

MVP Engenharia de Dados

Análise de atrasos de voos com Azure

2023

Pós-Graduação em Ciência de Dados e Analytics
Criado por: Muza Iwanow



MVP – Análise de atrasos de voos

Resumo

Neste relatório, propõe-se a implementação de um pipeline de dados na nuvem, utilizando tecnologias Azure, para analisar atrasos de voos no contexto das companhias aéreas brasileiras. O pipeline engloba a busca, coleta, modelagem, carga e análise dos dados, visando desenvolver um mínimo produto viável (MVP).

O cerne desta iniciativa é identificar padrões de atrasos e cancelamentos de voos no território nacional, utilizando os serviços de nuvem oferecidos pela plataforma Azure. Essa análise visa fornecer insights para aprimorar decisões operacionais e estratégicas.

Com ênfase no banco de dados SQL, a implementação do pipeline utilizará uma transformação de carga ETL para extrair as informações necessárias para responder questionamentos levantados.

“Azure, Aviação, Banco de dados, SQL, ETL, serviço de nuvem”

Objetivos

O objetivo principal do MVP é responder perguntas de negócios através de um banco de dados extraídos da ANAC (Agência Nacional de Aviação Civil) utilizando serviço de nuvem, realizando todas as etapas propostas do trabalho. Dentre as perguntas levantadas temos alguns pontos a serem considerados:

1. Identificar as companhias aéreas com mais atrasos e cancelamentos de voo
2. Determinar a taxa de pontualidade das companhias aéreas nacionais e estrangeiras que operam no brasil.
3. Analisar as rotas mais afetadas por atrasos e cancelamentos de voo
4. Analisar a distribuição de atrasos por dia da semana

Perguntas de negócios

1. **Qual é a taxa de pontualidade das companhias aéreas que operam no Brasil?**
2. **Quais as companhias que mais cancelam voos?**
3. **Quais aeródromos tiveram o maior número de voos no horário?**

Implementação do MVP

1. Busca pelos Dados

Foi escolhida a base de dados VRA (Voo Regular Ativo) uma base de dados composta por informações de voos das empresas de transporte aéreo que apresa os cancelamentos e horários em que os voos ocorreram.

A ANAC (Agência Nacional de Aviação Civil) disponibiliza esses dados públicos pelo portal em formato “.csv”.

Para acessar os dados diretamente no site da ANAC pelo endereço abaixo:

<https://www.gov.br/anac/pt-br/assuntos/dados-e-estatisticas/historico-de-voos>

Além disto, como os valores retornados são siglas ou códigos, também saremos as planilhas com as siglas com os nomes das empresas aéreas e a planilha com os nomes dos aeródromos, também disponibilizados pela ANAC.

2. Coleta e preparação dos ambientes

Uma vez definido o conjunto de dados, iniciou a configuração do ambiente nuvem para armazená-los.

O ambiente escolhido foi o Azure, uma plataforma que se destaca por ser user friendly e por oferecer um crédito para novos usuários explorarem o ambiente. Com a criação da conta gratuita foi realizado a configuração do Azure Storage Account para armazenamento dos dados.

Este armazenamento foi estabelecido em um novo grupo de recursos denominado “iwanowrg”, com uma conta de armazenamento denominada “iwanowstorage” situada na região (US) East US. As configurações da conta de armazenamento foram mantidas na versão standard e padrão, conforme ilustrado na figura 1.

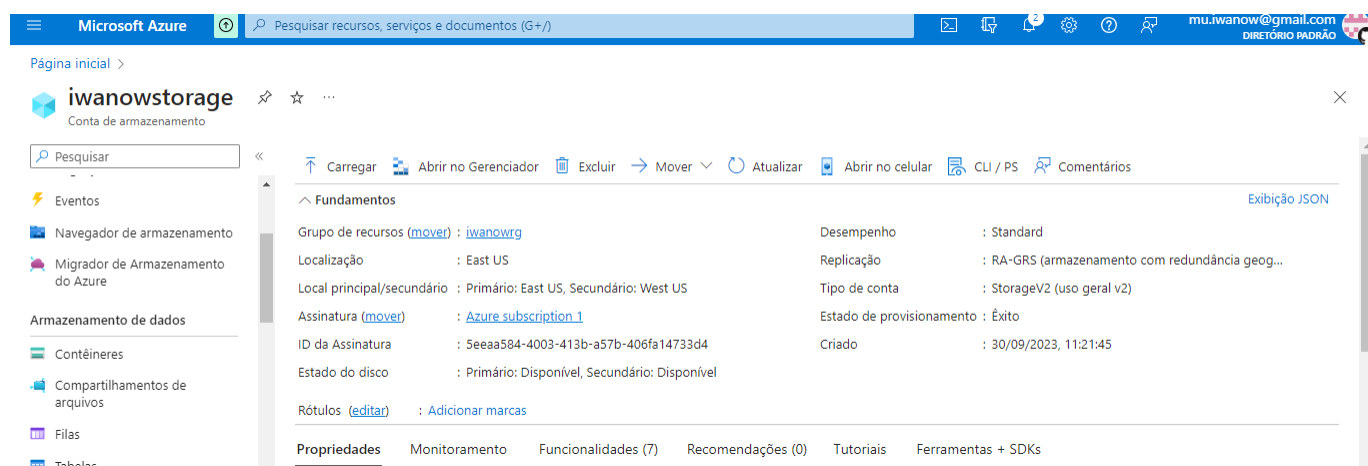


Figura 1 - configuração ambiente Azure - storage

Com o storage devidamente criado, realizou-se o upload dos arquivos provenientes do site da ANAC que foram alocados no container denominado “iwanowcontainer”.

Posteriormente, avançou-se para a fase de criação do SQL database, onde foi configurado o servidor SQL (ver figura 2) com o nome “iwanowsql”, localizado em (US) East US. A autenticação foi realizada por meio da SQL authentication, permitindo a configuração de um usuário e senha para o login no servidor.

Essa abordagem estratégica no Azure proporciona não apenas a robustez do armazenamento, as também a flexibilidade e escalabilidade do SQL data base, ambos configurados para atender as demandas específicas do projeto.

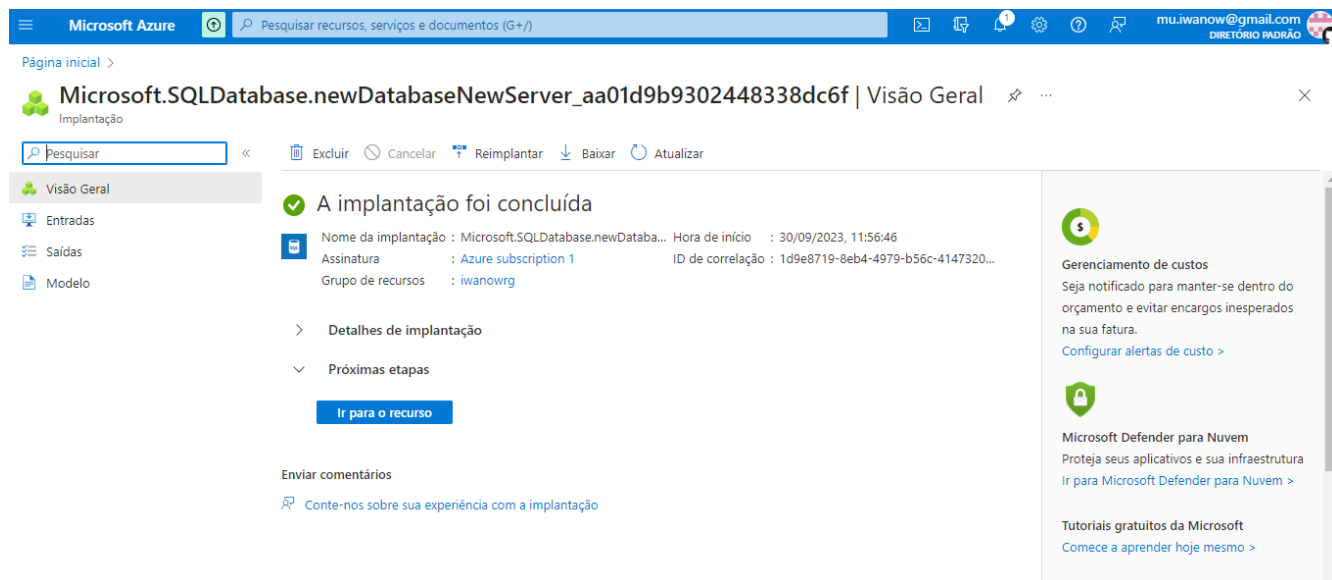


Figura 2 - Servidor SQL

Para finalizar a preparação dos ambientes, foi criado a instancia no Azure Data Factory, a qual foi nomeada de “iwanowfactory” para o grupo de recursos conforme a imagem abaixo.

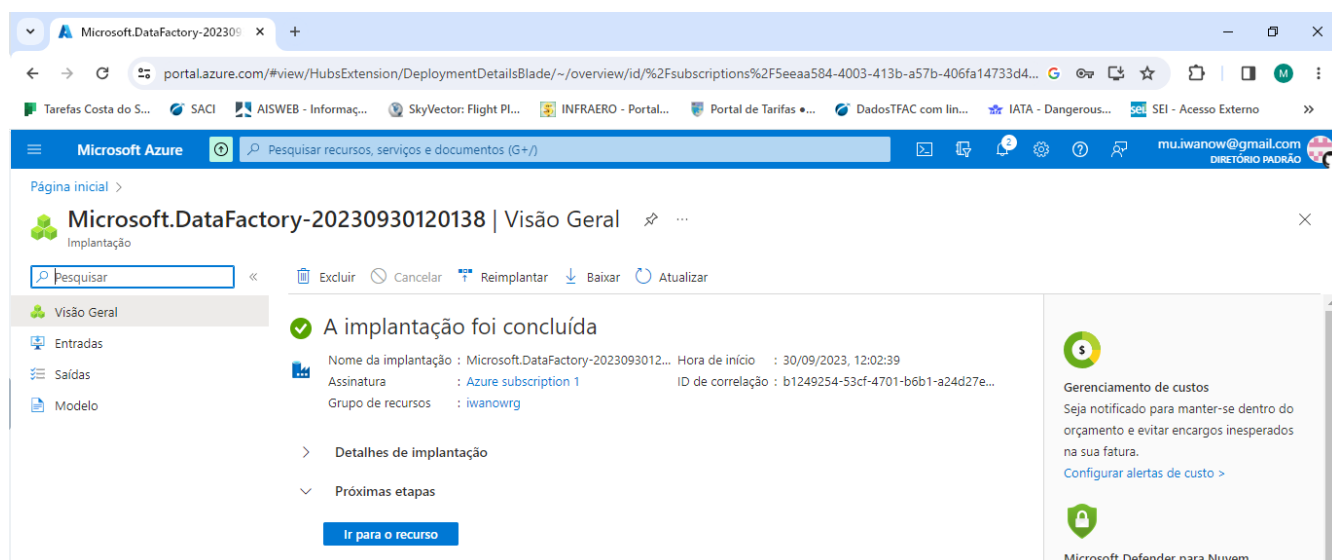
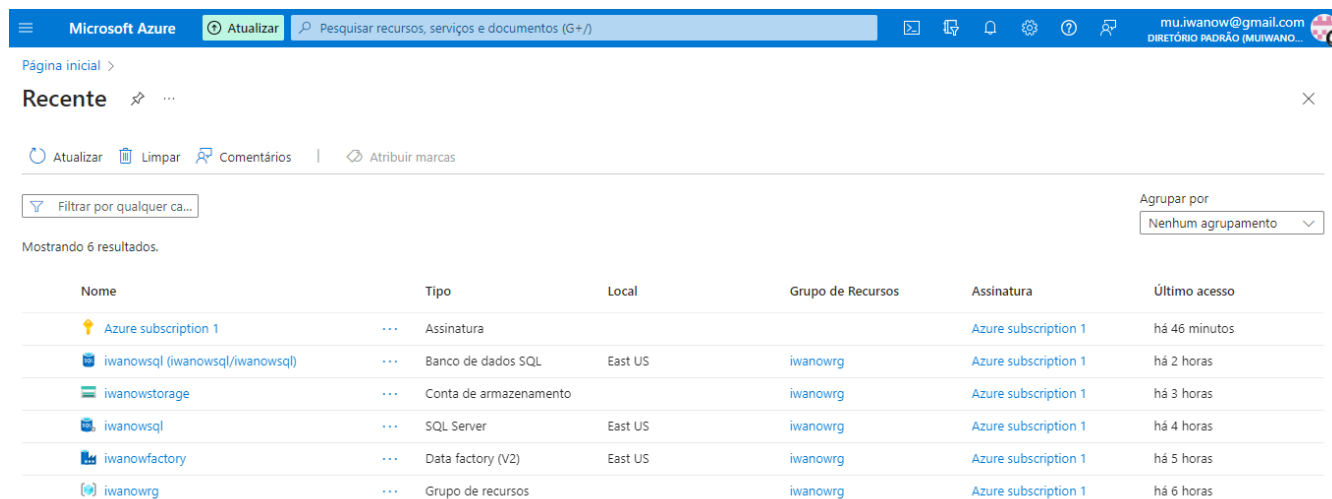


Figura 3 - Azure Data Factory

O Azure Data Factory, uma ferramenta integrada à plataforma Azure, oferece uma gama de recursos e opções para criação de pipelines de dados. Dentre suas características

destaca-se: Orquestração dos dados, Integração com fontes diversas, transformação de dados entre outros.

A figura 4 ilustra todos os ambientes que foram criados no Azure, marcando o ponto de partida para o refinamento dos dados da ANAC.



Nome	Tipo	Local	Grupo de Recursos	Assinatura	Último acesso
Azure subscription 1	Assinatura			Azure subscription 1	há 46 minutos
iwanowsq1 (iwanowsq1/iwanowsq1)	Banco de dados SQL	East US	iwanowrg	Azure subscription 1	há 2 horas
iwanowstorage	Conta de armazenamento		iwanowrg	Azure subscription 1	há 3 horas
iwanowsq1	SQL Server	East US	iwanowrg	Azure subscription 1	há 4 horas
iwanowfactory	Data factory (V2)	East US	iwanowrg	Azure subscription 1	há 5 horas
iwanowrg	Grupo de recursos		iwanowrg	Azure subscription 1	há 6 horas

Figura 4 - Ambiente Azure

3. Modelagem

Dando continuidade no processo de construção do MVP, após a criação do ambiente Azure, vamos adentrar na fase da modelagem, que desempenha papel fundamental na estruturação e organização dos dados, garantindo que estejam prontos para serem utilizados de maneira eficiente.

Inicialmente foi escolhido a modelagem de dados no esquema estrela, onde foi definido a tabela fato e as tabelas dimensões conforme o esquemático representado na figura a seguir (figura 5).

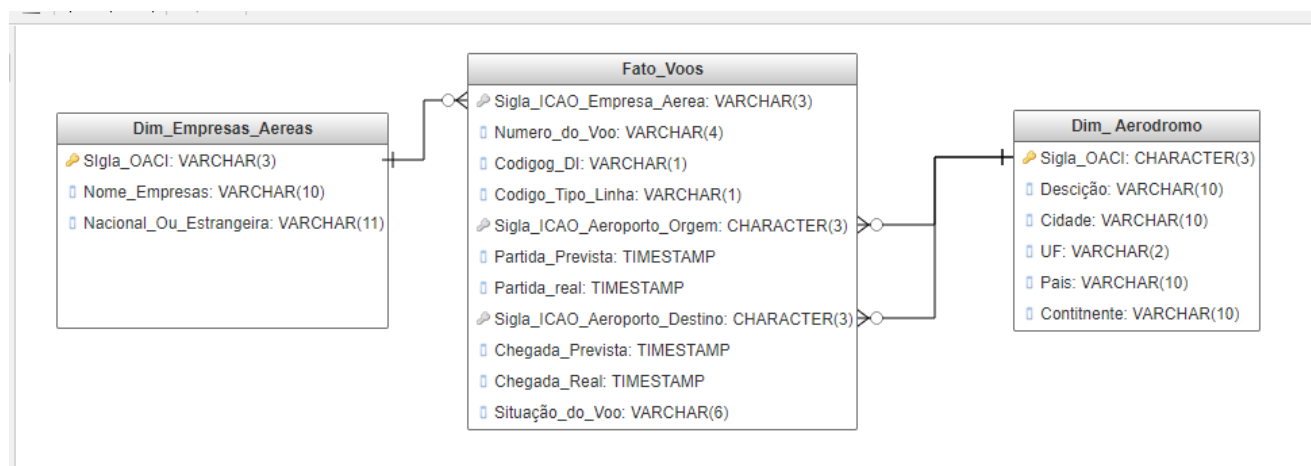


Figura 5 - Esquemático modelo estrela

Em seguida, desenvolvemos um catálogo de dados abrangente, oferecendo descrições detalhas que atuam como guia para a compreensão do banco de dados.

É relevante destacar que, embora o Azure proporcione soluções para diversas necessidades em ambiente de nuvem, atualmente não é mais possível criar novas contas do catálogo de dados do Azure. A plataforma oferece o recurso de catálogo dentro do serviço Microsoft Purview, que oferece governança de dados unificada, mas dado a complexidade do MVP a elaboração do catálogo foi uma etapa customizada visando fornecer informações para o entendimento do banco.

Segue abaixo o catálogo detalhado dos dados que serão utilizados neste projeto.

Tabela 1 - Catálogo VRA

VRA_2023_08			
Item	Atributo	Descrição do Atributo	Domínio
1	Sigla ICAO Empresa Aérea	A sigla ICAO (Organização da Aviação Civil Internacional) da empresa aérea responsável pelo voo	String alfanumérica
2	Número do Voo	Número único associado ao voo para identificação pública e operacional.	String alfanumérica
3	Código DI	Código numérico que referencia no sistema a variável Dígito Identificador	0 (zero) - Etapa Regular 2 (dois) - Etapa Extra 3 (três) - Etapa de Retorno 4 (quatro) - Inclusão de Etapa 6 (seis) - Etapa Não Remunerada Sem Transporte de Objetos

			7 (sete) -Etapa de Voo de Fretamento 9(nove) -Etapa de Voo Charter D -Etapa de Voo Duplicada E - Etapa Não Remunerada Com Transporte de Objetos
4	Código Tipo Linha	Código que classifica o tipo de linha aérea, como doméstica, internacional, regional, etc.	N – Doméstica Mista C – Doméstica Cargueira I – Internacional Mista G – Internacional Cargueira
5	Sigla ICAO Aeroporto Origem	A sigla ICAO do aeroporto de partida	String alfanumérica
6	Partida Prevista	Horário estimado para a decolagem da aeronave	Data e hora
7	Partida Real	Horário real da decolagem da aeronave	Data e hora
8	Sigla ICAO Aeroporto Destino	A sigla ICAO do aeroporto de chegada	String alfanumérica
9	Chegada Prevista	Horário estimado para a aterrissagem da aeronave	Data e hora
10	Chegada Real	Horário real da aterrissagem da aeronave	Data e hora
11	Situação do Voo	Estado atual do voo, podendo incluir status como “Cancelado”, “Realizado”	Cancelado, realizado

Tabela 2 - Glossário de aeródromos

glossario_de_aerodromo			
Item	Atributo	Descrição do Atributo	Domínio
1	Sigla OACI	A sigla OACI (Organização da Aviação Civil Internacional) atribuída ao aeroporto para identificação internacional	String alfanumérica
2	Descrição	O nome oficial do aeroporto	String alfanumérica
3	Cidade	Nome da cidade em que o aeroporto está localizado	String alfanumérica
4	UF	A unidade federativa (estado) à qual a cidade do aeroporto pertence	String alfanumérica
5	País	Nome do País em que o aeroporto está situado	String alfanumérica
6	Continente	O continente ao qual o País do aeroporto pertence	String alfanumérica

Tabela 3 - glossário de empresas aéreas

glossario_de_empresas_aereas			
Item	Atributo	Descrição do Atributo	Domínio
1	Sigla OACI	Sigla utilizada para identificar aeroportos internacionalmente, de acordo com as normas da Organização da Aviação Civil Internacional (OACI)	String alfanumérica
2	Nome Empresas	Refere-se ao nome das empresas aéreas envolvidas nas operações – companhias aéreas	String alfanumérica
3	Nacional ou Estrangeira	Indica se a empresa é de origem nacional (pertencente ao país onde está registrada) ou estrangeira pertencente a um país diferente)	Nacional, estrangeira

4. Carga

Após a configuração dos recursos, empregados a ferramenta Azure Data Factory para realizar a extração, junção e carga dos arquivos no banco de dados SQL precisamente configurado. Para a criação de um pipeline no Azure Data Factory iniciamos com a criação dos três dataset, um para cada arquivo que estava armazenado no Azure Blob Storage.

Para exemplificar, é evidenciado na figura abaixo a criação de um dos dataset com o arquivo em formato “.csv” contendo os dados dos voos (VRA_2023_08.csv). Este dataset atua como uma interface para a extração dos dados a fim de integrar o arquivo no pipeline de dados.

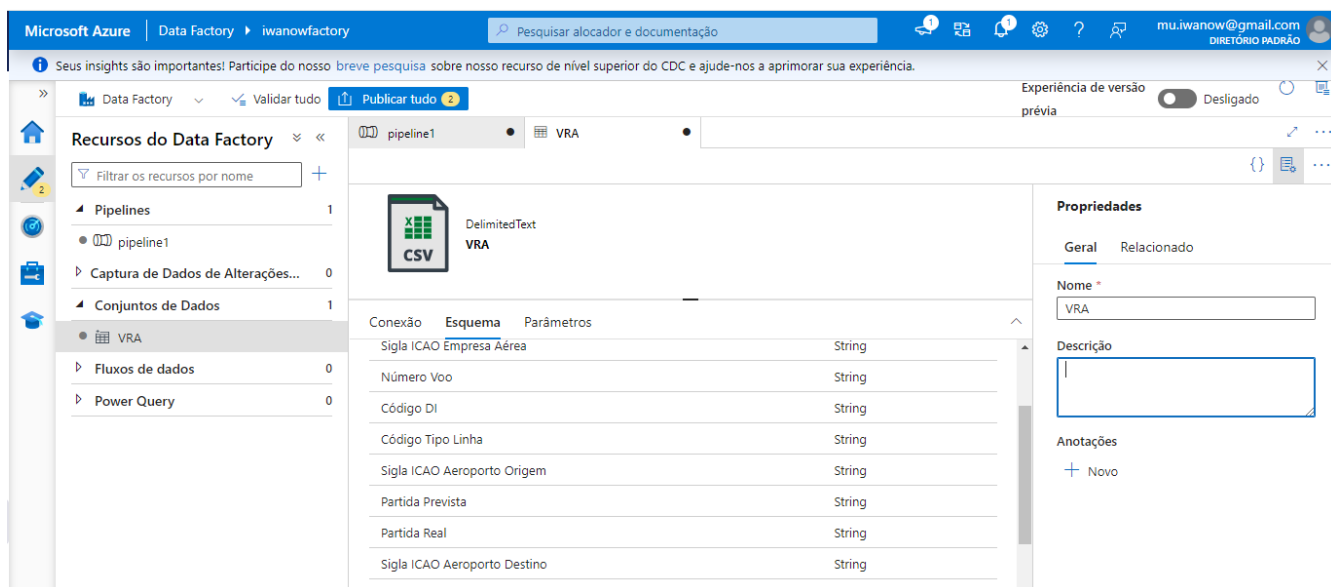


Figura 6 - dataset exemplo

Essa abordagem proporciona uma metodologia para gerenciar a movimentação dos dados de forma automatizada, estabelecendo uma conexão direta entre o Azure Blob Storage e o banco de dados SQL, simplificando o processo de carga e a integridade dos dados durante o fluxo de trabalho.

Para as duas planilhas e formato “xls” o processo foi o mesmo do arquivo “.csv”, destaca-se apenas a necessidade de informar os campos da planilha que contém os dados, por exemplo no caso na planilha glossário de companhias aéreas foi A1:C174, deste modo é possível verificar que existem 174 companhias aéreas que operam no Brasil, esta informação é importante para verificação do banco de dados na próxima etapa.

Além disto, é importante marcar a opção: “primeira linha é um cabeçalho” na hora de executar a conexão, deste modo ao visualizar o esquema o nome das colunas será mostrado de forma correta conforme as figuras 7 e 8.

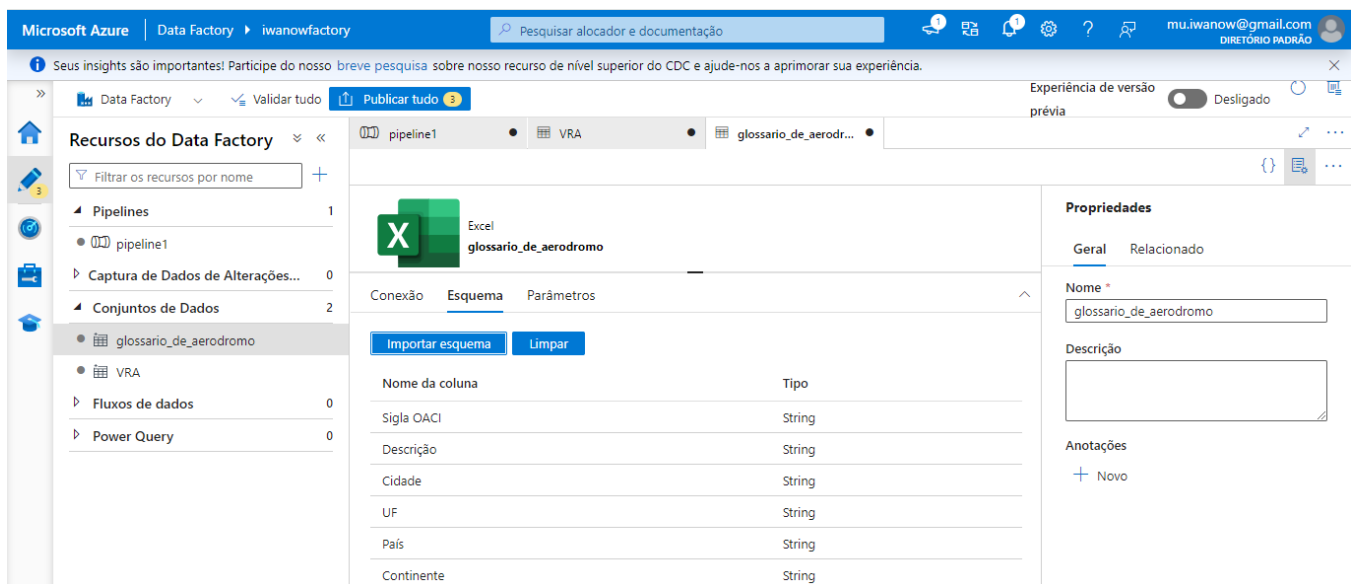


Figura 7 - dataset glossário de aeródromos

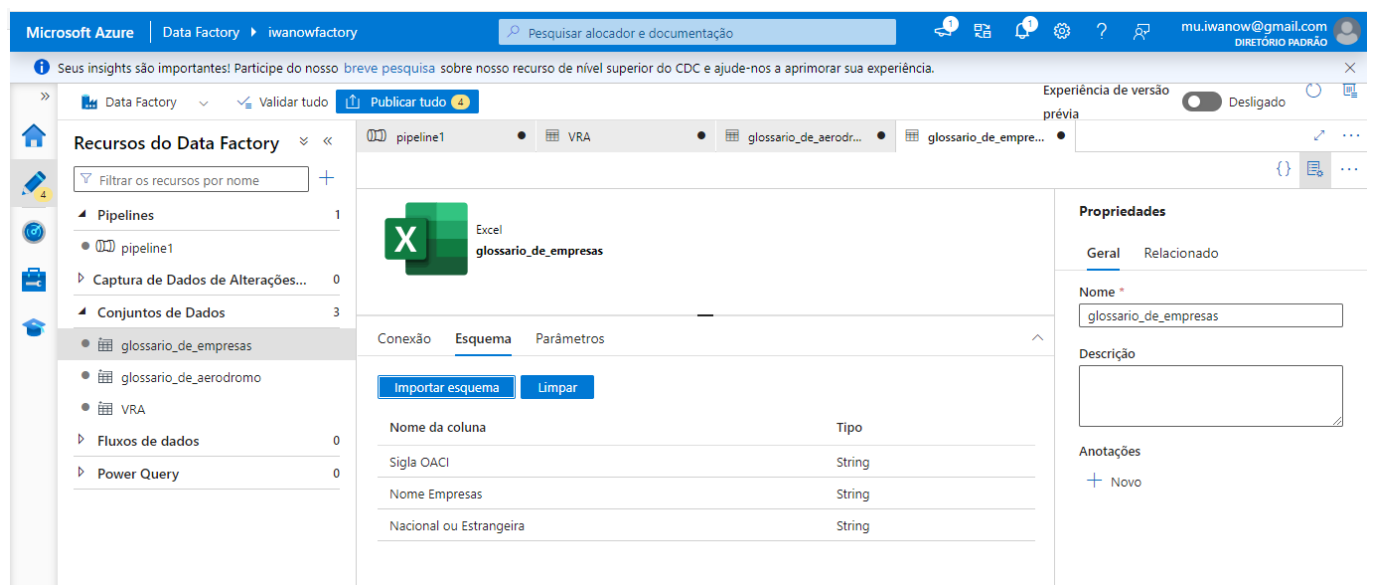


Figura 8 - dataset glossário de empresas aéreas

No fluxo de dados, foi realizado dois processos de junção (join). Para a integração eficaz dos dados, foi necessário realizar a conversão das colunas relacionadas aos horários de partida e chegada previstos e reais para o tipo timestamp. Essa transformação visa garantir a uniformidade dos dados e facilitar a análise temporal durante o processo.

Após a transformação, foi criado um sink no Azure Data Factory direcionando os dados para o banco de dados SQL. Este processo resultou na criação de uma nova tabela denominada “voos”. Esta tabela, representando a tabela fato em nosso esquema estrela, é agora o repositório central para os dados consolidados de voos conforme podemos ver na figura 9 a seguir.

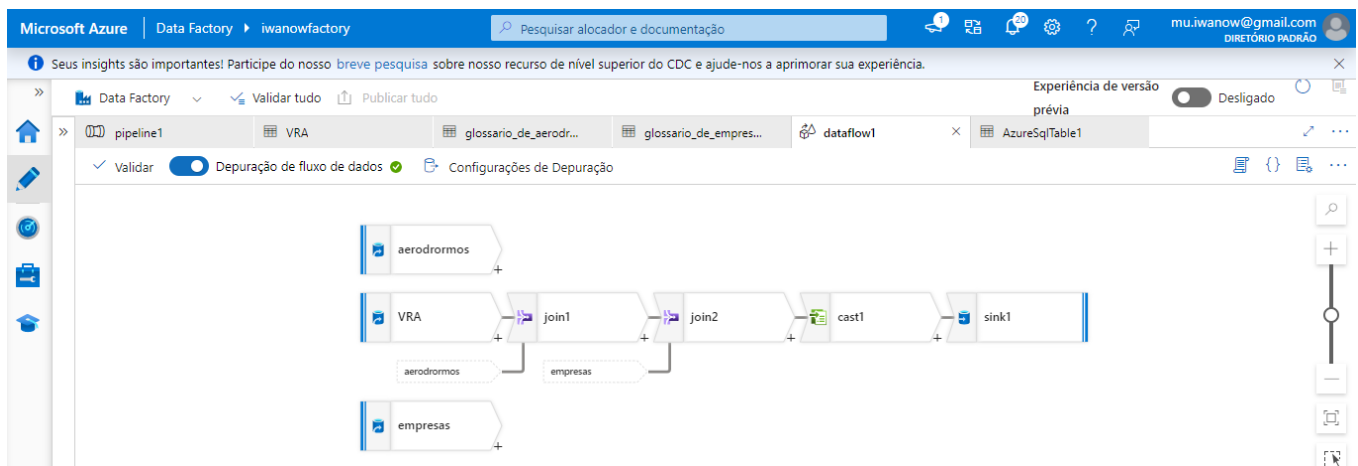


Figura 9 – dataflow

Após a configuração do dataflow, procedemos à montagem do pipeline para estabelecer a conexão com o Azure data Studio, que é uma ferramenta de gerenciamento de dados da Microsoft. Essa etapa permite a análise e consulta dos dados consolidados no ambiente Azure SQL Database.

O pipeline atua como uma ponte entre o processo de ETL (extração, transformação e carga) realizado no Data Factory e a interface do Data Studio.

Ao configurar o dataflow por meio deste pipeline (ver figura 10), criamos um fluxo contínuo de dados, que permite que as perguntas do projeto sejam respondidas.

Nome da atividade	Status da atividade	Tipo de atividade	Início da execução	Duração
dataflow1	Bem-sucedido	Fluxo de dados	9/30/2023, 4:35:42 PM	49s

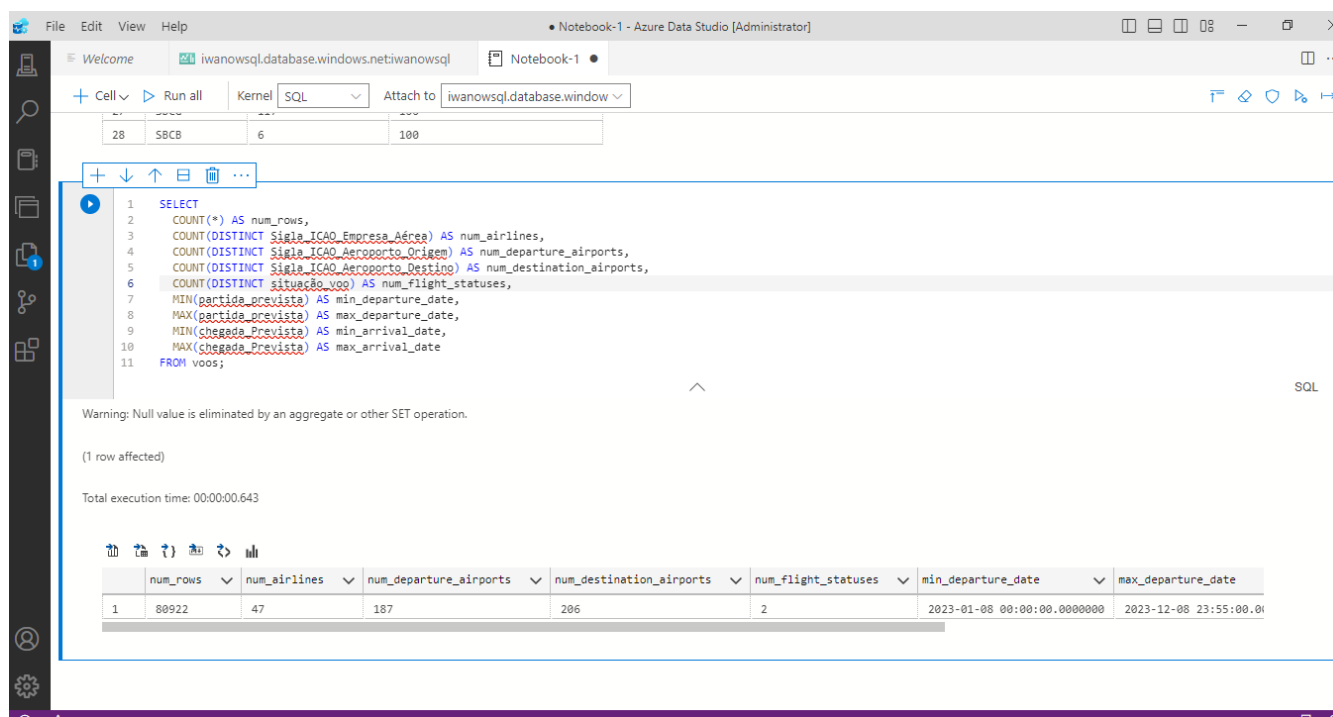
Figura 10 – pipeline

5. Análise

Após a execução bem-sucedida do pipeline, foi estabelecido a conexão entre o banco de dados e o Azure Data Studio para a execução das queries analíticas. A verificação do status “bem-sucedido” indica que o processo de ETL foi concluído sem contratempos. No entanto, é necessário garantir a integridade desses dados.

Para tal, foi feita uma análise nos dados da planilha, identificando os erros ou inconsistências para assegurar que os dados no banco estejam prontos para receber as consultas.

Na figura 11 abaixo ilustra a query para verificar os dados no banco de dados. Essa query oferece uma visão dos dados armazenados permitindo analisar a informação após o processo de ETL.



The screenshot shows the Azure Data Studio interface with a SQL query executed in a notebook. The query is a SELECT statement that aggregates data from a table named 'voos'. It includes counts for various attributes and minimum/maximum dates for flight status transitions.

```
1 SELECT
2   COUNT(*) AS num_rows,
3   COUNT(DISTINCT Sigla_ICAO_Empresa_Aerea) AS num_airlines,
4   COUNT(DISTINCT Sigla_ICAO_Aeroporto_Origem) AS num_departure_airports,
5   COUNT(DISTINCT Sigla_ICAO_Aeroporto_Destino) AS num_destination_airports,
6   COUNT(DISTINCT situacao_voo) AS num_flight_statuses,
7   MIN(partida_previsa) AS min_departure_date,
8   MAX(partida_previsa) AS max_departure_date,
9   MIN(chegada_previsa) AS min_arrival_date,
10  MAX(chegada_previsa) AS max_arrival_date
11 FROM voos;
```

Below the query, a warning message states: "Warning: Null value is eliminated by an aggregate or other SET operation." and "(1 row affected)". The total execution time is 00:00:00.643.

	num_rows	num_airlines	num_departure_airports	num_destination_airports	num_flight_statuses	min_departure_date	max_departure_date
1	80922	47	187	206	2	2023-01-08 00:00:00.0000000	2023-12-08 23:55:00.0000000

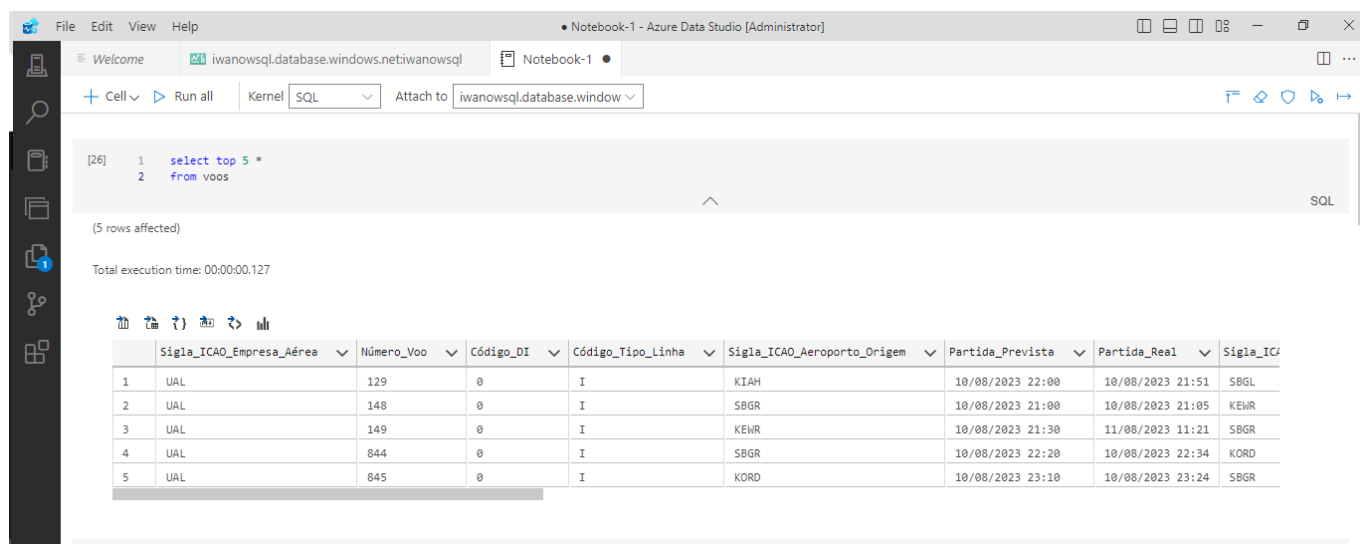
Figura 11 - Data Studio - Query consistência dos dados

Esse procedimento de verificação e consulta no Azure Data Studio procurou garantir a confiabilidade dos dados antes de prosseguir para análises.

A query realizada englobou pontos da tabela, proporcionando uma visão abrangente e estatística do banco de dados fornecendo os resultados:

1. Número total de linhas na tabela, proporcionando uma contagem dos voos registrados;
2. Número de companhias aéreas distintas, como dito anteriormente não poderiam passar de 174, contabilizando o número de companhias aéreas distintas presente na tabela “voos”;
3. Número de aeroportos de origem distintos;
4. Contagem de situações de voo distintas: uma vez que foi levantado que a situação de voo só poderia ter o status “realizado” e “cancelado, foi observado apenas duas situações distintas como era de se esperar.
5. As datas mais recentes e mais antigas tanto da partida quanto da chegada prevista para verificação dos registros de data e hora, uma vez que o banco de dados era referente ao mês de agosto de 2023.

Essa consulta forneceu uma estatística do banco para verificar os principais registros da tabela “voo”. Em seguida, foi realizada uma consulta para verificar os 5 primeiros registros da planilha afim de verificar se todos os campos estão corretos conforme a figura 12.



SQL

(5 rows affected)

Total execution time: 00:00:00.127

	Sigla_ICAO_Empresa_Aérea	Número_Voo	Código_OI	Código_Tipo_Linha	Sigla_ICAO_Aeroporto_Origem	Partida_Prevista	Partida_Real	Sigla_ICD
1	UAL	129	0	I	KIAH	10/08/2023 22:00	10/08/2023 21:51	SBGL
2	UAL	148	0	I	SBGR	10/08/2023 21:00	10/08/2023 21:05	KEWR
3	UAL	149	0	I	KEWR	10/08/2023 21:30	11/08/2023 11:21	SBGR
4	UAL	844	0	I	SBGR	10/08/2023 22:20	10/08/2023 22:34	KORD
5	UAL	845	0	I	KORD	10/08/2023 23:10	10/08/2023 23:24	SBGR

Figura 12 - query top 5 da tabela voos

Após as verificações, uma vez que não foi encontrada inconsistências iniciou a pesquisa no banco para responder as perguntas elaboradas no objetivo do MVP.

1. Quais companhias têm mais voos atrasados e a quantidade de voos no horário dessas companhias?

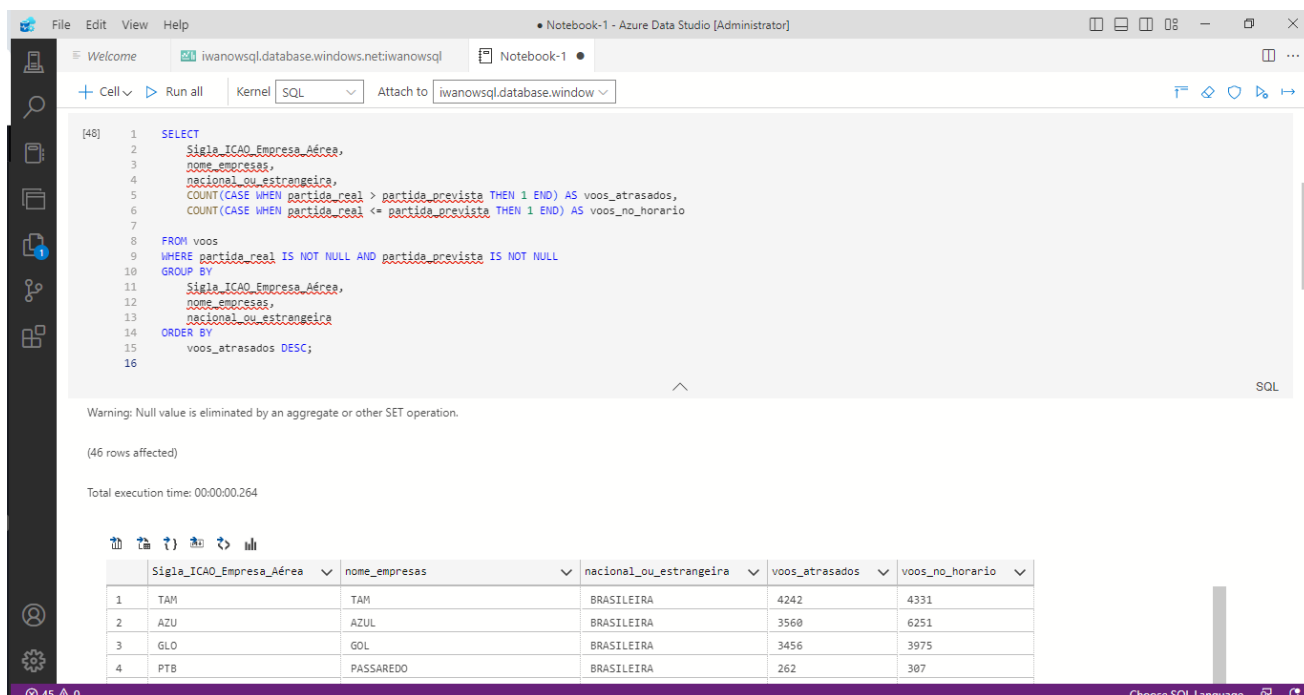


Figura 13 - Query questionamento atraso de voo

Graficamente para melhor visualização podemos observar que a TAM é a companhia que mais atrasou voos no período, seguido da Azul e Gol.

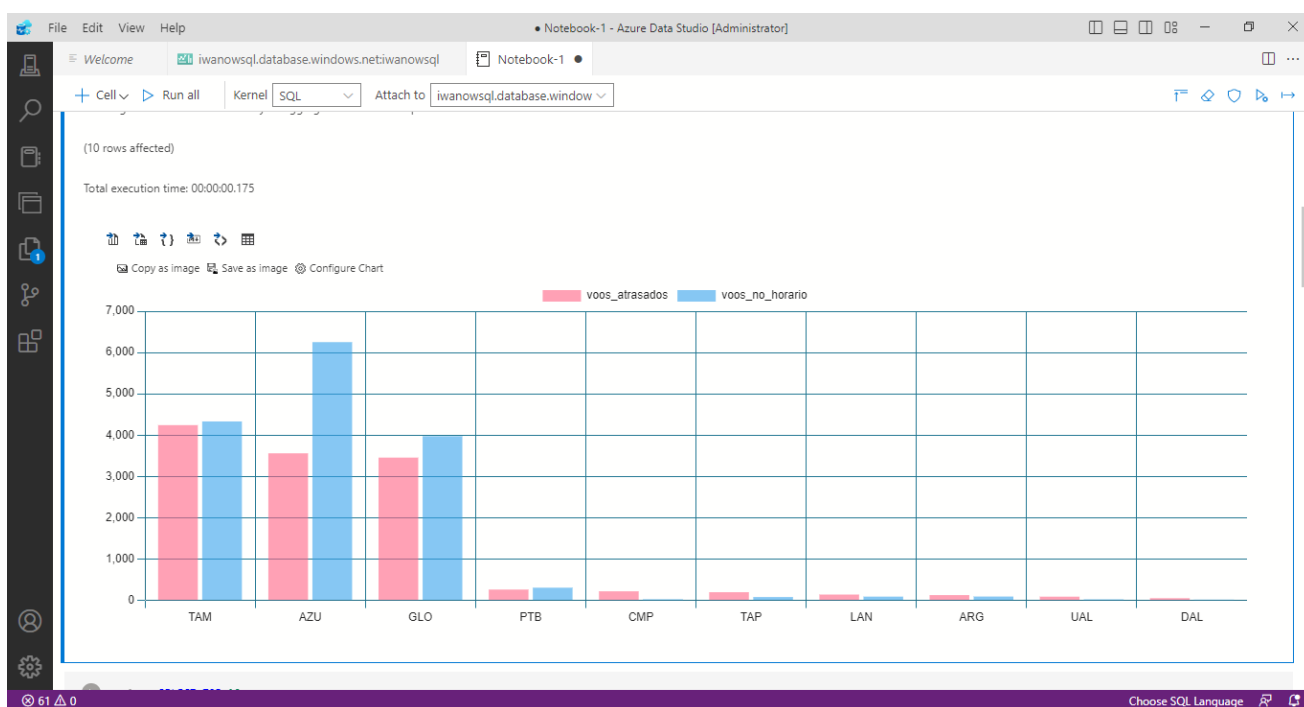


Figura 14 - gráfico atraso de voo e voos no horário

2. Companhias aéreas que mais tiveram seus voos cancelados?

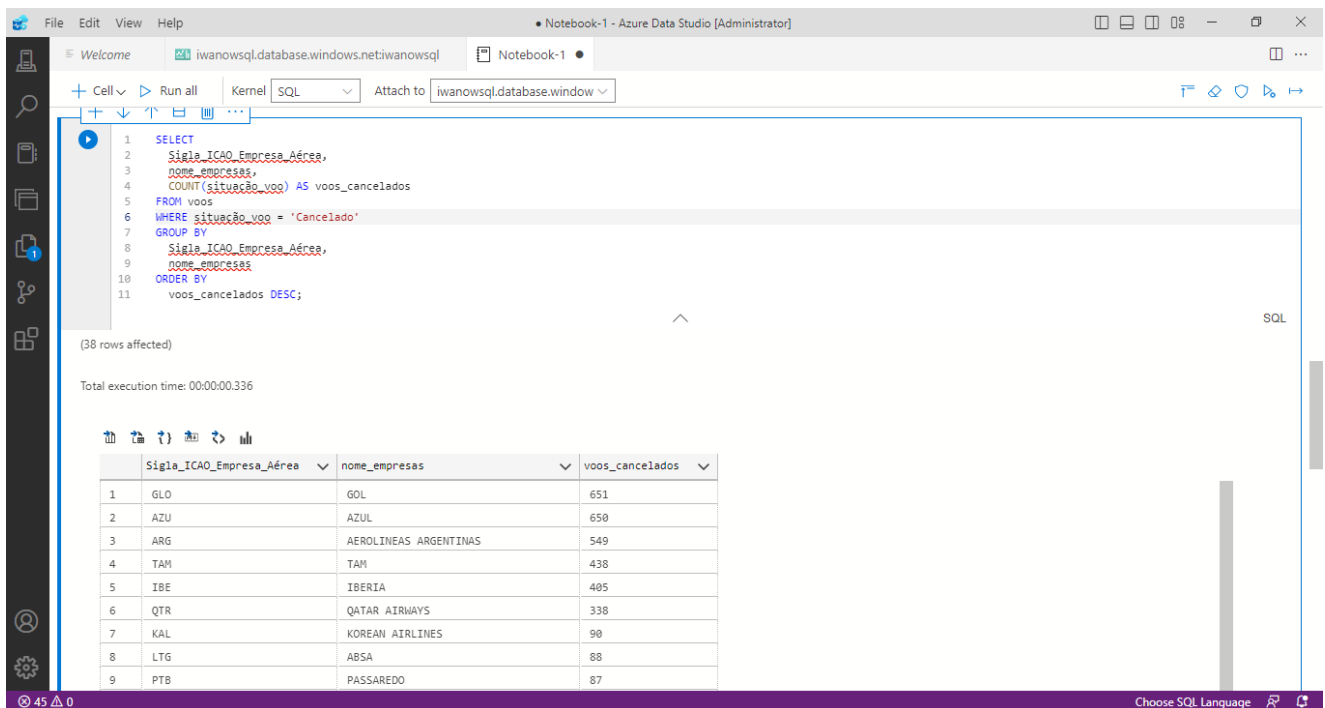


Figura 15 - Query cancelamento de voo

Na figura 16, abaixo, podemos ver o gráfico plotado da informação da query referente as companhias aéreas que mais tiveram voos cancelados no período, conforme ilustrado na figura 15 a cima.

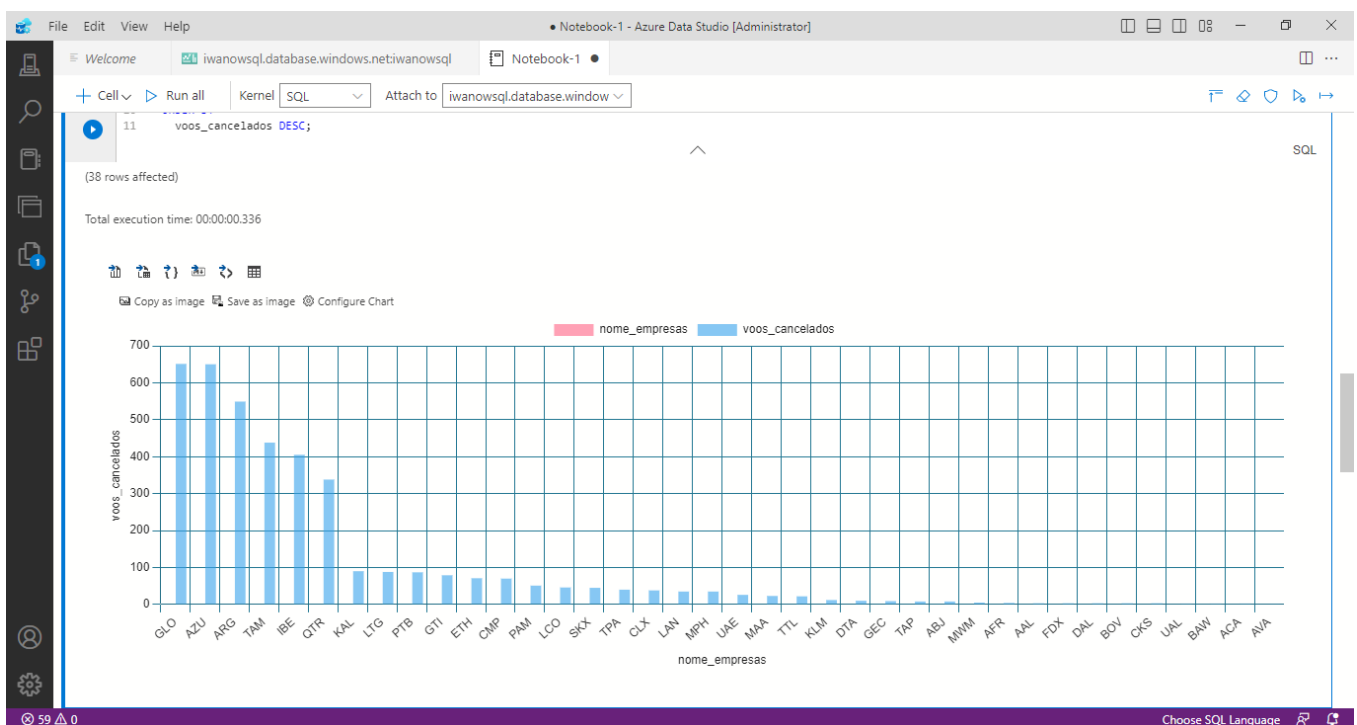


Figura 16 - Gráfico companhia aérea com maior número de cancelamentos

Podemos observar que as três companhias que mais tiveram voos cancelados foram a Gol, a Azul e companhia aérea argentina, Aerolineas Argentinas.

3. Quais aeródromos tiveram o maior número de voos no horário?

Para responder esse questionamento foi feito uma query selecionando as siglas dos aeródromos e verificado quantos voos saíram no horário previsto.

O aeródromo SBSR – Professor Eriberto Manoel Reino, localizado em São Jose do Rio Preto/ SP foi o aeródromo que mais teve voos no horário, conforme podemos observar na figura 17.

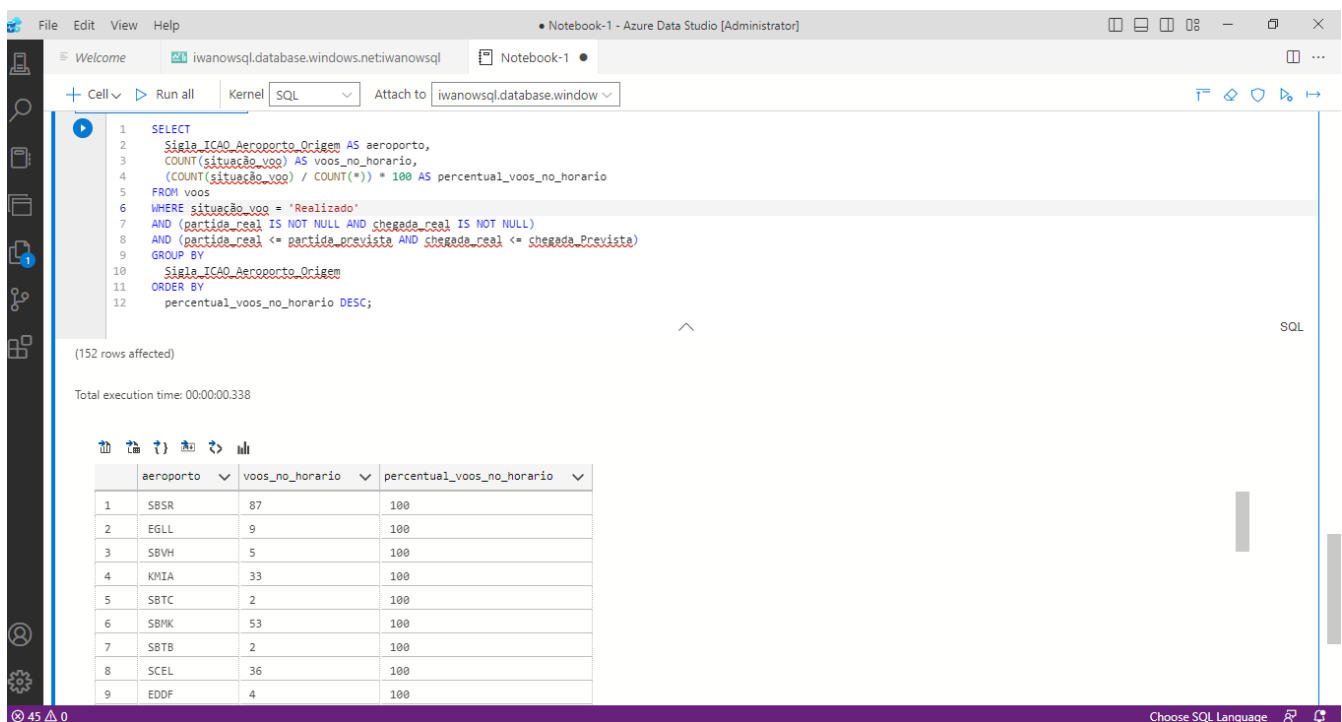


Figura 17 - Aeródromos com maior número de voos dentro do previsto

Análise de Custo do Azure

Uma vez que os serviços do Azure são cobrados, verifiquei a necessidade de fazer um levantamento dos custos do projeto afim de exemplificar quais recursos são os que acarretam a maior cobrança, sendo necessário sempre uma verificação e análise dos recursos ainda disponíveis.

A própria plataforma disponibiliza de forma gráfica todas as informações sobre cobrança, o que torna o controle dos recursos mais fácil de administrar.

Nas figuras 18 e 19 ilustra o consumo dos créditos que foram disponibilizados pela Azure para o projeto.

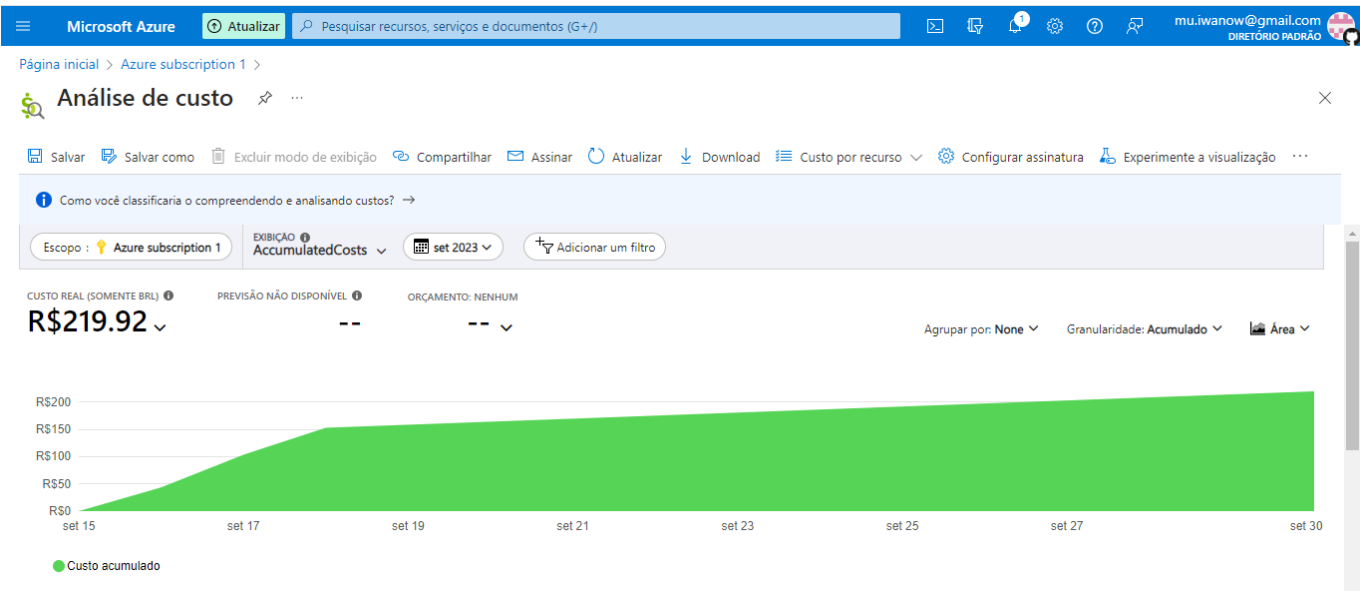


Figura 18-Análise de custo do projeto ao longo do tempo

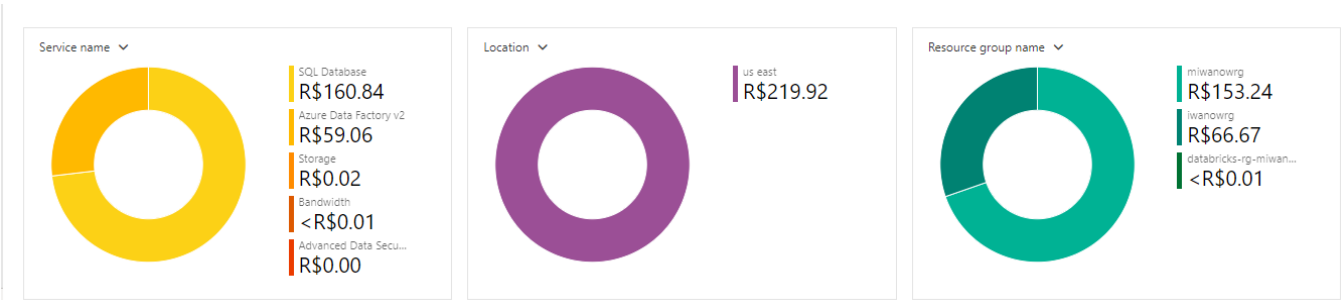


Figura 19 - Análise dos custos por serviço

Autoavaliação

Ao iniciar o MVP de engenharia de dados, me deparei com um cenário desafiador que exigiu aprendizado significativo em vários domínios. Essa foi sem dúvida a etapa da primeira sprint da Pós-Graduação de Ciência de dados e Analytic que mais me proporcionou uma visão de como funciona um banco de dados e nuvem e aplicação dos conceitos vistos no material de estudo, mas também apresentou obstáculos que merecem reflexão.

Como desafio enfrentado, destaco a modelagem do sistema, que foi uma tarefa complexa, pois envolvia a compreensão dos dados e a tradução dos requisitos para um esquema de banco de dados. A necessidade de alinhar a estrutura do banco com as perguntas proposta e a dificuldade de entender os conceitos por trás do esquema relacional foram as maiores dificuldades dessa etapa de modelagem.

Além disto, enfrentei limitações com a linguagem SQL, pois minha experiência inicial era limitada, o que refletiu nos desafios enfrentados durante a implementação. Foi necessário aprofundar os conhecimentos em SQL para manipulação dos dados e entender o funcionamento da ferramenta Azure Data Studio para gerar as informações de forma gráfica sem precisar utilizar outro recurso como importar uma biblioteca gráfica ou exportar para um BI.

Quanto ao uso e escolha da plataforma Azure, foi crucial para o funcionamento do projeto, mesmo não tendo familiaridade, com a ajuda dos orientadores da disciplina do MVP foi possível realizar as configurações necessárias para implementar o banco de dados corretamente na nuvem.

O maior desafio foi a parte de transformação de carga, pois os conceitos associados a ETL (Extract, Transform, Load) exigiram uma absorção rápida de conhecimento, que ainda não está satisfatória, deixando este ponto para aprendizado.

Apesar das inúmeras dificuldades e desafios na configuração da plataforma, foi possível responder algumas das perguntas propostas. A análise dos dados proporcionou alguns insights sobre o setor da aviação fazendo com que o objetivo do MVP fosse atendido.

A preocupação com os custos dos recursos da Azure foi constantemente considerada, por isso foi implementado medidas para otimizar os usos dos recursos para garantir uma abordagem econômica do projeto. O maior custo foi proveniente do SQL Data base, que acabou ficando ligado por alguns dias gerando custo significativo.

Com relação a projetos futuros e melhoria contínua, reconheço a necessidade contínua de aprimorar os conhecimentos em linguagem SQL e aprofundar nos recursos da plataforma Azure ou de outra plataforma de nuvem.

Além disto, pretendo configurar o catálogo de dados e aumentar e aumentar o banco de dados para abranger um período mais extensos de análise de dados de voo. Essa ampliação será necessária para responder questionamentos ao longo do tempo e explorar a relação entre os meses do ano e fatores externos, tais como estações do ano, período de chuva, períodos de férias escolares, feriados entre outros, dando uma visão mais abrangente das tendencias relacionadas ao desempenho dos voos.