Master's Degree in Informatics Engineering
MASSIVE DATA PROCESSING
CORES PRADO, FERNANDO
ACTIVITY HADOOP - MAPREDUCE
KIN TAT, TAN

Universitat de Lleida

ACTIVITY HADOOP - MAPREDUCE

KIN TAT, TAN

Master's Degree in Informatics Engineering
MASSIVE DATA PROCESSING
CORES PRADO, FERNANDO
ACTIVITY HADOOP - MAPREDUCE
KIN TAT, TAN

Universitat de Lleida

Sumário

Master's Degree in Informatics Engineering
MASSIVE DATA PROCESSING
CORES PRADO, FERNANDO
ACTIVITY HADOOP - MAPREDUCE
KIN TAT, TAN

Universitat de Lleida

GitHub

https://github.com/tankintat/bigdataudl/tree/master/Hadoop%20-%20MapReduce%20-%20Act

Introduction

I created the artefacts in the same package, that means makes more easily to put debugger and test in the HDFS. In this document, I will show the results getting out in the debugger output and HDFS execution. Some input archives I remove some data because my computer brake in some processes.

Activity 1 – Trending Tops

The objective about this activity was extract all the existence hashtags in a input text file. To a Twitter is a most important to know how much comments are talking about during the day. So the implementation was about the classical architecture MapReduce and the example count words we got the solution.

hadoop jar Artefactos/ActMapReduce.jar eps.examples.mapreduce.Task1 /user/tan/Act1Input1/ /user/tan/Act1Output1

Result the activity 1

```
#vtop   2
#virgo  1
#vivavideo      1
#voice1242      1
#vs     1
#waitingfordonghae      3
#waitingforjejung       1
#webcam 1
#webcamsex      1
#welcometweet   1
#werk   1
#wewalkbyfaith  1
#wewantonedirectiondead 2
#whenyourmomisondiet    1
#whiteprivilege 1
#wisconsinprimary       1
#wisdom 1
#wonderwall     1
#workout        1
#wotblitz       1
#wowjobs        1
#wt20   2
#wuyifan        1
#ye     1
#yellow 1
#yemen  3
#yesterday      1
#ygfamily       2
#yokooorer      1
#yuzu   1
#zaynmalik1d    1
#zdjeciaktorenieprzestanamniesmieszyc   2
#zdjeciaktorezrylymipsychike    1
#zzp    1
```

Universitat de Lleida

Activity 2 – Text cleanup using chain mapper

The objective this activity was using a json file to filter the properties in order to get information. Therefore was using the chain mapper and reduce because the chain implementation is possible to makes a sequence process mapper and reduce, so are more easily to control each step by the extraction.

First Mapper Step: Separate the properties in the json input file. The objective its to ignore the useless properties and use just useful properties.

Second Mapper Step: Get just extract the defined language twitted.

Third Mapper Step: Assuring has a Hashtag and Text in the tweet.

Fourth Mapper Step: Text-lower case

First Reduce: To count the similar hashtags

hadoop jar Artefactos/ActMapReduce.jar eps.examples.mapreduce.Task2 /user/tan/Act2Input1/ /user/tan/Act2Output1

```
#louisphilippede…       1
rt @nct_smrookies: 160502\\nto all that've been waiting for nct mv commentary eng subs,here you go!\\nall thanks to subber❤\\n#nct
#nct_u \\nhttps:…       1
rt @newsbreakslive: #china #army rap to boost recruitment... '#kill! kill! kill!' https://t.co/czhuhxprld       1
rt @nickjonas: thank you, @topman. #metgala #manusxmachina https://t.co/mjpo2ast54      1
rt @onemusicph: jadine at #thistimepremierenight https://t.co/pb3u41cr9o        1
rt @pinoy_quotes: achieve your goals. trust yourself.\\n\\n#lipatnapoe  1
rt @posh_da_popper: \\dollhouse\\👁ft my sister #modd coming soon... #gooddoll🎀#baddoll👽#teamqueenp https://t.co/t6k91moyfp  1
rt @savagexbehavior: when you jump to conclusions .. you can miss a whole lot .. #justsaying    1
rt @seabirdcentre: #win! a long-weekend in #northberwick for #puffinfest thanks to our friends at gilsland park https://t.co/256k75q7kt
htt…    1
rt @southendtravel: need to get to or from #gatwick #heathrow #stanstead \\n#airport #cheapest #taxi #minicabs \\nhttps://t.co/
lvoplceym5 https…       1
rt @ucheguys: #nnamdi kanu will come to appeal court come 5/52016 we pray good news that day.@david_cameron https://t.co/citgukqnxg
1
rt @willynyamitwe: @anotherburundi and now zeid ra'ad hussein brings in an other story : ex-fab, tutsis, hunted down by hutus. hate
&amp;amp; divis…    1
rt @xxmantra: flawless #metgala https://t.co/73o4n8m8db 1
shopify online store $9/m @madebymissm made by miss m: turquoise &amp;amp; coral working together pretty amazing #mad...  https://t.co/
xxnc104kec      1
si la loo babylon, bow down to the richest man on earth! #billionaire   1
spin bob ... spin\\n#mortgage \\n#realestate\\n#homesforsale\\n#homeownership\\n#housing\\n@fanniemae\\n@freddiemac\\n$fnma\\n$fmcc
https://t.co/tb7goialtr 1
this is exactly the issue with the jews #freepalestine https://t.co/gsqqq9wcch  1
wearable technology ecosystem 2016 to 2030 - analysis of 7 devices and 73 countries https://t.co/gzkidyjshw #wearables #iot     1
were you feeling #kimye's #metgala look? catch up on all the hottest (and not so hot) looks from… https://t.co/g6lcutsceb       1
why are you excited to try hawaiian tropic® silk hydration weightless lotion suns... #sponsored #hawaiiantropiccrowd https://t.co/
hxkthojamx      1
wow! my twitter is worth $6,110.40 today. find out your twitter fantasy value free! click here ==&amp;gt; https://t.co/d621lr39iq #rt
#followme       1
wow. co-op (via crtg) didn't get collective purchasing nailed until 2002 - way behind sus. #nussc2016| more: https://t.co/us82lciqyu
1
yeayyyyy dh nk start 😊 #diasemanishoney        1
```

Activity 3 – The Top-N pattern

The objective this activity was using all methods about the Mapper and Reduce, so we can use the setup and cleanup in these classes to order to get more fast and clear code.

Universitat de Lleida

The idea its get the results getting in the exercise one to input this activity, so then each mapper get a part the document and appoint the more appears with TreeMap help and passing one reduce to couple all mappers results and put the Top N Hashtags. The solution was convert the input and output type to Text. I loose a lot of time in a error that no makes sense, compatibility input and output. So I just use the [job.setInputFormatClass(KeyValueTextInput.Format.class)] to convert input to Text.

hadoop jar Artefactos/ActMapReduce.jar eps.examples.mapreduce.Task3 /user/tan/Act1Output1/part* /user/tan/Act3Output1/ 15

```
 #iHeartAwards  780
 #BestFanArmy   668
#RT     261
 #Directioners  218
#ALDUBStoryContinues    217
 #ALDUBStoryContinues   190
 #EgyptAir      178
 #5SOSFam       175
 #MessagesOfMSG 173
#BestFanArmy    147
```

Activity 4 - Sentiment of hashtags

The objective this activity was in each tweet get the quantity of the positive and negative in the sentence, to know if the hashtags its positive comments ou negative.

Use the similar word count, but use more two inputs documents positive and negative list, to find in each sentence how many words appears. And finally use the Weighted arithmetic to find the results.

hadoop jar Artefactos/ActMapReduce.jar eps.examples.mapreduce.Task4 /user/tan/Act2Output1/ /user/tan/Act4Output1/ /user/tan/positive-words.txt /user/tan/negative-words.txt

Universitat de Lleida

```
#pti    0.0
#puffinfest     0.0
#quanticpost    1.0
#quote  -1.0
#rdlc   0.0
#reporting      0.0
#rittersellsfast        1.0
#riverview      0.0
#rockclock      0.0
#rt     0.549407114624506
#rtした人全員フォローする   0.0
#ryderfor33mainevent    0.0
#selenagomez    0.0
#shoes  0.0
#sleepingbeauties       2.0
#software       0.0
#sponsored      1.0
#sql    0.0
#stanstead      0.0
#stfm16 -1.0
#system 0.0
#taxi   0.0
#teamfollowback 0.0
#teamqueenp     0.0
#tech   0.0
#technology     0.0
#tfb    0.0
#tfw    0.0
#thistimepremierenight  0.0
#wattpad        0.0
#wearables      0.0
#weewx  0.0
#win!   0.0
```

Activity 5 - Join all previous mapreduce Job in a single application

The objective the last activity its join all the previous exercises in a one application. But the problem was use the ChainMapper concept, cause each exercise was makes a different way and I had with compatibility problem.