

UNIVERSIDADE DE BRASÍLIA

Faculdade do Gama

Sistemas de Banco de Dados 2

Trabalho Final (TF)

Tema D - Data Warehouse

Ivan Diniz Dobbin - 17/0013278

Rogério S. dos Santos Júnior - 17/0021751

Brasília, DF



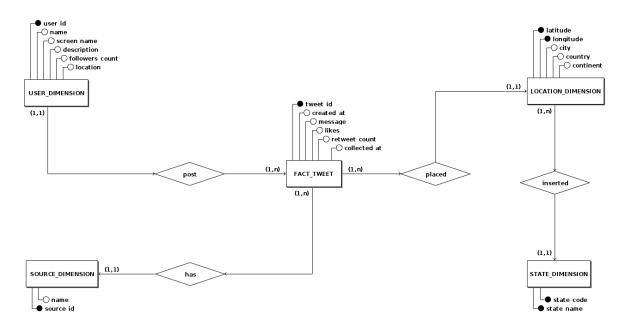
CONTEXTUALIZAÇÃO

Os dados utilizados foram extraídos do Kaggle (site dos dados). Essa base contém os dados dos tweets das eleições do Estados Unidos da América de 2020, mais especificamente de todas as pessoas que realizaram tweets com #JoeBiden e #Biden palavras-chave. Esses dados estavam em formato de arquivo CSV (hashtag_joebidden.csv), assim foi necessário realizar uma modelagem do banco MySQL (utilizando o star schema) de forma a normalizar os dados em suas respectivas tabelas. Em seguida, com o auxílio de um script Python, os dados foram inseridos na base, totalizando 155 mil tuplas. Foi gerado o dump da base modelada.

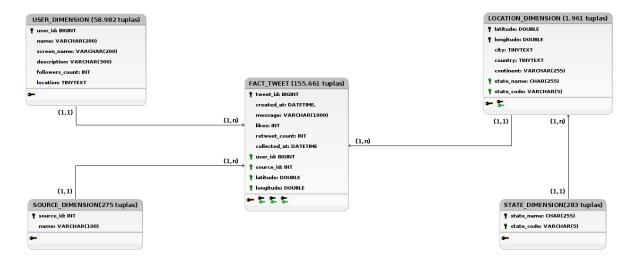
A base de dados escolhida foi a de tweets do Rogério, pois depois de considerações entre a dupla, chegou-se à conclusão que a transformação para o MySQL tinha sido melhor feita nesta base, na qual houve o tratamento de dados nulos e os dados tinham maior riqueza para construir consultas.

DOCUMENTAÇÃO

1) Diagrama Entidade Relacionamento (DE-R)



2) Diagrama Lógico



2) Dicionário de Dados





Entidade: USER_DIMENSION						
Descrição: Dados relacionados aos usuários do Twitter						
Atributo	Propriedade	Tipo de	Tamanho	Descrição		
, tiribato	do atributo	Dado	ramamo	Dodongao		
	chave primária			É o número de		
user_id	obrigatório	BIGINT	11	identificação do usuário		
	obligatorio		no twitter			
			200	Nome de usuário que		
name	obrigatório	VARCHAR	200	aparece na página de		
				perfil		
screen_name	obrigatório	VARCHAR	200	Nome de usuário único		
Soreen_name	obligatorio	VAROLIAR	200	precedido de @		
descritption	obrigatório	VARCHAR	300	Descrição do usuário		
				Quantidade de usuários		
followers_count	obrigatório	INT	4	que seguem este		
				usuário no twitter		
				Localização do usuário		
location	obrigatório	TINYTEXT	255	que inclui o país e pode		
				incluir o estado		

Entidade: STATE_DIMENSION								
Descri	Descrição: Dados relacionados a localização onde foram feitos os tweets.							
Atributo	Propriedade do Tipo de Dado Tamanho Descrição							
	atributo							
state_code	chave primária	CHAR	5	código do				
State_code	obrigatório	Of I/ (I C	0	estado				
state_name	chave primária	VARCHAR	255	nome do				
Stato_Harrie	obrigatório	VAROUAL	200	estado				





Entidade: LOCATION_DIMENSION						
Descrição: Dados relacionados a localização dos tweets						
Atributo	Propriedade	Tipo de	Tamanho	Descrição		
, teribato	do atributo	Dado	ramamo	Doconiquo		
	chave			Coordenada da latitude		
	primária	DOUBLE	8	baseada no perfil do usuário		
latitude	obrigatório			baseada no perm do asaano		
	chave			Coordenada da longitude		
longitude	primária	DOUBLE	8	baseada no perfil do usuário		
	obrigatório			baseada no perm do dsuano		
city	obrigatório	TINYTEXT	255	Nome de uma cidade		
country	obrigatório	TINYTEXT	255	Nome de um país		
continent	obrigatório	VARCHAR	200	Nome de um continente		
	chave			Código do estado onde foi		
state_code	estrangeira	CHAR	5	realizado o tweet		
	obrigatória					
	chave					
state_name	estrangeira	VARCHAR	255	Nome do estado onde foi		
State_Hairie	obrigatória	VANOLIAN		realizado o tweet		

Entidade: SOURCE_DIMENSION							
Descrição: Dados relacionados ao aparelho utilizado para realizar o tweet							
Atributo	Atributo Propriedade Tipo de Tamanho Descrição						
	do atributo	Dado					
	chave						
	primária	INT	4	Número de identificação do			
source_id	obrigatório	IIVI	4	aparelho utilizado.			
name	obrigatório	VARCHAR	100	Nome do aparelho utilizado.			





Entidade: FACT_TWEET

Descrição: Define dados relacionados aos tweets com #JoeBiden e #Biden como palavras-chave.

Atributo	Propriedade	Tipo de	Tamanho	Descrição
Allibulo	do atributo	Dado	ramamo	Descrição
tweet_id	chave			Número de
	primária	BIGINT	11	identificação
	obrigatória			do tweet
created_at				Data e hora de
	optativo	DATETIME		quando o tweet
			14	foi criado
				TOI CITAGO
	alandar (C. S	\/A DOLLA D		Mensagem
message	obrigatório	VARCHAR	1000	presente no
				tweet
				Número de
Likes	obrigatório	INT	4	likes do tweet
				Número de
		IN 17		compartilhame
retweet_count	obrigatório	INT	4	ntos de um
				tweet
				Número de
	chave			identificação
source_id	estrangeira	INT	4	do aparelho
	obrigatória			utilizado para
				realizar o tweet
	chave			É o número de
user_id	estrangeira	BIGINT	11	identificação
	obrigatória			do usuário



				criador do
				tweet
				Coordenada
				da latitude
	chave			baseada no
latitude	estrangeira	DOUBLE	8	perfil do
	obrigatória			usuário criador
				do tweet
				Coordenada
				da longitude
	chave			baseada no
longitude	estrangeira	DOUBLE	8	perfil do
	obrigatória			usuário criador
				do tweet
				Data e hora de
applead of	ontativo	DATETIME	14	quando os
collected_at	optativo	DATETIME	14	dados do tweet
				foram extraídos

	Visão: TWEETS_BY_STATE					
De	Descrição: Dados relacionados a quantidade de tweets por estado					
Atributo	Atributo Propriedade Tipo de Tamanho Descrição					
	do atributo	Dado				
	chave					
	primária	CHAR	5	Sigla do estado		
state	obrigatório	O1 11 ti t		orgia ao ootaao		





city	obrigatório	TINYTEXT	255	Nome da cidade.
quantity	obrigatório	INT	4	Quantidade de tweets do estado.

	Visão: USER_ANALYTICS						
Descrição: Dados relacionados aos usuários do Twitter com seus tweets							
Atributo	Propriedade do atributo	Tipo de Dado	Tamanho	Descrição			
id	chave primária obrigatório	BIGINT	11	É o número de identificação do usuário no twitter			
username	obrigatório	VARCHAR	200	Nome de usuário único precedido de @			
followers	obrigatório	INT	4	Quantidade de usuários que seguem este usuário no twitter			
qtd_tweets	obrigatório	INT	4	Quantidade de tweets do usuário			
likes_avg	obrigatório	DOUBLE	8	Quantidade de likes em média por tweet			
retweet_avg	obrigatório	DOUBLE	8	Quantidade de compartilhamentos em média por tweet			