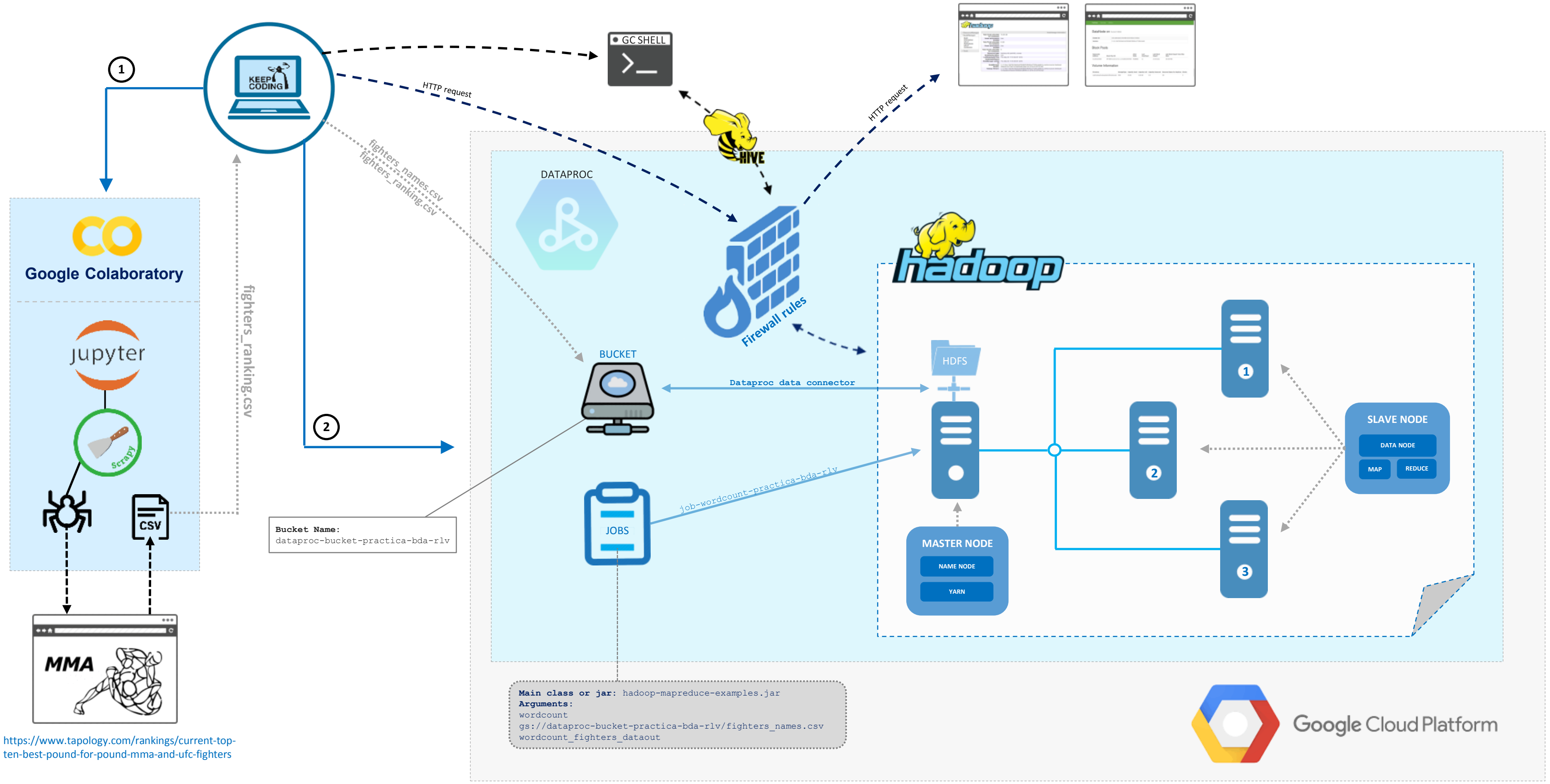
Práctica Big Data Architecture

Ramón Lerena Villarroel

II Big Data Bootcamp

Sprint 1 - Flujo de datos y herramientas utilizadas en la elaboración de la práctica.

Big Data Architecture
Ramón Lerena Villarroel



```
Bucket Name:
dataproc-bucket-practica-bda-rlv
```

```
Main class or jar: hadoop-mapreduce-examples.jar
Arguments:
wordcount
gs://dataproc-bucket-practica-bda-rlv/fighters_names.csv
wordcount_fighters_dataout
```

https://www.tapology.com/rankings/current-top-ten-best-pound-for-pound-mma-and-ufc-fighters

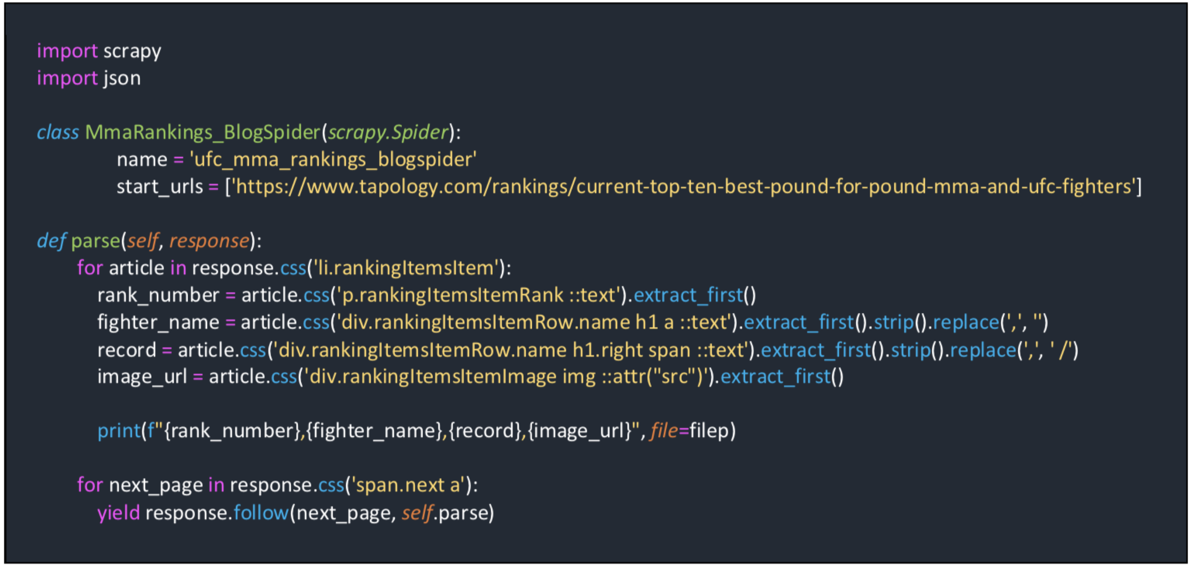
Sprint 2 - Creación de un crawler con scrapy en Google Colaboratory para descargar el ranking de luchadores libra por libra de MMA.

Origen de los datos → https://www.tapology.com/rankings/current-top-ten-best-pound-for-pound-mma-and-ufc-fighters

Reseñar que el contenido del fichero robots.txt no hacía referencia a no permitir hacer crawling de los datos.

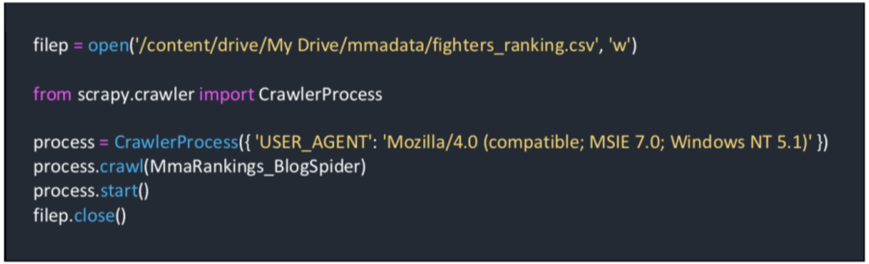
```
# See http://www.robotstxt.org/wc/norobots.html for documentation on how to use the robots.txt file
#
# To ban all spiders from the entire site uncomment the next two lines:
# User-Agent: *
# Disallow: /
```

- Clase MmaRankings_BlogSpider que recorre la web con el origen de los datos para obtener la información deseada



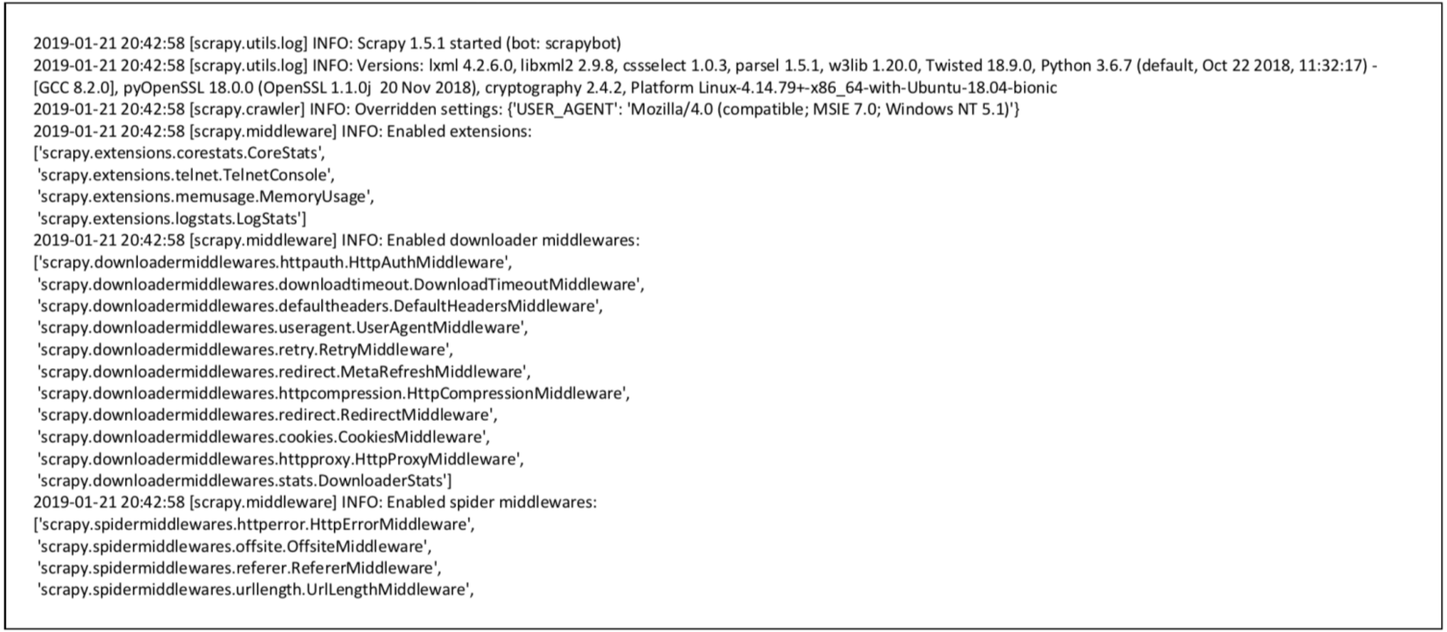
```
import scrapy
import json

class MmaRankings_BlogSpider(scrapy.Spider):
        name = 'ufc_mma_rankings_blogspider'
        start_urls = ['https://www.tapology.com/rankings/current-top-ten-best-pound-for-pound-mma-and-ufc-fighters']

def parse(self, response):
    for article in response.css('li.rankingItemsItem'):
      rank_number = article.css('p.rankingItemsItemRank ::text').extract_first()
      fighter_name = article.css('div.rankingItemsItemRow.name h1 a ::text').extract_first().strip().replace(',', '')
      record = article.css('div.rankingItemsItemRow.name h1.right span ::text').extract_first().strip().replace(',', ' /')
      image_url = article.css('div.rankingItemsItemImage img ::attr("src")').extract_first()

      print(f"{rank_number},{fighter_name},{record},{image_url}", file=filep)

    for next_page in response.css('span.next a'):
      yield response.follow(next_page, self.parse)
```

- Ejecución del crawler con la clase definida



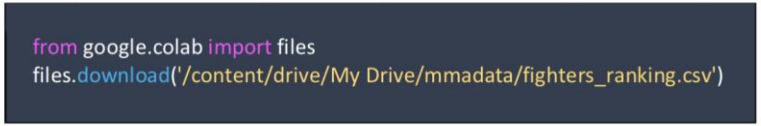
```
filep = open('/content/drive/My Drive/mmadata/fighters_ranking.csv', 'w')

from scrapy.crawler import CrawlerProcess

process = CrawlerProcess({ 'USER_AGENT': 'Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1)' })
process.crawl(MmaRankings_BlogSpider)
process.start()
filep.close()
```

- Log de la ejecución en Google Colaboratory



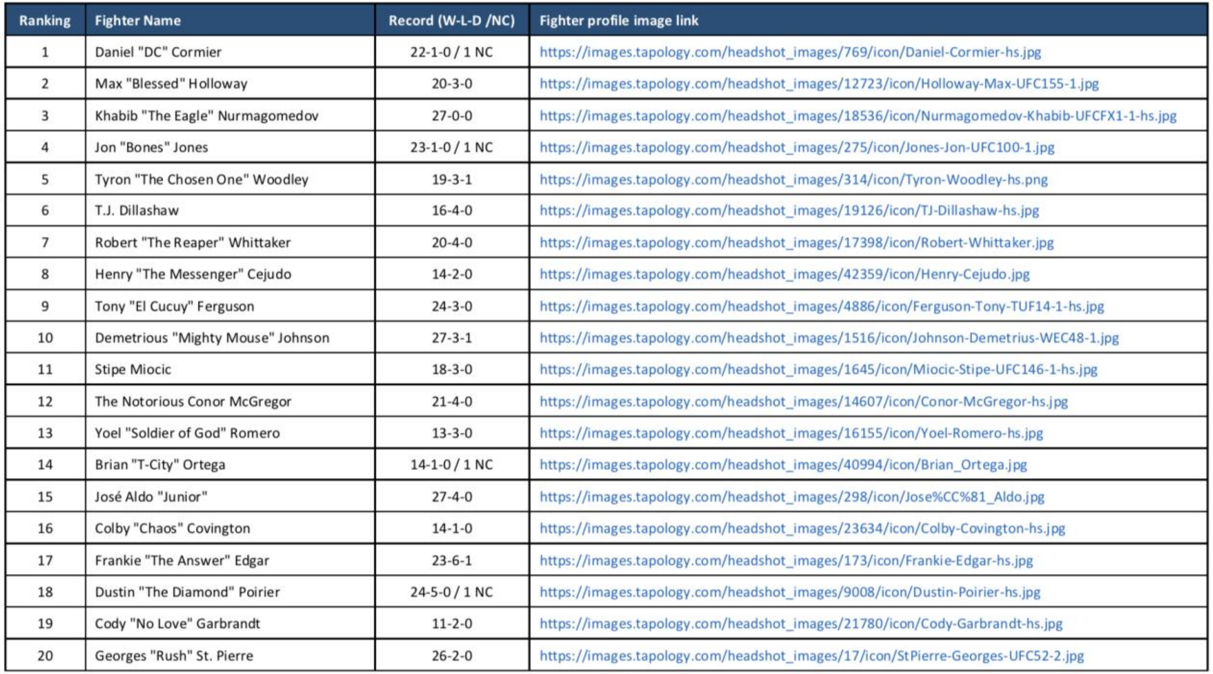
```
2019-01-21 20:42:58 [scrapy.utils.log] INFO: Scrapy 1.5.1 started (bot: scrapybot)
2019-01-21 20:42:58 [scrapy.utils.log] INFO: Versions: lxml 4.2.6.0, libxml2 2.9.8, cssselect 1.0.3, parsel 1.5.1, w3lib 1.20.0, Twisted 18.9.0, Python 3.6.7 (default, Oct 22 2018, 11:32:17) -
[GCC 8.2.0], pyOpenSSL 18.0.0 (OpenSSL 1.1.0j  20 Nov 2018), cryptography 2.4.2, Platform Linux-4.14.79+-x86_64-with-Ubuntu-18.04-bionic
2019-01-21 20:42:58 [scrapy.crawler] INFO: Overridden settings: {'USER_AGENT': 'Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1)'}
2019-01-21 20:42:58 [scrapy.middleware] INFO: Enabled extensions:
['scrapy.extensions.corestats.CoreStats',
 'scrapy.extensions.telnet.TelnetConsole',
 'scrapy.extensions.memusage.MemoryUsage',
 'scrapy.extensions.logstats.LogStats']
2019-01-21 20:42:58 [scrapy.middleware] INFO: Enabled downloader middlewares:
['scrapy.downloadermiddlewares.httpauth.HttpAuthMiddleware',
 'scrapy.downloadermiddlewares.downloadtimeout.DownloadTimeoutMiddleware',
 'scrapy.downloadermiddlewares.defaultheaders.DefaultHeadersMiddleware',
 'scrapy.downloadermiddlewares.useragent.UserAgentMiddleware',
 'scrapy.downloadermiddlewares.retry.RetryMiddleware',
 'scrapy.downloadermiddlewares.redirect.MetaRefreshMiddleware',
 'scrapy.downloadermiddlewares.httpcompression.HttpCompressionMiddleware',
 'scrapy.downloadermiddlewares.redirect.RedirectMiddleware',
 'scrapy.downloadermiddlewares.cookies.CookiesMiddleware',
 'scrapy.downloadermiddlewares.httpproxy.HttpProxyMiddleware',
 'scrapy.downloadermiddlewares.stats.DownloaderStats']
2019-01-21 20:42:58 [scrapy.middleware] INFO: Enabled spider middlewares:
['scrapy.spidermiddlewares.httperror.HttpErrorMiddleware',
 'scrapy.spidermiddlewares.offsite.OffsiteMiddleware',
 'scrapy.spidermiddlewares.referer.RefererMiddleware',
 'scrapy.spidermiddlewares.urllength.UrlLengthMiddleware',
 'scrapy.spidermiddlewares.depth.DepthMiddleware']
2019-01-21 20:42:58 [scrapy.middleware] INFO: Enabled item pipelines:
[]
2019-01-21 20:42:58 [scrapy.core.engine] INFO: Spider opened
2019-01-21 20:42:58 [scrapy.extensions.logstats] INFO: Crawled 0 pages (at 0 pages/min), scraped 0 items (at 0 items/min)
2019-01-21 20:42:58 [scrapy.extensions.telnet] DEBUG: Telnet console listening on 127.0.0.1:6023
2019-01-21 20:42:58 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.tapology.com/rankings/current-top-ten-best-pound-for-pound-mma-and-ufc-fighters> (referer: None)
2019-01-21 20:42:58 [scrapy.dupefilters] DEBUG: Filtered duplicate request: <GET https://www.tapology.com/rankings/current-top-ten-best-pound-for-pound-mma-and-ufc-fighters?page=2&ranking=1> - no more duplicates will be shown (see DUPEFILTER_DEBUG to show all duplicates)
2019-01-21 20:42:58 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.tapology.com/rankings/current-top-ten-best-pound-for-pound-mma-and-ufc-fighters?page=2&ranking=1> (referer: https://www.tapology.com/rankings/current-top-ten-best-pound-for-pound-mma-and-ufc-fighters)
2019-01-21 20:42:59 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.tapology.com/rankings/current-top-ten-best-pound-for-pound-mma-and-ufc-fighters?page=3&ranking=1> (referer: https://www.tapology.com/rankings/current-top-ten-best-pound-for-pound-mma-and-ufc-fighters?page=2&ranking=1)
2019-01-21 20:42:59 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.tapology.com/rankings/current-top-ten-best-pound-for-pound-mma-and-ufc-fighters?page=4&ranking=1> (referer: https://www.tapology.com/rankings/current-top-ten-best-pound-for-pound-mma-and-ufc-fighters?page=3&ranking=1)
2019-01-21 20:42:59 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.tapology.com/rankings/current-top-ten-best-pound-for-pound-mma-and-ufc-fighters?page=5&ranking=1> (referer: https://www.tapology.com/rankings/current-top-ten-best-pound-for-pound-mma-and-ufc-fighters?page=4&ranking=1)
2019-01-21 20:42:59 [scrapy.core.engine] INFO: Closing spider (finished)
2019-01-21 20:42:59 [scrapy.statscollectors] INFO: Dumping Scrapy stats:
{'downloader/request_bytes': 3484,
 'downloader/request_count': 5,
 'downloader/request_method_count/GET': 5,
 'downloader/response_bytes': 225390,
 'downloader/response_count': 5,
 'downloader/response_status_count/200': 5,
 'dupefilter/filtered': 4,
 'finish_reason': 'finished',
 'finish_time': datetime.datetime(2019, 1, 21, 20, 42, 59, 917198),
 'log_count/DEBUG': 7,
 'log_count/INFO': 7,
 'memusage/max': 164528128,
 'memusage/startup': 164528128,
 'request_depth_max': 4,
 'response_received_count': 5,
 'scheduler/dequeued': 5,
 'scheduler/dequeued/memory': 5,
 'scheduler/enqueued': 5,
 'scheduler/enqueued/memory': 5,
 'start_time': datetime.datetime(2019, 1, 21, 20, 42, 58, 228609)}
2019-01-21 20:42:59 [scrapy.core.engine] INFO: Spider closed (finished)
```

- Después de la correcta ejecución se procede a la descarga del fichero generado



```
from google.colab import files
files.download('/content/drive/My Drive/mmadata/fighters_ranking.csv')
```

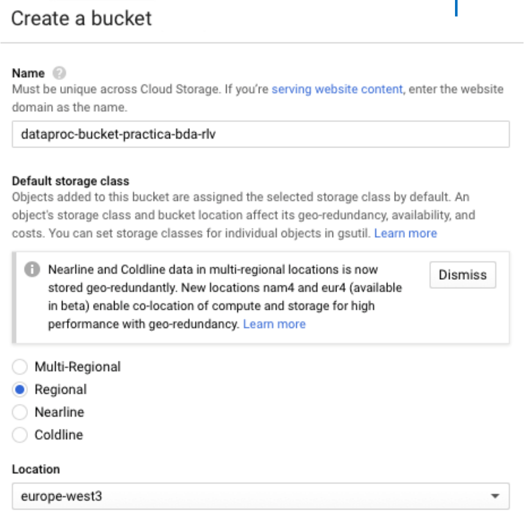
- Dando formato al fichero descargado (*fighters_ranking.csv*), obtenemos la siguiente lista con el ranking, nombre, récord y enlace a la foto correspondiente de cada luchador *(solo se mostrarán aquí los 20 primeros)*.



Ranking	Fighter Name	Record (W-L-D /NC)	Fighter profile image link
1	Daniel "DC" Cormier	22-1-0 / 1 NC	https://images.tapology.com/headshot_images/769/icon/Daniel-Cormier-hs.jpg
2	Max "Blessed" Holloway	20-3-0	https://images.tapology.com/headshot_images/12723/icon/Holloway-Max-UFC155-1.jpg
3	Khabib "The Eagle" Nurmagomedov	27-0-0	https://images.tapology.com/headshot_images/18536/icon/Nurmagomedov-Khabib-UFCFX1-1-hs.jpg
4	Jon "Bones" Jones	23-1-0 / 1 NC	https://images.tapology.com/headshot_images/275/icon/Jones-Jon-UFC100-1.jpg
5	Tyron "The Chosen One" Woodley	19-3-1	https://images.tapology.com/headshot_images/314/icon/Tyron-Woodley-hs.png
6	T.J. Dillashaw	16-4-0	https://images.tapology.com/headshot_images/19126/icon/TJ-Dillashaw-hs.jpg
7	Robert "The Reaper" Whittaker	20-4-0	https://images.tapology.com/headshot_images/17398/icon/Robert-Whittaker.jpg
8	Henry "The Messenger" Cejudo	14-2-0	https://images.tapology.com/headshot_images/42359/icon/Henry-Cejudo.jpg
9	Tony "El Cucuy" Ferguson	24-3-0	https://images.tapology.com/headshot_images/4886/icon/Ferguson-Tony-TUF14-1-hs.jpg
10	Demetrious "Mighty Mouse" Johnson	27-3-1	https://images.tapology.com/headshot_images/1516/icon/Johnson-Demetrius-WEC48-1.jpg
11	Stipe Miocic	18-3-0	https://images.tapology.com/headshot_images/1645/icon/Miocic-Stipe-UFC146-1-hs.jpg
12	The Notorious Conor McGregor	21-4-0	https://images.tapology.com/headshot_images/14607/icon/Conor-McGregor-hs.jpg
13	Yoel "Soldier of God" Romero	13-3-0	https://images.tapology.com/headshot_images/16155/icon/Yoel-Romero-hs.jpg
14	Brian "T-City" Ortega	14-1-0 / 1 NC	https://images.tapology.com/headshot_images/40994/icon/Brian_Ortega.jpg
15	José Aldo "Junior"	27-4-0	https://images.tapology.com/headshot_images/298/icon/Jose%CC%81_Aldo.jpg
16	Colby "Chaos" Covington	14-1-0	https://images.tapology.com/headshot_images/23634/icon/Colby-Covington-hs.jpg
17	Frankie "The Answer" Edgar	23-6-1	https://images.tapology.com/headshot_images/173/icon/Frankie-Edgar-hs.jpg
18	Dustin "The Diamond" Poirier	24-5-0 / 1 NC	https://images.tapology.com/headshot_images/9008/icon/Dustin-Poirier-hs.jpg
19	Cody "No Love" Garbrandt	11-2-0	https://images.tapology.com/headshot_images/21780/icon/Cody-Garbrandt-hs.jpg
20	Georges "Rush" St. Pierre	26-2-0	https://images.tapology.com/headshot_images/17/icon/StPierre-Georges-UFC52-2.jpg

Sprint 3 - Utilizar un proveedor de Cloud para montar un clúster de al menos 3 contenedores configurados correctamente.

Google Cloud Platform



Create a bucket

Name
Must be unique across Cloud Storage. If you're serving website content, enter the website domain as the name.

dataproc-bucket-practica-bda-rlv

Default storage class
Objects added to this bucket are assigned the selected storage class by default. An object's storage class and bucket location affect its geo-redundancy, availability, and costs. You can set storage classes for individual objects in gsutil. Learn more

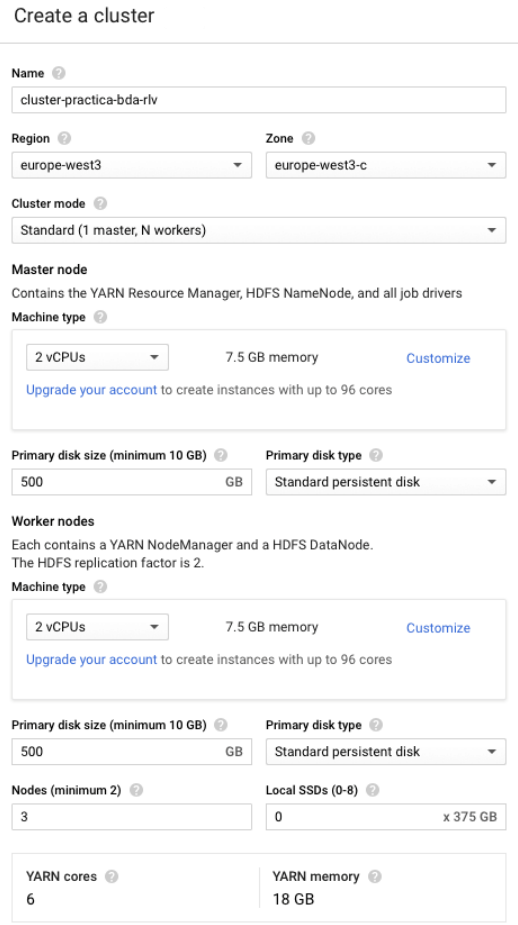
Nearline and Coldline data in multi-regional locations is now stored geo-redundantly. New locations nam4 and eur4 (available in beta) enable co-location of compute and storage for high performance with geo-redundancy. Learn more

Dismiss

○ Multi-Regional
● Regional
○ Nearline
○ Coldline

Location
europe-west3

Creamos primero el bucket para poder elegir el nombre y luego asignarlo al clúster en su creación



Create a cluster

Name
cluster-practica-bda-rlv

Region
europe-west3

Zone
europe-west3-c

Cluster mode
Standard (1 master, N workers)

Master node
Contains the YARN Resource Manager, HDFS NameNode, and all job drivers

Machine type
2 vCPUs — 7.5 GB memory — Customize

Upgrade your account to create instances with up to 96 cores

Primary disk size (minimum 10 GB)
500 GB

Primary disk type
Standard persistent disk

Worker nodes
Each contains a YARN NodeManager and a HDFS DataNode.
The HDFS replication factor is 2.

Machine type
2 vCPUs — 7.5 GB memory — Customize

Upgrade your account to create instances with up to 96 cores

Primary disk size (minimum 10 GB)
500 GB

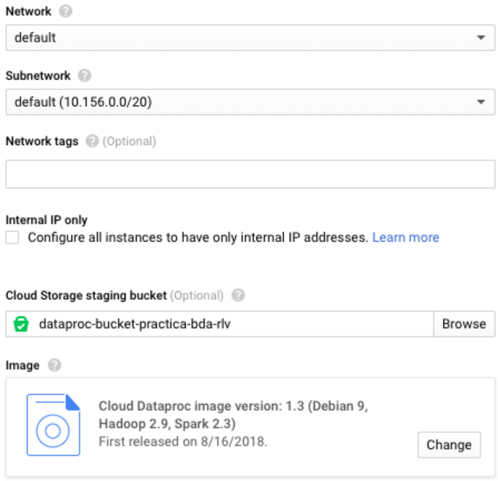
Primary disk type
Standard persistent disk

Nodes (minimum 2)
3

Local SSDs (0-8)
0 x 375 GB

YARN cores
6

YARN memory
18 GB



Network
default

Subnetwork
default (10.156.0.0/20)

Network tags (Optional)

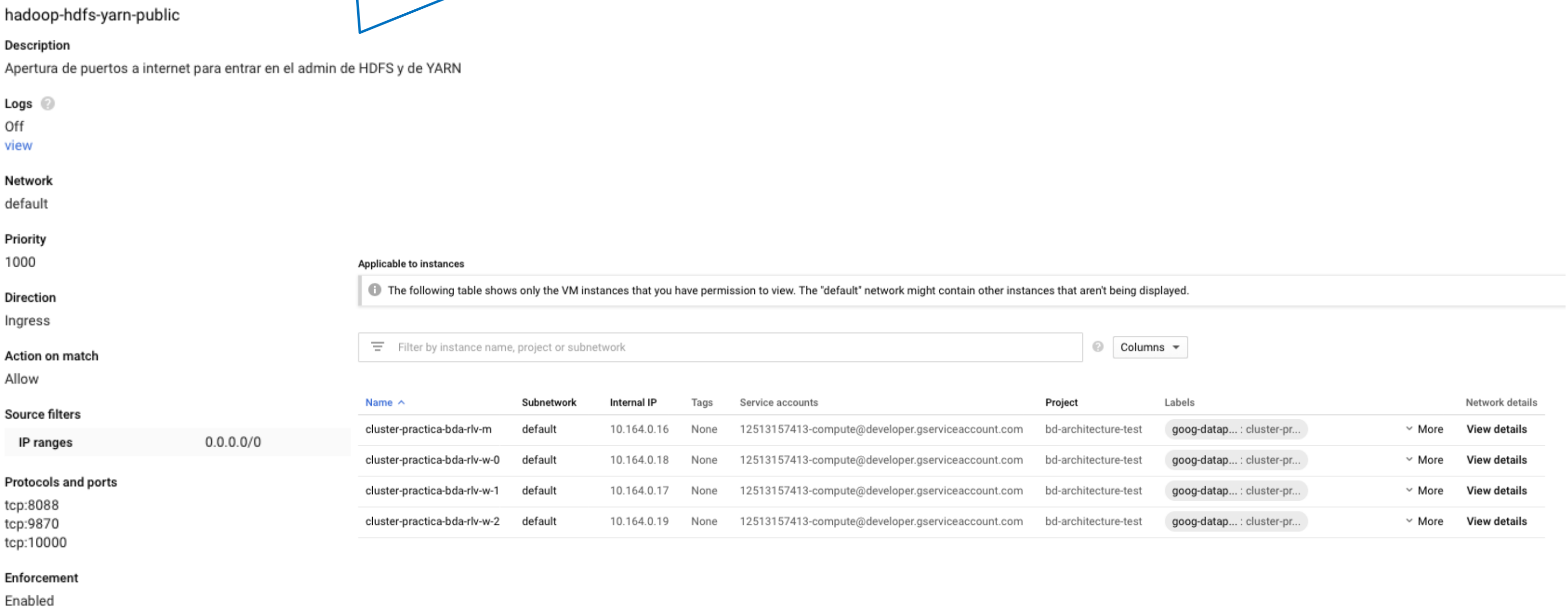
Internal IP only
□ Configure all instances to have only internal IP addresses. Learn more

Cloud Storage staging bucket (Optional)
dataproc-bucket-practica-bda-rlv — Browse

Image
Cloud Dataproc image version: 1.3 (Debian 9, Hadoop 2.9, Spark 2.3)
First released on 8/16/2018.

Change

Reglas del firewall en las que se "abren" los puertos 8088, 9870 y 10000 para cualquier IP



hadoop-hdfs-yarn-public

Description
Apertura de puertos a internet para entrar en el admin de HDFS y de YARN

Logs
Off
view

Network
default

Priority
1000

Direction
Ingress

Action on match
Allow

Source filters

IP ranges	0.0.0.0/0

Protocols and ports
tcp:8088
tcp:9870
tcp:10000

Enforcement
Enabled

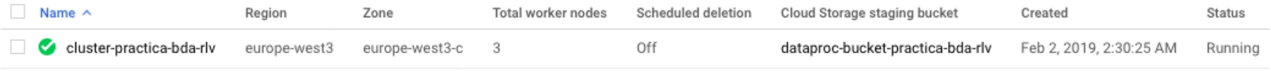
Applicable to instances

The following table shows only the VM instances that you have permission to view. The "default" network might contain other instances that aren't being displayed.

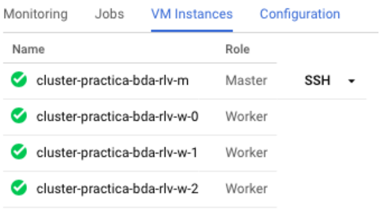
Filter by instance name, project or subnetwork

Name	Subnetwork	Internal IP	Tags	Service accounts	Project	Labels		Network details
cluster-practica-bda-rlv-m	default	10.164.0.16	None	12513157413-compute@developer.gserviceaccount.com	bd-architecture-test	goog-datap... : cluster-pr...	More	View details
cluster-practica-bda-rlv-w-0	default	10.164.0.18	None	12513157413-compute@developer.gserviceaccount.com	bd-architecture-test	goog-datap... : cluster-pr...	More	View details
cluster-practica-bda-rlv-w-1	default	10.164.0.17	None	12513157413-compute@developer.gserviceaccount.com	bd-architecture-test	goog-datap... : cluster-pr...	More	View details
cluster-practica-bda-rlv-w-2	default	10.164.0.19	None	12513157413-compute@developer.gserviceaccount.com	bd-architecture-test	goog-datap... : cluster-pr...	More	View details

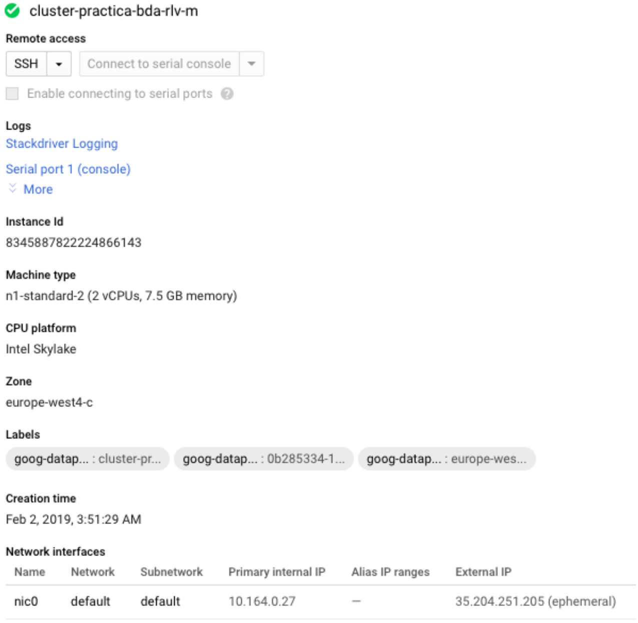
Comprobamos el estado y detalles del clúster creado



□	Name	Region	Zone	Total worker nodes	Scheduled deletion	Cloud Storage staging bucket	Created	Status
□	✓ cluster-practica-bda-rlv	europe-west3	europe-west3-c	3	Off	dataproc-bucket-practica-bda-rlv	Feb 2, 2019, 2:30:25 AM	Running



Name	Role	
✓ cluster-practica-bda-rlv-m	Master	SSH
✓ cluster-practica-bda-rlv-w-0	Worker	
✓ cluster-practica-bda-rlv-w-1	Worker	
✓ cluster-practica-bda-rlv-w-2	Worker	



✓ cluster-practica-bda-rlv-m

Remote access

□ Enable connecting to serial ports

Logs

Stackdriver Logging

Serial port 1 (console)

More

Instance Id

8345887822224866143

Machine type

n1-standard-2 (2 vCPUs, 7.5 GB memory)

CPU platform

Intel Skylake

Zone

europe-west4-c

Labels

goog-datap... : cluster-pr... | goog-datap... : 0b285334-1... | goog-datap... : europe-wes...

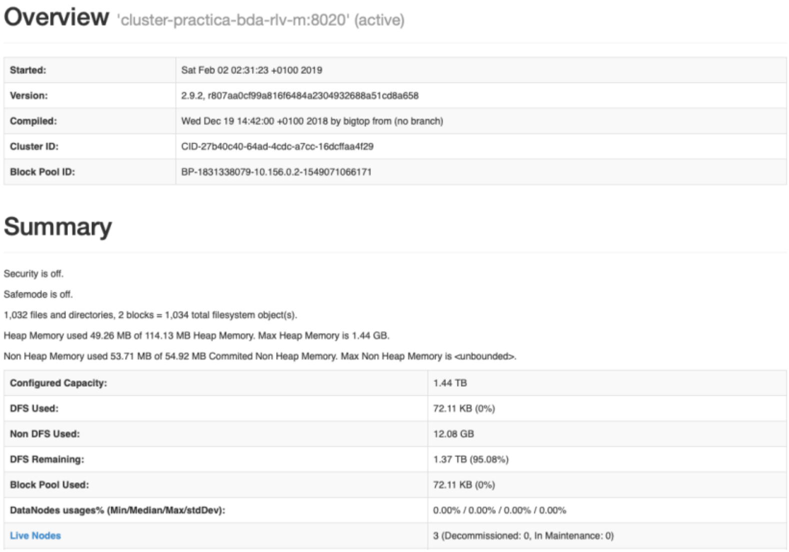
Creation time

Feb 2, 2019, 3:51:29 AM

Network interfaces

Name	Network	Subnetwork	Primary internal IP	Alias IP ranges	External IP
nic0	default	default	10.164.0.27	–	35.204.251.205 (ephemeral)

http://35.204.251.205:9870/dfshealth.html#tab-overview



Overview 'cluster-practica-bda-rlv-m:8020' (active)

Started:	Sat Feb 02 02:31:23 +0100 2019
Version:	2.9.2, r807aa0cf99a816f6484a2304932688a51cd8a658
Compiled:	Wed Dec 19 14:42:00 +0100 2018 by bigtop from (no branch)
Cluster ID:	CID-27b40c40-64ad-4cdc-a7cc-16dcffaa4f29
Block Pool ID:	BP-1831338079-10.156.0.2-1549071066171

Summary

Security is off.

Safemode is off.

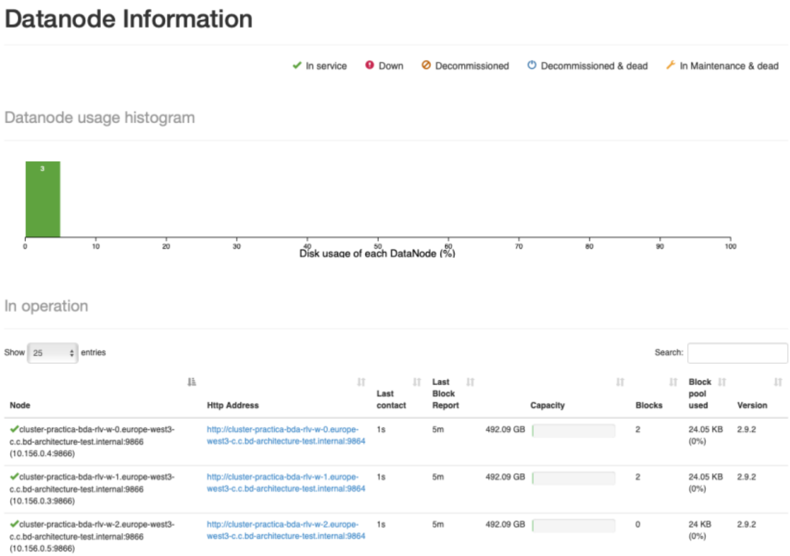
1,032 files and directories, 2 blocks = 1,034 total filesystem object(s).

Heap Memory used 49.26 MB of 114.13 MB Heap Memory. Max Heap Memory is 1.44 GB.

Non Heap Memory used 53.71 MB of 54.92 MB Commited Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	1.44 TB
DFS Used:	72.11 KB (0%)
Non DFS Used:	12.08 GB
DFS Remaining:	1.37 TB (95.08%)
Block Pool Used:	72.11 KB (0%)
DataNodes usages% (Min/Median/Max/stdDev):	0.00% / 0.00% / 0.00% / 0.00%
Live Nodes	3 (Decommissioned: 0, In Maintenance: 0)

http://35.204.251.205:9870/dfshealth.html#tab-datanode



Datanode Information

✓ In service ● Down ⊘ Decommissioned ○ Decommissioned & dead ✎ In Maintenance & dead

Datanode usage histogram

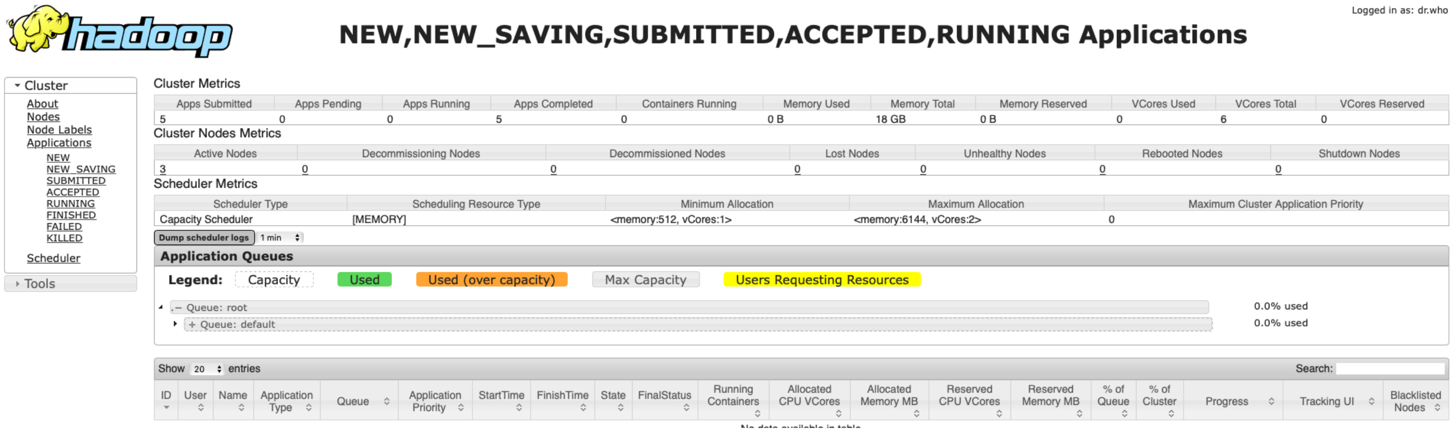
In operation

Show 25 entries

Search:

Node	Http Address	Last contact	Last Block Report	Capacity	Blocks	Block pool used	Version
✓cluster-practica-bda-rlv-w-0.europe-west3-c.c.bd-architecture-test.internal:9866 (10.156.0.4:9866)	http://cluster-practica-bda-rlv-w-0.europe-west3-c.c.bd-architecture-test.internal:9864	1s	5m	492.09 GB	2	24.05 KB (0%)	2.9.2
✓cluster-practica-bda-rlv-w-1.europe-west3-c.c.bd-architecture-test.internal:9866 (10.156.0.3:9866)	http://cluster-practica-bda-rlv-w-1.europe-west3-c.c.bd-architecture-test.internal:9864	1s	5m	492.09 GB	2	24.05 KB (0%)	2.9.2
✓cluster-practica-bda-rlv-w-2.europe-west3-c.c.bd-architecture-test.internal:9866 (10.156.0.5:9866)	http://cluster-practica-bda-rlv-w-2.europe-west3-c.c.bd-architecture-test.internal:9864	1s	5m	492.09 GB	0	24 KB (0%)	2.9.2

http://35.204.251.205:8088/cluster/scheduler



NEW,NEW_SAVING,SUBMITTED,ACCEPTED,RUNNING Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved
5	0	0	5	0	0 B	18 GB	0 B	0	6	0

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes	Shutdown Nodes
3	0	0	0	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	Maximum Cluster Application Priority
Capacity Scheduler	[MEMORY]	<memory:512, vCores:1>	<memory:6144, vCores:2>	0

Application Queues

Legend: Capacity 🟩 Used 🟧 Used (over capacity) Max Capacity 🟨 Users Requesting Resources

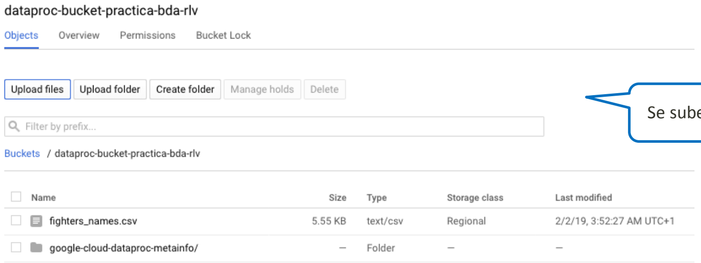
- Queue: root 0.0% used
+ Queue: default 0.0% used

Show 20 entries Search:

ID	User	Name	Application Type	Queue	Application Priority	StartTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	Reserved CPU VCores	Reserved Memory MB	% of Queue	% of Cluster	Progress	Tracking UI	Blacklisted Nodes

No data available in table

Sprint 4 - Proveer resultados de una tarea de procesamiento.

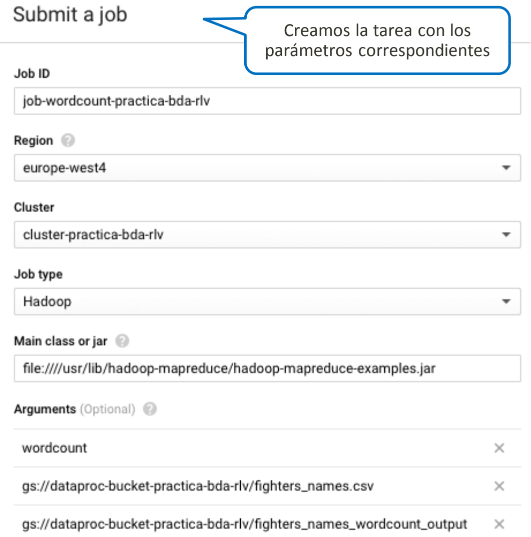


dataproc-bucket-practica-bda-rlv

Buckets / dataproc-bucket-practica-bda-rlv

□	Name	Size	Type	Storage class	Last modified
□	fighters_names.csv	5.55 KB	text/csv	Regional	2/2/19, 3:52:27 AM UTC+1
□	google-cloud-dataproc-metainfo/	–	Folder	–	–

Se sube el fichero a tratar (modificación del obtenido en el crawler del sprint 2)



Submit a job

Job ID: job-wordcount-practica-bda-rlv

Region: europe-west4

Cluster: cluster-practica-bda-rlv

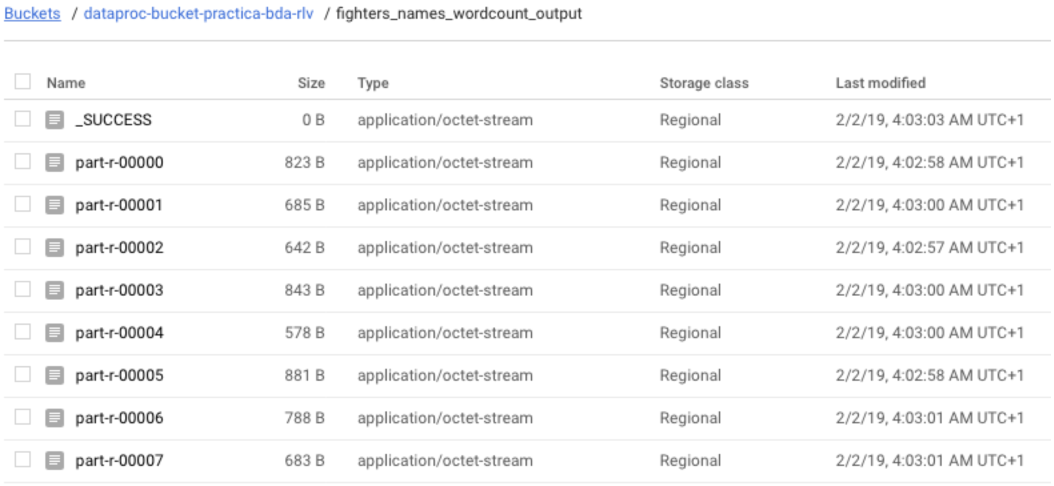
Job type: Hadoop

Main class or jar: file:///usr/lib/hadoop-mapreduce/hadoop-mapreduce-examples.jar

Arguments (Optional)

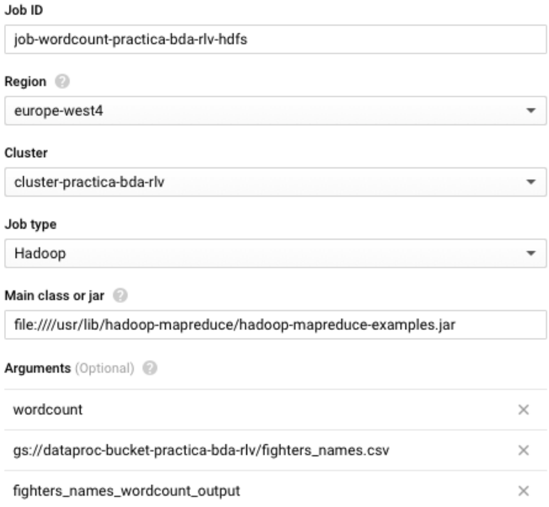
- wordcount
- gs://dataproc-bucket-practica-bda-rlv/fighters_names.csv
- gs://dataproc-bucket-practica-bda-rlv/fighters_names_wordcount_output

Creamos la tarea con los parámetros correspondientes



Buckets / dataproc-bucket-practica-bda-rlv / fighters_names_wordcount_output

□	Name	Size	Type	Storage class	Last modified
□	_SUCCESS	0 B	application/octet-stream	Regional	2/2/19, 4:03:03 AM UTC+1
□	part-r-00000	823 B	application/octet-stream	Regional	2/2/19, 4:02:58 AM UTC+1
□	part-r-00001	685 B	application/octet-stream	Regional	2/2/19, 4:03:00 AM UTC+1
□	part-r-00002	642 B	application/octet-stream	Regional	2/2/19, 4:02:57 AM UTC+1
□	part-r-00003	843 B	application/octet-stream	Regional	2/2/19, 4:03:00 AM UTC+1
□	part-r-00004	578 B	application/octet-stream	Regional	2/2/19, 4:03:00 AM UTC+1
□	part-r-00005	881 B	application/octet-stream	Regional	2/2/19, 4:02:58 AM UTC+1
□	part-r-00006	788 B	application/octet-stream	Regional	2/2/19, 4:03:01 AM UTC+1
□	part-r-00007	683 B	application/octet-stream	Regional	2/2/19, 4:03:01 AM UTC+1



Job ID
job-wordcount-practica-bda-rlv-hdfs

Region
europe-west4

Cluster
cluster-practica-bda-rlv

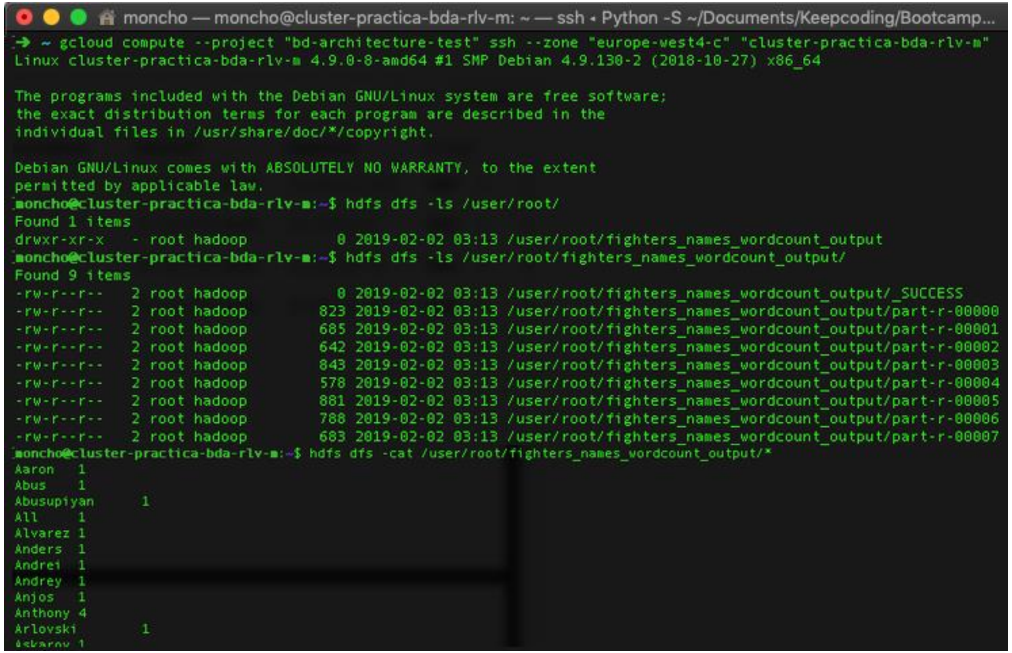
Job type
Hadoop

Main class or jar
file:///usr/lib/hadoop-mapreduce/hadoop-mapreduce-examples.jar

Arguments (Optional)
wordcount
gs://dataproc-bucket-practica-bda-rlv/fighters_names.csv
fighters_names_wordcount_output

En esta ocasión no generamos el resultado en Google Storage

Comandos Hadoop

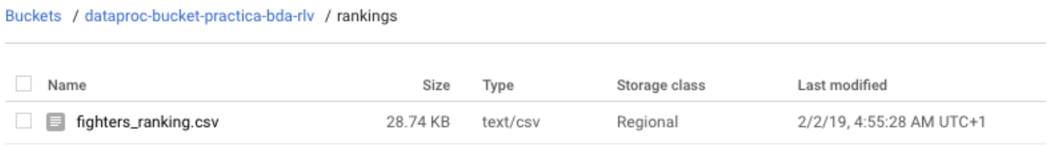


moncho — moncho@cluster-practica-bda-rlv-m: ~ — ssh ◂ Python -S ~/Documents/Keepcoding/Bootcamp...

```
➜ ~ gcloud compute --project "bd-architecture-test" ssh --zone "europe-west4-c" "cluster-practica-bda-rlv-m"
Linux cluster-practica-bda-rlv-m 4.9.0-8-amd64 #1 SMP Debian 4.9.130-2 (2018-10-27) x86_64

The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
moncho@cluster-practica-bda-rlv-m:~$ hdfs dfs -ls /user/root/
Found 1 items
drwxr-xr-x   - root hadoop          0 2019-02-02 03:13 /user/root/fighters_names_wordcount_output
moncho@cluster-practica-bda-rlv-m:~$ hdfs dfs -ls /user/root/fighters_names_wordcount_output/
Found 9 items
-rw-r--r--   2 root hadoop          0 2019-02-02 03:13 /user/root/fighters_names_wordcount_output/_SUCCESS
-rw-r--r--   2 root hadoop        823 2019-02-02 03:13 /user/root/fighters_names_wordcount_output/part-r-00000
-rw-r--r--   2 root hadoop        685 2019-02-02 03:13 /user/root/fighters_names_wordcount_output/part-r-00001
-rw-r--r--   2 root hadoop        642 2019-02-02 03:13 /user/root/fighters_names_wordcount_output/part-r-00002
-rw-r--r--   2 root hadoop        843 2019-02-02 03:13 /user/root/fighters_names_wordcount_output/part-r-00003
-rw-r--r--   2 root hadoop        578 2019-02-02 03:13 /user/root/fighters_names_wordcount_output/part-r-00004
-rw-r--r--   2 root hadoop        881 2019-02-02 03:13 /user/root/fighters_names_wordcount_output/part-r-00005
-rw-r--r--   2 root hadoop        788 2019-02-02 03:13 /user/root/fighters_names_wordcount_output/part-r-00006
-rw-r--r--   2 root hadoop        683 2019-02-02 03:13 /user/root/fighters_names_wordcount_output/part-r-00007
moncho@cluster-practica-bda-rlv-m:~$ hdfs dfs -cat /user/root/fighters_names_wordcount_output/*
Aaron   1
Abus    1
Abusupiyan      1
All     1
Alvarez 1
Anders  1
Andrei  1
Andrey  1
Anjos   1
Anthony 4
Arlovski        1
Askarov 1
```

- ➢ `gcloud compute --project "bd-architecture-test" ssh --zone "europe-west4-c" "cluster-practica-bda-rlv-m”`
- ➢ `hdfs dfs -ls /user/root/fighters_names_wordcount_output/`
- ➢ `hdfs dfs -cat /user/root/fighters_names_wordcount_output/*`

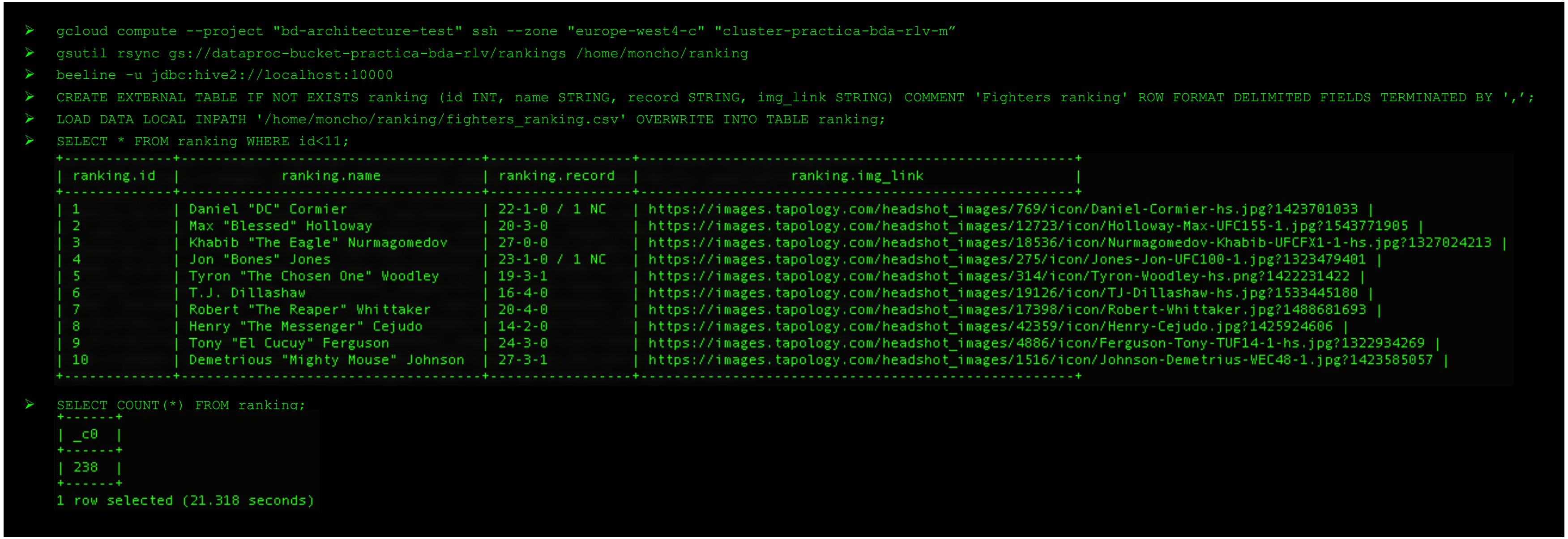
BONUS - Utilizar HIVE para categorizar los datos y hacer un par de queries con el command line para extraer datos a un fichero.



Buckets / dataproc-bucket-practica-bda-rlv / rankings

□ Name	Size	Type	Storage class	Last modified
□ fighters_ranking.csv	28.74 KB	text/csv	Regional	2/2/19, 4:55:28 AM UTC+1

Subimos al bucket el fichero obtenido con el crawler en el Sprint 2



```
➤ gcloud compute --project "bd-architecture-test" ssh --zone "europe-west4-c" "cluster-practica-bda-rlv-m”
➤ gsutil rsync gs://dataproc-bucket-practica-bda-rlv/rankings /home/moncho/ranking
➤ beeline -u jdbc:hive2://localhost:10000
➤ CREATE EXTERNAL TABLE IF NOT EXISTS ranking (id INT, name STRING, record STRING, img_link STRING) COMMENT 'Fighters ranking' ROW FORMAT DELIMITED FIELDS TERMINATED BY ',’;
➤ LOAD DATA LOCAL INPATH '/home/moncho/ranking/fighters_ranking.csv' OVERWRITE INTO TABLE ranking;
➤ SELECT * FROM ranking WHERE id<11;
  +-------------+-----------------------------------+-----------------+-----------------------------------------------------+
  | ranking.id  |           ranking.name            | ranking.record  |                  ranking.img_link                   |
  +-------------+-----------------------------------+-----------------+-----------------------------------------------------+
  | 1           | Daniel "DC" Cormier               | 22-1-0 / 1 NC   | https://images.tapology.com/headshot_images/769/icon/Daniel-Cormier-hs.jpg?1423701033 |
  | 2           | Max "Blessed" Holloway            | 20-3-0          | https://images.tapology.com/headshot_images/12723/icon/Holloway-Max-UFC155-1.jpg?1543771905 |
  | 3           | Khabib "The Eagle" Nurmagomedov   | 27-0-0          | https://images.tapology.com/headshot_images/18536/icon/Nurmagomedov-Khabib-UFCFX1-1-hs.jpg?1327024213 |
  | 4           | Jon "Bones" Jones                 | 23-1-0 / 1 NC   | https://images.tapology.com/headshot_images/275/icon/Jones-Jon-UFC100-1.jpg?1323479401 |
  | 5           | Tyron "The Chosen One" Woodley    | 19-3-1          | https://images.tapology.com/headshot_images/314/icon/Tyron-Woodley-hs.png?1422231422 |
  | 6           | T.J. Dillashaw                    | 16-4-0          | https://images.tapology.com/headshot_images/19126/icon/TJ-Dillashaw-hs.jpg?1533445180 |
  | 7           | Robert "The Reaper" Whittaker     | 20-4-0          | https://images.tapology.com/headshot_images/17398/icon/Robert-Whittaker.jpg?1488681693 |
  | 8           | Henry "The Messenger" Cejudo      | 14-2-0          | https://images.tapology.com/headshot_images/42359/icon/Henry-Cejudo.jpg?1425924606 |
  | 9           | Tony "El Cucuy" Ferguson          | 24-3-0          | https://images.tapology.com/headshot_images/4886/icon/Ferguson-Tony-TUF14-1-hs.jpg?1322934269 |
  | 10          | Demetrious "Mighty Mouse" Johnson | 27-3-1          | https://images.tapology.com/headshot_images/1516/icon/Johnson-Demetrius-WEC48-1.jpg?1423585057 |
  +-------------+-----------------------------------+-----------------+-----------------------------------------------------+

➤ SELECT COUNT(*) FROM ranking;
  +------+
  | _c0  |
  +------+
  | 238  |
  +------+
  1 row selected (21.318 seconds)
```