



CURSO DE ESPECIALIZACIÓN EN INTELIGENCIA ARTIFICIAL Y BIGDATA

BIGDATA APLICADO

TAREA EVALUABLE 5.1

Autor: Carlos Sánchez Recio.
20 / 02 / 2025

Índice

Apartado 1: Integridad de los datos	1
Apartado 2: Herramientas internas de monitorización de Hadoop	2
Apartado 3: Ganglia	6
Apartado 4: Apache Ambari	9

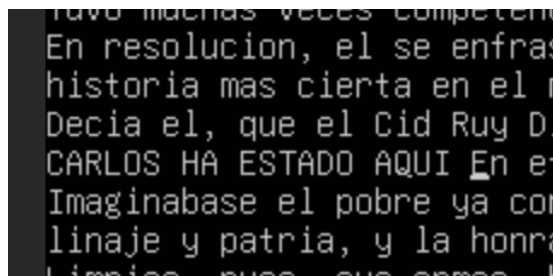
Apartado 1: Integridad de los datos

En una máquina con Linux (puede ser la Cloudera QuickStart VM) copia el archivo [quijote.txt](#) que ya hemos empleado en anteriores entregas.

Ejecuta las órdenes md5sum, sha256sum y sha512sum para obtener las sumas de verificación y realiza una captura de pantalla donde se vean las tres sumas obtenidas.

```
carlos@carlos:~/Documents$ ls
quijote.txt
carlos@carlos:~/Documents$ md5sum quijote.txt
a11a03ab70fead939d3842c3d5a8bb64 quijote.txt
carlos@carlos:~/Documents$ sha256sum quijote.txt
a67c1e3fefac58d0d2ee9f7c11d3790d4d0bc772115d5a77b964d733e9c1404d quijote.txt
carlos@carlos:~/Documents$ sha512sum quijote.txt
f39ac8fb273e9db8ac51ce1b3dfc27b88f2f3ddd67d260bd792987f7664f24ea855c829361e86e8a9c5ab581ce74d3b83a477dd97063178d9cf9f4d9494341d3 quijote.txt
carlos@carlos:~/Documents$ _
```

A continuación, haz alguna pequeña modificación en el archivo [quijote.txt](#) y vuelve a generar las tres sumas de verificación. Haz otra captura de pantalla en la que se vean las tres.



```
carlos@carlos:~/Documents$ md5sum quijote.txt
040720571f43044c97800c1d518cf7a6 quijote.txt
carlos@carlos:~/Documents$ sha256sum quijote.txt
d94d476df53c163888a461574a945fbd36875334b6b344c79df8264659eeb32 quijote.txt
carlos@carlos:~/Documents$ sha512sum quijote.txt
40af45fa97f8949beedf0de008115048a0fd350ca04afe18aa4389285cb908f5ec4d6bc301f32cfad21a65592344f279f9f54e2f35c0660003f90e0e11c587a7 quijote.txt
carlos@carlos:~/Documents$ _
```

Di si después de realizar el cambio, las sumas de verificación han cambiado o no.

Como se puede apreciar en la primera imagen y en la última, tras realizar las operaciones de hash en los archivos original y modificado, las sumas de verificación resultan diferentes siendo éste el resultado esperado.

Apartado 2: Herramientas internas de monitorización de Hadoop

En la Cloudera QuickStart VM, realiza las siguientes capturas de pantalla e insértelas en tu documento:

- **HDFS NameNode:** <http://quickstart.cloudera:50070> (apartados Overview y Summary). ¿Qué espacio (y porcentaje) del DFS está siendo utilizado?

OVERVIEW quickstart.cloudera:50070 (active)

Started:	Tue Jan 07 13:02:21 +0100 2025
Version:	2.6.0-cdh5.13.0, r42e8860b182e55321bd5f5605264da4adc8882be
Compiled:	Wed Oct 04 20:08:00 +0200 2017 by jenkins from Unknown
Cluster ID:	CID-a24185f9-a545-40fe-9553-84c3fda489f
Block Pool ID:	BP-1067413441-127.0.0.1-1508775264580

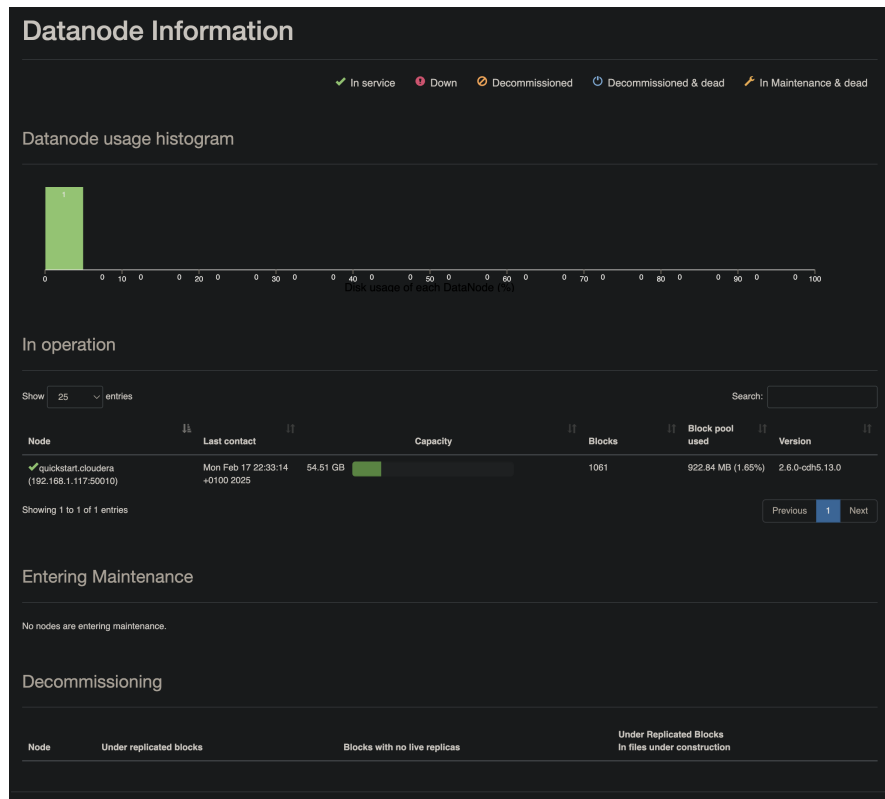
Summary

Security is off.
 Safemode is off.
 1,450 files and directories, 1,063 blocks = 2,513 total filesystem object(s).
 Heap Memory used 67.15 MB of 134 MB Heap Memory. Max Heap Memory is 889 MB.
 Non Heap Memory used 46.46 MB of 68.5 MB Committed Non Heap Memory. Max Non Heap Memory is 130 MB.

Configured Capacity:	54.51 GB
DFS Used:	922.84 MB (1.65%)
Non DFS Used:	9.12 GB
DFS Remaining:	41.47 GB (76.07%)
Block Pool Used:	922.84 MB (1.65%)
DataNodes usages% (Min/Median/Max/stdDev):	1.65% / 1.65% / 1.65% / 0.00%
Live Nodes	1 (Decommissioned: 0, In Maintenance: 0)
Dead Nodes	0 (Decommissioned: 0, In Maintenance: 0)
Decommissioning Nodes	0
Entering Maintenance Nodes	0
Total Datanode Volume Failures	0 (0 B)
Number of Under-Replicated Blocks	5
Number of Blocks Pending Deletion	0
Block Deletion Start Time	Tue Jan 07 13:02:21 +0100 2025
Last Checkpoint Time	Mon Feb 17 22:14:09 +0100 2025

Como se puede ver en la imagen anterior, se está utilizando un 1.65% del DFS lo que en mi caso equivale a un espacio de 922.84 MB.

- **HDFS DataNode information:** <http://quickstart.cloudera:50070/dfshealth.html#tab-datanode>. ¿Cuántos bloques contiene el DataNode actualmente?



Como se puede observar en la imagen, en la parte inferior en la tabla la columna *Blocks* indica que hay 1061 bloques actualmente.

- Datos de memoria del JMX. ¿Cuál es el máximo de memoria que se puede utilizar para el *heap*?

```

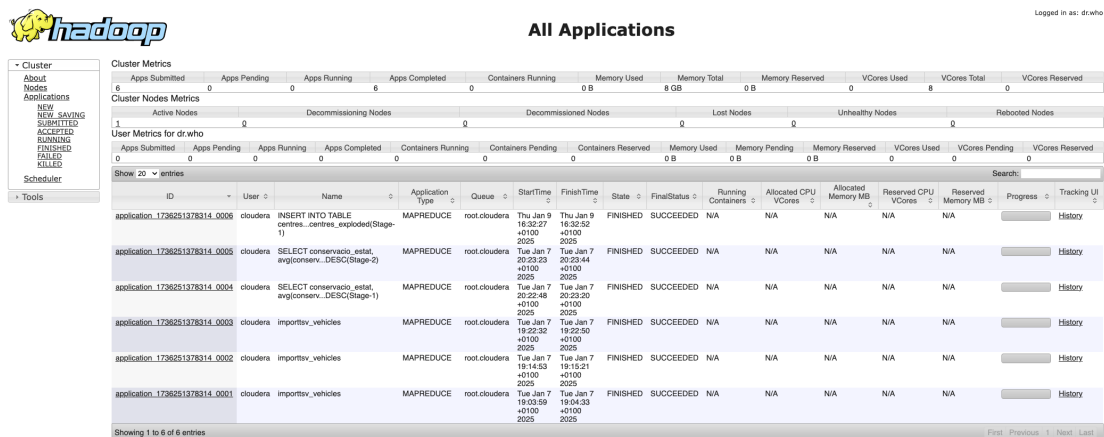
192.168.1.117
IEDIB
IEDIB repos
Gmail
Google Drive
Carlos
Apunts CE_5075 5.1: Cas pr...
CE_5075: Tasca avaluable C...
Getting Started on Kaggle [...
New Tab
192.168.1.117:50070/jmx

{
  "beans" : [ {
    "name" : "java.lang:type=Memory",
    "modelerType" : "sun.management.MemoryImpl",
    "Verbose" : false,
    "HeapMemoryUsage" : {
      "committed" : 146800640,
      "init" : 30749312,
      "max" : 932184064,
      "used" : 53951944
    },
    "NonHeapMemoryUsage" : {
      "committed" : 71892992,
      "init" : 24576000,
      "max" : 136314880,
      "used" : 48761872
    },
    "ObjectPendingFinalizationCount" : 0,
    "ObjectName" : "java.lang:type=Memory"
  }, {
    "name" : "java.lang:type=MemoryPool,name=PS Eden Space",
    "modelerType" : "sun.management.MemoryPoolImpl",
    "CollectionUsage" : {
      "committed" : 84934656,
      "init" : 8388608,
      "max" : 342360064,
      "used" : 0
    },
    "CollectionUsageThreshold" : 0,
    "CollectionUsageThresholdCount" : 0,
    "MemoryManagerNames" : [ "PS MarkSweep", "PS Scavenge" ],
    "PeakUsage" : {
      "committed" : 145752064,
      "init" : 8388608,
      "max" : 347602944,
      "used" : 145752064
    },
    "Usage" : {
      "committed" : 84934656,
      "init" : 8388608,
      "max" : 342360064,
      "used" : 19196600
    }
  } ]
}

```

Como se puede observar en la imagen, en la propiedad *committed* de *HeapMemoryUsage*, se pueden utilizar 146800640 bytes, lo cual equivale a unos 147 MB.

- **YARN ResourceManager:** <http://quickstart.cloudera:8088/cluster>



All 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
6	0	0	0	0	0 B	8 GB	0 B	0	8	0

User Metrics for dr.who

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Containers Pending	Containers Reserved	Memory Used	Memory Pending	Memory Reserved	VCores Used	VCores Pending	VCores Reserved
0	0	0	0	0	0	0 B	0 B	0 B	0 B	0	0	0

Showing 1 to 6 of 6 entries

- **YARN NodeManager:** <http://quickstart.cloudera:8042/node>



NodeManager information

Total Vmem allocated for Containers: 16.80 GB

Vmem enforcement enabled: false

Total Pmem allocated for Containers: 0 GB

Pmem enforcement enabled: true

Total Vcores allocated for Containers: 8

NodeHealthStatus: true

LastNodeHealthTime: Tue Feb 18 11:58:46 PST 2025

NodeHealthReport


NodeManager Version: 2.6.0-cdh5.13.0 from 42e880b182a55321bd5f5605264da4dc8882be by jenkins source checksum 80d31502ca6c524a0559051ad0221 on 2017-10-04T18:16Z

Hadoop Version: 2.6.0-cdh5.13.0 from 42e880b182a55321bd5f5605264da4dc8882be by jenkins source checksum 5e84c18598a2158e2b0e4b895311 on 2017-10-04T18:08Z

Ejecuta un trabajo en Pig e incluye también la captura del YARN ResourceManager. Indica si existe algún cambio respecto a la captura anterior de ResourceManager.

Para este apartado, utilizaré un archivo .pig el cual fue utilizado en entregas anteriores. En este caso para cargar los datos del archivo quijote.txt.

```
[cloudera@quickstart section1]$ cat 1.pig
-- Load all the data from the file
quijote = LOAD 'quijote/quijote.txt' AS (line:chararray);
-- Split lines into words
words = FOREACH quijote GENERATE FLATTEN(TOKENIZE(line)) as word;
-- Filter the words with REGEX
regex = FILTER words BY word MATCHES '^[A-Z].*a$';
-- Count the number of occurrences
group_words = GROUP regex BY word;
count_words = FOREACH group_words GENERATE group as word, COUNT(regex) AS count;
-- Sort results by number of occurrences (desc) and alphabetically (asc)
sort_words = ORDER count_words BY count DESC, word ASC;
-- Dump the results
DUMP sort_words;
[cloudera@quickstart section1]$ pig 1.pig
log4j:WARN No appenders could be found for logger (org.apache.hadoop.util.Shell).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.
2025-02-18 12:37:35,706 [main] INFO org.apache.pig.Main - Apache Pig version 0.12.0-cdh5.13.0 (rexpoted
) compiled Oct 04 2017, 11:09:03
2025-02-18 12:37:35,706 [main] INFO org.apache.pig.Main - Logging error messages to: /home/cloudera/pig_
files/block2/section1/pig_1739911055681.log
2025-02-18 12:37:36,713 [main] INFO org.apache.pig.impl.util.Utils - Default bootstrap file /home/cloudera/
.pigbootstrap not found
2025-02-18 12:37:36,850 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracke
r is deprecated. Instead, use mapreduce.jobtracker.address
2025-02-18 12:37:36,850 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is
 deprecated. Instead, use fs.defaultFS
```



Cluster

About

Nodes

Applications

NEW

SAVING

SUBMITTED

ACCEPTED

RUNNING

FINISHED

FAILED

KILLED

Scheduler

Tools

Cluster Metrics

Apps Submitted

Apps Pending

Apps Running

Apps Completed

Containers Running

Memory Used

Memory Total

Memory Reserved

VCoers Used

VCoers Total

VCoers Reserved

Cluster Nodes Metrics

Active Nodes

Decommissioning Nodes

Decommissioned Nodes

Lost Nodes

Unhealthy Nodes

Rebooted Nodes

User Metrics for dr:who

Apps Submitted

Apps Pending

Apps Running

Apps Completed

Containers Running

Containers Pending

Containers Reserved

Memory Used

Memory Pending

Memory Reserved

VCoers Used

VCoers Pending

VCoers Reserved

Show 20 entries

ID

User

Name

Application Type

Queue

StartTime

FinishTime

State

FinalStatus

Running Containers

Allocated CPU VCoers

Allocated Memory MB

Reserved CPU VCoers

Reserved Memory MB

Progress

Tracking UI

application_1736251378314_0007

cloudera

PigLatin1.pig

MAPREDUCE

root.cloudera

Tue Feb 18 21:37:52 +0100

N/A

ACCEPTED

UNDEFINED

1


1

2048

0

0

UNASSIGNED



Cluster

About

Nodes

Applications

NEW

SAVING

SUBMITTED

ACCEPTED

RUNNING

FINISHED

FAILED

KILLED

Scheduler

Tools

Cluster Metrics

Apps Submitted

Apps Pending

Apps Running

Apps Completed

Containers Running

Memory Used

Memory Total

Memory Reserved

VCoers Used

VCoers Total

VCoers Reserved

Cluster Nodes Metrics

Active Nodes

Decommissioning Nodes

Decommissioned Nodes

Lost Nodes

Unhealthy Nodes

Rebooted Nodes

User Metrics for dr:who

Apps Submitted

Apps Pending

Apps Running

Apps Completed

Containers Running

Containers Pending

Containers Reserved

Memory Used

Memory Pending

Memory Reserved

VCoers Used

VCoers Pending

VCoers Reserved

Show 20 entries

ID

User

Name

Application Type

Queue

StartTime

FinishTime

State

FinalStatus

Running Containers

Allocated CPU VCoers

Allocated Memory MB

Reserved CPU VCoers

Reserved Memory MB

Progress

Tracking UI

application_1736251378314_0009

cloudera

PigLatin1.pig

MAPREDUCE

root.cloudera

Tue Feb 18 21:38:59 +0100

Tue Feb 18 21:39:17 +0100

FINISHED

SUCCEEDED

N/A

N/A

N/A

N/A

N/A

History

application_1736251378314_0008

cloudera

PigLatin1.pig

MAPREDUCE

root.cloudera

Tue Feb 18 21:38:28 +0100

Tue Feb 18 21:38:51 +0100

FINISHED

SUCCEEDED

N/A

N/A

N/A

N/A

N/A

History

application_1736251378314_0007

cloudera

PigLatin1.pig

MAPREDUCE

root.cloudera

Tue Feb 18 21:37:52 +0100

Tue Feb 18 21:38:20 +0100

FINISHED

SUCCEEDED

N/A

N/A

N/A

N/A

N/A

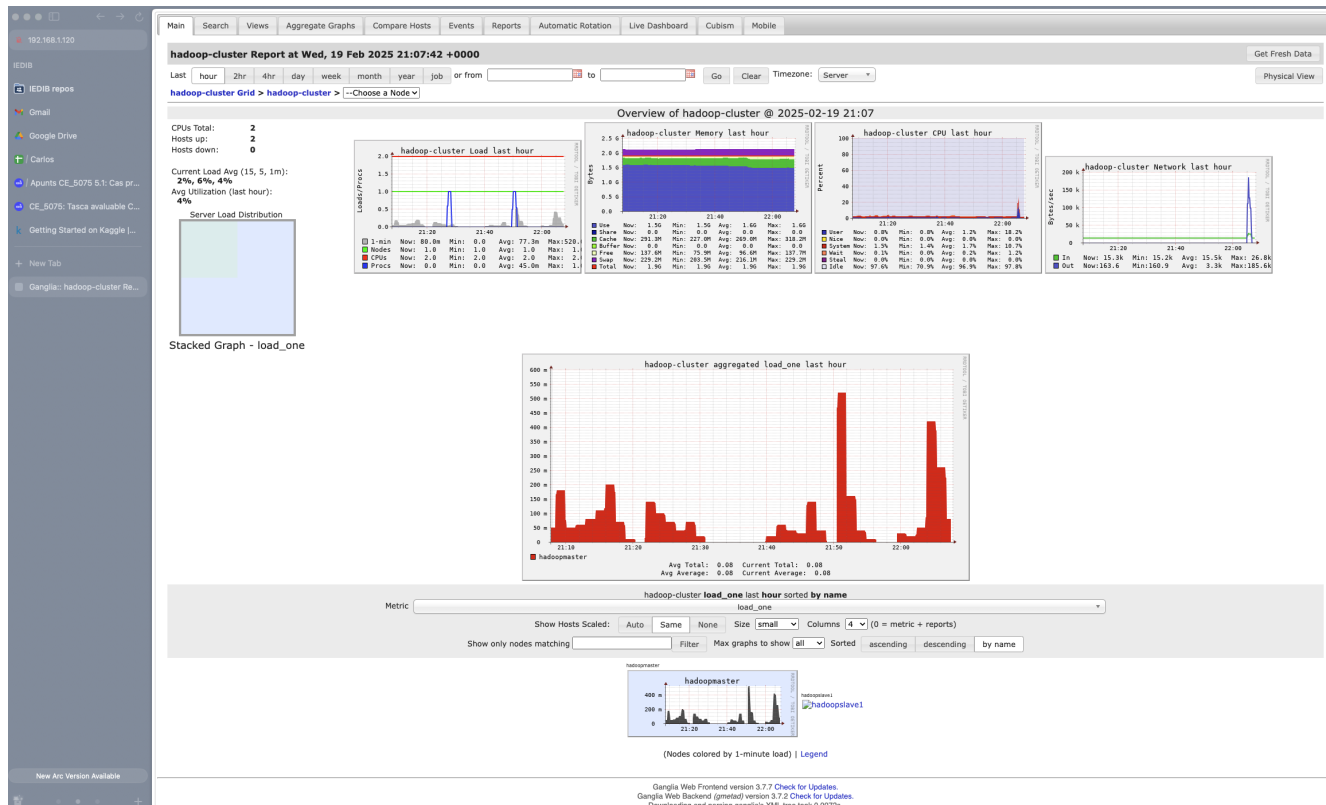
History

Como se puede ver en las imágenes anteriores, el trabajo Pig se puede ver reflejado en el panel, tanto cuando se ejecuta como cuando finaliza, el trabajo ejecutado.

Apartado 3: Ganglia

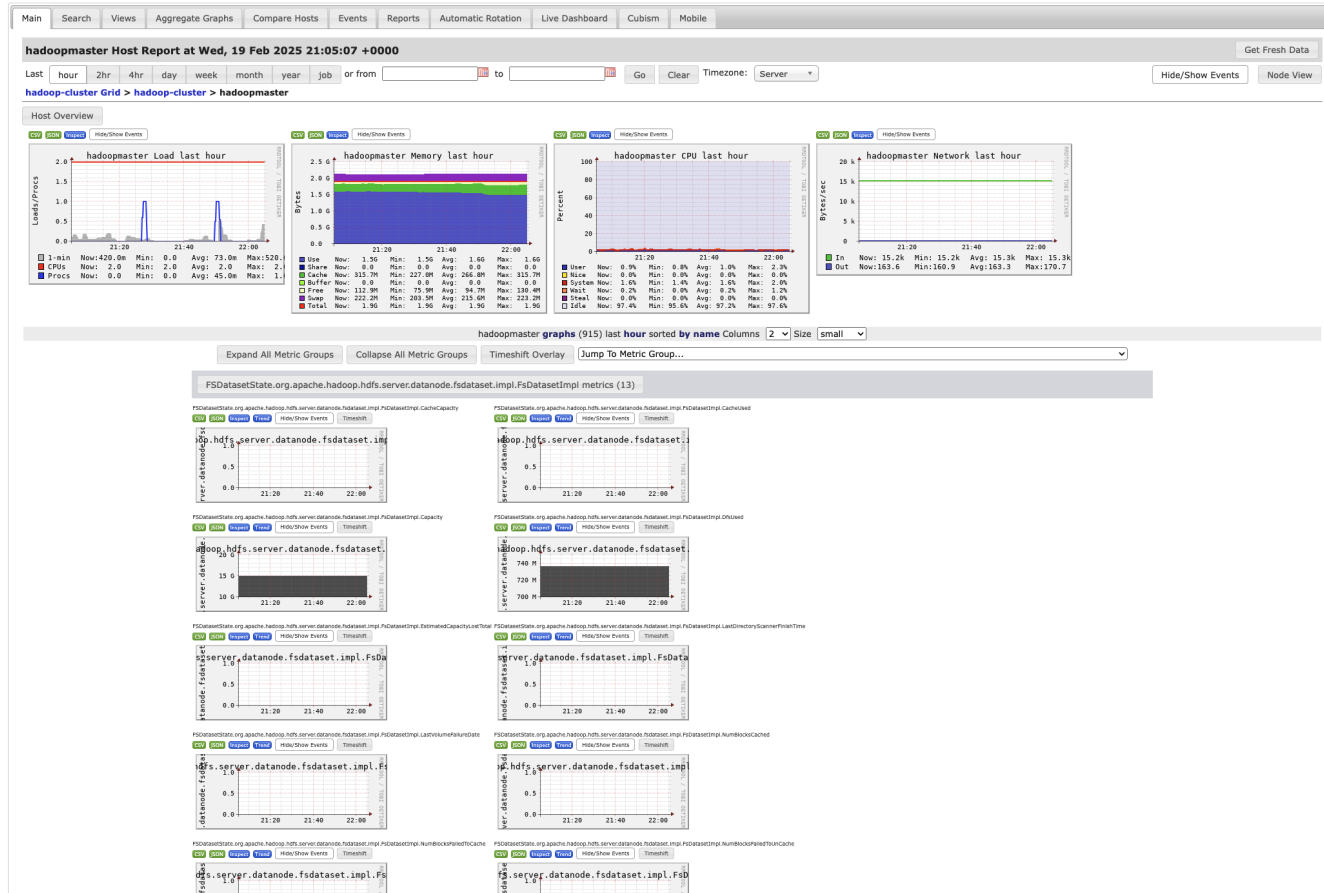
Con el clúster Hadoop de la Entrega 1 (si lo prefieres, basta con dos nodos, un maestro y un esclavo), haz las siguientes capturas de pantalla de la interfaz web de Ganglia e insértelas en tu documento:

- Página inicial de la interfaz web de Ganglia. ¿Cuál ha sido el uso máximo de memoria en la última hora (en MB)? ¿Y cuál ha sido el uso más alto de CPU por el sistema en la última hora (en %)?
- Gráficas de la métrica `cpu_idle` en los 3 nodos del clúster.

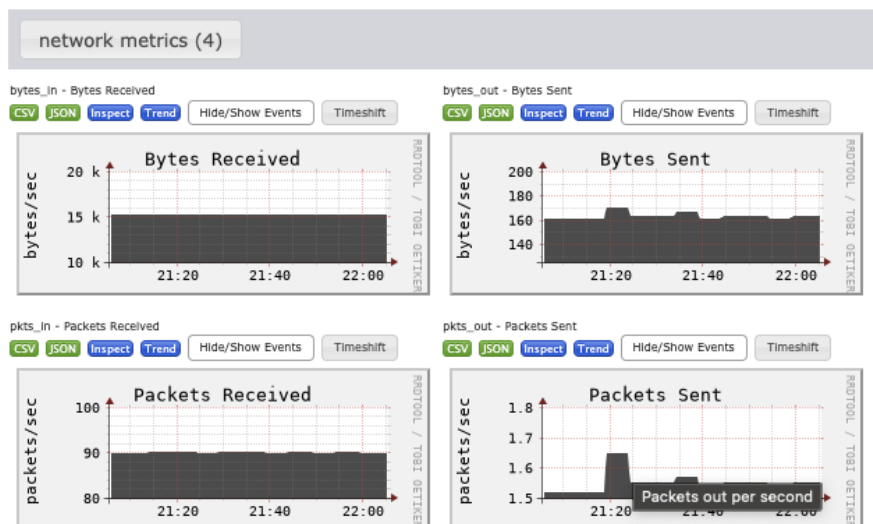


Estas estadísticas se pueden ver en el gráfico central izquierdo (*hadoop-cluster Memory last hour* para la memoria) y para el central derecho (*hadoop-cluster CPU last hour* para la CPU), siendo las métricas de 1.6GB y de 18.2% respectivamente.

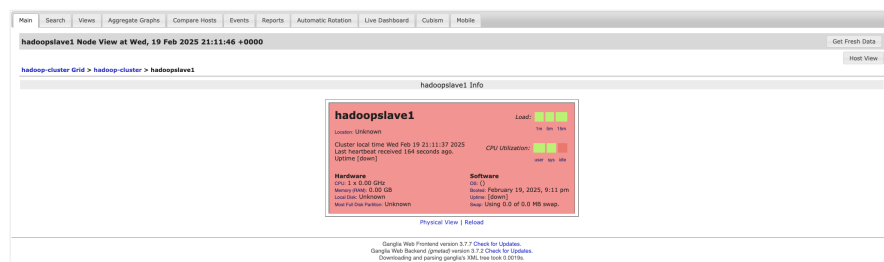
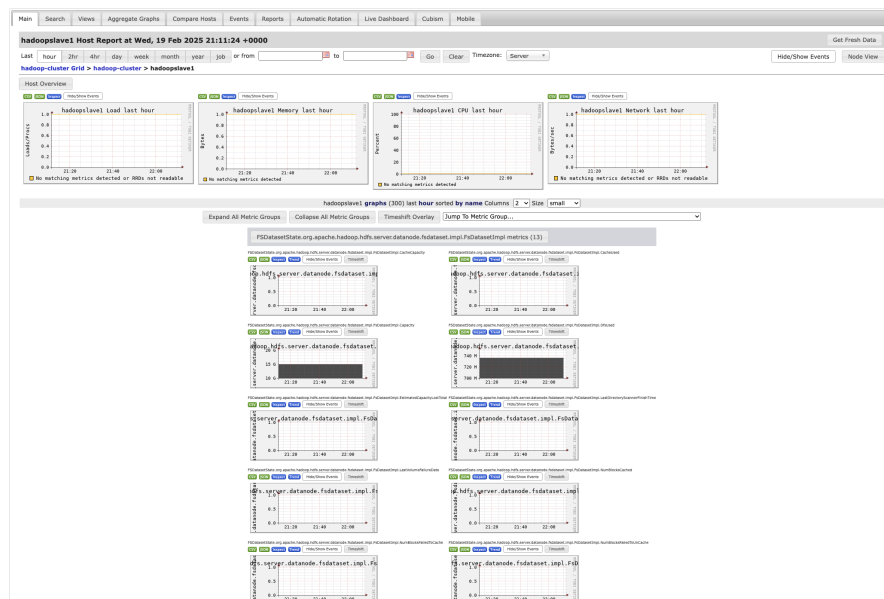
- Visión del host maestro.



- Gráficas de las métricas de red del host maestro.



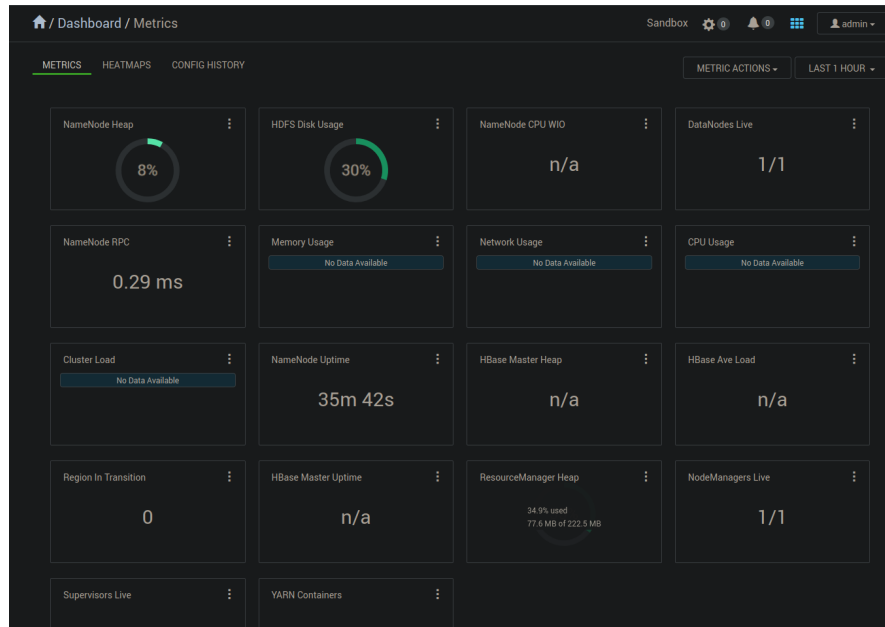
Carlos Sánchez Recio



Apartado 4: Apache Ambari

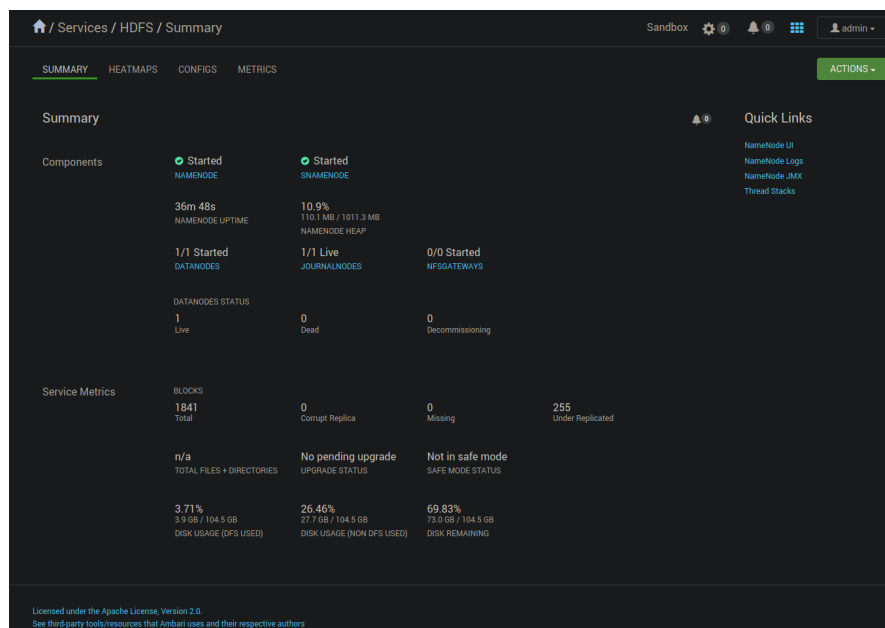
En el HortonWorks Sandbox HDP (o cualquier otra máquina con Hadoop y Apache Ambari instalados), haz las siguientes capturas de pantalla de la interfaz web de Ambari e insértelas en tu documento:

- Vista general del Dashboard. ¿En qué porcentaje están el heap de NameNode y el uso de disco HDFS?



En el momento de la captura de pantalla del dashboard estaban a un 8% y 30% respectivamente.

- Resumen de las métricas del servicio HDFS.



- Vista de las alertas del sistema.

Alerts

Status	Alert Definition Name	Service	Last Status Changed	State
CRIT	HBase Master Process	HBase	6 years ago	Enabled
CRIT	HBase RegionServer Process	HBase	6 years ago	Enabled
CRIT	Storm Web UI	Storm	6 years ago	Enabled
CRIT	Supervisor Process	Storm	6 years ago	Enabled
CRIT	Nimbus Process	Storm	6 years ago	Enabled
CRIT	DRPC Server Process	Storm	6 years ago	Enabled
CRIT	Druid Broker Process	Druid	6 years ago	Enabled
CRIT	Druid Historical Process	Druid	6 years ago	Enabled
CRIT	Druid Overlord Web UI	Druid	6 years ago	Enabled
CRIT	Druid Coordinator Web UI	Druid	6 years ago	Enabled

Items per page: 10 1 - 10 of 92

Licensed under the Apache License, Version 2.0.
See third-party tools/resources that Ambari uses and their respective authors