Proyecto de clase

- Jean Carlo Soto
- Aaron Merino

Objetivo General

El objetivo de este trabajo es aprender cómo usar la aplicación de **Map/Reduce** en un sistema de archivos distribuidos de **Hadoop**

Recordemos conceptos importantes

Qué es hadoop?

Hadoop es una estructura de software de código abierto para almacenar datos y ejecutar aplicaciones en clústeres de hardware comercial. Proporciona almacenamiento masivo para cualquier tipo de datos, enorme poder de procesamiento y la capacidad de procesar tareas o trabajos concurrentes virtualmente ilimitados.

Qué es MapReduce?

MapReduce es un framework que proporciona un sistema de procesamiento de datos paralelo y distribuido. Su nombre se debe a las funciones principales que son **Map** y **Reduce**

Map

La función Map recibe como parámetros un par de (clave, valor) y devuelve una lista de pares. Esta función se encarga del mapeo y se aplica a cada elemento de la entrada de datos, por lo que se obtendrá una lista de pares por cada llamada a la función Map. Después se agrupan todos los pares con la misma clave de todas las listas, creando un grupo por cada una de las diferentes claves generadas.

Reduce

La función Reduce se aplica en paralelo para cada grupo creado por la función Map(). La función Reduce se llama una vez para cada clave única de la salida de la función Map. Junto con esta clave, se pasa una lista de todos los valores asociados con la clave para que pueda realizar alguna fusión para producir un conjunto más pequeño de los valores.

Que utilizamos para desarrollar el proyecto?

Python

Hadoop & Hadoop MapReduce

Java

Nuestro dataset

ghtorrent-2019-05-20.csv (16.15 GB)

Detail Compact Column

10 of 11 columns

actor_login	actor_id	comment_id	comment	repo
houndci-bot 2% codacy-bot 1% Other (80068154) 98%	1 — 51.7m	192m — 290m	4396695 unique values	elasticsearch kibana Other (75522868)
EleisonC	37538393	197673683	That has been rectified.	openmrs-ocl-clien
cooltey	2354559	289490974	Please use `setActualImageResource` instead of `setImageResource`	apps-android-wikipedia
felixfbecker	8865192	250903696	My suggestion would be collapsing both properties to `{ updated: undefined | true | typeof LOADING |...	sourcegraph
slimsag	4019296	252913723	I think that if you encapsulate the success logic like this it makes the code harder to read	sourcegraph

Por qué python?

Para seleccionar una columna especifica de nuestro dataset adquirido como un .csv, posteriormente copiando estos datos a un .txt

Programa de python

```python

import csv
from datetime import datetime

start_time = datetime.now()

with open('/Users/jeancasoto/Downloads/ghtorrent-2019-05-20.csv', 'rb') as f:
    reader = csv.reader(f)
    next(reader)  # Ignoramos nombre de la columna

    with open('/Users/jeancasoto/Downloads/pullRequestsComments.txt', 'w') as nf:
        for row in reader:
            nf.write(row[3]+'\n')

end_time = datetime.now()
print('Duration: {}'.format(end_time - start_time))

```

output

```
jeancasoto@Jeans-MacBook-Pro ~ % cd Downloads/
jeancasoto@Jeans-MacBook-Pro Downloads % python csvToTxt.py
Duration: 0:08:22.167532
```

Resultado obtenido

pullRequestsComments.txt — showing 128 MB of 8.06 GB Open with TextEdit

```
That has been rectified.
Please use `setActualImageResource` instead of `setImageResource`
My suggestion would be collapsing both properties to `{ updated: undefined | true | typeof LOADING | ErrorLike }`
I think that if you encapsulate the success logic like this it makes the code harder to read
Ya its still not clear to me what exactly the behavior should be which is i guess the bigger question i should have
asked. \\One path is:\* throwing - Any exception that occurs that would cause pipeline failure is rethrown. User code
exceptions are
Agree. Updated it. Will merge/deploy when green.
You could make this a const and then below assign top/bottom directly, note that you don't even need the if
placementObj['style'] this way:\\const margins = {\   top: parseInt((placementObj['style'] || {})['top_m'], 10)  ||
undefined,\   bottom =
that sounds good
does this need to be 'conf' instead?
Renamed these all to be more consistent after doing more research into standard KEM patterns.
I think that using curly braces + field names is a better style in general (it's safer, more readable, and it's
slightly easier to search for some things), although here it does not make much of a difference. For consistency I
would also use the same sty
```suggestion\* [Get started](https://docs.improbable.io/unreal/latest/content/get-started/introduction) (on the
SpatialOS documentation website)\```
Change tabs to spaces
Whoops
Maybe this should be in the root directory and not in SpatialGDK. I think it's a little strange having it in a
subfolder when there is SpatialGDK and SpatialGDKEditorToolbar.
Given that it's specified in `em` is a sign that it's something I haven't updated to the new system yet, so removing
the value should be fine. We should consider using CSS grids for this specific layout anyway.
> share and [.] their know-how
'$prompt' is already declared in the upper scope   no-shadow
'$' is not defined
no-undef
There should be a line space between each test case.
You can use <>
below (one l)
Can't we move it outside for loop in order not to allocate memory each time?
all other keys here seem to be snake_case. could this be changed to payment_authorization_id?
Not working :-(
Won't the main context be cancel when legitimately stopping the cluster-agent, hence should this be an `Error`?
Let it crash. Allowed types validated on request, so unsupported typed shouldn't be here
Should this be using the baseFragment that you have put in place in another PR?
Indeed, I don't see why this was needed! And it seems that it has been added long time ago.
Fixed
I've added a 'lenient' strictness to the test class to ignore `UnnecessaryStubbingException`s because in this case
the mock stub isn't unnecessary.
If I'm reading things correctly this needs to be a little less strong (e.g. `looks up non-cached start timestamps in
batches`) in case someone calls this with more than `TRANSACTION_TIMESTAMP_LOAD_BATCH_LIMIT` uncached timestamps
```
```

Por qué Java?

Para la limpieza de la data recolectada anteriormente por el programa de python, limpiandola así de stop words, urls, etc…

Asimismo realizar el **Frequency Analysis**

Ejemplo Stop words

- a
- about
- above
- after
- again
- against
- all
- am
- and
- any
- are
- arent
- as
- at
- be
- because
- been
- before
- being
- below
- between
- both
- but
- by
- can
- cant
- cannot
- could
- couldnt
- did
- didnt
- do
- does
- doesnt
- doing
- dont
- down
- during
- each
- few
- for
- from
- further
- had
- hadn;t
- has
- hasnt
- have
- havent
- having
- he
- hed
- hell
- hes
- her
- here
- heres
- hers
- herself
- him
- himself
- his
- how
- hows
- i
- id
- ill
- im
- ive
- if
- in
- into
- is
- isnt
- it
- its
- its
- itslef
- me
- more
- most
- my
- myself
- no
- not
- of
- off
- on
- once
- only
- or
- other
- ought
- our
- ours
- out
- over
- own
- same
- she
- shed
- shell
- shes
- should
- shildnt
- so
- some
- such
- than
- that
- thats
- the
- their
- theirs
- them
- themselves
- then
- there
- theres
- these
- they
- theyd
- theyll
- theyre
- theyve
- this
- those
- through
- to
- too
-

entre otras ...

Por qué Hadoop & Hadoop MapReduce?

Para realizar la aplicación **WordCount**. Esta aplicación básicamente recibe un archivo de texto y devuelve otro archivo que enumera cada palabra encontrada en el archivo de entrada y la cantidad de veces que dicha palabra apareció.

Código utilizado en Map & Reduce

Mapper

```
public static class TokenizerMapper
     extends Mapper<Object, Text, Text, IntWritable>{

  private final static IntWritable one = new IntWritable(1);
  private Text word = new Text();

  public void map(Object key, Text value, Context context
                  ) throws IOException, InterruptedException {
    StringTokenizer itr = new StringTokenizer(value.toString());
    while (itr.hasMoreTokens()) {
      word.set(itr.nextToken());
      context.write(word, one);
    }
  }
}
```

Reduce Word Count

```
public static class IntSumReducer
     extends Reducer<Text,IntWritable,Text,IntWritable> {
  private IntWritable result = new IntWritable();

  public void reduce(Text key, Iterable<IntWritable> values,
                     Context context
                     ) throws IOException, InterruptedException {
    int sum = 0;
    for (IntWritable val : values) {
      sum += val.get();
    }
    result.set(sum);
    context.write(key, result);
  }
}
```

Mapper 2 Word Frequency

```java
public static class MAPPER extends Mapper<Object, Text, Text, IntWritable> {

    IntWritable one = new IntWritable(1);
    Text word = new Text();

    public void map(Object key, Text value, Context context) throws IOException, InterruptedException {
        String[] palabras = value.toString().split("\\s+");
        String str1 = null;
        String str2;
        if (palabras.length != 0) {
            str1 = palabras[0];
        }
        for (int i = 1; i < palabras.length; i++) {
            str2 = palabras[i];
            word.set(str1 + " , " + str2);
            context.write(word, one);
            str1 = str2;
        }
    }
}
```

Reduce 2 Word Frequency

```
public static class REDUCER extends Reducer<Text, IntWritable, Text, IntWritable> {

    private IntWritable result = new IntWritable();

    public void reduce(Text key, Iterable<IntWritable> values, Context context) throws IOException, InterruptedException {
        int suma = 0;
        for (IntWritable val : values) {
            suma += val.get();
        }
        result.set(suma);
        context.write(key, result);
    }
}
```

HDFS

Application application_161 | Browsing HDFS

localhost:9870/explorer.html#/

Hadoop | Overview | Datanodes | Datanode Volume Failures | Snapshot | Startup Progress | Utilities

Browse Directory

/ Go!

Show 25 entries

Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	hdoop	supergroup	0 B	Mar 24 08:49	0	0 B	WordCount
drwxr-xr-x	hdoop	supergroup	0 B	Mar 24 07:24	0	0 B	WordCount2Freq
drwx------	hdoop	supergroup	0 B	Mar 24 08:17	0	0 B	tmp

Showing 1 to 3 of 3 entries

Previous 1 Next

Hadoop, 2021.

HDFS/WordCount

Application application_161 | Browsing HDFS

localhost:9870/explorer.html#/WordCount

Hadoop | Overview | Datanodes | Datanode Volume Failures | Snapshot | Startup Progress | Utilities

Browse Directory

/WordCount | Go!

Show 25 entries | Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	hdoop	supergroup	0 B	Mar 24 08:43	0	0 B	Input
drwxr-xr-x	hdoop	supergroup	0 B	Mar 24 08:31	0	0 B	Output0
drwxr-xr-x	hdoop	supergroup	0 B	Mar 24 08:51	0	0 B	OutputUsernames

Showing 1 to 3 of 3 entries

Previous 1 Next

Hadoop, 2021.

HDFS/WordCount/Input

Application application_161 | Browsing HDFS

localhost:9870/explorer.html#/WordCount/Input

Hadoop | Overview | Datanodes | Datanode Volume Failures | Snapshot | Startup Progress | Utilities

Browse Directory

/WordCount/Input Go!

Show 25 entries Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hdoop	supergroup	5.23 GB	Mar 24 08:08	1	128 MB	cleanDataComplete.txt
-rw-r--r--	hdoop	supergroup	771.01 MB	Mar 24 08:43	1	128 MB	usernamePullRequests.txt

Showing 1 to 2 of 2 entries

Previous 1 Next

Hadoop, 2021.

WordCount2Freq

Application application_161 | Browsing HDFS

localhost:9870/explorer.html#/WordCount2Freq

Hadoop | Overview | Datanodes | Datanode Volume Failures | Snapshot | Startup Progress | Utilities

Browse Directory

/WordCount2Freq Go!

Show 25 entries Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	hdoop	supergroup	0 B	Mar 24 07:39	0	0 B	Input
drwxr-xr-x	hdoop	supergroup	0 B	Mar 24 07:30	0	0 B	Output0
drwxr-xr-x	hdoop	supergroup	0 B	Mar 24 02:14	0	0 B	Output1
drwxr-xr-x	hdoop	supergroup	0 B	Mar 24 02:33	0	0 B	Output2
drwxr-xr-x	hdoop	supergroup	0 B	Mar 24 07:11	0	0 B	Output3

Showing 1 to 5 of 5 entries

Previous 1 Next

Hadoop, 2021.

HDFS/WordCount2Freq/Input

Application application_161 | Browsing HDFS

localhost:9870/explorer.html#/WordCount2Freq/Input

Hadoop | Overview | Datanodes | Datanode Volume Failures | Snapshot | Startup Progress | Utilities

Browse Directory

/WordCount2Freq/Input Go!

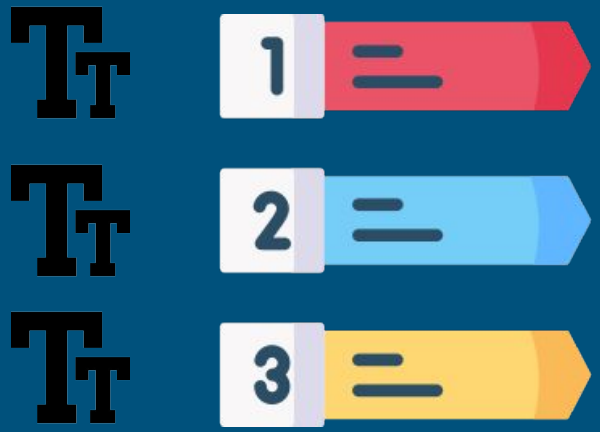
Show 25 entries Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hdoop	supergroup	1.5 GB	Mar 23 22:15	1	128 MB	cleanDataSegment0.txt
-rw-r--r--	hdoop	supergroup	1.5 GB	Mar 23 22:15	1	128 MB	cleanDataSegment1.txt
-rw-r--r--	hdoop	supergroup	1.5 GB	Mar 23 22:16	1	128 MB	cleanDataSegment2.txt
-rw-r--r--	hdoop	supergroup	745.23 MB	Mar 23 22:17	1	128 MB	cleanDataSegment3.txt

Showing 1 to 4 of 4 entries

Previous 1 Next

Hadoop, 2021.



Word Count Results

Data sin limpiar

pullRequestsComments.txt — showing 128 MB of 8.06 GB Open with TextEdit

```
That has been rectified.
Please use `setActualImageResource` instead of `setImageResource`
My suggestion would be collapsing both properties to `{ updated: undefined | true | typeof LOADING | ErrorLike }`
I think that if you encapsulate the success logic like this it makes the code harder to read
Ya its still not clear to me what exactly the behavior should be which is i guess the bigger question i should have
asked. \\One path is:\* throwing - Any exception that occurs that would cause pipeline failure is rethrown. User code
exceptions are
Agree. Updated it. Will merge/deploy when green.
You could make this a const and then below assign top/bottom directly, note that you don't even need the if
placementObj['style'] this way:\\const margins = {\   top: parseInt((placementObj['style'] || {})['top_m'], 10)  ||
undefined,\   bottom =
that sounds good
does this need to be 'conf' instead?
Renamed these all to be more consistent after doing more research into standard KEM patterns.
I think that using curly braces + field names is a better style in general (it's safer, more readable, and it's
slightly easier to search for some things), although here it does not make much of a difference. For consistency I
would also use the same sty
```suggestion\* [Get started](https://docs.improbable.io/unreal/latest/content/get-started/introduction) (on the
SpatialOS documentation website)\```
Change tabs to spaces
Whoops
Maybe this should be in the root directory and not in SpatialGDK. I think it's a little strange having it in a
subfolder when there is SpatialGDK and SpatialGDKEditorToolbar.
Given that it's specified in `em` is a sign that it's something I haven't updated to the new system yet, so removing
the value should be fine. We should consider using CSS grids for this specific layout anyway.
> share and [.] their know-how
'$prompt' is already declared in the upper scope   no-shadow
'$' is not defined
no-undef
There should be a line space between each test case.
You can use <>
below (one l)
Can't we move it outside for loop in order not to allocate memory each time?
all other keys here seem to be snake_case. could this be changed to payment_authorization_id?
Not working :-(
Won't the main context be cancel when legitimately stopping the cluster-agent, hence should this be an `Error`?
Let it crash. Allowed types validated on request, so unsupported typed shouldn't be here
Should this be using the baseFragment that you have put in place in another PR?
Indeed, I don't see why this was needed! And it seems that it has been added long time ago.
Fixed
I've added a 'lenient' strictness to the test class to ignore `UnnecessaryStubbingException`s because in this case
the mock stub isn't unnecessary.
If I'm reading things correctly this needs to be a little less strong (e.g. `looks up non-cached start timestamps in
batches`) in case someone calls this with more than `TRANSACTION_TIMESTAMP_LOAD_BATCH_LIMIT` uncached timestamps
????
??????? ??? ??? ??? ???? ?? ????? ?????????? ? `try-catch`
Not sure what you mean by short-circuit here.
form???reset????????????
We can use `when-first` here
ditto
```
```

Data limpia

cleanDataComplete copy.txt — showing 128 MB of 5.23 GB Open with TextEdit

```
   rectified
please  setactualimageresource instead  setimageresource
 suggestion   collapsing  properties   updated undefined    typeof loading  errorlike
 think    encapsulate  success logic    makes  code harder  read
ya  still  clear    exactly  behavior      guess  bigger question    asked one path  throwing    exception  occurs
cause pipeline failure  rethrown user code exceptions
agree updated   mergedeploy  green
     const    assign topbottom directly note        placementobjstyle  wayconst margins     top
parseintplacementobjstyle  top  10   undefined   bottom
 sounds good
     conf instead
renamed      consistent    research  standard kem patterns
 think  using curly braces  field names   better style  general  safer  readable   slightly easier  search   things
although         difference  consistency       sty
suggestion  started    spatialos documentation website
change tabs  spaces
whoops
      root directory    spatialgdk  think   little strange     subfolder    spatialgdk  spatialgdkeditortoolbar
given   specified  em   sign      updated   new system   removing  value   fine   consider using css grids   specific
layout anyway
 share
prompt  already declared   upper scope    shadow
   defined
undef
    line space   test case

 one
  move  outside  loop  order   allocate memory  time
  keys  seem   snake case    changed  payment authorization
 working
wont  main context  cancel  legitimately stopping  cluster agent hence     error
let  crash allowed types validated  request  unsupported typed shouldnt
   using  basefragment      place  another pr
indeed   see    needed   seems     added long time ago
fixed
 added  lenient strictness   test class  ignore unnecessarystubbingexceptions    case  mock stub  unnecessary
  reading things correctly  needs    little less strong eg looks  non cached start timestamps  batches  case someone
calls     transaction timestamp load batch limit uncached timestamps

          catch
 sure   mean  short circuit
formreset
    first
ditto
```

Tiempo de ejecución

Output - CleaningData (run)

```
run:
Inicio
374
Fin del programa
BUILD SUCCESSFUL (total time: 997 minutes 24 seconds)
```

Word Count

part-r-00000 Open with TextEdit

```
think   6276779
add     3705480
one     3440064
code    3241206
change  3157049
test    3117405
done    2713711
line    2637672
remove  2593754
instead 2554466
comment 2483788
sure    2478451
see     2475946
name    2385235
case    2344548
good    2319332
file    2284008
using   2283486
function        2253006
please  2218245
check   2179603
new     2012252
method  1953456
type    1903463
way     1878404
pr      1875514
right   1873839
better  1837596
error   1794193
return  1775045
value   1746298
suggestion      1715985
set     1684078
thanks  1659232
tests   1617600
seems   1613099
might   1601558
yes     1586034
probably        1545927
issue   1527789
call    1523985
added   1515435
fixed   1511188
default 1511107
still   1510712
already 1490402
```

Conclusión respecto a word count

1. Think
2. Add
3. One

Code

Change

Tenemos las palabras que más se repiten como Think, Add & One en los primeros tres puestos, podemos concluir que comentarios usando “I Think” es bastante común encontrarlos en los pull request.

< **ghtorrent-2019-01-07.csv** (16.4 GB)

etail Compact Column

		make me realize there was a small discrepancy between this docstri...
88	206420673	????????
42	226449691	This would be less confusing if the sections were \"Deployment Topology within a Single Region\" and...
779	250428232	did you accidentally leave this? TODO w/ issue if not?
178	258912689	i think this is the same instruction as above
74	198412164	I would omit even this line
53	209871618	Microservice **is** responsible
178	264274590	```suggestion\Displays highlighted message with warning icon.\```
6	205990106	i think we should keep the stack trace for the logs.\\(not blocking)
024	233532503	ok

Two Word Frequency

part-r-00000 — showing 128 MB of 220.3 MB

Open with TextEdit

```
issue , found   166551
codacy , issue  166184
good , catch    87167
good , point    83476
please , add    82677
add , comment   78762
line , long     78352
good , idea     76574
please , remove 49247
error , message 48088
sounds , good   45331
default , value 44168
think , better  43608
test , case     39753
return , value  37843
separate , pr   37621
found , avoid   37071
quoted , strings        34988
empty , line    34522
add , test      34425
trailing , whitespace   33102
comment , unexported    31913
blank , line    31543
metricslinelength , line        30985
unit , tests    30754
add , new       30125
unit , test     29902
commented , code        29815
one , line      28757
string , interpolation  28703
another , pr    28146
top , level     28118
stylestringliterals , prefer    27349
prefer , single 27292
single , quoted 26842
strings , string        26323
create , new    26280
see , comment   26099
think , good    25907
interpolation , special 25833
every , time    24971
looks , good    24702
take , look     24684
good , call     24383
either , way    24155
new , line      23783
better , way    22367
clang , format  22118
think , might   21691
might , better  21575
better , name   21460
remove , commented      21409
```

Part #1

Two Word Frequency

part-r-00000 — showing 128 MB of 220.3 MB

Open with TextEdit

```
issue , found   166930
codacy , issue  166558
good , catch    86944
good , point    83624
please , add    82579
line , long     78716
add , comment   78266
good , idea     76944
please , remove 49031
error , message 47399
sounds , good   45341
default , value 44234
think , better  43623
test , case     39698
return , value  37895
separate , pr   37543
found , avoid   37162
quoted , strings        34973
add , test      34602
empty , line    34522
trailing , whitespace   33335
comment , unexported    31787
blank , line    31022
metricslinelength , line        31016
unit , tests    30619
add , new       30395
unit , test     30013
commented , code        29585
one , line      28917
string , interpolation  28827
top , level     28274
another , pr    27863
prefer , single 27460
stylestringliterals , prefer    27344
single , quoted 27008
create , new    26472
strings , string        26446
see , comment   26362
think , good    26007
interpolation , special 25996
every , time    25212
take , look     24633
looks , good    24513
good , call     24416
new , line      23986
either , way    23455
clang , format  22597
better , way    22263
might , better  21774
think , might   21626
remove , commented      21360
```

Part #2

Two Word Frequency

part-r-00000 — showing 128 MB of 220.2 MB

Open with TextEdit

```
issue , found   165628
codacy , issue  165224
good , catch    86492
good , point    83072
please , add    82789
add , comment   78561
line , long     78255
good , idea     77125
please , remove 49105
error , message 47588
sounds , good   45089
default , value 43988
think , better  43929
test , case     39995
return , value  37937
separate , pr   37506
found , avoid   37032
quoted , strings        35360
add , test      34663
empty , line    34589
trailing , whitespace   33547
comment , unexported    32387
blank , line    31155
metricslinelength , line        30779
unit , tests    30568
add , new       30358
unit , test     29924
commented , code        29820
one , line      29235
string , interpolation  29103
another , pr    27950
top , level     27900
stylestringliterals , prefer    27767
prefer , single 27672
single , quoted 27238
strings , string        26790
create , new    26324
interpolation , special 26242
see , comment   26103
think , good    25891
every , time    25438
looks , good    24750
take , look     24666
good , call     24400
new , line      24056
either , way    23871
better , way    22614
clang , format  22510
might , better  22027
think , might   21518
better , name   21472
```

Part #3

Two Word Frequency

part-r-00000 — showing 128 MB of 182.9 MB

Open with TextEdit

```
issue , found   79028
codacy , issue  78829
good , catch    41890
good , point    40185
please , add    40012
add , comment   38080
line , long     37741
good , idea     36862
please , remove 23713
error , message 23162
sounds , good   21891
think , better  21254
default , value 21217
test , case     19438
return , value  18365
separate , pr   18061
found , avoid   17524
empty , line    16889
add , test      16776
quoted , strings        16478
trailing , whitespace   16317
comment , unexported    15839
blank , line    15143
unit , tests    14803
metricslinelength , line        14706
add , new       14597
unit , test     14347
commented , code        14087
one , line      13904
string , interpolation  13645
top , level     13606
another , pr    13500
prefer , single 12946
stylestringliterals , prefer    12891
single , quoted 12768
see , comment   12700
create , new    12598
strings , string        12507
think , good    12421
every , time    12314
interpolation , special 12279
take , look     12034
looks , good    11959
good , call     11866
new , line      11642
either , way    11589
clang , format  10902
better , way    10827
80 , characters 10562
might , better  10463
better , name   10407
think , might   10366
```

Part #4

Conclusión respecto a 2 word frequency

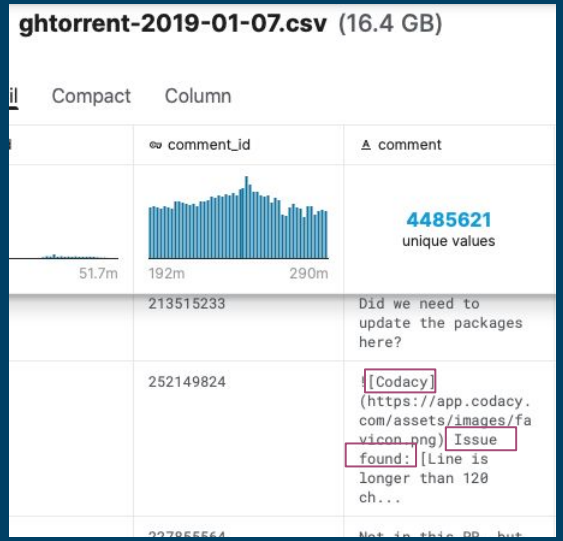
1. issue , found
2. codacy , issue
3. good , catch

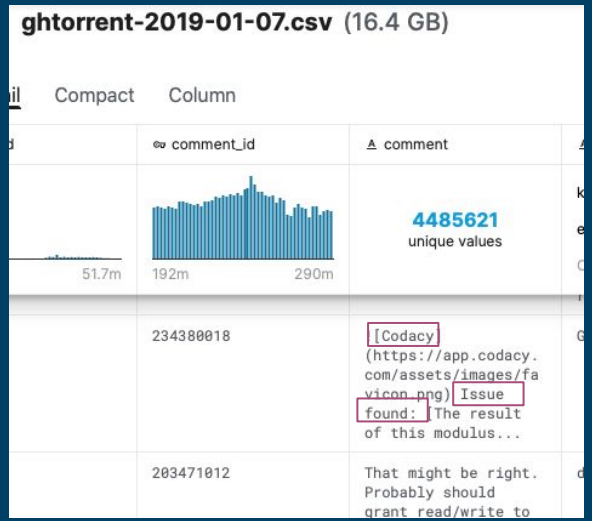
please , add

add , comment

Tenemos los conjuntos de palabras que más se repiten como issue found, codacy issue, good catch, basado en esta información concluimos que la resolución de issues es bastante frecuente en los pull request.

También las solicitudes de agregar comentarios o bien revision de codigo.

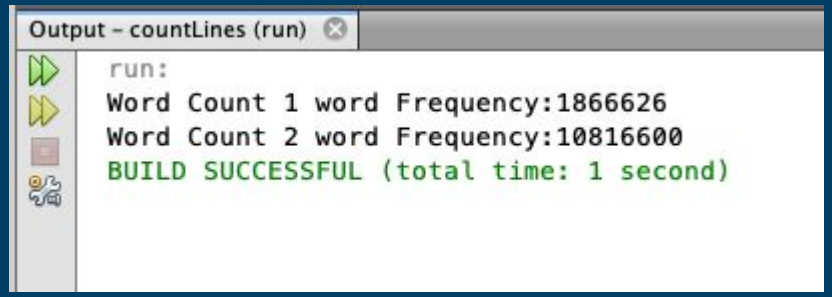




CODACY

Codacy es una herramiento que verifica la calidad del código y realiza seguimientos de su deuda técnica.

Frequency analyzer



Fuimos un poco más allá

Cuales fueron esos usuarios con mayor frecuencia de pull request?

Data sin limpiar

usernamePullRequest.txt — showing 128 MB of 771 MB

Open with TextEdit

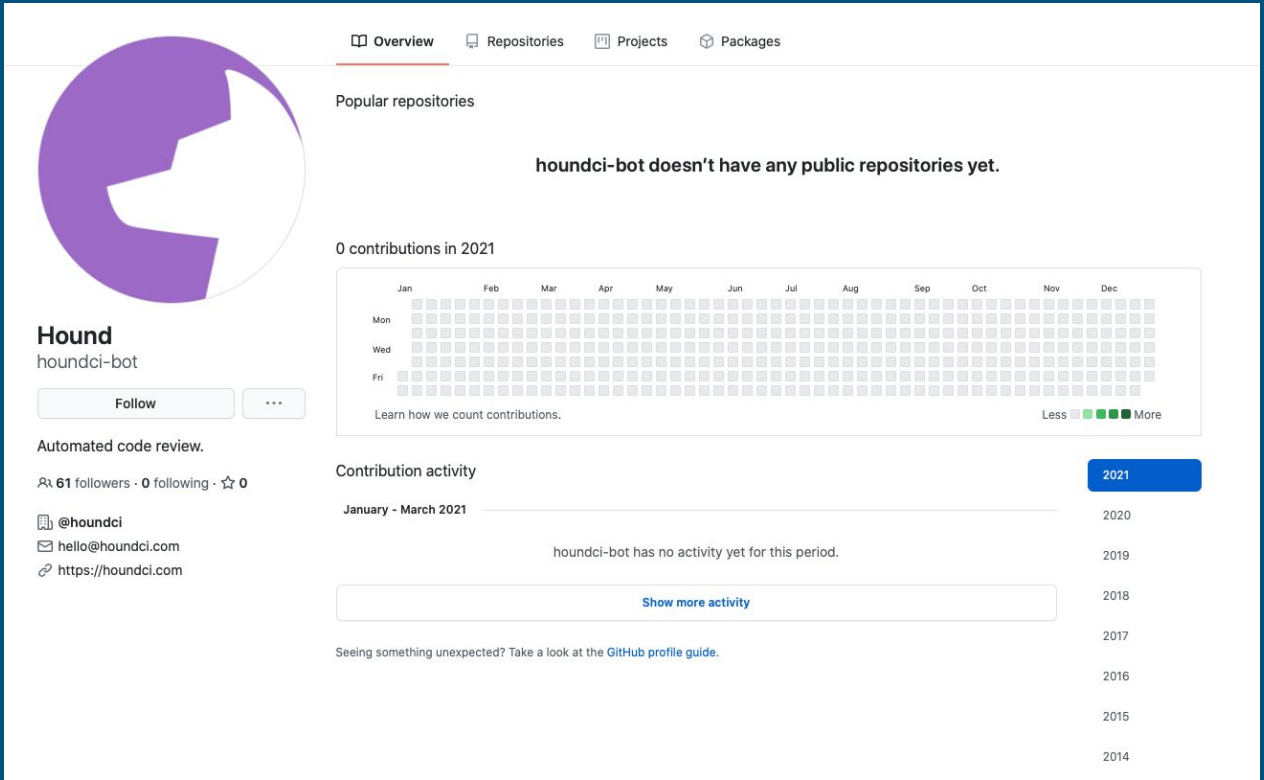
```
EleisonC
cooltey
felixfbecker
slimsag
alangenfeld
iHiD
keithwrightbos
lannka
eadpearce
andrewhop
virgil-serbanuta
anne-edwards
WVerlaek
Vatyx
joshuahuburn
pichfl
marcoow
houndci-bot
houndci-bot
JaphethCA
david-sautet
blinkov
KochetovNicolai
Azfletch
abulte
mfpierre
PavelVesnin
cquemin
TurboGit
blarz
BSaunders95
jeremyk-91
kylemh
wKich
juliantrueflynn
miyagit
alumi
athos
nazarhussain
mjingle
bmoore-msft
bmoore-msft
bmoore-msft
manishshar
sqren
EdwinGuzman
cqliu1
pon182
```

Username pull request frequency

part-r-00000

```
houndci-bot     1384907
codacy-bot      577009
egonelbre       311548
ywelsch 274487
mgorny  272067
daosbuild1      268466
ezyang  261099
philbucher      238835
MartinHjelmare  228083
aduth   214065
golangcibot     201360
DaveCTurner     198340
gziolo  186295
gchanan 176968
loumalouomega   171304
apaszke 168256
azasypkin       159868
CyrusNajmabadi  156867
nik9000 153045
stephentoub     150511
SirReal 150276
zdevito 146872
jaymode 137985
youknowriad     137811
chriseth        134178
tyxla   130215
pavel06081991   129819
jasontedor      128193
sqren   125026
william-gr      123906
bleskes 115242
rjernst 114923
Therzok 113988
cjcenizal       113320
Mic92   111322
arnikola        109956
jreback 109790
jrose-apple     107309
simison 102777
RiccardoRossi   99176
kaloyan-raev    98655
```

Top 1 & 2 son bots



1

El primer usuario humano

Egon Elbre
egonelbre

Follow ···

280 followers · **48** following · ☆ **443**

Estonia, Tartu
egonelbre@gmail.com
egonelbre.com

Highlights
Arctic Code Vault Contributor

Organizations

Overview | Repositories 71 | Projects | Packages

Pinned

gophers
Free gophers
Go ☆ 2.3k 109

spexs2
an exhaustive sequence pattern search tool
Go ☆ 43

exp
Experiments that do not fit into a separate repository.
Go ☆ 31 4

1,524 contributions in 2020

Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Mon Wed Fri
Learn how we count contributions. Less More

2021 2020 2019 2018 2017 2016 2015 2014

@storj @gioui @loov More

Activity overview

Contributed to **storj/storj**, **storj/uplink**, **storj/common** and 5 other repositories

11%
Code review

El segundo usuario humano

Overview | Repositories 13 | Projects | Packages

Yannick Welsch
ywelsch

Follow ···

27 followers · 2 following · 4

Elastic
Luxembourg
yannick@welsch.lu

Highlights

Arctic Code Vault Contributor

Pinned

elastic/elasticsearch
Free and Open, Distributed, RESTful Search Engine
Java 54.2k 19.5k

gradle/gradle
Adaptable, fast automation for all
Groovy 11.6k 3.4k

elastic/elasticsearch-formal-models
Formal models of core Elasticsearch algorithms
Isabelle 160 19

gradle-bash-tools
Shell 8 1

517 contributions in 2020

Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Mon Wed Fri

Learn how we count contributions. Less More

Contribution activity

November - December 2020

2021
2020
2019

ywelsch had no activity during this period.

El tercer usuario humano

Overview | Repositories | Projects | Packages

daosbuild1
daosbuild1

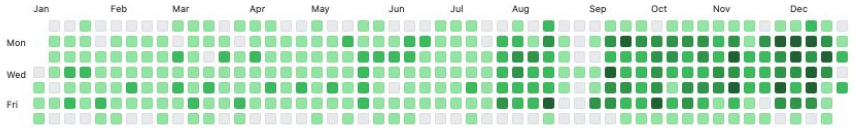
Follow ···

1 follower · 0 following · ☆ 2

Popular repositories

daosbuild1 doesn't have any public repositories yet.

1,392 contributions in 2020



Learn how we count contributions. Less More

Contribution activity

2021
2020
2019

December 2020

Reviewed 237 pull requests in 2 repositories

daos-stack/daos 236 pull requests

daos-stack/code_review 1 pull request

Show more activity

Seeing something unexpected? Take a look at the GitHub profile guide.

Ejecución del programa

1. Creamos la carpeta donde estará localizado el input en el HDFS (Hadoop Distributed File System):

   hadoop dfs -mkdir /WordCount/Input

2. Copiamos el input a la carpeta del HDFS:

   hadoop dfs -put /home/hdoop/WordCount/Input/cleanData0.txt

3. Ejecutamos Hadoop:

   hadoop jar /home/hdoop/WordCount/WCJAR.jar WordCount /WordCount/Input/cleanData0.txt /WordCount/Output

Hallazgos Interesantes

- La manera en que se debe de hacer uso de la memoria, de forma a que sea coherente con las necesidades de procesamiento que presenta Hadoop.

- Encontramos que la cantidad de data a ser procesada debe ser de acuerdo a la capacidad del sistema de archivos local.

- Para que un programa de mapeo & reducción tenga tiempos de ejecución eficiente, la proporción de la memoria que se le asigna al trabajo de reducción debe ser el doble que la del trabajo de mapeo.
    - La cantidad de memoria que se le asigna a heap de Java, es recomendable que sea de 1GB menos que lo que se le asignó al mapeo.

Conclusiones

Aprendimos a usar la aplicación **Map/Reduce** en un sistema de archivos distribuido de **Hadoop.**

Desarrollamos una aplicación capaz de procesar Millones de datos, por medio de **Map/Reduce & Hadoop**

Los datos como tal, deben ser analizados para tener sentido.

Bibliografía

- https://hadoop.apache.org
- https://blogs.solidq.com/es/business-analytics/que-es-mapreduce/
- Material de apoyo proporcionado por el docente

Anexo

Link del dataset

https://www.kaggle.com/stephangarland/ghtorrent-pull-requests

Gracias por su atencion!