CA 3

Data Model

Für:

Big Data Engineering & Analysis (SS22)

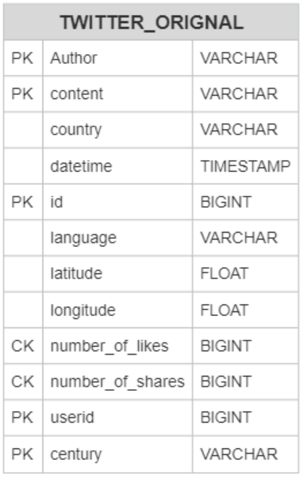
Prof. Dr. Oliver Hummel

Bei:

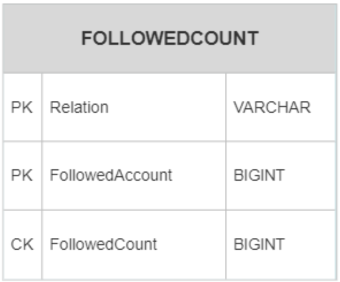
Milad Afshar Jahanshahi

2150426

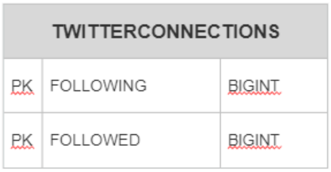
Data Model



TWITTER_ORIGNAL		
PK	Author	VARCHAR
PK	content	VARCHAR
	country	VARCHAR
	datetime	TIMESTAMP
PK	id	BIGINT
	language	VARCHAR
	latitude	FLOAT
	longitude	FLOAT
CK	number_of_likes	BIGINT
CK	number_of_shares	BIGINT
PK	userid	BIGINT
PK	century	VARCHAR



FOLLOWEDCOUNT		
PK	Relation	VARCHAR
PK	FollowedAccount	BIGINT
CK	FollowedCount	BIGINT



TWITTERCONNECTIONS		
PK	FOLLOWING	BIGINT
PK	FOLLOWED	BIGINT

The main reason for the above data model is :

In Cassandra, Instead of creating complex queries on a single table, it is preferred to create the schema according to the need. So as in our case, the schema is created by keeping the use case in mind of the queries.

Twitter_Orignal table is created which will hold the tweets and related data. One custom column is added named 'Century' holding the ceuntry of the tweet from the date in the datetime column. This is created so that we could make it part of key and then partition the data according to it as tweets from previous century are not required, so it will get all the data in the partition and by the help of it, the cluster keys will be fully functional which will sort the tweets the according to the popularity. The popularity is judged on the basis of the number of likes and shares, so the data is also sorted in the order of popularity, the reason to do this to select the top most popular tweets required in the queries.

Another thing which is added in this table is a secondary index on the content column in order to make the LIKE statement effective while searching the data.

Another Table TwitterConenctions is made which will hold all the data for the accounts like which account follows which account vice versa.

The third table FollowedCount table is created to hold the calculated data of the number of followers or following of the accounts. Either the data is of the follower or Followed account, is distinguished on the basis of relation column in this table. Another main reason to have this column is to partition data in order to make order by effectively on the selected partition.