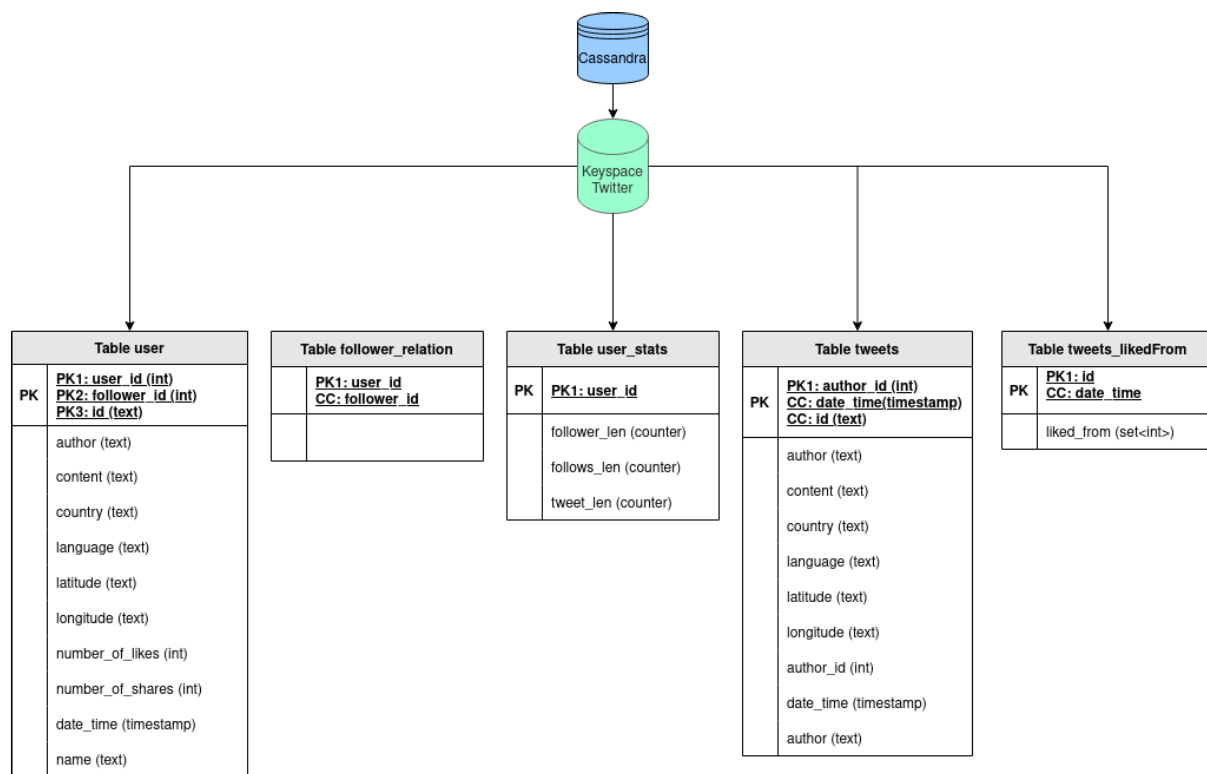


# Table Structure



- When first importing the data into Cassandra we ran into the following problem:  
Failed to import 1 rows: ParseError - Failed to parse 5.34896E+17 :  
invalid literal for int() with base 10: '5.34896E+17', given up  
without retries 'builtin\_function\_or\_method' object has no attribute  
'error' . In order to fix this we had to manipulate the data before the import  
manually.
- We had to change the data schema a lot of times and try out different  
combination to make the queries work. We end up to use the relation table  
with the relationship between user\_id, follower\_id and tweet\_id which build the  
primary key. So the data will be saved multiple times for each id and get around

45 times larger than the original data. Additionally we also use a stats table with counter to get a faster query for the length of follower or follows. Because of the "world-search"-query (task 6) we also add the tweets in a separate table to be able to run an index on the content col and filter with the LIKE-keyword.

- UDTs: We tried to load the tweet as UDTs to the above data schema. To perform this we updated the structure of the combined csv-file after [this-stackoverflow post](#) but always get an error for columns mismatch. We assume the content in the UDTs was not quoted so , in the content section caused these errors.
- The only materialized view we used is for the exercise 5 to get a fanout-like style, that updates after a new tweet is added.
- SASI-index are (like most of the things) experimental and are not recommended for production use. However we had to use it for the exercise 6 to enable the searching with the LIKE -statement. We assume that Cassandra makes an internal table with each word associated with the tweets row in which it appears, since the index takes around 5 minutes to create.