**Assignment 2**

**Basic Sentiment Analysis using Hive**

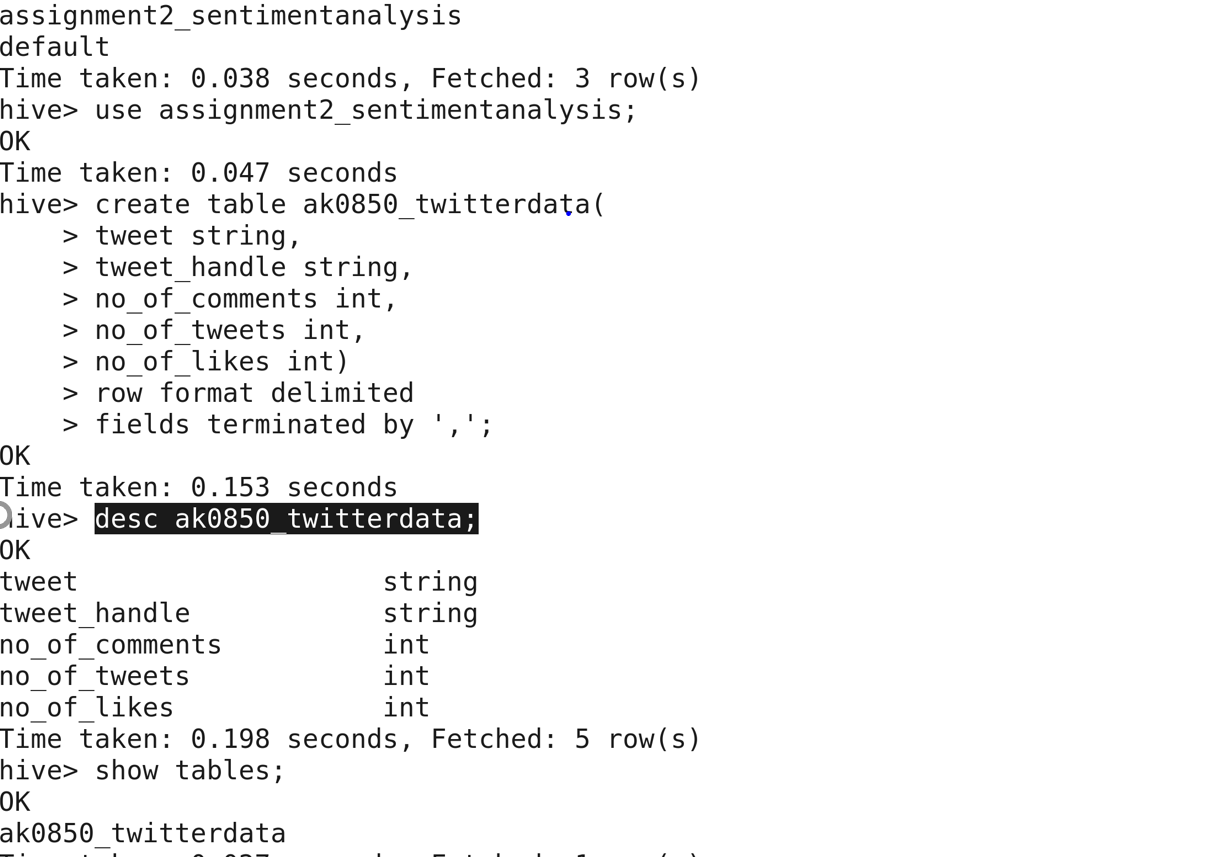
**Approach towards performing sentiment analysis:**

To perform sentiment analysis, twitter data has been collected which contains tweets, likes, tweet handler, comments and retweets in csv file. As the excel file is creating null values after importing the file to hive table, data has been collected in a “comma separated value” file format. While importing this data into hive table, if the table in hive is created with “fields terminated by ‘,’”, there is a chance of getting null values in the table if the tweet has any commas. These commas in tweets needs to be handled manually by replacing with any text. Once the data is successfully loaded into hive table, given “Dictionary.txt” is loaded into a hive table. Dictionary table contain tokens and scores. In order to compare the tweets with the tokens in the dictionary, each tweet is split into words. And these words are compared with the tokens in the dictionary to get the scores grouping by tweet handler. Sum of the all the scores for every user is calculated to determine whether they have positive or negative or neutral sentiment. In this assignment, these sentiments have been loaded into a new table with tweet handler, corresponding score and sentiment. If the sum of score is greater than 0, then the tweet is considered as positive sentiment and is sum of score is less than 0, it is considered as negative sentiment. Ans if the sum of score is equal to 0, then the tweet is considered as neutral.

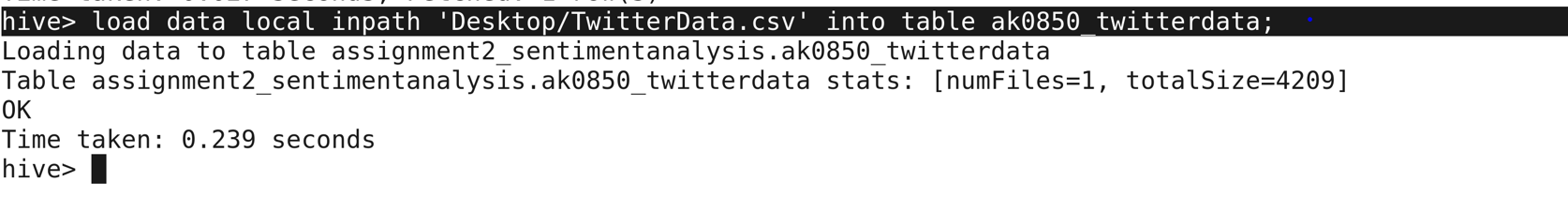
**Tables Creation and Insertion for Sentiment Analysis in Hive:**

New database assignment2\_sentimentanalysis has been created in hive. All the tables for sentiment analysis are created in this database.

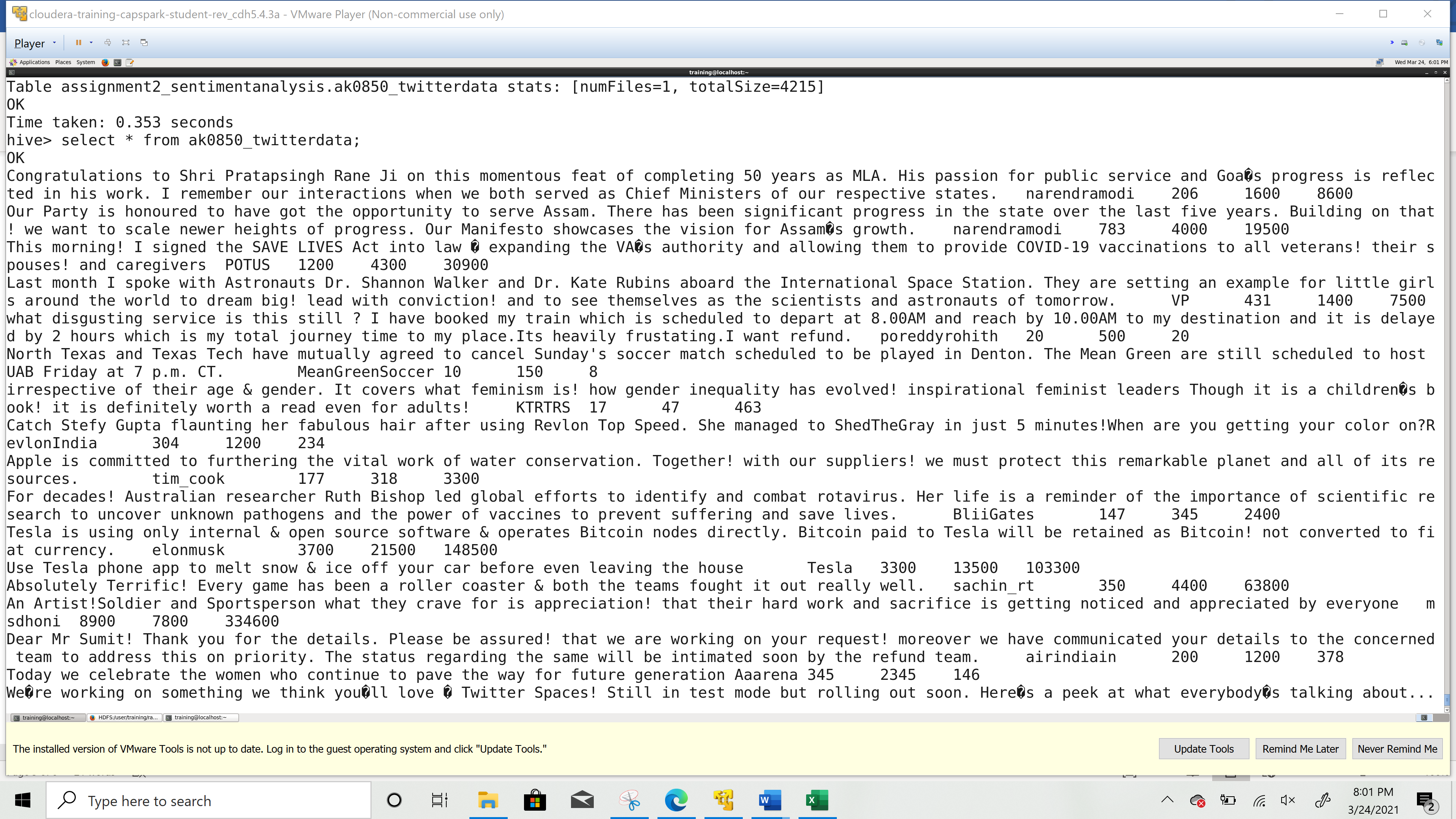
Creation and description of table **“ak0850\_twitterdata”** with fields terminated by as “,” as the data is loaded from csv file:



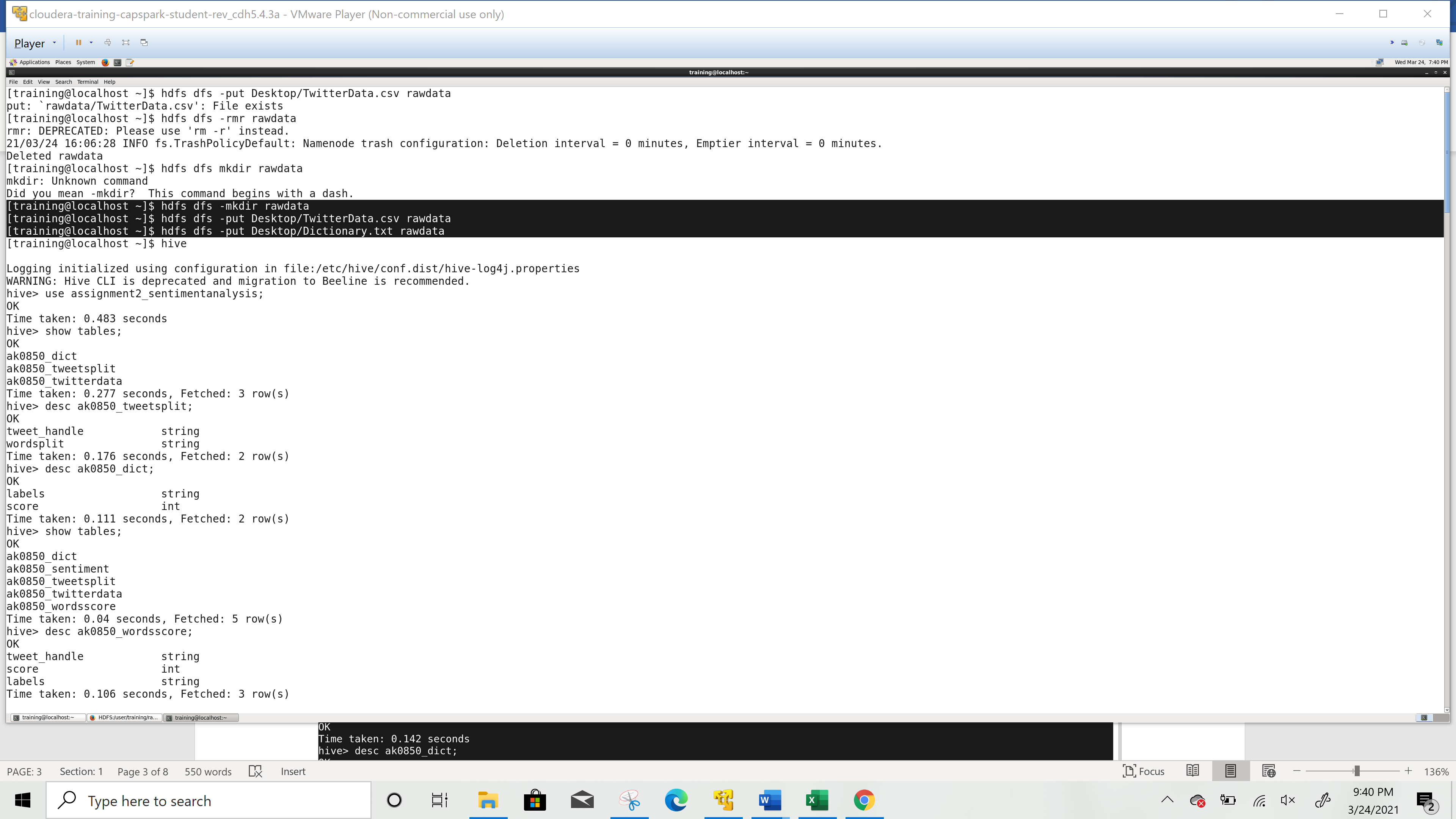
The data is loaded into above table from csv file which is copied to vmware instance and below is the snippet of code:

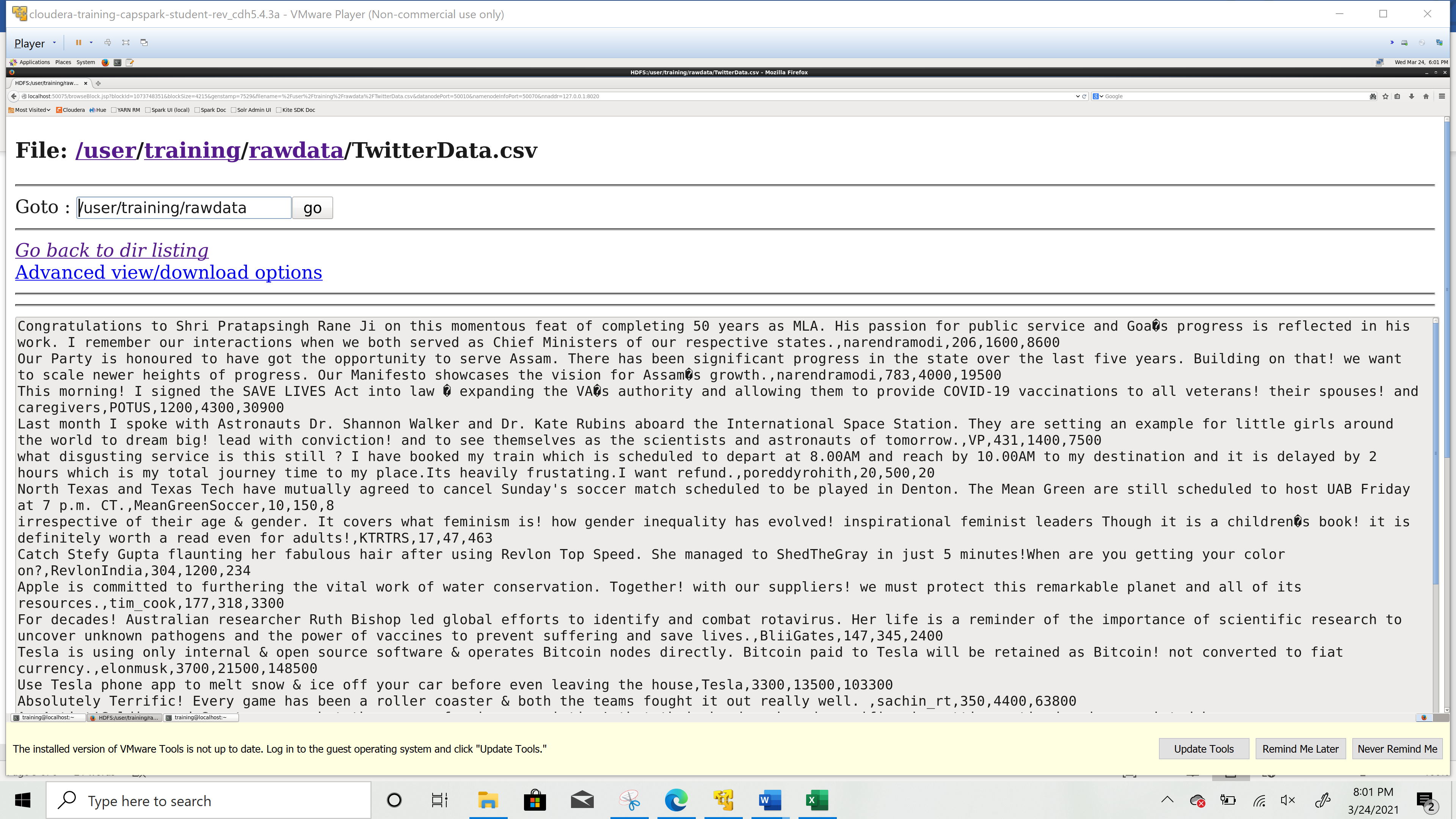


Below is the screen shot of data that is loaded into hive table:



TwitterData and Dictionary is loaded into hdfs and looks like below:





Dictionary file looks like below:

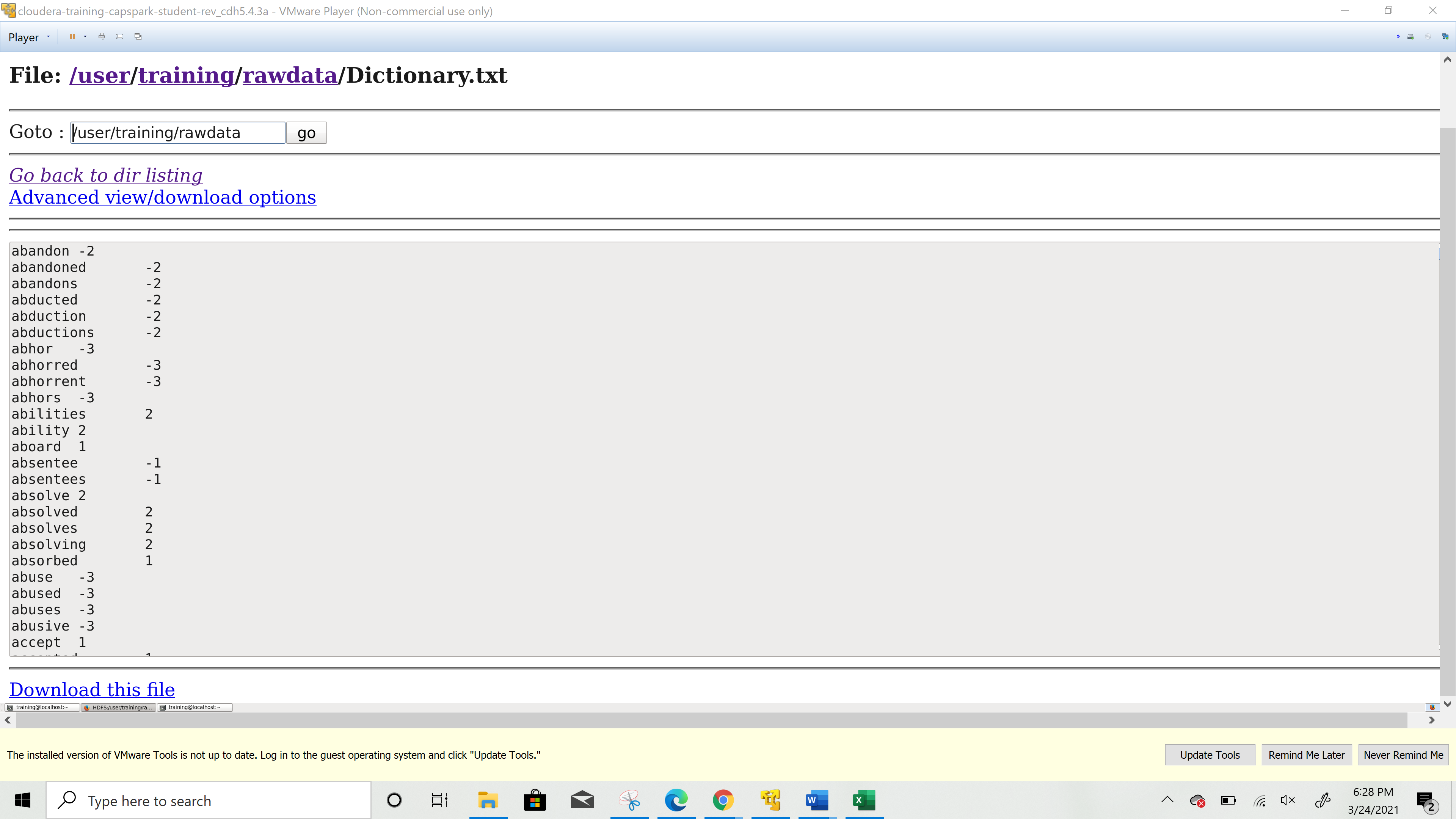
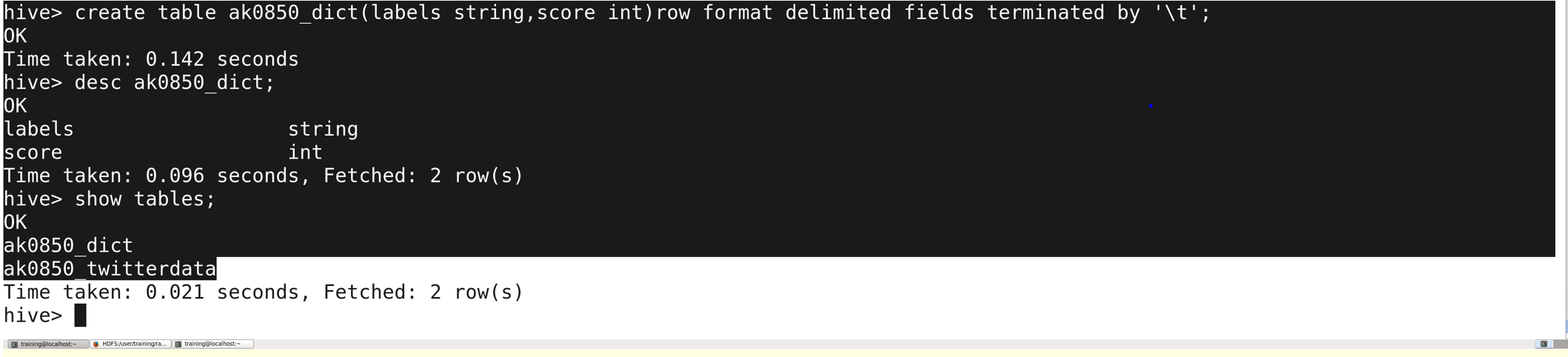
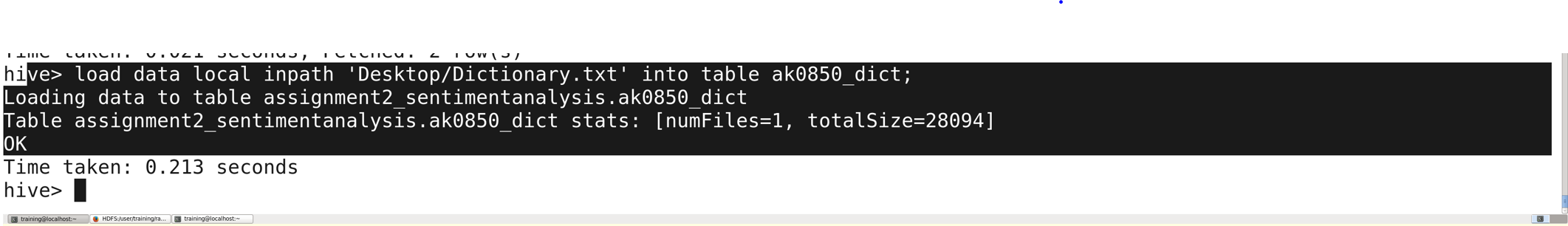


Table “ak0850\_dict” creation and description of table to load Dictionary.txt file .



Dictionary Data has been loaded into table :



Below is the snippet of table query result that holds the dictionary text values:

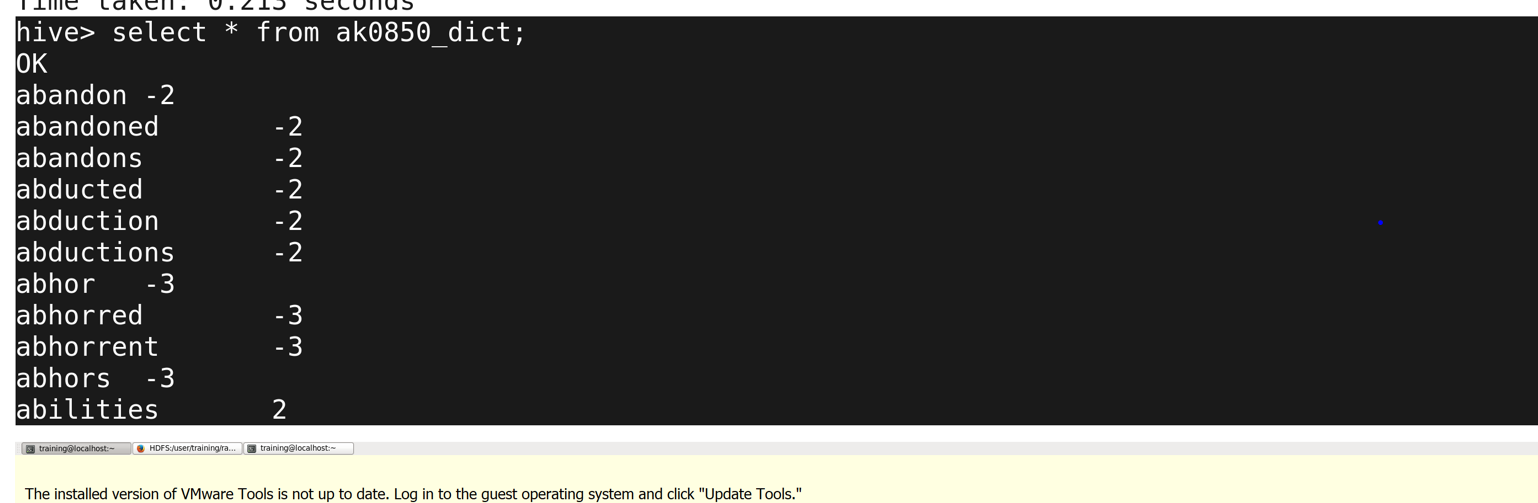
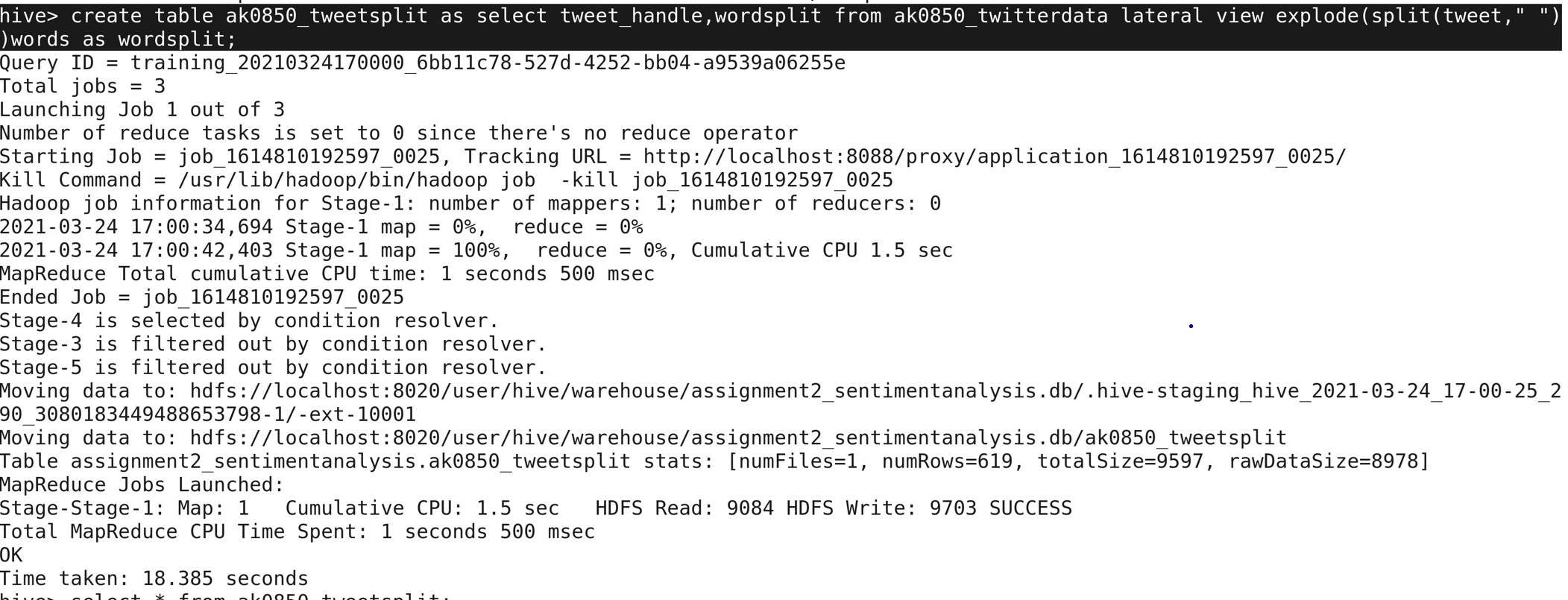
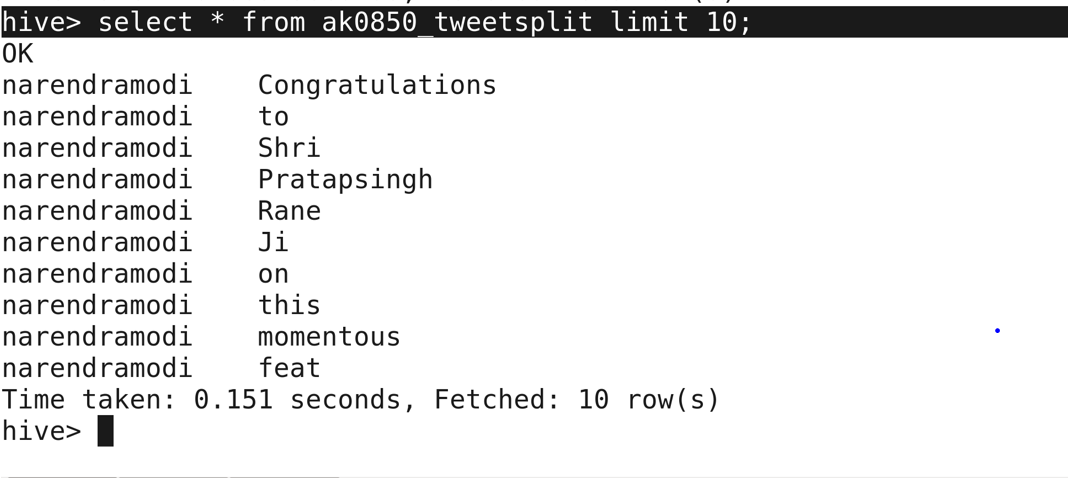


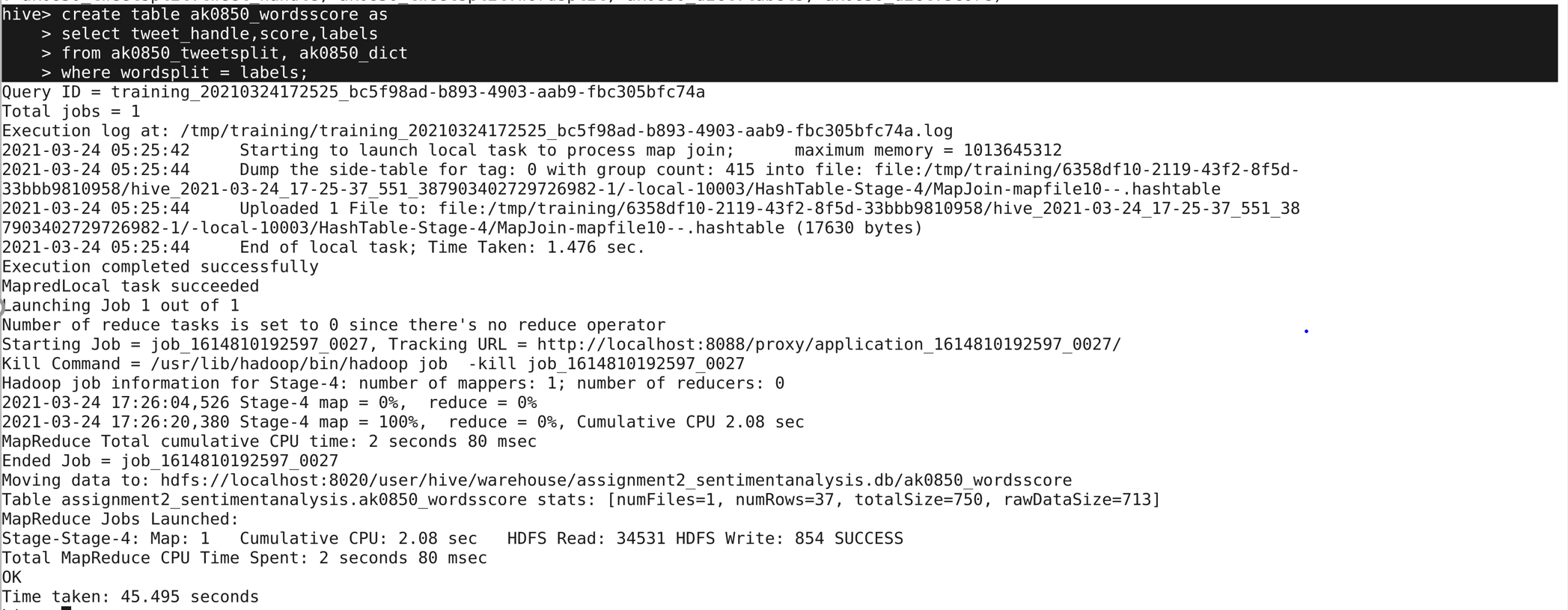
Table creation “ak0850\_tweetsplit” and loading data for tokenization of each tweet. This table contain words which are obtained by splitting each tweet and the tweet handler.



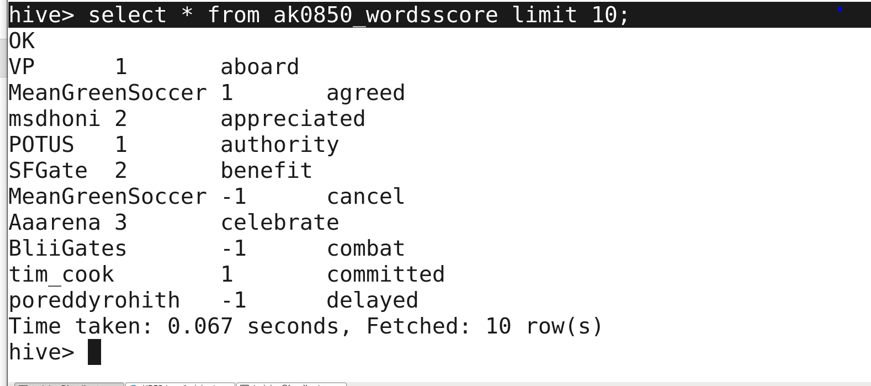
Split table data looks like below:



New table “ak0850\_wordsscore” has been created and loaded with tweet\_handle, scores and labels when the tokens in the split table matches with the labels in the dictionary table.



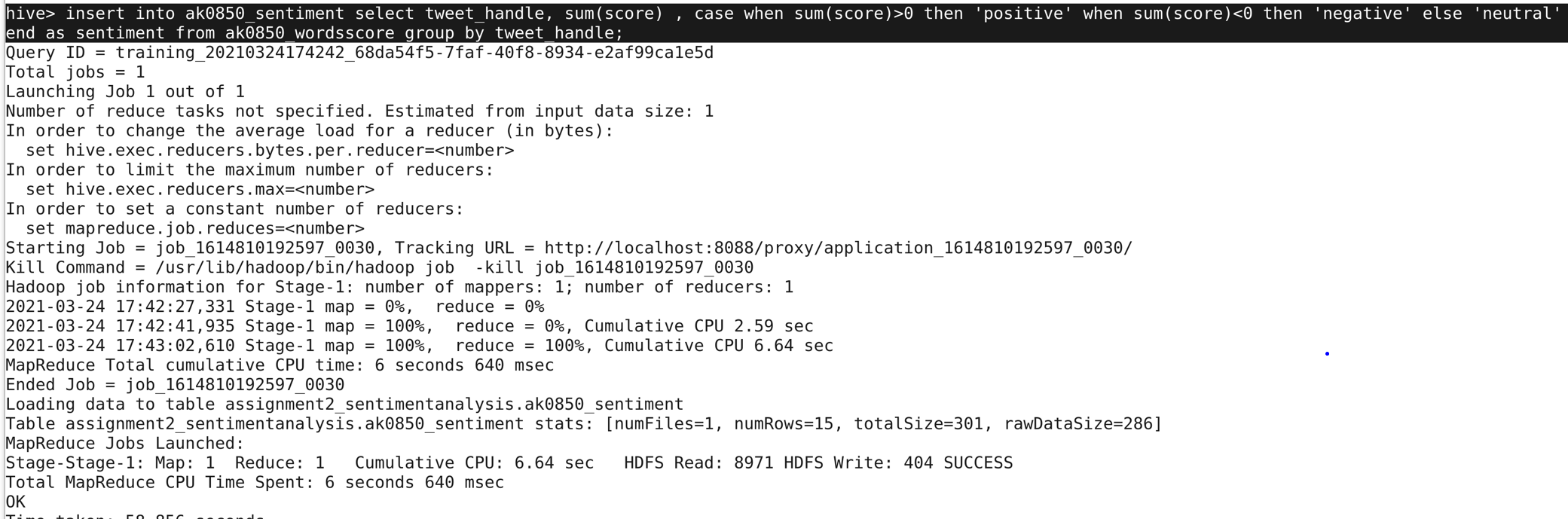
Sample of 10 records from ak0850\_wordsscore:



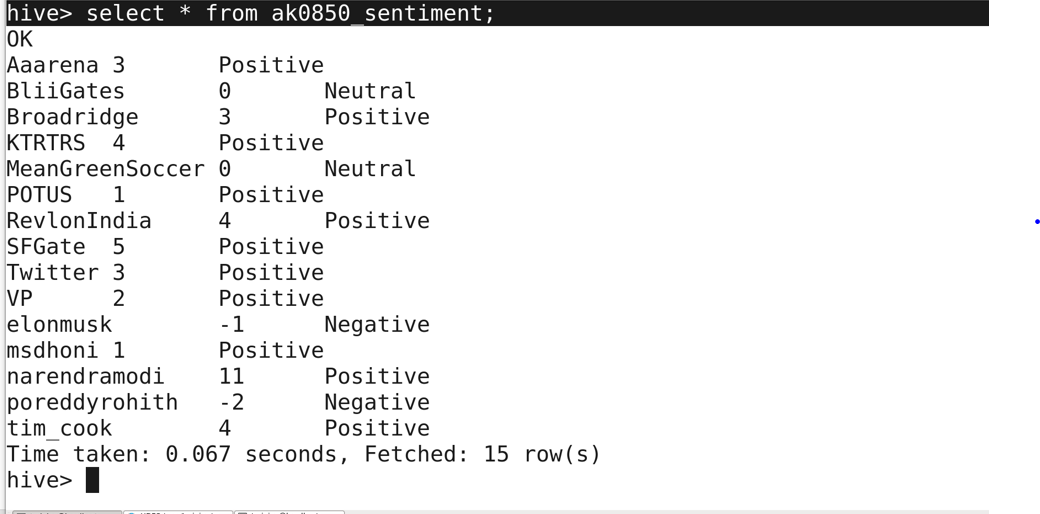
“ak0850\_sentiment” has been created to store the score, sentiment and tweet handler:



Above table ak0850\_sentiment is inserted with the score and the sentiment based on the sum of score value. If sum of score is greater then 0, it is considered as positive sentiment. If sum of score is less than 0, it is considered as negative sentiment else neutral sentiment.



Below is the snippet of data in ak0850\_sentiment table which displays the tweet handler, score and sentiment.



**Drawbacks:**

The main drawback of this approach is that if the data on which sentiment analysis needs to be performed doesn’t contain the words that match with dictionary tokens the accuracy will be decreased.

In order to improve the accuracy, it is recommended for the scholars to check the dictionary before proceeding with analysis and create a new list if required.

**Conclusion:**

Sentiment analysis helps to find out the number of people who are in favor of or against an issue. And also, the social interest and social pattern of every user. Sentiment analysis is considered for classification process. This dictionary-based approach lacks the ability to find the opinion words with domain. A new strategy has been proposed by Que and Hi i.e., an advertising strategy to improve user experience and relevance where they used dictionary based approach to identify sentiment based sentences. By using syntactic parsing and sentiment dictionary, they proposed rule based approach in order to tackle the topic word extraction and determining customer’s attitude in advertising keyword extraction. Using this approach they worked on web forums and the results demonstrated increase in accuracy and effectiveness using advertising keyword extraction and selection method. (Medhat, 2014) (Ingle, 2015)

**References:**

1. Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, *5*(4), 1093-1113.

|  |
| --- |
|  |
| 1. Ingle, A., Kante, A., Samak, S., & Kumari, A. (2015). Sentiment analysis of twitter data using hadoop. *International Journal of Engineering Research and General Science*, *3*(6), 144-147. |