Universidade do Minho
Escola de Engenharia

Licenciatura em Engenharia e Gestão de Sistemas de Informação

Assignment 1 - MySQL

Analysis of flights from and to Brazil in the year 2020

Data Bases 2

Ano Letivo de 2023/2024

e11273, Tomaso Stefanizzi

Guimarães, Março de 2024

Index

1 Introduction

This assignment consists in finding a dataset online and executing some queries in order to answer some chosen analytical questions. In the context of this project, I wanted to analyze the flights from and to Brazil in the year 2020. This data is publicly available in the website of Agência Nacional de Aviação Civil (ANAC) [1], the regulatory agency responsible for overseeing and regulating civil aviation activities within the country of Brazil. The airlines must share with ANAC several information in order to ensure safety, compliance with regulations, and protection of the interests of both the industry and passengers. In the following sections I'll provide the data source, the data wrangling procedure, the relational and conceptual model of the dataset, a data dictionary, as well as some analytical questions and answers.

2 MySQL

2.1 Chosen dataset

As I anticipated in the introduction, I chose to work with flights data from and to Brazil. The chosen dataset was found on kaggle (click here to access it). The original data contains the hystorical data of flights from the year 2000 until 2021.
The original dataset was subdivided in six tables:

- **DW_ARPT_DEST**: it contains the destinations of the flights.
- **DW_ARPT_ORIGEM**: it contains the airports of departure for the flights.
- **DW_EMPRESA**: it provides the informations about the airlines.
- **DW_EQPT**: it contains models of the planes used for the flights.
- **DW_TIPO_LINHA**: it describes some informations about the route, like its purpose and if it contains passengers
- **DW_VOOS**: each row represents a single flight. this table contains all the flights from the year 2000 to the year 2021, as well as some informations about them.

2.2 Data Wrangling

In order to work with a relatively concise dataset, I decided to filter the data, by considering only the flights taken in the year 2020. The flights table has been drastically re-scaled, from 5GB of data to approximately 55Mb (still, it's a fair amount of data points, since we have 367'029 rows, hence single flights). As a consequence, I just kept the airports and airlines which where present in this sub-set of the original dataset.

Moreover, the original dataset presented some inconsistencies that needed to be solved before proceeding with the population of the SQL Database. Some examples are duplicated rows for the same airport, or even wrong airport identifiers (both iata and icao identifiers).

Lastly, I decided to merge the tables **DW_ARPT_ORIGEM** and **DW_ARPT_DEST** in a unique table called **ariport**, since it didn't make sense to consider the origin and the destination as different entities, and finally I dropped some of the columns from the tables that where not that interesting or that contained too many null values.

The processed table are now:

- **airport**: as I said, it's a merge between the two tables from destinations and origin
- **company**: cleaned version of `DW_EMPRESA`
- **equipment**: cleaned version of `DW_EQPT`
- **line**: cleaned version of `DW_TIPO_LINHA`
- **flights**: cleaned and filtered version of `DW_VOOS`

2.3 Conceptual and Relational Database Model

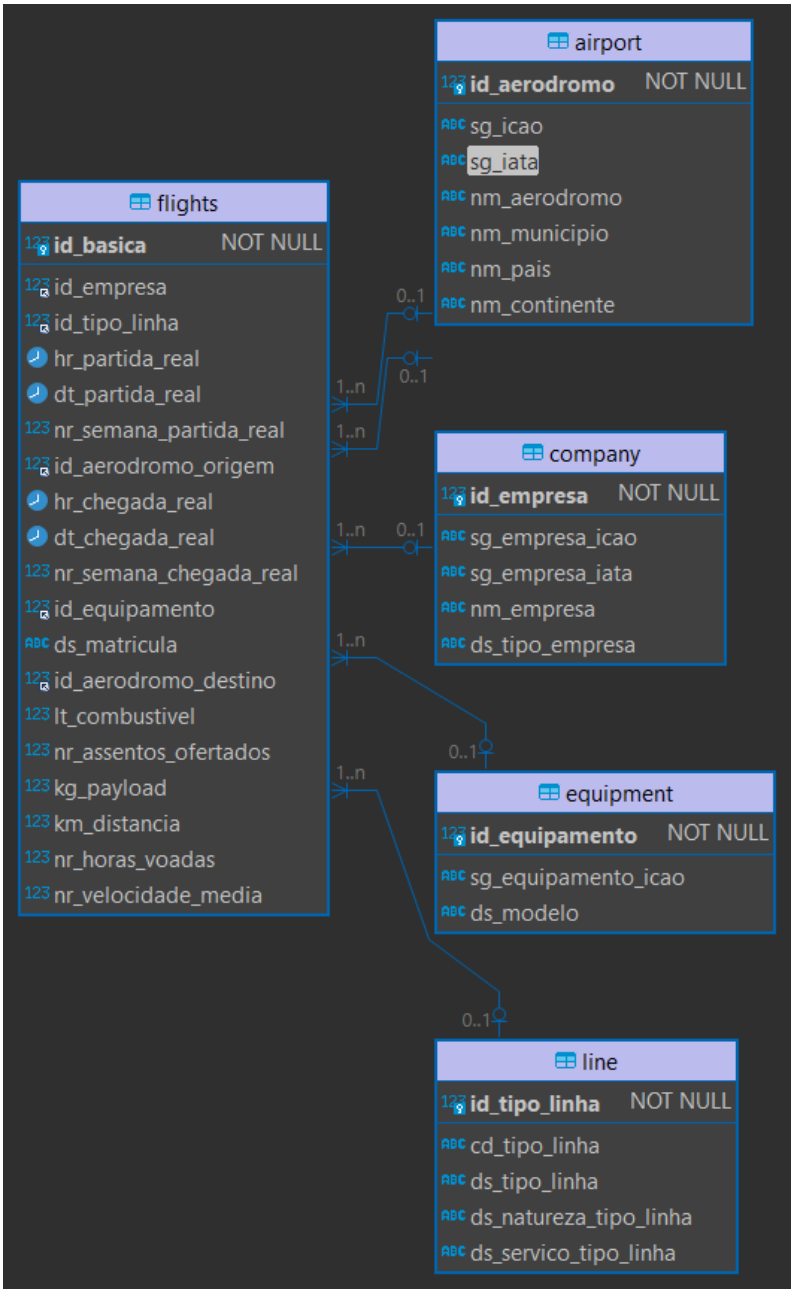

Figura 2.1: ER-Diagram

The conceptual model is the following:

Table `airport`

- `id_aerodromo` (Primary Key)
- `sg_icao`
- `sg_iata`
- `nm_aerodromo`
- `nm_municipio`

- nm_pais
- nm_continente

Table `line`

- id_tipo_linha (Primary Key)
- cd_tipo_linha
- ds_tipo_linha
- ds_natureza_tipo_linha
- ds_servico_tipo_linha

Table `equipment`

- id_equipamento (Primary Key)
- sg_equipamento_icao
- ds_modelo

Table `company`

- id_empresa (Primary Key)
- sg_empresa_icao
- sg_empresa_iata
- nm_empresa
- ds_tipo_empresa

Table `flights`

- id_basica (Primary Key)
- id_aerodromo_origem (References `airport.id_aerodromo`)
- id_aerodromo_destino (References `airport.id_aerodromo`)
- id_empresa (References `company.id_empresa`)
- ds_natureza_etapa (References `line.ds_natureza_tipo_linha`)
- id_equipamento (References `equipment.id_equipamento`)
- hr_partida_real
- dt_partida_real
- nr_semana_partida_real
- hr_chegada_real
- dt_chegada_real
- nr_semana_chegada_real

- ds_matricula
- lt_combustivel
- nr_assentos_ofertados
- kg_payload
- km_distancia
- nr_horas_voadas
- nr_velocidade_media

2.4 Data Dictionary

Table `airport`

Attribute	Data Type	Description
id_aerodromo	Integer	Unique identifier for an airport.
sg_icao	Varchar(4)	ICAO code for the airport.
sg_iata	Varchar(3)	IATA code for the airport.
nm_aerodromo	Varchar(50)	Name of the airport.
nm_municipio	Varchar(50)	Municipality where the airport is located.
nm_pais	Varchar(50)	Country where the airport is situated.
nm_continente	Varchar(30)	Continent where the airport is situated.

Table `line`

Attribute	Data Type	Description
id_tipo_linha	Integer	Unique identifier for a line type.
cd_tipo_linha	Char	Code representing the type of line.
ds_tipo_linha	Varchar(30)	Description of the line type.
ds_natureza_tipo_linha	Varchar(15)	Nature of the line type.
ds_servico_tipo_linha	Varchar(15)	Description of the service referring to the type of line (Passageiro/Cargueiro).

Table `equipment`

Attribute	Data Type	Description
id_equipamento	Integer	Code identifying the aircraft model.
sg_equipamento_icao	Varchar(4)	ICAO designator of the aircraft model (“Type Designator”)
ds_modelo	Varchar(50)	Description of the aircraft model.

Table `company`

Attribute	Data Type	Description
id_empresa	Integer	Unique identifier for a company.
sg_empresa_icao	Varchar(3)	ICAO acronym of the airline. Refers to the designator of the air transport company obtained from the ICAO (International Civil Aviation Organization).
sg_empresa_iata	Varchar(2)	IATA acronym for the airline. Refers to the air transport company designator obtained from IATA (International Air Transport Association).
nm_empresa	Varchar(100)	Airline name.
ds_tipo_empresa	Text	Description of the type of company. Refers to the description of the type of company in relation to the service performed.

Table `flights`

Key	Data Type	Description
id_basica	Integer	Unique identifier for a flight.
id_aerodromo_origem	Integer	foreign key
id_aerodromo_destino	Integer	foreign key
id_empresa	Integer	foreign key
id_tipo_linha	Integer	foreign key
id_equipamento	Integer	foreign key
hr_partida_real	TIME(0)	Actual departure time.
dt_partida_real	Datetime	Actual departure date.
nr_semana_partida_real	Integer	Week number of actual departure.
hr_chegada_real	TIME(0)	Actual arrival time.
dt_chegada_real	Datetime	Actual arrival date.
nr_semana_chegada_real	Integer	Week number of actual arrival.
ds_matricula	Varchar(3)	Aircraft registration code.
lt_combustivel	Integer	Fuel capacity in liters.
nr_assentos_ofertados	Integer	Number of offered seats.
kg_payload	Integer	Payload weight in kilograms.
km_distancia	Integer	Flight distance in kilometers.
nr_horas_voadas	Real	flying hours (for a single flight).
nr_velocidade_media	Real	Average flight speed.

2.5 Analytical Questions

In this section I provide three analytical questions that I find interesting in the context of flights data.

2.5.1 Query 1 - Airplane stats

For each plane model, print some informations as it's avg speed, avg liters consumed etc.

```
SELECT
  e.ds_modelo AS plane_model,
  COUNT(*) AS flights_count,
  AVG(f.nr_velocidade_media) AS avg_speed,
  AVG(f.lt_combustivel) as avg_liters,
  AVG(f.nr_horas_voadas) as avg_hours_per_flight,
  AVG(f.km_distancia)as avg_distance_in_km,
  AVG(f.kg_payload) as avg_payload_in_kg
FROM
  bd2.flights f
JOIN
  bd2.equipment e ON f.id_equipamento = e.id_equipamento
WHERE f.lt_combustivel != 0 and
      f.nr_horas_voadas != 0 and
      f.kg_payload != 0 and
      f.nr_velocidade_media != 0 and
      f.km_distancia != 0
GROUP BY
  e.ds_modelo
ORDER BY
  flights_count DESC
```

plane_model	flights_count	avg_speed	avg_liters	avg_hours_per_flight	avg_distance_in_km	avg_payload_in_kg
AIRBUS A320-100/200	90.084	561,491697	5.540,8644	2,0339746068	1.220,0161	17.251,4186
BOEING 737-800	86.355	544,216963	5.681,546	2,0630639117	1.210,5085	19.800
EMBRAER 195/ERJ-190-200	43.362	468,956514	3.167,1831	1,3261827007	649,5523	11.047,2702
AEROSPATIALE/ALENIA ATR 72 201/202	36.999	294,171877	1.264,1717	1,3666500288	417,8461	6.691,5954
AIRBUS A319	25.125	464,351876	3.701,9059	1,3999867835	695,8428	16.399,8622
BOEING 737-700	22.659	500,879962	4.080,4701	1,6938170158	912,0162	15.947
AIRBUS A321-100/200	17.567	595,518005	7.993,2556	2,3254919468	1.458,5914	26.881,2525
CESSNA 208 CARAVAN	9.839	230,705179	276,1319	1,5356083117	367,7004	647,0647
BOEING 767-300	8.119	670,879665	23.037,8916	4,4885123872	3.215,4756	46.701,2369
EMBRAER E195-E2/ERJ-190-400	7.466	479,375885	2.052,1547	1,404786142	697,9502	13.892,9293
EMBRAER 190/ERJ-190-100	3.123	481,289686	3.353,8236	1,4270786593	708,9705	10.622,0199
AIRBUS A321NEO	2.071	592,598812	5.434,6045	2,1737647842	1.382,9522	22.941,071
AIRBUS A330-200	1.980	731,356672	39.220,8798	5,8818264949	4.403,1263	45.137,8111
BOEING 777-300ER PAX	1.748	772,388221	72.020,5824	8,2364418089	6.582,1276	63.805,6104
BOEING 737-400	1.623	625,354572	8.221,87	3,1004107431	2.012,7184	18.278,5644
BOEING 727-200	1.565	431,750562	5.517,2856	1,1681682524	526,0128	22.402,278
AIRBUS A350-900	1.168	801,594358	70.581,6858	9,4846600856	7.694,4872	52.309
AEROSPATIALE/ALENIA ATR 42-300 / 320	1.032	361,502713	1.014,7878	1,4495962994	534,6366	4.536,4312
AEROSPATIALE/ALENIA ATR 42-500	894	358,218523	1.031,7506	1,3762676454	502,3949	4.697,3848
AIRBUS A330-900NEO	793	761,114943	22.168,5839	7,6357294451	5.897,0757	44.548,6154
BOEING 737-300	535	613,453159	8.581,8748	2,8896883607	1.843,9215	17.037,9421

2.5.2 Query 2 - Airline stats

For each company, display their total distance traveled and total hours of flight

```
SELECT
  c.nm_empresa,
  SUM(f.km_distancia) AS total_distance,
  SUM(f.nr_horas_voadas) as total_hours
FROM bd2.company c
JOIN bd2.flights f ON c.id_empresa = f.id_empresa
```

```
GROUP BY c.nm_empresa
ORDER BY total_distance DESC
```

nm_empresa	total_distance	total_hours
TAM LINHAS AÉREAS S.A.	139.666.212	226.523,4187290089
GOL LINHAS AÉREAS S.A. (EX- VRG LINHAS AÉREAS S.A.)	125.625.460	217.564,8337770038
AZUL LINHAS AÉREAS BRASILEIRAS S/A	116.701.938	219.485,6532709143
ABSA - AEROLINHAS BRASILEIRAS S.A.	10.144.439	14.282,232209
PASSAREDO TRANSPORTES AÉREOS S.A.	3.665.057	11.264,250263
TWO TÁXI AÉREO LTDA.	3.426.548	14.532,916866
MODERN TRANSPORTE AÉREO DE CARGA S.A.	1.843.708	2.950,516526
MAP TRANSPORTES AÉREOS LTDA.	1.155.101	3.338,533464
TOTAL LINHAS AÉREAS S.A.	964.627	2.232,299926
ASTA LINHAS AÉREAS LTDA (EX - AMÉRICA DO SUL LINHAS AÉREAS LTDA.)	210.066	844,633297
CONNECT LINHAS AÉREAS S.A. (ANTIGA CONNECT TÁXI AÉREO LTDA.)	183.634	277,799963
OMNI TÁXI AÉREO S.A.	168.344	451,365951
SIDERAL LINHAS AÉREAS LTDA.	4.174	7,583333

2.5.3 Query 3 - Flight Routes

Find all the routes in the database (hence, for year 2020). Also, count how many times they have been done.

```
SELECT
  a_origem.nm_aerodromo AS origin_airport,
  a_origem.sg_iata as IATA_origin,
  a_destino.nm_aerodromo AS destination_airport,
  a_destino.sg_iata as IATA_destination,
  COUNT(*) AS route_count
FROM
  bd2.flights f
JOIN
  bd2.airport a_origem ON f.id_aerodromo_origem = a_origem.id_aerodromo
JOIN
  bd2.airport a_destino ON f.id_aerodromo_destino = a_destino.id_aerodromo
GROUP BY
  f.id_aerodromo_origem,
  a_origem.nm_aerodromo,
  f.id_aerodromo_destino,
  a_destino.nm_aerodromo
ORDER BY
  route_count DESC
```

origin_airport	IATA_origin	destination_airport	IATA_destination	route_count
CONGONHAS	CGH	SANTOS DUMONT	SDU	6.592
SANTOS DUMONT	SDU	CONGONHAS	CGH	6.569
GUARULHOS - GOVERNADOR ANDRÉ FRANCO MONTORO	GRU	SALGADO FILHO	POA	3.907
SALGADO FILHO	POA	GUARULHOS - GOVERNADOR ANDRÉ FRANCO MONTORO	GRU	3.880
GUARULHOS - GOVERNADOR ANDRÉ FRANCO MONTORO	GRU	GUARARAPES - GILBERTO FREYRE	REC	3.129
GUARARAPES - GILBERTO FREYRE	REC	GUARULHOS - GOVERNADOR ANDRÉ FRANCO MONTORO	GRU	3.049
DEPUTADO LUÍS EDUARDO MAGALHÃES	SSA	GUARULHOS - GOVERNADOR ANDRÉ FRANCO MONTORO	GRU	2.932
GUARULHOS - GOVERNADOR ANDRÉ FRANCO MONTORO	GRU	DEPUTADO LUÍS EDUARDO MAGALHÃES	SSA	2.929
SANTOS DUMONT	SDU	PRESIDENTE JUSCELINO KUBITSCHEK	BSB	2.703
TANCREDO NEVES	CNF	CONGONHAS	CGH	2.696
PRESIDENTE JUSCELINO KUBITSCHEK	BSB	SANTOS DUMONT	SDU	2.680
TANCREDO NEVES	CNF	GUARULHOS - GOVERNADOR ANDRÉ FRANCO MONTORO	GRU	2.677
CONGONHAS	CGH	TANCREDO NEVES	CNF	2.677
GUARULHOS - GOVERNADOR ANDRÉ FRANCO MONTORO	GRU	TANCREDO NEVES	CNF	2.666
PRESIDENTE JUSCELINO KUBITSCHEK	BSB	GUARULHOS - GOVERNADOR ANDRÉ FRANCO MONTORO	GRU	2.615
GUARULHOS - GOVERNADOR ANDRÉ FRANCO MONTORO	GRU	PRESIDENTE JUSCELINO KUBITSCHEK	BSB	2.606

BRIG. EDUARDO GOMES	ZZ9	JUÍNA	JIA	1
JOÃO DURVAL CARNEIRO	N/I	VIRACOPOS	VCP	1
GLAUBER DE ANDRADE ROCHA	VDC	BAHIA - JORGE AMADO	IOS	1
GLAUBER DE ANDRADE ROCHA	VDC	AEROPORTO ESTADUAL DE JUNDIAÍ	QDV	1

3 Conclusions

This project consisted in analyzing an existing dataset and perform some queries, in order to respond at some analytical questions.
The dataset was quite large, so I had to perform some filtering in order to reduce the initial dimensions. SQL is an easy way to extract interesting insights from this dataset, and I didn't find too many difficulties in doing so. Perhaps, given the nature of the data, it might be more interesting to perform the queries and visualize some informations (such the routes) with a graph based technology.

Bibliography

[1] Agência Nacional de Aviação Civil (ANAC). https://www.anac.gov.br/. Accessed: March 2024.