Universidade de Brasília Faculdade UnB Gama

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

2020

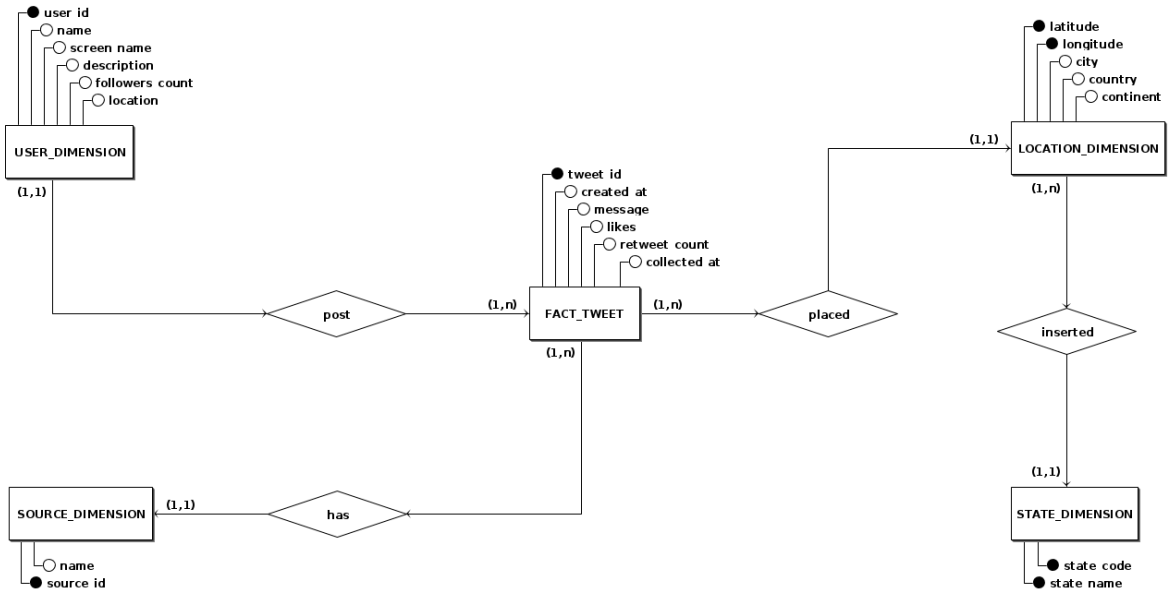
CONTEXTUALIZAÇÃO

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 como 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.

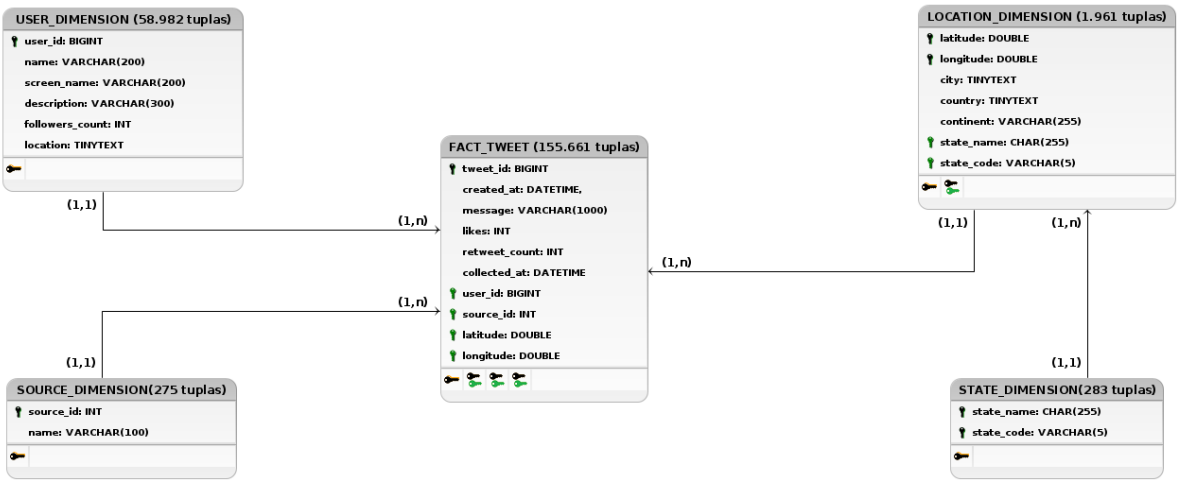
A base de dados escolhida foi a de tweets do Rogério, pois depois de considerações entre a dupla, chegou-se à conclusão que a transformação para o MySQL tinha sido melhor feita nesta base, na qual houve o tratamento de dados nulos e os dados tinham maior riqueza para construir consultas.

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 do atributo	Tipo de Dado	Tamanho	Descrição
user_id	chave primária obrigatório	BIGINT	11	É o número de identificação do usuário no twitter
name	obrigatório	VARCHAR	200	Nome de usuário que aparece na página de perfil
screen_name	obrigatório	VARCHAR	200	Nome de usuário único precedido de @
descritption	obrigatório	VARCHAR	300	Descrição do usuário
followers_count	obrigatório	INT	4	Quantidade de usuários que seguem este usuário no twitter
location	obrigatório	TINYTEXT	255	Localização do usuário que inclui o país e pode incluir o estado

Entidade: STATE_DIMENSION				
Descrição: Dados relacionados a localização onde foram feitos os tweets.				
Atributo	Propriedade do atributo	Tipo de Dado	Tamanho	Descrição
state_code	chave primária obrigatório	CHAR	5	código do estado
state_name	chave primária obrigatório	VARCHAR	255	nome do estado

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

Entidade: SOURCE_DIMENSION				
Descrição: Dados relacionados ao aparelho utilizado para realizar o tweet				
Atributo	Propriedade do atributo	Tipo de Dado	Tamanho	Descrição
source_id	chave primária obrigatório	INT	4	Número de identificação do 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 do atributo	Tipo de Dado	Tamanho	Descrição
tweet_id	chave primária obrigatória	BIGINT	11	Número de identificação do tweet
created_at	optativo	DATETIME	14	Data e hora de quando o tweet foi criado
message	obrigatório	VARCHAR	1000	Mensagem presente no tweet
Likes	obrigatório	INT	4	Número de likes do tweet
retweet_count	obrigatório	INT	4	Número de compartilhamentos de um tweet
source_id	chave estrangeira obrigatória	INT	4	Número de identificação do aparelho utilizado para realizar o tweet
user_id	chave estrangeira obrigatória	BIGINT	11	É o número de identificação do usuário

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

Visão: TWEETS_BY_STATE				
Descrição: Dados relacionados a quantidade de tweets por estado				
Atributo	Propriedade do atributo	Tipo de Dado	Tamanho	Descrição
state	chave primária obrigatório	CHAR	5	Sigla do estado

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