Universidad de Granada

Cloud Computing: Servicios y Aplicaciones

Computación Distribuida y Escalable con Hadoop

Marvin Matías Agüero Torales

maguero@correo.ugr.es

Curso 2016-2017

Sumario

Objetivo

El objetivo de esta práctica es realizar programas escalables para mejorar la eficiencia en entornos Big Data. Para ello, haremos uso del entorno que se ha convertido en un estándar de facto como es Hadoop, utilizando HDFS como sistema de archivos distribuido y Hadoop-MapReduce como mecanismo de ejecución.

Enunciado

Para constatar el manejo de la herramienta anterior, el alumno deberá realizar las tareas que se describen a continuación y entregar documentación describiendo las tareas realizadas. Se recomienda seguir el tutorial asociado en la página https://github.com/manuparra/MasterDegreeCC_Practice/ (Parra, 5 de marzo de 2017/2017)

Tareas

Utilizando como base el conjunto de datos ECBDL14 situado en la carpeta */tmp/BDCC/datasets/ECBDL14/ECBDL14_10tst.data* obtener los siguientes datos estadísticos descriptivos («Ejercicios HADOOP: Implementación y análisis de funciones básicas sobre conjuntos de datos BigData», 2017):

1. Calcula el valor mínimo de la variable (columna) 5

   ```
   hadoop jar Stat.jar oldapi.Stat /tmp/BDCC/datasets/ECBDL14/ECBDL14_10tst.data ./Min/output/ Min 5
   ```

2. Calcula el valor máximo de la variable (columna) 5

   ```
   hadoop jar Stat.jar oldapi.Stat /tmp/BDCC/datasets/ECBDL14/ECBDL14_10tst.data ./Max/output/ Max 5
   ```

3. Calcula al mismo tiempo los valores máximo y mínimo de la variable 5

   ```
   hadoop jar Stat.jar oldapi.Stat /tmp/BDCC/datasets/ECBDL14/ECBDL14_10tst.data ./MaxMin/output/ MaxMin 5
   ```

4. Calcula los valores máximo y mínimo de todas las variables (salvo la última, que es la etiqueta de clase)

   ```
   hadoop jar Stat.jar oldapi.Stat /tmp/BDCC/datasets/ECBDL14/ECBDL14_10tst.data ./MaxMin/output/ MaxMin -1
   ```

5. Realizar la media de la variable 5

   ```
   hadoop jar Stat.jar oldapi.Stat /tmp/BDCC/datasets/ECBDL14/ECBDL14_10tst.data ./Avg/output/ Avg 5
   ```

6. Obtener la media de todas las variables (salvo la clase)

   ```
   hadoop jar Stat.jar oldapi.Stat /tmp/BDCC/datasets/ECBDL14/ECBDL14_10tst.data ./Avg/output/ Avg -1
   ```

7. Comprobar si el conjunto de datos ECBDL es balanceado o no balanceado, es decir, que el ratio entre las clases sea menor o mayor que 1.5 respectivamente.

   Se puede ver que el ratio es mucho mayor (58.58 ...) a 1.5, entonces podemos concluir que el conjunto de datos no es balanceado.

   ```
   hadoop jar Stat.jar oldapi.Stat /tmp/BDCC/datasets/ECBDL14/ECBDL14_10tst.data
   ./Bal/output/ Bal 10
   ```

8. Cálculo del coeficiente de correlación entre todas las parejas de variables

   ```
   hadoop jar Stat.jar oldapi.Stat /tmp/BDCC/datasets/ECBDL14/ECBDL14_10tst.data
   ./Corr/output/ Corr -1
   ```

Adicionales

9. Cálculo de los estadísticos descriptivos.

   Nuestro objetivo ahora es actualizar el código para realizar las siguientes tareas:

   9.1. Parametrizar la columna sobre la que se quiere calcular el estadístico

   Se realizó desde la Tarea 1, parametrizando, el tipo de operaciones además del número de columna, como se puede ver en las tareas anteriores.

   9.2. Combinar el cálculo de todos los estadísticos en una única función

   ```
   hadoop jar Stat.jar oldapi.Stat /tmp/BDCC/datasets/ECBDL14/ECBDL14_10tst.data
   ./Stats/output/ Stats 4
   ```

   9.3. Calcular los estadísticos sobre todas las columnas

   ```
   hadoop jar Stat.jar oldapi.Stat /tmp/BDCC/datasets/ECBDL14/ECBDL14_10tst.data
   ./Stats/output/ Stats -1
   ```

   9.4. Repite el proceso sobre un conjunto de mayor volumen (Ej: /user/isaac/datasets/higgs...” ¿Hay grandes diferencias de tiempo?

   El conjunto de datos ECBDL14 cuenta con alrededor de 3.000.000 de filas, y se obtienen unos tiempos levemente mayores que para el conjunto de datos de HiggsImg10, compuesto aproximadamente de 500.000 de filas (como se puede ver en los Resultados). Qué la diferencia sea tan leve, nos hace ver que el tiempo no es directamente proporcional al tamaño, depende también del número de mappers o reducers, como de columnas.

   ```
   hadoop jar Stat.jar oldapi.Stat /user/isaac/datasets/higgsImb10-5-fold/higgsImb10.data
   ./Stats/compare/ Stats -1
   ```

   9.5. Acelera el proceso de cómputo descargando al Reducer de parte de la tarea.

   Se comprueba que el tiempo para el cálculo del mínimo en el código original es mucho mayor que con el uso de cleanup (ver en Resultados).

   ```
   hadoop jar StatCleanup.jar oldapi.StatCleanup
   /tmp/BDCC/datasets/ECBDL14/ECBDL14_10tst.data ./StatCleanup/output/
   ```

Resultados

```
1 [mcc4423998@hadoop-master stat]$ hdfs dfs -cat Stat/output/*
  Col5        -11.0
```

```
2 [mcc4423998@hadoop-master stat]$ hdfs dfs -cat Max/output/*
  Col5        9.0
```

```
3 [mcc4423998@hadoop-master stat]$ hdfs dfs -cat MaxMin/output/*
  Max-Col5  9.0
  Min-Col5  -11.0
```

```
4 [mcc4423998@hadoop-master stat]$ hdfs dfs -cat MaxMin/output/*
  Max-Col0  0.768
  Min-Col0  0.094
  Max-Col1  0.154
  Min-Col1  0.0
  Max-Col2  10.0
  Min-Col2  -12.0
  Max-Col3  8.0
  Min-Col3  -11.0
  Max-Col4  9.0
  Min-Col4  -12.0
  Max-Col5  9.0
  Min-Col5  -11.0
  Max-Col6  9.0
  Min-Col6  -13.0
  Max-Col7  9.0
  Min-Col7  -12.0
  Max-Col8  7.0
  Min-Col8  -12.0
  Max-Col9  10.0
  Min-Col9  -13.0
```

```
5 [mcc4423998@hadoop-master stat]$ hdfs dfs -cat Avg/output/*
  Col5        -1.282261707288373
```

```
6 [mcc4423998@hadoop-master stat]$ hdfs dfs -cat Avg/output/*
  Col0        0.25496195991787296
  Col1        0.05212776590953057
  Col2        -2.188240380935686
  Col3        -1.408876789776933
  Col4        -1.7528724942777865
  Col5        -1.282261707288373
  Col6        -2.293434905140485
  Col7        -1.5875789403216172
  Col8        -1.7390052924221087
  Col9        -1.6989002790625127
```

```
7 [mcc4423998@hadoop-master stat]$ hdfs dfs -cat Bal/output/*
  Col5        58.582560602010815
```

```
8 [mcc4423998@hadoop-master stat]$ hdfs dfs -cat Corr/output/*
  0, 3        0.07005931837274204
  1, 4        0.058856701859578545
  2, 5        0.024182999250758484
  3, 6        0.025952003813569456
  4, 7        0.01984291578033614
  5, 8        0.015183324110128226
  6, 9        0.1071360896407867
  0, 4        0.04742917822713238
  1, 5        0.014659977642218205
  2, 6        0.041153841377462724
  3, 7        0.01879122854336587
  4, 8        0.01224584385595619
  5, 9        0.023068393377281653
  0, 5        0.12916572715633357
  1, 6        -0.03183255332422876
  2, 7        0.03814283037771738
  3, 8        0.016130402799924542
  4, 9        0.014041854998880898
  0, 6        0.19252517589227605
  1, 7        -1.7503662130016114E-5
  2, 8        0.025077384911599235
  3, 9        0.01817123896585364
  0, 7        0.1792126656307003
  1, 8        0.015894103465096773
  2, 9        0.027549270387458427
  0, 8        0.06624560107321993
  1, 9        -0.0167306234595493
```

```
0, 9        0.1382708996670605
0, 1        -0.13589916868840649
1, 2        -0.003036453944885367
2, 3        -0.01726247486762999
3, 4        0.015754379166559307
4, 5        0.07125079800784533
5, 6        0.03200113594875155
6, 7        0.11488805268078417
7, 8        -0.3292179447994215
8, 9        0.1084960047958963
0, 2        0.09143593110662
1, 3        0.009438349446753204
2, 4        0.018191261366109063
3, 5        0.016128930425374947
4, 6        0.018264386288745375
5, 7        0.03297998768398484
6, 8        0.07783431570283235
7, 9        0.08936167755929571
```

9.1 Se realizó desde la Tarea 1

9.2

```
[mcc4423998@hadoop-master stat]$ hdfs dfs -cat Stats/output/*
Max-Col4  9.0
Min-Col4  -12.0
Avg-Col4  -1.7528724942777865
```

9.3

```
[mcc4423998@hadoop-master stat]$ hdfs dfs -cat Stats/output/*
Max-Col0  0.768
Min-Col0  0.094
Avg-Col0  0.25496195991782294
Max-Col1  0.154
Min-Col1  0.0
Avg-Col1  0.05212776590922098
Max-Col2  10.0
Min-Col2  -12.0
Avg-Col2  -2.188240380935686
Max-Col3  8.0
Min-Col3  -11.0
Avg-Col3  -1.408876789776933
Max-Col4  9.0
Min-Col4  -12.0
Avg-Col4  -1.7528724942777865
Max-Col5  9.0
Min-Col5  -11.0
Avg-Col5  -1.282261707288373
Max-Col6  9.0
Min-Col6  -13.0
Avg-Col6  -2.293434905140485
Max-Col7  9.0
Min-Col7  -12.0
Avg-Col7  -1.5875789403216172
Max-Col8  7.0
Min-Col8  -12.0
Avg-Col8  -1.7390052924221087
Max-Col9  10.0
Min-Col9  -13.0
Avg-Col9  -1.6989002790625127
```

9.4

```
[mcc4423998@hadoop-master stat]$ hdfs dfs -cat Stats/compare/*
Max-Col12         2.214872121810913
Min-Col120.0
Avg-Col121.0535079698182337
Max-Col23         6.010560989379883
Min-Col230.14747904241085052
Avg-Col231.0485020424771379
Max-Col0  11.673967361450195
Min-Col0  0.2746966481208801
Avg-Col0  1.0203453246822671
Max-Col13         9.598233222961426
Min-Col130.26360762119293213
Avg-Col130.9844488157520382
Max-Col24         12.891449928283691
Min-Col240.28217247128486633
Avg-Col241.0236238593077998
Max-Col1  2.4348678588867188
Min-Col1  -2.434976100921631
Avg-Col1  -0.0025247084598987494
Max-Col14         2.730008840560913
Min-Col14-2.7296628952026367
Avg-Col144.665200984794381E-4
```

```
Max-Col25	17.76285171508789
Min-Col25	0.05431479960680008
Avg-Col25	1.0553339343831032
Max-Col15	1.7428839206695557
Min-Col15	-1.7420687675476074
Avg-Col15	-4.820412042296641E-4
Max-Col2	1.7432359457015991
Min-Col2	-1.7425082921981812
Avg-Col2	-5.103753660642641E-4
Max-Col26	8.657637596130371
Min-Col26	0.34091895818710327
Avg-Col26	1.0589104783534258
Max-Col16	2.548224449157715
Min-Col16	0.0
Avg-Col16	1.0275992439451727
Max-Col27	6.482466697692871
Min-Col27	0.38276857137680054
Avg-Col27	1.0005302882229659
Max-Col3	9.579188346862793
Min-Col3	8.573559462092817E-4
Avg-Col3	1.0594176750387583
Max-Col17	11.418229103088379
Min-Col17	0.36535415053367615
Avg-Col17	0.967026475575443
Max-Col4	1.7432570457458496
Min-Col4	-1.7439435720443726
Avg-Col4	0.0013947659926604294
Max-Col18	2.498008966445923
Min-Col18	-2.497264862060547
Avg-Col18	0.0019268480147912228
Max-Col5	8.641400337219238
Min-Col5	0.1375940442085266
Avg-Col5	0.9643217926883103
Max-Col19	1.7433723211288452
Min-Col19	-1.7426908016204834
Avg-Col19	0.0011555539273784253
Max-Col6	2.9696741104125977
Min-Col6	-2.9697251319885254
Avg-Col6	-9.416839608677337E-4
Max-Col7	1.741453766822815
Min-Col7	-1.7412374019622803
Avg-Col7	-0.0012207837065991904
Max-Col8	2.1730761528015137
Min-Col8	0.0
Avg-Col8	1.0133794418470266
Max-Col9	11.64708137512207
Min-Col9	0.18898114562034607
Avg-Col9	0.9812130859741491
Max-Col20	3.101961374282837
Min-Col20	0.0
Avg-Col20	0.9810118929382996
Max-Col10	2.913209915161133
Min-Col10	-2.9130895137786865
Avg-Col10	6.822083683996807E-4
Max-Col21	26.093395233154297
Min-Col21	0.09788843244314194
Avg-Col21	1.0264101705980098
Max-Col11	1.7431747913360596
Min-Col11	-1.742371678352356
Avg-Col11	0.002825531704890823
Max-Col22	13.304651260375977
Min-Col22	0.2818564474582672
Avg-Col22	1.0154354222187747
```

[mcc4423998@hadoop-master stat]$ hadoop jar Stat.jar oldapi.Stat /tmp/BDCC/datasets/ECBDL14/ECBDL14_10tst.data ./Stats/output/ Stats -1	[mcc4423998@hadoop-master stat]$ hadoop jar Stat.jar oldapi.Stat /user/isaac/datasets/higgsImb10-5-fold/higgsImb10.data ./Stats/compare/ Stats -1
17/05/21 16:45:22 INFO client.RMProxy: Connecting to ResourceManager at hadoop-master/192.168.10.1:8032 17/05/21 16:45:22 INFO client.RMProxy: Connecting to ResourceManager at hadoop-master/192.168.10.1:8032 17/05/21 16:45:23 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this. 17/05/21 16:45:23 INFO mapred.FileInputFormat: Total input paths to process : 1 17/05/21 16:45:23 INFO mapreduce.JobSubmitter: number of splits:2	17/05/21 17:06:36 INFO client.RMProxy: Connecting to ResourceManager at hadoop-master/192.168.10.1:8032 17/05/21 17:06:36 INFO client.RMProxy: Connecting to ResourceManager at hadoop-master/192.168.10.1:8032 17/05/21 17:06:36 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this. 17/05/21 17:06:36 INFO mapred.FileInputFormat: Total input paths to process : 1 17/05/21 17:06:36 INFO mapreduce.JobSubmitter: number of splits:3

```
17/05/21 16:45:24 INFO mapreduce.JobSubmitter: Submitting tokens
for job: job_1494408081774_0418
17/05/21 16:45:24 INFO impl.YarnClientImpl: Submitted application
application_1494408081774_0418
17/05/21 16:45:24 INFO mapreduce.Job: The url to track the job:
http://hadoop.ugr.es:8088/proxy/application_1494408081774_0418/
17/05/21 16:45:24 INFO mapreduce.Job: Running job:
job_1494408081774_0418
17/05/21 16:45:28 INFO mapreduce.Job: Job
job_1494408081774_0418 running in uber mode : false
17/05/21 16:45:28 INFO mapreduce.Job:  map 0% reduce 0%
17/05/21 16:45:38 INFO mapreduce.Job:  map 30% reduce 0%
17/05/21 16:45:41 INFO mapreduce.Job:  map 43% reduce 0%
17/05/21 16:45:44 INFO mapreduce.Job:  map 56% reduce 0%
17/05/21 16:45:47 INFO mapreduce.Job:  map 67% reduce 0%
17/05/21 16:45:50 INFO mapreduce.Job:  map 73% reduce 0%
17/05/21 16:45:53 INFO mapreduce.Job:  map 79% reduce 0%
17/05/21 16:45:56 INFO mapreduce.Job:  map 85% reduce 0%
17/05/21 16:45:59 INFO mapreduce.Job:  map 100% reduce 0%
17/05/21 16:46:03 INFO mapreduce.Job:  map 100% reduce 25%
17/05/21 16:46:06 INFO mapreduce.Job:  map 100% reduce 75%
17/05/21 16:46:07 INFO mapreduce.Job:  map 100% reduce 100%
17/05/21 16:46:08 INFO mapreduce.Job: Job
job_1494408081774_0418 completed successfully
17/05/21 16:46:08 INFO mapreduce.Job: Counters: 49
		File System Counters
				FILE: Number of bytes read=47832813
				FILE: Number of bytes written=73505433
				FILE: Number of read operations=0
				FILE: Number of large read operations=0
				FILE: Number of write operations=0
				HDFS: Number of bytes read=102749934
				HDFS: Number of bytes written=570
				HDFS: Number of read operations=54
				HDFS: Number of large read operations=0
				HDFS: Number of write operations=32
		Job Counters
				Launched map tasks=2
				Launched reduce tasks=16
				Rack-local map tasks=2

				Total time spent by all maps in occupied slots
(ms)=401534
				Total time spent by all reduces in occupied
slots (ms)=3136049
				Total time spent by all map tasks (ms)=57362
				Total time spent by all reduce tasks
(ms)=64001
				Total vcore-seconds taken by all map
tasks=57362
				Total vcore-seconds taken by all reduce
tasks=64001
				Total megabyte-seconds taken by all map
tasks=401534000
				Total megabyte-seconds taken by all reduce
tasks=3200050000
		Map-Reduce Framework
				Map input records=2897917
				Map output records=28979170
				Map output bytes=376729210
				Map output materialized bytes=23773475
				Input split bytes=234
				Combine input records=0
				Combine output records=0
				Reduce input groups=10
				Reduce shuffle bytes=23773475
				Reduce input records=28979170
				Reduce output records=30
				Spilled Records=86937510
				Shuffled Maps =32
				Failed Shuffles=0
				Merged Map outputs=32
				GC time elapsed (ms)=366
				CPU time spent (ms)=138580
				Physical memory (bytes)
snapshot=10551685120
				Virtual memory (bytes)
```

```
17/05/21 17:06:37 INFO mapreduce.JobSubmitter: Submitting tokens
for job: job_1494408081774_0424
17/05/21 17:06:37 INFO impl.YarnClientImpl: Submitted application
application_1494408081774_0424
17/05/21 17:06:37 INFO mapreduce.Job: The url to track the job:
http://hadoop.ugr.es:8088/proxy/application_1494408081774_0424/
17/05/21 17:06:37 INFO mapreduce.Job: Running job:
job_1494408081774_0424
17/05/21 17:06:42 INFO mapreduce.Job: Job
job_1494408081774_0424 running in uber mode : false
17/05/21 17:06:42 INFO mapreduce.Job:  map 0% reduce 0%
17/05/21 17:06:51 INFO mapreduce.Job:  map 15% reduce 0%
17/05/21 17:06:52 INFO mapreduce.Job:  map 48% reduce 0%
17/05/21 17:06:55 INFO mapreduce.Job:  map 68% reduce 0%
17/05/21 17:06:58 INFO mapreduce.Job:  map 83% reduce 0%
17/05/21 17:07:01 INFO mapreduce.Job:  map 100% reduce 0%
17/05/21 17:07:03 INFO mapreduce.Job:  map 100% reduce 6%
17/05/21 17:07:04 INFO mapreduce.Job:  map 100% reduce 38%
17/05/21 17:07:05 INFO mapreduce.Job:  map 100% reduce 81%
17/05/21 17:07:08 INFO mapreduce.Job:  map 100% reduce 94%
17/05/21 17:07:09 INFO mapreduce.Job:  map 100% reduce 100%
17/05/21 17:07:10 INFO mapreduce.Job: Job

job_1494408081774_0424 completed successfully
17/05/21 17:07:10 INFO mapreduce.Job: Counters: 50
		File System Counters
				FILE: Number of bytes read=107574956
				FILE: Number of bytes written=163358420
				FILE: Number of read operations=0
				FILE: Number of large read operations=0
				FILE: Number of write operations=0
				HDFS: Number of bytes read=377165903
				HDFS: Number of bytes written=2381
				HDFS: Number of read operations=57
				HDFS: Number of large read operations=0
				HDFS: Number of write operations=32
		Job Counters
				Launched map tasks=3
				Launched reduce tasks=16
				Data-local map tasks=2
				Rack-local map tasks=1
				Total time spent by all maps in occupied slots
(ms)=340480
				Total time spent by all reduces in occupied
slots (ms)=3013304
				Total time spent by all map tasks (ms)=48640
				Total time spent by all reduce tasks
(ms)=61496
				Total vcore-seconds taken by all map
tasks=48640
				Total vcore-seconds taken by all reduce
tasks=61496
				Total megabyte-seconds taken by all map
tasks=340480000
				Total megabyte-seconds taken by all reduce
tasks=3074800000
		Map-Reduce Framework
				Map input records=527863
				Map output records=14780164
				Map output bytes=201643666
				Map output materialized bytes=53677531
				Input split bytes=378
				Combine input records=0
				Combine output records=0
				Reduce input groups=28
				Reduce shuffle bytes=53677531
				Reduce input records=14780164
				Reduce output records=84
				Spilled Records=44340492
				Shuffled Maps =48
				Failed Shuffles=0
				Merged Map outputs=48
				GC time elapsed (ms)=515
				CPU time spent (ms)=144200
				Physical memory (bytes)
snapshot=13047373824
				Virtual memory (bytes)
```

```
snapshot=984100139008
                Total committed heap usage
(bytes)=21234188288
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=102749700
        File Output Format Counters
                Bytes Written=570
```

```
snapshot=991361331200
                Total committed heap usage
(bytes)=24056430592
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=377165525
        File Output Format Counters
                Bytes Written=2381
```

9.5

```
[mcc4423998@hadoop-master statCleanup]$ hdfs dfs -cat StatCleanup/output/*
Max        9.0
```

```
[mcc4423998@hadoop-master stat]$ hadoop jar Stat.jar oldapi.Stat
/tmp/BDCC/datasets/ECBDL14/ECBDL14_10tst.data ./Max/output/
Max 5
```

```
[mcc4423998@hadoop-master statCleanup]$ hadoop jar
StatCleanup.jar oldapi.StatCleanup
/tmp/BDCC/datasets/ECBDL14/ECBDL14_10tst.data
./StatCleanup/output/
```

```
17/05/21 21:12:51 INFO client.RMProxy: Connecting to
ResourceManager at hadoop-master/192.168.10.1:8032
17/05/21 21:12:51 INFO client.RMProxy: Connecting to
ResourceManager at hadoop-master/192.168.10.1:8032
17/05/21 21:12:52 WARN mapreduce.JobResourceUploader: Hadoop
command-line option parsing not performed. Implement the Tool
interface and execute your application with ToolRunner to remedy
this.
17/05/21 21:12:52 INFO mapred.FileInputFormat: Total input paths
to process : 1
17/05/21 21:12:52 INFO mapreduce.JobSubmitter: number of splits:2
17/05/21 21:12:52 INFO mapreduce.JobSubmitter: Submitting tokens
for job: job_1494408081774_0451
17/05/21 21:12:52 INFO impl.YarnClientImpl: Submitted application
application_1494408081774_0451
17/05/21 21:12:52 INFO mapreduce.Job: The url to track the job:
http://hadoop.ugr.es:8088/proxy/application_1494408081774_0451/
17/05/21 21:12:52 INFO mapreduce.Job: Running job:
job_1494408081774_0451
17/05/21 21:12:57 INFO mapreduce.Job: Job
job_1494408081774_0451 running in uber mode : false
17/05/21 21:12:57 INFO mapreduce.Job:  map 0% reduce 0%
17/05/21 21:13:07 INFO mapreduce.Job:  map 100% reduce 0%
17/05/21 21:13:12 INFO mapreduce.Job:  map 100% reduce 75%
17/05/21 21:13:13 INFO mapreduce.Job:  map 100% reduce 81%
17/05/21 21:13:15 INFO mapreduce.Job:  map 100% reduce 88%
17/05/21 21:13:16 INFO mapreduce.Job:  map 100% reduce 100%
17/05/21 21:13:17 INFO mapreduce.Job: Job
job_1494408081774_0451 completed successfully
17/05/21 21:13:17 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=2235110
                FILE: Number of bytes written=6656356
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=102749934
                HDFS: Number of bytes written=9
                HDFS: Number of read operations=54
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=32
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=16
                Rack-local map tasks=2
                Total time spent by all maps in occupied slots
(ms)=114716
                Total time spent by all reduces in occupied
slots (ms)=1873123
                Total time spent by all map tasks (ms)=16388
                Total time spent by all reduce tasks
(ms)=38227
                Total vcore-seconds taken by all map
tasks=16388
                Total vcore-seconds taken by all reduce
tasks=38227
```

```
17/05/21 21:06:13 INFO client.RMProxy: Connecting to
ResourceManager at hadoop-master/192.168.10.1:8032
17/05/21 21:06:13 WARN mapreduce.JobResourceUploader: Hadoop
command-line option parsing not performed. Implement the Tool
interface and execute your application with ToolRunner to remedy
this.


17/05/21 21:06:13 INFO input.FileInputFormat: Total input paths to
process : 1
17/05/21 21:06:14 INFO mapreduce.JobSubmitter: number of splits:1
17/05/21 21:06:14 INFO mapreduce.JobSubmitter: Submitting tokens
for job: job_1494408081774_0449
17/05/21 21:06:14 INFO impl.YarnClientImpl: Submitted application
application_1494408081774_0449
17/05/21 21:06:14 INFO mapreduce.Job: The url to track the job:
http://hadoop.ugr.es:8088/proxy/application_1494408081774_0449/
17/05/21 21:06:14 INFO mapreduce.Job: Running job:
job_1494408081774_0449
17/05/21 21:06:18 INFO mapreduce.Job: Job
job_1494408081774_0449 running in uber mode : false
17/05/21 21:06:18 INFO mapreduce.Job:  map 0% reduce 0%
17/05/21 21:06:26 INFO mapreduce.Job:  map 100% reduce 0%
17/05/21 21:06:30 INFO mapreduce.Job:  map 100% reduce 81%
17/05/21 21:06:31 INFO mapreduce.Job:  map 100% reduce 88%
17/05/21 21:06:34 INFO mapreduce.Job:  map 100% reduce 100%
17/05/21 21:06:35 INFO mapreduce.Job: Job
job_1494408081774_0449 completed successfully

17/05/21 21:06:35 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=746
                FILE: Number of bytes written=2064084
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=102747274
                HDFS: Number of bytes written=8
                HDFS: Number of read operations=51
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=32
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=16
                Rack-local map tasks=1
                Total time spent by all maps in occupied slots
(ms)=38206
                Total time spent by all reduces in occupied
slots (ms)=1746703
                Total time spent by all map tasks (ms)=5458
                Total time spent by all reduce tasks
(ms)=35647
                Total vcore-seconds taken by all map
tasks=5458
                Total vcore-seconds taken by all reduce
tasks=35647
```

Total megabyte-seconds taken by all map tasks=114716000 Total megabyte-seconds taken by all reduce tasks=1911350000 Map-Reduce Framework Map input records=2897917 Map output records=2897917 Map output bytes=37672921 Map output materialized bytes=2235300 Input split bytes=234 Combine input records=0 Combine output records=0 Reduce input groups=1 Reduce shuffle bytes=2235300 Reduce input records=2897917 Reduce output records=1 Spilled Records=5795834 Shuffled Maps =32 Failed Shuffles=0 Merged Map outputs=32 GC time elapsed (ms)=365 CPU time spent (ms)=38500 Physical memory (bytes) snapshot=8006270976 Virtual memory (bytes) snapshot=984105091072 Total committed heap usage (bytes)=18727043072 Shuffle Errors BAD_ID=0 CONNECTION=0 IO_ERROR=0 WRONG_LENGTH=0 WRONG_MAP=0 WRONG_REDUCE=0 File Input Format Counters Bytes Read=102749700 File Output Format Counters Bytes Written=9	Total megabyte-seconds taken by all map tasks=38206000 Total megabyte-seconds taken by all reduce tasks=1782350000 Map-Reduce Framework Map input records=2897917 Map output records=1 Map output bytes=12 Map output materialized bytes=426 Input split bytes=130 Combine input records=0 Combine output records=0 Reduce input groups=1 Reduce shuffle bytes=426 Reduce input records=1 Reduce output records=1 Spilled Records=2 Shuffled Maps =16 Failed Shuffles=0 Merged Map outputs=16 GC time elapsed (ms)=438 CPU time spent (ms)=14580 Physical memory (bytes) snapshot=7202598912 Virtual memory (bytes) snapshot=976933064704 Total committed heap usage (bytes)=18374721536 Shuffle Errors BAD_ID=0 CONNECTION=0 IO_ERROR=0 WRONG_LENGTH=0 WRONG_MAP=0 WRONG_REDUCE=0 File Input Format Counters Bytes Read=102747144 File Output Format Counters Bytes Written=8

Bibliografía consultada

Ejercicios HADOOP: Implementación y análisis de funciones básicas sobre conjuntos de datos BigData. (2017). Universidad de Granada.

Parra, M. (2017). *MasterDegreeCC_Practice: Taller del Máster Profesional de Informática UGR. Curso de CloudComputing*. Recuperado a partir de https://github.com/manuparra/MasterDegreeCC_Practice (Original work published 5 de marzo de 2017)

Anexos

Disponibles en https://github.com/mmaguero

Adjuntos

Se adjuntan el código fuente en Java para llevar a cabo las tareas propuestas

- README.P4 → Archivo con instrucciones en Bash para llevar a cabo la ejecución de las tareas.
- src

- stat → Código fuente con el main principal que valida los parámetros e invoca a los mappers o reducers requeridos para las operaciones. Todas las operaciones son validadas y ejecutadas desde un solo .jar totalmente parametrizable.
- statCleanup → Código fuente que utiliza un enfoque cleanup para calcular el máximo de la columna 5.