Lab3: Spark SQL

- do the code in google colab
- before u download the code , delete the section of uploading the file and also the pip install, this one:

  ```
  ! pip install pyspark
  import pandas as pd
  from google.colab import files
  uploaded = files.upload()
  ```

- And then change the path of the csv file to its placement in hdfs (since we will execute the code in spark)

  ```
  df = spark.read.csv("hdfs://hadoop-master:9000/input/ngram.csv", header=True, schema=schema,sep='\t').limit(100)
  ```

  → Note that the path of hdfs it depends on your configuration so to change it depends on yours , try to write this command `cat $HADOOP_HOME/etc/hadoop/core-site.xml` you will get somthing like:

  ```
  <?xml version="1.0"?>
  <configuration>
     <property>
        <name>fs.defaultFS</name>
        <value>hdfs://hadoop-master:9000/</value> ← get this as the path of hdfs!
     </property>
  </configuration>
  ```

- then copy the `ngram.csv` to hdfs and the `code.py` to the locall and then run your code

  ```
  spark-submit --master spark://b524f35852c2:7077 lab3_bigdata.py > output_lab3.txt
  ```

- If your code take a long time on the execution, you can do a limit of uploading the dataset in ur code like this

  ```
  df = spark.read.csv("hdfs://hadoop-master:9000/input/ngram.csv", header=True, schema=schema,sep='\t').limit(100)
  ```