clientes fazer as recomendações.

2. Filtragem quantitativa adicional:

Dentro do conjunto filtrado pelos clusters, aplicam-se critérios quantitativos para eliminar aeroportos com atraso médio acima de um limiar, condições climáticas desfavoráveis (ex: precipitação elevada) ou volume insuficiente de voos. Isso reforça a qualidade e relevância dos candidatos.

3. Modelagem preditiva por regressão:

Um modelo de regressão baseado em LightGBM é treinado para prever o atraso médio esperado para cada aeroporto candidato, considerando múltiplas variáveis operacionais e meteorológicas que poderão ser também variadas.

4. Ranking e recomendação:

Os aeroportos candidatos são ordenados com base na previsão do modelo, priorizando aqueles com menor atraso esperado. Essa ordenação resulta numa lista de recomendações otimizadas para apoiar decisões estratégicas e operacionais.

Modelo preditivo para a estimativa dos atrasos em voos

Para o desenvolvimento do sistema de regressão, utilizamos um data frame previamente construído *df_flights_clean*. Desta dataset usamos as features numéricas que consideramos relevantes.

O algoritmo usado foi **XGBRegressor** por ser otimizado e mais rápido que Random Forest tradicional e pelo grande tamanho do dataset utilizado. Foi também tentado utilizar outros algoritmos como o **LightGBM**, no entanto este apresentou resultados inferiores, e o **RandomForestRegressor**. Este último apresentava graves problemas de performance, apresentando-se extremamente ineficiente para o problema apresentado.

Para a utilização deste será necessário efetuar uma filtragem por aeroporto, dado pelo consumidor, e será necessário fornecer informações das condições climatéricas, data e hora do voo, e do seu destino para o mesmo criar uma previsão do atraso na partida.

Recomendação de rotas baseada no cluster do aeroporto

A construção do modelo de análise segue uma abordagem estruturada, começando pela identificação do cluster a que pertence o aeroporto em questão. De seguida, são selecionados todos os aeroportos que fazem parte do mesmo cluster, permitindo uma comparação dentro do mesmo grupo. A função agrupa os voos por rota, calculando a média dos atrasos à chegada. Esta informação é então ordenada de forma ascendente, destacando as rotas com menor tempo médio de atraso. Por fim, o modelo devolve as 10 rotas com melhor desempenho, oferecendo uma visão clara das rotas mais pontuais no contexto do cluster analisado.

Recomendação de companhias aéreas baseada no cluster do aeroporto

A função `recomendar_airlines` foi desenvolvida com o objetivo de sugerir as companhias aéreas mais pontuais associadas a um determinado cluster de aeroportos. O processo inicia-se com a verificação da existência do aeroporto indicado. Se o aeroporto for válido, é identificado o seu cluster, permitindo agrupar todos os aeroportos que partilham características semelhantes. Em seguida, são filtrados os voos cuja origem pertence a esse conjunto de aeroportos, focando a análise nas companhias aéreas que operam nesse contexto. A função agrupa então os dados por companhia aérea, calculando a média dos atrasos na partida. Estes resultados são ordenados de forma crescente, destacando as companhias com menor média de atrasos. Por fim, são retornadas as 10 companhias aéreas mais pontuais do cluster, fornecendo uma recomendação baseada em desempenho.

8.5 Validação do desempenho do modelo

Nos sistemas de recomendação de rotas, o principal objetivo não é fornecer diretamente uma solução, mas sim disponibilizar às entidades reguladoras ou gestoras do aeroporto um ponto de partida para a identificação de possíveis melhorias. Devido à complexidade e à elevada quantidade de dados envolvidos na ocorrência de atrasos, torna-se impraticável apontar uma única causa ou resolução. Por este motivo, o sistema oferece rotas semelhantes às que estão associadas ao aeroporto em análise, permitindo que os responsáveis estudem essas rotas e identifiquem os fatores que contribuem para os seus baixos níveis de atraso.

Por sua vez, o sistema de recomendação de companhias aéreas tem como objetivo apresentar ao consumidor as companhias com melhor desempenho operacional para o tipo de aeroporto em questão. Esta abordagem permite ao passageiro tomar decisões mais informadas e ponderadas ao escolher a companhia aérea cujos serviços irá adquirir, baseando-se no histórico de desempenho em aeroportos com características semelhantes às que irá utilizar.

Para a validação dos outros modelos desenvolvidos neste trabalho foram adotados dois métodos distintos, adequados às características de cada tipo de sistema de recomendação.

Para os modelos baseados em conteúdo, que realizam recomendações por meio de agrupamento e médias de desempenho, foi realizada uma validação qualitativa. Esta abordagem considerou a análise da coerência das recomendações, a comparação dos atrasos médios entre as companhias aéreas recomendadas e as demais, e a avaliação da relevância das sugestões geradas no contexto operacional dos aeroportos. Essa validação qualitativa permitiu aferir a utilidade prática e a aderência do modelo à realidade do setor.

Já para os modelos preditivos, incluindo o modelo híbrido em cascata que utiliza LightGBM para estimar atrasos, foi utilizada uma validação quantitativa baseada em métricas padrão de regressão: RMSE (Root Mean Squared Error), MAE (Mean Absolute Error) e R² (Coeficiente de Determinação). Essas métricas forneceram uma avaliação rigorosa da precisão e robustez dos modelos, permitindo quantificar o erro médio das previsões e o grau de explicação da variância dos dados.

Dessa forma, a combinação dessas duas abordagens de validação garantiu uma análise abrangente do desempenho dos modelos, contemplando tanto aspectos técnicos quanto a aplicabilidade prática das recomendações.

8.6 Avaliação dos resultados

Modelo Preditivo

Resultados obtidos:

- **RMSE:** 88.10
- **MAE:** 18.08
- **R²:** 0.027

Embora os valores de **RMSE** e **MAE** estejam dentro de uma faixa aceitável para um conjunto de dados com alta variabilidade, o valor de **R²** indica que o modelo explica apenas uma pequena parte da variância nos dados. Isso sugere que o modelo tem desempenho limitado na previsão precisa do atraso.

Sistema de Recomendação Cascade

Resultados obtidos:

- **RMSE:** 0.9694
- **MAE:** 0.3476
- **R²:** 0.9728

Este modelo apresentou resultados bastante promissores, com baixo erro médio (RMSE de 0.97 e MAE de 0.35) e um elevado R² de 0.97, indicando que explica 97% da variância dos atrasos. Esses indicadores confirmam a capacidade do modelo em prever com precisão os atrasos nos aeroportos, tornando-o uma ferramenta eficaz para apoiar decisões operacionais e melhorar a gestão dos aeroportos.

9. Conclusões e Trabalho Futuro

9.1 Conclusões

Este projeto revelou-se uma experiência extremamente rica em termos de aprendizagem, tanto a nível técnico como metodológico. Apesar dos objetivos traçados inicialmente, consideramos que o tema do nosso projeto foi **demasiado ambicioso** para o tempo e, nomeadamente, para os recursos disponíveis. A complexidade dos dados, aliada ao volume e à diversidade de fontes, colocou desafios que poderiam ter sido mitigados com um **planeamento mais rigoroso** e uma **análise exploratória mais profunda dos datasets logo numa fase inicial**.

A criação e transformação de dados foi, por vezes, feita com base em suposições ou simplificações que, em retrospectiva, poderiam ter sido substituídas por estratégias mais **robustas e realistas**, tanto na geração de dados sintéticos como na deteção e correção de inconsistências.

Além disso, o **sistema de povoamento**, embora funcional, apresentou **problemas de desempenho** em certos pontos críticos do pipeline, especialmente ao lidar com grandes volumes de dados e integrações mais complexas (e.g., junção com dados climáticos). A validação e o controlo de qualidade dos dados exigiram um esforço considerável, o que expôs limitações na arquitetura inicial adotada.

Por outro lado, a implementação de múltiplas camadas de staging e histórico, bem como a separação clara entre dados válidos e inválidos, permitiu garantir uma maior rastreabilidade e controle sobre os fluxos de informação. Este aspeto é claramente um ponto forte do trabalho desenvolvido.

Em retrospectiva, deveríamos ter abdicado do grande volume de dados que tínhamos e ter utilizado diretamente o data warehouse em vez dos ficheiros CSV. Esta abordagem retirou alguma nuance ao projeto e não demonstrou plenamente as capacidades do data warehouse implementado.

Serve também referir que accidentalmente utilizamos a hora como feature, infelizmente só nos apercebemos após as apresentações por isso referimos isso aqui mas no notebook enviado já está corrigido.

Em suma, apesar de alguns constrangimentos técnicos e metodológicos, foi possível alcançar resultados e identificar áreas de valor analítico, nomeadamente na análise de clusters de voos.

9.2 Trabalho Futuro

Com base nas dificuldades encontradas e nas oportunidades identificadas ao longo do projeto, destacam-se diversas linhas de trabalho futuro que poderiam contribuir para melhorar significativamente a qualidade, eficiência e abrangência do sistema:

- **Melhoria do sistema de integração:** A atual arquitetura pode ser otimizada para lidar melhor com **datasets volumosos** e com diferentes formatos. A introdução de **ferramentas de ETL mais robustas** ou a utilização de **frameworks paralelas (como Apache Spark ou Dask)** poderá melhorar substancialmente o desempenho e a escalabilidade do sistema de povoamento.
- **Clusterização com novas variáveis:** Uma direção promissora seria aprofundar a análise de clusters, explorando **agrupamentos baseados nas companhias aéreas**, **nas características das rotas**, ou até na **pontualidade e satisfação dos passageiros**. Isso permitiria caracterizações mais granulares e insights orientados a intervenções operacionais.
- **Aprofundamento dos sistemas de recomendação:** O módulo de recomendação, ainda em fase preliminar, pode ser expandido com a introdução de **modelos colaborativos, baseados em conteúdo, ou híbridos**, e reforçado com feedback real (ou simulado) de utilizadores, promovendo recomendações mais personalizadas e eficazes.
- **Otimização da performance do sistema de povoamento:** Algumas operações complexas, nomeadamente joins com dados externos (e.g., clima), poderiam ser repensadas ou reestruturadas com base em **indexação geoespacial, materialização de views** ou **uso de bases de dados especializadas**, como **PostGIS** para dados geográficos.
- **Aquisição de dados adicionais:** Algumas **dimensões revelaram-se insuficientemente detalhadas**, nomeadamente a dimensão **Aeronave**, para a qual seria benéfico obter dados adicionais como tipo de motor, capacidade, fabricante, idade média da frota, entre outros. Estes dados enriqueceriam as análises e permitiriam novas correlações.

- **Monitorização e automação de qualidade de dados:** Implementar **pipelines automáticos de validação de dados**, com alertas em tempo real e mecanismos de autocorreção ou feedback assistido, poderia reforçar a fiabilidade do sistema e reduzir significativamente o esforço manual.
- **Interface gráfica de monitorização:** A criação de um **painel de controlo visual** para acompanhar o estado do povoamento, regtos em erro, e indicadores de qualidade permitiria um acompanhamento contínuo e mais intuitivo por parte dos utilizadores e administradores do sistema.
- **Diferentes técnicas de recomendação:** A introdução de informações e histórico de passageiros iria melhorar o sistema de recomendação existente pois iria permitir recomendações personalizadas não só para aeroportos mas também sugestões únicas para cada cliente.

9. Bibliografia

<https://www.ibm.com/think/topics/data-warehouse>

Kaufman, L., & Rousseeuw, P. J. (2005). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley.

Apache NiFi Documentation. (n.d.). *The Apache Software Foundation*. Disponível em:

<https://nifi.apache.org/docs.html>

Lista de Siglas e Acrónimos

BD	Base de Dados
DW	Data Warehouse
OLAP	<i>On-Line Analytical Processing</i>
ETL	<i>Extração, Transformação e Carregamento</i>
RMSE	<i>Raiz do Erro Quadrático Médio</i>
MAE	<i>Erro Médio Absoluto</i>
R^2	<i>Coeficiente de Determinação</i>

I. Anexo 1

- Ficheiro sql para criação da base de dados e tabelas

```
CREATE DATABASE FONTE;
CREATE DATABASE RAW;
CREATE DATABASE ERROS;
CREATE DATABASE RDY;
CREATE DATABASE HST;

USE FONTE;

CREATE TABLE Companhia_AereaFONTE (
    Airline_ID INT PRIMARY KEY,
    Name VARCHAR(150) NOT NULL,
    Country VARCHAR(50)
);

CREATE TABLE AeronaveFONTE (
    Model VARCHAR(100) PRIMARY KEY,
    Company VARCHAR(100)
);

CREATE TABLE AeroportoFONTE (
    ID_aeroporto INT PRIMARY KEY,
    Nome VARCHAR(100),
    Pais VARCHAR(100),
    Cidade VARCHAR(100),
    Latitude DECIMAL(15, 10),
    Longitude DECIMAL(15, 10)
);
USE FONTE;
SHOW TABLES;

USE AUD;

CREATE TABLE Companhia_AereaAUD (
    Airline_ID INT PRIMARY KEY,
    Name VARCHAR(150) NOT NULL,
    Country VARCHAR(50),
    Operacao VARCHAR(50) NOT NULL,
    Etiqueta VARCHAR(50) NOT NULL,
    Utilizador VARCHAR(50) NOT NULL,
    FOREIGN KEY (Airline_ID) REFERENCES
FONTE.Companhia_AereaFONTE(Airline_ID) ON DELETE CASCADE
);

CREATE TABLE AeronaveAUD (
```

```

Model VARCHAR(100) PRIMARY KEY,
Company VARCHAR(100),
Operacao VARCHAR(50) NOT NULL,
Etiqueta VARCHAR(50) NOT NULL,
Utilizador VARCHAR(50) NOT NULL
);

CREATE TABLE AeroportoAUD (
    ID_aeroporto INT PRIMARY KEY,
    Nome VARCHAR(100),
    Pais VARCHAR(100),
    Cidade VARCHAR(100),
    Latitude DECIMAL(15, 10),
    Longitude DECIMAL(15, 10),
    Operacao VARCHAR(50) NOT NULL,
    Etiqueta VARCHAR(50) NOT NULL,
    Utilizador VARCHAR(50) NOT NULL,
    FOREIGN KEY (ID_aeroporto) REFERENCES
FONTE.AeroportoFONTE(ID_aeroporto) ON DELETE CASCADE
);

USE FONTE;

DELIMITER //
CREATE TRIGGER trg_FONTE_to_AUD_Companhia_AereaINSERT
AFTER INSERT ON FONTE.Companhia_AereaFONTE
FOR EACH ROW
BEGIN
    INSERT INTO AUD.Companhia_AereaAUD (
        Airline_ID,
        Name,
        Country,
        Operacao,
        Etiqueta,
        Utilizador
    ) VALUES (
        NEW.Airline_ID,
        NEW.Name,
        NEW.Country,
        'INSERT',
        NOW(),
        USER()
    );
END //
DELIMITER ;

DELIMITER //
CREATE TRIGGER trg_FONTE_to_AUD_Companhia_AereaUPDATE
AFTER UPDATE ON FONTE.Companhia_AereaFONTE
FOR EACH ROW
BEGIN

```

```

INSERT INTO AUD.Companhia_AereaAUD (
    Airline_ID,
    Name,
    Country,
    Operacao,
    Etiqueta,
    Utilizador
) VALUES (
    NEW.Airline_ID,
    NEW.Name,
    NEW.Country,
    'UPDATE',
    NOW(),
    USER()
);
END //
DELIMITER ;

DELIMITER //
CREATE TRIGGER trg_FONTE_to_AUD_Companhia_AereaDELETE
AFTER DELETE ON FONTE.Companhia_AereaFONTE
FOR EACH ROW
BEGIN
    INSERT INTO AUD.Companhia_AereaAUD (
        Airline_ID,
        Name,
        Country,
        Operacao,
        Etiqueta,
        Utilizador
    ) VALUES (
        OLD.Airline_ID,
        OLD.Name,
        OLD.Country,
        'DELETE',
        NOW(),
        USER()
    );
END //
DELIMITER ;

#-----#
DELIMITER //
CREATE TRIGGER trg_FONTE_to_AUD_AeronaveINSERT
AFTER INSERT ON FONTE.AeronaveFONTE
FOR EACH ROW
BEGIN
    INSERT INTO AUD.AeronaveAUD (
        Model,
        Company,

```

```

Operacao,
Etiqueta,
Utilizador
) VALUES (
    NEW.Model,
    NEW.Company,
    'INSERT',
    NOW(),
    USER()
);
END //
DELIMITER ;

DELIMITER //
CREATE TRIGGER trg_FONTE_to_AUD_AeronaveUPDATE
AFTER UPDATE ON FONTE.AeronaveFONTE
FOR EACH ROW
BEGIN
    INSERT INTO AUD.AeronaveAUD (
        Model,
        Company,
        Operacao,
        Etiqueta,
        Utilizador
    ) VALUES (
        NEW.Model,
        NEW.Company,
        'UPDATE',
        NOW(),
        USER()
    );
END //
DELIMITER ;

DELIMITER //
CREATE TRIGGER trg_FONTE_to_AUD_AeronaveDELETE
AFTER DELETE ON FONTE.AeronaveFONTE
FOR EACH ROW
BEGIN
    INSERT INTO AUD.AeronaveAUD (
        Model,
        Company,
        Operacao,
        Etiqueta,
        Utilizador
    ) VALUES (
        OLD.Model,
        OLD.Company,
        'DELETE',
        NOW(),
        USER()
    );
END //
DELIMITER ;

```

```

        USER()
    );
END //
DELIMITER ;

#-----
USE FONTE;

DELIMITER //
CREATE TRIGGER trg_FONTE_to_AUD_AeroportoINSERT
AFTER INSERT ON FONTE.AeroportoFONTE
FOR EACH ROW
BEGIN
    INSERT INTO AUD.AeroportoAUD (
        ID_aeroporto,
        Nome,
        Pais,
        Cidade,
        Latitude,
        Longitude,
        Operacao,
        Etiqueta,
        Utilizador
    ) VALUES (
        NEW.ID_aeroporto,
        NEW.Nome,
        NEW.Pais,
        NEW.Cidade,
        NEW.Latitude,
        NEW.Longitude,
        'INSERT',
        NOW(),
        USER()
    );
END //
DELIMITER ;

DELIMITER //
CREATE TRIGGER trg_FONTE_to_AUD_AeroportoUPDATE
AFTER UPDATE ON FONTE.AeroportoFONTE
FOR EACH ROW
BEGIN
    INSERT INTO AUD.AeroportoAUD (
        ID_aeroporto,
        Nome,
        Pais,
        Cidade,
        Latitude,
        Longitude,
        Operacao,

```

```

        Etiqueta,
        Utilizador
    ) VALUES (
        NEW.ID_aeroporto,
        NEW.Nome,
        NEW.Pais,
        NEW.Cidade,
        NEW.Latitude,
        NEW.Longitude,
        'UPDATE',
        NOW(),
        USER()
    );
END //
DELIMITER ;

DELIMITER //
CREATE TRIGGER trg_FONTE_to_AUD_AeroportoDELETE
AFTER DELETE ON FONTE.AeroportoFONTE
FOR EACH ROW
BEGIN
    INSERT INTO AUD.AeroportoAUD (
        ID_aeroporto,
        Nome,
        Pais,
        Cidade,
        Latitude,
        Longitude,
        Operacao,
        Etiqueta,
        Utilizador
    ) VALUES (
        OLD.ID_aeroporto,
        OLD.Nome,
        OLD.Pais,
        OLD.Cidade,
        OLD.Latitude,
        OLD.Longitude,
        'DELETE',
        NOW(),
        USER()
    );
END //
DELIMITER ;

USE RAW;

CREATE TABLE Companhia_AereaRAW (
    Airline_ID INT PRIMARY KEY,
    Name VARCHAR(150) NOT NULL,

```

```

        Country VARCHAR(50),
        Operacao VARCHAR(50) NOT NULL,
        Etiqueta VARCHAR(50) NOT NULL,
        Utilizador VARCHAR(50) NOT NULL
    );

CREATE TABLE AeronaveRAW (
    Model VARCHAR(100),
    Company VARCHAR(100),
    Operacao VARCHAR(50) NOT NULL,
    Etiqueta VARCHAR(50) NOT NULL,
    Utilizador VARCHAR(50) NOT NULL
);

CREATE TABLE AeroportoRAW (
    ID_aeroporto INT PRIMARY KEY,
    Nome VARCHAR(100),
    Pais VARCHAR(100),
    Cidade VARCHAR(100),
    Latitude DECIMAL(15,10),
    Longitude DECIMAL(15,10),
    Operacao VARCHAR(50) NOT NULL,
    Etiqueta VARCHAR(50) NOT NULL,
    Utilizador VARCHAR(50) NOT NULL
);

USE ERROS;

CREATE TABLE Companhia_AereaERRO (
    Airline_ID INT PRIMARY KEY,
    Name VARCHAR(50),
    Country VARCHAR(50),
    Operacao VARCHAR(50) NOT NULL,
    Etiqueta VARCHAR(50) NOT NULL,
    Utilizador VARCHAR(50) NOT NULL,
    Tipo VARCHAR(100) NOT NULL,
    Descricao VARCHAR(100) NOT NULL
);

CREATE TABLE AeronaveERRO (
    Model VARCHAR(100) PRIMARY KEY,
    Company VARCHAR(100),
    Operacao VARCHAR(50) NOT NULL,
    Etiqueta VARCHAR(50) NOT NULL,
    Utilizador VARCHAR(50) NOT NULL,
    Tipo VARCHAR(100) NOT NULL,
    Descricao VARCHAR(100) NOT NULL
);

CREATE TABLE AeroportoERRO (
    ID_aeroporto INT PRIMARY KEY,
    Nome VARCHAR(100),

```

```

        Cidade VARCHAR(100),
        País VARCHAR(100),
        Latitude DECIMAL(15, 10),
        Longitude DECIMAL(15, 10),
        Operacao VARCHAR(50) NOT NULL,
        Etiqueta VARCHAR(50) NOT NULL,
        Utilizador VARCHAR(50) NOT NULL,
        Tipo VARCHAR(100) NOT NULL,
        Descrição VARCHAR(100) NOT NULL
    );

CREATE TABLE VooQUA (
    id_voo INT AUTO_INCREMENT PRIMARY KEY,
    nr_voo INT,
    data_partida DATE,
    nome_companhia_aerea VARCHAR(100),
    aeroporto_origem CHAR(3),
    aeroporto_destino CHAR(3),
    hora_partida_real DECIMAL(5,1),
    hora_partida_esperada DECIMAL(5,1),
    hora_chegada_real DECIMAL(5,1),
    hora_chegada_esperada DECIMAL(5,1),
    cancelado DECIMAL(6,1),
    desviado DECIMAL(6,1),
    distancia DECIMAL(6,1),
    atraso_companhia_aerea DECIMAL(5,1),
    atraso_meteorologia DECIMAL(5,1),
    atraso_SNA DECIMAL(5,1),
    atraso_seguranca DECIMAL(5,1),
    atraso_voo_anterior DECIMAL(5,1),
    Modelo_Avião VARCHAR(50) NOT NULL,
    Tipo VARCHAR(100) NOT NULL,
    Descricao VARCHAR(100) NOT NULL,
    avaliacao DECIMAL(5,1)
);

USE RDY;

CREATE TABLE Companhia_Aerea (
    id_companhia_aerea INT PRIMARY KEY,
    nome VARCHAR(150) UNIQUE NOT NULL,
    país VARCHAR(50) NOT NULL
);

CREATE TABLE Aeronave (
    Modelo VARCHAR(100) PRIMARY KEY,
    Fabricante VARCHAR(100)
);

CREATE TABLE País (
    pais VARCHAR(100) PRIMARY KEY
)

```

```

);

CREATE TABLE Cidade (
    id_cidade INT PRIMARY KEY,
    cidade VARCHAR(100)
);

CREATE TABLE Aeroporto (
    ID_aeroporto INT PRIMARY KEY,
   Codigo_IATA CHAR(3) UNIQUE,
    Nome VARCHAR(100),
    Cidade VARCHAR(100),
    Pais VARCHAR(100),
    Latitude DECIMAL(15,10),
    Longitude DECIMAL(15,10)
);

CREATE TABLE clima (
    id_clima INT PRIMARY KEY,
    temperatura DECIMAL(5,5),
    precipitacao DECIMAL(5,5),
    probabilidade DECIMAL(5,5),
    visibilidade DECIMAL(5,5),
    cobertura DECIMAL(5,5),
    velocidade DECIMAL(5,5),
    direcao DECIMAL(5,5)
);

CREATE TABLE distancia (
    id_distancia INT AUTO_INCREMENT PRIMARY KEY,
    valor DECIMAL(5,1)
);

CREATE TABLE Voo (
    id_voo INT AUTO_INCREMENT PRIMARY KEY,
    nr_voo INT,
    id_distancia INT,
    hora_partida_real DECIMAL(5,1),
    hora_partida_esperada DECIMAL(5,1),
    hora_chegada_real DECIMAL(5,1),
    hora_chegada_esperada DECIMAL(5,1),
    FOREIGN KEY (id_distancia) REFERENCES RDY.distancia(id_distancia)
ON DELETE CASCADE
);

CREATE TABLE Viagem (
    id_viagem INT AUTO_INCREMENT PRIMARY KEY,
    avaliacao DECIMAL(5,1),

```

```

cancelado DECIMAL(6,1),
desviado DECIMAL(6,1),
atraso_companhia_aerea DECIMAL(5,1),
atraso_metereologia DECIMAL(5,1),
atraso_SNA DECIMAL(5,1),
atraso_seguranca DECIMAL(5,1),
atraso_voo_anterior DECIMAL(5,1),
id_voo INT,
nome_companhia_aerea VARCHAR(150),
aeroporto_origem CHAR(3),
aeroporto_destino CHAR(3),
Modelo_Avião VARCHAR(100),
id_clima INT,
FOREIGN KEY (id_voo) REFERENCES RDY.Voo(id_voo) ON DELETE
CASCADE,
FOREIGN KEY (nome_companhia_aerea) REFERENCES
RDY.Companhia_Aerea(nome) ON DELETE CASCADE,
FOREIGN KEY (aeroporto_origem) REFERENCES
RDY.Aeroporto(Codigo_IATA) ON DELETE CASCADE,
FOREIGN KEY (aeroporto_destino) REFERENCES
RDY.Aeroporto(Codigo_IATA) ON DELETE CASCADE,
FOREIGN KEY (Modelo_Avião) REFERENCES RDY.Aeronave(Modelo) ON
DELETE CASCADE,
FOREIGN KEY (id_clima) REFERENCES RDY.clima(id_clima) ON DELETE
CASCADE
);

```

```

CREATE TABLE VooRDY (
    id_voo INT AUTO_INCREMENT PRIMARY KEY,
    nr_voo INT,
    data_partida DATE,
    nome_companhia_aerea VARCHAR(100),
    aeroporto_origem CHAR(3),
    aeroporto_destino CHAR(3),
    hora_partida_real DECIMAL(5,1),
    hora_partida_esperada DECIMAL(5,1),
    hora_chegada_real DECIMAL(5,1),
    hora_chegada_esperada DECIMAL(5,1),
    cancelado DECIMAL(6,1),
    desviado DECIMAL(6,1),
    distancia DECIMAL(6,1),
    atraso_companhia_aerea DECIMAL(5,1),
    atraso_metereologia DECIMAL(5,1),
    atraso_SNA DECIMAL(5,1),
    atraso_seguranca DECIMAL(5,1),
    atraso_voo_anterior DECIMAL(5,1),
    Modelo_Avião VARCHAR(100),
    avaliacao DECIMAL(5,1)
);

```

```
#DROP TABLE Voo;

USE HST;

CREATE TABLE Companhia_AereaHST (
    id_companhia_aerea INT PRIMARY KEY,
    nome VARCHAR(150) NOT NULL,
    país VARCHAR(50) NOT NULL,
    Modificação VARCHAR(100),
    DataHoraModificação VARCHAR(100)
);

CREATE TABLE AeronaveHST (
    Modelo VARCHAR(100),
    Fabricante VARCHAR(100),
    Modificação VARCHAR(100),
    DataHoraModificação VARCHAR(100)
);

CREATE TABLE AeroportoHST (
    ID_aeroporto INT PRIMARY KEY,
    Código_IATA CHAR(3),
    Nome VARCHAR(100),
    Cidade VARCHAR(100),
    País VARCHAR(100),
    Latitude DECIMAL(15,10),
    Longitude DECIMAL(15,10),
    Modificação VARCHAR(100),
    DataHoraModificação VARCHAR(100)
);
```

II. Anexo 2 Histogramas

