**PRACTICE CASE**

**SOCIAL MEDIA SENTIMENT ANALYSIS WITH SPARK**

1. **Overview**

The purpose of this practice case is analyzing social media especially twitter with Spark. The dataset of this use case is clean\_data.csv (can be found on Jupyter Notebook folder). Because of numpy library can not be installed on this Spark, I will tell about data exploration with Spark only.

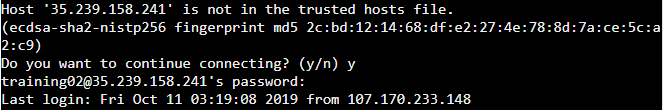
1. **Getting Started**

Before we go into the step of analysis, we should login on Spark :

* Open <https://bellard.org/jslinux/>, then choose Windows 2000.
* Login with the command below :



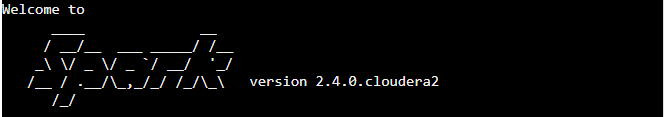
* If successful, then input the password Cl0ud3r4\*. The successful login status can be seen below.



* Type the command bellow to use Spark :



* If successful, the Spark will be shown as below :



1. **Steps**

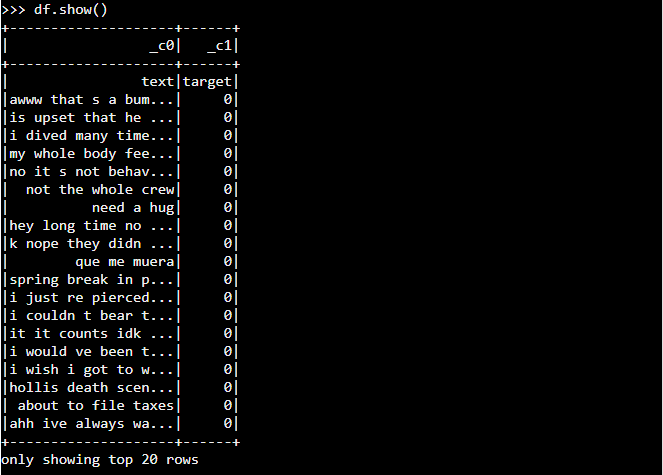
The step of analysis social media with Spark are :

1. First of all, import library to load dataframe on Spark.

/user/cloudera/clean\_tweet.csv is the file directory of dataset.

D:\ADSB-Batch-2\Spark\practice\3.PNG

1. Show dataset by using df.show()



1. Count the row of dataset use df.count()

D:\ADSB-Batch-2\Spark\practice\count total data.PNG

From the above we can know that dataset has 1600K rows.

1. We need to drop duplicates in dataset, use df.dropna()

D:\ADSB-Batch-2\Spark\practice\drop null and count after drop.PNG

After we drop duplicates, the dataset now has 1.596.754 rows.