INTRODUÇÃO


Apache Airflow
Apache Spark
PostgreSQL
redis
ORACLE DATABASE
Apache kafka
MySQL
Microsoft SQL Server
aws
MariaDB
SQLite

CAPÍTULO 1

DADOS E SUAS ORGANIZAÇÕES

COMO ORGANIZAR DADOS?

UMA LISTA TELEFÔNICA BEM RUIM

[Rodrigues, Silva, Soares, Sousa, Sousa]

[Felipe, Maria, João, João, Marina]

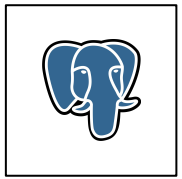
[0001234, 0001834, 0101534, 0011244, 0234644]

SOBRENOME	NOME	TELEFONE
Rodrigues	Felipe	0001234
Silva	Maria	0001834
Soares	João	0101534
Sousa	João	0011244
Sousa	Marina	0234644

MODELOS DE DADOS

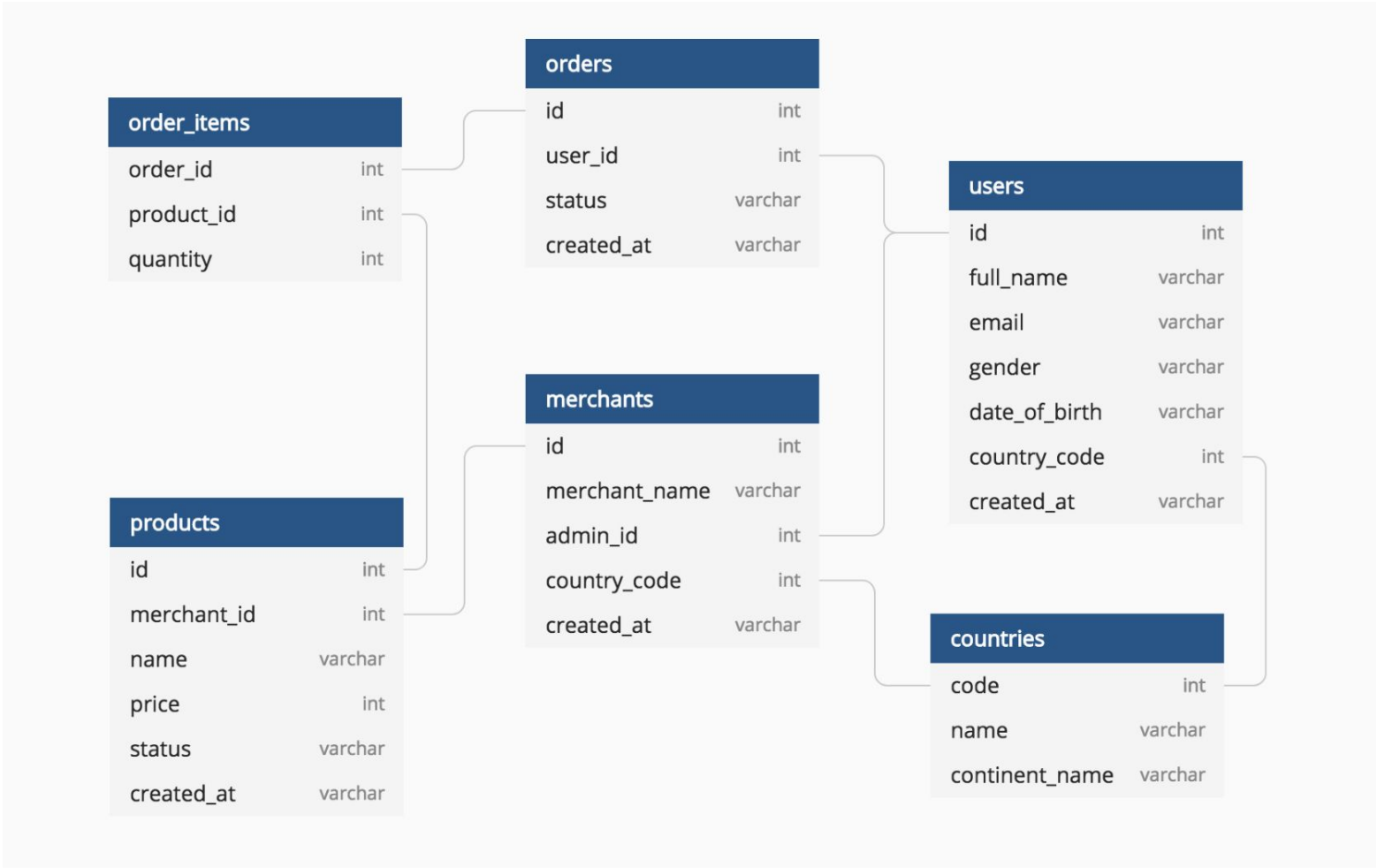
RELACIONAIS

+ Modelo mais conhecido de bancos de dados, baseado no modelo relacional de Edgar Codd em 1970.
+ Em geral aceitam a conhecida "linguagem de busca estruturada", ou SQL (Structured Query Language)
+ Representam dados em tabelas, com linhas e colunas.



NoSQL

+ "Not only SQL"
+ Modelos de armazenamento de dados que não são correlatos com o modelo relacional, e portanto não tem a noção de tabelas/linhas/colunas.
+ Flexibilidade no "schema" e localidade dos dados
+ Operações de busca especializadas por caso de uso!


order_items
order_id int
product_id int
quantity int
orders
id int
user_id int
status varchar
created_at varchar
users
id int
full_name varchar
email varchar
gender varchar
date_of_birth varchar
country_code int
created_at varchar
products
id int
merchant_id int
name varchar
price int
status varchar
created_at varchar
merchants
id int
merchant_name varchar
admin_id int
country_code int
created_at varchar
countries
code int
name varchar
continent_name varchar

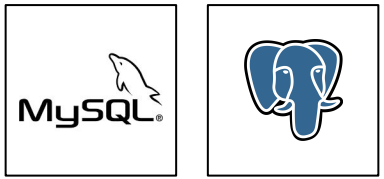
LINHAS (SQL) - COLUNAS (UM TIPO DE NOSQL)

Lineares (row oriented)

+ Dados são armazenados linha a linha.
+ Otimizados para escrita e leitura de linhas de dados.
+ Mantém todos os dados de uma mesma linha juntos na memória.

Colunares (column oriented)

+ Dados são armazenados coluna a coluna
+ São otimizados para leitura de colunas, bem como de computações feitas sobre ela.
+ Mantém todos os dados de uma mesma coluna juntos na memória.



NOSQL - COLUNAS

+ Essas duas plataformas são “queridinhas” em aplicações big data.
+ São soluções de “datawarehousing”, um conceito que veremos mais para frente

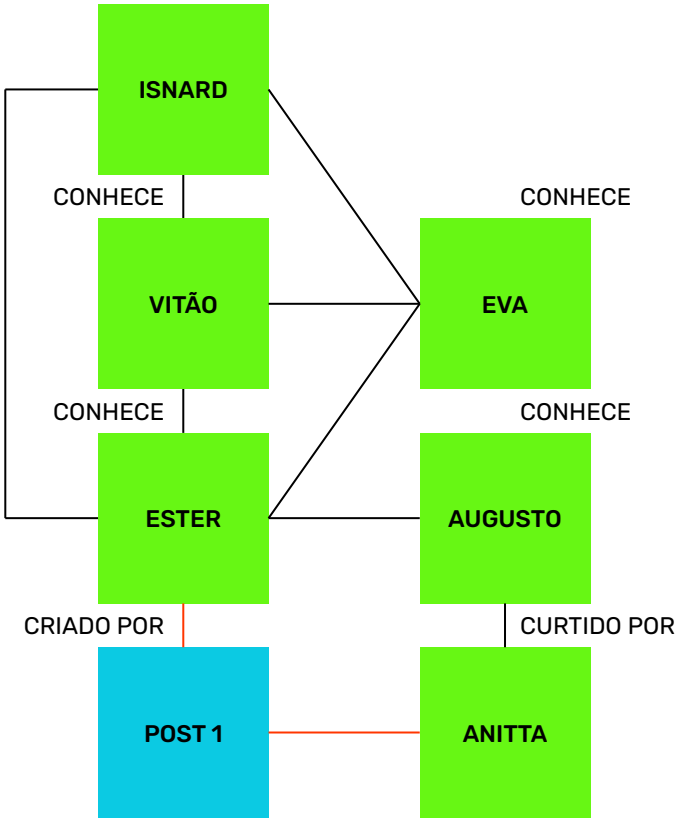
NOSQL - DOCUMENTOS

+ Dados são armazenados no formato de "documentos"
+ Analogia com pasta de arquivo
+ Documentos não precisam ter "schema" definido, sendo muito flexíveis.
+ Dificuldade em juntar informações de documentos diferentes.


```
{
  "user_id": 251,
  "first_name": "Bill",
  "last_name": "Gates",
  "positions": [
    {
     "job_title": "Co-chair",
     "organization": "Bill & Melinda Gates Found."
    },
    {
     "job_title": "Co-founder, Chairman",
     "organization": "Microsoft"
    }
  ],
  "contact_info": {
    "blog": "http://thegatesnotes.com",
    "twitter": "http://twitter.com/BillGates"
  }
}
```

NOSQL - GRAFOS

+ Dados são armazenados no formato de "grafos"
+ Usados para modelar relações entre entidade.
+ Podem conter nós heterogêneos (como posts e likes)
+ Usados em redes sociais, sistemas de recomendação, detecção de fraude



NOSQL - CHAVE-VALOR

+ Uma das muitas formas de guardar informações na Amazon
+ Alta escalabilidade e totalmente gerenciado
+ No modelo chave-valor, os dados são armazenados em pares guiados pelas chaves, permitindo buscas bem rápidas

DIFERENTES SERVIDORES
DIFERENTES ESTRUTURAS

OUTROS “STORAGES” FAMOSOS

CHAVE-VALOR EM MEMÓRIA

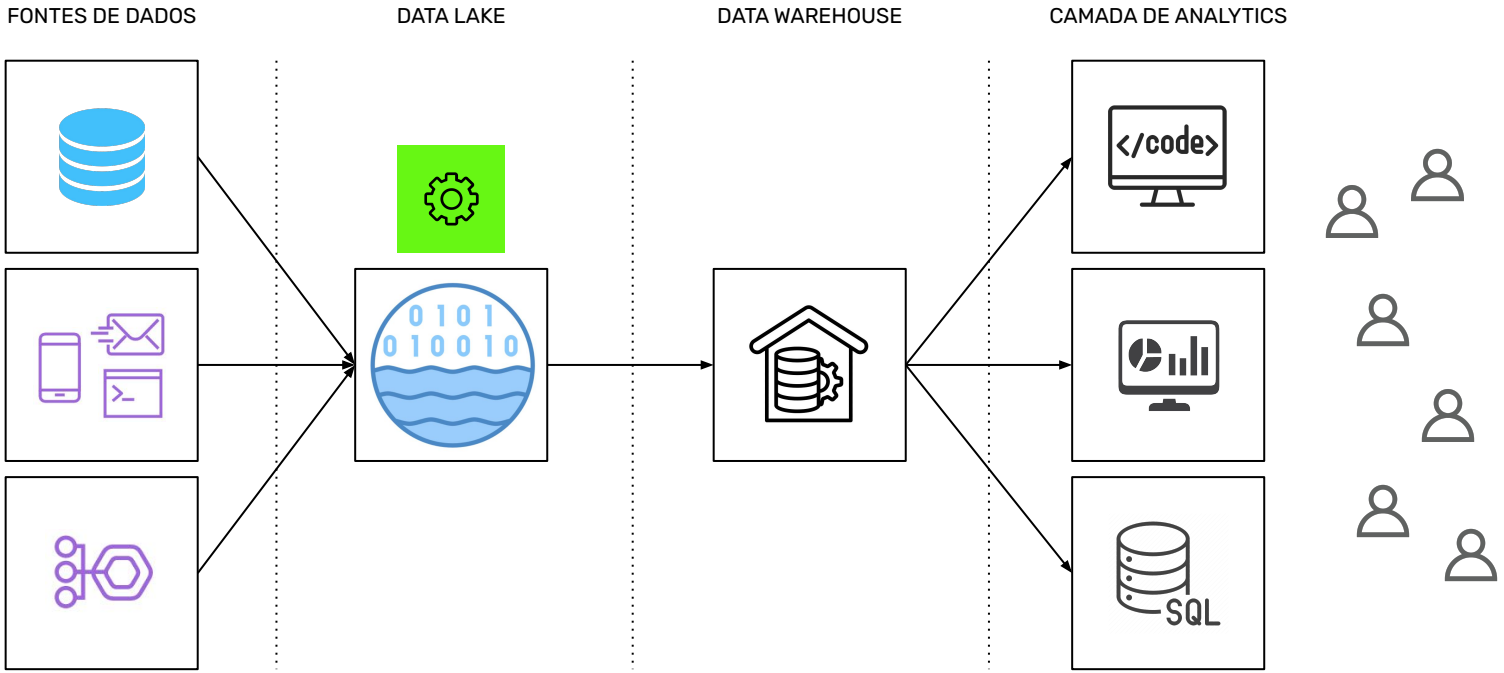
+ Normalmente mais usado pelos times de desenvolvimento do que pelos times de “data”
+ É um armazenamento de dados muito, muito rápido porém efêmero - seus dados são armazenados na memória RAM e não no disco
+ Seu principal caso de uso é fazer cache de informações temporárias, colocando o redis na frente de um banco de dados normal - e mais lento

OBJECT STORAGE

+ Até aqui nós observamos diversas formas de guardar dados textuais - informações de pessoas, relações entre elas, chaves, valores, etc. Mas e arquivos binários como fotos e vídeos?
+ Normalmente essas entidades ficam fora dos bancos comuns por diversas questões - principalmente custo e performance.
+ Arquivos binários são normalmente guardados em plataformas chamadas de Object Storage, como o S3 e o Cloud Storage

CAPÍTULO 2

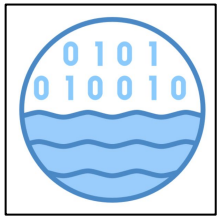
ARQUITETURA



DATA LAKE

Repositório central dos seus dados, onde reside toda a informação

+ Funciona como um HD “infinito”, com armazenamento muito barato (2 cents de dólar por mês por gb)
+ Dado cru, não processado/agregado/filtrado
+ Conjuntos grandes e variados de dados



DATA WAREHOUSE

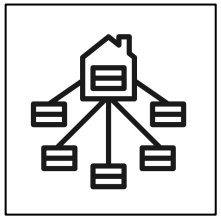
Repositório orientado à performance e de interface amigável

+ Dado estruturado e ordenado, pronto para ser consultado e manipulado
+ Mais caro que o Data Lake, recebendo (geralmente) os dados já processados e de menor volume
+ Otimizado para consulta: leituras rápidas, escritas lentas

DATA MART

Subconjunto de dados de um Data Warehouse

+ Subconjunto temático de um Data Warehouse
+ Escopo claro e definido, com dados de propósito específico
+ Geralmente usados para guardar dados de áreas específicas, diminuindo a complexidade do todo

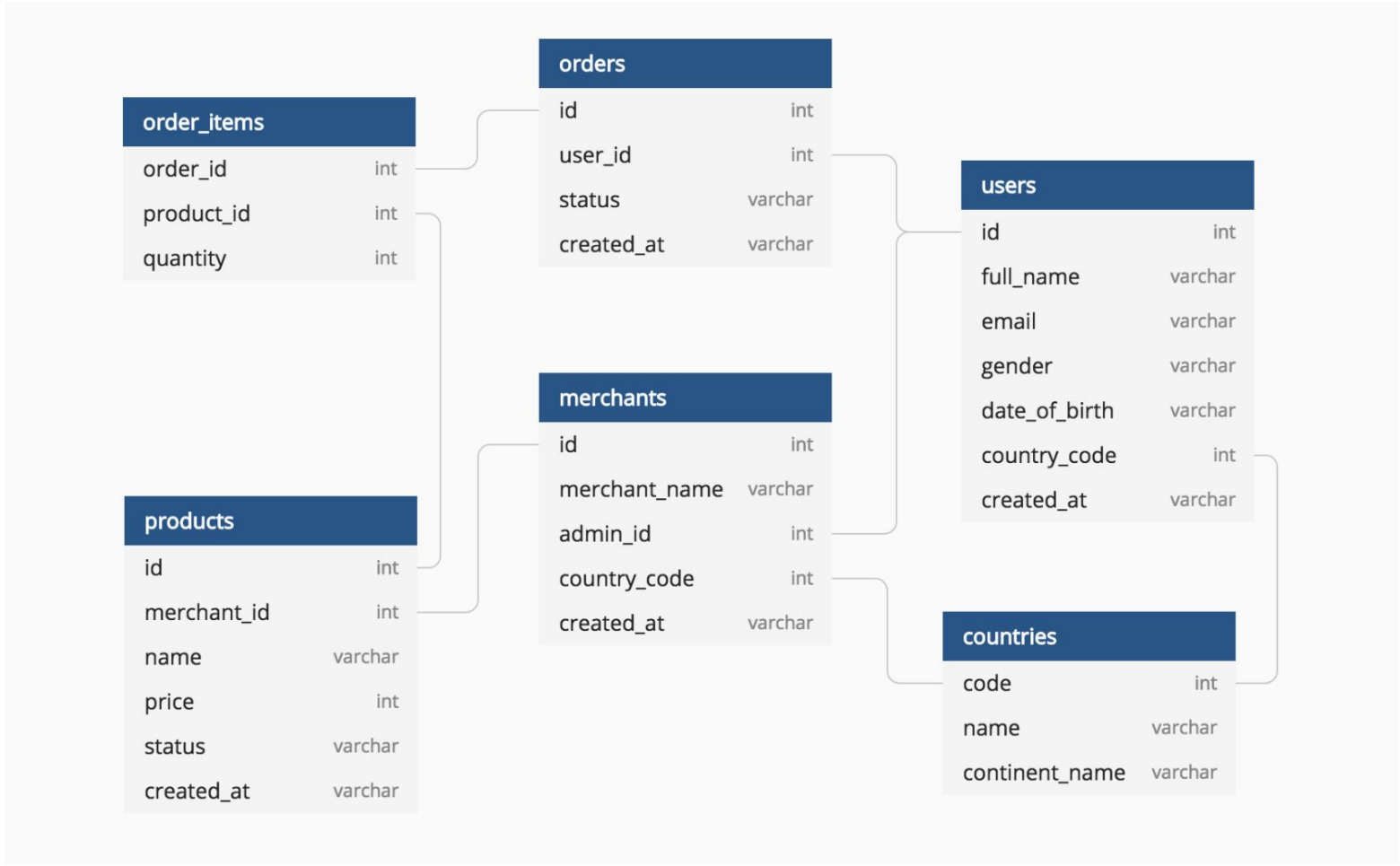


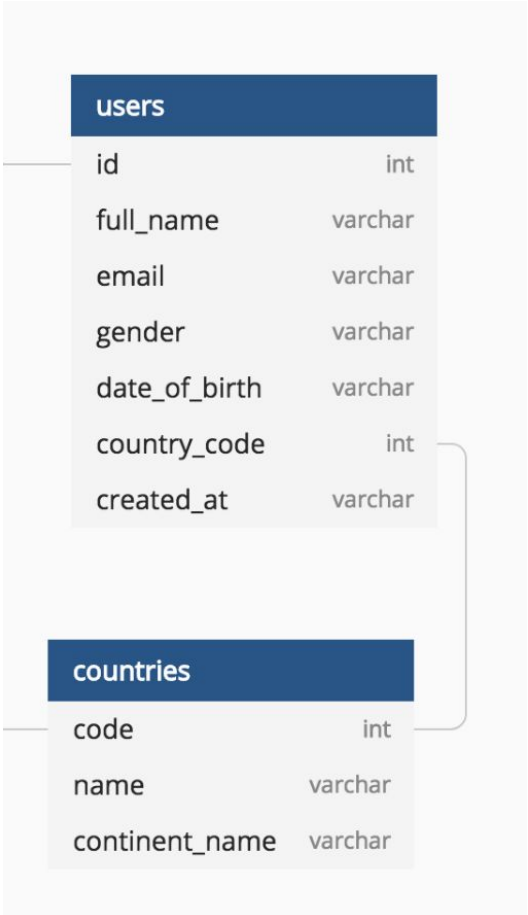
CAPÍTULO 3

MUITOS SERVIDORES POSSÍVEIS

MAS TODOS FALAM A “MESMA” LÍNGUA

	CustomerId	FirstName	LastName	Company	Address	City	State
	Filtro	Filtro	Filtro	Filtro	Filtro	Filtro	Filtro
1	1	Luís	Gonçalves	Embraer - Empresa ...	Av. Brigadeiro Faria Lima, ...	São José do...	SP
2	2	Leonie	Köhler	*NULL*	Theodor-Heuss-Straße 34	Stuttgart	*NULL*
3	3	François	Tremblay	*NULL*	1498 rue Bélanger	Montréal	QC
4	4	Bjørn	Hansen	*NULL*	Ullevålsveien 14	Oslo	*NULL*
5	5	František	Wichterlová	JetBrains s.r.o.	Klanova 9/506	Prague	*NULL*
6	6	Helena	Holý	*NULL*	Rilská 3174/6	Prague	*NULL*
7	7	Astrid	Gruber	*NULL*	Rotenturmstraße 4, 1010 ...	Vienne	*NULL*
8	8	Daan	Peeters	*NULL*	Grétrystraat 63	Brussels	*NULL*
9	9	Kara	Nielsen	*NULL*	Sønder Boulevard 51	Copenhagen	*NULL*
10	10	Eduardo	Martins	Woodstock Discos	Rua Dr. Falcão Filho, 155	São Paulo	SP
11	11	Alexandre	Rocha	Banco do Brasil S.A.	Av. Paulista, 2022	São Paulo	SP
12	12	Roberto	Almeida	Riotur	Praça Pio X, 119	Rio de Janeiro	RJ
13	13	Fernanda	Ramos	*NULL*	Qe 7 Bloco G	Brasília	DF
14	14	Mark	Philips	Telus	8210 111 ST NW	Edmonton	AB
15	15	Jennifer	Peterson	Rogers Canada	700 W Pender Street	Vancouver	BC
16	16	Frank	Harris	Google Inc.	1600 Amphitheatre Parkway	Mountain ...	CA


order_items
order_id int
product_id int
quantity int
orders
id int
user_id int
status varchar
created_at varchar
users
id int
full_name varchar
email varchar
gender varchar
date_of_birth varchar
country_code int
created_at varchar
products
id int
merchant_id int
name varchar
price int
status varchar
created_at varchar
merchants
id int
merchant_name varchar
admin_id int
country_code int
created_at varchar
countries
code int
name varchar
continent_name varchar


users
id int
full_name varchar
email varchar
gender varchar
date_of_birth varchar
country_code int
created_at varchar
countries
code int
name varchar
continent_name varchar

RELACIONAMENTOS & PERFORMANCE

USUÁRIOS

full_name	country
Felipe	BR
Aline	br
Matheus	Brasil
Flávia	brazil
Osvaldo	Br

BRASIL É...

BR

Br

br

Brasil

brazil

DESIGN “NORMAL”

full_name	country_id
Felipe	1
Aline	1
Matheus	1
Flávia	1
Osvaldo	1

id	country	continent
1	Brazil	South America

GARANTIAS DE INTEGRIDADE

UM BANCO RELACIONAL IMPEDE

+ Quebra de tipos de dados
+ Registros com campos obrigatórios faltantes
+ Registros com referências inválidas