**Lab 3 (questions words frequency)**

**Name: Zeyad Mohamed ID:21ZTEM**

**Using PIG Latin:**

**Step 1: Copy the whquestion.txt file into a hdfs location.**

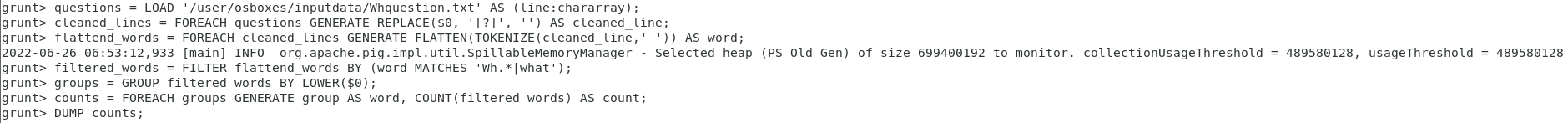
****

**Step 2: open grunt shell using pig command in terminal.**

**Text

Description automatically generated with medium confidence**

**Step 3: write required commands.**

****

**Output:**

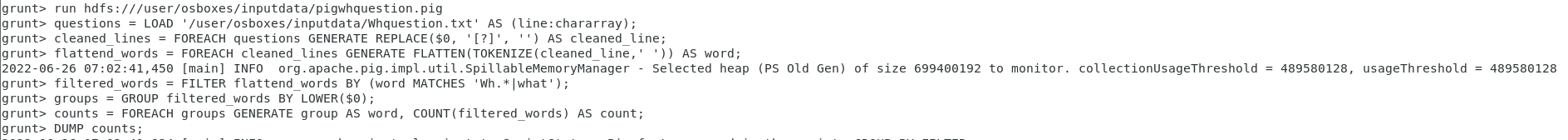
**A picture containing graphical user interface

Description automatically generated**

**Step 4: Save pig commands into pig file and copy the pig file from local drive into hdfs.**

****

**Step 5: run pig file using run command.**

****

**Output:**

**Text

Description automatically generated with medium confidence**

**Using Hive:**

**Step 1: open hive.**

****

**Step 2: Create table**

****

**Step 3: Load data from text file into table and verify the data.**

**Text, letter

Description automatically generated**

**Step 4: Run required command to get wh words frequancies.**

****

**Output:**

**Text, letter

Description automatically generated**

**Hive commands:**

**CREATE TABLE IF NOT EXISTS questions(question STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\n';**

**LOAD DATA INPATH '/user/osboxes/inputdata/Whquestion.txt' INTO TABLE questions;**

**SELECT \* FROM questions;**

**SELECT LOWER(wh\_word), SUM(wh\_count) FROM (SELECT REPLACE(SPLIT(question,' ')[0],'?','') AS wh\_word, 1 wh\_count FROM questions WHERE question REGEXP "^Wh\*" OR question LIKE "%what") AS count\_problem GROUP BY (wh\_word);**

**My findings:**

**The time to execute hive commands was small regarding to pig and we solve that problem in one SELECT command and that’s why hive was faster, and I notice that hive is different from pig where hive sort the output records alphabetically but the output in PIG was not ordered, also hive was better than PIG in case of memory where in PIG we required to save every variable, but in hive we just view results and we didn’t require any memory, and the most frequented word in hive and PIG was ‘What’ , and the least frequented words were ‘When’, ‘Where’, ‘Which’ , ’Whom’ 95 words.**