



Academia SQL + Teradata

Antes de realizarmos transformações nos dados, é importante estabelecermos processos que apontem erros de qualidade nos dados que estamos trabalhando, dessa forma é possível ter clareza das inconsistências comuns e assim criar formas de melhorar a qualidade dos dados. Pensando nisso, a primeira atividade planejada é criarmos em cada tabela colunas adicionais reportando o tipo de inconsistência que encontrarmos.

Obtendo os dados

Dados previamente carregados no database “marcos.nagato@capgemini.com” no ambiente Lab da Teradata com os nomes de tabela *airports*, *planes* e *flights*.

Criar as estruturas no database do seu próprio usuário, e copiar os dados das 3 tabelas. As permissões apropriadas foram previamente atribuídas.

Airports Dataset



Dicionário

faa (string): Identificador do aeroporto determinado pela Federal Aviation Administration. Formato: 3-5 caracteres alfanuméricos.

name (string): Nome do aeroporto.

lat (float): Latitude do aeroporto. Intervalo de valores $[-180, 180]$.

lon (float): Longitude do aeroporto. Intervalo de valores $[-180, 180]$.

alt (int): Altitude do aeroporto. Unidade de medida em pés. Intervalo de valores $[-100, +\infty]$.

tz (float): Fuso horário baseado no deslocamento de horas a partir de UTC/GMT. Intervalo de valores $[-11, +14]$. Pode ser fuso fracionário.

dst (category): Horário de verão. Descrição dos possíveis valores:

E (Europe)

A (US/Canada)

S (South America)

O (Australia)

Z (New Zealand)

N (None)

U (Unknown)

Tarefas

1. Crie a coluna `qa_faa` e aponte inconsistências da coluna `faa` de acordo com as regras abaixo:
M : Indica que está com dado faltante.
F : Indica que não respeita o formato de 3-5 caracteres alfanuméricos.
2. Crie a coluna `qa_name` e aponte inconsistências da coluna `name` de acordo com as regras abaixo:
M : Indica que está com dado faltante.
3. Crie a coluna `qa_lat` e aponte inconsistências da coluna `lat` de acordo com as regras abaixo:
M : Indica que está com dado faltante.
I : Indica que o valor excede o intervalo $[-180, 180]$.

4. Crie a coluna `qa_lon` e aponte inconsistências da coluna `lon` de acordo com as regras abaixo:
 - M : Indica que está com dado faltante.
 - I : Indica que o valor excede o intervalo $[-180, 180]$.
5. Crie a coluna `qa_alt` e aponte inconsistências da coluna `alt` de acordo com as regras abaixo:
 - M : Indica que está com dado faltante.
 - I : Indica que o valor excede o intervalo $[-100, +\infty]$.
6. Crie a coluna `qa_tz` e aponte inconsistências da coluna `tz` de acordo com as regras abaixo:
 - M : Indica que está com dado faltante.
 - I : Indica que o valor excede o intervalo $[-11, +14]$.
7. Crie a coluna `qa_dst` e aponte inconsistências da coluna `dst` de acordo com as regras abaixo:
 - M : Indica que está com dado faltante.
 - C : Indica que o valor não pertence a nenhuma das categorias esperadas: E, A, S, O, Z, N, U.
 - N : Indica que o valor é numérico.

Planes Dataset



Dicionário

tailnum (string): Identificação do avião. Formato "**N-Number**", é composto por 5-6 caracteres.

Primeira letra é sempre "N".

De 1 a 4 dígitos seguidos por 1 caractere (ex. N1234Z).

De 1 a 3 dígitos seguidos por 2 caracteres (ex. N123AZ).

Não deve conter 0 (zero) como primeiro dígito, e não deve conter as letras "I" ou "O".

ano (int): Ano de fabricação do avião. Intervalo de valores [1950,+∞].

tipo (string): Tipo do avião.

manufacturer (string): Nome do fabricante.

model (string): Modelo do avião

engines (int): Número de motores. Intervalo de valores [1, 4].

seats (int): Número de assentos. Intervalo de valores [2, 500].

speed (int): Velocidade média de cruzeiro. Unidade de medida em milhas. Intervalo de valores [50, 150].

engine (category): Tipo de motor.

Tarefas

1. Crie a coluna `qa_tailnum` e aponte inconsistências da coluna `tailnum` de acordo com as regras abaixo:
 - M : Indica que está com dado faltante.
 - S : Indica que não tem exatamente 5 ou 6 caracteres.
 - FN : Indica que não inicia com a letra "N".
 - FE : Indica que contém caracteres inválidos ("I", "O", ou 0 como primeiro dígito).
 - FD : Indica caracteres não numéricos em posições onde tem que haver dígito.
2. Crie a coluna `qa_ano` e aponte inconsistências da coluna `ano` de acordo com as regras abaixo:
 - M : Indica que não possui informação de ano.
 - I : Indica que o valor excede o intervalo [1950,+∞).
3. Crie a coluna `qa_tipo` e aponte inconsistências da coluna `tipo` de acordo com as regras abaixo:
 - M : Indica que está com dado faltante.
 - C : Indica que o valor não pertence a nenhuma categoria esperada:
 - Fixed wing multi engine
 - Fixed wing single engine
 - Rotorcraft
4. Crie a coluna `qa_manufacturer` e aponte inconsistências da coluna `manufacturer` de acordo com as regras abaixo:
 - M : Indica que está com dado faltante.
 - C : Indica que o valor não pertence a nenhuma categoria esperada:

- AIRBUS
 - BOEING
 - BOMBARDIER
 - CESSNA
 - EMBRAER
 - SIKORSKY
 - CANADAIR
 - PIPER
 - MCDONNELL DOUGLAS
 - CIRRUS
 - BELL
 - KILDALL GARY
 - LAMBERT RICHARD
 - BARKER JACK
 - ROBINSON HELICOPTER
 - GULFSTREAM
 - MARZ BARRY
5. Crie a coluna `qa_model` e aponte inconsistências da coluna `model` de acordo com as regras abaixo:
 - M : Indica que está com dado faltante.
 - F : Indica que não respeita o formato esperado
 - Modelos AIRBUS devem começar com "A"
 - Modelos BOEING devem começar com "7"
 - Modelos BOMBARDIER e CANADAIR devem começar com "CL"
 - Modelos MCDONNELL DOUGLAS devem começar com "MD" ou "DC"
 6. Crie a coluna `qa_engines` e aponte inconsistências da coluna `engines` de acordo com as regras abaixo:
 - M : Indica que está com dado faltante.
 - I : Indica que o valor excede o intervalo [1, 4].
 7. Crie a coluna `qa_seats` e aponte inconsistências da coluna `seats` de acordo com as regras abaixo:
 - M : Indica que está com dado faltante.
 - I : Indica que o valor excede o intervalo [2, 500].
 8. Crie a coluna `qa_speed` e aponte inconsistências da coluna `speed` de acordo com as regras abaixo:
 - M : Indica que está sem informação de velocidade.
 - I : Indica que o valor excede o intervalo [50, 150].
 9. Crie a coluna `qa_engine` e aponte inconsistências da coluna `engine` de acordo com as regras abaixo:
 - M : Indica que está com dado faltante.
 - C : Indica que o valor não pertence a nenhuma categoria esperada:
 - Turbo-fan
 - Turbo-jet
 - Turbo-prop
 - Turbo-shaft
 - 4 Cycle

Flights Dataset



Dicionário

ano (int), **mes**(int), **dia** (int): Ano, Mês, Dia de partida.

dep_time (string): Horário real de partida do voo no horário local. Formato: HHMM ou HMM.

dep_delay (int): Atraso de partida do voo em minutos. Valores negativos representam partidas antecipadas.

arr_time (string): Horário real de chegada do voo no horário local. Formato: HHMM ou HMM.

arr_delay (int): Atraso de chegada do voo em minutos. Valores negativos representam chegadas antecipadas.

carrier (string): Identificador da empresa aérea.

tailnum (string): Identificador do avião. Veja dataset planes.

flight (string): Identificador do voo. Formato: 4 dígitos (preenchidos com zero a esquerda caso necessário).

origin (string) e **dest** (string): Identificadores faa dos aeroportos de origem e destino.

air_time (int): Tempo de voo. Unidade de medida em minutos. Intervalo de dados [20, 500].

distance (int): Distancia entre aeroportos. Unidade de medida em milhas. Intervalo de valores [50, 3000].

hora (int), **minuto** (int): Hora e Minuto agendada para partida.

Tarefas

Considere o dataset `flights.csv` para realizar as seguintes tarefas:

1. Crie a coluna `qa_year_month_day` e aponte inconsistências das colunas `ano`, `mes`, `dia` de acordo com as regras abaixo:

MY : Indica que está com dado faltante no ano.

MM : Indica que está com dado faltante no mes.

MD : Indica que está com dado faltante no dia.

IY : Indica que o valor excede o intervalo [1950,+∞) no ano.

IM : Indica que o valor excede o intervalo [1, 12] no mês.

- ID : Indica que o valor excede o intervalo esperado para o dia conforme o mês (para fevereiro usar o limite de 29 dias).
2. Crie a coluna `qa_hour_minute` e aponte inconsistências das colunas `hora` e `minuto` de acordo com as regras abaixo:
 - MH : Indica que está com dado faltante na hora.
 - MM : Indica que está com dado faltante no minuto.
 - IH : Indica que o valor excede o intervalo [0, 23] na hora.
 - IM : Indica que o valor excede o intervalo [0, 59] no minuto.
 3. Crie a coluna `qa_dep_arr_time` e aponte inconsistências da coluna `dep_time` e `arr_time` de acordo com as regras abaixo:
 - MD : Indica que está com dado faltante no `dep_time`.
 - MA : Indica que está com dado faltante no `arr_time`.
 - FD : Indica que não respeita o formato esperado no `dep_time` (3 ou 4 dígitos).
 - FA : Indica que não respeita o formato esperado no `arr_time` (3 ou 4 dígitos).
 - ID : Indica que o `dep_time` não é um horário válido (HHMM ou HMM).
 - IA : Indica que o `arr_time` não é um horário válido (HHMM ou HMM).
 4. Crie a coluna `qa_dep_arr_delay` e aponte inconsistências da coluna `dep_delay` e `arr_delay` de acordo com as regras abaixo:
 - MD : Indica que está com dado faltante no `dep_delay`.
 - MA : Indica que está com dado faltante no `arr_delay`.
 - FD : Indica que há caracteres no campo `dep_delay`.
 - FA : Indica que há caracteres no campo `arr_delay`.
 5. Crie a coluna `qa_carrier` e aponte inconsistências da coluna `carrier` de acordo com as regras abaixo:
 - M : Indica que está com dado faltante.
 - F : Indica que não respeita o formato esperado (2 caracteres alfanuméricos).
 6. Crie a coluna `qa_tailnum` e aponte inconsistências da coluna `tailnum` de acordo com as regras abaixo:
 - M : Indica que está com dado faltante.
 - R : Indica que a aeronave está ausente no dataset `planes`.
 7. Crie a coluna `qa_flight` e aponte inconsistências da coluna `flight` de acordo com as regras abaixo:
 - M : Indica que está com dado faltante.
 - F : Indica que o campo não está com formato numérico.
 8. Crie a coluna `qa_origin_dest` e aponte inconsistências da coluna `origin`, `dest` de acordo com as regras abaixo:
 - MO : Indica que está com dado faltante no `origin`.
 - MD : Indica que está com dado faltante no `dest`.
 - RO : Indica que o aeroporto origem está ausente no dataset `airports`.
 - RD : Indica que o aeroporto destino está ausente no dataset `airports`.
 9. Crie a coluna `qa_air_time` e aponte inconsistências da coluna `air_time` de acordo com as regras abaixo:
 - M : Indica que está com dado faltante.
 - I : Indica que o valor excede o intervalo [20, 500].
 - F : Indica que o valor não é numérico.
 10. Crie a coluna `qa_distance` e aponte inconsistências da coluna `distance` de acordo com as regras abaixo:
 - M : Indica que está com dado faltante.
 - I : Indica que o valor excede o intervalo [50, 3000].
 11. Crie a coluna `qa_distance_airtime` e aponte inconsistências entre as colunas `distance` e `air_time` de acordo com as regras abaixo:
 - M : Indica que está com `distance` ou `air_time` faltante.
 - TL : Indica que a viagem é longa de acordo com a condição: $\text{air_time} \geq \text{distance} \times 0.1 + 30$.

TS : Indica que a viagem é curta de acordo com a condição: $\text{air_time} \leq \text{distance} \times 0.1 + 10$.

TR : Indica que a viagem é normal caso as duas anteriores não sejam verdade.