

Tarea evaluable

CE_5075 5.1

Big data aplicado



Índice

APARTADO 1 - Integridad de los datos	2
APARTADO 2 - Herramientas internas de monitoreo de Hadoop	3

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 usado en entregas anteriores.

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

A continuación, realiza una pequeña modificación en el archivo quijote.txt y vuelve a generar las tres sumas de verificación. Toma otra captura de pantalla donde se vean las tres.

Indica si, después de hacer el cambio, las sumas de verificación han cambiado o no.

```
cloudera@quickstart.cloudera ~ (0.678s)
md5sum quijote.txt
a11a03ab70fead939d3842c3d5a8bb64 quijote.txt

cloudera@quickstart.cloudera ~ (0.324s)
sha256sum quijote.txt
a67c1e3fefac58d0d2ee9f7c11d3790d4d0bc772115d5a77b964d733e9c1404d quijote.txt

cloudera@quickstart.cloudera ~ (0.14s)
sha512sum quijote.txt
f39ac8fb273e9db8ac51ce1b3dfc27b88f2f3ddd67d260bd792987f7664f24ea855c829361e86
e8a9c5ab581ce74d3b83a477dd97063178d9cf9f4d9494341d3 quijote.txt
```

```
GNU nano 2.0.9 File: quijote.txt

DON QUIJOTE DE LA MANCHA
Miguel de Cervantes Saavedra

Esto es una modificacion del Quijote original

PRIMERA PARTE
CAPITULO 1: Que trata de la condicion y ejercici
En un lugar de la Mancha, de cuyo nombre no quiero
```

```

cloudera@quickstart.cloudera ~ (0.134s)
md5sum quijote.txt
838d726b1656c9c01a90abc1a2f85984  quijote.txt

cloudera@quickstart.cloudera ~ (0.198s)
sha256sum quijote.txt
783de8be6501fe9dddb3e3a1c16bf092d52b4712bb46e19abcb980fd4cfe7a6  quijote.txt

cloudera@quickstart.cloudera ~ (0.128s)
sha512sum quijote.txt
78e9e2fe0fdaee73f9e62be6f9b53139dd473a55fe51b0c7b53255a798cb409b8af5b46132ad5
84237ba78e2d12fde1b7aeeba3001ec7ce76b59ccfc50d1ee4c  quijote.txt

```

Podemos comprobar que las sumas de verificación sí que han cambiado después de añadir esa línea de texto en “quijote.txt”.

APARTADO 2 - Herramientas internas de monitoreo de Hadoop

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

Overview 'quickstart.cloudera:8020' (active)

Started:	Wed Feb 19 08:16:48 -0800 2025
Version:	2.6.0-cdh5.12.0, rdba647c5a8bc5e09b572d76a8d29481c78d1a0dd
Compiled:	Thu Jun 29 04:32:00 -0700 2017 by jenkins from Unknown
Cluster ID:	CID-0faeffb7-c487-4469-ab9d-c35dc4d1b32f
Block Pool ID:	BP-1037806989-127.0.0.1-1500499492170

Summary

Security is off.

Safemode is off.

1,378 files and directories, 1,065 blocks = 2,443 total filesystem object(s).

Heap Memory used 121.61 MB of 265.5 MB Heap Memory. Max Heap Memory is 889 MB.

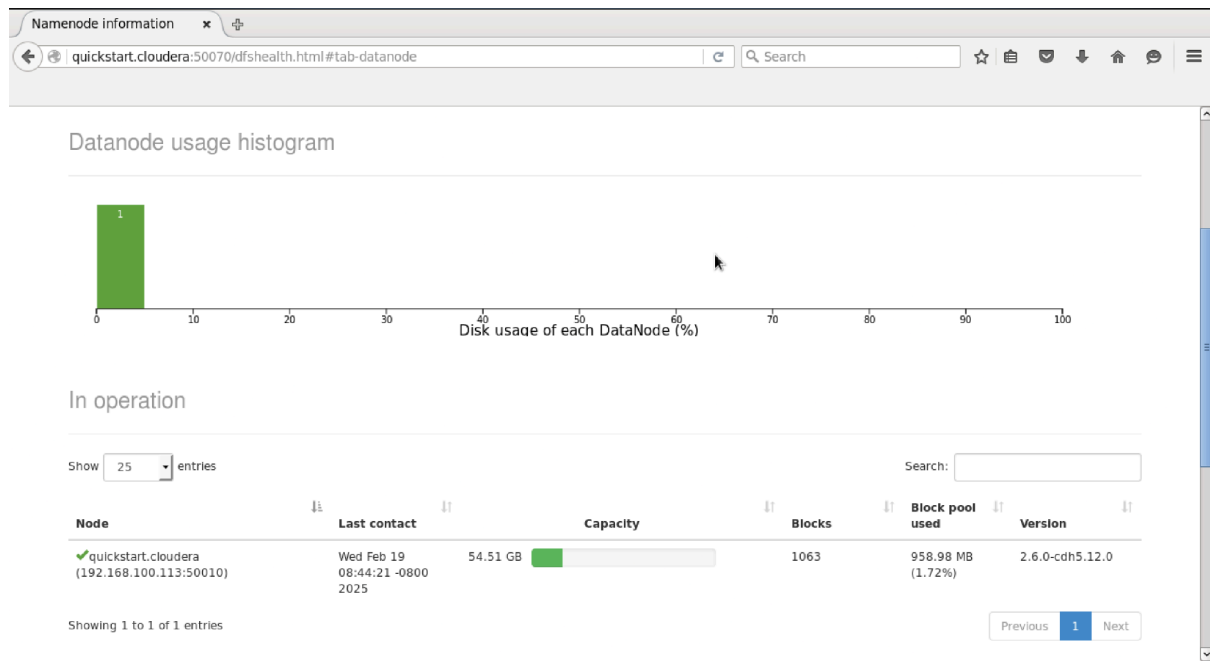
Non Heap Memory used 35.63 MB of 36.94 MB Committed Non Heap Memory. Max Non Heap Memory is 130 MB.

Configured Capacity:	54.51 GB
DFS Used:	958.98 MB (1.72%)
Non DFS Used:	8.26 GB
DFS Remaining:	42.29 GB (77.57%)
Block Pool Used:	958.98 MB (1.72%)
DataNodes usages% (Min/Median/Max/stdDev):	1.72% / 1.72% / 1.72% / 0.00%
Live Nodes	1 (Decommissioned: 0, In Maintenance: 0)
Dead Nodes	0 (Decommissioned: 0, In Maintenance: 0)
Decommissioning Nodes	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	Wed Feb 19 08:16:48 -0800 2025
Last Checkpoint Time	Wed Feb 19 07:53:31 -0800 2025

- **HDFS NameNode:** Accede a <http://quickstart.cloudera:50070> y captura las secciones Overview y Summary. ¿Cuánto espacio (y porcentaje) del DFS está siendo utilizado?

Se están utilizando 958.98MB (1.72%) del DFS



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

Actualmente el DataNode contiene 1063 bloques

```
{
  "beans" : [ {
    "name" : "java.lang:type=Memory",
    "modelerType" : "sun.management.MemoryImpl",
    "HeapMemoryUsage" : {
      "committed" : 278396928,
      "init" : 95887616,
      "max" : 932184064,
      "used" : 148296808
    }
  } ]
}
```

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

Accediendo a "quickstart.cloudera:50070/jmx" vemos que el máximo de memoria para usar en el Heap es 932184064 bytes

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
0	0	0	0	0	0 B	8 GB	0 B	0	8	0

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes
1	0	0	0	0	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	0 B	0 B	0 B	0	0	0

Showing 0 to 0 of 0 entries

- **YARN ResourceManager:** Accede a <http://quickstart.cloudera:8088/cluster> y captura la pantalla.

NodeManager information

Total Vmem allocated for Containers	16.80 GB
Vmem enforcement enabled	false
Total Pmem allocated for Container	8 GB
Pmem enforcement enabled	true
Total VCores allocated for Containers	8
NodeHealthyStatus	true
LastNodeHealthTime	Wed Feb 19 10:26:58 PST 2025
NodeHealthReport	
Node Manager Version:	2.6.0-cdh5.12.0 from dba647c5a8bc5e09b572d76a8d29481c78d1a0dd by jenkins source checksum 90d315b02cac6c524a5f659051adb221 on 2017-06-29T11:40Z
Hadoop Version:	2.6.0-cdh5.12.0 from dba647c5a8bc5e09b572d76a8d29481c78d1a0dd by jenkins source checksum 7c45ae7a4592ce5af86bc4598c5b4 on 2017-06-29T11:32Z

- **YARN NodeManager:** Accede a <http://quickstart.cloudera:8042/node> y captura la pantalla.

```

cloudera@quickstart.cloudera ~ (3m 18.70s)
pig -x mapreduce quijote.pig

(cerrasen,1)
(cervices,1)
(ce?dula,3)
(chapi?n,1)
(chicoria,1)
(chimenea,2)
(ciani?s,2)
(ciencias,2)
(ciertas?,1)
(cintura.,1)
(cintura;,1)
(cin?ese,1)
(cin?o?,1)
(cipre?s,1)
(ciriales,1)
(ciudades,5)
(claridad,5)
(clarines,1)
(claustro,2)
(clavaron,1)
(clerical,1)
(cobardes,2)

```

The screenshot shows the Hadoop All Applications web interface. The top navigation bar includes 'All Applications' and a search bar. The main content area displays 'Cluster Metrics' and 'Cluster Nodes Metrics'. Below these, there is a table of applications. The application 'application_1739981964863_0001' is shown with a status of 'FINISHED' and 'SUCCEEDED'.

Cluster Metrics										
Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved
1	0	0	1	0	0 B	8 GB	0 B

Cluster Nodes Metrics										
Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes				
1	0	0	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	...
0	0	0	0	0	0	0	0 B	0 B	0 B	...

ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus	...
application_1739981964863_0001	cloudera	PigLatin:quijote.pig	MAPREDUCE	root.cloudera	Wed Feb 19 10:42:59 -0800 2025	Wed Feb 19 10:45:23 -0800 2025	FINISHED	SUCCEEDED	N/A

- **Ejecuta un trabajo en Pig** y captura nuevamente el YARN ResourceManager. **Indica si ha habido algún cambio** respecto a la captura anterior del ResourceManager.

Así es, después de ejecutar el script de Pig llamado quijote.pig (desarrollado en temas anteriores del curso), el estado del YARN ResourceManager ha cambiado. Ahora, en el historial, aparece registrado el script de Pig que he ejecutado. Esto indica que el trabajo ha sido procesado correctamente y está reflejado en el sistema de gestión de recursos de YARN.