



M013: Projecte de Desenvolupament d'Aplicacions Multiplataforma



Implantar una solució de BigData amb un clúster Apache Hadoop



Curs: 2019-20

CFGS: DAM2

Alumne: Arnau Subirós Puigarnau

Data: 02-06-2020



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

INDEX

pàgina

1.	Introducció al projecte	4
1.1.	Introducció a Big Data	
1.1.1.	Tecnologia Big Data :Apache Hadoop	
1.2.	Apache Hadoop	
1.2.1.	HDFS	
1.2.2.	Processos	
1.2.3.	Altres productes implicats (ecosistema Hadoop)	
1.2.3.1.	HDFS	
1.2.3.2.	Processos	
1.2.3.3.	Altres productes implicats (ecosistema Hadoop)	
1.2.4.	Distribucions Hadoop	
2.	Introducció al Clúster Hadoop	16
2.1.	Tamany del clúster	
3.	Clúster Hadoop (CentOS 6.10)	18
3.1.	Cluster Hadoop pseudodistribuit (1 node).....	19
3.1.1.	Requisits del hardware	
3.1.2.	Version del software (per tema de compatibilitats)	
3.1.3.	Instal·lació de CentOS 6.10	
3.1.4.	Instal·lació i configuració de Hadoop 2.10.0	
3.1.4.1.	Configuració de les dades	
3.1.4.1.1.	Exemple pràctic amb HDFS	
3.1.4.1.2.	HDFS Snapshot	
3.1.4.2.	Configuració dels processos (Yarn)	
3.1.4.2.1.	Exemples Mapreduce	
<input type="checkbox"/> Utilitzant el jar d'exemples Mapreduce		
<input type="checkbox"/> Utilitzant codi Java		
<input type="checkbox"/> Utilitzant codi Python		
3.2.	Cluster Hadoop (diversos nodes).....	56
3.2.1.	Configuracions prèvies	
3.2.2.	Configuració de la xarxa	
3.2.3.	Configuració de SSH	
3.2.4.	Configuració del clúster	
3.2.5.	Exemples Mapreduce	
3.2.5.1.	Exemple 1 - Llençar un procés Mapreduce contra el clúster	
3.2.5.2.	Exemple 2 - Llençar un procés Mapreduce contra el clúster	
3.2.5.3.	Exemple 3 - Llençar un procés en streaming amb comandos Shell de Linux.	
3.2.6.	Yarn Schelduler	
3.2.6.1.	Configuracio de Yarn Schelduler	
3.2.6.1.1.	Exemple : llençar un job	
3.2.7.	Ecosistema Hadoop	
3.2.7.1.	Instal·lació i configuració de Hive	
3.2.7.2.	Instal·lació i configuració de Spark	
<input type="checkbox"/> Exemple amb Spark Shell		
<input type="checkbox"/> Exemple amb Pyspark		
<input type="checkbox"/> Exemple amb Spark Shell interactuant amb HDFS		

Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

3.3.	Cluster Hadoop utilitzant Ambari (desde 0).....	126
3.3.1.	Connexions entre el clúster :Xarxa interna	
3.3.2.	SSH : Claus públiques entre els nodes	
3.3.3.	Instal·lació i configuració del servidor Ambari	
3.3.3.1.	Configuració de Hive	
3.3.3.2.	Visualització desde Windows 10	
4.	Annexe :Cluster Hadoop amb Rasperri Pi Desktop(Testing).....	147
4.1.	Intent amb Raspberry Pi Desktop (Debian 9 i Debian 10)	
4.2.	Intent de S.O.hibrid (master: Centos i 2 slaves: Raspberry Pi Desktop)	
4.3.		
5.	Conclusions finals.....	151
6.	Bibliografia.....	152



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

1. Introducció al projecte

Inicialment aquest projecte consistia amb un clúster Hadoop utilitzant 3 Raspberry Pi. Malauradament degut al confinament del COVID-19 no ha pogut ser així, i s'ha optat per buscar alternatives.

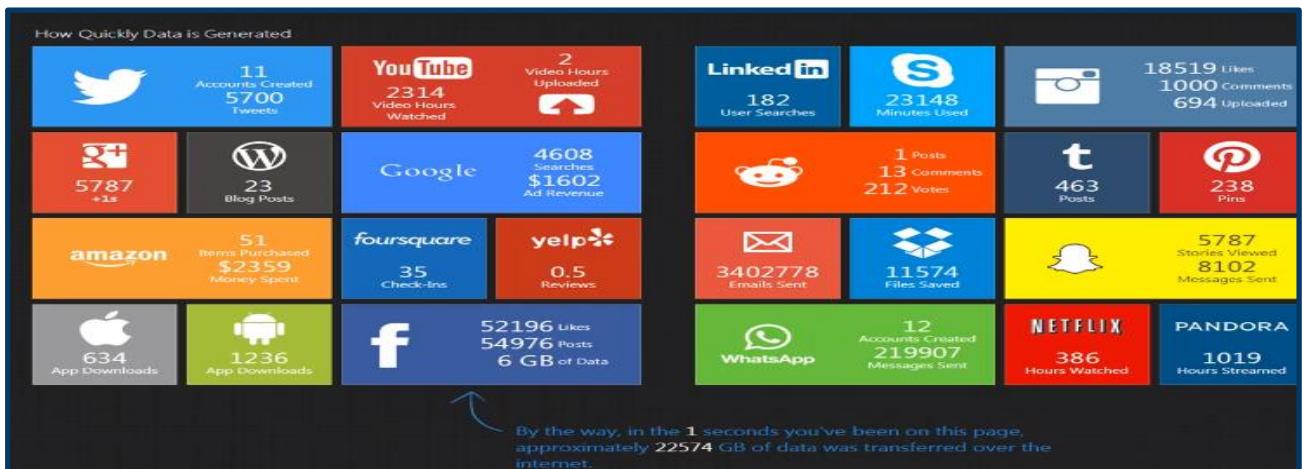
- Com que el portàtil té 16GB RAM s'ha pensat en crear 3 màquines virtuals de 4GB de RAM cadascuna ,deixant limitadament 4GB al host amfitrió (pot funcionar mentre el host carregui mol la memòria, ja que es bloquejaria).
 - En aquesta opció primer es fa fer amb sistema operatiu CentOS, després es volia fer amb Raspbian (simulant la Raspberri Pi) però té arquitectura ARM incompatible amb VirtualBox i VMWare en què m'he instal·lat Raspberry Pi Desktop (una distribució de Debian amb interfície de Raspberry Pi dissenyat per poder-ho emular en maquines virtuals, però amb limitacions). S'ha fet proves però s'ha trobat moltes incompatibilitats de sistema i problemes de diferents repositoris.
- També s'ha intentat fer-ho amb Dockers. Primer des de el sistema operatiu actual (Window 10). Que com que no disposo de la llicència Windows 10 Pro, Dockers s'havia utilitzat Dockers Toolbox on el servidor Dockers estaria ubicat en una màquina de Linux (una versió lleugera)en VirtualBox. S'ha fet proves amb altres sistemes operatius (Ubuntu 20.04 core i Centos 6) instal·lats en maquines virtuals. Però en intentar crear un clúster de diverses nodes, no s'ha aconseguit i s'ha descartat l'opció.
- Hem centrat en el sistema operatiu Centos6. On s'ha creat 3 tipus de clúster
 - El bàsic per practicar on només hi havia 1 node (màster-slave)
 - Un clúster hadoop d'1 master i 2 slaves o s'ha ampliat l'ecosistema Hadoop, instal·lant i configurant Hive,Spark i Ambari(no operatiu)
 - Clúster Hadoop d'un 1 master i 2 slaves utilitzant Ambari (des de 0), ja que en l'anterior tot i comentar les variables d'entorn, com que ja hi havia configuracions estableties, generava errors.

Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

1.1. Introducció a Big Data

Big Data és la convergència d'enormes quantitats de dades tan estructurades (ex: Base de dades relacionals en forma de taules Oracle, MySQL, etc.) com sense estructurar (dades de tweets, de Facebook, de blogs...)

- Es creen *Petabytes*¹ de dades diàriament
 - Xarxes socials
 - Mòbils
 - Sensors
 - Dades científiques ...



Hi ha una enorme quantitat de dades dels quals volem processar i extreure la informació. Quan parlem de Big Data parlem de les 3 V:

- Volum (Terabytes, Petabytes...)
- Varietat (d'informació: estructurat, no estructurat)
- Velocitat

Com que les tecnologies tradicionals no poden fer front aquesta quantitat d'informació, és necessari utilitzar noves tecnologies. De grans servidors a entorns distribuïts.

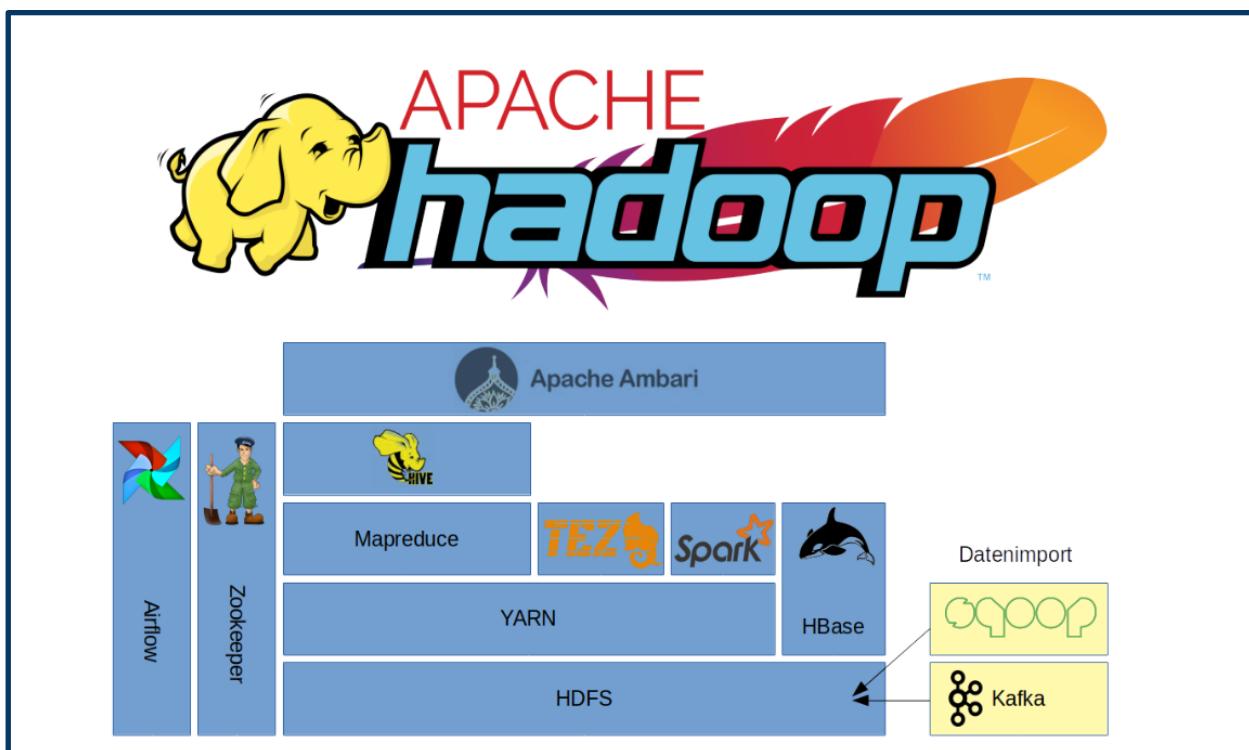
¹ Nota: 1 Petabyte són 1.000.000 GB o dit d'una altra manera, 1e+6GB



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

1.1.1. Tecnologia Big Data : Apache Hadoop

Apache Hadoop és un framework (basat en Java) que permet el processament distribuït de grans conjunts de dades en grups de computadores utilitzant models de programació simples. Està dissenyat per a escalar des de servidors individuals a milers de màquines, cadascuna de les quals ofereix computació i emmagatzematge local.





Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

1.2. Apache Hadoop

Apache Hadoop és un entorn distribuït on el seu “core” està format per 2 components bàsics:

- *dades*
 - *processos*
-
- És un sistema de tipus clúster que es pot escalar d’acord amb les nostres necessitats(de pocs nodes a milers de màquines), a més es pot utilitzar un hardware relativament econòmic.
 - Implementa processament en paral·lel a través de nodes de dades en un sistema de fitxers distribuït.
 - Està dissenyat per executar-se en servidors de baix cost i disposa una gran tolerància a les fallades.
 - Subministra llibreries Open Source per la computació distribuïda utilitzant diversos components:
 - *Hadoop Common*
 - *MapReduce*
 - *Hadoop Distributed File System (HDFS)*

1.2.1. HDFS

HDFS és el sistema d’arxius que utilitza Hadoop com a sistema d’emmagatzematge per als arxius.

- És un sistema distribuït, escalable i portàtil
- Permet connectar nodes (màquines de baix cost) les quals formen els clústers, en aquests nodes s’emmagatzemarà la informació de forma distribuïda
- Les aplicacions o comandos s’executaran seguint el model MapReduce. Pel fet que Hadoop tracta amb grans volums de dades es veurà afectat per la velocitat de transferència, sigui de xarxa o de velocitat d’escritura en disc.



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

- Fa possible que els càlculs s'executin en la màquina on resideixen les dades en lloc de moure les dades on es realitzen la petició d'aquests.

Característiques generals dels Hadoop HDFS

- Una escriptura i moltes lectures.
- Tolerància a fallades i recuperació automàtica en cas de fallada.
- Coherència en les dades.
- Replicació de dades.
- Processament de les dades on estan emmagatzemats, en lloc de moure'ls.
- Facilitat de processament a través de sistemes heterogenis de maquinari i sistema operatiu.
- Escalabilitat per a processar i emmagatzemar grans quantitats de dades.

HDFS gestiona l'emmagatzematge en el clúster, dividint els fitxers en blocs i emmagatzemat copies duplicades a través dels nodes (mida del bloc 128 MB).

Per defecte es replicarà en 3 nodes diferents.

En crear un Clúster Hadoop hi haurà:

- Un node que actuarà com a “màster” de dades. Només té metadades i gestionarà les dades dels nodes “slaves”.
- La resta de nodes són “slaves” que contenen les dades.

Disposa d'uns fitxers (metadades) que gestionen els canvis que es produueixen al clúster de HDFS que són:

- edits_000xxxx → Son els canvis que es produueixen dins la base de dades de HDFS
- edits_inprogress_xxxx → A on s'escriuen les dades en temps real
- fsimage_00000xxxx → Conte una còpia d'un moment en el temps en el sistema de fitxers

Quan s'arrenca el **HDFS** es llegeix l'últim **fsimage** que es carrega a la memòria del **Namenode**, no es toca sinó que es construeix el fitxer **edits_inprogress** (es canvien les dades en memòria i no al **fsimage** original per guanyar rendiment)



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

- Cada cert temps se sincronitza amb el **Secondary Namenode** (la qual cosa s'ha anat guardant amb el que hi havia en aquella imatge)
- Aquests fitxers no es toquen

```
[hadoop@node1 current]$ ls -ls
total 2080
  4 -rw-rw-r-- 1 hadoop hadoop      42  5 mai 10:52 edits_00000000000000000001-00000000000000000002
  4 -rw-rw-r-- 1 hadoop hadoop      42  5 mai 11:53 edits_00000000000000000003-00000000000000000004
1024 -rw-rw-r-- 1 hadoop hadoop 1048576  5 mai 11:53 edits_00000000000000000005-00000000000000000005
1024 -rw-rw-r-- 1 hadoop hadoop 1048576  5 mai 12:00 edits_inprogress_00000000000000000006
  4 -rw-rw-r-- 1 hadoop hadoop     325  5 mai 10:39 fsimage_00000000000000000000
  4 -rw-rw-r-- 1 hadoop hadoop      62  5 mai 10:39 fsimage_00000000000000000000.md5
  4 -rw-rw-r-- 1 hadoop hadoop     325  5 mai 11:53 fsimage_00000000000000000004
  4 -rw-rw-r-- 1 hadoop hadoop      62  5 mai 11:53 fsimage_00000000000000000004.md5
  4 -rw-rw-r-- 1 hadoop hadoop       2  5 mai 12:00 seen_txid
  4 -rw-rw-r-- 1 hadoop hadoop    215  5 mai 10:39 VERSIÓN
[hadoop@node1 current]$
```

1.2.2. PROCESSOS

És el component de processament de Hadoop. Consisteix en un framework de programació (llibreries i entorn d'execució) que treballa sobre HDFS i es basa en l'ús de dos tipus de funcions:

- **Map** – Divideix la tasca d'entrada en subtasques i les executa entre diferents nodes.
- **Reduce** – La funció “Reduce” recull les respostes de les subtasques en cada subnode i les combina i agrupa per a obtenir la resposta final.

Resource Manager(segons les versions de Hadoop)

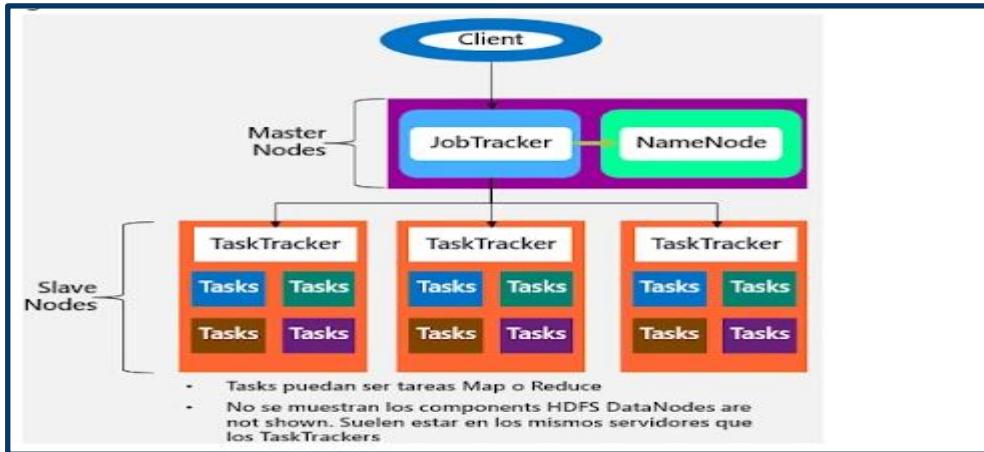
- **Hadoop 1.x :** Utilitza Map Reduce V1

Cada màquina d'un clúster Hadoop té un servidor MapReduce que es diu TaskTracker. Al seu torn, hi ha un gestor de Jobs per cada clúster, el JobTracker, que s'encarrega de dividir cada procés a realitzar en subprocessos, i distribuir la computació d'aquests subprocessos entre diferents màquines del clúster, enviànt als TaskTrackers de cadascuna d'elles el job que li correspon realitzar.

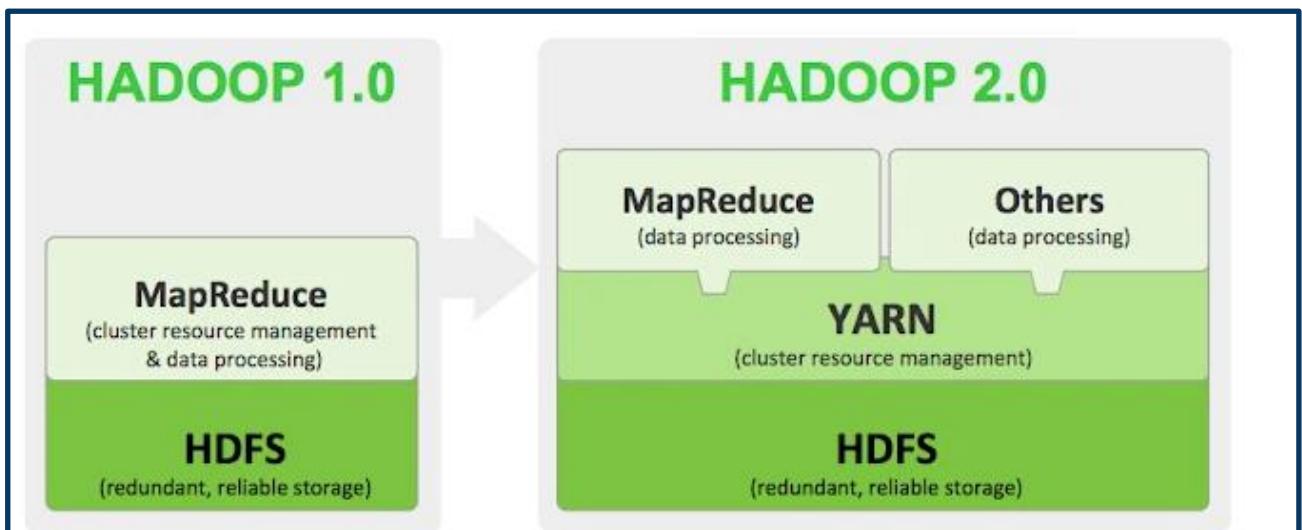
Nota: si queia el Job Tracker o se saturava, el clúster deixava de funcionar.



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020



- **Hadoop 2.x**²: A partir de l'any 2011 amb la versió 2 incorpora YARN (Yet Another Resource Manager), un gestor del clúster millorat que separa les tasques que realitzava el Job Tracker.



En la versió 2 es va desenvolupar una nova versió de **MapReduce**,³ molt diferent de l'anterior, ja que va deixar d'estar integrada en el Core (s'executa com una aplicació independent). Gràcies a això, ara és possible l'accés al sistema de fitxers de Hadoop des d'altres entorns de programació i execució

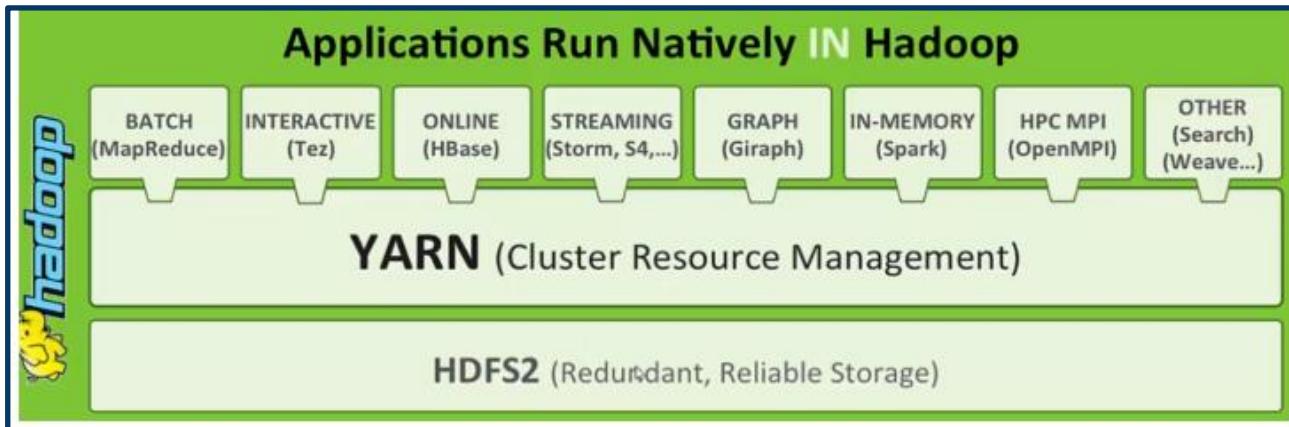
² És la versió que he utilitzat

³ És un algoritme de processament de dades que implementa processos en paral·lel. De forma simple distribueix les tasques a través dels nodes d'un clúster executant la funció "map".

- La funció "map" estudia el problema, ho divideix en parts i les envia a diferents màquines perquè totes les parts puguin executar-se de forma simultanea
- Els resultats d'aquest procés paral·lel són recollits i es distribueixen a través dels diferents serveis que executen una funció "reduce" que agafa els resultats de les parts i les recompta per obtenir una resposta simple



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020



- **Hadoop 3.x** : La primera versió 3.0.3 va aparèixer l'any 2018. Per a un clúster, la gestió de recursos utilitza YARN, amb totes les característiques.

1.2.3. Altres productes implicats (ecosistema Hadoop)

1.2.3.1. Hbase: Una base de dades orientada a valors/clau que s'executa sobre HDFS. Aquesta base de dades no relacional és de tipus columnar i de codi obert. Distribuïda i escalable utilitzat en l'emmagatzemant de Big Data

1.2.3.2. Hive : Permet accedir a HDFS com si fos una Base de dades

Hive implementa una variant al SQL, anomenada HQL (Hive QL). Quan s'executen consultes HQL, Hive converteix la consulta HQL a un treball MapReduce que és executat per a obtenir les dades.

- Simplifica el desenvolupament i gestió amb Hadoop
- HIVE té un metastore que és un repositori central per les metadades de Hive.

Hi ha 3 tipus HIVE:

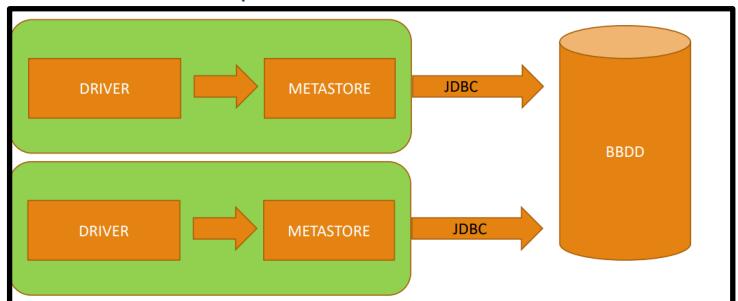
- **Encastat**: Tot està a la mateixa maquina virtual JAVA , però només ho pot utilitzar 1 usuari)



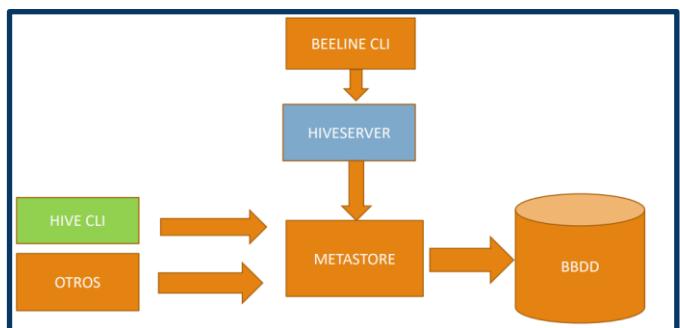
Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020



- **Local:** Més potent que el metode "Encastat". JVM executa el programa + el servei de metastore i després via JDBC accedim a la base de bades, per exemple MySQL. Aquest mètode ens permet tenir múltiples usuaris però per defecte tot està a la mateixa maquina .



- **Remot** On tenim un servei metastore que es accedit per un servidor anomenat Hive Server. Aquest Hive Server pot ser accedit per determinats clients utilitzant beeline o hive cli.



1.2.3.3. **Pig:** Llenguatge de alt nivell (de tipus scripting) per gestionar els fluxes de dades i execucions d'aplicacions sobre Hadoop.

- Treballa amb paralel·lel i permet gestionar gran quantitat d'informació
- Es un compilador que genera comandos MapReduce



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

1.2.3.4. **Mahout:** Entorn d'aprenentatge de màquines implementat en Hadoop.

- Machine Learning
- Data Mining

1.2.3.5. **Zookeeper :** és un servei per mantenir la configuració , coordinació i aprovisionament d'aplicacions distribuïdes .No només funciona amb Hadoop(no és un producte de Hadoop, sino de Apache), però és molt útil en aquesta arquitectura

- Elimina la complexitat de la gestió distribuïda en la plataforma
- Funciona amb Java
- Indispensable si es vol utilitzar Apache Ambari

1.2.3.6. **Sqoop:**Eines dissenyades per transferir dades massives des de Hadoop a altres entorns com base de dades relacionals (gestors de dades estructurades).

- Sqoop ofereix connectors per integrar Hadoop amb altres sistemes com Oracle o SqlServer

1.2.3.7. **Flume:** és un servei distribuït per distribuir, afegir i recollir grans quantitats de informació (especialment registres) com fitxers de logs, paquets de Twitter,etc.

- Té una arquitectura de tipus streaming amb un flux de dades molt potent i personalitzable

1.2.3.8. **Cassandra:** és un sistema de gestió de base de dades NoSQL distribuït de tipus columnar.

- Dissenyat per a manejar grans quantitats de dades en molts servidors bàsics, proporcionant alta disponibilitat sense un sol punt de falla.

1.2.3.9. **Ambari :** és una eina molt potent que serveix per instal·lar, configurar, mantenir i monitoritzar Hadoop des d'una plataforma gràfica.

- Podrem instal·lar Hadoop des de 0, configurar-ho, gestionar-lo i monitorar-lo
- Simplifica enormement la gestió d'Hadoop

1.2.3.10. **HUE :** És un entorn que ens permet treballar de forma més amigable amb un entorn Hadoop. És una eina gràfica amb una sèrie de característiques

- Té editor per treballar amb Hive,Impala,Pig

Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

- Té un Dashboard
- Te un Scheduler (Planificador)
- Es un navegador de components
- Podem navegar pels directoris HDFS
- Accedir al metastore de Hive

1.2.3.11. **Spark :** És un motor eficient de processament de dades a gran escala.

- Implementa processament en temps real al contrari que MapReduce
- Es molt més ràpid que MapReduce
- Treballa de forma massiva en memòria
- Pot funcionar stand-alone (no necessita Hadoop)
- Spark pot treballar tant en Java,Scala ,Python i R
- Permet l'anàlisis de gran quantitat de dades
- Integra diferents entorns com Base de Dades NoSQL, Real Time, machine learning, anàlisis de grafos
- Tambe podríem mesclar aplicacions Spark i MapReduce para batch i Real Time.
- Soporta diferents fonts de dades
 - hive
 - json
 - cassandra
 - csv
 - rdbms, etc
- Spark constitueix un “core” i un conjunt de llibreries com a més importants:
 - Spark SQL
 - Spark Streaming
 - Spark Mlib
 - GraphX

1.2.3.11.2. **Spark CORE:** És el motor base per al processament en escala i distribuït.

- Esta programat amb Scala però hi ha APIS per Python, JAVA i R
- S'encarrega de:
 - Gestió de la memòria
 - Recuperació envers les fallades
 - Planificació, distribució de treballs del clúster
 - Motorització del treball



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

Accedeixes als sistemes d'emmagatzemant

- Utilitza una estructura de dades especial denominada RDD:

- Realitzen processos
 - Són col·leccions de registres immutables i particionades que a més a més poden ser utilitzades en paral·lel
 - Poden contenir qualsevol classe d'objectes de Java, Python, Scala o personalitzats

1.2.4. Distribucions Hadoop

Hi ha diferents empreses que ofereixen solucions empaquetades per Hadoop, entre les més importants hi ha:

Distribucions Hadoop	
Nom	Producte
Amazon Web Service (AWS)	Amazon Elastic MapReduce
Cloudera	Cloudera Enterprise
Hortonworks	Hortonworks Data Platform
Intel	Intel Distribution for Apache Hadoop
MapR Technologies	MapR M3 - MapR M7
Microsoft	Windows Azure HDInsight
Pivotal Software	Pivotal HD
Teradata	Teradata Open Distribution Hadoop (TDH)



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

2. Introducció al Clúster Hadoop

Un clúster de Hadoop no és més que un grup de computadores connectades entre si a través de LAN. Ho fem servir per a emmagatzemar i processar grans conjunts de dades.

- Els clústers d' Hadoop tenen una sèrie de maquinari bàsic connectat entre si.
- Es comuniquen amb una màquina d'alta gamma que actua com un mestre.
- Aquests mestres i esclaus implementen la computació distribuïda sobre l'emmagatzematge de dades distribuïdes.

El clúster de Hadoop té una arquitectura “*master-slave*”

➤ **Master** : és una màquina amb una bona configuració de memòria i CPU. Té 2 “demons (serveis executant-se en segon pla) que són:

- **NameNode**
 - Administra l'espai de noms del sistema d'arxius
 - Regula l'accés als arxius per part del client
 - Emmagatzema metadades de dades reals (ruta d'arxius, núm.de blocs, ubicació, etc.)
 - Executa operacions d'espai de noms del sistema d'arxius com obrir, tancar, reanomenar arxius i directoris.
- **Resource Manager**
 - Arbitra recursos entre nodes competidors
 - Realitza un seguiment dels nodes vius i morts

➤ **Esclau** : és una màquina de configuració normal on hi ha 2 demons executant-se en màquines Slave que són:



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

- **DateNode**
 - Emmagatzema les dades
 - Realitza operacions de lectura, escriptura i processament de dades
 - Després de la instrucció d'un mestre, crea, elimina i replica blocs de dades.
- **NodeManager**
 - Executa serveis en el node per a verificar el seu estat i informa el mateix a Resource Manager.
 - Podem escalar fàcilment el clúster d' Hadoop agregant-li més nodes. Per tant, ho diem un clúster amb escala lineal. Cada node agregat augmenta el rendiment del clúster.

2.1. Tamany del clúster

- **Cluster pseudistribuit (1 node)** s'implementa tot en una sola màquina i els clústers de múltiples nodes s'implementen en diverses màquines.
 - Tots els “daemons” com NameNode, DataNode s'executen en la mateixa màquina. En un clúster de Hadoop d'un sol node.
 - Tots els processos s'executen en una instància de JVM. L'usuari no necessita realitzar cap ajust de configuració. L'usuari d' Hadoop només necessita establir la variable JAVA_HOME. El factor predeterminat per al clúster Hadoop de node únic és un.
- **Clúster Hadoop de múltiples nodes**, té una arquitectura “master-slave”

Màquina Master s'executa :

- “daemon” NameNode
- “daemon” ResourceManager
- La màquina ha de ser un servidor potent

Màquina Slave s'executa :

- “daemon”DataNode
- “daemon” NodeManager
- Les màquines poden ser simples i barates
- En múltiples nodes, les màquines “slaves” poden estar presents en qualsevol ubicació, independentment d'ubicació física del servidor mestre



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

3. Clúster Hadoop (CentOS 6)



En aquest apartat mostrarem la part pràctica amb exemples. S'ha utilitzat VirtualBox on s'ha virtualitzat diverses màquines amb S.O CentOS 6

S'ha pensat que era més didàctic i entenedor fer-ho per fases.

- Clúster pseudodistribuït (1 node)
- Clúster real (1 master i 2 slaves) : Exemples pràctics
- Clúster real (1 master i 2 slaves) amb l'ampliació ecosistema Hadoop (Hive, Spark i Ambari) . Exemples pràctics
- Clúster Hadoop (1 master i 2 slaves) utilitzant Ambari (visualització des del navegador de Windows [utilitza la mateixa xarxa local])



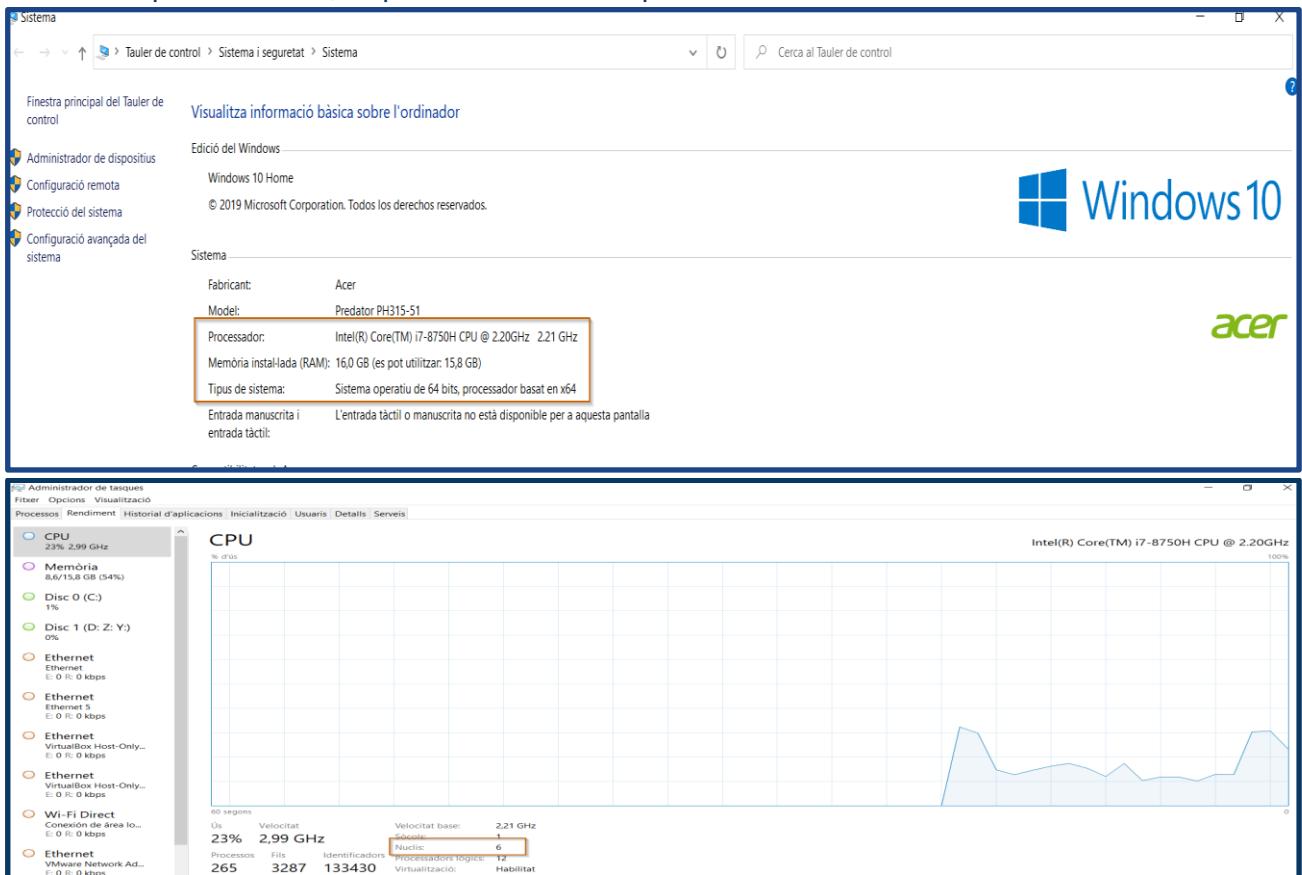
Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

3.1. Clúster Hadoop pseudodistribuït ⁴

3.1.1. Requisits de hardware

Abans de començar la instal·lació hem de saber quanta memòria RAM i quants nuclis té el host amfitrió (Windows 10) abans de crear la màquina virtual CentOS.

- PC amfitrió (Windows 10) disposa de 16 GB de memòria RAM i 6 nuclis. Farem servir **4 GB**⁵ de memòria RAM i 2 nuclis. Revisem que la unitat D: on estaran ubicades les màquines virtuals, disposa de suficient espai



3.1.2. Versions del software (per tema de compabilitats)

⁴ Aquest clúster Hadoop se li diu pseudodistribuït ja només hi haurà 1 node on una part farà de master i l'altra de slave

⁵ En aquest cas podríem fer servir més memòria RAM, però com ja es veurà posteriorment aquesta màquina es clonarà dos cops i ens quedarà 4 GB pel host amfitrió.



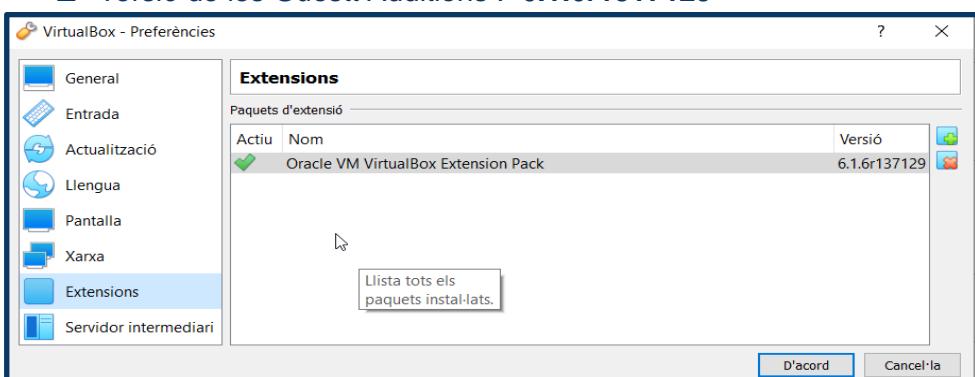
Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

A continuació mostraré les versions dels programes que s'han instal·lat (incloent-hi VirtualBox, ja que pot ser que en versions diferents s'hagi de fer modificacions)

versió del VirtualBox : 6.1

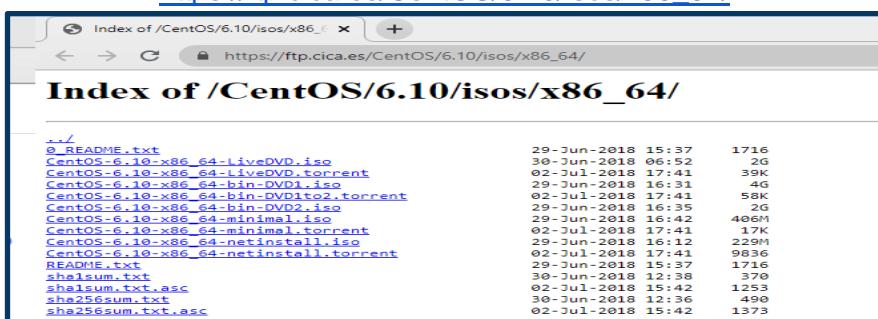


versió de les Guest Additions : 6.1.6r1317129



Versió del sistema operatiu: CentOS 6.10 de 64 bits

https://ftp.cica.es/CentOS/6.10/isos/x86_64/



Versió de Hadoop : 2.10.0

<https://hadoop.apache.org/release/2.10.0.html>

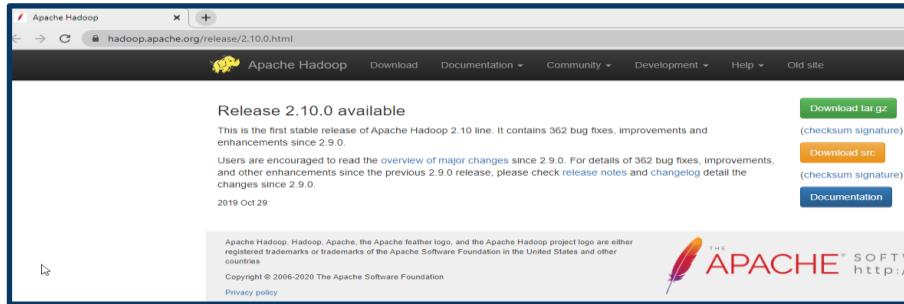


Nom i Cognoms

Arnaud Subirós Puigarnau

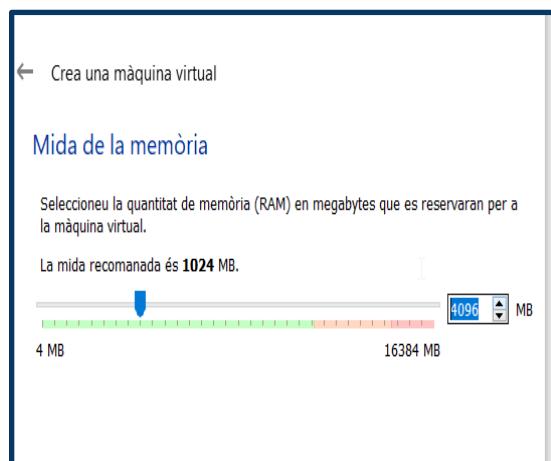
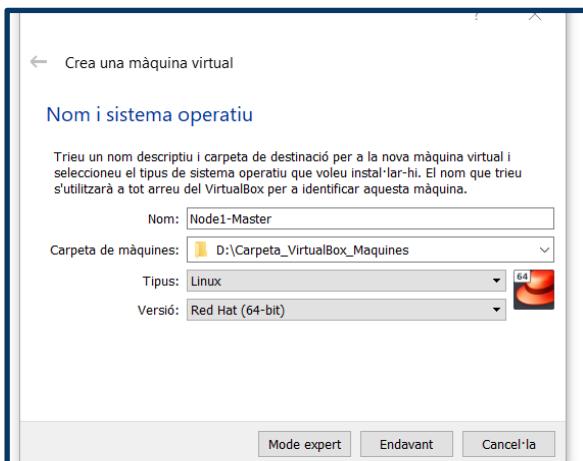
Data

02-06-2020



3.1.3. Instal·lació de CentOS 6.10

A continuació mostrarem les captures de com hem configurat la màquina i la instal·lació del seu sistema operatiu , afegint posteriorment les “*Guest Additions*” per poder tenir una millor resolució de pantalla.





Nom i Cognoms

Arnau Subirós Puigarnau

Data

02-06-2020

The image displays five windows from the Oracle VM VirtualBox Manager:

- Node1-Master - Paràmetres (Sistema):** Shows CPU settings with 1 CPU selected from 12 available. Other options include PAE/NX and VT-x/AMD-V.
- Node1-Master [S'està executant] - Oracle VM VirtualBox (Fitxer):** Shows the name "node1" being set for the running machine.
- Node1-Master [S'està executant] - Oracle VM VirtualBox (Fitxer):** A "Crea un usuari" (Create User) dialog is open, prompting for user information: Nom d'usuari: hadoop, Nom complet: hadoop, Contrasenya: (redacted), and Confirmeu la contrasenya: (redacted).
- Node1-Master [S'està executant] - Oracle VM VirtualBox (Fitxer):** A "Kdump" configuration dialog is open, showing memory usage (3959 MB total, 128 MB reserved) and a confirmation message about rebooting the system. The "Sí" button is highlighted with a red box.
- Node1-Master [S'està executant] - Oracle VM VirtualBox (Fitxer):** A login screen for the "node1" machine is shown, featuring the Hadoop logo and the "hadoop" user account.



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

- ☐ Afegim les Guest Additions (abans haurem d'actualitzar el repositori [yum](#) i instal·lar alguns paquets com es pot veure a continuació.

```
hadoop@node1:/media/VBox_GAs_6.1.6
Fitxer Edita Visualitza Cerca Terminal Ajuda
VirtualBox Guest Additions: Kernel headers not found for target kernel
2.6.32-754.el6.x86_64. Please install them and execute
  /sbin/rcvboxadd setup
modprobe vboxguest failed
The log file /var/log/vboxadd-setup.log may contain further information.
Building the VirtualBox Guest Additions kernel modules. This may take a while.
To build modules for other installed kernels, run
  /sbin/rcvboxadd quicksetup <version>
or
  /sbin/rcvboxadd quicksetup all
Kernel headers not found for target kernel 2.6.32-754.el6.x86_64. Please
install them and execute
  /sbin/rcvboxadd setup
vboxadd-service.sh: Starting VirtualBox Guest Addition service.
VirtualBox Additions module not loaded!
[root@node1 VBox_GAs_6.1.6]# yum install kernel-devel-2.6.32-754.el6.x86_64
Connectors carregats: fasttestmirror, refresh-packagekit, security
S'està preparant el procés d'instal·lació
base                                         3.7 KB   00:00
base/primary_db                                4.7 MB   00:00
extras                                         3.4 KB   00:00
extras/primary_db                               29 KB   00:00
updates                                         3.4 KB   00:00
```

```
hadoop@node1:/media/VBox_GAs_6.1.6
Fitxer Edita Visualitza Cerca Terminal Ajuda
modules. This may take a while.
VirtualBox Guest Additions: To build modules for other installed kernels, run
VirtualBox Guest Additions:  /sbin/rcvboxadd quicksetup <version>
VirtualBox Guest Additions: or
VirtualBox Guest Additions:  /sbin/rcvboxadd quicksetup all
VirtualBox Guest Additions: Building the modules for kernel
2.6.32-754.el6.x86_64.

This system is currently not set up to build kernel modules.
Please install the gcc make perl packages from your distribution.
modprobe vboxguest failed
The log file /var/log/vboxadd-setup.log may contain further information.
Building the VirtualBox Guest Additions kernel modules. This may take a while.
To build modules for other installed kernels, run
  /sbin/rcvboxadd quicksetup <version>
or
  /sbin/rcvboxadd quicksetup all
Building the modules for kernel 2.6.32-754.el6.x86_64.

This system is currently not set up to build kernel modules.
Please install the gcc make perl packages from your distribution.
vboxadd-service.sh: Starting VirtualBox Guest Addition service.
VirtualBox Additions module not loaded!
[root@node1 VBox_GAs_6.1.6]# yum install gcc make perl
```

Executem el script per instal·lar les Guest Additions i posteriorment reiniciem la màquina virtual

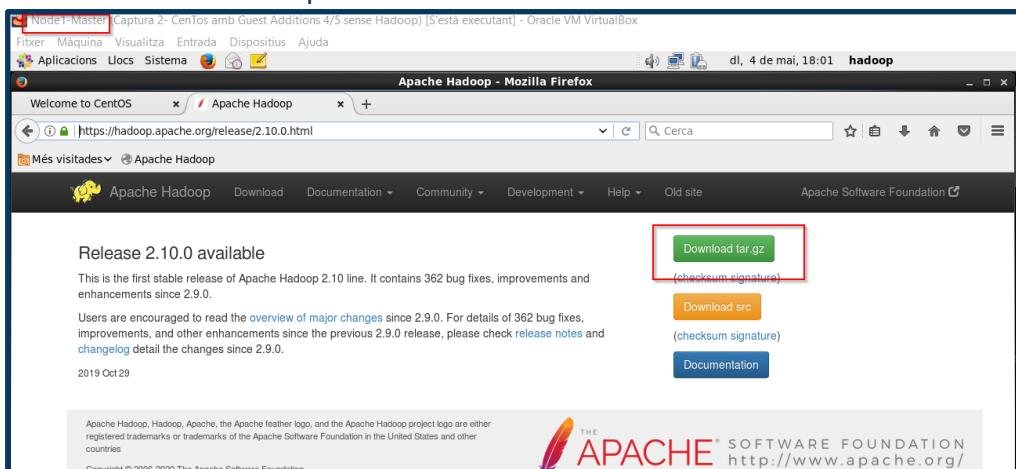
```
hadoop@node1:/media/VBox_GAs_6.1.6
Fitxer Edita Visualitza Cerca Terminal Ajuda
[hadoop@node1 VBox_GAs_6.1.6]$ ./VBoxLinuxAdditions.run
```



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

3.1.4. Instal·lació i configuració Hadoop 2.10.0

- Desde la màquina virtual tenint connexió a Internet accedim al navegador Firefox per instal·lar Hadoop 2.10.0



- Crearem un directori anomenat “hadoop” dins del directori `/opt`⁶ i donant-li permisos a l'usuari hadoop per poder-hi accedir.

⁶ Podríem instal·lar-ho en un altre directori, però per tema de bones pràctiques s'ha instal·lat a `/Opt`, ja que és a on instal·lem les aplicacions opcionals manualment.



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

```
root@node1:/opt/hadoop
[hadoop@node1 opt]$ su - root
Contrasenya:
[root@node1 ~]# cd /opt
[root@node1 opt]# mkdir hadoop
[root@node1 opt]# ls -ls
total 12
4 drwxr-xr-x. 2 root root 4096 4 mai 18:08 hadoop
4 drwxr-xr-x. 2 root root 4096 4 oct 2017 rh
4 drwxr-xr-x. 8 root root 4096 4 mai 17:38 VBoxGuestAdditions-6.1.6
[root@node1 opt]# chown hadoop /opt/hadoop
[root@node1 opt]# ls -ls
total 12
4 drwxr-xr-x. 2 hadoop root 4096 4 mai 18:08 hadoop
4 drwxr-xr-x. 2 root root 4096 4 oct 2017 rh
4 drwxr-xr-x. 8 root root 4096 4 mai 17:38 VBoxGuestAdditions-6.1.6
[root@node1 opt]#
```

- Descomprimim l'arxiu [hadoop-2.10.0.tar.gz](#) al directori que hem creat `/opt/hadoop`

```
jari@node1:~/Baixades
[hadoop@node1 Baixades]$ ls
hadoop-2.10.0.tar.gz
[hadoop@node1 Baixades]$
```

```
hadoop@node1:/opt/hadoop
[hadoop@node1 hadoop]$ tar xvf /home/hadoop/Baixades/hadoop-2.10.0.tar.gz
```

- Un cop fet això hem de revisar quina versió de JAVA tenim (i si és JRE o JDK) utilitzant la comanda `"java -version"`

Important : Utilitzant Hadoop 2.x necessitem mínim JDK 7, però utilitzarem JDK 8(que és el mínim si utilitzéssim Hadoop 3.x)

```
root@node1:/opt
[hadoop@node1 hadoop]$ java -version
java version "1.7.0_181"
OpenJDK Runtime Environment (rhel-2.6.14.10.el6-x86_64 u181-b00)
OpenJDK 64-Bit Server VM (build 24.181-b00, mixed mode)

Hi ha 2 programes que proveeixen 'java'.
Selecció      Ordre
-----+-----+
*+ 1          /usr/lib/jvm/jre-1.7.0-openjdk.x86_64/bin/java
          /usr/lib/jvm/jre-1.6.0-openjdk.x86_64/bin/java

Premeu la tecla de retorn per mantenir la selecció actual[+], o teclegeu el número de la selecció: ^?^C
[hadoop@node1 hadoop]$
```



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

- Veient que no tenim el JDK, anem a la pàgina oficial d'Oracle i ens descarreguem el JDK 8 en format RPM

The screenshot shows a Linux desktop environment with a file manager window showing 'hadoop' and 'hadoop-2.10.0' folders. A Firefox browser window is open to the Oracle Java SE Development Kit 8 Downloads page. The user has selected the 'Linux x64 RPM Package' link, which is highlighted with a red box. A modal dialog box is displayed, asking for acceptance of the Oracle Technology Network License Agreement. The 'Required' checkbox is checked, and the text 'You must accept the Oracle Technology Network License Agreement for Oracle Java SE to download this software.' is visible. Below the dialog, it says 'You will be redirected to the login screen in order to download the file.' and there is a 'Download jdk-8u251-linux-x64.rpm' button.

```
[hadoop@node1 Baixades]$ ls
hadoop-2.10.0.tar.gz  jdk-8u251-linux-x64.rpm
[hadoop@node1 Baixades]$
```

- Instal·lem el paquet RPM amb el comando “*rpm -ivh jdk-8u251-linux-x64.rpm*”

```
[hadoop@node1:~/Baixades]
[hadoop@node1 Baixades]$ ls
hadoop-2.10.0.tar.gz  jdk-8u251-linux-x64.rpm
[hadoop@node1 Baixades]$
```



```
root@node1:/home/hadoop/Baixades
root@node1:/opt/hadoop
[hadoop@node1 Baixades]# ls
hadoop-2.10.0.tar.gz  jdk-8u251-linux-x64.rpm
[hadoop@node1 Baixades]# rpm -ivh jdk-8u251-linux-x64.rpm
avis: jdk-8u251-linux-x64.rpm: Capçalera V3 RSA/SHA256 Signature, key ID ec551f03: NOKEY
S'està preparant... #####
1:jdk1.8 #####
[100%] ( 4%)
[hadoop@node1 Baixades]#
```

- Tornem a revisar les versions disponibles de JAVA i confirmem que la que ens interessa està en ús.



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

```

root@node1:/home/hadoop/Baixades
Fitxer Edita Visualitza Cerca Terminal Pestanyes Ajuda
hadoop@node1:/opt/hadoop
[root@node1 Baixades]# java --version
Unrecognized option: --version
Error: Could not create the Java Virtual Machine.
Error: A fatal exception has occurred. Program will exit.
[root@node1 Baixades]# java -version
java version "1.8.0_251"
Java(TM) SE Runtime Environment (build 1.8.0_251-b08)
Java HotSpot(TM) 64-Bit Server VM (build 25.251-b08, mixed mode)
[root@node1 Baixades]# alternatives --config java

Hi ha 3 programes que proveeixen 'java'.

  Selecció      Ordre
-----  -----
  1          /usr/lib/jvm/jre-1.7.0-openjdk.x86_64/bin/java
  2          /usr/lib/jvm/jre-1.6.0-openjdk.x86_64/bin/java
*+ 3          /usr/java/jdk1.8.0_251-amd64/jre/bin/java

Premeu la tecla de retorn per mantenir la selecció actual[+], o teclegeu el número de la selecció: 3@█

```

- Posteriorment degut a que hem instal·lat JDK de 64 bits ens apareixeran warnings al executar comandos HDFS, ja que les llibreries natives de Hadoop per defecte en aquesta versió és de 32 bits.

Nota: En el nostre cas hem fet una prova per eliminar-los de forma manual però no ho hem afegir a l'arxiu `~/.bashrc` i obrim un nou terminal ens apareixerà novament. Per aquest motiu a continuació es podran veure captures amb o sense warnings

```

hadoop@node1:~/Escriptori
Fitxer Edita Visualitza Cerca Terminal Ajuda
[hadoop@node1 Escriptori]$ hdfs dfs -ls /
20/05/25 12:27:12  WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 7 items
drwxr-xr-x  - hadoop supergroup          0 2020-05-05 14:02 /dades
drwxr-xr-x  - hadoop supergroup          0 2020-05-05 18:50 /llibres
drwxr-xr-x  - hadoop supergroup          0 2020-05-05 18:58 /sortida_llibreria
drwxr-xr-x  - hadoop supergroup          0 2020-05-05 19:55 /sortida_log
drwxr-xr-x  - hadoop supergroup          0 2020-05-05 20:24 /sortida_log2
drwxr-xr-x  - hadoop supergroup          0 2020-05-05 13:06 /temporal
drwx-----  - hadoop supergroup          0 2020-05-05 18:58 /tmp
[hadoop@node1 Escriptori]$
[hadoop@node1 Escriptori]$ export HADOOP_HOME_WARN_SUPPRESS=1
[hadoop@node1 Escriptori]$ export HADOOP_ROOT_LOGGER="WARN,DRFA"
[hadoop@node1 Escriptori]$
[hadoop@node1 Escriptori]$ hdfs dfs -ls /
Found 7 items
drwxr-xr-x  - hadoop supergroup          0 2020-05-05 14:02 /dades
drwxr-xr-x  - hadoop supergroup          0 2020-05-05 18:50 /llibres
drwxr-xr-x  - hadoop supergroup          0 2020-05-05 18:58 /sortida_llibreria
drwxr-xr-x  - hadoop supergroup          0 2020-05-05 19:55 /sortida_log
drwxr-xr-x  - hadoop supergroup          0 2020-05-05 20:24 /sortida_log2
drwxr-xr-x  - hadoop supergroup          0 2020-05-05 13:06 /temporal
drwx-----  - hadoop supergroup          0 2020-05-05 18:58 /tmp

```



```

hadoop@node1:~/Escriptori
Fitxer Edita Visualitza Cerca Terminal Ajuda
[hadoop@node1 Escriptori]$ hdfs dfs -ls
20/05/25 12:29:09  WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
ls: `.' No such file or directory
[hadoop@node1 Escriptori]$

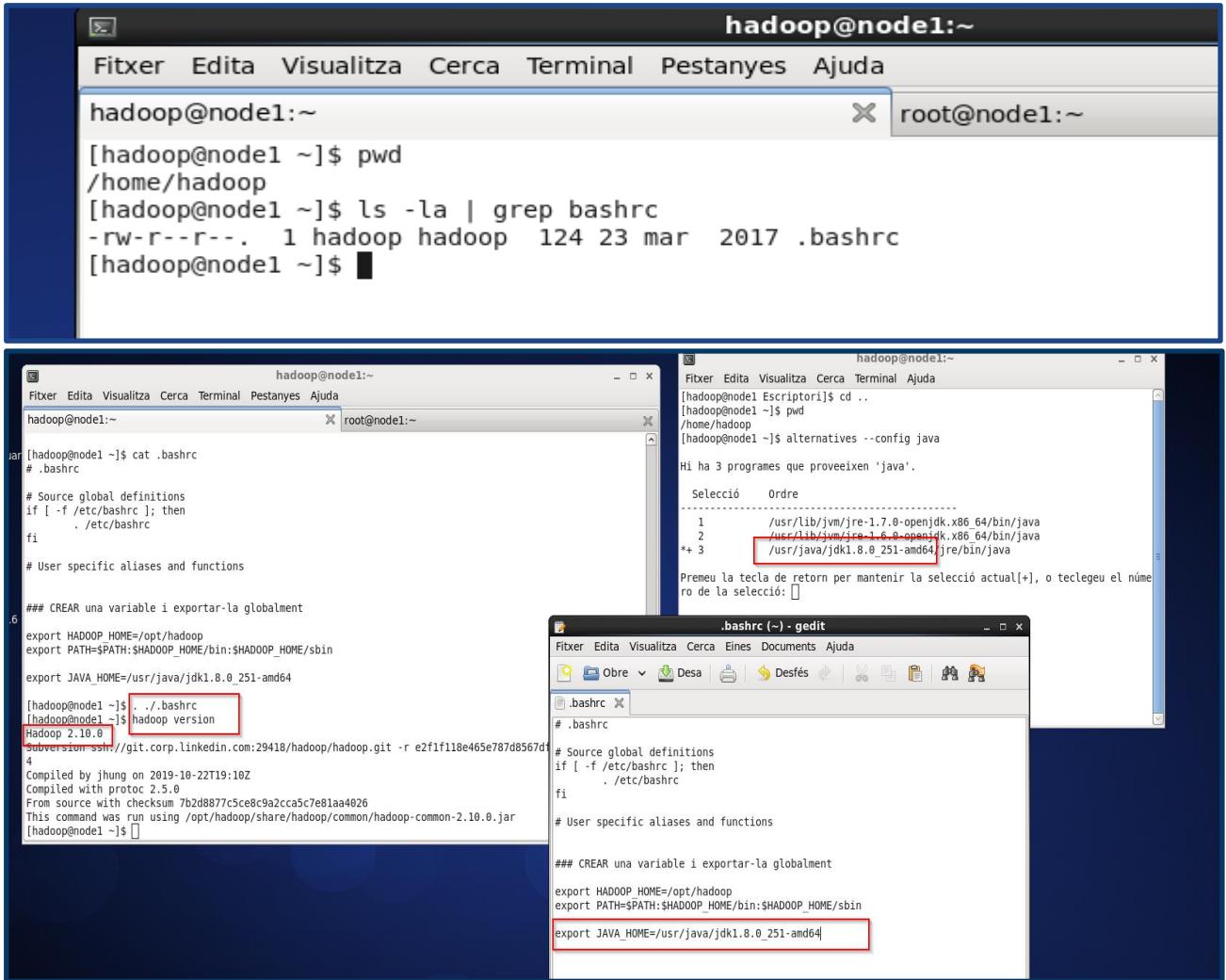
```

- Un cop instal·lat Hadoop hem de tenir en compte els directoris més importants.
 - /bin** : Estan ubicats els binaris per llençar els processos a Hadoop
 - /etc** : És el directori de configuració (que posteriorment haurem de modificar)

Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

- **/lib** : Conté les llibreries natives
- **/sbin** : Scripts que em permeten parar, arrencar hdfs...
- **/share** : On hi ha la paqueteria de Hadoop, exemples...

- El següent punt important és configurar *les variables d'entorn de Hadoop i Java* per poder treballar de manera directa amb els executables. Per això haurem de modificar l'arxiu ocult anomenat `~/.bashrc`



```

hadoop@node1:~$ pwd
/home/hadoop
[hadoop@node1 ~]$ ls -la | grep bashrc
-rw-r--r--. 1 hadoop hadoop 124 23 mar 2017 .bashrc
[hadoop@node1 ~]$ 

hadoop@node1:~$ cat .bashrc
# Source global definitions
if [ -f /etc/bashrc ]; then
    . /etc/bashrc
fi

# User specific aliases and functions

## CREAR una variable i exportar-la globalment
export HADOOP_HOME=/opt/hadoop
export PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin
export JAVA_HOME=/usr/java/jdk1.8.0_251-amd64

[hadoop@node1 ~]$ . ./bashrc
[hadoop@node1 ~]$ hadoop version
Hadoop 2.10.0
Subversion svn://git.corp.linkedin.com:29418/hadoop/hadoop.git -r e2f1f18e465e787d8567d
4
Compiled by jhung on 2019-10-22T19:10Z
Compiled with protoc 2.5.0
From source with checksum 7b2d8877c5ce8c9a2cca5c7e81aa4026
This command was run using /opt/hadoop/share/hadoop/common/hadoop-common-2.10.0.jar
[hadoop@node1 ~]$ 

```

hadoop@node1:~\$ alternatives --config java

Hi ha 3 programes que proveeixen 'java'.

Selecció	Ordre
1	/usr/lib/jvm/jre-1.7.0-openjdk.x86_64/bin/java
2	/usr/lib/jvm/jre-1.6.0-openjdk.x86_64/bin/java
*+ 3	/usr/java/jdk1.8.0_251-amd64/jre/bin/java

Premeu la tecla de retorn per mantenir la selecció actual[+], o tecleu el número de la selecció: []

.bashrc (~) - gedit

```

# .bashrc

# Source global definitions
if [ -f /etc/bashrc ]; then
    . /etc/bashrc
fi

# User specific aliases and functions

## CREAR una variable i exportar-la globalment
export HADOOP_HOME=/opt/hadoop
export PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin
export JAVA_HOME=/usr/java/jdk1.8.0_251-amd64

```

- Respecte a la xarxa, en aquest cas d'un sol node, farem servir el mode NAT, però així i tot haurem d'accendir a l'arxiu */etc/hosts*, ja que no tenim servidor DNS i quan utilitzem SSH ho voldrem amb el nom i no la IP



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

```
[root@node1 ~]# ip a
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
    inet 127.0.0.1/8 scope host lo
        inet6 ::1/128 scope host
            valid_lft forever preferred_lft forever
2: eth0: <NO-CARRIER,BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc pfifo_fast
    link/ether 08:00:00:07:dd:56 brd ff:ff:ff:ff:ff:ff
    inet 10.0.2.15/24 brd 10.0.2.255 scope global eth0
        inet6 fe80::a00:2ff:fedd:56d1/64 scope link
            valid_lft forever preferred_lft forever
[root@node1 ~]# hostname
node1
[root@node1 ~]# 

[root@node1 ~]# ping node1
PING node1 (10.0.2.15) 56(84) bytes of data.
64 bytes from node1 (10.0.2.15): icmp_seq=1 ttl=64 time=0.027 ms
64 bytes from node1 (10.0.2.15): icmp_seq=2 ttl=64 time=0.051 ms
^C
--- node1 ping statistics ---
2 packets transmitted, 2 received, 0% packet loss, time 1177ms
rtt min/avg/max/mdev = 0.027/0.039/0.051/0.012 ms
[root@node1 ~]# 
```

- Un cop modificat l'arxiu `/etc/hosts` hem de configurar el SSH, ens interessa accedir-hi sense password, per aquest motiu haurem de generar un parell de claus (pública i privada) i crear l'arxiu "`authorized_keys`" on dipositem la clau pública.

```
[hadoop@node1 ~]$ ssh-keygen
Generating public/private rsa key pair.
Enter file in which to save the key (/home/hadoop/.ssh/id_rsa):
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /home/hadoop/.ssh/id_rsa.
Your public key has been saved in /home/hadoop/.ssh/id_rsa.pub.
The key's randomart image is:
++ [ RSA 2048] ++
|...|
|o* . + |
|B B o . |
|.B . . |
|o.o S . |
|.o. . . |
|. . . . |
| . E . . |
| .. . . . |
[hadoop@node1 ~]$ 

[hadoop@node1 ~]$ cd /home/hadoop/.ssh
[hadoop@node1 .ssh]$ ls -ls
total 8
4 -rw-r----- 1 hadoop hadoop 1675 4 mai 20:05 id_rsa
4 -rw-r--r-- 1 hadoop hadoop 394 4 mai 20:05 id_rsa.pub
[hadoop@node1 .ssh]$ 
[hadoop@node1 .ssh]$ cat id_rsa.pub
ssh-rsa AAAAB3MzIzaC1yc2EAAQABlwAAQEA1+Zbu76E0KoXLWhxQXM1bbLajlmktUT895EEf6d06Mcq
InxuksOssmM6CbstDorYXk6GmcrcryxaTzTdg7TMoxWbD+IiciaMmCN/egN77xGyBLKH1/cDhJq+0
yxqlxEahj4R4Nq1p6RLRIDz6ymlm/6rys0DDtqGUb0h+2HD1Y2NuwgXmh3h3+L/uZGVvAhqggCHkFK
z4oT1AaTwfuPmVSofgYHAttftEH017Dyahah8lrmivAtCFMjJlJb7fqM505g0ghutpA8jPsZYVhxp/pw4
v0gTrJ+sxHYd32ofx0Cu3lK2Fzp31bMneA3PZWIffWqps4ft70WdALwQ== hadoop@node1
[hadoop@node1 .ssh]$ 
[hadoop@node1 .ssh]$ cp id_rsa.pub authorized_keys
[hadoop@node1 .ssh]$ cat authorized_keys
ssh-rsa AAAAB3MzIzaC1yc2EAAQABlwAAQEA1+Zbu76E0KoXLWhxQXM1bbLajlmktUT895EEf6d06Mcq
InxuksOssmM6CbstDorYXk6GmcrcryxaTzTdg7TMoxWbD+IiciaMmCN/egN77xGyBLKH1/cDhJq+0
yxqlxEahj4R4Nq1p6RLRIDz6ymlm/6rys0DDtqGUb0h+2HD1Y2NuwgXmh3h3+L/uZGVvAhqggCHkFK
z4oT1AaTwfuPmVSofgYHAttftEH017Dyahah8lrmivAtCFMjJlJb7fqM505g0ghutpA8jPsZYVhxp/pw4
v0gTrJ+sxHYd32ofx0Cu3lK2Fzp31bMneA3PZWIffWqps4ft70WdALwQ== hadoop@node1
[hadoop@node1 .ssh]$ 
```

- Fem una prova per accedir per SSH sense contrasenya (des de la mateixa maquina).

⁷ Aquest punt és molt important en la creació d'un clúster en diversos nodes (ja es mostrarà més endavant)



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

```
hadoop@node1:~/ssh
[...] $ ssh-keygen -t rsa
[...] $ cp id_rsa.pub authorized_keys
[...] $ cat authorized_keys
[...] $ ssh node1
The authenticity of host 'node1 (10.0.2.15)' can't be established.
RSA key fingerprint is e6:7d:43:d6:f6:4c:35:22:e3:c9:b4:ab:b3:47:33:4d.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'node1,10.0.2.15' (RSA) to the list of known hosts.
[...] $ exit
[...] $ logout
[...] $ Connection to node1 closed.
[...] $
```

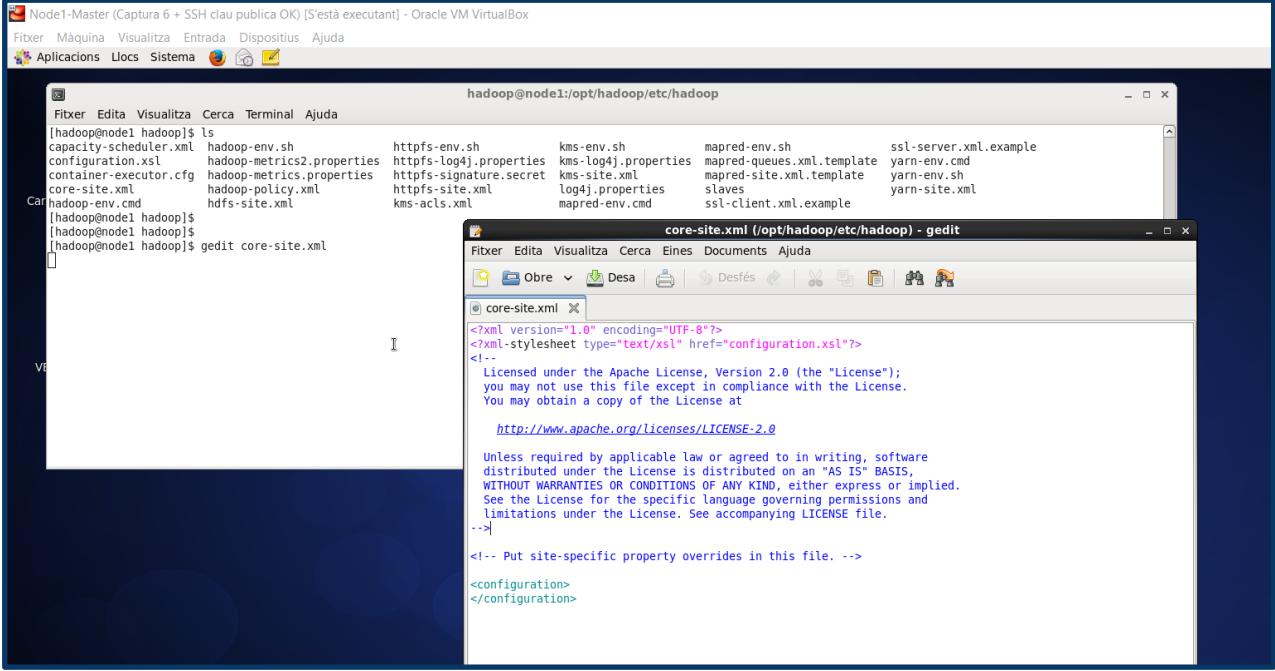
3.1.4.1. Configuració de les dades

Com ja hem comentat abans els 2 components bàsics d'un clúster Hadoop són les “dades” i els “processos”

- Primer farem la configuració de les dades amb algun exemple pràctic (creació d'un directori HDFS)
- Seguidament haurem de fer la configuració dels processos amb algun exemple amb MapReduce
- Per la configuració de les dades i els processos ens interessa modificar 4 arxius XML ubicats a **/opt/hadoop/etc/hadoop**
 - **core-site.xml** : Arxiu de configuració general del Clúster
 - **hdfs-site.xml** : Arxiu de configuració per les dades HDFS
 - **mapred-site.xml.template** : Arxiu de configuració per MapReduce
 - **yarn-site.xml** : Arxiu de configuració del mode de processos Yarn

Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

- Arxiu **core-site.xml**⁸ : Li direm la configuració bàsica pel nostre Clúster Hadoop



The screenshot shows a terminal window titled "Node1-Master (Captura 6 + SSH clau pública OK) [S'està executant] - Oracle VM VirtualBox". Inside the terminal, the user is navigating through the Hadoop configuration directory:

```

[hadoop@node1:~/opt/hadoop/etc/hadoop]$ ls
capacity-scheduler.xml      hadoop-env.sh          httpfs-env.sh      kms-env.sh        mapred-env.sh      ssl-server.xml.example
configuration.xsl            hadoop-metrics2.properties  httpfs-log4j.properties  kms-log4j.properties  mapred-queues.xml.template  yarn-env.cmd
container-executor.cfg       hadoop-metrics.properties   httpfs-signature.secret  kms-site.xml       mapred-site.xml.template  yarn-env.sh
core-site.xml                hadoop-policy.xml      httpfs-site.xml      log4j.properties   slaves           yarn-site.xml
hadoop-env.cmd              hdfs-site.xml       kms-acls.xml       mapred-env.cmd     ssl-client.xml.example
[hadoop@node1:~/opt/hadoop/etc/hadoop]$ gedit core-site.xml
[hadoop@node1:~/opt/hadoop/etc/hadoop]$ gedit core-site.xml
[hadoop@node1:~/opt/hadoop/etc/hadoop]$ gedit core-site.xml

```

Below the terminal, a file browser window titled "core-site.xml (/opt/hadoop/etc/hadoop) - gedit" is open, displaying the XML content of the core-site.xml file. The XML includes standard Apache license information and a placeholder for site-specific overrides.

```

<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
Licensed under the Apache License, Version 2.0 (the "License");
you may not use this file except in compliance with the License.
You may obtain a copy of the License at

http://www.apache.org/licenses/LICENSE-2.0

Unless required by applicable law or agreed to in writing, software
distributed under the License is distributed on an "AS IS" BASIS,
WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
See the License for the specific language governing permissions and
limitations under the License. See accompanying LICENSE file.
-->
<!-- Put site-specific property overrides in this file. -->
<configuration>
</configuration>

```

farem una configuració bàsica

⁸ Per defecte aquest arxiu està buit

Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

*core-site.xml (/opt/hadoop/etc/hadoop) - gedit

Fitxer Edita Visualitzà Cerca Eines Documents Ajuda

Obre Desa Desfés

*core-site.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
Licensed under the Apache License, Version 2.0 (the "License");
you may not use this file except in compliance with the License.
You may obtain a copy of the License at

http://www.apache.org/licenses/LICENSE-2.0

Unless required by applicable law or agreed to in writing, software
distributed under the License is distributed on an "AS IS" BASIS,
WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
See the License for the specific language governing permissions and
limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<!-- NAMENODE -->

<configuration>
    <property>
        <name>fs.defaultFS</name>    <!--aquesta propietat ens indica el sistema de fitxers que
                                         utilitzarem per a Hadoop -->
        <value>hdfs://node1:9000</value> <!--especificuem el servidor "master" on hi han les metadades -->
    </property>
</configuration>
```

- Arxiu `hdfs-site.xml`⁹ : Farem la configuració bàsica per al nostre clúster Hadoop

The screenshot shows a terminal window on the left and a gedit editor window on the right. The terminal window displays the command `hadoop@node1:~$ ls` followed by a list of files in the `/opt/hadoop/etc/hadoop` directory. The gedit editor window shows the `hdfs-site.xml` file with its XML structure and Apache License 2.0 header.

```
[hadoop@node1 hadoop]$ ls
capacity-scheduler.xml  hadoop-env.sh      httpfs-env.sh      kms-env.sh
configuration.xsl        hadoop-metrics2.properties  httpfs-log4j.properties  kms-log4j.properties
container-executor.cfg   hadoop-metrics.properties  httpfs-signature.secret  kms-site.xml
core-site.xml            hadoop-policy.xml     httpfs-site.xml      log4j.properties
hadoop-env.cmd          hdfs-site.xml       kms-acls.xml       mapred-env.cmd

[hadoop@node1 hadoop]$
[hadoop@node1 hadoop]$
[hadoop@node1 hadoop]$ gedit core-site.xml
[hadoop@node1 hadoop]$ gedit hdfs-site.xml
[hadoop@node1 hadoop]$
```

hdfs-site.xml (/opt/hadoop/etc/hadoop) - gedit

```
<?xml version="1.0" encoding="UTF-8"?>
<xm...-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
 Licensed under the Apache License, Version 2.0 (the "License");
 you may not use this file except in compliance with the License.
 You may obtain a copy of the License at

 http://www.apache.org/licenses/LICENSE-2.0

 Unless required by applicable law or agreed to in writing, software
 distributed under the License is distributed on an "AS IS" BASIS,
 WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
 See the License for the specific language governing permissions and
 limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. --&gt;

&lt;configuration&gt;
&lt;/configuration&gt;</pre>
```

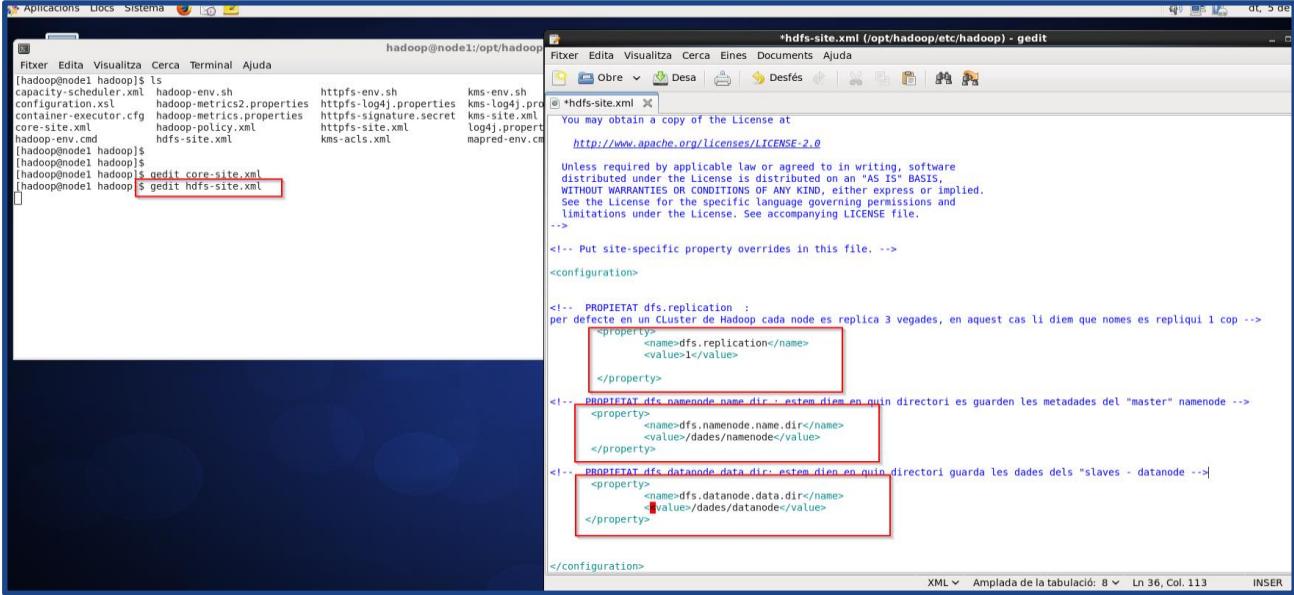
⁹ Per defecte aquest arxiu està buit

Nom i Cognoms

Arnau Subirós Puigarnau

Data

02-06-2020



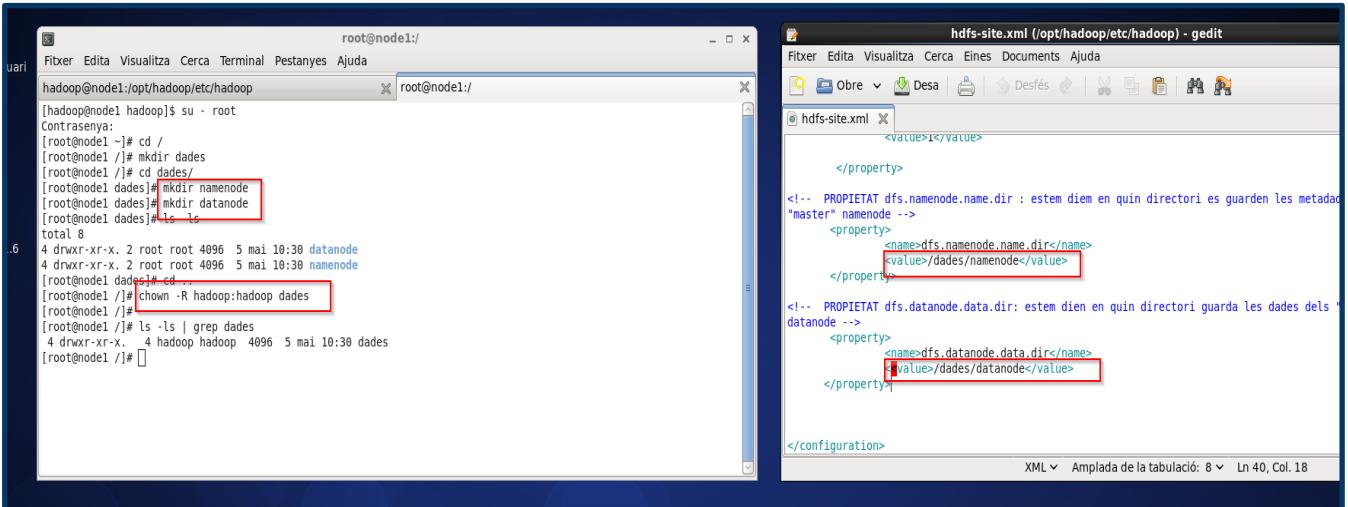
The screenshot shows two windows side-by-side. On the left is a terminal window titled 'hadoop@node1:/opt/hadoop'. It displays the command 'ls' followed by several files: cap-safety-filter.xml, hadoop-env.sh, configuration.xml, container-executor.cfg, core-site.xml, hadoop-env.cmd, hadoop-env.sh, hadoop-metrics2.properties, hadoop-metrics.properties, hadoop-policy.xml, httpfs-env.sh, httpfs-signature.secret, httpfs-site.xml, kms-acls.xml, kms-site.xml, log4j.properties, mapred-env.cmd, and mapred-site.xml. The file 'hdfs-site.xml' is highlighted with a red box.

On the right is a text editor window titled '*hdfs-site.xml (/opt/hadoop/etc/hadoop) - gedit'. It shows the XML configuration for HDFS. Three specific properties are highlighted with red boxes:

- <property> <name>dfs.replication</name> <value>1</value> </property>
- <property> <name>dfs.namenode.name.dir</name> <value>/dades/namenode</value> </property>
- <property> <name>dfs.datanode.data.dir</name> <value>/dades/datanode</value> </property>

- Un cop configurats aquests 2 arxius haurem d'arrencar HDFS.

Primer haurem de crear els directoris que hem dit en les propietats del NameNode i DataNode **/dades/namenode** i **/dades/datanode** i donar-li permisos a l'usuari



The screenshot shows two windows. On the left is a terminal window titled 'root@node1:/'. It shows the user switching to root ('su - root'), navigating to the '/dades' directory ('cd /'), creating the 'dades/namenode' and 'dades/datanode' directories ('mkdir namenode' and 'mkdir datanode'), listing the contents ('ls'), and then listing files containing 'dades' ('ls | grep dades').

On the right is a text editor window titled 'hdfs-site.xml (/opt/hadoop/etc/hadoop) - gedit'. It shows the XML configuration for HDFS. The same three properties from the previous screenshot are highlighted with red boxes:

- <property> <name>dfs.replication</name> <value>1</value> </property>
- <property> <name>dfs.namenode.name.dir</name> <value>/dades/namenode</value> </property>
- <property> <name>dfs.datanode.data.dir</name> <value>/dades/datanode</value> </property>

- Després utilitzant el comando "**hdfs namenode -format**" crearem les metadades al directori **/dades/namenode**

Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

```
mai 10:30 datanode
mai
[hadoop@node1 ~]$ hadoop@node1:/opt/hadoop/etc/hadoop
Fitxer Edita Visualitzar Cerca Terminal Ajuda
        at org.apache.hadoop.conf.Configuration.loadResource(Configuration.java:2796)
        ... 8 more
[hadoop@node1 hadoop]$ clear

[hadoop@node1 hadoop]$ hdfs namenode -format
20/05/05 10:39:27 INFO namenode.NameNode: STARTUP_MSG:
/*****STARTUP_MSG: Starting NameNode
STARTUP_MSG: host = node1/10.0.2.15
STARTUP_MSG: args = [-format]
STARTUP_MSG: version = 2.10.0
STARTUP_MSG: classpath = /opt/hadoop/etc/hadoop:/opt/hadoop/share/hadoop/common/lib/httpcore-4.4.4.jar:/opt/hadoop/share/hadoop/common/lib/slf4j-api-1.7.25.jar:/opt/hadoop/share/hadoop/common/lib/java-xmlbuilder-0.4.jar:/opt/hadoop/share/hadoop/common/lib/apacheds-i18n-2.0.0-M15.jar:/opt/hadoop/share/hadoop/common/lib/hadoop-auth-2.10.0.jar:/opt/hadoop/share/hadoop/common/lib/jetty-util-6.1.26.jar:/opt/hadoop/share/hadoop/common/lib/jersey-core-1.9.jar:/opt/hadoop/share/hadoop/common/lib/zookeeper-3.4.9.jar:/opt/hadoop/share/hadoop/common/lib/jersey-server-1.9.jar:/opt/hadoop/share/hadoop/common/lib/commons-math3-3.1.1.jar:/opt/hadoop/share/hadoop/common/lib/jackson-xc-1.9.13.jar:/opt/hadoop/share/hadoop/common/lib/hadoop-annotations-2.10.0.jar:/opt/hadoop/share/hadoop/common/lib/jetty-sslengine-6.1.26.jar:/opt/hadoop/share/hadoop/common/lib/jsch-0.1.54.jar:/opt/hadoop/share/hadoop/common/lib/junit-4.11.jar:/opt/hadoop/share/hadoop/common/lib/servlet-api-2.5.jar:/opt/hadoop/share/hadoop/common/lib/commons-digester-1.8.jar:/opt/hadoop/share/hadoop/common/lib/protobuf-java-2.5.0.jar:/opt/hadoop/share/hadoop/common/lib/netty-3.10.6.Final.jar:/opt/hadoop/share/hadoop/common/lib/jackson-mapper-asl-1.9.13.jar:/opt/hadoop/share/hadoop/c
datanode
ai
[hadoop@node1 ~]$ hadoop@node1:/opt/hadoop/etc/hadoop
Fitxer Edita Visualitzar Cerca Terminal Ajuda
[hadoop@node1 ~]$ hadoop@node1:/dades/namenode/current
hadoop@node1:/dades/namenode/current
Fitxer Edita Visualitzar Cerca Terminal Ajuda
20/05/05 10:39:29 INFO namenode.FSImage: Allocated new BlockPoolId: BP-1065493658-10.0.2.15-1588667969587
20/05/05 10:39:29 INFO common.Storage: Storage directory /dades/namenode has been successfully formatted.
20/05/05 10:39:29 INFO namenode.FSImageFormatProtobuf: Saving image file /dades/namenode/current/fsimage.ckpt_00000000000000000000 using no compression
20/05/05 10:39:29 INFO namenode.FSImageFormatProtobuf: Image file /dades/namenode/current/fsimage.ckpt_00000000000000000000 of size 325 bytes saved in 0 seconds.
20/05/05 10:39:29 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
20/05/05 10:39:29 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid = 0 when meet shutdown.
20/05/05 10:39:29 INFO namenode.NameNode: SHUTDOWN_MSG:
/*****SHUTDOWN_MSG: Shutting down NameNode at node1/10.0.2.15
*****
[hadoop@node1 hadoop]$ cd /dades/namenode/
[hadoop@node1 namenode]$ ls -ls
total 4
4 drwxrwxr-x. 2 hadoop hadoop 4096 5 mai 10:39 current
[hadoop@node1 current]$ cd current/
[hadoop@node1 current]$ ls -ls
total 16
4 -v--v-- 1 hadoop hadoop 325 5 mai 10:39 fsimage_00000000000000000000000000000000
4 -rw-rw-r--. 1 hadoop hadoop 62 5 mai 10:39 fsimage_00000000000000000000000000000000.md5
4 -rw-rw-r--. 1 hadoop hadoop 2 5 mai 10:39 seen txid
4 -rw-rw-r--. 1 hadoop hadoop 215 5 mai 10:39 VERSION
[hadoop@node1 current]$
```

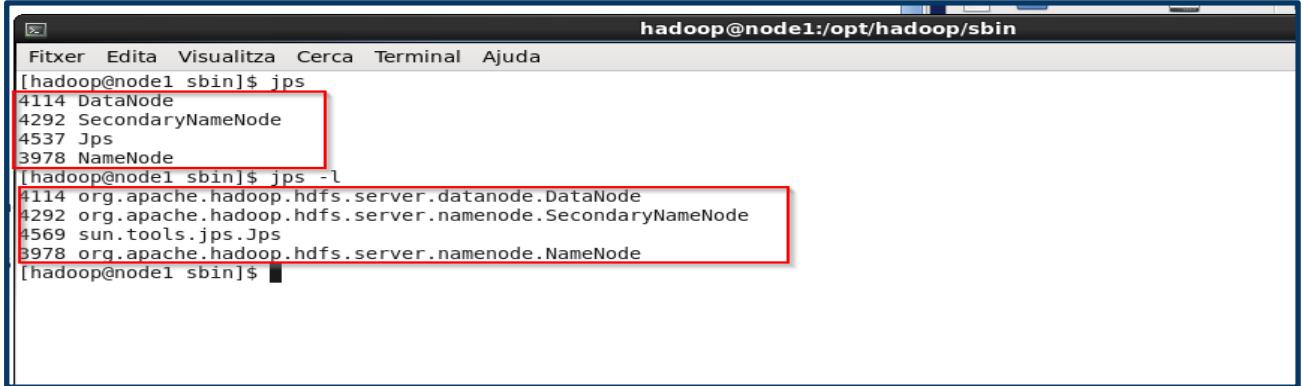
- Després utilitzant el script **start-dfs.sh**¹⁰ ubicat al directori `/opt/hadoop/sbin` arrencarà el clúster (només dades) segon hem especificat en l'arxiu de configuració. Podrem observar que arrencarà 3 processos: "*namenode*, *datanode* i *secondary namenode*".

```
# cd des1#
[des1# Fitxer Edita Visualitzar Cerca Terminal Ajuda
[des1# [hadoop@node1 ~]$ cd /opt/hadoop/sbin
[hadoop@node1 sbin]$ ls
distribute-exclude.sh  hdfs-config.sh      slaves.sh          start-dfs.sh    stop-all.sh    stop-yarn.cmd
[hadoop@node1 sbin]$ ls
distribute-exclude.sh  hdfs-config.sh      slaves.sh          start-dfs.sh    stop-all.sh    stop-yarn.cmd
[hadoop@node1 sbin]$ FederationStateStore   httpfs.sh        start-all.cmd   start-secure-dns.sh  stop-balancer.sh  stop-yarn.sh
[hadoop@node1 sbin]$ FederationStateStore   httpfs.sh        start-all.cmd   start-secure-dns.sh  stop-balancer.sh  stop-yarn.sh
[hadoop@node1 sbin]$ hadoop-daemon.sh       kms.sh          start-dfs.sh    start-yarn.cmd   stop-balancer.sh  stop-yarn.sh
[hadoop@node1 sbin]$ hadoop-daemons.sh      mr-jobhistory-daemon.sh  start-yarn.sh    start-yarn.sh    stop-dfs.cmd     yarn-daemon.sh
[hadoop@node1 sbin]$ hdfs-config.cmd       refresh-namenodes.sh  start-dfs.cmd   stop-all.cmd    stop-dfs.sh      yarn-daemons.sh
[hadoop@node1 sbin]$ [hadoop@node1 sbin]$ start-dfs.sh
[hadoop@node1 sbin]$ [hadoop@node1 sbin]$ start-dfs.sh
20/05/05 10:51:35 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Starting namenodes on [node1]
node1: starting namenode, logging to /opt/hadoop/logs/hadoop-hadoop-namenode-node1.out
The authenticity of host 'localhost (::1)' can't be established.
RSA key fingerprint is e6:7d:43:d6:f6:4c:35:22:e3:c9:b4:ab:b3:47:33:4d.
Are you sure you want to continue connecting (yes/no)? yes
Local address: [permanently added] "localhost" (RSA) to the list of known hosts.
localhost: starting datanode, logging to /opt/hadoop/logs/hadoop-hadoop-datanode-node1.out
Starting secondary namenode [0.0.0.0]
Starting secondary namenode [0.0.0.0]
The authenticity of host 'localhost ([0.0.0.0])' can't be established.
RSA key fingerprint is e6:7d:43:d6:f6:4c:35:22:e3:c9:b4:ab:b3:47:33:4d.
Are you sure you want to continue connecting (yes/no)? yes
0.0.0.0: Warning: permanently added "[0.0.0.0]" (RSA) to the list of known hosts.
0.0.0.0: starting secondarynamenode, logging to /opt/hadoop/logs/hadoop-secondarynamenode-node1.out
20/05/05 10:52:03 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
[hadoop@node1 sbin]$
```

- Hi ha un comando de Java (dins del JDK) molt útil anomenat **jps** per saber quins processos s'estan executant a la màquina virtual Java (JVM)
 - Utilitzant **jps -l** ens dóna més informació de la classe .java que s'està executant

¹⁰ L'arrencada del clúster es fa desde el "master" i via SSH enviarà comandes als "slaves".

Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020



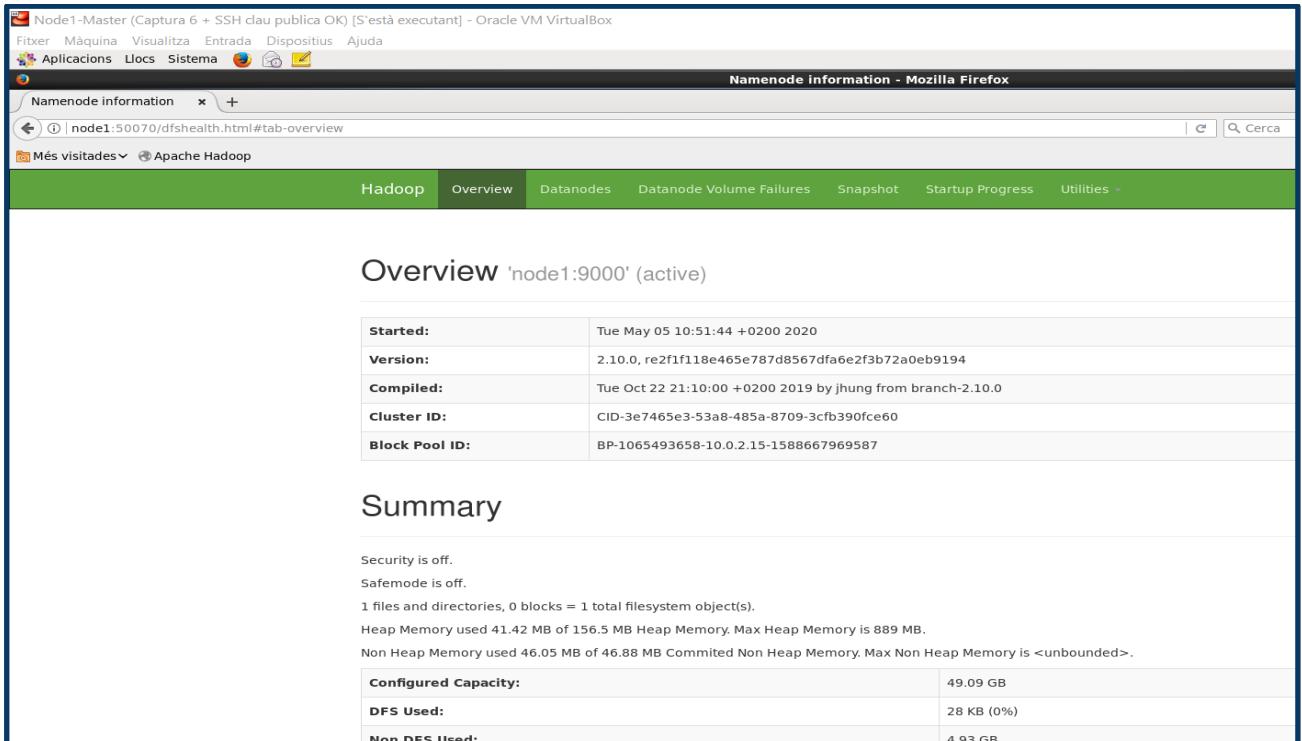
```

hadoop@node1:/opt/hadoop/sbin
Fitxer Edita Visualitza Cerca Terminal Ajuda
[hadoop@node1 sbin]$ jps
4114 DataNode
4292 SecondaryNameNode
4537 Jps
3978 NameNode
[hadoop@node1 sbin]$ jps -l
4114 org.apache.hadoop.hdfs.server.datanode.DataNode
4292 org.apache.hadoop.hdfs.server.namenode.SecondaryNameNode
4569 sun.tools.jps.Jps
3978 org.apache.hadoop.hdfs.server.namenode.NameNode
[hadoop@node1 sbin]$

```

- Obrim un navegador per veure la interfície d'usuari web del **Namenode**¹¹ (per defecte utilitzà el port 50070 . Com que anteriorment ja havíem associat el nom de la màquina amb la IP estàtica a l'arxiu [/etc/hosts](#) accedirem pel nom del host i el port.

node1:50070



Overview 'node1:9000' (active)

Started:	Tue May 05 10:51:44 +0200 2020
Version:	2.10.0, re2f1f118e465e787d8567dfa6e2f3b72a0eb9194
Compiled:	Tue Oct 22 21:10:00 +0200 2019 by jhung from branch-2.10.0
Cluster ID:	CID-3e7465e3-53a8-485a-8709-3cfb390fce60
Block Pool ID:	BP-1065493658-10.0.2.15-1588667969587

Summary

Security is off.
 Safemode is off.
 1 files and directories, 0 blocks = 1 total filesystem object(s).
 Heap Memory used 41.42 MB of 156.5 MB Heap Memory. Max Heap Memory is 889 MB.
 Non Heap Memory used 46.05 MB of 46.88 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	49.09 GB
DFS Used:	28 KB (0%)
Non DFS Used:	4.93 GB

- Parem el clúster utilitzant el comando [stop-dfs.sh](#)

¹¹ Si utilitzem una versió de [Hadoop 3.x](#), el port per defecte canvia, seria el 9870



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

```
hadoop@node1:/dades/namenode/current
Fitxa Edita Visualitza Cerca Terminal Ajuda
4 -rw-rw-r--. 1 hadoop hadoop      42  5 mai 10:52 edits_00000000000000000001-00000000000000000002
4 -rw-rw-r--. 1 hadoop hadoop      42  5 mai 11:53 edits_00000000000000000003-00000000000000000004
1024 -rw-rw-r--. 1 hadoop hadoop 1048576 5 mai 11:53 edits_inprogress_00000000000000000005
4 -rw-rw-r--. 1 hadoop hadoop      325  5 mai 10:59 fsimage_00000000000000000000
4 -rw-rw-r--. 1 hadoop hadoop      62   5 mai 10:39 fsimage_00000000000000000000.md5
4 -rw-rw-r--. 1 hadoop hadoop      325  5 mai 11:53 fsimage_00000000000000000004
4 -rw-rw-r--. 1 hadoop hadoop      62   5 mai 11:53 fsimage_00000000000000000004.md5
4 -rw-rw-r--. 1 hadoop hadoop      2    5 mai 11:53 seen_txid
4 -rw-rw-r--. 1 hadoop hadoop     215  5 mai 10:39 VERSION
[hadoop@node1 current]$ stop-dfs.sh
20/05/05 11:59:07 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
using builtin-java classes where applicable
Stopping namenodes on [node1]
node1: stopping namenode
localhost: stopping datanode
Stopping secondary namenodes [0.0.0.0]
0.0.0.0: stopping secondarynamenode
20/05/05 11:59:36 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using built-in-java classes where applicable
[hadoop@node1 current]$ clear

[hadoop@node1 current]$ start-dfs.sh
20/05/05 12:00:42 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using built-in-java classes where applicable
Starting namenodes on [node1]
node1: starting namenode, logging to /opt/hadoop/logs/hadoop-hadoop-namenode-node1.out
localhost: starting datanode, logging to /opt/hadoop/logs/hadoop-hadoop-datanode-node1.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /opt/hadoop/logs/hadoop-hadoop-secondarynamenode-node1.out
20/05/05 12:01:04 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using built-in-java classes where applicable
[hadoop@node1 current]$
[hadoop@node1 current]$ ls -ls
total 2080
4 -rw-rw-r--. 1 hadoop hadoop      42  5 mai 10:52 edits_00000000000000000001-00000000000000000002
4 -rw-rw-r--. 1 hadoop hadoop      42  5 mai 11:53 edits_00000000000000000003-00000000000000000004
1024 -rw-rw-r--. 1 hadoop hadoop 1048576 5 mai 11:53 edits_00000000000000000005-00000000000000000005
1024 -rw-rw-r--. 1 hadoop hadoop 1048576 5 mai 12:00 edits_inprogress_00000000000000000006
4 -rw-rw-r--. 1 hadoop hadoop      325  5 mai 10:59 fsimage_00000000000000000000
4 -rw-rw-r--. 1 hadoop hadoop      62   5 mai 10:39 fsimage_00000000000000000000.md5
4 -rw-rw-r--. 1 hadoop hadoop      325  5 mai 11:53 fsimage_00000000000000000004
4 -rw-rw-r--. 1 hadoop hadoop      62   5 mai 11:53 fsimage_00000000000000000004.md5
4 -rw-rw-r--. 1 hadoop hadoop      2    5 mai 12:00 seen_txid
4 -rw-rw-r--. 1 hadoop hadoop     215  5 mai 10:39 VERSION
[hadoop@node1 current]$
```

3.1.4.1.1. Exemples pràctics amb HDFS



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

- Crearem un directori HDFS anomenat "temporal" i utilitzant la comanda "put" pujarem un fitxer de txt del sistema d'arxius local al directori temporal de HDFS. Posteriorment ho podrem visualitzar des de la interfície d'usuari web del Namenode on seleccionarem l'opció "Browse the file system" ubicat en la pestanya "Utilities"

- hdfs dfs -mkdir /temporal
- hdfs dfs -put prova.txt /temporal

The figure consists of three vertically stacked screenshots of the Apache Hadoop Web UI, specifically the 'Browse Directory' section.

Screenshot 1: Shows the 'Browse Directory' interface for the '/sortida_log' directory. The 'Utilities' tab is selected, and a dropdown menu shows 'Browse the file system' and 'Logs'. A red box highlights the 'Browse the file system' option.

Screenshot 2: Shows the terminal window of a Node1 host where commands are being run. It includes:

- [hadoop@node1 bin]\$ echo Hola >> prova.txt
- [hadoop@node1 bin]\$ hdfs dfs -mkdir /temporal
- [hadoop@node1 bin]\$ hdfs dfs -put prova.txt /temporal

A red box highlights the 'hdfs dfs -put prova.txt /temporal' command. Another red box highlights the output line 'WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using built-in Java classes where applicable'.

Screenshot 3: Shows the 'File information - prova.txt' modal dialog. It displays details about the file, including its block information. The 'Block 0' tab is selected, showing:

- Block ID: 1073741825
- Block Pool ID: BP-1065493658-10.0.2.15-1588667969587
- Generation Stamp: 1001
- Size: 5
- Availability: node1

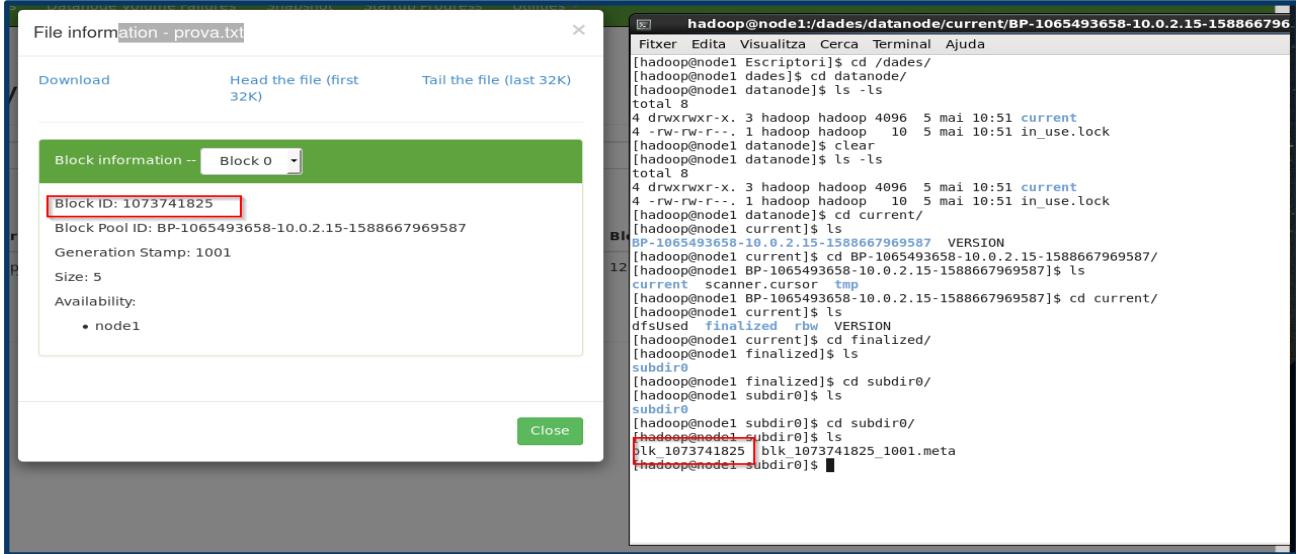
El ID del block i la seva ubicació en el sistema d'arxius local dins del directori [/dades/datanode](#)

Nom i Cognoms

Arnau Subirós Puigarnau

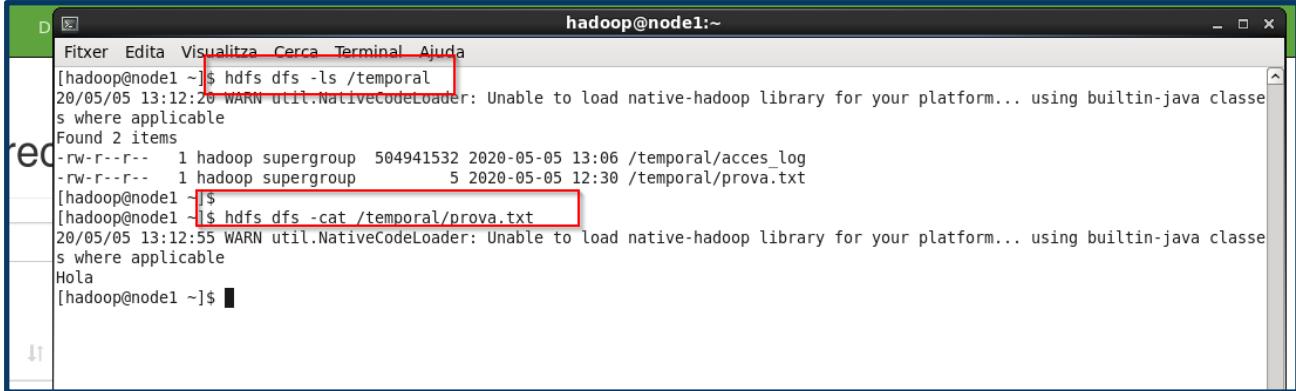
Data

02-06-2020

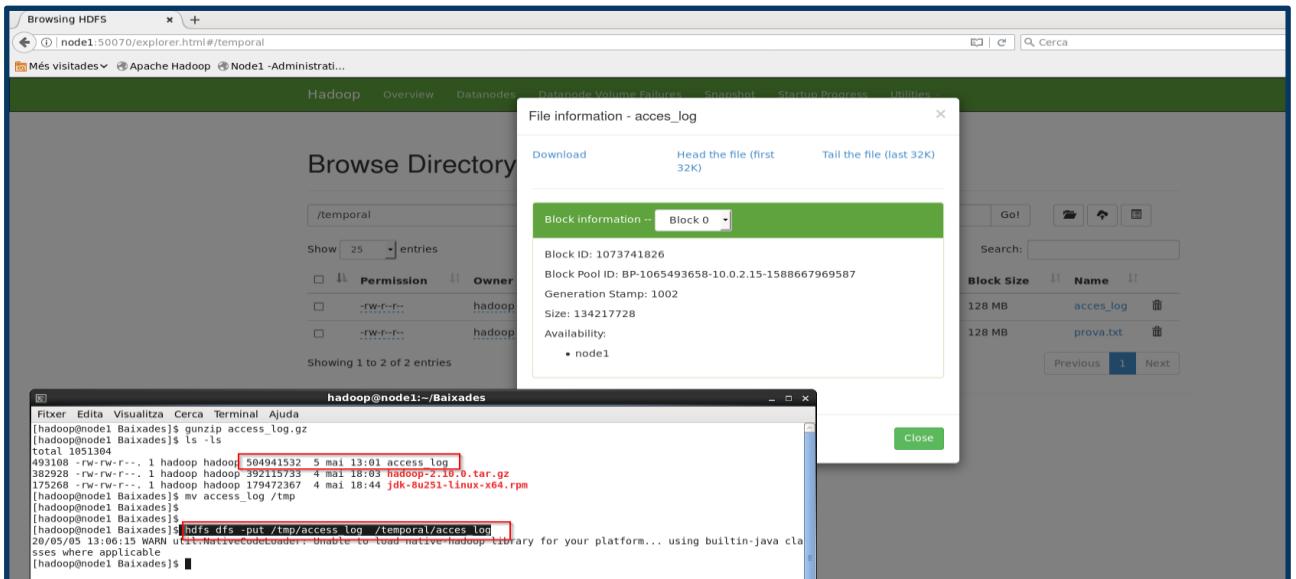


The screenshot shows two terminal windows. The left window is titled "File information - prova.txt" and displays details about a file. It shows the Block ID (1073741825), Block Pool ID (BP-1065493658-10.0.2.15-1588667969587), Generation Stamp (1001), Size (5), and Availability (node1). The right window shows a command-line session on node1. The user runs "hadoop dfs -ls /temporal" and "hadoop dfs -cat /temporal/prova.txt". The output shows a directory listing for /temporal and the contents of prova.txt.

- Farem una prova, amb un fitxer més gran (> de 500 MB) on ja no es guarda en 1 block(128 MB) sino en 4.



The screenshot shows a terminal window titled "hadoop@node1:~". The user runs "hadoop dfs -ls /temporal" and "hadoop dfs -cat /temporal/prova.txt". The output shows a directory listing for /temporal and the contents of prova.txt. The file is 5 MB in size and is split into four blocks of 128 MB each.



The screenshot shows a browser-based HDFS file browser titled "Browsing HDFS" and a terminal window titled "hadoop@node1:~/Baixades". In the browser, the user browses the "/temporal" directory and views the "File information" for "acces_log". The file is 128 MB in size and is stored in one block. In the terminal, the user compresses "acces_log" into "acces_log.gz" using "gunzip", transfers it to the browser using "hdfs dfs -put", and then decompresses it back to "acces_log" using "tar -xvf". The terminal also shows a warning message about the NativeCodeLoader library.



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

File information - acces_log

Download Head the file (first 32K) Tail the file (last 32K)

Block information --

Block ID: 107374182
Block Pool ID: BP-10
Generation Stamp: 1
Size: 134217728
Availability:
• node1

Block 0

Block 1
Block 2
Block 3

0.2.15-1588667969587

- A continuació mostrarem una sèrie de comandes útils per l'administració del clúster (via terminal)

- **hdfs dfsadmin -report** : On ens diu la informació bàsica el sistema d'arxius i estadístiques

Nota: també ho podem veure des de la interfície gràfica via web a la URL : node1:50070

```
[hadoop@node1 tmp]$ hdfs dfsadmin -report
20/05/05 13:23:11 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Configured Capacity: 52710469632 (49.09 GB)
Present Capacity: 44229148672 (41.19 GB)
DFS Remaining: 43720196096 (40.72 GB)
DFS Used: 508952576 (485.38 MB)
DFS Used%: 1.15%
Under replicated blocks: 0
Blocks with corrupt replicas: 0
Missing blocks: 0
Missing blocks (with replication factor 1): 0
Pending deletion blocks: 0

-----
Live datanodes (1):

Name: 10.0.2.15:50010 (node1)
Hostname: node1
Decommission Status : Normal
Configured Capacity: 52710469632 (49.09 GB)
DFS Used: 508952576 (485.38 MB)
Non DFS Used: 5796966400 (5.40 GB)
DFS Remaining: 43720196096 (40.72 GB)
DFS Used%: 0.97%
DFS Remaining%: 82.94%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 1
Last contact: Tue May 05 13:23:11 CEST 2020
Last Block Report: Tue May 05 12:00:51 CEST 2020

[hadoop@node1 tmp]$
```

- **hdfs fsck /** : Ens fa un check del sistema de fitxers des de l'arrel



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

```
[hadoop@node1 tmp]$ hdfs fsck /
20/05/05 13:25:29 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Connecting to namenode via http://node1:50070/fsck?ugi=hadoop&path=%2F
FSCK started by hadoop (auth:SIMPLE) from /10.0.2.15 for path / at Tue May 05 13:25:30 CEST 2020
..Status: HEALTHY
Total size: 504941537 B
Total dirs: 2
Total files: 2
Total symlinks: 0
Total blocks (validated): 5 (avg. block size 100988307 B)
Minimally replicated blocks: 5 (100.0 %)
Over-replicated blocks: 0 (0.0 %)
Under-replicated blocks: 0 (0.0 %)
Mis-replicated blocks: 0 (0.0 %)
Default replication factor: 1
Average block replication: 1.0
Corrupt blocks: 0
Missing replicas: 0 (0.0 %)
Number of data-nodes: 1
Number of racks: 1
FSCK ended at Tue May 05 13:25:30 CEST 2020 in 4 milliseconds

The filesystem under path '/' is HEALTHY
[hadoop@node1 tmp]$
```

- **hdfs dfsadmin -printTopology** : Ens indica el nombre de nodes que tenim dins la màquina i a quin rack (físic) pertanyen

```
[hadoop@node1 tmp]$ hdfs dfsadmin -printTopology
20/05/05 13:29:00 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Rack: /default-rack
  10.0.2.15:50010 (node1)

[hadoop@node1 tmp]$
```

3.1.4.1.2. HDFS Snapshot

- ❑ És una foto del sistema de fitxers HDFS en un moment determinat.
- ❑ Serveix per fer Backups, recuperacions, manteniment d'un històric, etc.
- ❑ A continuació farem un exemple pràctic de HDFS Snapshot
- ❑ Crearem un directori anomenat **dades** i inserirem un arxiu anomenat **f1.txt**
 - **hdfs dfs -mkdir /dades**
 - **hdfs dfs -put f1.txt /dades**



Nom i Cognoms

Arnau Subirós Puigarnau

Data

02-06-2020

```

hadoop@node1:~$ hdfs dfs -mkdir /dades
20/05/05 13:37:28 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
hadoop@node1:~$ hdfs dfs -ls /
20/05/05 13:37:45 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 2 items
drwxr-xr-x - hadoop supergroup 0 2020-05-05 13:37 /dades
drwxr-xr-x - hadoop supergroup 0 2020-05-05 13:06 /temporal
[hadoop@node1 ~]$ 
[hadoop@node1 ~]$ echo Això es una prova > f1.txt
Això es una prova
[hadoop@node1 ~]$ hdfs dfs -put f1.txt /dades
20/05/05 13:39:08 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
[hadoop@node1 ~]$ hdfs dfs -ls /dades
20/05/05 13:39:33 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 1 items
-rw-r--r-- 1 hadoop supergroup 18 2020-05-05 13:39 /dades/f1.txt
[hadoop@node1 ~]$ 

hadoop@node1:~$ hdfs fsck /dades/f1.txt -blocks -locations
20/05/05 13:42:15 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Connecting to namenode via http://node1:50070/fsck?ugi=hadoop&blocks=1&location=1&path=%2Fdades%2Ff1.txt
FSCK started by hadoop (auth:SIMPLE) from /10.0.2.15 for path /dades/f1.txt at Tue May 05 13:42:14 CEST 2020
.Status: HEALTHY
.Total size: 18 B
.Total dirs: 0
.Total files: 1
.Total symlinks: 0
.Total blocks (validated): 1 (avg. block size 18 B)
.Minimally replicated blocks: 1 (100.0 %)
.Over-replicated blocks: 0 (0.0 %)
.Under-replicated blocks: 0 (0.0 %)
.Mis-replicated blocks: 0 (0.0 %)
.Default replication factor: 1
.Average block replication: 1.0
.Corrupt blocks: 0
.Missing replicas: 0 (0.0 %)
.Number of data-nodes: 1
.Number of racks: 1
FSCK ended at Tue May 05 13:42:14 CEST 2020 in 1 milliseconds

The filesystem under path '/dades/f1.txt' is HEALTHY
[hadoop@node1 ~]$ 

```

Activem snapshot al directori "dades"

- `hdfs dfsadmin -allowSnapshot /dades`

```

hadoop@node1:~$ hdfs dfs -ls /dades
20/05/05 13:47:42 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 1 items
-rw-r--r-- 1 hadoop supergroup 18 2020-05-05 13:39 /dades/f1.txt
[hadoop@node1 ~]$ hdfs dfsadmin -allowSnapshot /dades
20/05/05 13:48:04 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Allowing snapshot on /dades succeeded
[hadoop@node1 ~]$ 

```

Creacio del Snapshot anomenat "snap1"

- `hdfs dfsadmin -createSnapshot /dades snap1`

```

hadoop@node1:~$ hdfs dfs -ls /dades
20/05/05 13:49:44 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 1 items
-rw-r--r-- 1 hadoop supergroup 18 2020-05-05 13:39 /dades/f1.txt
[hadoop@node1 ~]$ 
[hadoop@node1 ~]$ hdfs dfs -createSnapshot /dades snap1
20/05/05 13:50:01 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
created snapshot /dades/.snapshot/snap1
[hadoop@node1 ~]$ 
[hadoop@node1 ~]$ hdfs dfs -ls /dades/.snapshot
20/05/05 13:52:01 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 1 items
drwxr-xr-x - hadoop supergroup 0 2020-05-05 13:50 /dades/.snapshot/snap1
[hadoop@node1 ~]$ 
[hadoop@node1 ~]$ hdfs dfs -ls /dades/.snapshot/snap1
20/05/05 13:53:29 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 1 items
-rw-r--r-- 1 hadoop supergroup 18 2020-05-05 13:39 /dades/.snapshot/snap1/f1.txt
[hadoop@node1 ~]$ 

```



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

- Ja tenim una còpia. Ara farem una prova i eliminarem l'arxiu f1.txt.
- Seguidament visualitzem el contingut del snapshot creat on hi ha l'arxiu eliminat
- Finalment fem una còpia de l'arxiu i ja el tornem a tenir com abans.
 - hdfs dfs -rm /dades/f1.txt
 - hdfs dfs -ls /dades/.snapshot/snap1
 - hdfs dfs -cp /dades/.snapshot/snap1/f1.txt /dades

```
[hadoop@node1 ~]$ hdfs dfs -ls /dades/
20/05/05 13:55:17 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 1 items
-rw-r--r-- 1 hadoop supergroup 18 2020-05-05 13:39 /dades/f1.txt
[hadoop@node1 ~]$
[hadoop@node1 ~]$ hdfs dfs -rm /dades/f1.txt
20/05/05 13:55:43 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Deleted /dades/f1.txt
[hadoop@node1 ~]$ hdfs dfs -ls /dades/
20/05/05 13:55:49 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
[hadoop@node1 ~]$ hdfs dfs -ls /dades/.snapshot/snap1
20/05/05 13:57:10 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 1 items
-rw-r--r-- 1 hadoop supergroup 18 2020-05-05 13:39 /dades/.snapshot/snap1/f1.txt
[hadoop@node1 ~]$ hdfs dfs -cp /dades/.snapshot/snap1/f1.txt /dades
20/05/05 13:58:07 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
[hadoop@node1 ~]$
[hadoop@node1 ~]$ hdfs dfs -ls /dades/
20/05/05 13:58:18 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 1 items
-rw-r--r-- 1 hadoop supergroup 18 2020-05-05 13:58 /dades/f1.txt
[hadoop@node1 ~]$
```

3.1.4.2. Configuració dels processos (Yarn)

- Primer de tot haurem de parar el clúster (dades)

```
[hadoop@node1 ~]$ stop-dfs.sh
Stopping namenodes on [node1]
node1: stopping namenode
localhost: stopping datanode
Stopping secondary namenodes [0.0.0.0]
0.0.0.0: stopping secondarynamenode
[hadoop@node1 ~]$
```



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

- Accedim als fitxers de configuració ubicats a /opt/hadoop/etc/hadoop
- En aquest cas haurem de modificar 2 arxius: **yarn-site.xml** i **mapred-site.xml.template**
- Els arxius **templates** (plantilles) no els hem de modificar, haurem de copiar-lo canviant el nom de l'arxiu (sense acabar amb template :**mapred-site.xml**)
- Editem l'arxiu **mapred-site.xml** que per defecte està buit i afegim la propietat **mapreduce.framework.name** per indicar que el motor serà de tipus Yarn

hadoop@node1:~

```
[hadoop@node1 ~]$ stop-dfs.sh
Stopping namenodes on [node1]
node1: stopping namenode
localhost: stopping datanode
Stopping secondary namenodes [0.0.0.0]
0.0.0.0: stopping secondarynamenode
[hadoop@node1 ~]$
```

hadoop@node1:/opt/hadoop/etc/hadoop

```
[hadoop@node1 hadoop]$ ls
capacity-scheduler.xml      hadoop-metrics.properties  kms-acls.xml
configuration.xml           hadoop-policy.xml        kms-env.sh
container-executor.cfg       hdfs-site.xml          kms-log4j.properties
core-site.xml                hdfs-site.xml~          kms-site.xml
core-site.xml~              httpfs-env.sh          log4j.properties
hadoop-env.cmd              httpfs-log4j.properties  mapred-env.cmd
hadoop-env.sh               httpfs-signature.secret  mapred-env.sh
hadoop-metrics2.properties  httpfs-site.xml        mapred-queues.xml.template
[hadoop@node1 hadoop]$
[hadoop@node1 hadoop]$ cp mapred-site.xml.template mapred-site.xml
[hadoop@node1 hadoop]$
[hadoop@node1 hadoop]$ ls
capacity-scheduler.xml      hadoop-metrics.properties  kms-acls.xml
configuration.xml           hadoop-policy.xml        kms-env.sh
container-executor.cfg       hdfs-site.xml          kms-log4j.properties
core-site.xml                hdfs-site.xml~          kms-site.xml
core-site.xml~              httpfs-env.sh          log4j.properties
hadoop-env.cmd              httpfs-log4j.properties  mapred-env.cmd
hadoop-env.sh               httpfs-signature.secret  mapred-env.sh
hadoop-metrics2.properties  httpfs-site.xml        mapred-queues.xml.template
[hadoop@node1 hadoop]$
[hadoop@node1 hadoop]$ gedit mapred-site.xml
```

mapred-site.xml (/opt/hadoop/etc/hadoop) - gedit

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
Licensed under the Apache License, Version 2.0 (the "License");
you may not use this file except in compliance with the License.
You may obtain a copy of the License at
http://www.apache.org/licenses/LICENSE-2.0

Unless required by applicable law or agreed to in writing, software
distributed under the License is distributed on an "AS IS" BASIS,
WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
See the License for the specific language governing permissions and
limitations under the License. See accompanying LICENSE file.
-->
<!-- Put site-specific property overrides in this file. --&gt;
&lt;configuration&gt;

&lt;!-- PROPIETAT mapreduce.framework.name ||i indiquem el motor sera de tipus YARN --&gt;
&lt;property&gt;
  &lt;name&gt;mapreduce.framework.name&lt;/name&gt;
  &lt;value&gt;yarn&lt;/value&gt;
&lt;/property&gt;

&lt;/configuration&gt;</pre>
```



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

□ Editem l'arxiu **yarn-site.xml** que per defecte està buit.

```

[hadoop@node1 hadoop]$ ls
capacity-scheduler.xml  hadoop-policy.xml  kms-log4j.properties
configuration.xsl        hdfs-site.xml    kms-site.xml
container-executor.cfg   log4j.properties
core-site.xml            httpfs-env.sh    mapred-env.cmd
core-site.xml-           httpfs-log4j.properties
hadoop-env.cmd          httpfs-signature.secret
hadoop-env.sh            hdfs-site.xml
hadoop-metrics2.properties  kms-acls.xml
hadoop-metrics.properties  kms-env.sh
[hadoop@node1 hadoop]$ gedit yarn-site.xml

```

The text editor shows the `yarn-site.xml` file with the following content:

```

<!-- PROPIETAT 1 : Quin és el nom de la màquina on està el gestor del YARN -->
<property>
  <name>yarn.resourcemanager.hostname</name>
  <value>node1</value>
</property>

<!-- PROPIETAT 2 : El gestor de serveis auxiliars que gestionara el mapreduce : MAPREDUCE SHUFFLE -->
<property>
  <name>yarn.nodemanager.aux-services</name>
  <value>mapreduce_shuffle</value>
</property>

<!-- PROPIETAT 3 : Indiquem el valor de mapreduce.shuffle.class -->
<property>
  <name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name>
  <value>org.apache.hadoop.mapred.ShuffleHandler</value>
</property>

```

□ Guardem els canvis i arrenquem les 2 parts el clúster, primer la part de les dades i després la part dels processos. i visualitzem els processos actius de JVM.

- **start-dfs.sh**
- **start-yarn.sh**
- **jps**

```

[hadoop@node1 ~]$ stop-dfs.sh
Stopping namenodes on [node1]
node1: stopping namenode
localhost: stopping datanode
Stopping secondary namenodes [0.0.0.0]
0.0.0.0: stopping secondarynamenode
[hadoop@node1 ~]$
[hadoop@node1 ~]$ start-dfs.sh
Starting namenodes on [node1]
node1: starting namenode, logging to /opt/hadoop/logs/hadoop-hadoop-namenode-node1.out
localhost: starting datanode, logging to /opt/hadoop/logs/hadoop-hadoop-datanode-node1.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /opt/hadoop/logs/hadoop-hadoop-secondarynamenode-node1.out
[hadoop@node1 ~]$
[hadoop@node1 ~]$ start-yarn.sh
starting yarn daemons
starting resourcemanager, logging to /opt/hadoop/logs/yarn-hadoop-resourcemanager-node1.out
localhost: starting nodemanager, logging to /opt/hadoop/logs/yarn-hadoop-nodemanager-node1.out
[hadoop@node1 ~]$
[hadoop@node1 ~]$ jps
9186 SecondaryNameNode
9477 NodeManager
9037 DataNode
9374 ResourceManager
8894 NameNode
9822 Jps
[hadoop@node1 ~]$

```



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

- Si volem el historic dels "jobs" llençats haurem d'activar un altre servei.

- `mr-jobhistory-daemon.sh start historyserver`

```
[hadoop@node1 Escriptori]$ mr-jobhistory-daemon.sh start historyserver
starting historyserver, logging to /opt/hadoop/logs/mapred-hadoop-historyserver-node1.out
[hadoop@node1 Escriptori]$ ^C
[hadoop@node1 Escriptori]$ jps
9186 SecondaryNameNode
17573 Jps
9477 NodeManager
11943 JobHistoryServer
9037 DataNode
9374 ResourceManager
8894 NameNode
[Hadoop Configuration Editor]
```

- Accedim a la interfície web de Resource Manager per defecte el port és el 8088

Browsing HDFS All Applications +

node1:8088/cluster

Més visitades: Apache Hadoop Node1 -Administració...

hadoop

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	Shutdown Nodes
1	0	0	0	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	Maximum Cluster Application P
Capacity Scheduler	[<name=memory-mb default-unit=Mi type=COUNTABLE>, <name=vcores default-unit= type=COUNTABLE>] <memory:1024, vcores:1> <memory:8192, vcores:4>	<memory:1024, vcores:1>	<memory:8192, vcores:4>	0

Show 20 entries

ID	User	Name	Application Type	Queue	Application Priority	StartTime	LaunchTime	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
No data available in table																				

Showing 0 to 0 of 0 entries

First Previous Next



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

3.1.4.2.1. Exemples Mapreduce

□ Utilitzant el jar d'exemples Mapreduce

- Ens descarreguem "El Quijote" <https://www.gutenberg.org/ebooks/2000>

- Crearem un directori HDFS anomenat "llibres" i pujarem l'arxiu

- `hdfs dfs -mkdir /llibres`
- `hdfs dfs -put quijote.txt /llibres`

```

hadoop@node1:~/Baixades$ hadoop dfs -ls
Found 2 items
drwxr-xr-x - hadoop supergroup 0 2020-05-05 14:02 /dades
drwxr-xr-x - hadoop supergroup 0 2020-05-05 13:06 /temporal
[hadoop@node1 ~]$ hdfs dfs -mkdir /llibres
[hadoop@node1 ~]$ hdfs dfs -ls /
Found 3 items
drwxr-xr-x - hadoop supergroup 0 2020-05-05 14:02 /dades
drwxr-xr-x - hadoop supergroup 0 2020-05-05 18:49 /llibres
drwxr-xr-x - hadoop supergroup 0 2020-05-05 13:06 /temporal
[hadoop@node1 ~]$ hdfs dfs -ls /llibres
Found 1 items
-rw-r--r-- 1 hadoop supergroup 2198927 2020-05-05 18:50 /llibres/quijote.txt
[hadoop@node1 ~]$ 

```

- Un cop hem pujat l'arxiu el quijote.txt, crearem un nou fitxer perquè indiqui el número de paraules que té i quantes vegades apareix cada paraula

Nom i Cognoms	Data
Arna Subirós Puigarnau	02-06-2020

- Utilitzem el jar d'exemples de mapreduce ubicat a `/opt/hadoop/share/hadoop/mapreduce` anomenat `hadoop-mapreduce-examples-2.10.jar` (un dels exemples és el `wordcount`) que busca el contingut del directori seleccionat deixant les paraules a un altre directori destí(si no està creat ,es crea)

- `hdfs jar hadoop-mapreduce-examples-2.10.jar wordcount /llibres /sortida_llibres`

The screenshot shows a Firefox browser with several tabs open. The main content area displays the Apache Hadoop cluster interface, which includes sections for Cluster Metrics, Cluster Nodes Metrics, and Scheduler Metrics. A terminal window titled 'hadoop@node1:/opt/hadoop/share/hadoop/mapreduce' is also visible, showing a list of JAR files and some log output related to the mapreduce-examples application.

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed
1	0	1	0

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes
1	0

Scheduler Metrics

Scheduler Type	Schedule Capacity
Capacity Scheduler	<name=memory-mb default-unit=Mi type=COUNTABLE>

MapReduce Applications

ID	User	Name	Application Type	Queue	App Priority
application_1588694315492_0001	hadoop	word count	MAPREDUCE	default	0

Terminal Session (hadoop@node1)

```
hadoop@node1:/opt/hadoop/share/hadoop/mapreduce$ cd ~  
[hadoop@node1 ~]$ clear  
[hadoop@node1 ~]$ cd /opt/hadoop/share/  
[hadoop@node1 share]$ ls  
dec hadoop  
[hadoop@node1 share]$ cd hadoop/mapreduce/  
[hadoop@node1 mapreduce]$ ls  
hadoop-mapreduce-client-app-2.10.0.jar  
hadoop-mapreduce-client-common-2.10.0.jar  
hadoop-mapreduce-client-core-2.10.0.jar  
hadoop-mapreduce-client-hs-2.10.0.jar  
hadoop-mapreduce-client-hs-plugins-2.10.0.jar  
hadoop-mapreduce-client-jobclient-2.10.0.jar  
hadoop-mapreduce-client-jobclient-2.10.0-tests.jar  
hadoop-mapreduce-client-shuffle-2.10.0.jar  
hadoop-mapreduce-examples-2.10.0.jar  
java  
lib  
lib-examples  
sources  
[hadoop@node1 mapreduce]$ hadoop jar hadoop-mapreduce-examples-2.10.0.jar wordcount /llibres /sortida_llibrés  
20/05/05 18:50:25 WARN util.NativeCodeLoader: Unable to load native hadoop libraries for your platform... using builtin-java classes where applicable  
20/05/05 18:50:25 INFO client.RMProxy: Connecting to ResourceManager at node1/10.0.2.15:8083  
20/05/05 18:50:27 INFO input.FileInputFormat: Total input files to process : 1  
20/05/05 18:50:28 INFO mapreduce.JobSubmissionHandler: number of splits:1  
20/05/05 18:50:28 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead use system-metrics-public-publisher.enabled  
20/05/05 18:50:28 INFO mapreduce.JobSubmitter: Submitting application master job: job_1588694315492_0001  
20/05/05 18:50:28 INFO conf.Configuration: resource-types.xml not found  
20/05/05 18:50:28 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.  
20/05/05 18:50:28 INFO resource.ResourceUtils: Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE  
20/05/05 18:50:28 INFO resource.ResourceUtils: Adding resource type - name = vcores, units = , type = COUNTABLE
```

- ❑ Confirmem que l'arxiu de sortida s'ha creat al nou directori **sortida_llibres** i farem un “get” per baixar-lo i poder-lo llegir (li direm **paraules_quijote.txt** i l'ubiquem al directori Baixades)

- `dfs dfs -ls /sortida_llibres`
 - `hdfs dfs -get /sortida_llibres/part-r-00000 /home/hadoop/Baixades/paraules_quijote.txt`
 - `cat /home/hadoop/Baixades/paraules_quijote.txt | more`

```
hadoop@node1:~/Baixades
```

Fitxes Edits Visualitzar Cerca Terminal Ajuda

```
[hadoop@node1 Baixades]$ pwd  
/home/hadoop/Baixades  
[hadoop@node1 Baixades]$ ls  
hadoop-p-1.2.1-80251-linux-x64.rpm paraulas_quijoe.txt pg2000.txt quijote.txt  
[hadoop@node1 Baixades]$ cat paraulas_quijoe.txt | more
```

Fitxes Edits Visualitzar Cerca Terminal Ajuda

```
hadoop@node1:/opt/hadoop/share/hadoop/mapreduce
```

```
[hadoop@node1 mapreduce]$ hdfs dfs -ls /sortida_llibres  
20/05/05 19:02:36 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using built  
lasses where applicable  
output 2 items  
-rw-r--r-- 1 hadoop supergroup 0 2020-05-05 18:58 /sortida_llibres/SUCCESS  
-rwxr--r-- 1 hadoop supergroup 448894 2020-05-05 18:58 /sortida_llibres/part-r-00000  
[hadoop@node1 mapreduce]$  
[hadoop@node1 mapreduce]$ hdfs dfs -get /sortida_llibres/part-r-00000 /home/hadoop/Baixades/paraulas_quijoe.txt  
20/05/05 19:05:59 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using built  
lasses where applicable  
[hadoop@node1 mapreduce]$
```



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

□ Utilitzant codi JAVA

- Farem un exemple de MapReduce desde codi JAVA amb l'arxiu anomenat ComptarParaules.java

```
*ComptarParaules.java
import java.io.IOException;
import java.util.StringTokenizer;
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import org.apache.hadoop.util.GenericOptionsParser;
/**
 * <p>Comptem el número de paraules que apareixen en un document utilitzant MapReduce.
 * El codi té un mapper, reducer, i el programa principal.</p>
 */
public class ComptarParaules {
    /**
     * <p>
     * El mapper s'estre de la interface org.apache.hadoop.mapreduce.Mapper.
     * al executar ejecutar Hadoop , es rep cada línia del fitxer d'entrada com un input
     * La funció map retorna per cada paraula(word) un (word,1) com sortida </p>
     */
    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);
            }
        }
    }
}
```

```
*ComptarParaules.java
}

/*
La funció reduce rep tots els valors que té la mateixa clau com entrada i retorna la clau i el
número de ocurrences com a sortida */
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);
    }
}
/**
 * La entrada es qualsevol fitxer
 */
public static void main(String[] args) throws Exception {
    Configuration conf = new Configuration();
    String[] otherArgs = new GenericOptionsParser(conf, args).getRemainingArgs();
    if (otherArgs.length != 2) {
        System.err.println("Us: ComptarParaules <in> <out>");
        System.exit(2);
    }
    Job job = Job.getInstance(conf, "word count");
    job.setJarByClass(ComptarParaules.class);
    job.setMapperClass(TokenizerMapper.class);

    ****
    //job.setCombinerClass(IntSumReducer.class);

    job.setReducerClass(IntSumReducer.class);
    job.setOutputKeyClass(Text.class);
    job.setOutputValueClass(IntWritable.class);
    FileInputFormat.addInputPath(job, new Path(otherArgs[0]));
    FileOutputFormat.setOutputPath(job, new Path(otherArgs[1]));
    System.exit(job.waitForCompletion(true) ? 0 : 1);
}
}
```



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

- ❑ Com que no tenim IDE de compilació de JAVA (Eclipse, Netbeans,etc) farem la compilació de l'arxiu amb hadoop que ens generarà 3 classes (1 pel Main, un altre pel Mapper i l'altre pel Reduccer)
 - hadoop com.sun.tools.javac.Main ComptarParaules.java
- ❑ Hadoop només admet fitxers jar. Haurem de crear un fitxer jar
 - jar cf ComptarParaules.jar ComptarParaules.class
- ❑ Executem el programa ComptarParaules i el resultat el dipositarà a /temporal/acces_log/sortida_log
 - hadoop jar ComptarParaules.jar ComptarParaules /temporal/acces_log/sortida_log

```

hadoop@node1:~/Escriptori/practiques
[...] 
Starting historyserver... Logging to /opt/hadoop/logs/mapped-hadoop-historyserver-node1.out
[hadoop@node1 ~]$ ./mr-jobhistory-daemon.sh start historyserver
[hadoop@node1 ~]$ hadoop jar ComptarParaules.jar ComptarParaules /temporal/acces_log/sortida_log
20/05/05 19:50:23 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
20/05/05 19:50:23 INFO Client: RMP proxy: Connecting to ResourceManager at node1/10.0.2.15:8082
20/05/05 19:50:23 INFO InputFormat: Total input files to process : 1
20/05/05 19:50:24 INFO mapreduce.JobSubmitter: number of splits:4
20/05/05 19:50:24 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
20/05/05 19:50:25 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1588694315492_0002
20/05/05 19:50:25 INFO conf.Configuration: resource-types.xml not found
20/05/05 19:50:25 INFO ResourceResourceUtils: Adding resource type - name = memory_mb, units = Mi, type = COUNTABLE
20/05/05 19:50:25 INFO ResourceResourceUtils: Adding resource type - name = vcores, units = , type = COUNTABLE
20/05/05 19:50:25 INFO YarnClientImpl: Submitted application application_1588694315492_0002
20/05/05 19:50:25 INFO mapreduce.Job: The url to track the job: http://node1:8088/proxy/application_1588694315492_0002/
20/05/05 19:50:25 INFO mapreduce.Job: Running job: job_1588694315492_0002
^@20/05/05 19:50:33 INFO mapreduce.Job: Job job_1588694315492_0002 running in uber mode : false
20/05/05 19:50:33 INFO mapreduce.Job: map 0% reduce 0%
20/05/05 19:51:02 INFO mapreduce.Job: map 20% reduce 0%
20/05/05 19:51:08 INFO mapreduce.Job: map 22% reduce 0%
20/05/05 19:51:14 INFO mapreduce.Job: map 35% reduce 0%
20/05/05 19:51:22 INFO mapreduce.Job: map 38% reduce 0%
20/05/05 19:51:28 INFO mapreduce.Job: map 48% reduce 0%
20/05/05 19:51:29 INFO mapreduce.Job: map 53% reduce 0%
20/05/05 19:51:35 INFO mapreduce.Job: map 58% reduce 0%
20/05/05 19:51:40 INFO mapreduce.Job: map 59% reduce 0%
20/05/05 19:51:41 INFO mapreduce.Job: map 65% reduce 0%
20/05/05 19:51:47 INFO mapreduce.Job: map 68% reduce 0%
20/05/05 19:51:57 INFO mapreduce.Job: map 70% reduce 0%
20/05/05 19:51:58 INFO mapreduce.Job: map 71% reduce 0%
20/05/05 19:52:05 INFO mapreduce.Job: map 73% reduce 0%
20/05/05 19:52:11 INFO mapreduce.Job: map 76% reduce 0%
20/05/05 19:52:17 INFO mapreduce.Job: map 78% reduce 0%
20/05/05 19:52:23 INFO mapreduce.Job: map 80% reduce 0%

```

```

hadoop@node1:~/Escriptori/practiques
[...] 
20/05/05 19:50:25 INFO resource.ResourceUtils: Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE
20/05/05 19:50:25 INFO resource.ResourceUtils: Adding resource type - name = vcores, units = , type = COUNTABLE
20/05/05 19:50:25 INFO impl.YarnClientImpl: Submitted application application_1588694315492_0002
20/05/05 19:50:25 INFO mapreduce.Job: The url to track the job: http://node1:8088/proxy/application_1588694315492_0002/
20/05/05 19:50:25 INFO mapreduce.Job: Running job: job_1588694315492_0002
^@20/05/05 19:50:33 INFO mapreduce.Job: Job job_1588694315492_0002 running in uber mode : false
20/05/05 19:50:33 INFO mapreduce.Job: map 0% reduce 0%
20/05/05 19:51:02 INFO mapreduce.Job: map 20% reduce 0%
20/05/05 19:51:08 INFO mapreduce.Job: map 22% reduce 0%
20/05/05 19:51:14 INFO mapreduce.Job: map 35% reduce 0%
20/05/05 19:51:22 INFO mapreduce.Job: map 38% reduce 0%
20/05/05 19:51:28 INFO mapreduce.Job: map 48% reduce 0%
20/05/05 19:51:29 INFO mapreduce.Job: map 53% reduce 0%
20/05/05 19:51:35 INFO mapreduce.Job: map 58% reduce 0%
20/05/05 19:51:40 INFO mapreduce.Job: map 59% reduce 0%
20/05/05 19:51:41 INFO mapreduce.Job: map 65% reduce 0%
20/05/05 19:51:47 INFO mapreduce.Job: map 68% reduce 0%
20/05/05 19:51:57 INFO mapreduce.Job: map 70% reduce 0%
20/05/05 19:51:58 INFO mapreduce.Job: map 71% reduce 0%
20/05/05 19:52:05 INFO mapreduce.Job: map 73% reduce 0%
20/05/05 19:52:11 INFO mapreduce.Job: map 76% reduce 0%
20/05/05 19:52:17 INFO mapreduce.Job: map 78% reduce 0%
20/05/05 19:52:23 INFO mapreduce.Job: map 80% reduce 0%

```

- ❑ L'opció "history" la tindrem activa, ja que anteriorment hem iniciat el servei start historyserver)
 - mr-jobhistory-daemon.sh start historyserver



Nom i Cognoms

Arnaud Subirós Puigarnau

Data

02-06-2020

Node1-Master (Captura 9- Cluster 1 node HDFS +YARN) [S'està executant] - Oracle VM VirtualBox

Fitxer Mànica Visualitzar Entrada Dispositius Ajuda

Aplicacions Llocs Sistema

Browsing HDFS Application application_1588694315492_0002 - Mozilla Firefox

node1:8088/cluster/app/application_1588694315492_0002

Més visitades Apache Hadoop Node1 -Administraci... Yarn

Application application_1588694315492_0002

Cluster

- About
- Nodes
- Node Labels
- Applications
 - NEW
 - NEW_SAVING
 - SUBMITTED
 - ACCEPTED
 - BURNING
 - FINISHED
 - FAILED
 - KILLED
- Scheduler

Tools

User: hadoop
Name: word count
Application Type: MAPREDUCE
Application Tags:
Application Priority: 0 (Higher Integer value indicates higher priority)
YarnApplicationState: FINISHED
Queue: default
FinalStatus Reported by AM: SUCCEEDED
Started: dt. de maig 05 19:50:25 +0200 2020
Launched: dt. de maig 05 19:50:25 +0200 2020
Finished: dt. de maig 05 19:55:04 +0200 2020
Elapsed: 4mins, 38sec
Tracking URL: History
Log Aggregation Status: DISABLED

Application Timeout (Remaining Time): Unlimited
Diagnostics:
Unmanaged Application: false
Application Node Label expression: <Not set>
AM container Node Label expression: <DEFAULT_PARTITION>

Total Resource Preempted: <memory:0, vCores:0>

Node1-Master (Captura 9- Cluster 1 node HDFS +YARN) [S'està executant] - Oracle VM VirtualBox

Fitxer Mànica Visualitzar Entrada Dispositius Ajuda

Aplicacions Llocs Sistema

Browsing HDFS MapReduce Job job_1588694315492_0002 - Mozilla Firefox

node1:19888/jobhistory/job/job_1588694315492_0002

Més visitades Apache Hadoop Node1 -Administraci... Yarn

MapReduce Job job_1588694315492_0002

Application

Job

- Overview
- Counters
- Configuration
- Map tasks
- Reduce tasks

Tools

Job Name: word count
User Name: hadoop
Queue: default
State: SUCCEEDED
Uberized: false
Submitted: Tue May 05 19:50:25 CEST 2020
Started: Tue May 05 19:50:31 CEST 2020
Finished: Tue May 05 19:55:03 CEST 2020
Elapsed: 4mins, 31sec
Diagnostics:
Average Map Time 3mins, 4sec
Average Shuffle Time 36sec
Average Merge Time 0sec
Average Reduce Time 33sec

ApplicationMaster

Attempt Number	Start Time	Node
1	Tue May 05 19:50:27 CEST 2020	node1:8042

Task Type	Total	Complete
Map	4	4
Reduce	1	1

Attempt Type	Failed	Killed	Successful
Maps	0	0	4
Reduces	0	0	1



Nom i Cognoms

Arnau Subirós Puigarnau

Data

02-06-2020

Browsing HDFS - Mozilla Firefox

Configuration for MapR... node1:50070/explorer.html#

Més visitades Apache Hadoop Node1 -Administrati... Yarn

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities

Browse Directory

/

Show 25 entries Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	hadoop	supergroup	0 B	May 05 14:02	0	0 B	dades
drwxr-xr-x	hadoop	supergroup	0 B	May 05 18:50	0	0 B	llibres
drwxr-xr-x	hadoop	supergroup	0 B	May 05 18:58	0	0 B	sortida_llibres
drwxr-xr-x	hadoop	supergroup	0 B	May 05 19:55	0	0 B	sortida_log
drwxr-xr-x	hadoop	supergroup	0 B	May 05 13:06	0	0 B	temporal
drwx-----	hadoop	supergroup	0 B	May 05 18:58	0	0 B	tmp

Showing 1 to 6 of 6 entries Previous 1 Next

Hadoop, 2019.

The screenshot shows the HDFS browser interface in Mozilla Firefox. It displays the root directory with six entries: dades, llibres, sortida_llibres, sortida_log, temporal, and tmp. The sortida_log file is highlighted with a red border. The browser has a standard header with tabs for Configuration for MapR..., node1:50070/explorer.html#, and a search bar. Below the header is a navigation bar with links for Més visitades, Apache Hadoop, Node1 -Administrati..., and Yarn. The main content area is titled 'Browse Directory' and shows a table of file entries with columns for Permission, Owner, Group, Size, Last Modified, Replication, Block Size, and Name. The 'sortida_log' entry is the second item in the list.

Browsing HDFS - Mozilla Firefox

Configuration for MapR... node1:50070/explorer.html#/sortida_log

Més visitades Apache Hadoop Node1 -Administrati... Yarn

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities

Browse Directory

/sortida_log

Show 25 entries Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hadoop	supergroup	0 B	May 05 19:55	1	128 MB	_SUCCESS
-rw-r--r--	hadoop	supergroup	55.1 MB	May 05 19:55	1	128 MB	part-r-00000

Showing 1 to 2 of 2 entries Previous 1 Next

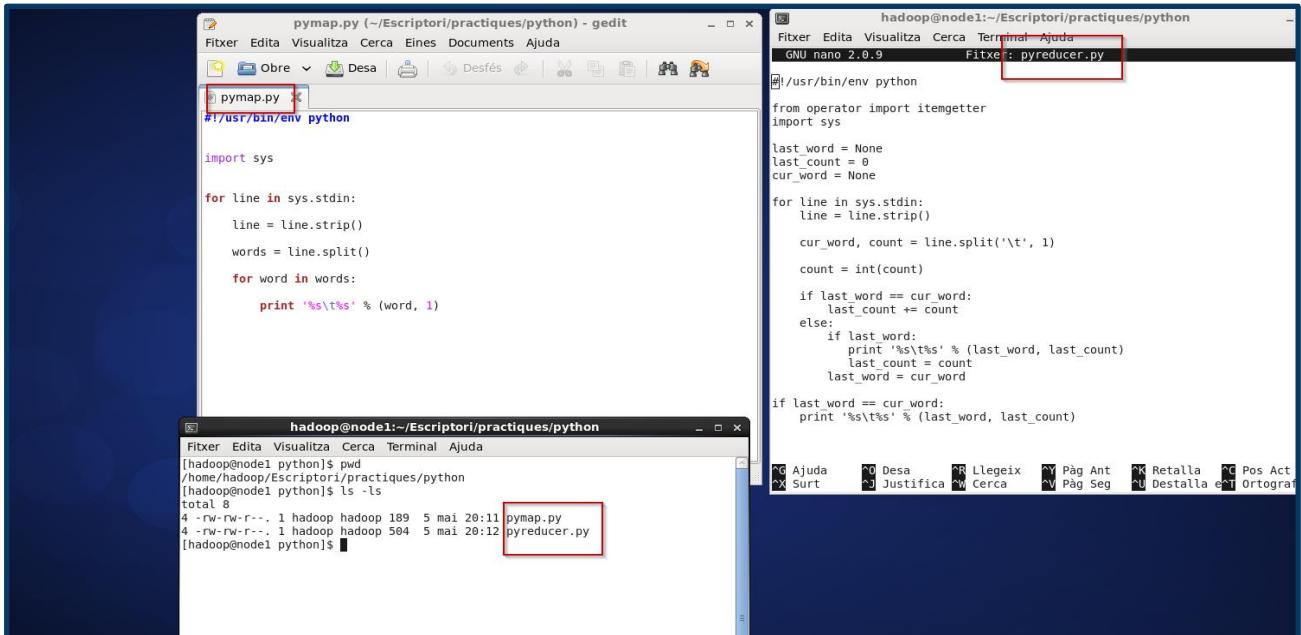
Hadoop, 2019.

The screenshot shows the HDFS browser interface in Mozilla Firefox, specifically viewing the contents of the 'sortida_log' directory. It displays two entries: '_SUCCESS' and 'part-r-00000'. The '_SUCCESS' file is a small file (0 B) from May 05 19:55, while 'part-r-00000' is a larger file (55.1 MB) from the same date. The browser interface is identical to the first screenshot, with a header, navigation bar, and a table-based list of files in the main content area.

Utilitzant codi Python

Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

- ❑ Farem un exemple de MapReduce desde utilitzant codi Python amb *Hadoop streaming* ¹²



```

pymap.py (~/Escriptori/practiques/python) - gedit
Fitxer Edita Visualitza Cerca Eines Documents Ajuda
[...] Obre Desa Desfés | Vorear | Ajuda
pymap.py
#!/usr/bin/env python

import sys

for line in sys.stdin:
    line = line.strip()
    words = line.split()
    for word in words:
        print '%s\t%s' % (word, 1)

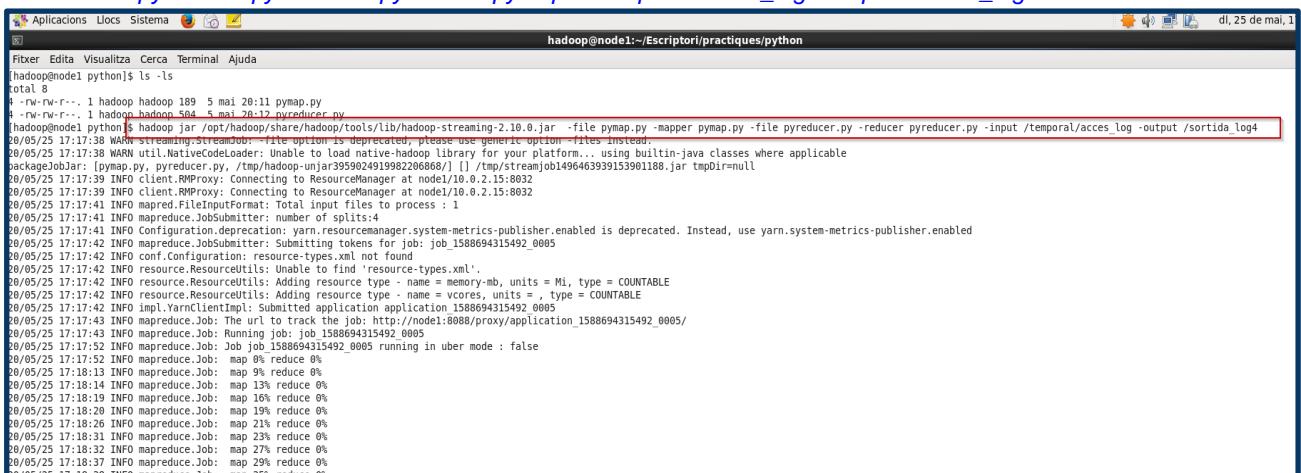
hadoop@node1:~/Escriptori/practiques/python
Fitxer Edita Visualitza Cerca Terminal Ajuda
[hadoop@node1 python]$ pwd
/home/hadoop/Escriptori/practiques/python
[hadoop@node1 python]$ ls -ls
total 8
4 -rw-rw-r--. 1 hadoop hadoop 189 5 mai 20:11 pymap.py
4 -rw-rw-r--. 1 hadoop hadoop 504 5 mai 20:12 pyreducer.py
[hadoop@node1 python]$ 

hadoop@node1:~/Escriptori/practiques/python
Fitxer Edita Visualitza Cerca Terminal Ajuda
[hadoop@node1 python]$ ls
total 8
4 -rw-rw-r--. 1 hadoop hadoop 189 5 mai 20:11 pymap.py
4 -rw-rw-r--. 1 hadoop hadoop 504 5 mai 20:12 pyreducer.py
[hadoop@node1 python]$ 

```

- ❑ Executem l'aplicació amb Hadoop streaming especificant: el mapper ,el reducer, l'arxiu d'origen de les dades i a quina carpeta volem ubicar el resultat(el directori destí no fa falta que existeixi, es crea automàticament)

- *hadoop jar /opt/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.10.0.jar -file pymap.py -mapper pymap.py -file pyreducer.py -reducer pyreducer.py -input /temporal/acces_log -output /sortida_log4*



```

hadoop@node1:~/Escriptori/practiques/python
Fitxer Edita Visualitza Cerca Terminal Ajuda
[hadoop@node1 python]$ ls -ls
total 8
4 -rw-rw-r--. 1 hadoop hadoop 189 5 mai 20:11 pymap.py
4 -rw-rw-r--. 1 hadoop hadoop 504 5 mai 20:12 pyreducer.py
[hadoop@node1 python]$ hadoop jar /opt/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.10.0.jar -file pymap.py -mapper pymap.py -file pyreducer.py -reducer pyreducer.py -input /temporal/acces_log -output /sortida_log4
20/05/25 17:17:38 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
20/05/25 17:17:38 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform.. using builtin-java classes where applicable
packageJobJar: [pymap.py, pyreducer.py, /tmp/hadoop-unja3959024919982206868/] [] /tmp/streamjob1496463939153901188.jar tmpDir=null
20/05/25 17:17:39 INFO client.RMProxy: Connecting to ResourceManager at node1/10.0.2.15:8083
20/05/25 17:17:39 INFO client.RMProxy: Connecting to ResourceManager at node1/10.0.2.15:8083
20/05/25 17:17:41 INFO mapred.FileInputFormat: Total input files to process : 1
20/05/25 17:17:41 INFO mapred.JobClient: number of splits:4
20/05/25 17:17:41 INFO mapred.JobClient: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
20/05/25 17:17:42 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1588694315492_0005
20/05/25 17:17:42 INFO conf.Configuration: resource-types.xml not found
20/05/25 17:17:42 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
20/05/25 17:17:42 INFO resource.ResourceUtils: Adding resource type - name = memory_mb, units = Mi, type = COUNTABLE
20/05/25 17:17:42 INFO resource.ResourceUtils: Adding resource type - name = vcores, units = , type = COUNTABLE
20/05/25 17:17:42 INFO impl.YarnClientImpl: Submitted application application_1588694315492_0005
20/05/25 17:17:43 INFO mapreduce.Job: The url to track the job: http://node1:8088/proxy/application_1588694315492_0005/
20/05/25 17:17:43 INFO mapreduce.Job: Running job: job_1588694315492_0005
20/05/25 17:17:52 INFO mapreduce.Job: Job job_1588694315492_0005 running in uber mode : false
20/05/25 17:17:52 INFO mapreduce.Job: map 0% reduce 0%
20/05/25 17:17:53 INFO mapreduce.Job: map 9% reduce 0%
20/05/25 17:18:01 INFO mapreduce.Job: map 13% reduce 0%
20/05/25 17:18:10 INFO mapreduce.Job: map 16% reduce 0%
20/05/25 17:18:19 INFO mapreduce.Job: map 16% reduce 0%
20/05/25 17:18:20 INFO mapreduce.Job: map 19% reduce 0%
20/05/25 17:18:26 INFO mapreduce.Job: map 21% reduce 0%
20/05/25 17:18:31 INFO mapreduce.Job: map 23% reduce 0%
20/05/25 17:18:32 INFO mapreduce.Job: map 27% reduce 0%
20/05/25 17:18:37 INFO mapreduce.Job: map 29% reduce 0%
20/05/25 17:18:38 INFO mapreduce.Job: map 35% reduce 0%

```

¹² Ens permet fer mappers i reducer amb un altre llenguatge que no sigui Java. S'uneix el mapper i el reducer i ho llença com si fos un "proces Java"

Nom i Cognoms

Arnau Subirós Puigarnau

Data

02-06-2020

```

Aplicacions Llocs Sistema ☰ 🌐 🎯
Fitxer Edita Visualitza Cerca Terminal Ajuda
hadoop@node1:~/Escriptori/practiques/python

20/05/25 17:21:51 INFO mapreduce.Job: map 100% reduce 84%
20/05/25 17:21:57 INFO mapreduce.Job: map 100% reduce 85%
20/05/25 17:22:03 INFO mapreduce.Job: map 100% reduce 87%
20/05/25 17:22:09 INFO mapreduce.Job: map 100% reduce 88%
20/05/25 17:22:22 INFO mapreduce.Job: map 100% reduce 90%
20/05/25 17:22:28 INFO mapreduce.Job: map 100% reduce 91%
20/05/25 17:22:46 INFO mapreduce.Job: map 100% reduce 95%
20/05/25 17:22:52 INFO mapreduce.Job: map 100% reduce 97%
20/05/25 17:22:58 INFO mapreduce.Job: map 100% reduce 99%
20/05/25 17:23:01 INFO mapreduce.Job: map 100% reduce 100%
20/05/25 17:23:03 INFO mapreduce.Job: Job job_1588694315492_0005 completed successfully
20/05/25 17:23:03 INFO mapreduce.Job: Counters: 49
    File System Counters
        FILE: Number of bytes read=1368141470
        FILE: Number of bytes written=2053257276
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        HDFS: Number of bytes read=504954172
        HDFS: Number of bytes written=57779668
        HDFS: Number of read operations=15
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=2
    Job Counters
        Launched map tasks=4
        Launched reduce tasks=1
        Data-local map tasks=4
        Total time spent by all maps in occupied slots (ms)=511429
        Total time spent by all reduces in occupied slots (ms)=205251
        Total time spent by all map tasks (ms)=511429
        Total time spent by all reduce tasks (ms)=205251
        Total vcore-milliseconds taken by all map tasks=511429
        Total vcore-milliseconds taken by all reduce tasks=205251
        Total megabyte-milliseconds taken by all map tasks=523703296
        Total megabyte-milliseconds taken by all reduce tasks=210177024
Map-Reduce Framework
    
```

```

Fitxer Edita Visualitza Cerca Terminal Ajuda
hadoop@node1:~/Escriptori/practiques/python

    HDFS: Number of read operations=15
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=2
Job Counters
    Launched map tasks=4
    Launched reduce tasks=1
    Data-local map tasks=4
    Total time spent by all maps in occupied slots (ms)=511429
    Total time spent by all reduces in occupied slots (ms)=205251
    Total time spent by all map tasks (ms)=511429
    Total time spent by all reduce tasks (ms)=205251
    Total vcore-milliseconds taken by all map tasks=511429
    Total vcore-milliseconds taken by all reduce tasks=205251
    Total megabyte-milliseconds taken by all map tasks=523703296
    Total megabyte-milliseconds taken by all reduce tasks=210177024
Map-Reduce Framework
    Map input records=4477843
    Map output records=44778500
    Map output bytes=594506083
    Map output materialized bytes=684070690
    Input split bytes=352
    Combine input records=0
    Combine output records=0
    Reduce input groups=2475645
    Reduce shuffle bytes=684070690
    Reduce input records=44778500
    Reduce output records=2475645
    Spilled Records=134335500
    Shuffled Maps =4
    Failed Shuffles=0
    Merged Map outputs=4
    GC time elapsed (ms)=844
    CPU time spent (ms)=313860
    Physical memory (bytes) snapshot=1256296448
    Virtual memory (bytes) snapshot=10497081344
    Total committed heap usage (bytes)=865075200
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=504953820
File Output Format Counters
    Bytes Written=57779668
20/05/25 17:23:03 INFO streaming.StreamJob: Output directory: /sortida_log4
[hadoop@node1 python] 

```



Nom i Cognoms

Arnau Subirós Puigarnau

Data

02-06-2020

Welcome to CentOS | All Applications | All Applications

node1:8088/cluster

Més visitades Apache Hadoop Node1 -Administraci... Yarn

hadoop

All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved
5	0	1	4	5	6 GB	8 GB	0 B

Cluster Nodes Metrics

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

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation
Capacity Scheduler	<name=memory-mb default-unit=Mi type=COUNTABLE>, <name=vcores default-unit= type=COUNTABLE>	<memory:1024, vCores:1>	<memory:8192, vCores:1>

Show 20 entries

ID	User	Name	Application Type	Queue	Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU	Allocated Memory MB	Reserved CPU
application_1588694315492_0005	hadoop	streamjob1496463939153901188.jar	MAPREDUCE	default	0	Mon May 25 17:17:43 +0200 2020	Mon May 25 17:17:43 +0200 2020	N/A	RUNNING	UNDEFINED	5	5	6144	0

Welcome to CentOS | All Applications | Application application_1588694315492_0005

node1:8088/cluster/app/application_1588694315492_0005

Més visitades Apache Hadoop Node1 -Administraci... Yarn

hadoop

Application application_1588694315492_0005

Kill Application

User: hadoop
Application Type: MAPREDUCE
Application Token: [View](#)
Application Priority: 0 (Higher Integer value indicates higher priority)
YarnApplicationState: RUNNING: AM has registered with RM and started running.
Queue: default
FinalStatus Reported by AM: Application has not completed yet.
Started: Mon May 25 17:17:42 +0200 2020
Launched: Mon May 25 17:17:43 +0200 2020
Finished: N/A
Elapsed: 2mins, 20sec
Tracking URL: ApplicationMaster
Log Aggregation Status: DISABLED
Application Timeout (Remaining Time): Unlimited
Unmanaged Application: false
Application Node Label expression: <Not set>
AM container Node Label expression: <DEFAULT_PARTITION>

Total Resource Preempted: <memory:0, vCores:0>
Total Number of Non-AM Containers Preempted: 0
Total Number of AM Containers Preempted: 0
Resource Preempted: <memory:0, vCores:0>
Number of Non-AM Containers Preempted from Current Attempt: 0
Aggregate Resource Allocation: 826121 MB-seconds, 663 vcore-seconds
Aggregate Preempted Resource Allocation: 0 MB-seconds, 0 vcore-seconds

Show 20 entries

Attempt ID	Started	Node	Logs	Nodes blacklisted by the app	Nodes blacklisted by the system
appattempt_1588694315492_0005_000001	Mon May 25 17:17:42 +0200 2020	http://node1:8042	Logs	0	0

Welcome to CentOS | All Applications | MapReduce Application application_1588694315492_0005

node1:8088/proxy/application_1588694315492_0005/

Més visitades Apache Hadoop Node1 -Administraci... Yarn

hadoop

MapReduce Application application_1588694315492_0005

Active Jobs

Job ID	Name	State	Map Progress	Maps Total	Maps Completed	Reduce Progress	Reduces Total	Reduces Completed
job_1588694315492_0005	streamjob1496463939153901188.jar	RUNNING	4	4	4	1	0	0

Showing 1 to 1 of 1 entries

Nom i Cognoms

Arnau Subirós Puigarnau

Data

02-06-2020

Welcome to CentOS | All Applications | MapReduce Job job_15... | +

node1:8088/proxy/application_1588694315492_0005/mapreduce/job/job_1588694315492_0005

Més visitades Apache Hadoop Node1 -Administrati... Yarn

MapReduce Job job_1588694315492_0005

Job Name: streamjob1496463939153901188.jar
 User Name: hadoop
 Queue Name: default
 State: RUNNING
 User ID: hadoop
 Started: Mon May 25 17:17:50 CEST 2020
 Elapsed: 3mins, 3sec

ApplicationMaster

Attempt Number	Start Time	Node	Logs
1	Mon May 25 17:17:45 CEST 2020	node1:8042	

Task Type

Map	Reduce	Progress	Total	Pending	Running	Complete
Maps	Reduces	0	1	0	1	0

Attempt Type

New	Running	Failed	Killed	Successful
0	0	0	0	0

node1:50070/explorer.html#/

Més visitades Apache Hadoop Node1 -Administrati... Yarn

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities -

Browse Directory

/

Show 25 entries Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	hadoop	supergroup	0 B	May 05 14:02	0	0 B	dades
drwxr-xr-x	hadoop	supergroup	0 B	May 05 18:50	0	0 B	llibres
drwxr-xr-x	hadoop	supergroup	0 B	May 05 18:58	0	0 B	sortida_llibres
drwxr-xr-x	hadoop	supergroup	0 B	May 05 19:55	0	0 B	sortida_log
drwxr-xr-x	hadoop	supergroup	0 B	May 05 20:24	0	0 B	sortida_log2
drwxr-xr-x	hadoop	supergroup	0 B	May 25 17:16	0	0 B	sortida_log3
drwxr-xr-x	hadoop	supergroup	0 B	May 25 17:22	0	0 B	sortida_log4
drwxr-xr-x	hadoop	supergroup	0 B	May 05 13:06	0	0 B	temporal
drwx-----	hadoop	supergroup	0 B	May 05 18:58	0	0 B	tmp

Showing 1 to 9 of 9 entries Previous 1 Next

node1:50070/explorer.html#/sortida_log4

Més visitades Apache Hadoop Node1 -Administrati... Yarn

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities -

Browse Directory

/sortida_log4

Show 25 entries Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hadoop	supergroup	0 B	May 25 17:22	1	128 MB	SUCCESS
-rw-r--r--	hadoop	supergroup	55.1 MB	May 25 17:22	1	128 MB	part-00000

Showing 1 to 2 of 2 entries Previous 1 Next

Hadoop, 2019.

Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

3.2. Clúster Hadoop (diversos nodes)

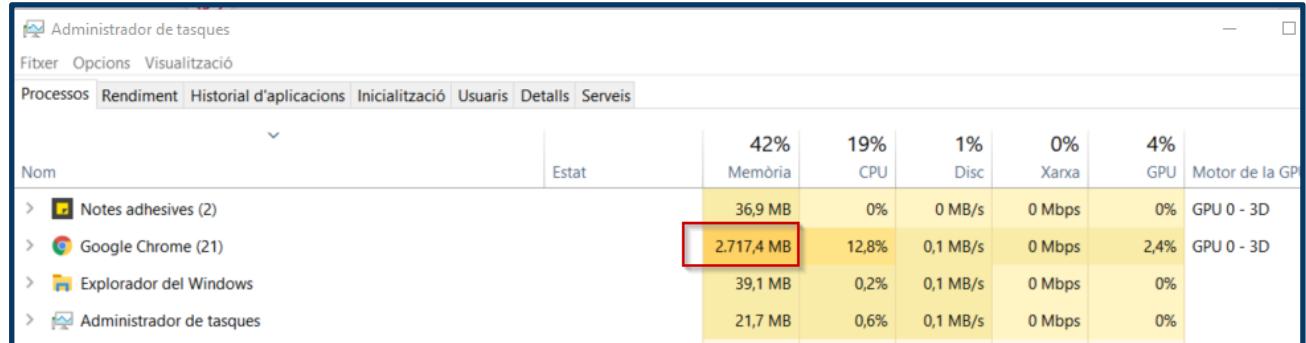
3.2.1. Configuracions prèvies

Per no tornar a instal·lar el sistema operatiu, i diversos paquets que anteriorment hem hagut de descarregar, clonarem la màquina virtual on havíem fet les proves amb el clúster d' Hadoop pseudodistribuit però haurem de tenir en compte:

- Canviar el hostname
- Configurar la xarxa
- Configurar el SSH amb claus públiques

Nota: Hem de tenir en compte que quan tinguem les 3 màquines virtuals enceses (seran 4 GB /host), només tindrem 4 GB disponibles en el host amfitrió, o sigui que tindrem que limitar la memòria RAM. Si obrim el navegador, hauria de ser el indispensable. (Tot i que normalment utilitzo Google Chrome consumeix més memòria que altres navegadors com Firefox)

Actualment tenint 6 pàgines de Google Chrome obertes em consumeix 2.7 GB de memòria RAM



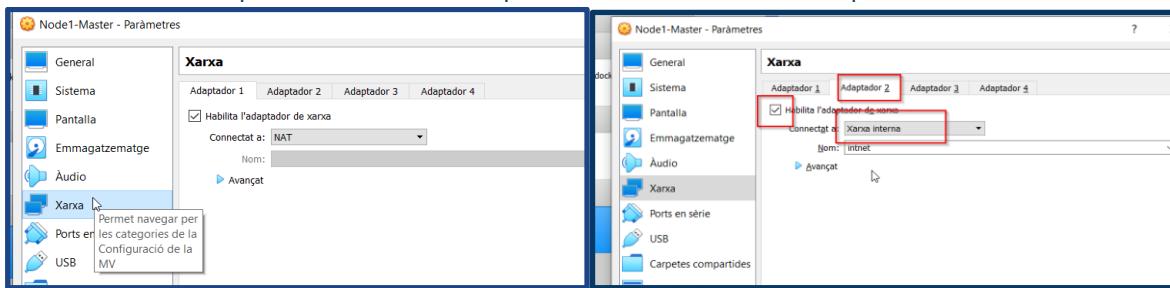


Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

- ❑ Hem de tenir en compte que abans de clonar les màquines hem de parar el clúster de Hadoop (tant dades com processos) i parar la màquina virtual.

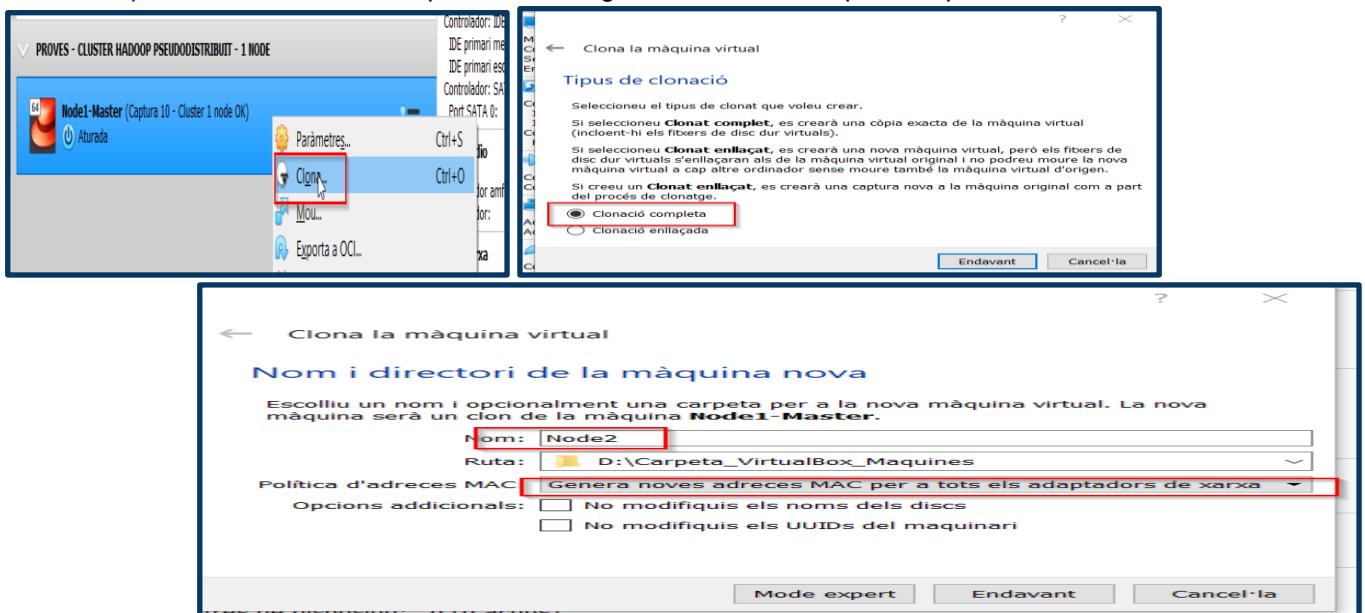
 - [root@node1~]# stop-yarn.sh
 - [root@node1~]# stop-dfs.sh
 - [root@node1~]# init 0

- ❑ Abans afegirem un nou adaptador de xarxa.
 - Adaptador 1 : NAT per conèixer-se a Internet
 - Adaptador 2 : Xarxa interna per la comunicació entre les màquines del clúster.



- ❑ Accedirem al VirtualBox i farem la clonació de la màquina node1 ,3 cops per tenir una màquina "template" i poder fer altres clonacions en cas d'errors o modificacions.

Important: Al clonar les màquines hem de generar noves MAC pels adaptadors de xarxa



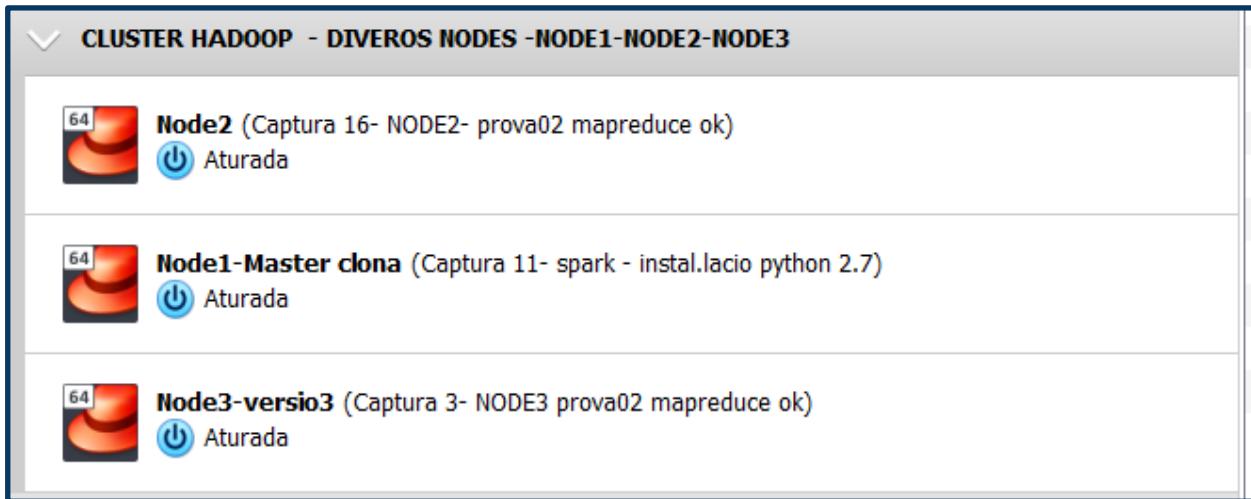


Nom i Cognoms

Arnau Subirós Puigarnau

Data

02-06-2020



3.2.2. Configuració de la xarxa

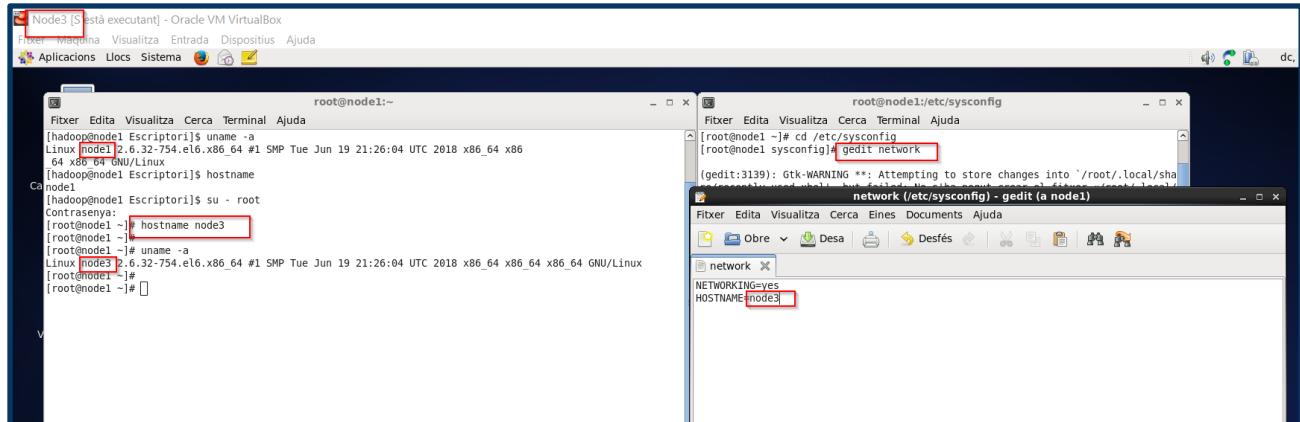
- Un cop ja tenim les màquines clonades, haurem de canviar el nom de les màquines associant-la a una IP estàtica (de la xarxa interna), ja que no disposem de servidor DNS

NODE 2

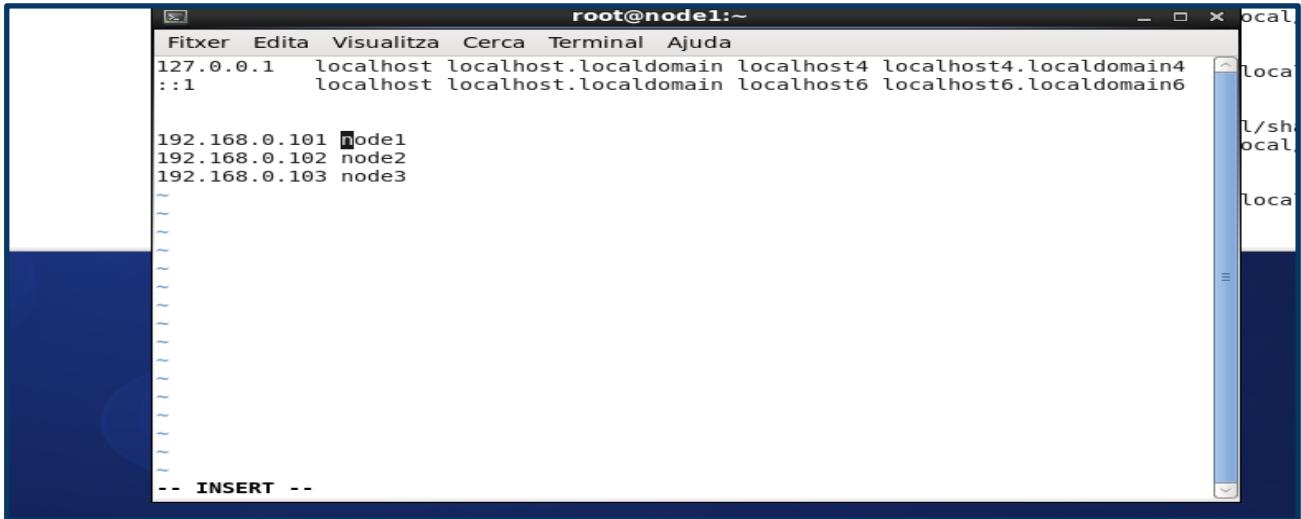
- [root@node1~]# hostname node2
- [root@node2~]# gedit /etc/sysconfig/network
- [root@node2~]# gedit /etc/hosts

NODE 3

- [root@node1~]# hostname node3
- [root@node3~]# gedit /etc/sysconfig/network
- [root@node1~]# gedit /etc/hosts

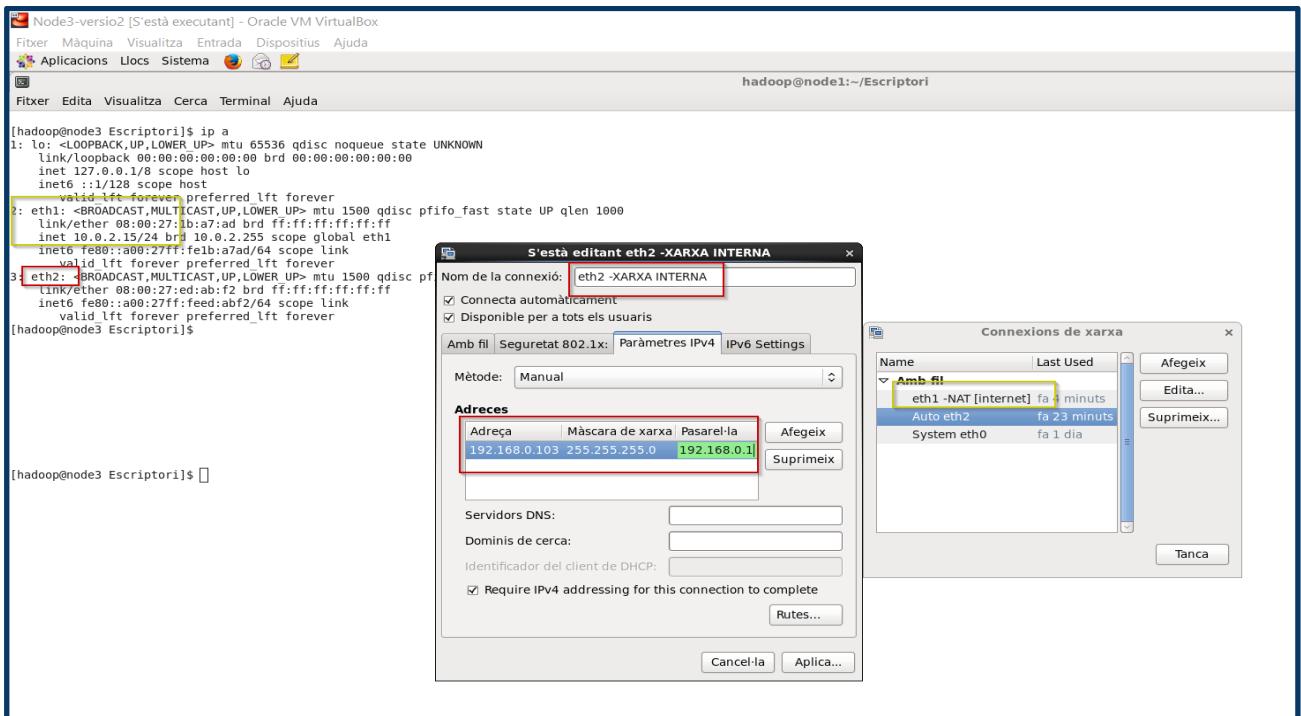


Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020



- Accedim a la interfície de Network Manager per afegir manualment les IP estàtiques . La IP de cada node serà la que anteriorment hem configurat a l'arxiu `/etc/hosts`

NODE 3

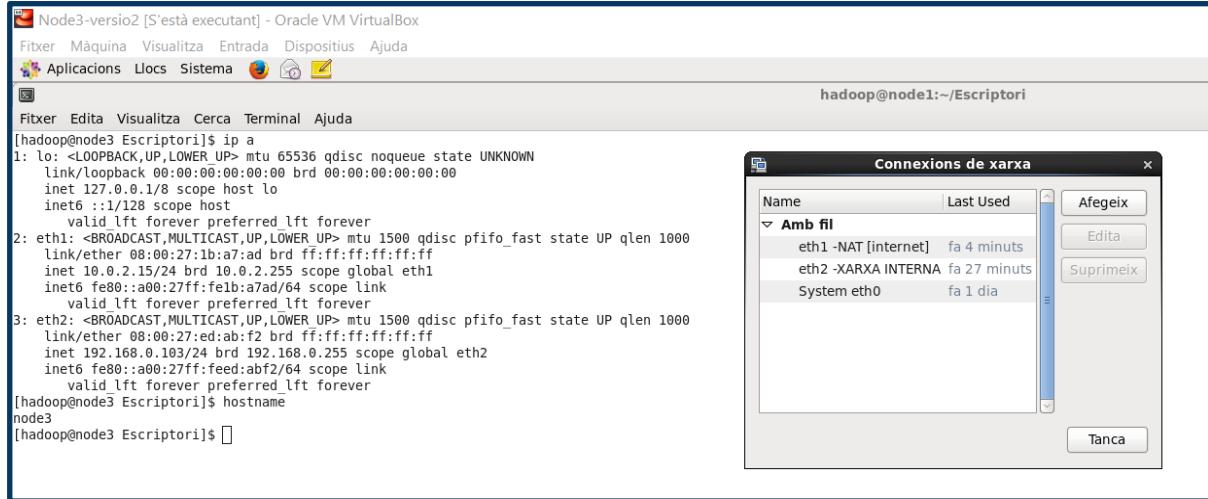


Nom i Cognoms

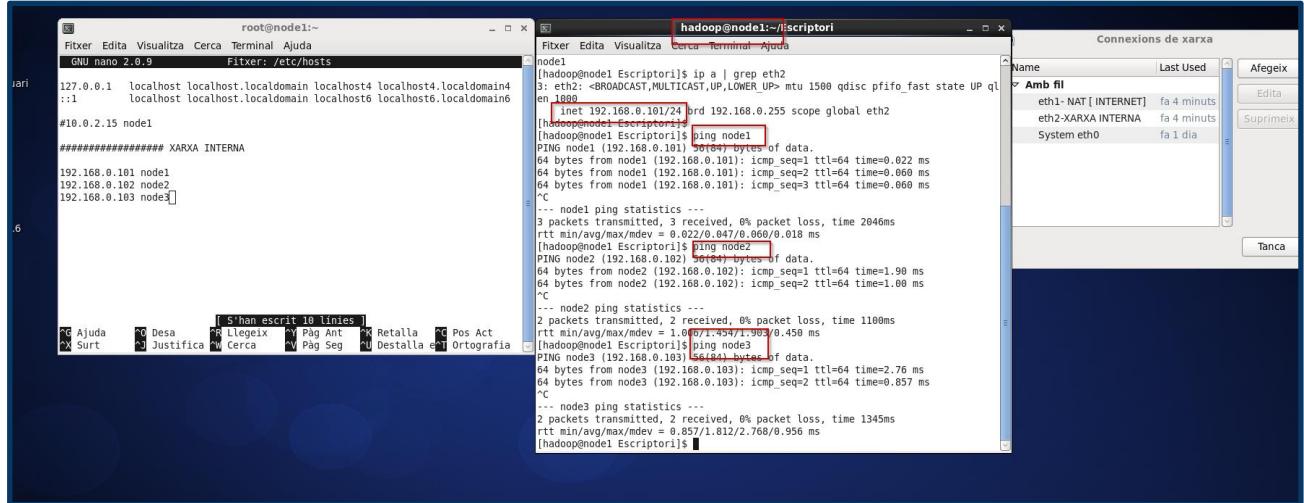
Arnau Subirós Puigarnau

Data

02-06-2020

NODE 3


- Posteriorment fem ping entre les màquines amb el nom del host.



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

3.2.3. Configuració del SSH

- ❑ En la carpeta oculta `~/.ssh` eliminem tots els fitxers per evitar errors (ja que les màquines estan clonades) i ho tornem a generar pels 3 nodes.

NODE1

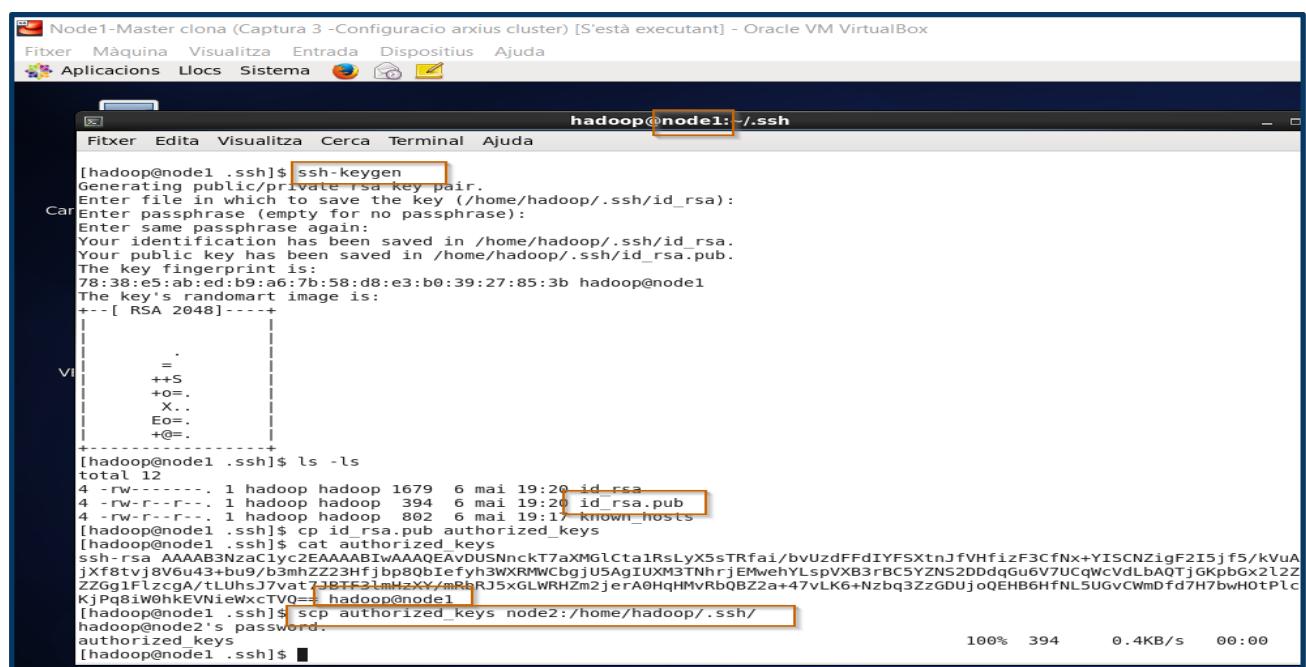
```
[hadoop@node1 .ssh]# rm *
[hadoop@node1 .ssh]# ssh-keygen
[hadoop@node1 .ssh]# cp id_rsa.pub authorized_keys
[hadoop@node1 .ssh]# cat authorized_keys
[hadoop@node1 .ssh]# scp 13authorized_keys node2:/hadoop/.ssh
```

NODE2

```
[hadoop@node2 .ssh]# rm *
[hadoop@node1 .ssh]# ssh-keygen
[hadoop@node1 .ssh]# cat id_rsa.pub >> authorized_keys
[hadoop@node1 .ssh]# cat authorized_keys
[hadoop@node1 .ssh]# scp authorized_keys node3:/hadoop/.ssh
```

NODE3

```
[hadoop@node3 .ssh]# rm *
[hadoop@node3 .ssh]# ssh-keygen
[hadoop@node3 .ssh]# cat id_rsa.pub >> authorized_keys
[hadoop@node3 .ssh]# cat authorized_keys
[hadoop@node3 .ssh]# scp authorized_keys node2:/hadoop/.ssh
[hadoop@node3 .ssh]# scp authorized_keys node1:/hadoop/.ssh
```



¹³ **scp** el fem servir per fer una copia d'un fitxer utilitzant SSH

Nom i Cognoms

Arnau Subirós Puigarnau

Data

02-06-2020

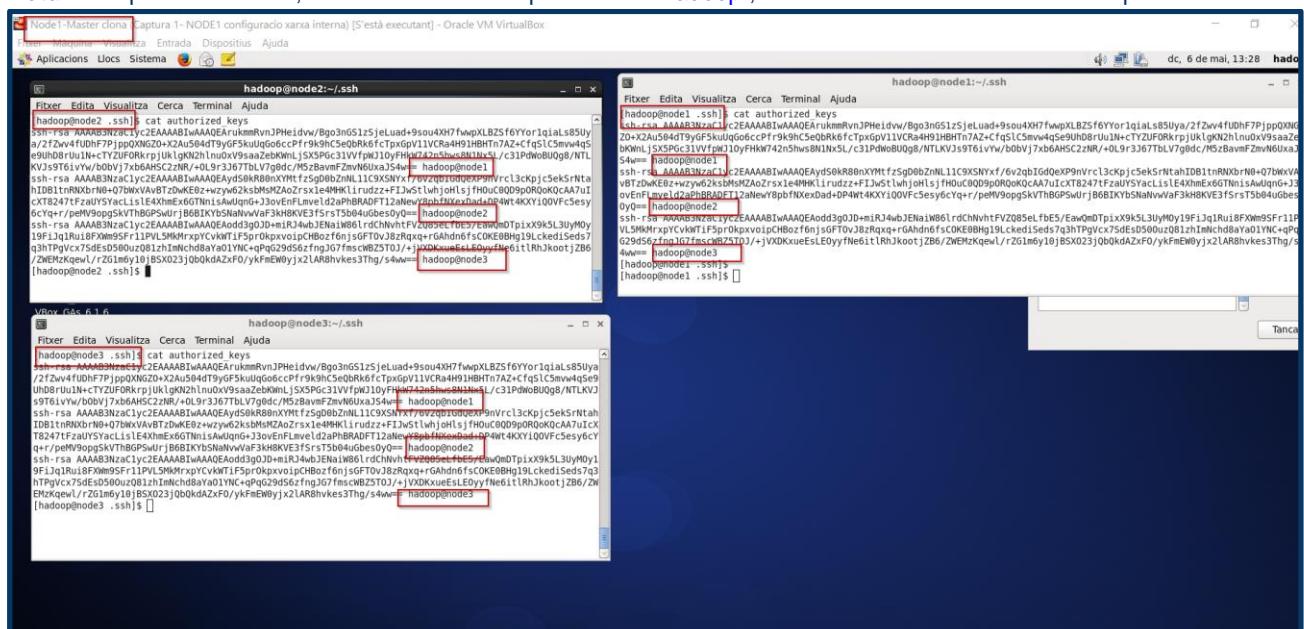
```

Nx+Y
JCQW
NL5U
[hadoop@node3 ~.ssh]$ ls -ls
4 -rw-r--r--. 1 hadoop hadoop 788 6 mai 19:28 authorized_keys
ssh-rsa AAAAB3NzaC1yc2EAAA8IwAAQEAyDUSNckT7aXMGltalRsLyX5sTRfai/bvUzDFFdIYFSXtNjFvHfzfZ3CfNx+YISCNZigF2I5jf5/kVuA5xjXf8tvj8V6u43
+bu9/b3mhZ2Z3Hfjbpb8QbIefyh3wXRMWCbgjU5AgIUXM3TNhrjEMwheYLspVB3rBc5YZNS2DDqGu6V7UCqWcVdLbAQ7jGKpb6x2l2ZkqZGg1Flzcq/tiUhs7J7B
F3lMhZXY/mrBRJ5xGLWRHZmZjera0A0HqHMvRbQbz2a+47vLk6+Nzbq3ZzGDUjoQEHB6HFNL5UGvCwmDfd7H7bwHtPrcsRKjPq81w0hKEVNieWxctV0=hadoop@node1
ssh-rsa AAAAB3NzaC1yc2EAAA8IwAAQEAyVnkUrRnzsm0Wznci=4CoqqMhJNNsQZAEGIMXVmN8BM1t+kNawDZnLRIVRBj0xScxaTi+q051A8RzxqSaLLHf8G07bTPb1
yhZb+qbPgWkfL/X3o6sXhWQhkLM1Zc9itHRNVI72+YJ+0mWGMidIyXb3E97nb7YN56Fq9atP3AKlIaRSWdh81gZT22nQpk1zfsiuTTAsZpxdywsir/F40k49m5b20
0WeckrZHDeP1Serkr7ZtOcp/Hhv/mmaca9DU3G3MeJtca4MhMvhGTCAXcMIZu4zMrI0ohVwn2dNeibT06qZdo0EbrXTClArh/iscXZc7Hs3B0w=hadoop@node2
[hadoop@node3 ~.ssh]$
[hadoop@node3 ~.ssh]$ ssh-keygen
Generating public/private rsa key pair.
Enter file in which to save the key (/home/hadoop/.ssh/id_rsa):
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /home/hadoop/.ssh/id_rsa.
Your public key has been saved in /home/hadoop/.ssh/id_rsa.pub.
The key fingerprint is:
38:c9:b0:64:ad:92:9b:a2:ed:d0:ae:86:e2:52:1e:df hadoop@node3
The key's randomart image is:
+---[ RSA 2048]---+
| |
| |
| + .
| + = o
| o o = S
| .o+ .
| +o+o .
| *=. . E
| B++
+-----+
[hadoop@node3 ~.ssh]$ cat id_rsa.pub >> authorized_keys
[hadoop@node3 ~.ssh]$ scp authorized_keys node2:/home/hadoop/.ssh/
The authenticity of host 'node2' (192.168.0.102) can't be established.
RSA key fingerprint is e6:7d:43:d6:f6:4c:35:22:e3:c9:b4:ab:b3:47:33:4d.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'node2,192.168.0.102' (RSA) to the list of known hosts.
hadoop@node2's password:
authorized_keys
[hadoop@node3 ~.ssh]$ 
```

100% 1182 1.2KB/s 00:00

- Revisem en cada node tenim tots les claus públiques a l'arxiu “**authorized_keys**” i podem establir connexió via SSH entre nodes sense password.

Nota: en aquest casnomés, només està habilitat per l'usuari “**hadoop**”, els altres usuaris necessitarien el password.



```

Node1-Master clona Captura 1-NODE1 configuració xaria interna [5 està executant] - Oracle VM VirtualBox
Filets. Imatges. Visualitzar Entrada Dispositius Ajuda
Aplicacions Llocs Sistema
hadoop@node2:~.ssh$ cat authorized_keys
Filter Editar Visualitzar cerca Terminal Ajuda
hadoop@node2 ~.ssh$ cat authorized_keys
ssh-rsa AAAAB3NzaC1yc2EAAA8IwAAQEAyDUSNckT7aXMGltalRsLyX5sTRfai/bvUzDFFdIYFSXtNjFvHfzfZ3CfNx+YISCNZigF2I5jf5/kVuA5xjXf8tvj8V6u43
+bu9/b3mhZ2Z3Hfjbpb8QbIefyh3wXRMWCbgjU5AgIUXM3TNhrjEMwheYLspVB3rBc5YZNS2DDqGu6V7UCqWcVdLbAQ7jGKpb6x2l2ZkqZGg1Flzcq/tiUhs7J7B
F3lMhZXY/mrBRJ5xGLWRHZmZjera0A0HqHMvRbQbz2a+47vLk6+Nzbq3ZzGDUjoQEHB6HFNL5UGvCwmDfd7H7bwHtPrcsRKjPq81w0hKEVNieWxctV0=hadoop@node1
ssh-rsa AAAAB3NzaC1yc2EAAA8IwAAQEAyVnkUrRnzsm0Wznci=4CoqqMhJNNsQZAEGIMXVmN8BM1t+kNawDZnLRIVRBj0xScxaTi+q051A8RzxqSaLLHf8G07bTPb1
yhZb+qbPgWkfL/X3o6sXhWQhkLM1Zc9itHRNVI72+YJ+0mWGMidIyXb3E97nb7YN56Fq9atP3AKlIaRSWdh81gZT22nQpk1zfsiuTTAsZpxdywsir/F40k49m5b20
0WeckrZHDeP1Serkr7ZtOcp/Hhv/mmaca9DU3G3MeJtca4MhMvhGTCAXcMIZu4zMrI0ohVwn2dNeibT06qZdo0EbrXTClArh/iscXZc7Hs3B0w=hadoop@node2
[hadoop@node3 ~.ssh]$
[hadoop@node3 ~.ssh]$ cat id_rsa.pub >> authorized_keys
[hadoop@node3 ~.ssh]$ scp authorized_keys node2:/home/hadoop/.ssh/
The authenticity of host 'node2' (192.168.0.102) can't be established.
RSA key fingerprint is e6:7d:43:d6:f6:4c:35:22:e3:c9:b4:ab:b3:47:33:4d.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'node2,192.168.0.102' (RSA) to the list of known hosts.
hadoop@node2's password:
authorized_keys
[hadoop@node3 ~.ssh]$ 
```



```

hadoop@node2:~.ssh$ cat authorized_keys
Filter Editar Visualitzar cerca Terminal Ajuda
hadoop@node2 ~.ssh$ cat authorized_keys
ssh-rsa AAAAB3NzaC1yc2EAAA8IwAAQEAyDUSNckT7aXMGltalRsLyX5sTRfai/bvUzDFFdIYFSXtNjFvHfzfZ3CfNx+YISCNZigF2I5jf5/kVuA5xjXf8tvj8V6u43
+bu9/b3mhZ2Z3Hfjbpb8QbIefyh3wXRMWCbgjU5AgIUXM3TNhrjEMwheYLspVB3rBc5YZNS2DDqGu6V7UCqWcVdLbAQ7jGKpb6x2l2ZkqZGg1Flzcq/tiUhs7J7B
F3lMhZXY/mrBRJ5xGLWRHZmZjera0A0HqHMvRbQbz2a+47vLk6+Nzbq3ZzGDUjoQEHB6HFNL5UGvCwmDfd7H7bwHtPrcsRKjPq81w0hKEVNieWxctV0=hadoop@node1
ssh-rsa AAAAB3NzaC1yc2EAAA8IwAAQEAyVnkUrRnzsm0Wznci=4CoqqMhJNNsQZAEGIMXVmN8BM1t+kNawDZnLRIVRBj0xScxaTi+q051A8RzxqSaLLHf8G07bTPb1
yhZb+qbPgWkfL/X3o6sXhWQhkLM1Zc9itHRNVI72+YJ+0mWGMidIyXb3E97nb7YN56Fq9atP3AKlIaRSWdh81gZT22nQpk1zfsiuTTAsZpxdywsir/F40k49m5b20
0WeckrZHDeP1Serkr7ZtOcp/Hhv/mmaca9DU3G3MeJtca4MhMvhGTCAXcMIZu4zMrI0ohVwn2dNeibT06qZdo0EbrXTClArh/iscXZc7Hs3B0w=hadoop@node2
[hadoop@node3 ~.ssh]$
[hadoop@node3 ~.ssh]$ cat id_rsa.pub >> authorized_keys
[hadoop@node3 ~.ssh]$ 
```



```

hadoop@node3:~.ssh$ cat authorized_keys
Filter Editar Visualitzar cerca Terminal Ajuda
hadoop@node3 ~.ssh$ cat authorized_keys
ssh-rsa AAAAB3NzaC1yc2EAAA8IwAAQEAyDUSNckT7aXMGltalRsLyX5sTRfai/bvUzDFFdIYFSXtNjFvHfzfZ3CfNx+YISCNZigF2I5jf5/kVuA5xjXf8tvj8V6u43
+bu9/b3mhZ2Z3Hfjbpb8QbIefyh3wXRMWCbgjU5AgIUXM3TNhrjEMwheYLspVB3rBc5YZNS2DDqGu6V7UCqWcVdLbAQ7jGKpb6x2l2ZkqZGg1Flzcq/tiUhs7J7B
F3lMhZXY/mrBRJ5xGLWRHZmZjera0A0HqHMvRbQbz2a+47vLk6+Nzbq3ZzGDUjoQEHB6HFNL5UGvCwmDfd7H7bwHtPrcsRKjPq81w0hKEVNieWxctV0=hadoop@node1
ssh-rsa AAAAB3NzaC1yc2EAAA8IwAAQEAyVnkUrRnzsm0Wznci=4CoqqMhJNNsQZAEGIMXVmN8BM1t+kNawDZnLRIVRBj0xScxaTi+q051A8RzxqSaLLHf8G07bTPb1
yhZb+qbPgWkfL/X3o6sXhWQhkLM1Zc9itHRNVI72+YJ+0mWGMidIyXb3E97nb7YN56Fq9atP3AKlIaRSWdh81gZT22nQpk1zfsiuTTAsZpxdywsir/F40k49m5b20
0WeckrZHDeP1Serkr7ZtOcp/Hhv/mmaca9DU3G3MeJtca4MhMvhGTCAXcMIZu4zMrI0ohVwn2dNeibT06qZdo0EbrXTClArh/iscXZc7Hs3B0w=hadoop@node2
[hadoop@node3 ~.ssh]$
[hadoop@node3 ~.ssh]$ 
```



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

3.2.4. Configuració del clúster

- Abans de començar la configuració del clúster, hem de tenir en compte:

- El mateix usuari en tots els nodes del clúster (en el nostre cas, no suposarà cap problema, ja que les màquines estan clonades i fan servir el mateix usuari "hadoop")
- Els 3 nodes han de ser d'accessibles via SSH sense contrasenya
- Tenir els mateixos directoris de Hadoop en cada node
- Tenir copiat el software de Hadoop en el mateix directori.
- Tenir creat el directori de dades en tots els nodes i els seus permisos
 - /dades/datanode

(no és obligatori però és recomanable, els servidors haurien de ser clons en l'àmbit de Hadoop no només en l'àmbit de software sinó també en l'àmbit de hardware per un millor rendiment de les màquines)

- Haurem de modificar el directori /dades, ja que anteriorment el node1 feia les funcions de "master" i "slave"

HOST (master) : node1

- En el directori /dades hi ha els directoris:
 - **namenode**¹⁴
 - **datanode** : Eliminem el directori i els arxius que hi hagi, ja que només pot estar en el node slave
 - [hadoop@node1 dades]# rm -rf datanode

HOST (slave) : node2

- En el directori /dades hi ha els directoris:
 - **namenode** : Eliminem el directori i els arxius que hi hagi, ja que només pot estar en el node master
 - [hadoop@node2 dades]# rm -rf namenode
 - **datanode** : En el datanode, com que és una còpia, eliminarem el directori current (faria referència al node1)
 - [hadoop@node2 datanode]# rm -rf current

HOST (slave) : node3

- En el directori /dades hi ha els directoris:
 - **namenode** : Eliminem el directori i els arxius que hi hagi, ja que només pot estar en el node master
 - [hadoop@node3 dades]# rm -rf namenode
 - **datanode** : En el datanode, com que és una còpia, eliminarem el directori current (faria referència al node1)
 - [hadoop@node3 datanode]# rm -rf current

- Haurem d'accedir al directori /opt/hadoop/etc/hadoop per modificar els arxius de configuració.

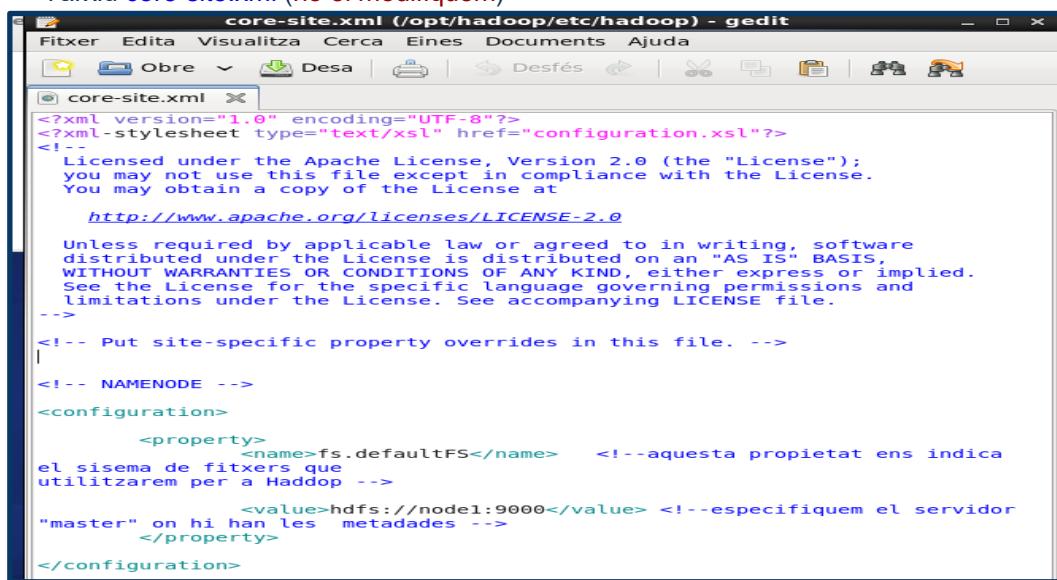
¹⁴ També podríem eliminar el contingut del **namenode**, però com que abans d'iniciar el clúster haurem de fer un format ja ho esborrarà.

Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

- core-site.xml
- hdfs-site.xml
- mapred-site.xml
- yarn-site.xml

HOST (master) : node1

- l'arxiu **core-site.xml** (no el modifiquem)



```

<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
Licensed under the Apache License, Version 2.0 (the "License");
you may not use this file except in compliance with the License.
You may obtain a copy of the License at

http://www.apache.org/licenses/LICENSE-2.0

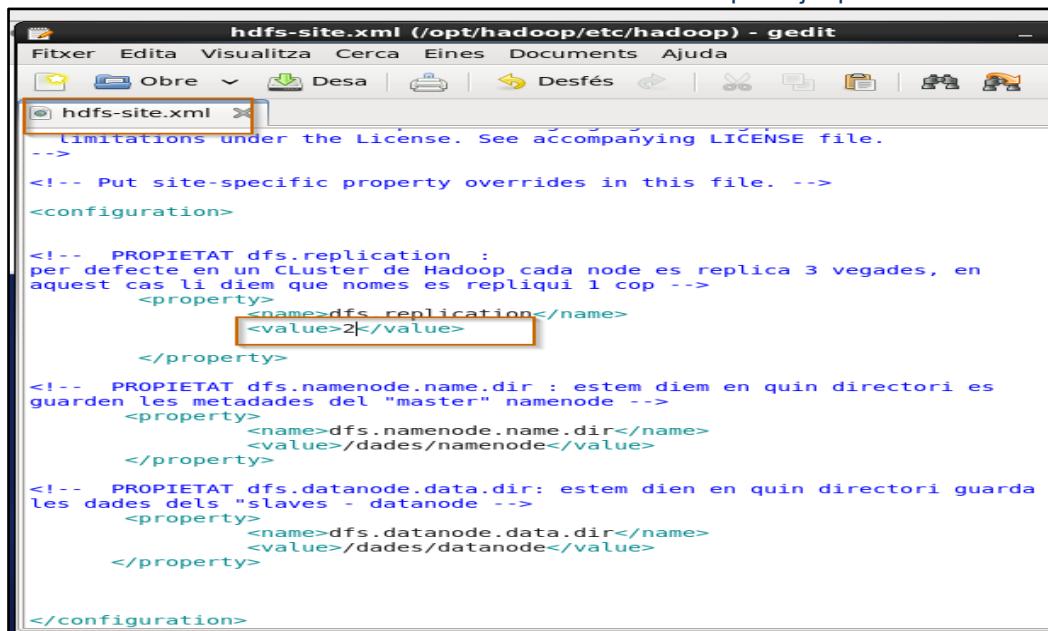
Unless required by applicable law or agreed to in writing, software
distributed under the License is distributed on an "AS IS" BASIS,
WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
See the License for the specific language governing permissions and
limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->
| 
<!-- NAMENODE -->

<configuration>
    <property>
        <name>fs.defaultFS</name>      <!--aquesta propietat ens indica
el sisema de fitxers que
utilitzarem per a Hadoop -->
        <value>hdfs://node1:9000</value> <!--especifiquem el servidor
"master" on hi han les metadades -->
    </property>
</configuration>

```

- l'arxiu **hdfs-site.xml** s'haurà de modificar el número de rèplica ja que tindrem 2 "slaves"



```

<!--
LIMITATIONS under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->
<configuration>

    <!-- PROPIETAT dfs.replication :
per defecte en un CLuster de Hadoop cada node es replica 3 vegades, en
aquest cas li diem que només es repliqui 1 cop -->
    <property>
        <name>dfs.replication</name>
        <value>2</value>
    </property>

    <!-- PROPIETAT dfs.namenode.name.dir : estem diem en quin directori es
guarden les metadades del "master" namenode -->
    <property>
        <name>dfs.namenode.name.dir</name>
        <value>/dades/namenode</value>
    </property>

    <!-- PROPIETAT dfs.datanode.data.dir: estem dien en quin directori guarda
les dades dels "slaves" - datanode -->
    <property>
        <name>dfs.datanode.data.dir</name>
        <value>/dades/datanode</value>
    </property>

</configuration>

```



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

- Un cop modificat l'arxiu hdfs-site.xml l'enviarem als altres nodes utilitzant el comando `scp`¹⁵

- [hadoop@node1 hadoop]# `scp hdfs-site.xml node2:/opt/hadoop/etc/hadoop`
- [hadoop@node1 hadoop]# `scp hdfs-site.xml node3:/opt/hadoop/etc/hadoop`

```
hadoop@node1:/opt/hadoop/etc/hadoop
Fitxer Edita Visualitza Cerca Terminal Ajuda
/opt/hadoop/etc/hadoop
[hadoop@node1 hadoop]$ gedit core-site.xml
[hadoop@node1 hadoop]$ gedit hdfs-site.xml
[hadoop@node1 hadoop]$ gedit hdfs-site.xml
[hadoop@node1 hadoop]$ 
[hadoop@node1 hadoop]$ scp hdfs-site.xml node2:/opt/hadoop/etc/hadoop/
hdfs-site.xml
[hadoop@node1 hadoop]$ scp hdfs-site.xml node3:/opt/hadoop/etc/hadoop/
100% 1460    1.4KB/s  00:00
```

- l'arxiu `mapred-site.xml` (no el modifiquem)

```
mapred-site.xml
<?xml version="1.0"?>
<?xmlstylesheet type="text/xsl" href="configuration.xsl"?>
<!--
Licensed under the Apache License, Version 2.0 (the "License");
you may not use this file except in compliance with the License.
You may obtain a copy of the License at

http://www.apache.org/licenses/LICENSE-2.0

Unless required by applicable law or agreed to in writing, software
distributed under the License is distributed on an "AS IS" BASIS,
WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
See the License for the specific language governing permissions and
limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>

<!-- PROPIETAT mapreduce.framework.name | li indiquem el motor sera de tipus
YARN -->
<property>
  <name>mapreduce.framework.name</name>
  <value>yarn</value>
</property>

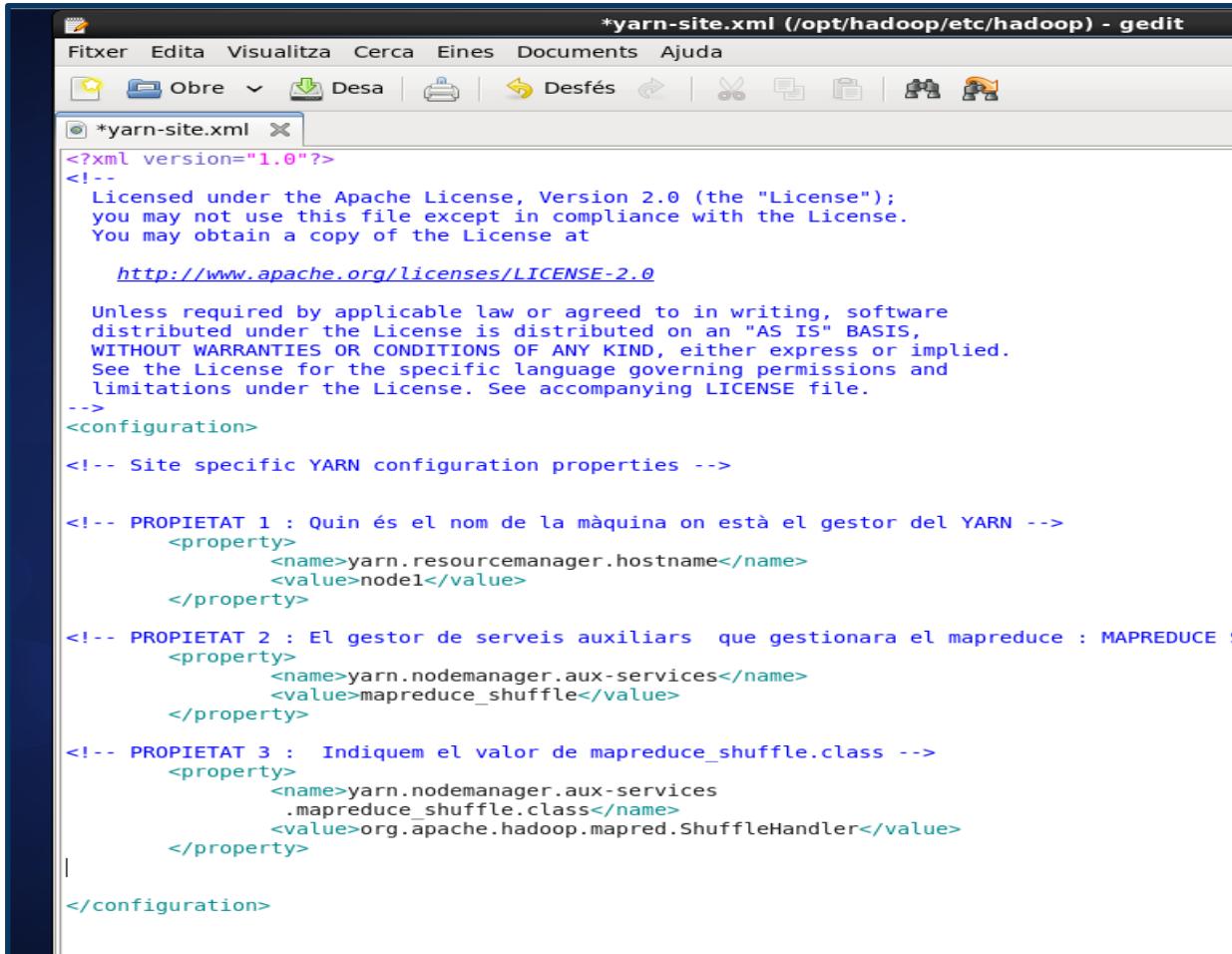
</configuration>
```

¹⁵ Farem una còpia de l'arxiu via SSH



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

- l'arxiu `yarn-site.xml` (no el modifiquem)



```
<?xml version="1.0"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

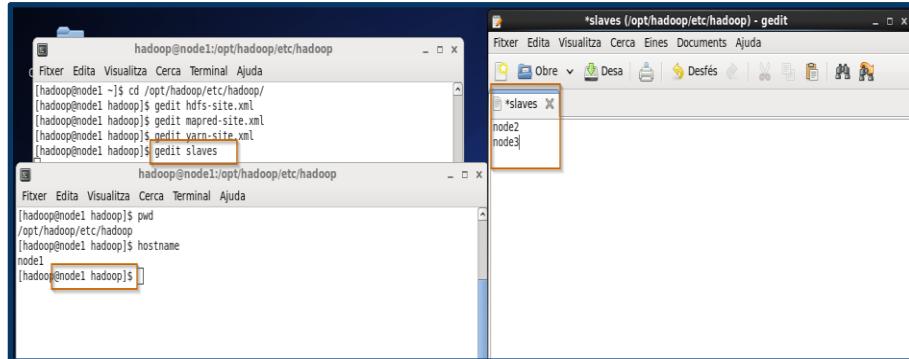
  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->
<configuration>
  <!-- Site specific YARN configuration properties -->

  <!-- PROPIETAT 1 : Quin és el nom de la màquina on està el gestor del YARN -->
  <property>
    <name>yarn.resourcemanager.hostname</name>
    <value>node1</value>
  </property>

  <!-- PROPIETAT 2 : El gestor de serveis auxiliars que gestionara el mapreduce : MAPREDUCE S -->
  <property>
    <name>yarn.nodemanager.aux-services</name>
    <value>mapreduce_shuffle</value>
  </property>

  <!-- PROPIETAT 3 : Indiquem el valor de mapreduce_shuffle.class -->
  <property>
    <name>yarn.nodemanager.aux-services
      .mapreduce_shuffle.class</name>
    <value>org.apache.hadoop.mapred.ShuffleHandler</value>
  </property>
| 
</configuration>
```

- Haurem de modificar l'arxiu "slaves" que per defecte només consta localhost
 - [hadoop@node1 hadoop]# gedit slaves



```
[hadoop@node1:~/opt/hadoop/etc/hadoop]$ cd /opt/hadoop/etc/hadoop/
[hadoop@node1 hadoop]$ gedit hdfs-site.xml
[hadoop@node1 hadoop]$ gedit mapred-site.xml
[hadoop@node1 hadoop]$ gedit yarn-site.xml
[hadoop@node1 hadoop]$ gedit slaves
[hadoop@node1 hadoop]$
```

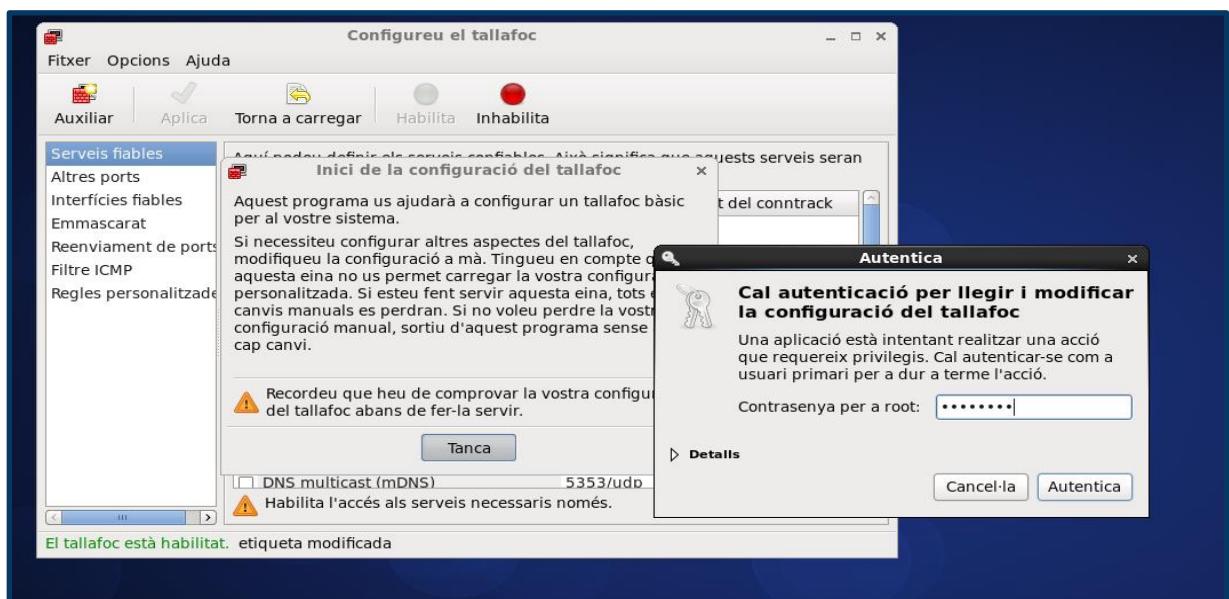
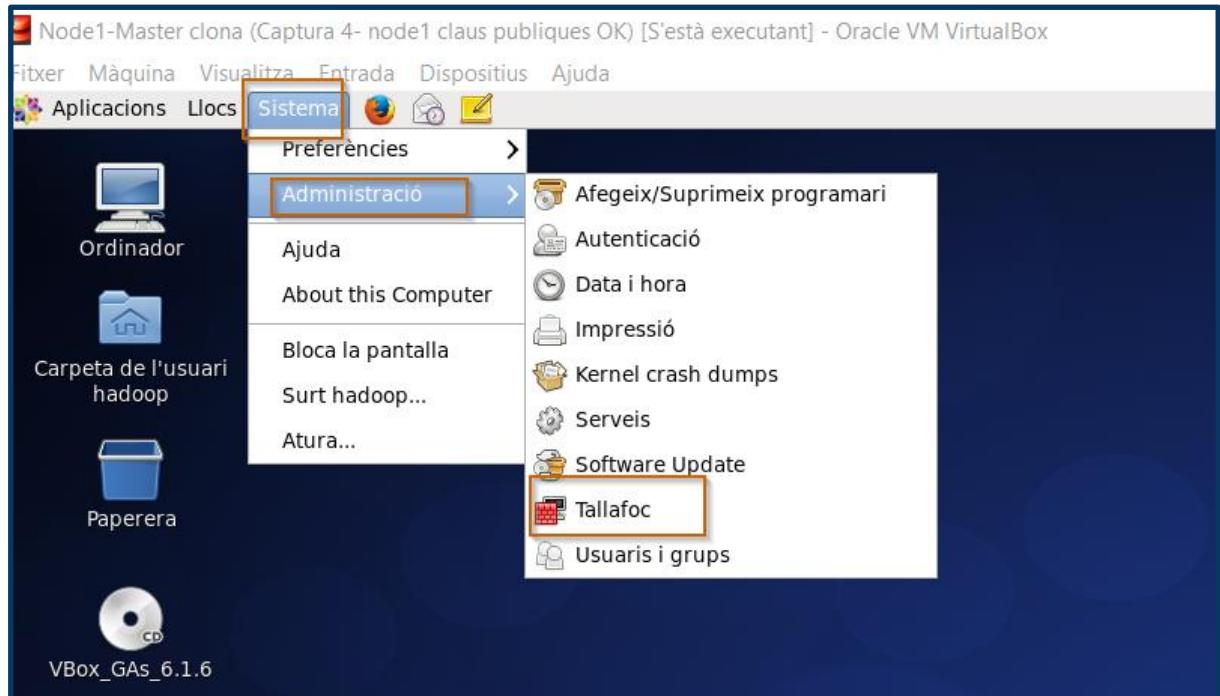
The slaves file content is shown in the terminal:

```
node2
node3
```

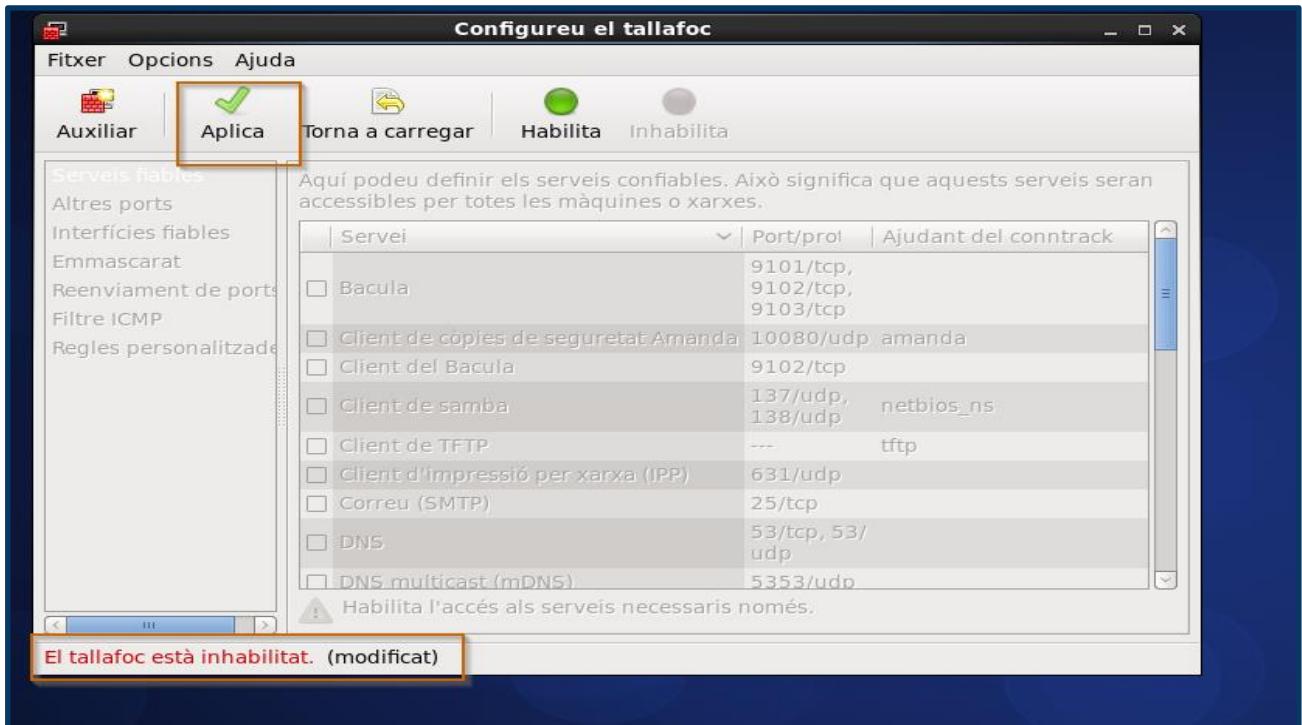


Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

- Hem de tenir en compte que el sistema operatiu CentOS per defecte té el **firewall** actiu, i només permet les connexions al port 22 (SSH). Desactivarem els **firewall** per aquestes pràctiques tot i que l'aconsellable seria crear regles en les **iptables** obrint els ports que ens interessin.



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020



□ Formatejar el **namenode**¹⁶

- [hadoop@node1 hadoop]# hdfs namenode -format

```
[x] hadoop@node1:dades  
Fitxa Edita Visualitza Cerca Terminal Ajuda  
[hadoop@node1 dades]$ pwd  
/dades  
[hadoop@node1 dades]$ hdfs namenode -format
```

```
hadoop@node1:/dades
```

Fitxer Edita Visualiza Cerca Terminal Ajuda

```
namenode/current/edits_000000000000000006-0000000000000000000007, /dades/namenode/current/edits_000000000000000053-0000000000000000153, /dades/namenode/current/edits_inprogress_00000000000000248, /dades/namenode/current/edits_000000000000000003-0000000000000000064, /dades/namenode/current/edits_000000000000000001-000000000000000002, /dades/namenode/current/edits_000000000000000008-0000000000000016, /dades/namenode/current/fsimage_00000000000000247.md5, /dades/namenode/current/edits_000000000000000050-000000000000000050, /dades/namenode/current/fsimage_00000000000000247, /dades/namenode/current/edits_000000000000000005-000000000000000005, /dades/namenode/current/seen_txid, /dades/namenode/current/edits_000000000000000051-0000000000000052]
```

20/05/06 20:23:10 INFO common.Storage: Storage directory /dades/namenode has been successfully formatted.

20/05/06 20:23:10 INFO namenode.FSImageFormatProtobuf: Saving image file /dades/namenode/current/fsimage.ckpt_0000000000000000 using no compression

20/05/06 20:23:10 INFO namenode.FSImageFormatProtobuf: Image file /dades/namenode/current/fsimage.ckpt_0000000000000000 of size 324 bytes saved in 0 seconds .

20/05/06 20:23:10 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0

20/05/06 20:23:10 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid = 0 when me et shutdown.

20/05/06 20:23:10 INFO namenode.NameNode: SHUTDOWN MSG:

SHUTDOWN MSG: Shutting down NameNode at node1/192.168.0.101

[hadoop@node1 dades]\$

¹⁶Està en el node mater (node1)



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

□ Arrencar el Namenode

[hadoop@node1 hadoop]# hdfs-dfs.sh

```
hadoop@node1:/dades/namenode
Fitxer Edita Visualitza Cerca Terminal Ajuda
[hadoop@node1 namenode]$ pwd
/dades/namenode
[hadoop@node1 namenode]$ ls
current
[hadoop@node1 namenode]$ start-dfs.sh
20/05/06 20:27:19 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Starting namenodes on [node1]
node1: starting namenode, logging to /opt/hadoop/logs/hadoop-namenode-node1.out
node2: starting datanode, logging to /opt/hadoop/logs/hadoop-hadoop-datanode-node2.out
node3: starting datanode, logging to /opt/hadoop/logs/hadoop-hadoop-datanode-node3.out
Starting secondary namenodes [0.0.0.0]
The authenticity of host '0.0.0.0 (0.0.0.0)' can't be established.
RSA key fingerprint is e6:7d:43:d6:f6:4c:35:22:e3:c9:b4:ab:b3:47:33:4d.
Are you sure you want to continue connecting (yes/no)? yes
0.0.0.0: Warning: Permanently added '0.0.0.0' (RSA) to the list of known hosts.
0.0.0.0: starting secondarynamenode, logging to /opt/hadoop/logs/hadoop-secondarynamenode-node1.out
20/05/06 20:27:47 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
[hadoop@node1 namenode]$ jps
5571 Jps
5242 NameNode
5452 SecondaryNameNode
[hadoop@node1 namenode]$
```

□ Confirmem utilitzant el comando jps que tenim els processos en els nodes "slaves"

[hadoop@node2 hadoop]# jps

```
Node2 (Captura 14 - node2 claus publiques OK) [S'està executant] - Oracle VM VirtualBox
Fitxer MÀquina Visualitza Entrada Dispositius Ajuda
Aplicacions Llocs Sistema
hadoop@node2:~
Fitxer Edita Visualitza Cerca Terminal Ajuda
[hadoop@node2 ~]$ jps
5515 Jps
5375 DataNode
[hadoop@node2 ~]$
```

[hadoop@node3 hadoop]# jps

```
Node3-versio3 (Captura 1 node3 claus publiques OK) [S'està executant] - Oracle VM VirtualBox
Fitxer MÀquina Visualitza Entrada Dispositius Ajuda
Aplicacions Llocs Sistema
hadoop@node3:~/Escriptori
Fitxer Edita Visualitza Cerca Terminal Ajuda
[hadoop@node3 Escriptori]$ jps
3816 DataNode
3963 Jps
[hadoop@node3 Escriptori]$
```

□ Hi han comandes de "yarn" molt útils Per exemple saber tots els nodes del clúster



Nom i Cognoms

Arnau Subirós Puigarnau

Data

02-06-2020

[hadoop@node1 ~]# yarn node -list

```
[hadoop@node1 ~]$ yarn node -list
20/05/07 19:00:15 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
20/05/07 19:00:15 INFO client.RMProxy: Connecting to ResourceManager at node1/192.168.0.101:8032
Total Nodes:2
  Node-Id          Node-State Node-Http-Address      Number-of-Running-Containers
  node2:41813      RUNNING     node2:8042                  0
  node3:37004      RUNNING     node3:8042                  0
[hadoop@node1 ~]$
```

[hadoop@node1 ~]

[hadoop@node1 ~]\$ yarn node -list -showDetails

20/05/07 19:03:02 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
20/05/07 19:03:02 INFO client.RMProxy: Connecting to ResourceManager at node1/192.168.0.101:8032
20/05/07 19:03:03 INFO conf.Configuration: resource-types.xml not found
20/05/07 19:03:03 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
20/05/07 19:03:03 INFO resource.ResourceUtils: Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE
20/05/07 19:03:03 INFO resource.ResourceUtils: Adding resource type - name = vcores, units = , type = COUNTABLE

Total Nodes:2

Node-Id	Node-State	Node-Http-Address	Number-of-Running-Containers
node2:41813	RUNNING	node2:8042	0

Detailed Node Information :

- Configured Resources : <memory:8192, vCores:8>
- Allocated Resources : <memory:0, vCores:0>
- Resource Utilization by Node : PMem:1673 MB, VMem:1673 MB, Vcores:0.0033300032
- Resource Utilization by Containers : PMem:0 MB, VMem:0 MB, Vcores:0.0
- Node-Labels :

node3:37004	RUNNING	node3:8042	0
-------------	---------	------------	---

Detailed Node Information :

- Configured Resources : <memory:8192, vCores:8>
- Allocated Resources : <memory:0, vCores:0>
- Resource Utilization by Node : PMem:1660 MB, VMem:1660 MB, Vcores:0.0066644456
- Resource Utilization by Containers : PMem:0 MB, VMem:0 MB, Vcores:0.0
- Node-Labels :

[hadoop@node1 ~]\$

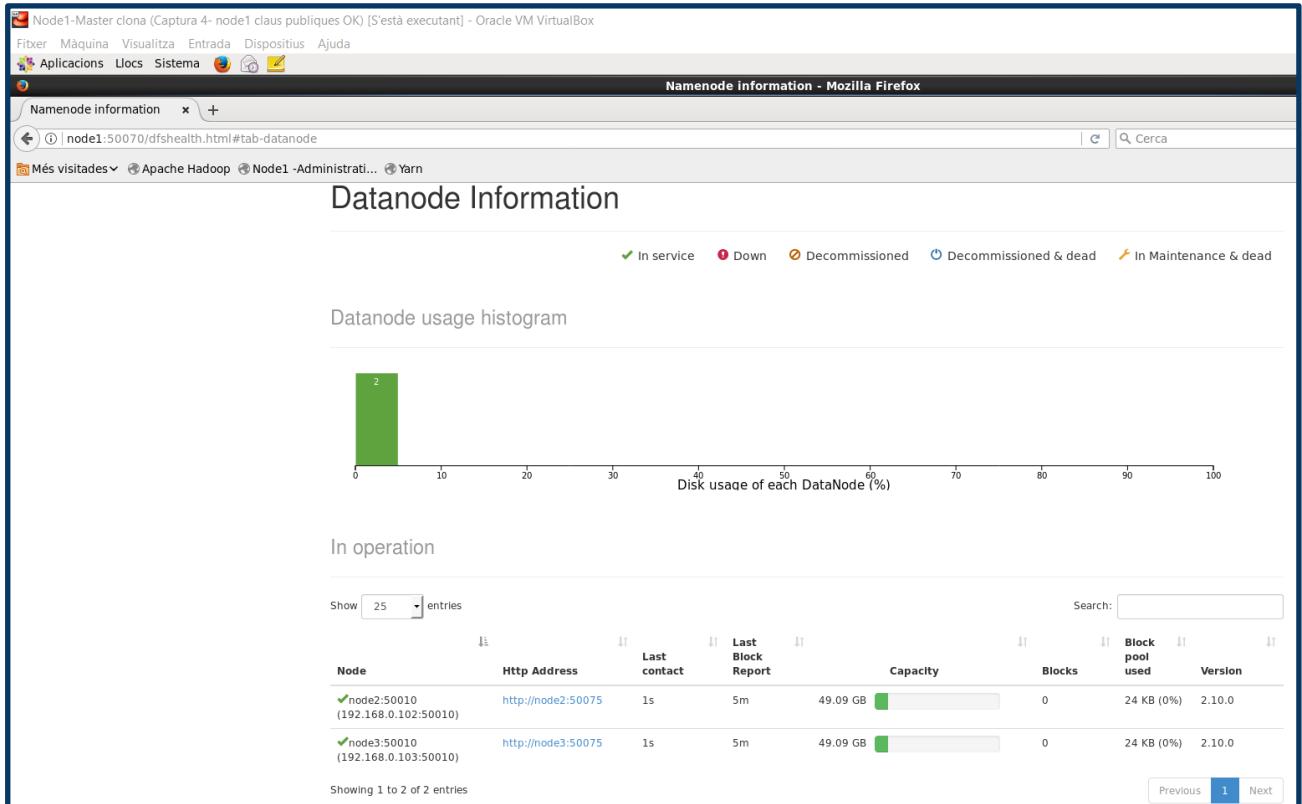
[hadoop@node1 ~]# yarn application -status application_1588845620923_0001

ID	User	Application
application_1588845620923_0001	hadoop	hadoop@node1 - [hadoop@node1 ~]# yarn application -status application_1588845620923_0001 20/05/07 19:07:25 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable 20/05/07 19:07:25 INFO client.RMProxy: Connecting to ResourceManager at node1/192.168.0.101:8032 20/05/07 19:07:25 INFO conf.Configuration: resource-types.xml not found 20/05/07 19:07:25 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'. 20/05/07 19:07:25 INFO resource.ResourceUtils: Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE 20/05/07 19:07:25 INFO resource.ResourceUtils: Adding resource type - name = vcores, units = , type = COUNTABLE Application-Id : application_1588845620923_0001 Application-Name : word count Application-Type : MAPREDUCE User : hadoop Queue : default Application Priority : 0 Start-Time : 1588845937796 Finish-Time : 1588846046114 Progress : 100% State : FINISHED Final-State : SUCCEEDED Tracking-URL : http://node2:19888/jobhistory/job/job_1588845620923_0001 RPC Port : 4222 AM Host : node2 Aggregate Resource Allocation : 531333 MB-seconds, 402 vcore-seconds Aggregate Resource Preempted : 0 MB-seconds, 0 vcore-seconds Log Aggregation Status : DISABLED Diagnostics : Unmanaged Application : false Application Node Label Expression : <Not set> AM container Node Label Expression : <DEFAULT PARTITION> TimeoutType : LIFETIME ExpiryTime : UNLIMITED RemainingTime : -1seconds
application_1588845620923_0002	hadoop	
application_1588845620923_0004	hadoop	
application_1588845620923_0005	hadoop	
application_1588845620923_0006	hadoop	
application_1588845620923_0007	hadoop	

[hadoop@node1 ~]\$

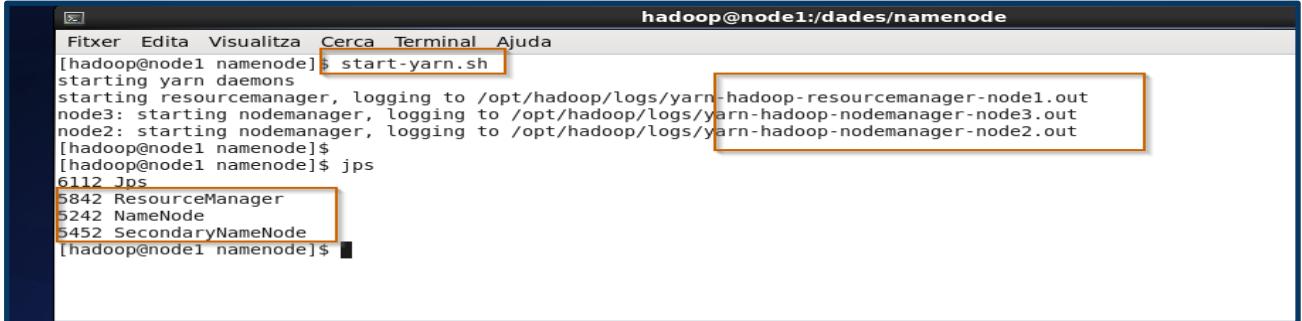
- Des de el node1 accedim a l'interfície web del Namenode (port 50070) i visualitzem la informació dels Datanodes.

Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020



☐ Arrencar el Yarn

```
[hadoop@node1 hadoop]# start-yarn.sh
```



```

[hadoop@node1 namenode]$ start-yarn.sh
starting yarn daemons
starting resourcemanager, logging to /opt/hadoop/logs/yarn-hadoop-resourcemanager-node1.out
node3: starting nodemanager, logging to /opt/hadoop/logs/yarn-hadoop-nodemanager-node3.out
node2: starting nodemanager, logging to /opt/hadoop/logs/yarn-hadoop-nodemanager-node2.out
[hadoop@node1 namenode]$ jps
5842 ResourceManager
5242 NameNode
5452 SecondaryNameNode
[hadoop@node1 namenode]$ 
```



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

- Desde el node1 accedim a l'interfície web del Yarn (port 8088) i visualitzem la informació dels Datanodes.

The screenshot shows the 'Nodes of the cluster' interface. On the left, there's a sidebar with 'Cluster Metrics' (0 Apps Submitted, 0 Apps Pending, 0 Apps Running, 0 Apps Completed, 0 Containers Running, 0 B Memory Used, 16 GB Memory Total, 0 B Memory Reserved, 0 VCores Used, 16 VCores Total, 0 VCores Available), 'Node Labels' (NEW, NEW_SAVING, SUBMITTED, ACCEPTED, RUNNING, FINISHED, FAILED, KILLED), and 'Scheduler' (Capacity Scheduler). The main area has tabs for 'Nodes of the cluster' and 'HDFS'. Under 'Nodes of the cluster', it shows 'Nodes Metrics' with 2 Active Nodes, 0 Decommissioning Nodes, 0 Decommissioned Nodes, 0 Lost Nodes, 0 Unhealthy Nodes, 0 Rebooted Nodes, and 0 Shutdown Nodes. It also shows 'Scheduler Metrics' for the Capacity Scheduler. A table lists nodes by label, rack, state, address, HTTP address, last health update, health report, containers, mem used, mem avail, vcores used, and vcores available. Two rows are shown: node2:46292 (node2:8042) and node3:38236 (node3:8042).

- Abans de fer algun exemple d'un procés MapReduce contra el clúster haurem d'activar un procés per poder veure l'històric dels processos de cada aplicació executada

```
[hadoop@node1 hadoop]# mr-jobhistory-daemon.sh start historyserver
```

The terminal window shows the command 'mr-jobhistory-daemon.sh start historyserver' being run. The output indicates that the historyserver is starting and logging to /opt/hadoop/logs/mapred-hadoop-historyserver-node1.out. The user then runs 'jps' to list running processes, which shows several Java processes: ResourceManager (5842), JobHistoryServer (6310), Jps (6409), NameNode (5242), and SecondaryNameNode (5452).



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

3.2.5. Exemples Mapreduce

3.2.5.1. Exemple 1 - Llençar un procés Mapreduce contra el clúster

- Des de el node “master” crearem un directori HDFS anomenat “prova01”

```
[hadoop@node1 hadoop]# hdfs dfs -mkdir /prova01
```

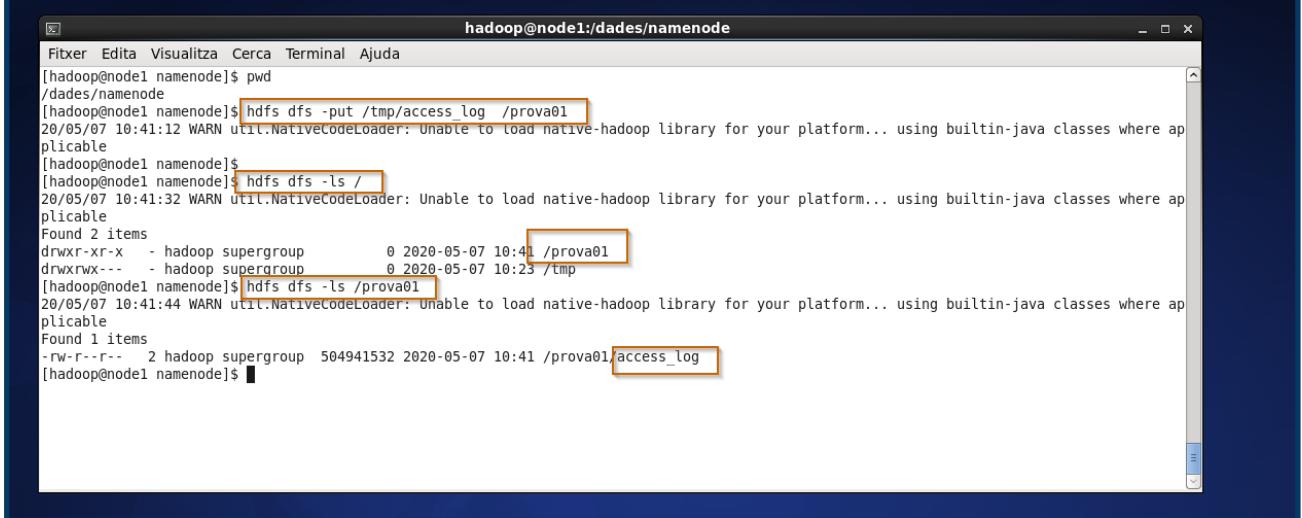
```
hadoop@node1:dades/namenode
Fitxer Edita Visualitza Cerca Terminal Ajuda
[hadoop@node1 namenode]$ hdfs dfs -ls /
20/05/07 10:29:34 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 1 items
drwxrwx--- - hadoop supergroup          0 2020-05-07 10:23 /tmp
[hadoop@node1 namenode]$ hdfs dfs -mkdir /prova01
20/05/07 10:30:28 WARN util.NativeCodeLoader: unable to load native-hadoop library for your platform... using builtin-java classes where applicable
[hadoop@node1 namenode]$ hdfs dfs -ls /
20/05/07 10:30:34 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 2 items
drwxr-xr-x - hadoop supergroup          0 2020-05-07 10:30 /prova01
drwxrwx--- - hadoop supergroup          0 2020-05-07 10:23 /tmp
[hadoop@node1 namenode]$
```

- Ens hem descarregat un fitxer d'exemple de logs de 504 MB i el dipositem a la carpeta /tmp. Posteriorment el pugem a HDFS

```
access_log (/tmp) - gedit
Fitxer Edita Visualitza Cerca Eines Documents Ajuda
Característiques: access_log
10.223.157.186 - - [15/Jul/2009:14:58:59 -0700] "GET / HTTP/1.1" 403 202
10.223.157.186 - - [15/Jul/2009:14:58:59 -0700] "GET /favicon.ico HTTP/1.1" 404 209
10.223.157.186 - - [15/Jul/2009:15:50:35 -0700] "GET / HTTP/1.1" 200 9157
10.223.157.186 - - [15/Jul/2009:15:50:35 -0700] "GET /assets/js/lightbox.js HTTP/1.1" 200 10469
10.223.157.186 - - [15/Jul/2009:15:50:35 -0700] "GET /assets/css/reset.css HTTP/1.1" 200 1014
10.223.157.186 - - [15/Jul/2009:15:50:35 -0700] "GET /assets/css/960_12.css HTTP/1.1" 200 6206
10.223.157.186 - - [15/Jul/2009:15:50:35 -0700] "GET /assets/css/the-associates.css HTTP/1.1" 200 15779
10.223.157.186 - - [15/Jul/2009:15:50:35 -0700] "GET /assets/js/the-associates.js HTTP/1.1" 200 4492
10.223.157.186 - - [15/Jul/2009:15:50:35 -0700] "GET /assets/js/lightbox.js HTTP/1.1" 200 25960
[hadoop@node1:tmp]
Fitxer Edita Visualitza Cerca Terminal Ajuda
[hadoop@node1 tmp]$ ls -ls | grep access_log
493108 -rw-rw-r-- 1 hadoop hadoop 504941532 5 mai 13:01 access_log
[hadoop@node1 tmp]$
```

Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

[hadoop@node1 hadoop]# hdfs dfs -put /tmp/acess_log /prova01

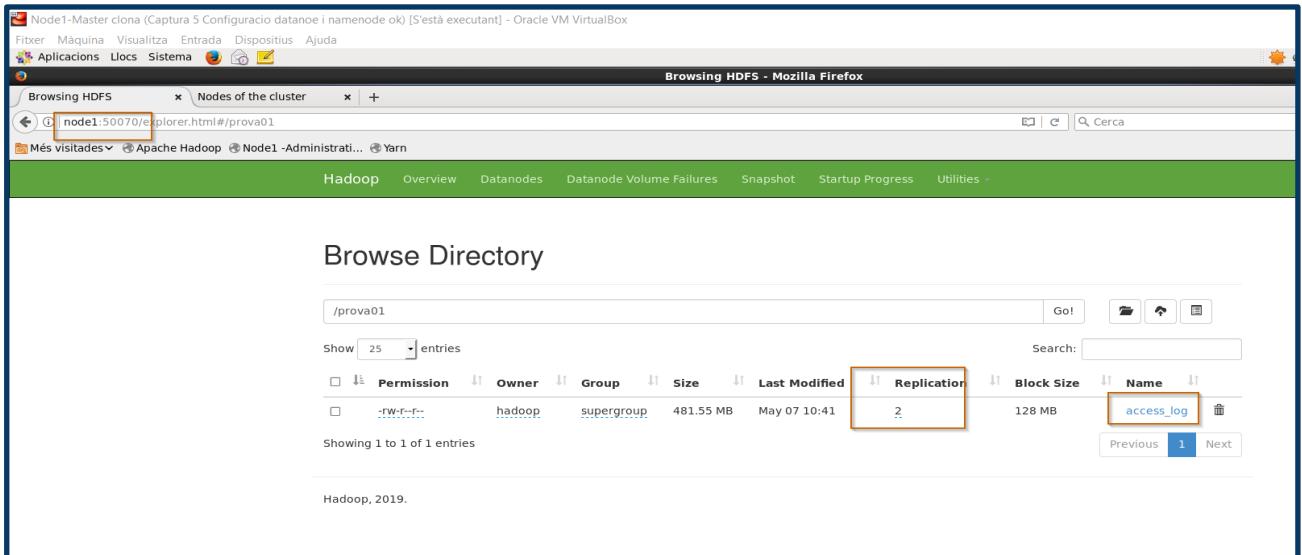


```

[hadoop@node1 namenode]$ pwd
/dades/namenode
[hadoop@node1 namenode]$ hdfs dfs -put /tmp/acess_log /prova01
20/05/07 10:41:12 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
[hadoop@node1 namenode]$
[hadoop@node1 namenode]$ hdfs dfs -ls /
20/05/07 10:41:32 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 2 items
drwxr-xr-x - hadoop supergroup 0 2020-05-07 10:41 /prova01
drwxrwx--- - hadoop supergroup 0 2020-05-07 10:23 /tmp
[hadoop@node1 namenode]$ hdfs dfs -ls /prova01
20/05/07 10:41:44 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 1 items
-rw-r--r-- 2 hadoop supergroup 504941532 2020-05-07 10:41 /prova01/access_log
[hadoop@node1 namenode]$

```

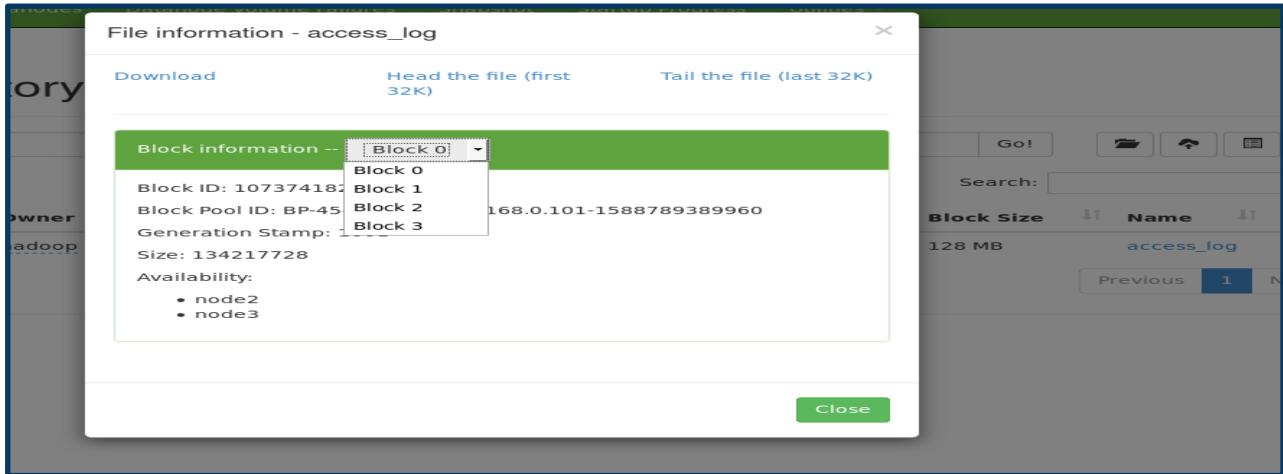
- Accedim a la interfície del Namenode i anem a la capeta a on hem pujat el fitxer. Podrem observar que el número de replicació és 2



Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hadoop	supergroup	481.55 MB	May 07 10:41	2	128 MB	access_log

Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

- Podrem observar que s'han creat 4 blocks, ja que la mida de l'arxiu és superior a 128 MB

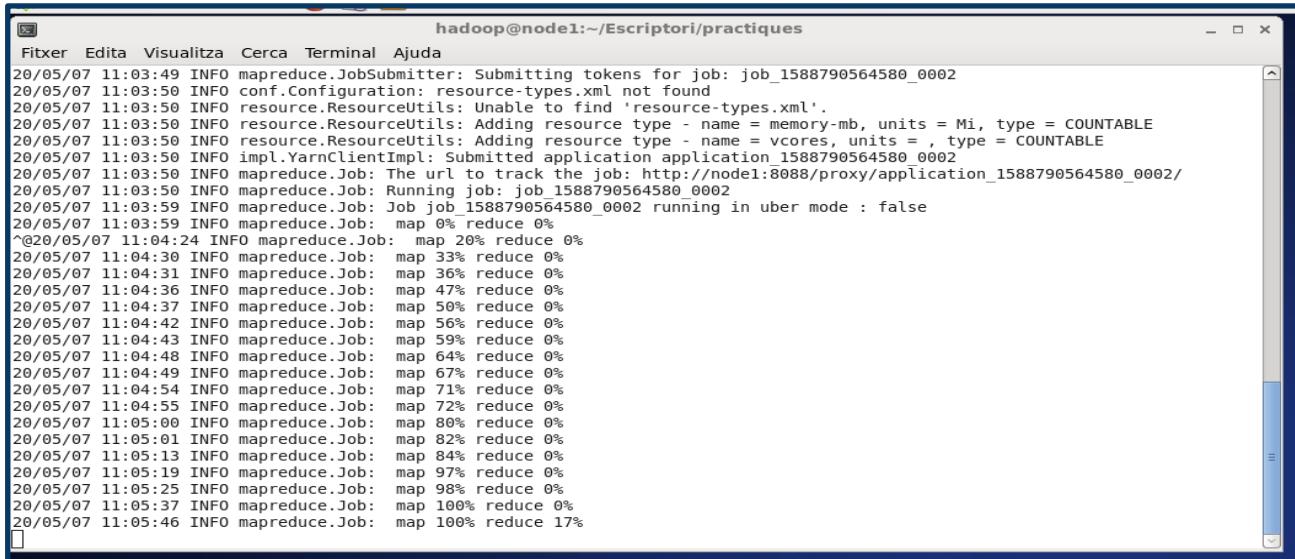


- Executem l'aplicació Java que vàrem utilitzar anteriorment “ComptarParaules” (en el clúster pseudodistribuït)

```
[hadoop@node1 hadoop]# hadoop jar ComptarParaules.jar ComptarParaules /prova01/access_log prova_access_log_sortida
```

```
[hadoop@node1 practiques]$ ls -ls
total 24
0 -rw-rw-r-- 1 hadoop hadoop 0 5 mai 19:36 ComptarParaules
0 -rw-rw-r-- 1 hadoop hadoop 1850 5 mai 19:36 ComptarParaules.class
0 -r--r--r-- 1 hadoop hadoop 1757 5 mai 19:36 ComptarParaules$IntSumReducer.class
0 -rw-rw-r-- 1 hadoop hadoop 3290 5 mai 19:39 ComptarParaules.jar
0 -rw-rw-r-- 1 hadoop hadoop 2761 5 mai 19:35 ComptarParaules.java
0 -rw-rw-r-- 1 hadoop hadoop 1754 5 mai 19:36 ComptarParaules$TokenizerMapper.class
0 drwxrwxr-x 2 hadoop hadoop 4096 5 mai 20:12 python
[hadoop@node1 practiques]$ hadoop jar ComptarParaules /prova01/access_log prova access log sortida
20/05/07 11:03:47 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
20/05/07 11:03:48 INFO client.RMProxy: Connecting to ResourceManager at node1/192.168.0.101:8032
20/05/07 11:03:49 INFO mapreduce.JobSubmitter: Total input files to process : 1
20/05/07 11:03:49 INFO mapreduce.JobSubmitter: number of splits:4
20/05/07 11:03:49 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
20/05/07 11:03:49 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1588790564580_0002
20/05/07 11:03:50 INFO conf.Configuration: resource-types.xml not found
20/05/07 11:03:50 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
20/05/07 11:03:50 INFO resource.ResourceUtils: Adding resource type - name = memory-MB, units = Mi, type = COUNTABLE
20/05/07 11:03:50 INFO resource.ResourceUtils: Adding resource type - name = vcores, units = , type = COUNTABLE
20/05/07 11:03:50 INFO impl.YarnClientImpl: Submitted application application_1588790564580_0002
20/05/07 11:03:50 INFO mapreduce.Job: The url to track the job: http://node1:8088/proxy/application_1588790564580_0002/
20/05/07 11:03:50 INFO mapreduce.Job: Running job: job_1588790564580_0002
20/05/07 11:03:59 INFO mapreduce.Job: Job job_1588790564580_0002 running in uber mode : false
20/05/07 11:03:59 INFO mapreduce.Job: map 0% reduce 0%
```

Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020



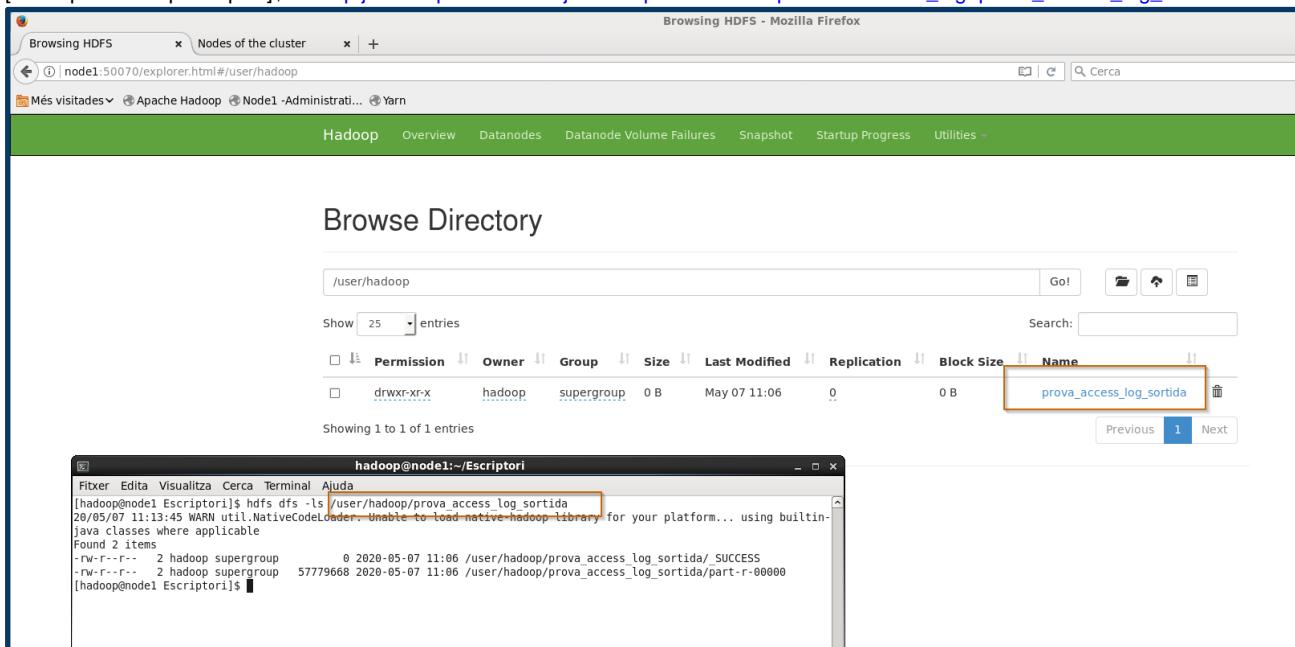
```

hadoop@node1:~/Escriptori/practiques
Fitxer Edita Visualitza Cerca Terminal Ajuda
20/05/07 11:03:49 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1588790564580_0002
20/05/07 11:03:50 INFO conf.Configuration: resource-types.xml not found
20/05/07 11:03:50 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
20/05/07 11:03:50 INFO resource.ResourceUtils: Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE
20/05/07 11:03:50 INFO resource.ResourceUtils: Adding resource type - name = vcores, units = , type = COUNTABLE
20/05/07 11:03:50 INFO impl.YarnClientImpl: Submitted application application_1588790564580_0002
20/05/07 11:03:50 INFO mapreduce.Job: The url to track the job: http://node1:8088/proxy/application_1588790564580_0002/
20/05/07 11:03:50 INFO mapreduce.Job: Running job: job_1588790564580_0002
20/05/07 11:03:59 INFO mapreduce.Job: Job job_1588790564580_0002 running in uber mode : false
20/05/07 11:03:59 INFO mapreduce.Job: map 0% reduce 0%
^@20/05/07 11:04:24 INFO mapreduce.Job: map 20% reduce 0%
20/05/07 11:04:30 INFO mapreduce.Job: map 33% reduce 0%
20/05/07 11:04:31 INFO mapreduce.Job: map 36% reduce 0%
20/05/07 11:04:36 INFO mapreduce.Job: map 47% reduce 0%
20/05/07 11:04:37 INFO mapreduce.Job: map 50% reduce 0%
20/05/07 11:04:42 INFO mapreduce.Job: map 56% reduce 0%
20/05/07 11:04:43 INFO mapreduce.Job: map 59% reduce 0%
20/05/07 11:04:48 INFO mapreduce.Job: map 64% reduce 0%
20/05/07 11:04:49 INFO mapreduce.Job: map 67% reduce 0%
20/05/07 11:04:54 INFO mapreduce.Job: map 71% reduce 0%
20/05/07 11:04:55 INFO mapreduce.Job: map 72% reduce 0%
20/05/07 11:05:00 INFO mapreduce.Job: map 80% reduce 0%
20/05/07 11:05:01 INFO mapreduce.Job: map 82% reduce 0%
20/05/07 11:05:13 INFO mapreduce.Job: map 84% reduce 0%
20/05/07 11:05:19 INFO mapreduce.Job: map 97% reduce 0%
20/05/07 11:05:25 INFO mapreduce.Job: map 98% reduce 0%
20/05/07 11:05:37 INFO mapreduce.Job: map 100% reduce 0%
20/05/07 11:05:46 INFO mapreduce.Job: map 100% reduce 17%

```

NOTA: com que en la comanda m'he oblidat l'arrel en el directori de sortida en la [/prova_access_log_sortida](#), es crearà a [/users/hadoop/prova_acces_log_sortida](#)

[hadoop@node1 practiques]\$hadoop jar ComptarParaules.jar ComptarParaules /prova01/access_log prova_access_log_sortida



Browsing HDFS - Mozilla Firefox

Browsing HDFS - Mozilla Firefox

Més visitades ▾ Apache Hadoop Node1 -Administraci... Yarn

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities ▾

Browse Directory

/user/hadoop

Show 25 entries

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	hadoop	supergroup	0 B	May 07 11:06	0	0 B	prova_access_log_sortida

Showing 1 to 1 of 1 entries

Go! |

Search:

Previous 1 Next

hadoop@node1:~/Escriptori

```

[hadoop@node1 Escriptori]$ hdfs dfs -ls /user/hadoop/prova access log sortida
20/05/07 11:13:45 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-jar classes where applicable
Found 2 items
-rw-r--r-- 2 hadoop supergroup 0 2020-05-07 11:06 /user/hadoop/prova access log sortida/_SUCCESS
-rw-r--r-- 2 hadoop supergroup 57779668 2020-05-07 11:06 /user/hadoop/prova access log sortida/part-r-00000
[hadoop@node1 Escriptori]$ 

```

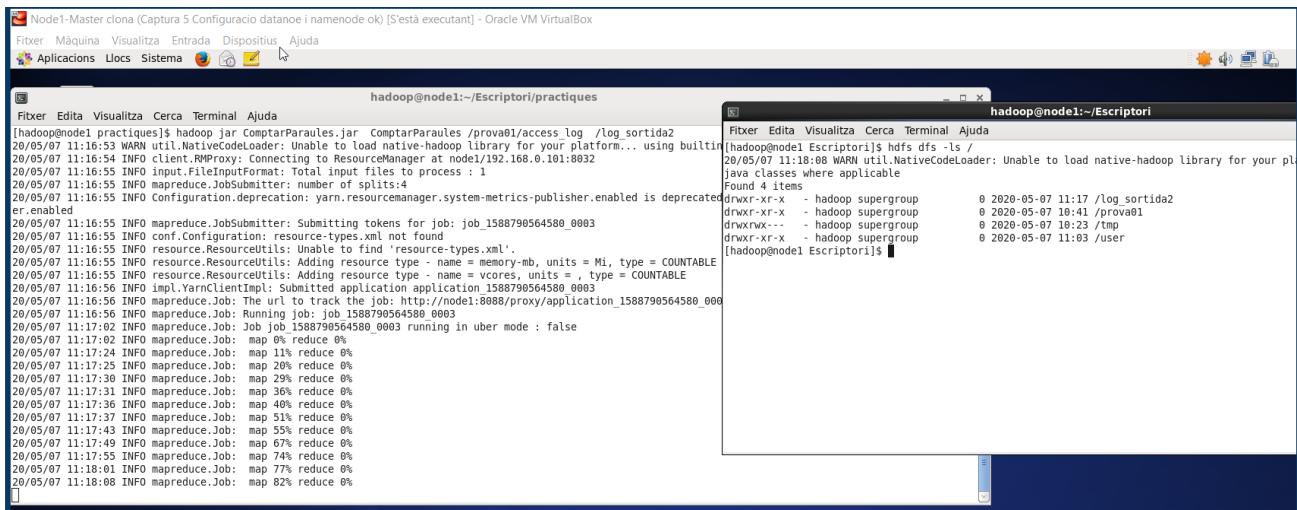
- Tornem a executar l'aplicació canviant el directori de sortida, li direm "["log_sortida2"](#)"

Nom i Cognoms

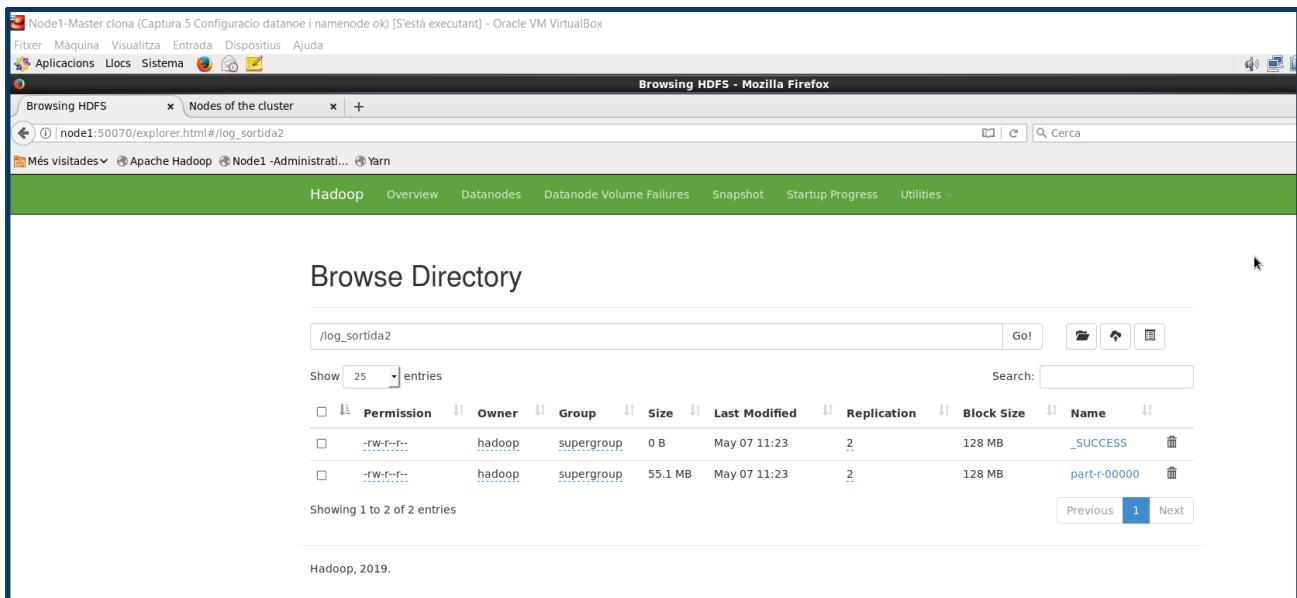
Arnaud Subirós Puigarnau

Data

02-06-2020



```
[hadoop@node1-practiques] hadoop jar ComptarpParaules.jar ComptarpParaules /prova01/access log /log_sortida2
20/05/07 11:16:53 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin
20/05/07 11:16:54 INFO client.RMProxy: Connecting to ResourceManager at node1/192.168.0.101:8032
20/05/07 11:16:55 INFO input.FileInputFormat: Total input files to process : 1
20/05/07 11:16:55 INFO mapreduce.JobSubmitter: number of splits:4
20/05/07 11:16:55 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated
20/05/07 11:16:55 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1588790564580_0003
20/05/07 11:16:55 INFO conf.Configuration: resource-types.xml not found
20/05/07 11:16:55 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
20/05/07 11:16:55 INFO resource.ResourceUtils: Adding resource type - name = memory_mb, units = Mi, type = COUNTABLE
20/05/07 11:16:55 INFO resource.ResourceUtils: Adding resource type - name = vcores, units = , type = COUNTABLE
20/05/07 11:16:56 INFO impl.YarnClientImpl: Submitted application application_1588790564580_0003
20/05/07 11:16:56 INFO mapreduce.Job: The url to track the job: http://node1:8088/proxy/application_1588790564580_0003
20/05/07 11:16:56 INFO mapreduce.Job: Running job: job_1588790564580_0003
20/05/07 11:17:02 INFO mapreduce.Job: Job job_1588790564580_0003 running in uber mode : false
20/05/07 11:17:02 INFO mapreduce.Job: map 0% reduce 0%
20/05/07 11:17:02 INFO mapreduce.Job: map 25% reduce 0%
20/05/07 11:17:25 INFO mapreduce.Job: map 20% reduce 0%
20/05/07 11:17:30 INFO mapreduce.Job: map 29% reduce 0%
20/05/07 11:17:31 INFO mapreduce.Job: map 36% reduce 0%
20/05/07 11:17:36 INFO mapreduce.Job: map 49% reduce 0%
20/05/07 11:17:37 INFO mapreduce.Job: map 51% reduce 0%
20/05/07 11:17:43 INFO mapreduce.Job: map 55% reduce 0%
20/05/07 11:17:49 INFO mapreduce.Job: map 67% reduce 0%
20/05/07 11:17:55 INFO mapreduce.Job: map 74% reduce 0%
20/05/07 11:18:01 INFO mapreduce.Job: map 77% reduce 0%
20/05/07 11:18:08 INFO mapreduce.Job: map 82% reduce 0%
[hadoop@node1-practiques]$ hdfs dfs -ls /
20/05/07 11:18:08 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin
java classes where applicable
Found 4 items
drwxr-xr-x - hadoop supergroup 0 2020-05-07 11:17 /log_sortida2
drwxr-xr-x - hadoop supergroup 0 2020-05-07 10:41 /prova01
drwxrwx--- - hadoop supergroup 0 2020-05-07 10:23 /tmp
drwxr-xr-x - hadoop supergroup 0 2020-05-07 11:03 /user
[hadoop@node1-practiques]$
```



Browsing HDFS - Mozilla Firefox

Browsing HDFS

Nodes of the cluster

node1:5007/explorer.html#/log_sortida2

Més visitades ▾ Apache Hadoop Node1-Administració Yarn

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities

Browse Directory

/log_sortida2

Show 25 entries

	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
□	-rw-r--r--	hadoop	supergroup	0 B	May 07 11:23	2	128 MB	_SUCCESS
□	-rw-r--r--	hadoop	supergroup	55.1 MB	May 07 11:23	2	128 MB	part-r-00000

Showing 1 to 2 of 2 entries

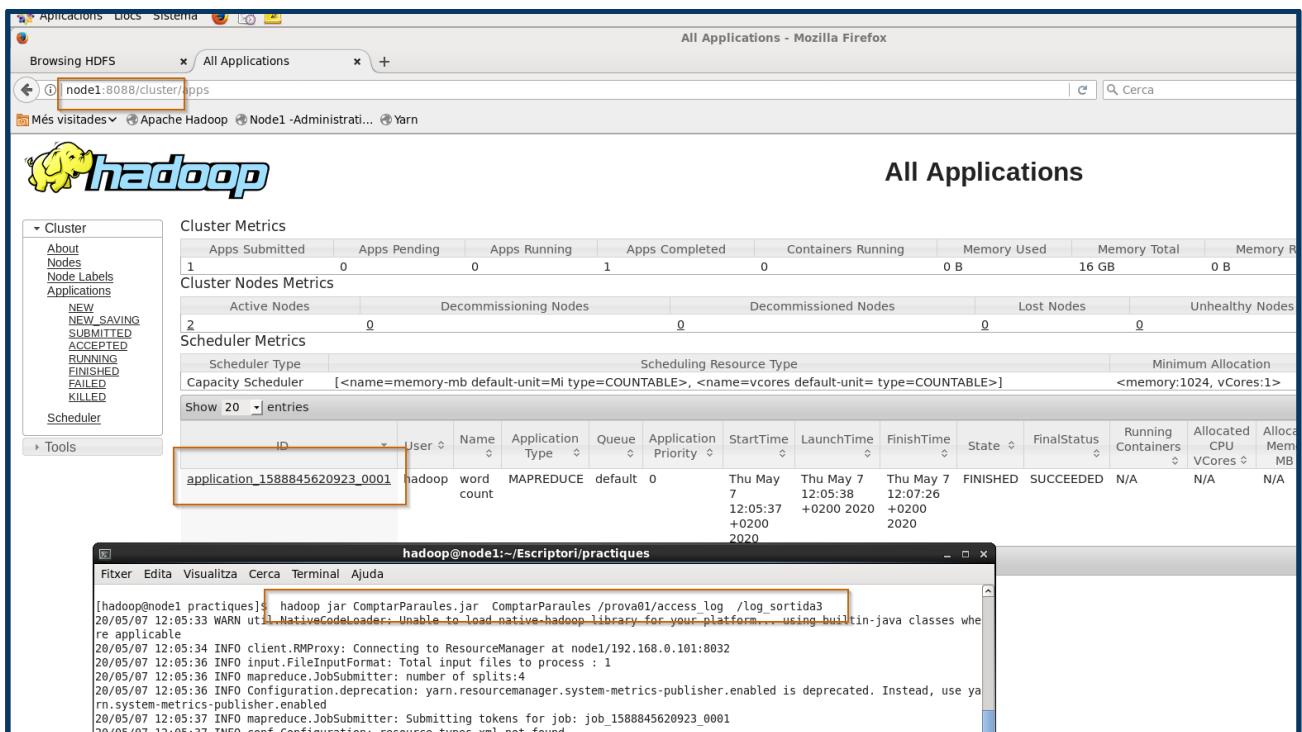
Previous 1 Next

Hadoop, 2019.

Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

- Tornem a executar l'aplicació per comptar paraules canviant el directori de destí, li direm `/log_sortida3`, ja que l'ordinador s'ha bloquejat i s'ha hagut de reiniciar parant-se les màquines virtuals (i els processos que estaven actius). Hem hagut d'iniciar novament els serveis.

```
[hadoop@node1 practiques]$ start-dfs.sh
[hadoop@node1 practiques]$ start-yarn.sh
[hadoop@node1 practiques]$ mr-jobhistory-daemon.sh start historyserver
[hadoop@node1 practiques]$ hadoop jar ComptarParaules.jar ComptarParaules /prova01/access_log /log_sortida3
```



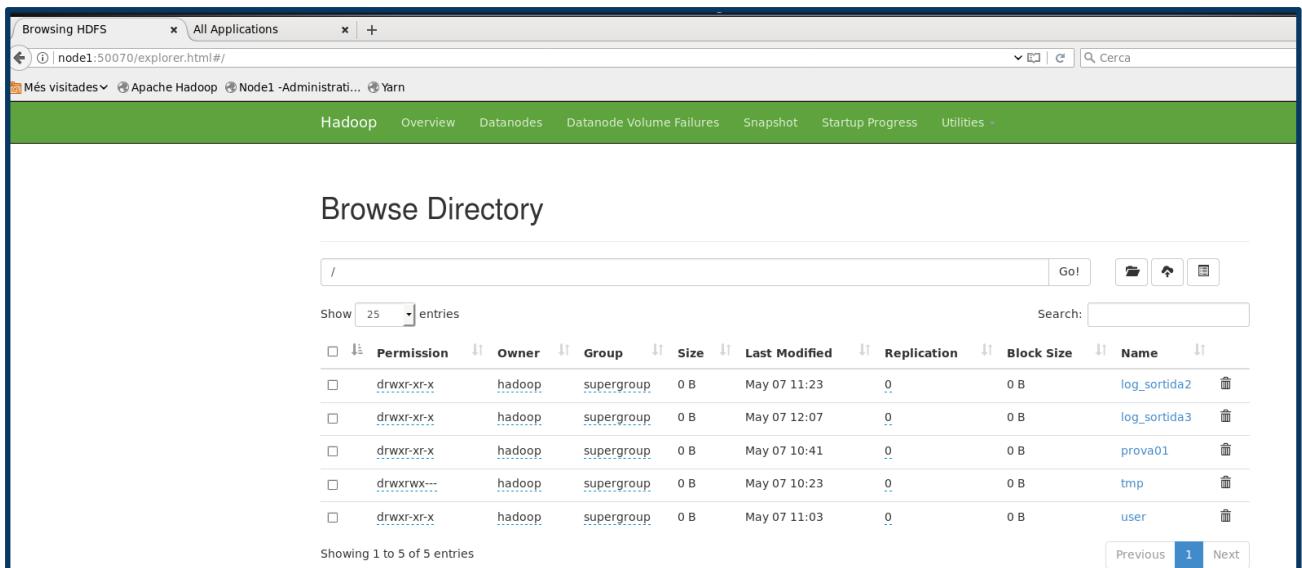
The screenshot shows two tabs open in a Firefox browser:

- Cluster Metrics:** Displays metrics for the cluster, including the number of nodes (2), active nodes (2), and various scheduler and resource allocation details.
- All Applications:** Shows a table of running applications. One row is highlighted for the job `application_1588845620923_0001`, which has the following details:

ID	User	Name	Application Type	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU	Allocated Vcores	Allocated Mem MB
application_1588845620923_0001	hadoop	word count	MAPREDUCE	default	0	Thu May 7 12:05:37 +0200 2020	+0200 2020	Thu May 7 12:07:26 2020	FINISHED	SUCCEEDED	N/A	N/A	N/A	

Below the browser window, a terminal window titled "hadoop@node1:~/Escriptori/practiques" shows the command and its output:

```
[hadoop@node1 practiques]$ hadoop jar ComptarParaules.jar ComptarParaules /prova01/access_log /log_sortida3
20/05/07 12:05:33 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using built-in java classes where applicable
20/05/07 12:05:34 INFO client.RMProxy: Connecting to ResourceManager at node1/192.168.0.101:8032
20/05/07 12:05:34 INFO input.FileInputFormat: Total input files to process : 1
20/05/07 12:05:34 INFO mapreduce.JobSubmitter: number of splits:4
20/05/07 12:05:34 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
20/05/07 12:05:37 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1588845620923_0001
20/05/07 12:05:37 INFO conf.Configuration: configuration resource types xml not found
```



The screenshot shows a Firefox browser window with the URL `node1:50070/explorer.html#` in the address bar. The page title is "Hadoop". The main content area is titled "Browse Directory" and displays a table of files in the root directory (/). The table includes columns for Name, Size, Last Modified, and Block Size. The files listed are:

Name	Size	Last Modified	Block Size
log_sortida2	0 B	May 07 11:23	0 B
log_sortida3	0 B	May 07 12:07	0 B
prova01	0 B	May 07 10:41	0 B
tmp	0 B	May 07 10:23	0 B
user	0 B	May 07 11:03	0 B

At the bottom of the table, it says "Showing 1 to 5 of 5 entries".



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

Browsing HDFS - Mozilla Firefox

All Applications node1:50070/explorer.html#/log_sortida3

Més visitades Apache Hadoop Node1 -Administrati... Yarn

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities

Browse Directory

/log_sortida3

Show 25 entries Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hadoop	supergroup	0 B	May 07 12:07	2	128 MB	_SUCCESS
-rw-r--r--	hadoop	supergroup	55.1 MB	May 07 12:07	2	128 MB	part-r-00000

Showing 1 to 2 of 2 entries Previous 1 Next

Hadoop, 2019.

Node1-Master clona (Captura 5 Configuracio datanode i namenode ok) [S'està executant] - Oracle VM VirtualBox

Fixer Màniga Visualitz Entrada Dispositius Ajuda

Aplicacions Llocs Sistema

Browsing HDFS / MapReduce Job job_15... +

node1:19888/johnhistory/job/job_1588845620923_0001

Més visitades Apache Hadoop Node1 -Administrati... Yarn

MapReduce Job job_1588845620923_0001

Job Name: word count
User Name: hadoop
Queue: default
State: SUCCEEDED
Uberized: false
Submitted: Thu May 07 12:05:37 CEST 2020
Started: Thu May 07 12:05:44 CEST 2020
Finished: Thu May 07 12:07:25 CEST 2020
Elapsed: 1mins, 41sec

Diagnostics:
Average Map Time 59sec
Average Shuffle Time 23sec
Average Merge Time 0sec
Average Reduce Time 20sec

ApplicationMaster

Attempt Number	Start Time	Node	Logs
1	Thu May 07 12:05:40 CEST 2020	node2:8042	logs

Task Type

Total	Complete
4	4
1	1

Attempt Type

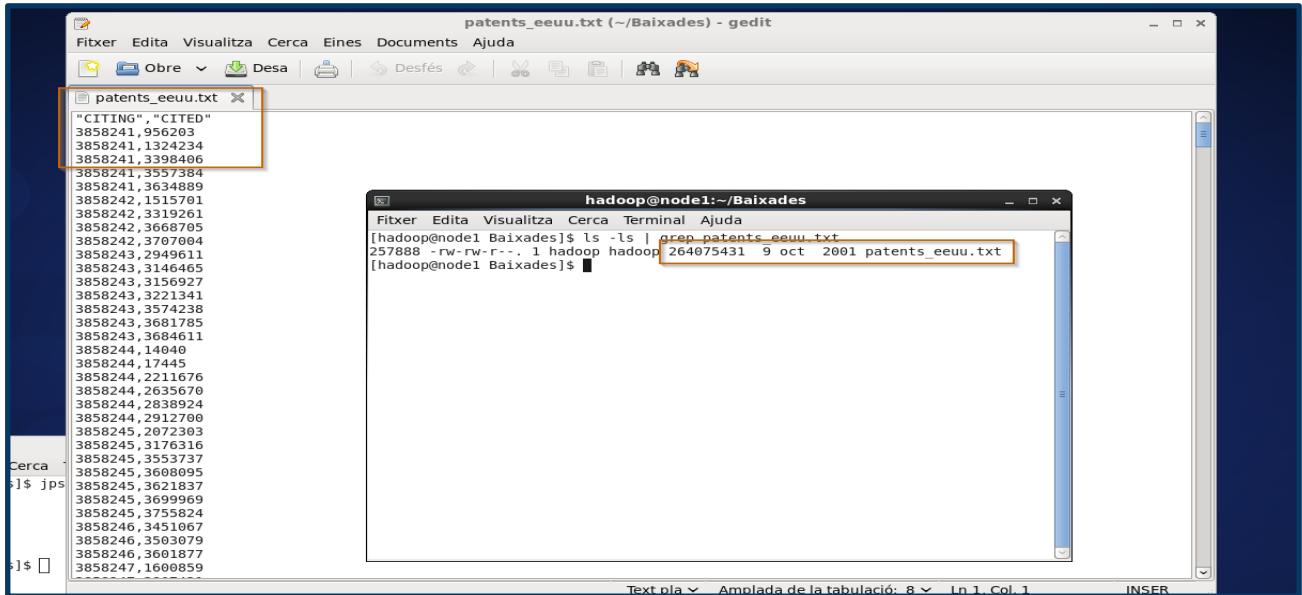
Maps	Reduces	Failed	Killed	Successful
0	0	0	0	4
0	0	0	0	1



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

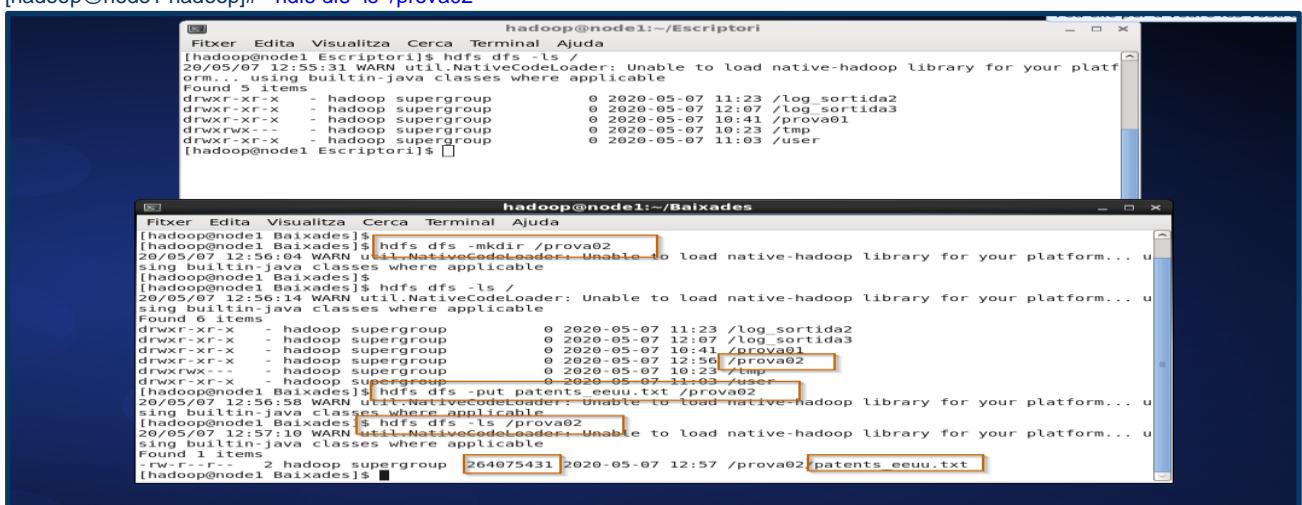
3.2.5.2. Exemple 2 - Llençar un procés Mapreduce contra el clúster

- Ens descarguem un arxiu d'exemple de patents dels E.E.U.U. anomenat "patents_eeuu.txt" de 264MB



- Desde el node "master" crearem un directori HDFS anomenat "prova02" i pujarem l'arxiu descarregat

```
[hadoop@node1 hadoop]# hdfs dfs -mkdir /prova02
[hadoop@node1 hadoop]# hdfs dfs -put patents_eeuu.txt /prova02
[hadoop@node1 hadoop]# hdfs dfs -ls /prova02
```



- Utilitzarem un programa Java anomenat "MyJob.java" per treballar amb aquest arxiu (ens agruparà les patents pel codi principal)



Nom i Cognoms

Arnaud Subirós Puigarnau

Data

02-06-2020

```
Fitxer Edita Visualitza Cerca Eines Documents Ajuda
MyJob.java X myjob.java (~/escritorio)

public class MyJob extends Configured implements Tool {
    public static class MapClass extends Mapper<LongWritable, Text, Text, Text> {
        public void map(LongWritable key, Text value, Context context) throws IOException, InterruptedException {
            String[] citation = value.toString().split(",");
            context.write(new Text(citation[1]), new Text(citation[0]));
        }
    }
    public static class Reduce extends Reducer<Text, Text, Text, Text> {
        public void reduce(Text key, Iterable<Text> values, Context context) throws IOException, InterruptedException {
            String csv = "";
            for (Text val : values) {
                if (csv.length() > 0) {
                    csv += ",";
                }
                csv += val.toString();
            }
            context.write(key, new Text(csv));
        }
    }
    public int run(String[] args) throws Exception {
        Configuration conf = getConf();
        Job job = new Job(conf, "MyJob");
        job.setJarByClass(MyJob.class);
        Path in = new Path(args[0]);
        Path out = new Path(args[1]);
        FileInputFormat.setInputPaths(job, in);
        FileOutputFormat.setOutputPath(job, out);
        job.setMapperClass(MapClass.class);
        job.setReducerClass(Reduce.class);
        job.setInputFormatClass(TextInputFormat.class);
        job.setOutputFormatClass(TextOutputFormat.class);
        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(Text.class);
        System.exit(job.waitForCompletion(true) ? 0 : 1);
        return 0;
    }
}
```

```
public static void main(String[] args) throws Exception {
    int res = ToolRunner.run(new Configuration(), new MyJob(), args);
    System.exit(res);
}
```



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

- Exportem la llibreria per localitzar-la en el CLASSPATH i com que no tenim IDE per compilar aplicacions Java, utilitzarem hadoop per compila-ho

```
[hadoop@node1 practica2]# export HADOOP_CLASSPATH=$JAVA_HOME/lib/tools.jar
[hadoop@node1 practica2]# hadoop com.sun.tools.javac.Main MyJob.java
[hadoop@node1 practica2]# jar cvf MyJob.jar My*
```

The terminal window shows the following command sequence:

```
hadoop@node1:~/Escriptori/practiques/practica2
Fitxer Edita Visualitza Cerca Terminal Ajuda
[hadoop@node1 practica2]$ echo exportem la llibreria per localitzar-la en el CLASSPATH
exportem la llibreria per localitzar-la en el CLASSPATH
[hadoop@node1 practica2]$
[hadoop@node1 practica2]$ export HADOOP_CLASSPATH=$JAVA_HOME/lib/tools.jar
[hadoop@node1 practica2]$
[hadoop@node1 practica2]$ echo compilem
compilem
[hadoop@node1 practica2]$
[hadoop@node1 practica2]$ hadoop com.sun.tools.javac.Main MyJob.java
Note: MyJob.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
[hadoop@node1 practica2]$
[hadoop@node1 practica2]$ hadoop com.sun.tools.javac.Main MyJob.java -Xlint
warning: [path] bad path element "/opt/hadoop/share/hadoop/common/lib/jaxb-api.jar": no such file or directory
warning: [path] bad path element "/opt/hadoop/share/hadoop/common/lib/activation.jar": no such file or directory
warning: [path] bad path element "/opt/hadoop/share/hadoop/common/lib/jsr173_1.0_api.jar": no such file or directory
warning: [path] bad path element "/opt/hadoop/share/hadoop/common/lib/jaxb1-impl.jar": no such file or directory
warning: [path] bad path element "/opt/hadoop/share/hadoop/yarn/lib/jaxb-api.jar": no such file or directory
warning: [path] bad path element "/opt/hadoop/share/hadoop/yarn/lib/activation.jar": no such file or directory
warning: [path] bad path element "/opt/hadoop/share/hadoop/yarn/lib/jsr173_1.0_api.jar": no such file or directory
warning: [path] bad path element "/opt/hadoop/share/hadoop/yarn/lib/jaxb1-impl.jar": no such file or directory
warning: [path] bad path element "/opt/hadoop/contrib/capacity-scheduler/*.jar": no such file or directory
MyJob.java:46: warning: [deprecation] Job(Configuration, String) in Job has been deprecated
    Job job = new Job(conf, "MyJob");
                           ^
10 warnings
[hadoop@node1 practica2]$ ls
MyJob.class MyJob.java MyJob$MapClass.class MyJob$Reduce.class
[hadoop@node1 practica2]$
```

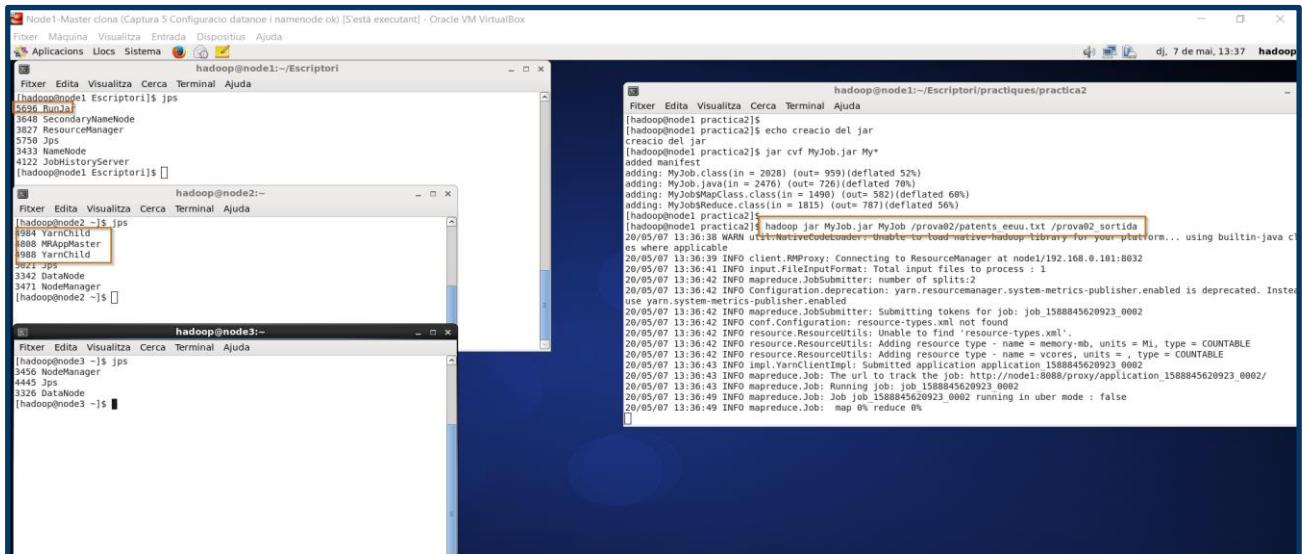
The terminal window shows the following command sequence:

```
hadoop@node1:~/Escriptori/practiques/practica2
Fitxer Edita Visualitza Cerca Terminal Ajuda
[hadoop@node1 practica2]$ ls -ls
total 16
4 -rw-rw-r-- 1 hadoop hadoop 2028 7 mai 13:23 MyJob.class
4 -rw-rw-r-- 1 hadoop hadoop 2476 7 mai 13:20 MyJob.java
4 -rw-rw-r-- 1 hadoop hadoop 1490 7 mai 13:23 MyJob$MapClass.class
4 -rw-rw-r-- 1 hadoop hadoop 1815 7 mai 13:23 MyJob$Reduce.class
[hadoop@node1 practica2]$
[hadoop@node1 practica2]$ echo creacio del jar
creacio del jar
[hadoop@node1 practica2]$ jar cvf MyJob.jar My*
added manifest
adding: MyJob.class(in = 2028) (out= 959)(deflated 52%)
adding: MyJob.java(in = 2476) (out= 726)(deflated 70%)
adding: MyJob$MapClass.class(in = 1490) (out= 582)(deflated 60%)
adding: MyJob$Reduce.class(in = 1815) (out= 787)(deflated 56%)
[hadoop@node1 practica2]$
```

Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

- Executem el **jar** que hem creat de MyJob on l'origen de les dades serà **/prova02/patents_euuu.txt** i crearà en cas que no existeixi el directori destí anomenat **prova02_sortida** on és disposarà el resultat. Mentre s'està creant execuem **jps** per veure els processos de tots els nodes.

```
[hadoop@node1 practica2]# hadoop jar MyJob.jar /prova02/patents_euuu.txt /prova02_sortida
[hadoop@node1 practica2]# jps
[hadoop@node2 ~]# jps
[hadoop@node3 ~]# jps
```



- Accedim a l'interfície web del Namenode i visualitzem el directori creat.

node1:50070

The screenshot shows the HDFS browser interface in Mozilla Firefox, displaying the contents of the `/prova02_sortida` directory:

Name
log_sortida2
log_sortida3
prova01
prova02
prova02_sortida
tmp
user

Details of the `prova02_sortida` directory entry:

- Permission: drwxr-xr-x
- Owner: hadoop
- Group: supergroup
- Size: 0 B
- Last Modified: May 07 11:03
- Replication: 0
- Block Size: 0 B
- Name: prova02_sortida



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

Browsing HDFS - Mozilla Firefox

node1:50070/explorer.html#/prova02_sortida

Més visitades Apache Hadoop Node1 -Administrat... Yarn

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities

Browse Directory

/prova02_sortida

Show 25 entries

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hadoop	supergroup	0 B	May 07 13:37	2	128 MB	_SUCCESS
-rw-r--r--	hadoop	supergroup	150.76 MB	May 07 13:37	2	128 MB	part-r-00000

Showing 1 to 2 of 2 entries

Previous 1 Next

Hadoop, 2019.

File information - part-r-00000

Download Head the file (first 32K) Tail the file (last 32K)

Block information --

Block ID: 107374186

Block Pool ID: BP-454387204-192.168.0.101-1588789389960

Generation Stamp: 1044

Size: 134217728

Availability:

- node3
- node2

Close

- Accedim a la interfície web del Yarn i visualitzem el directori creat.



Nom i Cognoms

Arnau Subirós Puigarnau

Data

02-06-2020

node1:8088

The screenshot shows the Apache Hadoop YARN Application page for application_1588845620923_0002. The page displays various configuration parameters and resource usage statistics.

Application Configuration:

- User: hadoop
- Name: Myjob
- Application Type: MAPREDUCE
- Application Tags: (empty)
- Application Priority: 0 (Higher integer value indicates higher priority)
- YarnApplicationState: FINISHED
- Queue: default
- FinalStatus Reported by AM: SUCCEEDED
- Started: dj. de maig 07 13:36:42 +0200 2020
- Launched: dj. de maig 07 13:36:44 +0200 2020
- Finished: dj. de maig 07 13:37:51 +0200 2020
- Elapsed: 1mins, 8sec
- Tracking URL: History
- Log Aggregation Status: DISABLED
- Application Timeout (Remaining Time): Unlimited
- Diagnostics: (empty)
- Unmanaged Application: false
- Application Node Label expression: <Not set>
- AM container Node Label expression: <DEFAULT_PARTITION>

Total Resource Preempted:

- Total Number of Non-AM Containers Preempted: 0
- Total Number of AM Containers Preempted: 0
- Resource Preempted from Current Attempt: <memory:0, vCores:0>
- Number of Non-AM Containers Preempted from Current Attempt: 0
- Aggregate Resource Allocation: 253918 MB-seconds, 171 vcore-seconds
- Aggregate Preempted Resource Allocation: 0 MB-seconds, 0 vcore-seconds

Logs:

Show 20 entries

Search:

The screenshot shows the Apache Hadoop MapReduce Job page for job_1588845620923_0002. The page displays job configuration and task execution details.

Job Configuration:

- Job Name: Myjob
- User Name: hadoop
- Queue: default
- State: SUCCEEDED
- Uberized: false
- Submitted: Thu May 07 13:36:42 CEST 2020
- Started: Thu May 07 13:36:48 CEST 2020
- Finished: Thu May 07 13:37:51 CEST 2020
- Elapsed: 1mins, 2sec
- Diagnostics:
- Average Map Time: 37sec
- Average Shuffle Time: 2sec
- Average Merge Time: 0sec
- Average Reduce Time: 18sec

ApplicationMaster:

Attempt Number	Start Time	Node	Logs
1	Thu May 07 13:36:45 CEST 2020	node2:8042	logs

Task Summary:

Task Type	Total	Complete
Map	2	2
Reduce	1	1

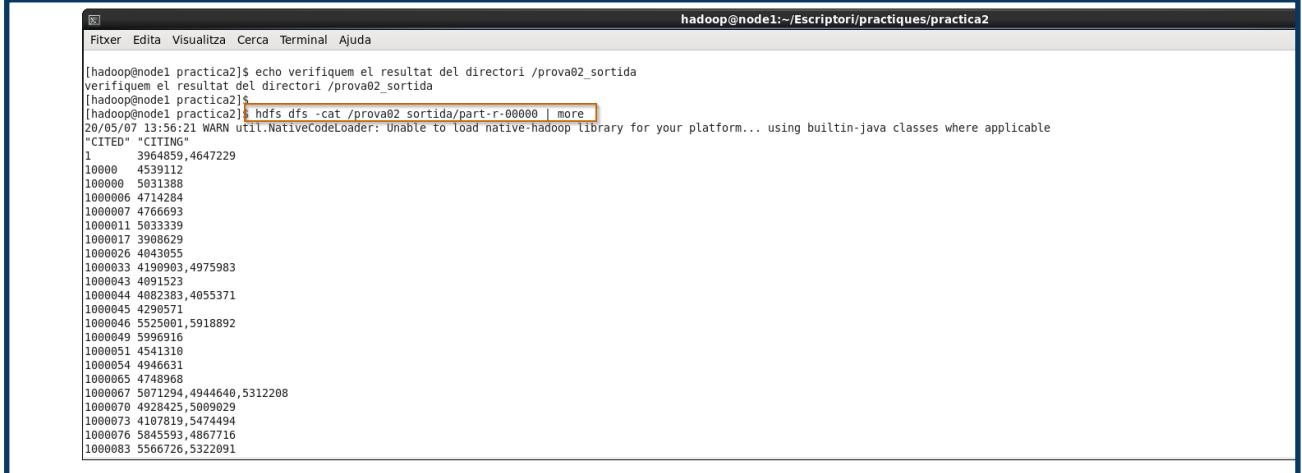
Attempt Type:

Maps	Failed	Killed	Successful
0	0	2	1
Reduces	0	0	1

Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

- No ens fa falta descarregar el resultat per poder-ho llegir, ja que és un text pla

[hadoop@node1 practica2]# hdfs dfs -cat /prova02_sortida/part-r-00000 | more



```

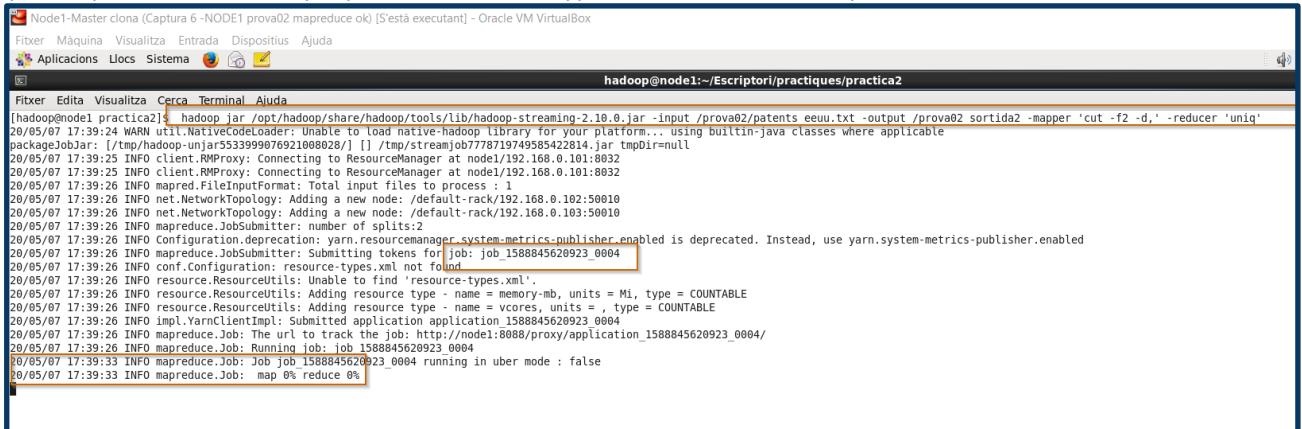
[hadoop@node1 practica2]$ echo verifiquem el resultat del directori /prova02_sortida
verifiquem el resultat del directori /prova02_sortida
[hadoop@node1 practica2]$
[hadoop@node1 practica2]$ hdfs dfs -cat /prova02_sortida/part-r-00000 | more
20/05/07 13:56:21 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
"CITED" "CITING"
1 3964859,4647229
10000 4539112
100000 5031388
1000006 4714284
1000007 4766693
1000011 5033339
1000017 3908629
1000026 4043055
1000033 4199903,4975983
1000043 4091523
1000044 4082383,4055371
1000045 4290571
1000046 5525601,5918892
1000049 5996916
1000051 4541310
1000054 4946631
1000065 4748969
1000067 5071294,4944640,5312208
1000070 4928425,5090929
1000073 4107819,5474494
1000076 5845593,4867716
1000083 5566726,5322091

```

3.2.5.3. Exemple 3 - Llençar un procés en streaming amb comandos de Shell de Linux.

- En aquest exemple utilitzarem comandos de Shell de Linux per fer de Mapper i de Reducer, simulant el programa que hem fet en l'exemple 2.
- Utilitzarem la llibreria de **streaming**
- Li passarem el fitxer i en el Mapper extreuem les dades del camp2 amb el comando “cut” i eliminarem duplicats amb el “Reducer” i el comando “uniq”
- Mentre s'està creant execuem **jps** per veure els processos de tots els nodes.

[hadoop@node1 practica2]# hadoop jar /opt/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.10.0.jar -input /prova02/patents_eeuu.txt -output /prova02_sortida2 -mapper 'cut -f2 -d,' -reducer 'uniq'



```

[hadoop@node1 Master clona (Captura 6 -NODE1 prova02 mapreduce ok) [S'està executant] - Oracle VM VirtualBox
Fitxer MÀquina Visualitzar Entrada Dispositius Ajuda
Aplicacions Llocs Sistema
hadoop@node1:~/Escriptori/practiques/practica2
[hadoop@node1 practica2]$ hadoop jar /opt/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.10.0.jar -input /prova02/patents_eeuu.txt -output /prova02_sortida2 -mapper 'cut -f2 -d,' -reducer 'uniq'
20/05/07 17:39:24 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
packageJobJar: [/tmp/hadoop-unjar5533999076921008028/] [] /tmp/streamjob7778719749585422814.jar tmpDir=null
20/05/07 17:39:25 INFO Client.RMProxy: Connecting to ResourceManager at node1/192.168.0.101:8032
20/05/07 17:39:26 INFO mapred.FileInputFormat: Total input files to process : 1
20/05/07 17:39:26 INFO net.NetworkTopology: Adding a new node: /default-rack/192.168.0.102:50010
20/05/07 17:39:26 INFO net.NetworkTopology: Adding a new node: /default-rack/192.168.0.103:50010
20/05/07 17:39:26 INFO mapreduce.JobSubmitter: number of splits:2
20/05/07 17:39:26 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
20/05/07 17:39:26 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1588845620923_0004
20/05/07 17:39:26 INFO conf.Configuration: resource-types.xml not found
20/05/07 17:39:26 INFO resource.ResourcesUtils: Unable to find 'resource-types.xml'.
20/05/07 17:39:26 INFO resource.ResourcesUtils: Adding resource type - name = memory_mb, units = Mi, type = COUNTABLE
20/05/07 17:39:26 INFO resource.ResourcesUtils: Adding resource type - name = vcores, units = , type = COUNTABLE
20/05/07 17:39:26 INFO mapreduce.YarnClientImpl: Submitted application_1588845620923_0004
20/05/07 17:39:26 INFO mapreduce.Job: The url to track the job: http://node1:8088/proxy/application_1588845620923_0004/
20/05/07 17:39:33 INFO mapreduce.Job: Job job_1588845620923_0004 running in uber mode : false
20/05/07 17:39:33 INFO mapreduce.Job: map 0% reduce 0%

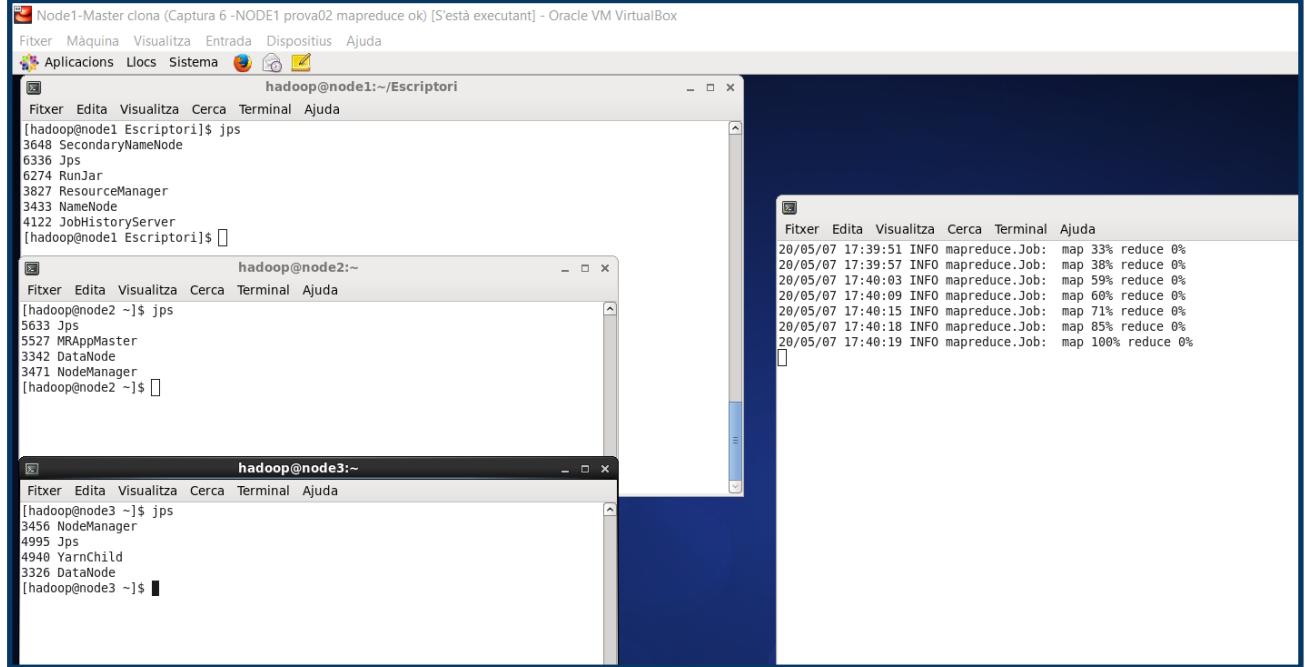
```

Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

```
[hadoop@node1 practica2]# jps
```

```
[hadoop@node2 ~]# jps
```

```
[hadoop@node3 ~]# jps
```

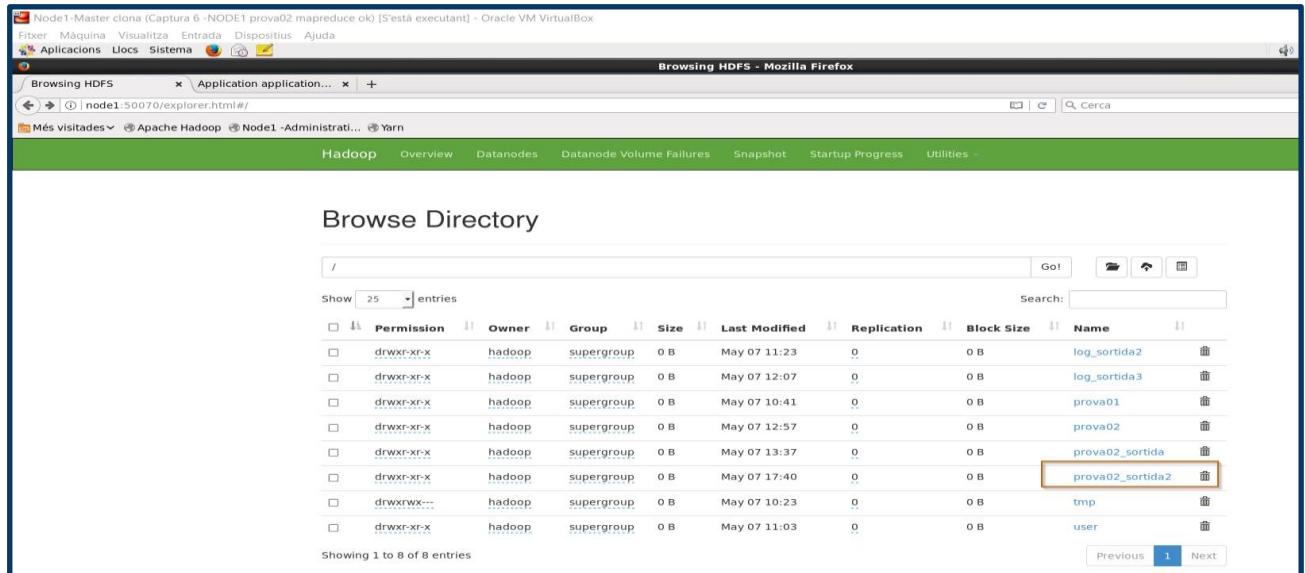


The screenshot shows three terminal windows side-by-side:

- Terminal 1 (node1):** Shows processes like SecondaryNameNode (3648), Jps (6336), ResourceManager (6274), NameNode (3827), and JobHistoryServer (4122).
- Terminal 2 (node2):** Shows processes like RunJar (5633), Jps (5527), MRAppMaster (3342), DataNode (3471), and NodeManager (4122).
- Terminal 3 (node3):** Shows processes like NodeManager (3456), Jps (4995), YarnChild (4940), DataNode (3326), and HDFS (4940).

- Accedim a la interfície web del Namenode i visualitzem el directori creat.

node1:50070



The screenshot shows the HDFS web interface with the following details:

- Title Bar:** Browsing HDFS - Mozilla Firefox
- Address Bar:** node1:50070/explorer.html#/
- Navigation Bar:** Back, Forward, Stop, Reload, Home, Search bar (Cerca), and a link to Apache Hadoop.
- Header:** Hadoop, Overview, Datanodes, Datanode Volume Failures, Snapshot, Startup Progress, Utilities.
- Section:** Browse Directory
- Table:** A list of directory entries under the root path (/). The table includes columns: Permission, Owner, Group, Size, Last Modified, Replication, Block Size, and Name.
- Entries:**

Name	Size	Last Modified	Replication
log_sortida2	0 B	May 07 11:23	0
log_sortida3	0 B	May 07 12:07	0
prova01	0 B	May 07 10:41	0
prova02	0 B	May 07 12:57	0
prova02_sortida	0 B	May 07 13:37	0
prova02_sortida2	0 B	May 07 17:40	0
tmp	0 B	May 07 10:23	0
user	0 B	May 07 11:03	0
- Page Footer:** Showing 1 to 8 of 8 entries, Previous, Next.



Nom i Cognoms

Arnau Subirós Puigarnau

Data

02-06-2020

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hadoop	supergroup	0 B	May 07 17:40	2	128 MB	_SUCCESS
-rw-r--r--	hadoop	supergroup	27.81 MB	May 07 17:40	2	128 MB	part-00000

Showing 1 to 2 of 2 entries

Hadoop, 2019.

- Accedim a la interfície web del Yarn i visualitzem el directori creat.

node1:8088

ID	User	Name	Application type	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU	Allocated Memory	Reserved CPU	Vcores
application_1588845620923_0004	hadoop	streamjob7778719749585422814.jar	MAPREDUCE	default	0	Thu May 7 17:39:26 2020	+0200	Thu May 7 17:40:52 2020	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A
application_1588845620923_0002	hadoop	Myjob	MAPREDUCE	default	0	Thu May 7 13:36:42 2020	+0200	Thu May 7 13:37:51 2020	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A
application_1588845620923_0001	hadoop	word count	MAPREDUCE	default	0	Thu May 7 12:05:37 2020	+0200	Thu May 7 12:07:26 2020	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A

Showing 1 to 3 of 3 entries



Nom i Cognoms

Arnaud Subirós Puigarnau

Data

02-06-2020

The screenshot shows the Apache Hadoop YARN Application page for application_1588845620923_0004. The main content area displays the following details:

- User: hadoop
- Name: streamjob7778719749585422814.jar
- Application Type: MAPREDUCE
- Application Tags: (empty)
- Application Priority: 0 (Higher Integer value indicates higher priority)
- YarnApplicationState: FINISHED
- Queue: default
- FinalStatus Reported by AM: SUCCEEDED
- Started: dj, de maig 07 17:39:26 +0200 2020
- Launched: dj, de maig 07 17:39:27 +0200 2020
- Finished: dj, de maig 07 17:40:52 +0200 2020
- Elapsed: 1mins, 26sec
- Tracking URL: History
- Log Aggregation Status: DISABLED
- Application Timeout (Remaining Time): Unlimited
- Diagnostics:
 - Unmanaged Application: false
 - Application Node Label expression: <Not set>
 - AM container Node Label expression: <DEFAULT_PARTITION>

Below this, resource statistics are listed:

- Total Resource Preempted: <memory:0, vCores:0>
- Total Number of Non-AM Containers Preempted: 0
- Total Number of AM Containers Preempted: 0
- Resource Preempted from Current Attempt: <memory:0, vCores:0>
- Number of Non-AM Containers Preempted from Current Attempt: 0
- Aggregate Resource Allocation: 314804 MB-seconds, 213 vcore-seconds
- Aggregate Preempted Resource Allocation: 0 MB-seconds, 0 vcore-seconds

The bottom section shows a table of application attempts:

Show	20	entries			
Attempt ID	Started	Node	Logs	Nodes blacklisted by the app	Nodes blacklisted by the system
appattempt_1588845620923_0004_00001	Thu May 7 17:39:26 +0200	http://node2:8042	Logs 0	0	0

The screenshot shows the Apache Hadoop MapReduce Job page for job_1588845620923_0004. The main content area displays the following details:

- Job Name: streamjob7778719749585422814.jar
- User Name: hadoop
- Queue: default
- State: SUCCEEDED
- Uberized: false
- Submitted: Thu May 07 17:39:26 CEST 2020
- Started: Thu May 07 17:39:31 CEST 2020
- Finished: Thu May 07 17:40:52 CEST 2020
- Elapsed: 1mins, 21sec

Below this, diagnostics are listed:

- Average Map Time: 44sec
- Average Shuffle Time: 2sec
- Average Merge Time: 0sec
- Average Reduce Time: 29sec

The bottom section shows a table of application master attempts:

ApplicationMaster	Attempt Number	Start Time	Node
1	Thu May 07 17:39:28 CEST 2020	node2:8042	

Below the table, task type statistics are shown:

Task Type	Total	Complete
Map	2	2
Reduce	1	1

At the bottom, attempt type statistics are shown:

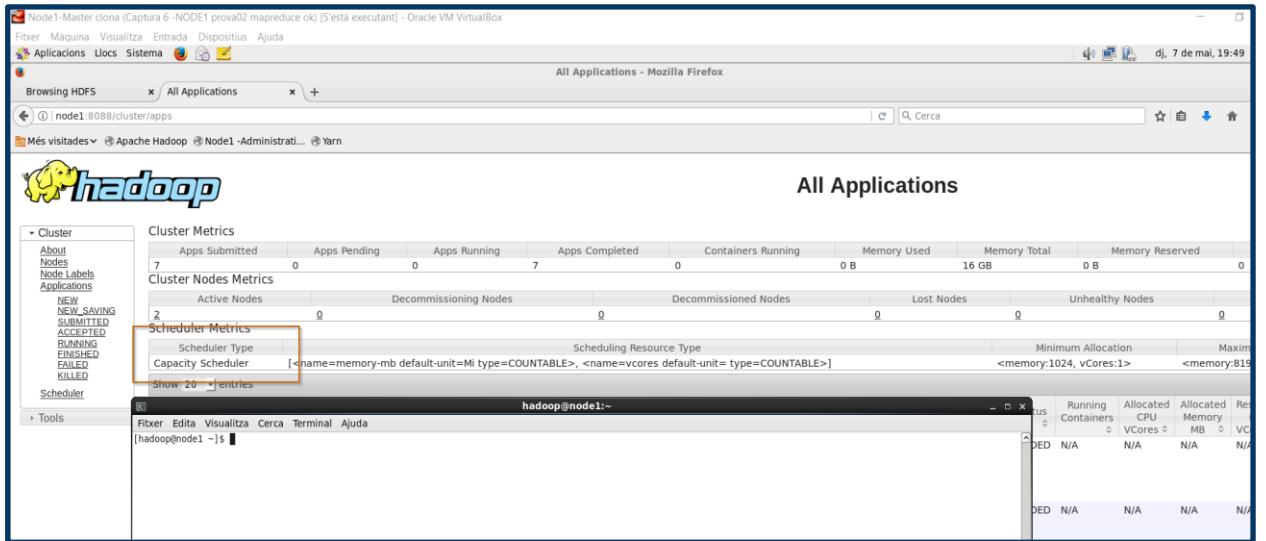
Attempt Type	Failed	Killed	Successful
Maps	0	0	2
Reduces	0	0	1

Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

3.2.6. Yarn Schelduler

- El planificador determina els recursos que tenen cada un dels components del clúster. Per defecte si no està configurat, tots els components tenen els mateixos recursos. Hi ha 2 tipus de planificadors:
 - **Fair Schelduler** : Tots els processos s'assignen de manera homogènia els recursos dins el clúster
 - utilitzat en les versions de Hadoop 1.x
 - **Capacity Schelduler** : En canvi aquest planificador utilitza una estructura de jerarquia. Si no especificuem res (no el tenim configurat), després de root, tindrem una cua de recursos anomenada “**default**” que està associada al 100% dels recursos, seria com si estiguessin fem servir **Fair Scheduler**
 - utilitzat en les versions de Hadoop 2.x i 3.x
- Com que estem utilitzant una versió de Hadoop 2.x estem utilitzant Capacity Schelduler però com que no està configurat és com si estiguéssim utilitzant Fair Schelduler. Accedim a l'interfície web del Yarn per veure la configuració “**default**” del Yarn Schelduler

node1:8088



The screenshot shows the Hadoop YARN web interface at node1:8088. The main page title is "All Applications". On the left, there's a sidebar with "Cluster" and "Scheduler" sections. The "Scheduler" section is expanded, showing "Scheduler Metrics" with a red box around it. Inside the box, "Scheduler Type" is set to "Capacity Scheduler". Below it, "Scheduling Resource Type" is listed as "<name=memory-mb default-unit=Mi type=COUNTABLE>, <name=vcores default-unit= type=COUNTABLE>". At the bottom of the sidebar, there's a "Tools" section.



Nom i Cognoms

Arnau Subirós Puigarnau

Data

02-06-2020

The screenshot shows the Apache Hadoop YARN web interface. On the left, there's a sidebar with navigation links like 'About', 'Nodes', 'Node Labels', 'Applications' (with sub-options: NEW, NEW_SAVING, SUBMITTED, ACCEPTED, RUNNING, FINISHED, FAILED, KILLED), and 'Scheduler'. The 'Scheduler' link is highlighted with a red box. Below the sidebar is a 'Tools' section. The main content area has a title 'All Applications - Mozilla Firefox'. It displays 'Cluster Metrics' with counts for Apps Submitted (7), Apps Pending (0), Apps Running (0), Apps Completed (7), Containers Running (0), and Memory (0 B). It also shows 'Cluster Nodes Metrics' with Active Nodes (2), Decommissioning Nodes (0), and Decommissioned Nodes (0). Under 'Scheduler Metrics', it shows the Scheduler Type as 'Capacity Scheduler' and the Scheduling Resource Type as '<name=memory-mb default-unit=Mi type=COUNTABLE>, <name=vcores default-unit= type=COUNTABLE>'. A table lists applications: application_1588845620923_0001 (hadoop, word count, MAPREDUCE, default, 0, Thu May 7 12:05:38 +0200 2020).

Podem observar que la cua del planificador per defecte es “default” configurat al 100%

This screenshot shows the 'Scheduler' section of the YARN web interface. The 'Scheduler' link in the sidebar is highlighted with a red box. The main area shows 'Dump scheduler logs' for the 'default' queue. A legend indicates 'Capacity' (grey), 'Used' (green), 'Used (over capacity)' (orange), 'Max Capacity' (yellow), and 'Users Requesting Resources' (yellow). The 'Used' tab is selected. A tree view shows 'Queue: root' expanded to show 'Queue: default'. A tooltip for 'Queue: default' provides detailed configuration information for the 'default' queue, including:

- Queue State: RUNNING
- Used Capacity: 0.0%
- Configured Capacity: 100.0%
- Configured Max Capacity: 100.0%
- Absolute Used Capacity: 0.0%
- Absolute Configured Capacity: 100.0%
- Absolute Configured Max Capacity: 100.0%
- Used Resources: <memory:0, vCores:0>
- Configured Max Application Master Limit: 10.0
- Max Application Master Resources: <memory:2048, vCores:1>
- Used Application Master Resources: <memory:0, vCores:0>
- Max Application Master Resources Per User: <memory:2048, vCores:1>
- Num Schedulable Applications: 0
- Num Non-Schedulable Applications: 0
- Num Containers: 0
- Max Applications: 10000
- Max Applications Per User: 10000
- Configured Minimum User Limit Percent: 100%
- Configured User Limit Factor: 1.0
- Accessible Node Labels: *
- Ordering Policy: FibroOrderingPolicy
- Preemption: disabled
- Intra-queue Preemption: disabled
- Default Node Label Expression: <DEFAULT_PARTITION>
- Default Application Priority: 0

- La informació de la cua també la podem visualitzar per terminal



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

[hadoop@node1 ~]# mapred queue -list

```

[hadoop@node1 ~]# mapred queue -list
[hadoop@node1 ~]$ mapred queue -list
20/05/07 20:03:59 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
20/05/07 20:03:59 INFO client.RMProxy: Connecting to ResourceManager at node1/192.168.0.101:8032

Queue Name : default
Queue State : running
Scheduling Info : Capacity: 100.0, MaximumCapacity: 100.0, CurrentCapacity: 0.0
[hadoop@node1 ~]$

```

3.2.6.1. Configuració de Yarn Scheduler

- L'arxiu de configuració com ja hem vist amb altres està ubicat a /opt/hadoop/etc/hadoop

[hadoop@node1 ~]# gedit capacity-schelduler.xml

```

[hadoop@node1 ~]# cd /opt/hadoop/etc/hadoop/
[hadoop@node1 hadoop]$ ls -ls | grep capacity-scheduler.xml
12 -rw-r--r--. 1 hadoop hadoop 8814 22 oct 2018 capacity-scheduler.xml
[hadoop@node1 hadoop]$

```

```

<!-- The default i.e. DefaultResourceCalculator only uses Memory while
DominantResourceCalculator uses dominant-resource to compare
multi-dimensional resources such as Memory, CPU etc.
-->
</property>
<property>
<name>yarn.scheduler.capacity.root.queues</name>
<value>default</value>
<description>
The queues at the this level (root is the root queue).
</description>
</property>
<property>
<name>yarn.scheduler.capacity.root.default.capacity</name>
<value>100</value>
<description>Default queue target capacity.</description>
</property>
<property>
<name>yarn.scheduler.capacity.root.default.user-limit-factor</name>
<value>1</value>

```

- Crearem noves cues : producció i desenvolupament. Les cues que no estiguin assignades aniran a default



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

```
<property>
  <name>yarn.scheduler.capacity.root.queues</name>
  <value>default,produccio,desenvolupament</value>
  <description>
    The queues at the this level (root is the root queue).
  </description>
</property>

<property>
  <name>yarn.scheduler.capacity.root.default.capacity</name>
  <value>100</value>
  <description>Default queue target capacity.</description>
</property>
```

- Després haurem de modificar la propietat de la capacitat, ja que per defecte el 100% està assignat a default. Li canbiarem els valors
 - queue default : 30%
 - queue produccio 50%
 - queue desenvolupament 20%

```
*capacity-scheduler.xml (/opt/hadoop/etc/yarn)
<?xml version="1.0" encoding="UTF-8"?>
<configuration>
  <!-- Root node -->
  <!-- ResourceCalculator implementation to be used to compare Resources in the scheduler. The default i.e. DefaultResourceCalculator only uses Memory while DominantResourceCalculator uses dominant-resource to compare multi-dimensional resources such as Memory, CPU etc. -->
  <!-- Queues at the this level (root is the root queue). -->
  <!-- Default queue target capacity. -->
  <!-- Desenvolupament queue target capacity. -->
```

- Un cop hem modificat l'arxiu i el guardem ens fa falta activar els canvis.
[hadoop@node1 ~]# yarn rmadmin -refreshQueues
- Accedim a la interfície web del Yarn per veure els canvis realitzats.



Nom i Cognoms

Arnaud Subirós Puigarnau

Data

02-06-2020

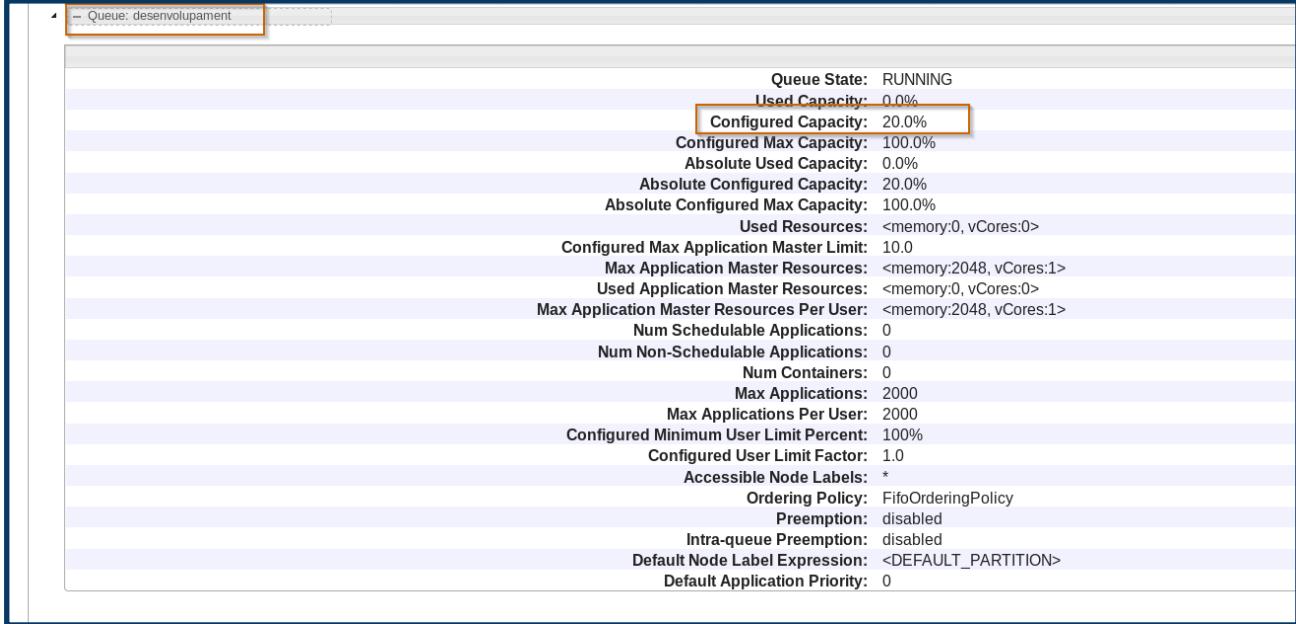
node1:8088

The screenshot shows the Apache YARN scheduler metrics interface. On the left, there's a sidebar with navigation links like 'Scheduler Metrics', 'Scheduler Type', 'Capacity Scheduler type=COUNTABLE', 'Dump scheduler logs 1 min', and 'Application Queues'. The main area displays 'Scheduler Metrics' with values for Capacity, Minimum Allocation, Maximum Allocation, and Maximum Cluster App Priority. Below this is the 'Application Queues' section, which lists four queues: 'root', 'default', 'desenvolupament', and 'produccio'. The 'default' queue is highlighted with an orange box. A legend indicates 'Capacity' (grey), 'Used' (green), 'Used (over capacity)' (orange), 'Max Capacity' (yellow), and 'Users Requesting Resources' (yellow). At the bottom, there's a table titled 'Aggregate scheduler counts' with columns for Total Container Allocations(count), Total Container Releases(count), Total Fulfilled Reservations(count), and Total Container Preemptions(count).

The screenshot shows the Apache YARN Queue configuration page for the 'Queue: default'. The configuration parameters listed include:

- Queue State: RUNNING
- Used Capacity: 0.0%
- Configured Capacity: 30.0%
- Configured Max Capacity: 100.0%
- Absolute Used Capacity: 0.0%
- Absolute Configured Capacity: 30.0%
- Absolute Configured Max Capacity: 100.0%
- Used Resources: <memory:0, vCores:0>
- Configured Max Application Master Limit: 10.0
- Max Application Master Resources: <memory:2048, vCores:1>
- Used Application Master Resources: <memory:0, vCores:0>
- Max Application Master Resources Per User: <memory:2048, vCores:1>
- Num Schedulable Applications: 0
- Num Non-Schedulable Applications: 0
- Num Containers: 0
- Max Applications: 3000
- Max Applications Per User: 3000
- Configured Minimum User Limit Percent: 100%
- Configured User Limit Factor: 1.0
- Accessible Node Labels: *
- Ordering Policy: FifoOrderingPolicy
- Preemption: disabled
- Intra-queue Preemption: disabled
- Default Node Label Expression: <DEFAULT_PARTITION>
- Default Application Priority: 0

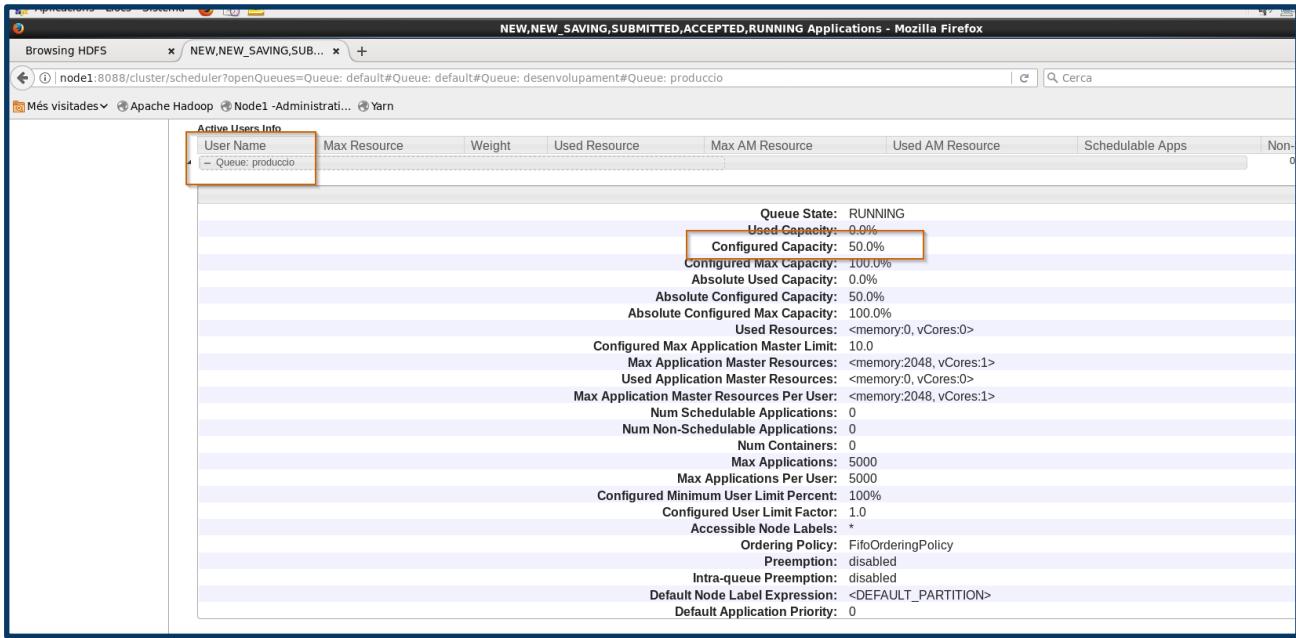
Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020



```

Queue State: RUNNING
Used Capacity: 0.0%
Configured Capacity: 20.0%
Configured Max Capacity: 100.0%
Absolute Used Capacity: 0.0%
Absolute Configured Capacity: 20.0%
Absolute Configured Max Capacity: 100.0%
Used Resources: <memory:0, vCores:0>
Configured Max Application Master Limit: 10.0
Max Application Master Resources: <memory:2048, vCores:1>
Used Application Master Resources: <memory:0, vCores:0>
Max Application Master Resources Per User: <memory:2048, vCores:1>
Num Schedulable Applications: 0
Num Non-Schedulable Applications: 0
Num Containers: 0
Max Applications: 2000
Max Applications Per User: 2000
Configured Minimum User Limit Percent: 100%
Configured User Limit Factor: 1.0
Accessible Node Labels: *
Ordering Policy: FifoOrderingPolicy
Preemption: disabled
Intra-queue Preemption: disabled
Default Node Label Expression: <DEFAULT_PARTITION>
Default Application Priority: 0

```



NEW,NEW_SAVING,SUBMITTED,ACCEPTED,RUNNING Applications - Mozilla Firefox

Browsing HDFS NEW,NEW_SAVING,SUB... +

node1:8088/cluster/scheduler?openQueues=Queue: default#Queue: desenvolupament#Queue: produccio

Més visitades ▾ Apache Hadoop Node1 -Administrati... Yarn

User Name	Max Resource	Weight	Used Resource	Max AM Resource	Used AM Resource	Schedulable Apps	Non-Schedulable Apps
- Queue: produccio							
Queue State: RUNNING Used Capacity: 0.0% Configured Capacity: 50.0% Configured Max Capacity: 100.0% Absolute Used Capacity: 0.0% Absolute Configured Capacity: 50.0% Absolute Configured Max Capacity: 100.0% Used Resources: <memory:0, vCores:0> Configured Max Application Master Limit: 10.0 Max Application Master Resources: <memory:2048, vCores:1> Used Application Master Resources: <memory:0, vCores:0> Max Application Master Resources Per User: <memory:2048, vCores:1> Num Schedulable Applications: 0 Num Non-Schedulable Applications: 0 Num Containers: 0 Max Applications: 5000 Max Applications Per User: 5000 Configured Minimum User Limit Percent: 100% Configured User Limit Factor: 1.0 Accessible Node Labels: * Ordering Policy: FifoOrderingPolicy Preemption: disabled Intra-queue Preemption: disabled Default Node Label Expression: <DEFAULT_PARTITION> Default Application Priority: 0							



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

- Com que ens ha funcionat, crearem 2 cues on la cua arrel no serà root sinó producció i després farem un exemple pràctic . Els valors de les noves cues seran:
- queue produccio 50%
 - producció1 40%
 - producció2 60%

[hadoop@node1 ~]# gedit capacity-schedulader.xml

```
<value>org.apache.hadoop.yarn.util.resource.DefaultResourceCalculator</value>
<description>
  The ResourceCalculator implementation to be used to compare
  Resources in the scheduler.
  The default i.e. DefaultResourceCalculator only uses Memory while
  DominantResourceCalculator uses dominant-resource to compare
  multi-dimensional resources such as Memory, CPU etc.
</description>
</property>

<property>
  <name>yarn.scheduler.capacity.root.queues</name>
  <value>default,produccio,desenvolupament</value>
  <description>
    The queues at the this level (root is the root queue).
  </description>
</property>

<property>
  <name>yarn.scheduler.capacity.root.produccio.queues</name>
  <value>produccio1,produccio2</value>
  <description>
    The queues at the this level (produccio is the root queue).
  </description>
</property>
```

```
<name>yarn.scheduler.capacity.root.default.capacity</name>
<value>30</value>
<description>Default queue target capacity.</description>
</property>

<property>
  <name>yarn.scheduler.capacity.root.produccio.capacity</name>
  <value>50</value>
  <description>Produccio queue target capacity.</description>
</property>

<property>
  <name>yarn.scheduler.capacity.root.desenvolupament.capacity</name>
  <value>20</value>
  <description>Desenvolupament queue target capacity.</description>
</property>

<property>
  <name>yarn.scheduler.capacity.root.produccio.produccio1.capacity</name>
  <value>40</value>
  <description>Produccio1 queue target capacity.</description>
</property>

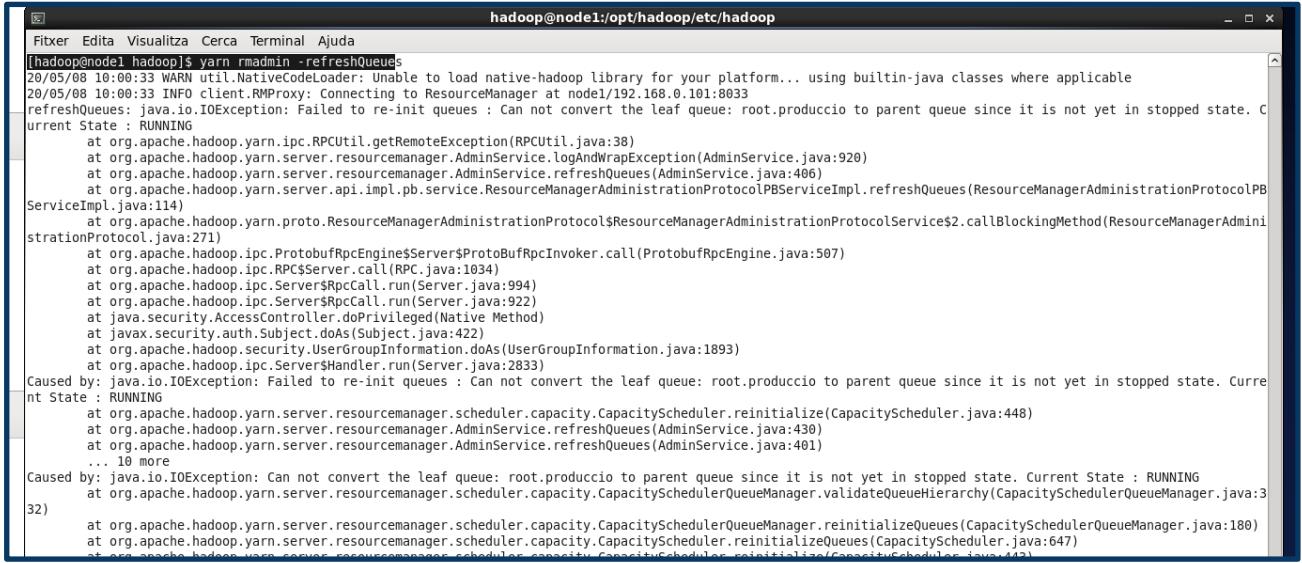
<property>
  <name>yarn.scheduler.capacity.root.produccio.produccio2.capacity</name>
  <value>60</value>
  <description>Produccio2 queue target capacity.</description>
</property>

<property>
  <name>yarn.scheduler.capacity.root.default.user-limit-factor</name>
  <value>1</value>
  <description>
```

Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

- Un cop hem modificat l'arxiu i el guardem ens fa falta activar els canvis.

[hadoop@node1 ~]# yarn rmadmin -refreshQueues



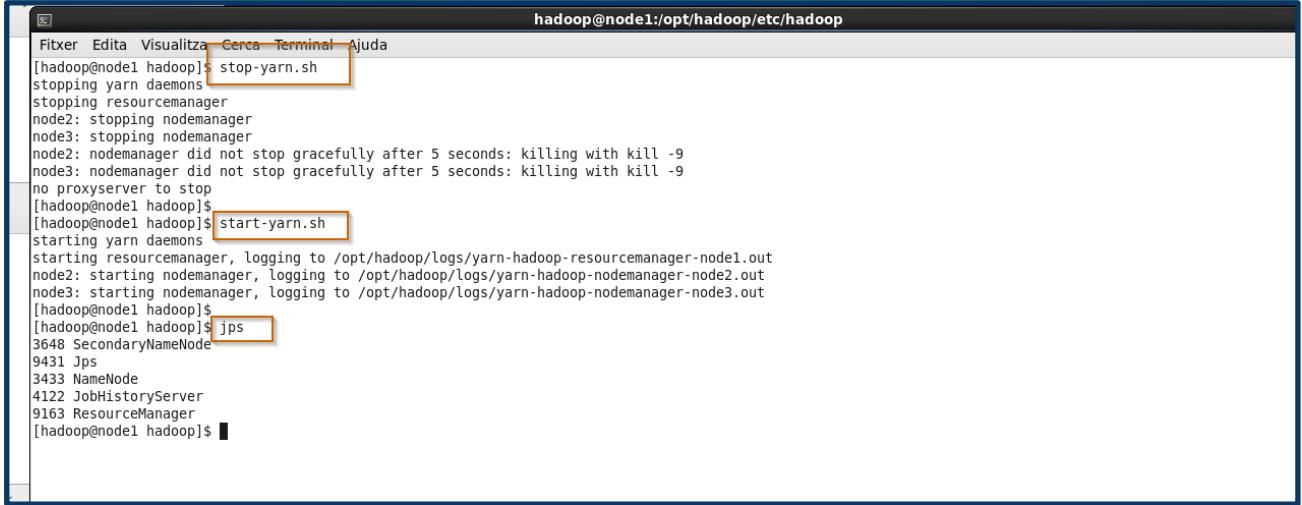
```

[hadoop@node1 hadoop]$ yarn rmadmin -refreshQueues
20/05/08 10:00:33 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
20/05/08 10:00:33 INFO client.RMProxy: Connecting to ResourceManager at node1/192.168.0.101:8033
refreshQueues: java.io.IOException: Failed to re-init queues : Can not convert the leaf queue: root.produccio to parent queue since it is not yet in stopped state. Current State : RUNNING
        at org.apache.hadoop.yarn.ipc.RPCUtil.getRemoteException(RPCUtil.java:38)
        at org.apache.hadoop.yarn.server.resourcemanager.AdminService.logAndWrapException(AdminService.java:920)
        at org.apache.hadoop.yarn.server.resourcemanager.AdminService.refreshQueues(AdminService.java:406)
        at org.apache.hadoop.yarn.server.api.impl.pb.service.ResourceManagerAdministrationProtocolPBServiceImpl.refreshQueues(ResourceManagerAdministrationProtocolPBServiceImpl.java:114)
        at org.apache.hadoop.yarn.proto.ResourceManagerAdministrationProtocol$ResourceManagerAdministrationProtocolService$2.callBlockingMethod(ResourceManagerAdministrationProtocol.java:271)
        at org.apache.hadoop.ipc.ProtobufRpcEngine$Server$ProtoBufRpcInvoker.call(ProtobufRpcEngine.java:507)
        at org.apache.hadoop.ipc.RPC$Server.call(RPC.java:1034)
        at org.apache.hadoop.ipc.ServersRpcCall.run(Server.java:994)
        at org.apache.hadoop.ipc.ServersRpcCall.run(Server.java:922)
        at java.security.AccessController.doPrivileged(Native Method)
        at javax.security.auth.Subject.doAs(Subject.java:422)
        at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1893)
        at org.apache.hadoop.ipc.ServerHandler.run(Server.java:2833)
Caused by: java.io.IOException: Failed to re-init queues : Can not convert the leaf queue: root.produccio to parent queue since it is not yet in stopped state. Current State : RUNNING
        at org.apache.hadoop.yarn.server.resourcemanager.scheduler.capacity.CapacityScheduler.reinitialize(CapacityScheduler.java:448)
        at org.apache.hadoop.yarn.server.resourcemanager.AdminService.refreshQueues(AdminService.java:430)
        at org.apache.hadoop.yarn.server.resourcemanager.AdminService.refreshQueues(AdminService.java:401)
        ... 10 more
Caused by: java.io.IOException: Can not convert the leaf queue: root.produccio to parent queue since it is not yet in stopped state. Current State : RUNNING
        at org.apache.hadoop.yarn.server.resourcemanager.scheduler.capacity.CapacitySchedulerQueueManager.validateQueueHierarchy(CapacitySchedulerQueueManager.java:32)
        at org.apache.hadoop.yarn.server.resourcemanager.scheduler.capacity.CapacitySchedulerQueueManager.reinitializeQueues(CapacitySchedulerQueueManager.java:180)
        at org.apache.hadoop.yarn.server.resourcemanager.scheduler.capacity.CapacityScheduler.reinitializeQueues(CapacityScheduler.java:647)
        at org.apache.hadoop.yarn.server.resourcemanager.scheduler.capacity.CapacityScheduler.reinitialize(CapacityScheduler.java:442)

```

- Ens dóna error, ja que hem assignat 2 cues filla a cues que ja estan funcionant i per tant no puc refrescar. Per solucionar el problema, hauré de parar el clúster (només la part de processos) i tornar-lo a reiniciar.

[hadoop@node1 ~]# stop-yarn.sh
 [hadoop@node1 ~]# start-yarn.sh



```

[hadoop@node1 hadoop]$ stop-yarn.sh
stopping yarn daemons
stopping resourcemanager
node2: stopping nodemanager
node3: stopping nodemanage
node2: nodemanager did not stop gracefully after 5 seconds: killing with kill -9
node3: nodemanager did not stop gracefully after 5 seconds: killing with kill -9
no proxyserver to stop
[hadoop@node1 hadoop]$
[hadoop@node1 hadoop]$ start-yarn.sh
starting yarn daemons
starting resourcemanager, logging to /opt/hadoop/logs/yarn-hadoop-resourcemanager-node1.out
node2: starting nodemanager, logging to /opt/hadoop/logs/yarn-hadoop-nodemanager-node2.out
node3: starting nodemanager, logging to /opt/hadoop/logs/yarn-hadoop-nodemanager-node3.out
[hadoop@node1 hadoop]$
[hadoop@node1 hadoop]$ jps
3648 SecondaryNameNode
9431 Jps
3433 NameNode
4122 JobHistoryServer
9163 ResourceManager
[hadoop@node1 hadoop]$

```



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

- Accedim a la interfície web del Yarn per veure els canvis realitzats.

node1:8088

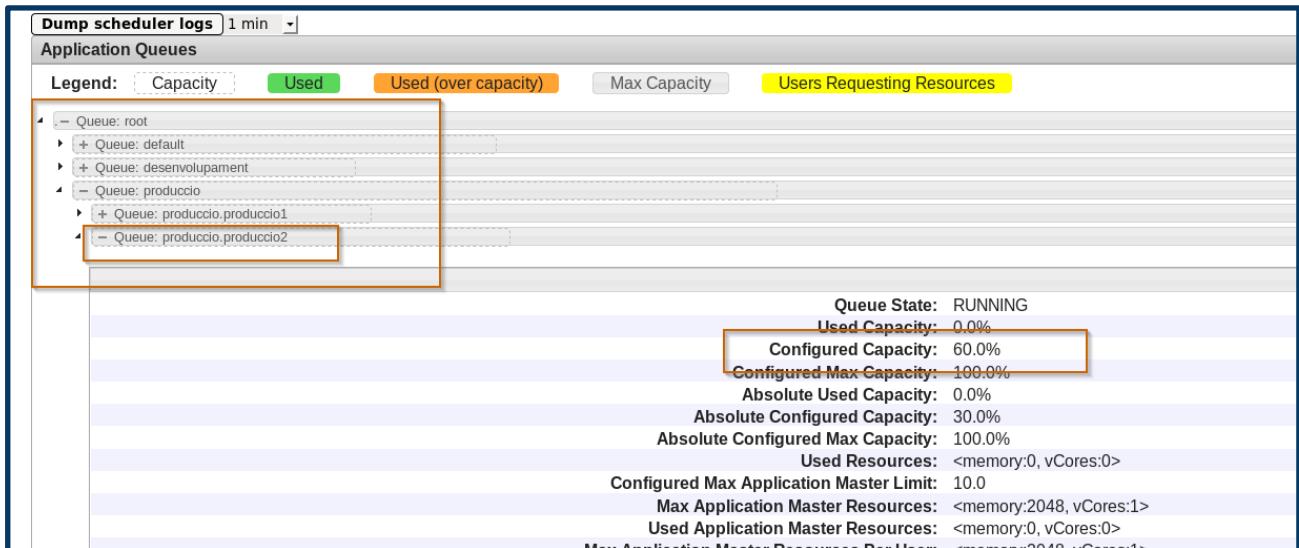
The screenshot shows the Apache Hadoop YARN web interface at node1:8088. It displays various metrics and application queue information. The 'Application Queues' section is highlighted with a red box, showing a tree view of queues: root, default, desenvolupament, produccio, produccio.produccio1, and produccio.produccio2. The 'produccio' queue is specifically highlighted with a red box.

The screenshot shows the Apache Hadoop YARN web interface at node1:8088, focusing on the 'Application Queues' section. A specific entry for the 'produccio' queue is highlighted with a red box. A tooltip provides detailed information about the queue's state and capacity:

```

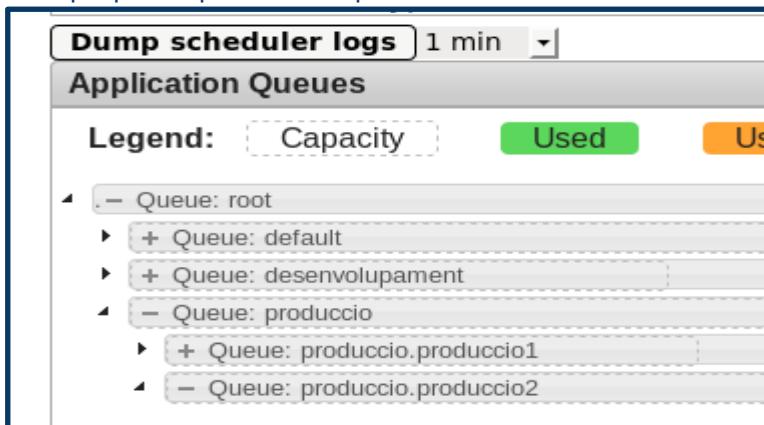
Queue State: RUNNING
Used Capacity: 0.0%
Configured Capacity: 40.0%
Configured Max Capacity: 100.0%
Absolute Used Capacity: 0.0%
Absolute Configured Capacity: 20.0%
Absolute Configured Max Capacity: 100.0%
Used Resources: <memory:0, vCores:0>
Configured Max Application Master Limit: 10.0
Max Application Master Resources: <memory:2048, vCores:1>
Used Application Master Resources: <memory:0, vCores:0>
Max Application Master Resources Per User: <memory:2048, vCores:1>
Num Schedulable Applications: 0
  
```

Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020



3.2.6.1.1. Exemple : llençar un job

- Un cop tenim configurat el Yarn Scheduler amb la jerarquia de cues i la seva capacitat, farem un exemple pràctic per verificar que funciona

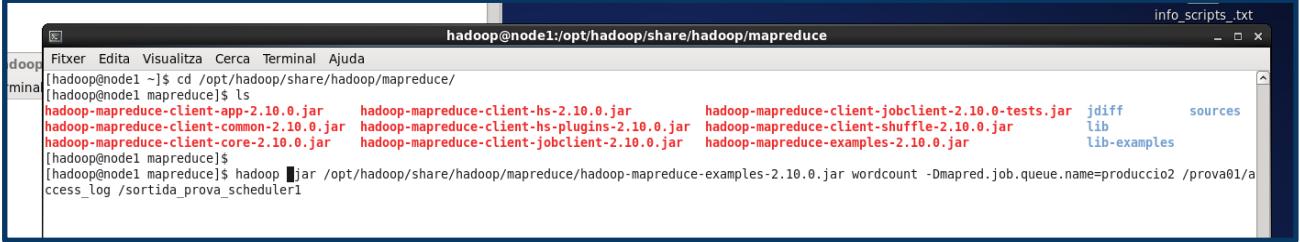


- Farem servir un exemple Mapreduce anomenat “wordcount”.

Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

- **-Dmapred.job.queue.name** :Especificuem la cua
- origen del fitxer : /prova01/access_log
- desti de resultat : /sortida_prova_scheduler1

```
[hadoop@node1 ~]# hadoop jar /opt/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.10.0.jar wordcount -Dmapred.job.queue.name=produccio2 /prova01/access_log /sortida_prova_scheduler1
```

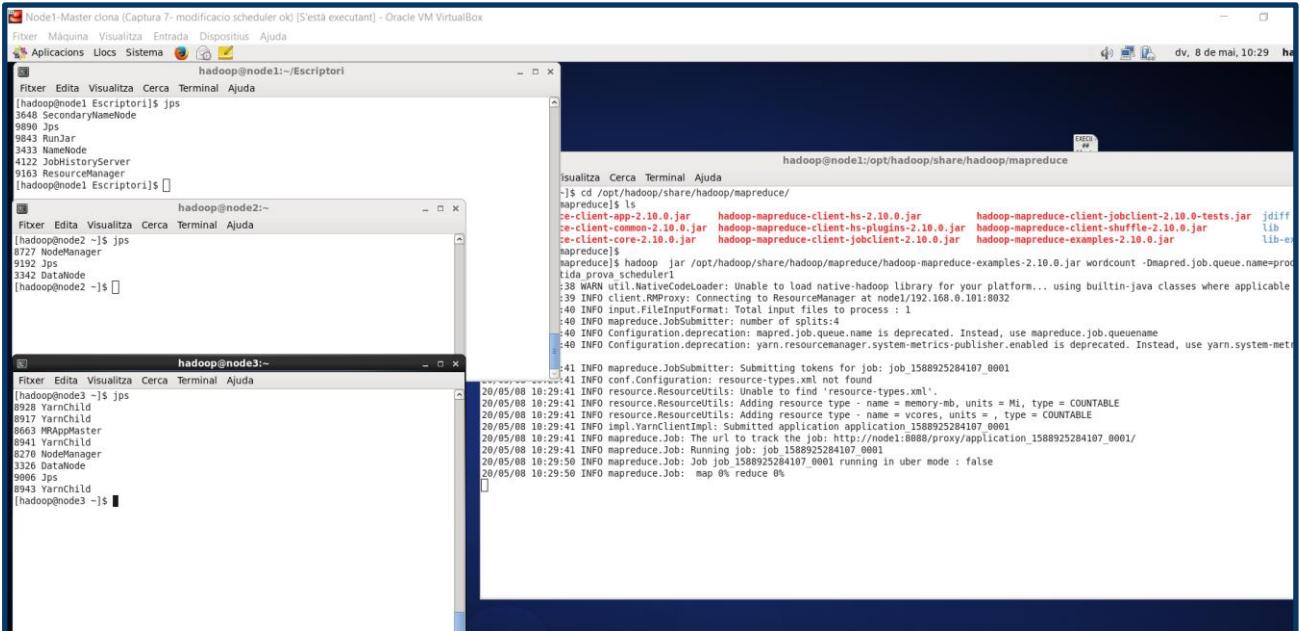


The terminal window title is "hadoop@node1:/opt/hadoop/share/hadoop/mapreduce". The command entered is:

```
[hadoop@node1 ~]$ hadoop jar /opt/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.10.0.jar wordcount -Dmapred.job.queue.name=produccio2 /prova01/access_log /sortida_prova_scheduler1
```

□ Executem jps en tots els nodes per veure els processos.

```
[hadoop@node1 ~]# jps
[hadoop@node2 ~]# jps
[hadoop@node3 ~]# jps
```



The terminal windows show the following outputs:

- Node1 (Master):**

```
[hadoop@node1 ~]# jps
3648 SecondaryNameNode
9890 Jps
9843 ResourceManager
9431 NameNode
4122 JobHistoryServer
9163 ResourceManager
[hadoop@node1 ~]#
```
- Node2 (DataNode):**

```
[hadoop@node2 ~]# jps
8727 NodeManager
9192 Jps
3342 DataNode
[hadoop@node2 ~]#
```
- Node3 (DataNode):**

```
[hadoop@node3 ~]# jps
8928 YarnChild
8917 YarnChild
8863 YarnChild
8941 YarnChild
8278 NodeManager
3326 DataNode
9006 Jps
8943 YarnChild
[hadoop@node3 ~]#
```



Nom i Cognoms

Arnau Subirós Puigarnau

Data

02-06-2020

- Accedim a l'interfície web de Yarn i observem el Schelduler mentre els processos estan funcionant

The screenshot shows the Apache Hadoop YARN scheduler interface. The top navigation bar includes links for Fitxer, Màquina, Visualitzar, Entrada, Dispositius, Ajuda, Aplicacions, Llocs, Sistemes, and a search bar. The main content area displays 'Cluster Metrics' and 'Cluster Nodes Metrics' tables. Below these are sections for 'Scheduler Metrics' and 'Application Queues'. The 'Application Queues' section includes a legend for Capacity, Used, Used (over capacity), Max Capacity, and Users Requesting Resources. It lists four queues: Queue: root, Queue: default, Queue: desenvolupament, and Queue: produccio. At the bottom, there is a table of 'Application Queues' with columns: ID, User, Name, Application Type, Queue, Application Priority, StartTime, LaunchTime, FinishTime, State, FinalStatus, Running Containers, Allocated CPU Vcores, and Allocated Mem MB. One entry is shown: application_1588925284107_0001, hadoop, word count, MAPREDUCE, produccio2, 0, Fri May 8 10:29:41 +0200 2020, Fri May 8 10:29:42 +0200 2020, N/A, RUNNING, UNDEFINED, 5, 5, 6144.

This screenshot is identical to the one above, showing the Apache Hadoop YARN scheduler interface. It displays the same cluster metrics, application queues, and the same specific application entry in the queue table. The application entry is application_1588925284107_0001, hadoop, word count, MAPREDUCE, produccio2, 0, Fri May 8 10:29:41 +0200 2020, Fri May 8 10:29:42 +0200 2020, N/A, RUNNING, UNDEFINED, 5, 5, 6144.



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

3.2.6.2. Ecosistema Hadoop

- A més a més del clúster Hadoop i els seus components bàsics. A continuació mostrem la instal·lació, configuració i algun exemple de Hive i Spark.
Nota: Tot i que vàrem instal·lar Ambari no ho mostrarem, ja que té un apartat propi on es crearà un nou clúster de 0 utilitzant Ambari i 3 màquines CentOS (les mateixes màquines clonades sense Hadoop)

3.2.6.2.1. Instal.lació i configuració de Hive

- Accedim a la pàgina oficial, per realitzar la descarrega

<https://hive.apache.org/downloads.html>

Aplicacions Llocs Sistema Downloads

Browsing HDFS NEW,NEW_SAVING,SUB... Downloads

https://hive.apache.org/downloads.html

Més visitades Apache Hadoop Node1 -Administrati... Yarn

HIVE

GENERAL

- Home
- Downloads
- License
- Privacy Policy

DOCUMENTATION

- Language Manual
- Javadoc
- Wiki

DOWNLOADS

Releases may be downloaded from Apache mirrors:

[Download a release now!](#)

On the mirror, all recent releases are available, but are not guaranteed to be stable. For stable releases, look in the stable directory.

News

Apache Download Mirrors - Mozilla Firefox

Browsing HDFS NEW,NEW_SAVING,SUB... Apache Download M... +

https://www.apache.org/dyn/closer.cgi/hive/

Més visitades Apache Hadoop Node1 -Administrati... Yarn

News About Make a Donation The Apache Way Join Us Downloads

THE APACHE SOFTWARE FOUNDATION 20TH ANNIVERSARY

COMMUNITY-LED DEVELOPMENT "THE APACHE WAY"

We suggest the following mirror site for your download:

<https://apache.brunneis.com/hive/>

Other mirror sites are suggested below.

It is essential that you verify the integrity of the downloaded file using the PGP signature (.asc file) or a hash (.md5 or .sha* file).

Please only use the backup mirrors to download KEYS, PGP signatures and hashes (SHA* etc) -- or if no other mirrors are working.

HTTP

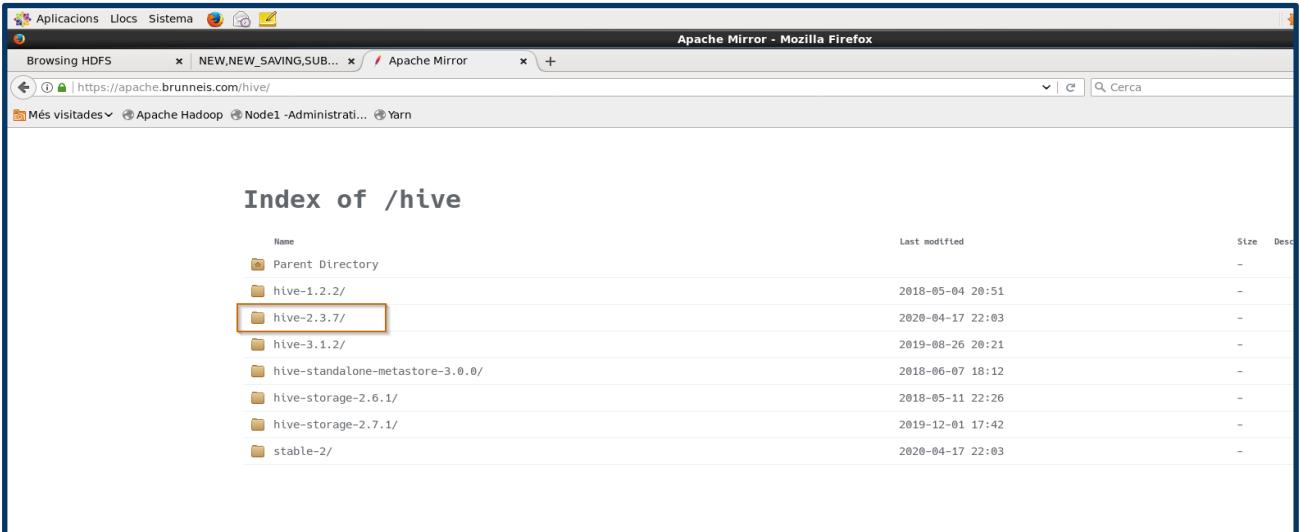
<http://apache.uvigo.es/hive/>

<https://apache.brunneis.com/hive/>

SUPPORT APACHE

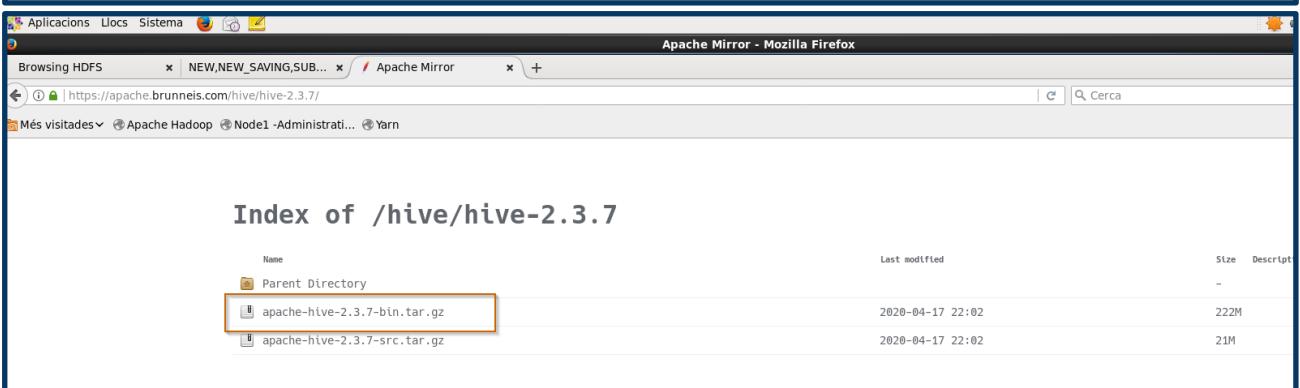
Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

□ Ens descarreguem la versió 2.3.7



Index of /hive

Name	Last modified	Size	Description
Parent Directory		-	
hive-1.2.2/	2018-05-04 20:51	-	
hive-2.3.7/	2020-04-17 22:03	-	
hive-3.1.2/	2019-08-26 20:21	-	
hive-standalone-metastore-3.0.0/	2018-06-07 18:12	-	
hive-storage-2.6.1/	2018-05-11 22:26	-	
hive-storage-2.7.1/	2019-12-01 17:42	-	
stable-2/	2020-04-17 22:03	-	

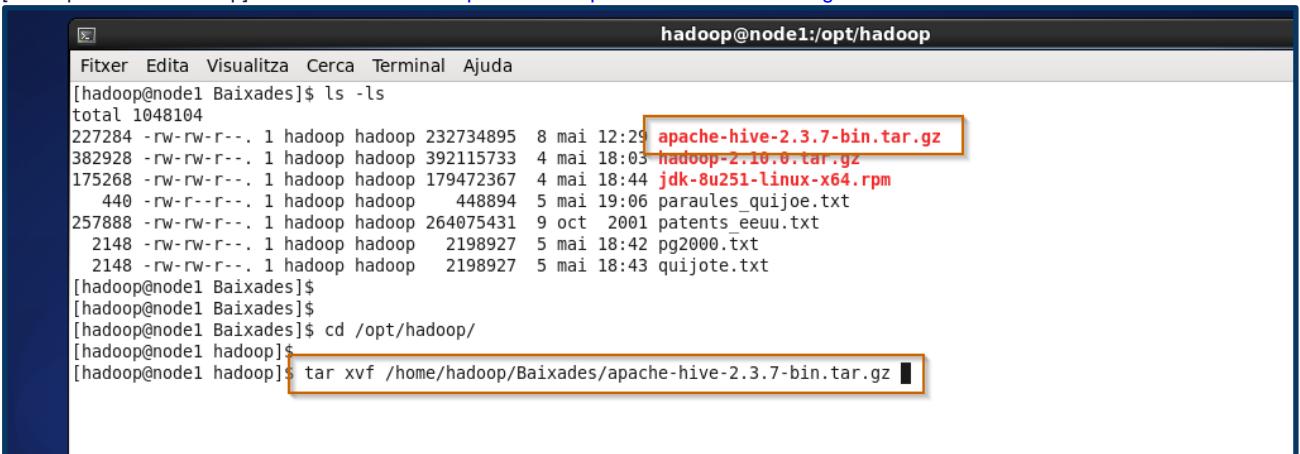


Index of /hive/hive-2.3.7

Name	Last modified	Size	Description
Parent Directory		-	
apache-hive-2.3.7-bin.tar.gz	2020-04-17 22:02	222M	
apache-hive-2.3.7-src.tar.gz	2020-04-17 22:02	21M	

□ Descomprimim l'arxiu i l'ubiquem dins el directori `/opt/hadoop`

```
[hadoop@node1 hadoop]# tar xvf /home/hadoop/Baixades/apache-hive-2.3.7-bin.tar.gz
```



```

[hadoop@node1 Baixades]$ ls -ls
total 1048104
227284 -rw-rw-r-- 1 hadoop hadoop 232734895 8 mai 12:29 apache-hive-2.3.7-bin.tar.gz
382928 -rw-rw-r-- 1 hadoop hadoop 392115733 4 mai 18:03 hadoop-2.10.0.tar.gz
175268 -rw-rw-r-- 1 hadoop hadoop 179472367 4 mai 18:44 jdk-8u251-linux-x64.rpm
440 -rw-r--r-- 1 hadoop hadoop 448894 5 mai 19:06 paraules_quijoe.txt
257888 -rw-rw-r-- 1 hadoop hadoop 264075431 9 oct 2001 patents_eeuu.txt
2148 -rw-rw-r-- 1 hadoop hadoop 2198927 5 mai 18:42 pg2000.txt
2148 -rw-rw-r-- 1 hadoop hadoop 2198927 5 mai 18:43 quijote.txt
[hadoop@node1 Baixades]$
[hadoop@node1 Baixades]$
[hadoop@node1 Baixades]$ cd /opt/hadoop/
[hadoop@node1 hadoop]$
[hadoop@node1 hadoop]$ tar xvf /home/hadoop/Baixades/apache-hive-2.3.7-bin.tar.gz

```



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

- Desplaçem els arxius del directori creat, ja que es molt llarg a un nou directori anomenat hive

[hadoop@node1 hadoop]# mv apache-hive-2.3.7-bin/ hive

```
hadoop@node1:/opt/hadoop
Fitxa Edita Visualitzar Cerca Terminal Ajuda
[hadoop@node1 hadoop]$ ls -ls
total 160
4 drwxrwxr-x. 10 hadoop hadoop 4096 8 mai 12:38 apache-hive-2.3.7-bin
4 drwxr-xr-x. 2 hadoop hadoop 4096 5 mai 12:26 bin
4 drwxr-xr-x. 3 hadoop hadoop 4096 22 oct 2019 etc
4 drwxr-xr-x. 2 hadoop hadoop 4096 22 oct 2019 include
4 drwxr-xr-x. 3 hadoop hadoop 4096 22 oct 2019 lib
4 drwxr-xr-x. 2 hadoop hadoop 4096 22 oct 2019 libexec
104 -rw-r--r--. 1 hadoop hadoop 106210 22 oct 2019 LICENSE.txt
4 drwxrwxr-x. 3 hadoop hadoop 4096 8 mai 10:07 logs
16 -rw-r--r--. 1 hadoop hadoop 15841 22 oct 2019 NOTICE.txt
4 -rw-r--r--. 1 hadoop hadoop 1366 22 oct 2019 README.txt
4 drwxr-xr-x. 3 hadoop hadoop 4096 22 oct 2019 sbin
4 drwxr-xr-x. 4 hadoop hadoop 4096 22 oct 2019 share
[hadoop@node1 hadoop]$ ^C
[hadoop@node1 hadoop]$ mv apache-hive-2.3.7-bin/ hive
[hadoop@node1 hadoop]$ ls -ls
total 160
4 drwxr-xr-x. 2 hadoop hadoop 4096 5 mai 12:26 bin
4 drwxr-xr-x. 3 hadoop hadoop 4096 22 oct 2019 etc
4 drwxrwxr-x. 10 hadoop hadoop 4096 8 mai 12:38 hive
4 drwxr-xr-x. 2 hadoop hadoop 4096 22 oct 2019 include
4 drwxr-xr-x. 3 hadoop hadoop 4096 22 oct 2019 lib
4 drwxr-xr-x. 2 hadoop hadoop 4096 22 oct 2019 libexec
104 -rw-r--r--. 1 hadoop hadoop 106210 22 oct 2019 LICENSE.txt
4 drwxrwxr-x. 3 hadoop hadoop 4096 8 mai 10:07 logs
16 -rw-r--r--. 1 hadoop hadoop 15841 22 oct 2019 NOTICE.txt
4 -rw-r--r--. 1 hadoop hadoop 1366 22 oct 2019 README.txt
4 drwxr-xr-x. 3 hadoop hadoop 4096 22 oct 2019 sbin
4 drwxr-xr-x. 4 hadoop hadoop 4096 22 oct 2019 share
[hadoop@node1 hadoop]$
```

- Haurem d'afegir les variables d'entorn de l'arxiu `~/.bashrc`

[hadoop@node1 hadoop]# gedit ~/.bashrc

```
hadoop@node1:/opt/hadoop
Fitxa Edita Visualitzar Cerca Terminal Ajuda
[hadoop@node1 hadoop]$ gedit ~/.bashrc
```

The .bashrc file contains the following code:

```
#!/bin/bash
# Source global definitions
if [ -f /etc/bashrc ]; then
    . /etc/bashrc
fi

# User specific aliases and functions

## CREAR una variable i exportar-la globalment
export HADOOP_HOME=/opt/hadoop
export PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin:$HIVE_HOME/bin
export JAVA_HOME=/usr/java/jdk1.8.0_251-amd64
export HIVE_HOME=/opt/hadoop/hive
```



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

```
hadoop@node1:/opt/hadoop/hive
```

Fitxer Edita Visualitza Cerca Terminal Ajuda

```
[hadoop@node1 hadoop]$ cd hive/
[hadoop@node1 hive]$ ls -ls
total 76
4 drwxrwxr-x. 3 hadoop hadoop 4096 8 mai 12:38 bin
4 drwxrwxr-x. 2 hadoop hadoop 4096 8 mai 12:38 binary-package-licenses
4 drwxrwxr-x. 2 hadoop hadoop 4096 8 mai 12:38 conf
4 drwxrwxr-x. 4 hadoop hadoop 4096 8 mai 12:38 examples
4 drwxrwxr-x. 7 hadoop hadoop 4096 8 mai 12:38 hcatalog
4 drwxrwxr-x. 2 hadoop hadoop 4096 8 mai 12:38 jdbc
16 drwxrwxr-x. 4 hadoop hadoop 16384 8 mai 12:38 lib
24 -rw-r--r--. 1 hadoop hadoop 20798 9 mar 17:09 LICENSE
4 -rw-r--r--. 1 hadoop hadoop 230 7 abr 20:57 NOTICE
4 -rw-r--r--. 1 hadoop hadoop 361 7 abr 21:37 RELEASE_NOTES.txt
4 drwxrwxr-x. 4 hadoop hadoop 4096 8 mai 12:38 scripts
[hadoop@node1 hive]$ ls -ls bin
total 40
4 -rwxr-xr-x. 1 hadoop hadoop 881 21 ago 2017 beeline
4 drwxrwxr-x. 3 hadoop hadoop 4096 8 mai 12:38 ext
12 -rwxr-xr-x. 1 hadoop hadoop 9838 7 abr 20:46 hive
4 -rwxr-xr-x. 1 hadoop hadoop 1900 16 des 2016 hive-config.sh
4 -rwxr-xr-x. 1 hadoop hadoop 885 16 des 2016 hiveserver2
4 -rwxr-xr-x. 1 hadoop hadoop 880 2 mai 2017 hql
4 -rwxr-xr-x. 1 hadoop hadoop 832 16 des 2016 metatool
4 -rwxr-xr-x. 1 hadoop hadoop 884 16 des 2016 schematool
[hadoop@node1 hive]$
```

- Seguidament haurem d'editar l'arxiu de configuració de Hive ubicat al directori </opt/hadoop/hive/conf>
Com es pot observar, la majoria son templates

```
hadoop@node1:/opt/hadoop/hive/conf
```

Fitxer Edita Visualitza Cerca Terminal Ajuda

```
[hadoop@node1 conf]$ pwd
/opt/hadoop/hive/conf
[hadoop@node1 conf]$ ls -ls
total 288
4 -rw-r--r--. 1 hadoop hadoop 1596 16 des 2016 beeline-log4j2.properties.template
252 -rw-r--r--. 1 hadoop hadoop 257573 7 abr 21:42 hive-default.xml.template
4 -rw-r--r--. 1 hadoop hadoop 2365 2 mai 2017 hive-env.sh.template
4 -rw-r--r--. 1 hadoop hadoop 2274 8 mar 2017 hive-exec-log4j2.properties.template
4 -rw-r--r--. 1 hadoop hadoop 2925 9 mar 17:09 hive-log4j2.properties.template
4 -rw-r--r--. 1 hadoop hadoop 2060 8 mar 2017 ivysettings.xml
4 -rw-r--r--. 1 hadoop hadoop 2719 9 mar 17:09 llap-cli-log4j2.properties.template
8 -rw-r--r--. 1 hadoop hadoop 7041 9 mar 17:09 llap-daemon-log4j2.properties.template
4 -rw-r--r--. 1 hadoop hadoop 2662 8 mar 2017 parquet-logging.properties
[hadoop@node1 conf]$
```



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

- Copiarem les templates que ens interessa modificar canviant el nom de l'arxiu (sense el template final)

- [hadoop@node1 conf]\$ cp hive-default.xml.template hive-site.xml
- [hadoop@node1 conf]\$ cp hive-env.sh.template hive-env.sh
- [hadoop@node1 conf]\$ cp hive-exec-log4j2.properties.template hive-exec-log4j2.properties
- [hadoop@node1 conf]\$ cp beeline-log4j2.properties.template beeline-log4j2.properties

The screenshot shows two terminal windows side-by-side. Both windows have a title bar 'hadoop@node1:/opt/hadoop/hive/conf'. The left window shows the user navigating to the directory and listing files:

```
[hadoop@node1 conf]$ pwd  
/opt/hadoop/hive/conf  
[hadoop@node1 conf]$ ls -ls  
total 288  
4 -rw-r--r--. 1 hadoop hadoop 1596 16 des 2016 beeline-log4j2.properties.template  
252 -rw-r--r--. 1 hadoop hadoop 257573 7 abr 21:42 hive-default.xml.template  
4 -rw-r--r--. 1 hadoop hadoop 2365 2 mai 2017 hive-env.sh.template  
4 -rw-r--r--. 1 hadoop hadoop 2274 8 mar 2017 hive-exec-log4j2.properties.template  
4 -rw-r--r--. 1 hadoop hadoop 2925 9 mar 17:09 hive-log4j2.properties.template  
4 -rw-r--r--. 1 hadoop hadoop 2060 8 mar 2017 ivysettings.xml  
4 -rw-r--r--. 1 hadoop hadoop 2719 9 mar 17:09 llap-cli-log4j2.properties.template  
8 -rw-r--r--. 1 hadoop hadoop 7041 9 mar 17:09 llap-daemon-log4j2.properties.template  
4 -rw-r--r--. 1 hadoop hadoop 2662 8 mar 2017 parquet-logging.properties  
[hadoop@node1 conf]$  
[hadoop@node1 conf]$
```

The right window shows the user running the cp command to copy the template files:

```
[hadoop@node1 Escriptori]$ cd /opt/hadoop/hive/conf/  
[hadoop@node1 conf]$ cp hive-default.xml.template hive-default.xml  
[hadoop@node1 conf]$ cp hive-env.sh.template hive-env.sh  
[hadoop@node1 conf]$ cp hive-exec-log4j2.properties.template hive-exec-log4j2.properties  
[hadoop@node1 conf]$ cp beeline-log4j2.properties.template beeline-log4j2.properties  
[hadoop@node1 conf]$ ls -ls  
total 552  
4 -rw-r--r--. 1 hadoop hadoop 1596 8 mai 13:10 beeline-log4j2.properties  
4 -rw-r--r--. 1 hadoop hadoop 1596 16 des 2016 beeline-log4j2.properties.template  
252 -rw-r--r--. 1 hadoop hadoop 257573 8 mai 13:07 hive-default.xml  
252 -rw-r--r--. 1 hadoop hadoop 257573 7 abr 21:42 hive-default.xml.template  
4 -rw-r--r--. 1 hadoop hadoop 2365 8 mai 13:08 hive-env.sh  
4 -rw-r--r--. 1 hadoop hadoop 2365 2 mai 2017 hive-env.sh.template  
4 -rw-r--r--. 1 hadoop hadoop 2274 8 mai 13:09 hive-exec-log4j2.properties  
4 -rw-r--r--. 1 hadoop hadoop 2274 8 mar 2017 hive-exec-log4j2.properties.template  
4 -rw-r--r--. 1 hadoop hadoop 2925 9 mar 17:09 hive-log4j2.properties.template  
4 -rw-r--r--. 1 hadoop hadoop 2060 8 mar 2017 ivysettings.xml  
4 -rw-r--r--. 1 hadoop hadoop 2719 9 mar 17:09 llap-cli-log4j2.properties.template  
8 -rw-r--r--. 1 hadoop hadoop 7041 9 mar 17:09 llap-daemon-log4j2.properties.template  
4 -rw-r--r--. 1 hadoop hadoop 2662 8 mar 2017 parquet-logging.properties  
[hadoop@node1 conf]$
```

In the right window, the commands being run are highlighted with orange boxes: 'cp hive-default.xml.template hive-default.xml', 'cp hive-env.sh.template hive-env.sh', and 'cp hive-exec-log4j2.properties.template hive-exec-log4j2.properties'.



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

□ Accedim a l'arxiu **hive-env.sh** per definir 2 variables:

- HADOOP_HOME
- HIVE_CONF_DIR

[hadoop@node1 conf]\$ gedit hive-env.sh

```

[hadoop@node1 conf]$ ls -ls | grep hive-env.sh
4 -rw-r--r--. 1 hadoop hadoop 2365 8 mai 13:08 hive-env.sh
4 -rw-r--r--. 1 hadoop hadoop 2365 2 mai 2017 hive-env.sh.template
[hadoop@node1 conf]$ gedit hive-env.sh

[hadoop@node1:/opt/hadoop/hive/conf] Fitxa Edita Visualitza Cerca Terminal Ajuda
[hadoop@node1 conf]$ ls -ls | grep hive-env.sh
4 -rw-r--r--. 1 hadoop hadoop 2365 8 mai 13:08 hive-env.sh
4 -rw-r--r--. 1 hadoop hadoop 2365 2 mai 2017 hive-env.sh.template
[hadoop@node1 conf]$ gedit hive-env.sh

[hadoop@node1:/opt/hadoop/hive/conf] *hive-env.sh
# Hive Client memory usage can be an issue if a large number of clients
# are running at the same time. The flags below have been useful in
# reducing memory usage:
#
# if [ "$SERVICE" = "cli" ]; then
#   if [ -z "$DEBUG" ]; then
#     export HADOOP_OPTS="$HADOOP_OPTS -XX:NewRatio=12 -Xms10m -XX:MaxHe
+UseParNewGC -XX:+UseGCOverheadLimit"
#   else
#     export HADOOP_OPTS="$HADOOP_OPTS -XX:NewRatio=12 -Xms10m -XX:MaxHe
UseGCOverheadLimit"
#   fi
# fi

# The heap size of the jvm started by hive shell script can be controlled
#
# export HADOOP_HEAPSIZE=1024
#
# Larger heap size may be required when running queries over large numbers
# By default hive shell scripts use a heap size of 256 (MB). Larger heap
# appropriate for hive server.

# Set HADOOP_HOME to point to a specific hadoop install directory
# HADOOP_HOME=$sbin/../../..;/hadoop
export HADOOP_HOME=/opt/hadoop

# Hive Configuration Directory can be controlled by:
# export HIVE_CONF_DIR
export HIVE_CONF_DIR=/opt/hadoop/hive/conf

```

□ Posteriorment crearem 2 directoris HDFS (**/tmp** i **/user/hive/warehouse**) i li assignem permisos, ja que és el lloc (per defecte) treballa Hive. Si no els tenim, ens podria fallar.

[hadoop@node1 conf]\$ hdfs dfs -mkdir /tmp
[hadoop@node1 conf]\$ hdfs dfs -chmod g+w /tmp
[hadoop@node1 conf]\$ hdfs dfs -mkdir -p /user/hive/warehouse
[hadoop@node1 conf]\$ hdfs dfs -chmod g+w /user/hive/warehouse

```

[hadoop@node1 conf]$ hdfs dfs -mkdir /tmp
[hadoop@node1 conf]$ hdfs dfs -chmod g+w /tmp
20/05/08 13:45:16 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
mkdir: `/tmp': File exists
[hadoop@node1 conf]$ hdfs dfs -mkdir -p /user/hive/warehouse
[hadoop@node1 conf]$ hdfs dfs -chmod g+w /user/hive/warehouse
20/05/08 13:45:35 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
[hadoop@node1 conf]$ hdfs dfs -chmod g+w /user/hive/warehouse
20/05/08 13:46:06 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

```

[hadoop@node1:~/Escriptori]

```

[hadoop@node1 Escriptori]$ hdfs dfs -ls /
20/05/08 13:49:42 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform.
. using builtin-java classes where applicable
Found 13 items
drwxr-xr-x  hadoop supergroup          0 2020-05-07 11:23 /log_sortida2
drwxr-xr-x  - hadoop supergroup          0 2020-05-07 12:07 /log_sortida3
drwxr-xr-x  - hadoop supergroup          0 2020-05-07 10:41 /prova01
drwxr-xr-x  - hadoop supergroup          0 2020-05-07 12:57 /prova02
drwxr-xr-x  - hadoop supergroup          0 2020-05-07 13:37 /prova02_sortida
drwxr-xr-x  - hadoop supergroup          0 2020-05-07 17:40 /prova02_sortida2
drwxr-xr-x  - hadoop supergroup          0 2020-05-07 18:00 /prova3_python
drwxr-xr-x  - hadoop supergroup          0 2020-05-07 18:28 /prova3_python02
drwxr-xr-x  - hadoop supergroup          0 2020-05-07 18:35 /prova3_python03
drwxr-xr-x  - hadoop supergroup          0 2020-05-07 18:47 /prova3_python04
drwxr-xr-x  - hadoop supergroup          0 2020-05-08 10:30 /sortida_prova_scheduler1
drwxrwx---  - hadoop supergroup          0 2020-05-07 10:23 /Tmp
drwxr-xr-x  - hadoop supergroup          0 2020-05-08 13:45 /user

```



Nom i Cognoms

Arnau Subirós Puigarnau

Data

02-06-2020

Browsing HDFS NEW,NEW_SAVING,SUB... Apache Mirror node1:50070/explorer.html#/user/hive/warehouse

Més visitades ▾ Apache Hadoop Node1 -Administraci... Yarn

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities ▾

Browse Directory

/user/hive/warehouse

Show 25 entries

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
No data available in table							

Showing 0 to 0 of 0 entries

Hadoop, 2019.

Browsing HDFS NEW,NEW_SAVING,SUB... Apache Mirror node1:50070/explorer.html#/
visitades ▾ Apache Hadoop Node1 -Administraci... Yarn

Browse Directory

Permission denied: user=dr.who, access=READ_EXECUTE, inode="/tmp":hadoop:supergroup:drwxrwx---

hadoop@node1:~/Baixades

Fitxer	Edita	Visualitza	Cerca	Terminal	Ajuda
[hadoop@node1 Baixades]\$ hdfs dfs -ls /tmp					
20/05/08 17:21:00 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable					
Found 1 items					
drwxrwx--- - hadoop supergroup 0 2020-05-07 10:23 /tmp/hadoop-yarn					
[hadoop@node1 Baixades]\$					

- Com es pot veure no li havia donat tots els permisos. Modifico els permisos
[hadoop@node1 conf]\$ hdfs dfs -chmod 755 /tmp

Browsing HDFS NEW,NEW_SAVING,SUB... Apache Mirror node1:50070/explorer.html#/
Més visitades ▾ Apache Hadoop Node1 -Administraci... Yarn

Browse Directory

/

Show 25 entries

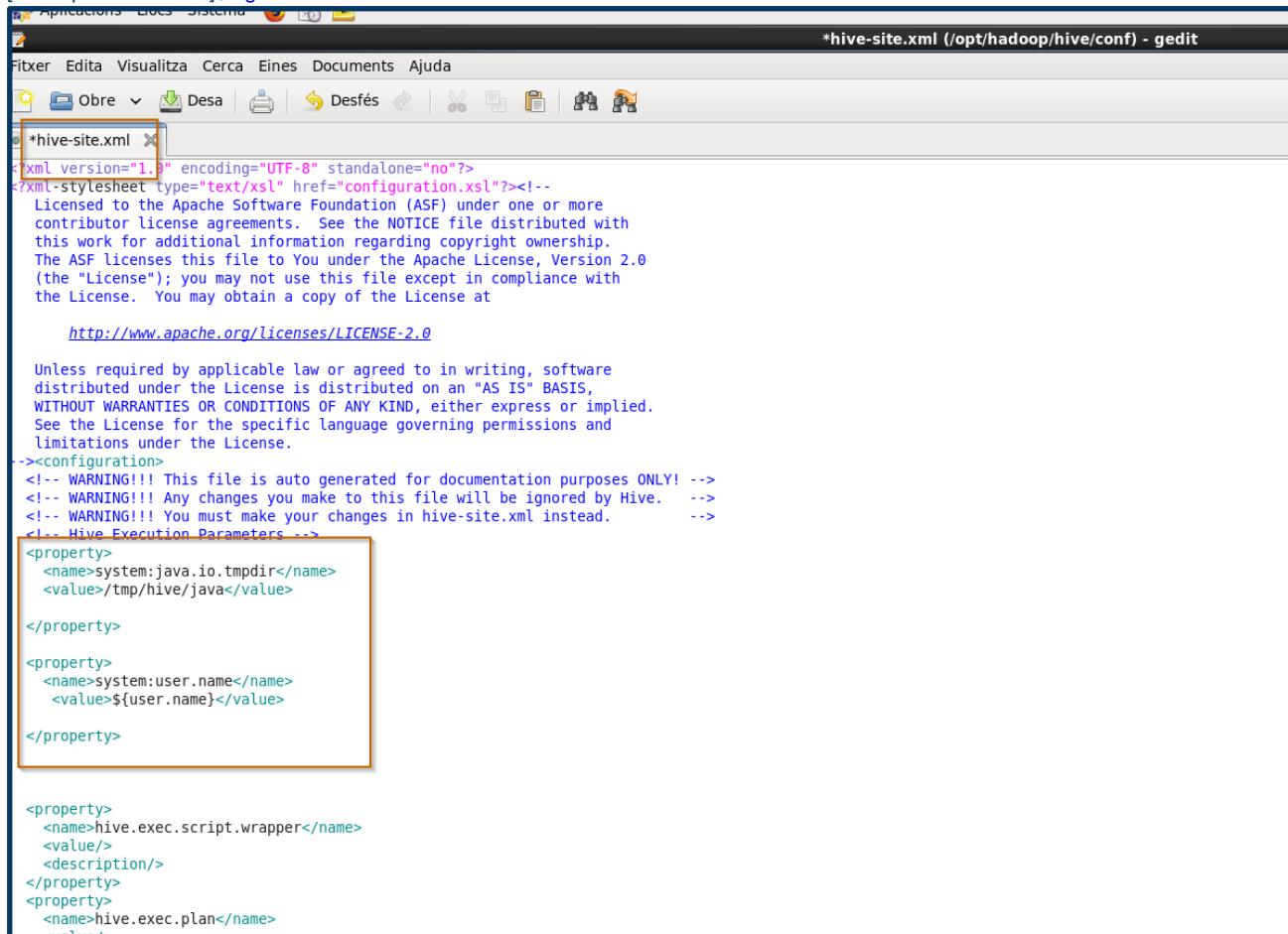
Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
Fitxer Edita Visualitza Cerca Terminal Ajuda							
[hadoop@node1 Baixades]\$ hdfs dfs -ls /tmp							
20/05/08 17:28:46 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable							
[hadoop@node1 Baixades]\$ hdfs dfs -ls /							
20/05/08 17:28:55 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable							
Found 13 items							
drwxr-xr-x	- hadoop supergroup	0	2020-05-07 11:23	/log	sortida2		
drwxr-xr-x	- hadoop supergroup	0	2020-05-07 12:07	/log	sortida3		
drwxr-xr-x	- hadoop supergroup	0	2020-05-07 10:41	/prova01			
drwxr-xr-x	- hadoop supergroup	0	2020-05-07 12:57	/prova02			
drwxr-xr-x	- hadoop supergroup	0	2020-05-07 13:37	/prova03	sortida		
drwxr-xr-x	- hadoop supergroup	0	2020-05-07 17:48	/prova04	sortida2		
drwxr-xr-x	- hadoop supergroup	0	2020-05-07 17:48	/prova05	sortida3		
drwxr-xr-x	- hadoop supergroup	0	2020-05-07 18:26	/prova06	prova01		
drwxr-xr-x	- hadoop supergroup	0	2020-05-07 18:35	/prova07	prova02		
drwxr-xr-x	- hadoop supergroup	0	2020-05-07 18:47	/prova08	prova03		
drwxr-xr-x	- hadoop supergroup	0	2020-05-08 10:30	/sortida	prova_scheduler1		
drwxr-xr-x	- hadoop supergroup	0	2020-05-07 10:23	/tmp			
drwxr-xr-x	- hadoop supergroup	0	2020-05-08 17:16	/user			

Showing 1 to 13 of 13 entries

Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

- Abans d'utilitzar un client Hive, haurem d'afegir unes propietats bàsiques en l'arxiu de configuració `hive-site.xml`

[hadoop@node1 conf]\$ gedit hive-site.xml



```

<?xml version="1.0" encoding="UTF-8" standalone="no"?><!--
  Licensed to the Apache Software Foundation (ASF) under one or more
  contributor license agreements. See the NOTICE file distributed with
  this work for additional information regarding copyright ownership.
  The ASF licenses this file to You under the Apache License, Version 2.0
  (the "License"); you may not use this file except in compliance with
  the License. You may obtain a copy of the License at

  http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License.
--><configuration>
<!-- WARNING!!! This file is auto generated for documentation purposes ONLY! -->
<!-- WARNING!!! Any changes you make to this file will be ignored by Hive. -->
<!-- WARNING!!! You must make your changes in hive-site.xml instead. -->
<!-- Hive Execution Parameters -->
<property>
  <name>system:java.io.tmpdir</name>
  <value>/tmp/hive/java</value>
</property>

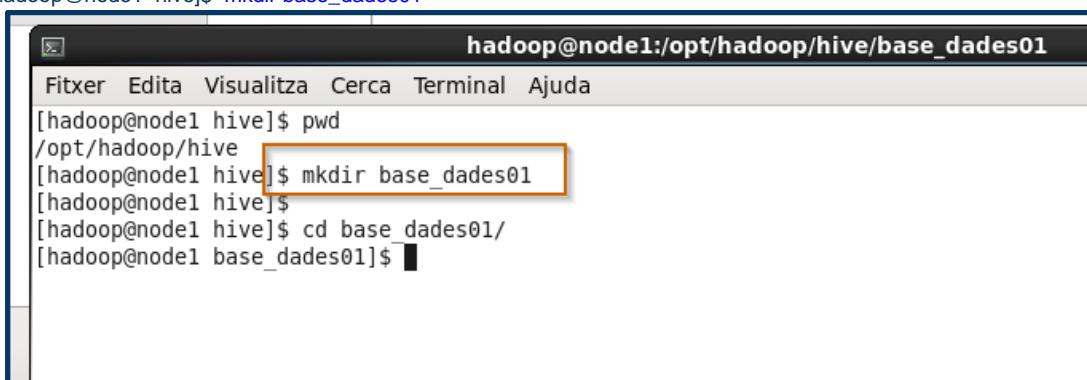
<property>
  <name>system:user.name</name>
  <value>${user.name}</value>
</property>

<property>
  <name>hive.exec.script.wrapper</name>
  <value/>
  <description/>
</property>
<property>
  <name>hive.exec.plan</name>
  <value/>
</property>

```

- Anirem al directori principal de Hive i crearem un directori anomenat `base_dades01` on treballarem

[hadoop@node1 hive]\$ mkdir base_dades01



```

hadoop@node1:/opt/hadoop/hive/base_dades01
Fitxer Edita Visualitza Cerca Terminal Ajuda
[hadoop@node1 hive]$ pwd
/opt/hadoop/hive
[hadoop@node1 hive]$ mkdir base_dades01
[hadoop@node1 hive]$
[hadoop@node1 hive]$ cd base_dades01/
[hadoop@node1 base_dades01]$

```

Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

- ❑ El tipus de Hive que farem servir és “encastat” on per defecte utilitza la BBDD Derby. Hem d'inicialitzar la base de dades en mode local.

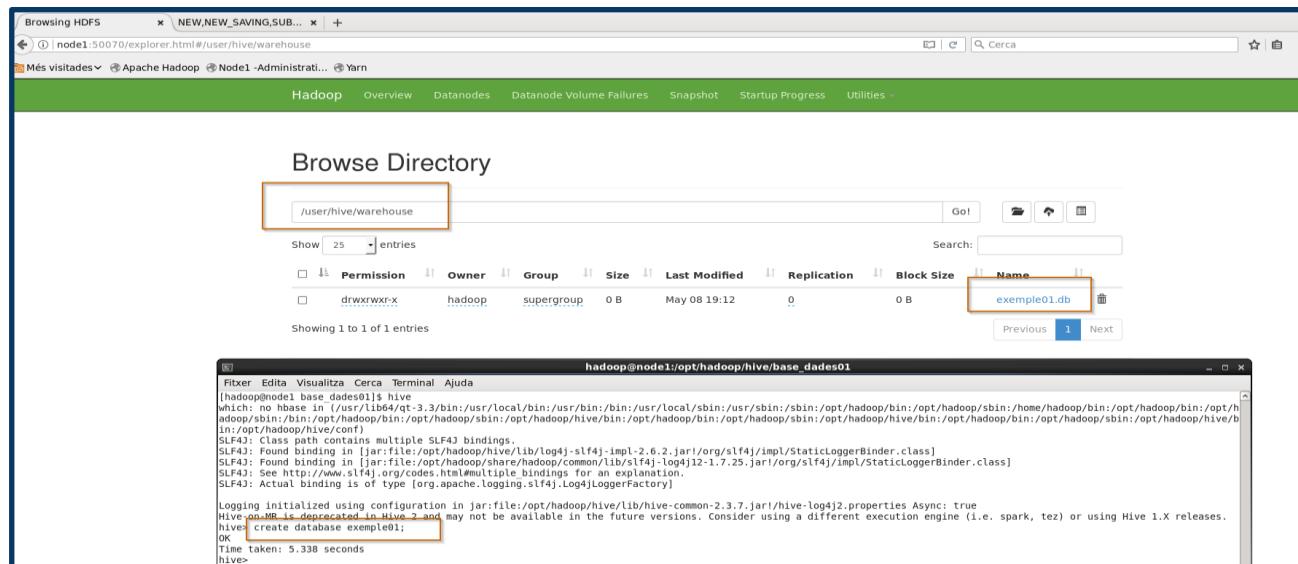
```
[hadoop@node1 ~]$ schematool -dbType derby -initSchema
```

Nota : Posteriorment s'hauria de configurar a l'arxiu hive-site.xml

```
[hadoop@node1 base_dades01]$ pwd  
/opt/hadoop/hive/base_dades01  
[hadoop@node1 base_dades01]$ schematool -dbType derby -initSchema  
SLF4J: Class path contains multiple SLF4J bindings.  
SLF4J: Found binding in [jar:file:/opt/hadoop/hive/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]  
SLF4J: Found binding in [jar:file:/opt/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]  
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.  
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]  
Metastore connection URL:      jdbc:derby:;databaseName=metastore_db;create=true  
Metastore Connection Driver :  org.apache.derby.jdbc.EmbeddedDriver  
Metastore connection User:     APP  
Starting metastore schema initialization to 2.3.0  
Initialization script hive-schema-2.3.0.derby.sql  
Initialization script completed  
schemaTool completed  
[hadoop@node1 base_dades01]$ ls -ls  
total 8  
4 -rw-rw-r-- 1 hadoop hadoop 669 8 mai 19:05 derby.log  
4 drwxrwxr-x. 5 hadoop hadoop 4096 8 mai 19:05 metastore_db  
[hadoop@node1 base_dades01]$
```

- Creem una base de dades anomenada `exemple01` i ho visualitzem per l'interfície web del Namenode

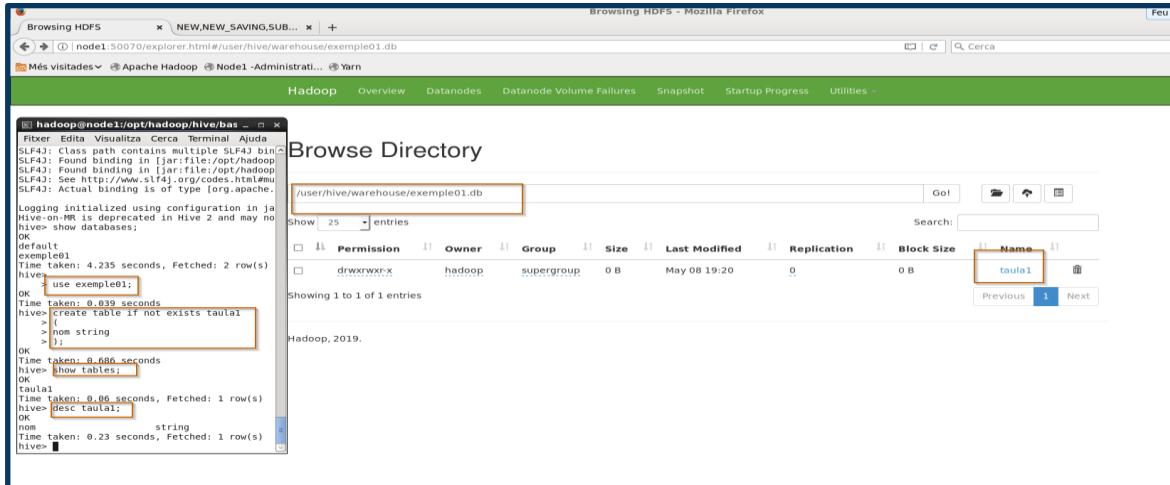
```
[hadoop@node1 ~]$ hive  
hive> create database exemple01  
node1:50070
```



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

- ☐ Farem un exemple senzill, creant una taula i ho visualitzem per l'interfície web

```
[hadoop@node1 hive]$ hive
hive> use exemplo01;
hive> create table if not exists taula1;
hive> desc taula1;
node1:50070
```



The terminal window shows the following Hive session:

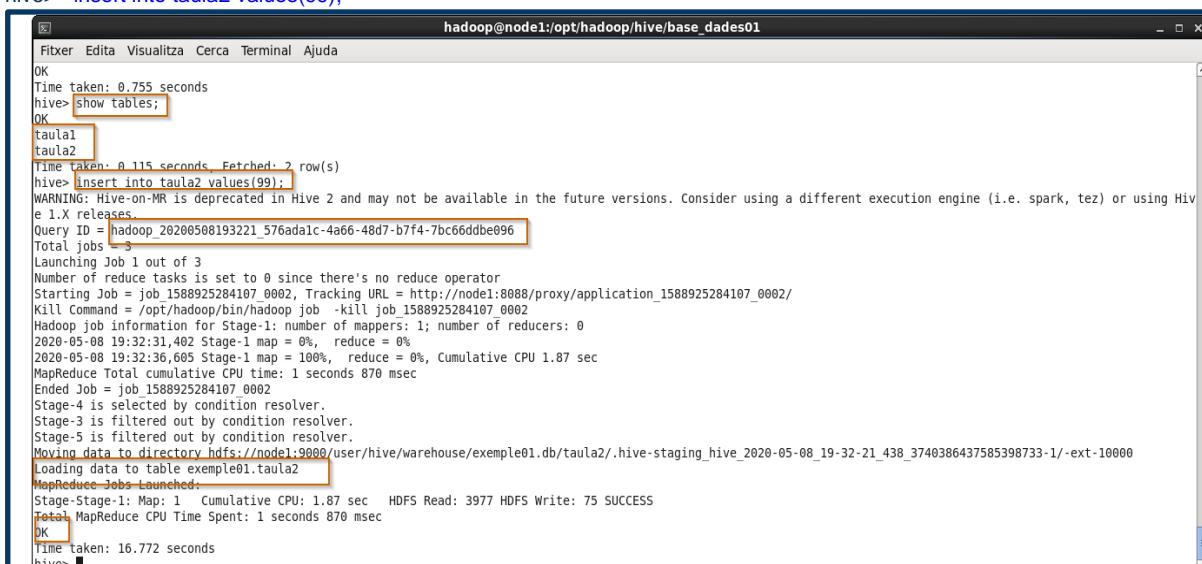
```
[hadoop@node1 ~]$ hadoop@node1:/opt/hadoop/hive/base_dades01
Fitxer Edits Visualitza Cerca Terminal Ajuda
SLF4J: Class path contains multiple SLF4J bindings
SLF4J: Found binding in [jar:/file:/opt/hadoop
SLF4J: Found binding in [jar:/file:/opt/hadoop
SLF4J: Found binding in [jar:/file:/opt/hadoop
SLF4J: Actual binding is of type [org.apache.
Logging initialized using configuration in ja
Hive-on-MR is deprecated in Hive 2 and may no
hive> show databases;
OK
default
example01
Time taken: 4.235 seconds, Fetched: 2 rows(s)
hive> use exemplo01;
OK
Time taken: 0.039 seconds
hive> create table if not exists taula1
> |
> > nom string
> );
OK
Time taken: 0.686 seconds
hive> show tables;
OK
taula1
Time taken: 0.06 seconds, Fetched: 1 row(s)
hive> desc taula1;
OK
nom
string
Time taken: 0.23 seconds, Fetched: 1 row(s)
hive> 
```

The browser window shows the HDFS file system at <http://node1:50070/explorer.html#/user/hive/warehouse/exemplo01.db>. It lists a single file named "taula1" with the following details:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxrwxr-x	hadoop	supergroup	0 B	May 08 19:20	0	0 B	taula1

- ☐ Crearem una altra taula i farem un insert on veurem que llença un job Mapreduce

```
hive> create table if not exists taula2;
hive> show tables;
hive> insert into taula2 values(99);
```



The terminal window shows the following session:

```
[hadoop@node1 ~]$ hadoop@node1:/opt/hadoop/hive/base_dades01
Fitxer Edits Visualitza Cerca Terminal Ajuda
OK
Time taken: 0.755 seconds
hive> show tables;
OK
taula1
taula2
Time taken: 0.115 seconds, Fetched: 2 row(s)
hive> insert into taula2 values(99);
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = hadoop_20200508193221_576adalc-a66-48d7-b7f4-7bc66ddbe096
Total jobs: 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1588925284107_0002, Tracking URL = http://node1:8088/proxy/application_1588925284107_0002/
Kill Command = /opt/hadoop/bin/hadoop job -kill job_1588925284107_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2020-05-08 19:32:31,402 Stage-1 map = 0%, reduce = 0%
2020-05-08 19:32:36,605 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 1.87 sec
MapReduce Total cumulative CPU time: 1 seconds 870 msec
Ended Job = job_1588925284107_0002
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to directory hdfs://node1:9000/user/hive/warehouse/exemplo01.db/taula2/.hive-staging_hive_2020-05-08_19-32-21_438_3740386437585398733-1/-ext-10000
Loading data to table exemplo01.taula2
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Cumulative CPU: 1.87 sec HDFS Read: 3977 HDFS Write: 75 SUCCESS
Total MapReduce CPU Time Spent: 1 seconds 870 msec
OK
Time taken: 16.772 seconds
hive> 
```



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

- Com que el resultat és text pla, desde HDFS ho podem visualitzar.

```
[hadoop@node1 Escriptori]$ hdfs dfs -cat /user/hive/warehouse/exemple01.db/taula2/000000_0
```

Browsing HDFS - Mozilla Firefox

Browsing HDFS

node1:50070/explorer.html#/user/hive/warehouse/exemple01.db/taula2

Més visitades Apache Hadoop Node1-Administraci... Yarn

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities

Browse Directory

	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
	-rwxrwxr-x	hadoop	supergroup	3 B	May 08 19:32	2	128 MB	000000_0

Showing 1 to 1 of 1 entries

```
hadoop@node1:~/Escriptori
Fitxer Edita Visualitza Cerca Terminal Ajuda
Had[hadoop@node1 Escriptori]$ hdfs dfs -cat /user/hive/warehouse/exemple01.db/taula2/000000_0
20/05/08 19:38:22 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
orm, using builtin-java classes where applicable
99
[hadoop@node1 Escriptori]$
```

- Hem de diferenciar entre taules internes i taules externes.

- Les **taules internes** tot el seu cicle de vida (informació que contenen, ho gestiona el mateix Hive. Si crees una taula amb dades, si elimino la taula , des de Hive, elimini els fitxers associats).
- Si la **taula** és **externa**, segurament ja tinc unes dades que s'han carregat d'una alta forma i segurament estan essent utilitzats per altres productes i Hive també pot que utilitzar-les .En aquest cas, no és Hive que controla el cicle de vida de la taula, per tant si esborro la taula, no s'esborren les dades.



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

□ Creació d'una taula interna

- Creació d'un arxiu de text (empleats.txt) pla on hi han els noms i les edats separats per comes
- Creació d'una taula anomenada "empleats" amb les variables "nom" i "edat"

```
[hadoop@node1 hadoop]$ gedit empleats.txt
[hadoop@node1 hadoop]$ hive
hive> use exemple01;
hive> create table empleats
> (
> nom string,
> edat integer,
> row format delimited
> fields terminated by ',';
```

The screenshot shows a terminal window titled 'hadoop@node1:/tmp/hadoop'. It has three tabs: 'hadoop@node1:/opt/hadoop/hive', 'hadoop@node1:/opt/hadoop/hive', and 'hadoop@node1:/tmp/hadoop'. The third tab contains the following command:

```
hive> use exemple01;
OK
Time taken: 4.179 seconds
hive> create table empleats
> (
> nom string,
> edat integer
> )
> row format delimited
> fields terminated by ',';
```

Below the terminal window, there is a file viewer window titled '*empleats.txt (/tmp/hadoop) - gedit'. It shows the contents of the 'empleats.txt' file:

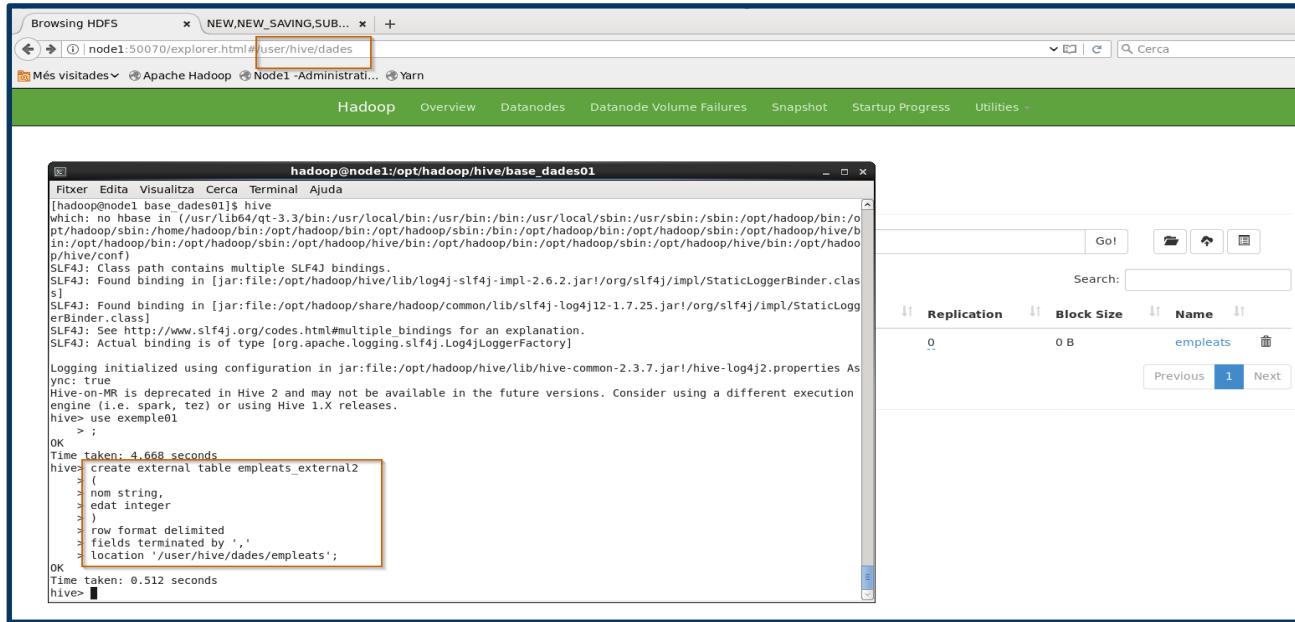
```
Rosa,50
Pere,45
Raul,47
Joan,40
Marta,36
```

- Seguidament carreguem la taula empleats i l'associem a l'arxiu empleats.txt ubicat a /tmp/hadoop . Posteriorment ho visualitzem en la interfície web del Namenode.
- També visualitzem el resultat en HDFS ja que és un text pla.

NOTA: Com que és una taula interna, si fes un "drop" no només eliminaria la taula sinó els fitxers HDFS associats a ella.

```
hive> load data local inpath '/tmp/hadoop/emplets.txt' into table empleats;
hive> select * from empleats;
[hadoop@node1 hadoop]$ hdfs dfs -cat /user/hive/warehouse/exemple01.db/empleats/empleats.txt
```

Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

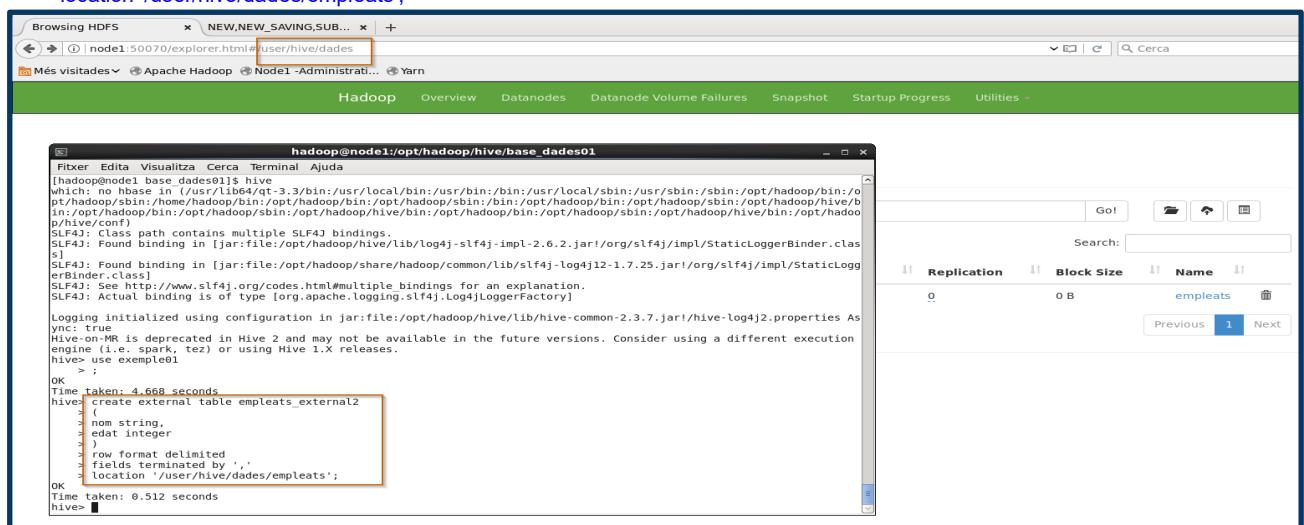


```
[hadoop@node1 hadoop]$ hive
hive> use exemple01;
hive> create external table empleats_external2
> (
> nom string,
> edat integer,
> )
> row format delimited
> fields terminated by ','
> location '/user/hive/dades/empleats';
OK
Time taken: 0.512 seconds
hive>
```

□ Creació d'una taula externa

- Creació d'una taula externa ubicada a [/user/hive/dades/empleats](#)
- Farem servir el mateix arxiu “emplats.txt”

```
[hadoop@node1 hadoop]$ hive
hive> use exemple01;
hive> create external table empleats_external2
> (
> nom string,
> edat integer,
> )
> row format delimited
> fields terminated by ','
> location '/user/hive/dades/empleats';
```

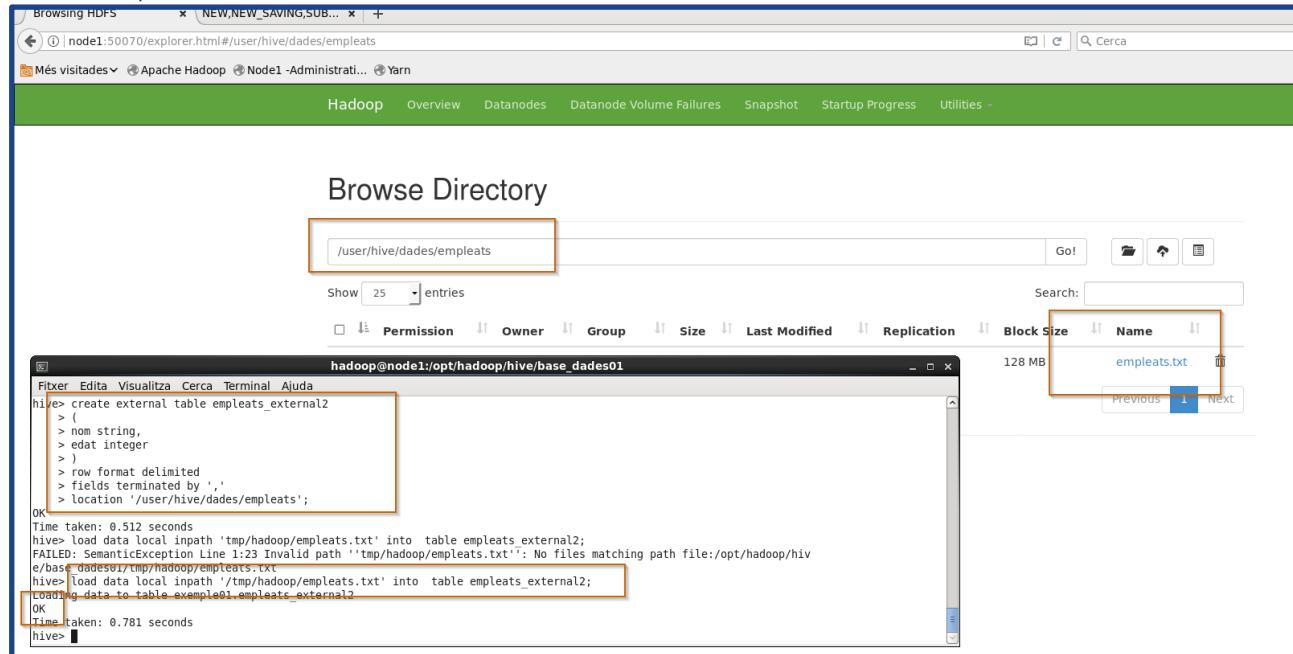


```
[hadoop@node1 hadoop]$ hive
hive> use exemple01;
hive> create external table empleats_external2
> (
> nom string,
> edat integer,
> )
> row format delimited
> fields terminated by ','
> location '/user/hive/dades/empleats';
OK
Time taken: 0.512 seconds
hive>
```

[hadoop@node1 hadoop]\$ load data local inpath '/tmp/hadoop/empleats.txt' into table empleats_external2;

Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

NOTA: En aquest cas si eliminem la taula, el directori i fitxers HDFS no s'esborren



The screenshot shows a terminal window and a browser window side-by-side.

Terminal Window (Left):

```

hive> create external table empleats_external
> (
>     nom string,
>     edat integer
> )
> row format delimited
> fields terminated by ','
> location '/user/hive/dades/empleats';
OK
Time taken: 0.512 seconds
hive> load data local inputpath '/tmp/hadoop/empleats.txt' into table empleats_external2;
FAILED: SemanticException Line 1:23 Invalid path ''/tmp/hadoop/empleats.txt'': No files matching path file:/opt/hadoop/hive/base_dades01/tmp/hadoop/empleats.txt
hive> load data local inputpath '/tmp/hadoop/empleats.txt' into table empleats_external2;
Loading data to table empleats01.empleats_external2
OK
Time taken: 0.781 seconds
hive>

```

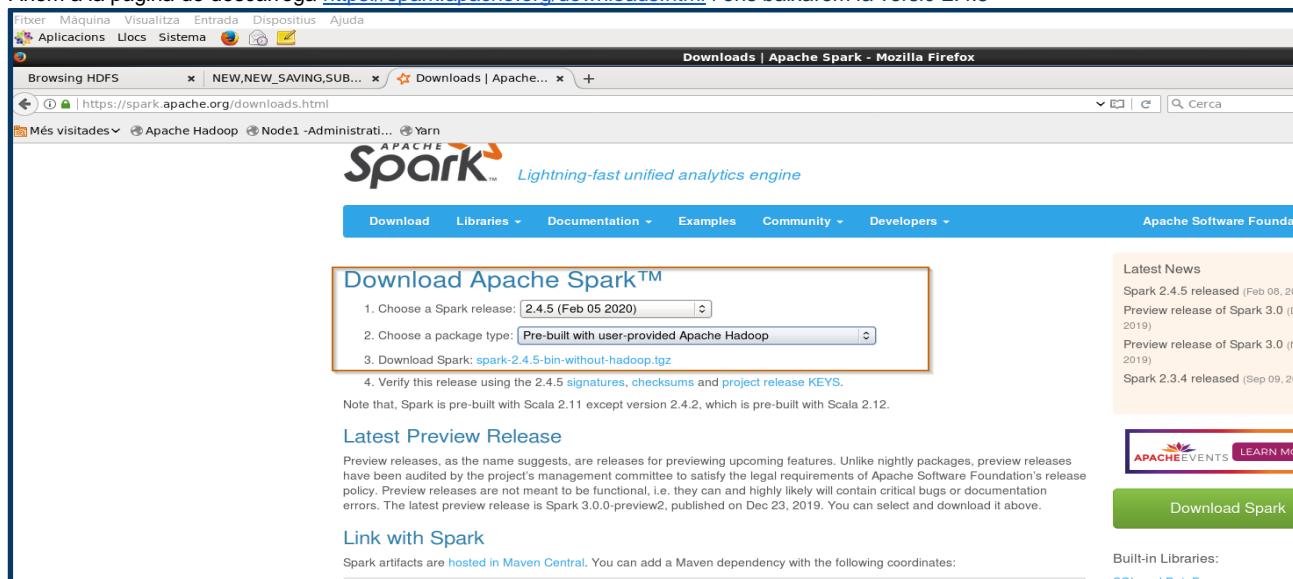
Browser Window (Right):

The browser shows the HDFS file system at `/user/hive/dades/empleats`. It lists a single file named `empleats.txt` with a block size of 128 MB. The URL in the address bar is `http://node1:50070/explorer.html#/user/hive/dades/empleats`.

3.2.6.2.2. Instal.lació i configuració de Spark

- En les opcions de descàrrega, com que ja tenim instal·lat Hadoop, seleccionarem el binari sense Hadoop per poder utilitzar-lo en el nostre clúster.
- Requisits mínims: Java 8 i Python 2.7

Anem a la pàgina de descàrrega <https://spark.apache.org/downloads.html> i ens baixarem la versió 2.4.5



The screenshot shows the Apache Spark download page on a Mozilla Firefox browser.

Page Headers:

- Fitxer
- Máquina
- Visualitzar
- Entrada
- Dispositius
- Ajuda
- Aplicacions
- Llocs
- Sistema

Page Content:

Download Apache Spark™

- Choose a Spark release: 2.4.5 (Feb 05 2020)
- Choose a package type: Pre-built with user-provided Apache Hadoop
- Download Spark: spark-2.4.5-bin-without-hadoop.tgz
- Verify this release using the 2.4.5 signatures, checksums and project release KEYS.

Note that, Spark is pre-built with Scala 2.11 except version 2.4.2, which is pre-built with Scala 2.12.

Latest Preview Release

Preview releases, as the name suggests, are releases for previewing upcoming features. Unlike nightly packages, preview releases have been audited by the project's management committee to satisfy the legal requirements of Apache Software Foundation's release policy. Preview releases are not meant to be functional, i.e. they can and highly likely will contain critical bugs or documentation errors. The latest preview release is Spark 3.0.0-preview2, published on Dec 23, 2019. You can select and download it above.

Link with Spark

Spark artifacts are hosted in [Maven Central](#). You can add a Maven dependency with the following coordinates:

Apache Software Foundation

Latest News

- Spark 2.4.5 released (Feb 08, 2020)
- Preview release of Spark 3.0 (Dec 23, 2019)
- Preview release of Spark 3.0 (Nov 26, 2019)
- Spark 2.3.4 released (Sep 09, 2019)

APACHE EVENTS [LEARN MORE](#)

Download Spark

Built-in Libraries:

- SQL and DataFrames



Nom i Cognoms

Arnaud Subirós Puigarnau

Data

02-06-2020

The screenshot shows a Firefox browser window with the Apache Download Mirrors page. A download dialog box is open, prompting the user to choose how to handle the file 'spark-2.4.5-bin-without-hadoop.tgz'. The 'Save the file' option is selected. The file is described as an arxiu gzip (160 MB) from https://ftp.cixug.es. The download path is https://downloads.apache.org/spark/spark-2.4.5/spark-2.4.5-bin-without-hadoop.tgz.

The screenshot shows a terminal window titled 'hadoop@node1:~/Baixades'. The command 'ls' is run, showing several files including 'spark-2.4.5-bin-without-hadoop.tgz', which is highlighted with a red box.

- Descomprimim l'arxiu a /opt/hadoop i posteriorment li canviarem el nom, ja que el nom del directori és molt llarg

```
[hadoop@node1 hadoop]$ tar xvf /home/hadoop/Baixades/spark-2.4.5-bin-without-hadoop.tgz  
[hadoop@node1 hadoop]$ mv spark-2.4.5-bin-without-hadoop/ spark
```

The screenshot shows a terminal window titled 'hadoop@node1:/opt/hadoop'. The command 'tar xvf /home/hadoop/Baixades/spark-2.4.5-bin-without-hadoop.tgz' is run, with the command line highlighted with a red box.

The screenshot shows a terminal window titled 'hadoop@node1:/opt/hadoop'. The command 'mv spark-2.4.5-bin-without-hadoop/ spark' is run, with the command line highlighted with a red box. This renames the directory 'spark-2.4.5-bin-without-hadoop' to 'spark'.

Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

- Haurem d'accedir a l'arxiu `~/.bashrc` per afegir les variables d'entorn de Spark

```
[hadoop@node1 spark]$ gedit ~/.bashrc
```

The screenshot shows a terminal window titled "hadoop" with the command `[hadoop@node1:~/opt/hadoop/spark] $ gedit ~/.bashrc`. The file content is a script named ".bashrc" containing environment variable definitions for Java, Hive, Hadoop, and Spark. A red box highlights the last two lines of the script.

```
# .bashrc

# Source global definitions
if [ -f /etc/bashrc ]; then
    . /etc/bashrc
fi

# User specific aliases and functions

### CREAR una variable i exportar-la globalment

export JAVA_HOME=/usr/java/jdk1.8.0_251-amd64
export HIVE_HOME=/opt/hadoop/hive
export HADOOP_HOME=/opt/hadoop

export PATH=$PATH:$HADOOP_HOME/bin:/opt/hadoop/sbin:$HIVE_HOME/bin:$HIVE_HOME/conf:/opt/hadoop/spark/sbin:/opt/hadoop/spark/bin
export SPARK_DIST_CLASSPATH=$(hadoop classpath)
```

Exemple amb Spark Shell

- Carreguem les variables d'entorn i obrim un terminal **Spark-Shell** on crea un context Spark i després crida l'entorn de línia d'ordres de Scala

```
[hadoop@node1 bin ]$ source ~/.bashrc  
[hadoop@node1 bin ]$ spark-shell
```

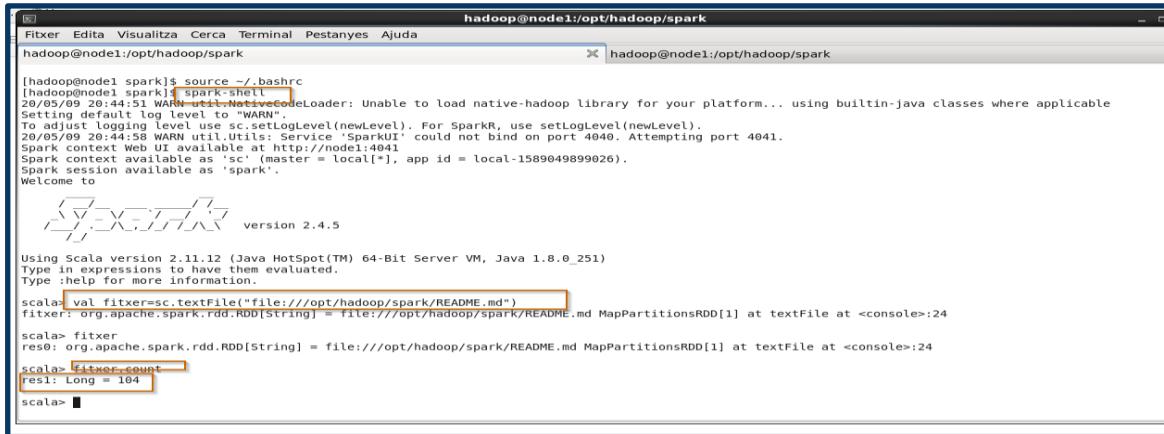
```
[hadoop@node1 ~]$ spark-shell
[Truncated]

```

- ❑ Farem un exemple senzill, llegirem un fitxer de text `readme` que ja ve incorporat amb el mateix Spark i contarem el número de línies
 - crearem una variable Scala anomenada `fitxer` (val `fitxer`)
 - A través del context(`sc`) afegirem un comando “`textFile`” per accedir a un fitxer de text
 - Com que ens interessa un fitxer local(per defecte ho busca a HDFS) hem d’indicar “`file://+path complet`”
 - Associem el fixer a la variable `fitxer`

```
[hadoop@node1 bin ]$ spark-shell  
scala> val fitxer=sc.textFile("file:///opt/hadoop/spark/README.md")  
scala> fitxer.count
```

Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020



```

hadoop@node1:/opt/hadoop/spark
Fitxer Edita Visualitzar Cerca Terminal Pestanyes Ajuda
hadoop@node1:/opt/hadoop/spark
[hadoop@node1 spark]$ source ~/.bashrc
[hadoop@node1 spark]$ spark-shell
20/05/09 20:44:51 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
20/05/09 20:44:51 WARN util.NativeCodeLoader: Native library version 2.4.5
Spark context Web UI available at http://node1:4040.
Spark context available as 'sc' (master = local[*], app id = local-1589049899026).
Spark session available as 'spark'.
Welcome to

  / \   / \   / \   / \
 / \ / \ / \ / \ / \ / \
version 2.4.5

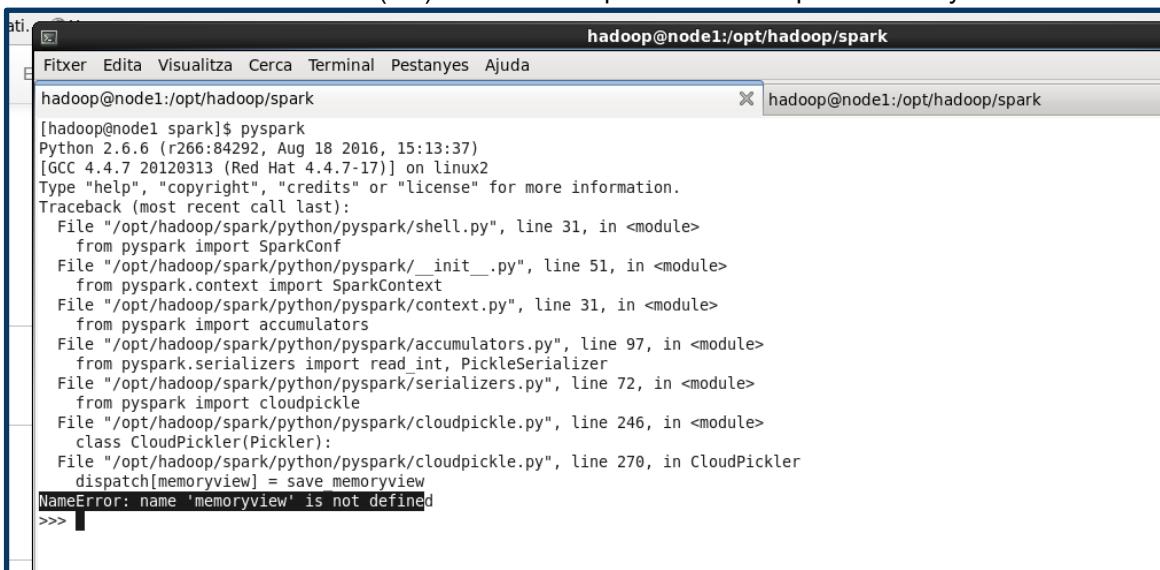
Using Scala version 2.11.12 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_251)
Type in expressions to have them evaluated.
Type :help for more information.

scala> val fixter=sc.textFile("file:///opt/hadoop/spark/README.md")
fixter: org.apache.spark.rdd.RDD[String] = file:///opt/hadoop/spark/README.md MapPartitionsRDD[1] at textFile at <console>:24
scala> fixter
res0: org.apache.spark.rdd.RDD[String] = file:///opt/hadoop/spark/README.md MapPartitionsRDD[1] at textFile at <console>:24
scala> fixter.count
res1: Long = 164
scala>

```

Exemple amb Pyspark

- Volem fer el mateix però amb Pyspark, en lloc de Scala utilitzant Python. Però em dóna error, ja que tinc una versió inferior a 2.7 (2.6) i es la versió que utilitzà els repositoris de "yum"



```

hadoop@node1:/opt/hadoop/spark
Fitxer Edita Visualitzar Cerca Terminal Pestanyes Ajuda
hadoop@node1:/opt/hadoop/spark
[hadoop@node1 spark]$ pyspark
Python 2.6.6 (r266:84292, Aug 18 2016, 15:13:37)
[GCC 4.4.7 20120313 (Red Hat 4.4.7-17)] on linux2
Type "help", "copyright", "credits" or "license" for more information.
Traceback (most recent call last):
  File "/opt/hadoop/spark/python/pyspark/shell.py", line 31, in <module>
    from pyspark import SparkConf
  File "/opt/hadoop/spark/python/pyspark/_init_.py", line 51, in <module>
    from pyspark.context import SparkContext
  File "/opt/hadoop/spark/python/pyspark/context.py", line 31, in <module>
    from pyspark import accumulators
  File "/opt/hadoop/spark/python/pyspark/accumulators.py", line 97, in <module>
    from pyspark.serializers import read_int, PickleSerializer
  File "/opt/hadoop/spark/python/pyspark/serializers.py", line 72, in <module>
    from pyspark import cloudpickle
  File "/opt/hadoop/spark/python/pyspark/cloudpickle.py", line 246, in <module>
    class CloudPickler(Pickler):
  File "/opt/hadoop/spark/python/pyspark/cloudpickle.py", line 270, in CloudPickler
    dispatch[memoryview] = save_memoryview
NameError: name 'memoryview' is not defined
>>>

```

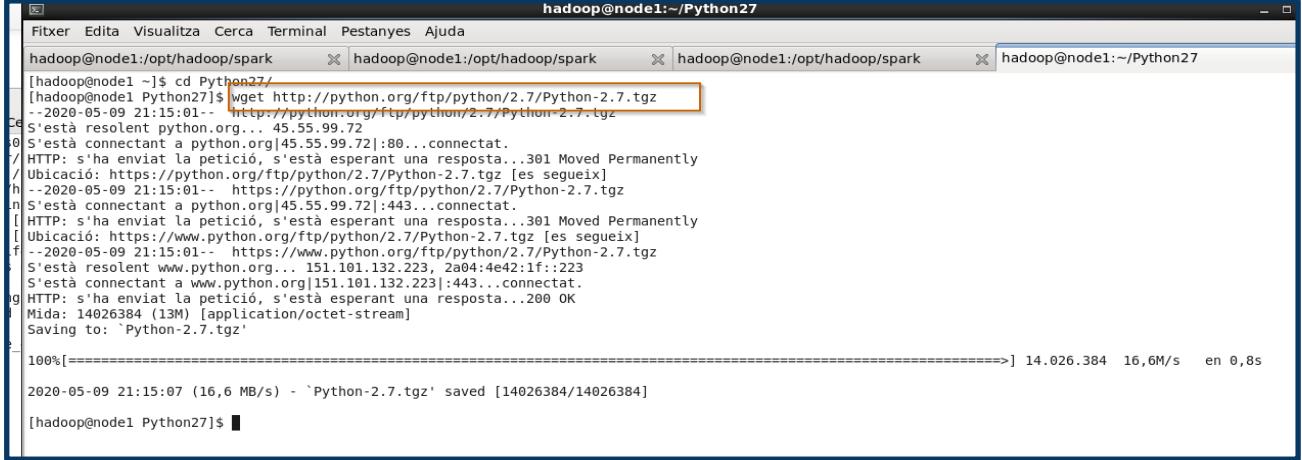
- Instal·lem Python 2.7 i els paquets necessaris. En acabar ens assegurem la versió de Python que estem utilitzant

```

[hadoop@node1 ~]$ yum -y groupinstall "Development Tools"
[hadoop@node1 ~]$ yum -y install openssl-devel* ncurses-devel* zlib*.x86_64
[hadoop@node1 ~]$ yum -y install bzip2 bzip2-devel bzip2-libs
[hadoop@node1 ~]$ mkdir Python27
[hadoop@node1 ~]$ cd Python27/
[hadoop@node1 Python27]$ wget http://python.org/ftp/python/2.7/Python-2.7.tgz
[hadoop@node1 Python27]$ tar xvf Python-2.7.tgz
[root @node1 ~]$ which python

```

Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020



```

hadoop@node1:~/Python27/
[hadoop@node1 Python27]$ cd Python27/
[hadoop@node1 Python27]$ wget http://python.org/ftp/python/2.7/Python-2.7.tgz
--2020-05-09 21:15:01--  http://python.org/ftp/python/2.7/Python-2.7.tgz
S'està resolent python.org... 45.55.99.72
S'està connectant a python.org[45.55.99.72]:80...connectat.
HTTP: s'ha enviat la petició, s'està esperant una resposta...301 Moved Permanently
Ubicació: https://www.python.org/ftp/python/2.7/Python-2.7.tgz [es segueix]
--2020-05-09 21:15:01--  https://www.python.org/ftp/python/2.7/Python-2.7.tgz
S'està resolent www.python.org... 151.101.132.223, 2a04:4e42:1f::223
S'està connectant a www.python.org[151.101.132.223]:443...connectat.
HTTP: s'ha enviat la petició, s'està esperant una resposta...200 OK
Mida: 14026384 (13M) [application/octet-stream]
Saving to: 'Python-2.7.tgz'

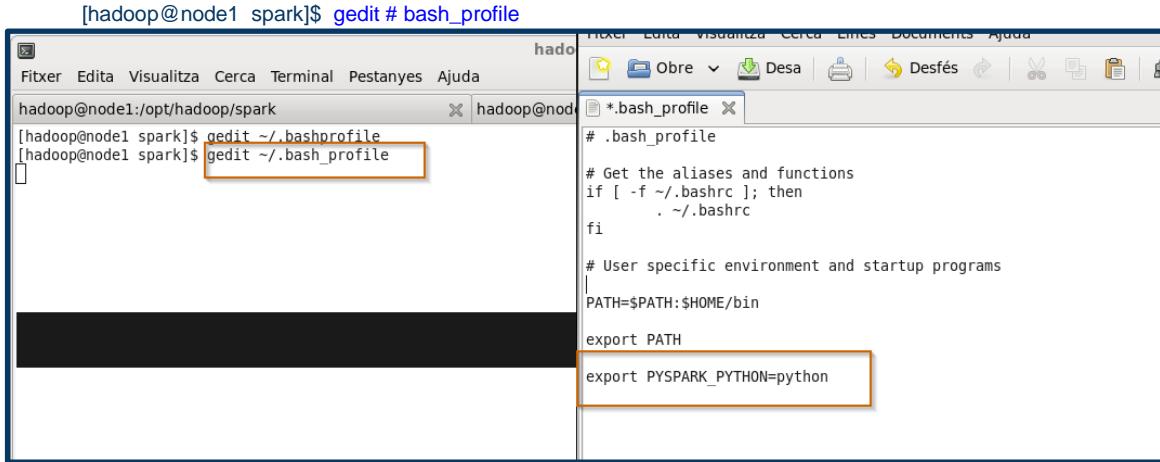
100%[=====] 14.026.384 16,6M/s   en 0,8s

2020-05-09 21:15:07 (16,6 MB/s) - `Python-2.7.tgz' saved [14026384/14026384]

[hadoop@node1 Python27]$ 

```

- Un cop instal.lat afegim una nova variable d'entorn però a [~/.bash_profile](#)



```

hadoop@node1:~/spark$ gedit # bash_profile
hadoop@node1:~/spark$ gedit ~./bash_profile
# .bash_profile

# Get the aliases and functions
if [ -f ~/.bashrc ]; then
    . ~/.bashrc
fi

# User specific environment and startup programs

PATH=$PATH:$HOME/bin

export PATH

export PYSPARK_PYTHON=python

```



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

☐ Ara tornem a executar Pyspark

```
[hadoop@node1:~/opt/hadoop/spark]$ pyspark
[hadoop@node1 spark]$ pyspark
Python 2.6.6 (r266:84792, Aug 18 2010, 15:13:37)
[GCC 4.4.7 20120313 (Red Hat 4.4.7-17)] on linux2
Type "help", "copyright", "credits" or "license" for more information.
Traceback (most recent call last):
  File "/opt/hadoop/python/pyspark/shell.py", line 31, in <module>
    from pyspark import SparkConf
  File "/opt/hadoop/spark/python/pyspark/_init_.py", line 51, in <module>
    from pyspark.context import SparkContext
  File "/opt/hadoop/spark/python/pyspark/context.py", line 31, in <module>
    from pyspark.context import SparkContext
  File "/opt/hadoop/spark/python/pyspark/accumulators.py", line 97, in <module>
    from pyspark.serializers import ReadIntPickleSerializer
  File "/opt/hadoop/spark/python/pyspark/serializers.py", line 72, in <module>
    from pyspark import CloudPickle
  File "/opt/hadoop/spark/python/pyspark/cloudpickle.py", line 246, in <module>
    class CloudPickler(Pickler):
  File "/opt/hadoop/spark/python/pyspark/cloudpickle.py", line 270, in CloudPickler
    dispatch(memoryview).save_memoryview
NameError: name 'memoryview' is not defined
>>>
>>> quit
Use quit() or Ctrl-D (i.e. EOF) to exit
>>> exit
Use exit() or Ctrl-D (i.e. EOF) to exit
>>>
[hadoop@node1 spark]$ pyspark
Python 2.7.0 (r27:82500, May  9 2020, 21:17:37)
[GCC 4.4.7 20120313 (Red Hat 4.4.7-23)] on linux2
Type "help", "copyright", "credits" or "license" for more information.
20/05/09 21:21:33 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
20/05/09 21:21:33 WARN util.Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
Welcome to

    / \ \
   /   \
  /     \
 /       \
/         \
 \       /
  \     /
   \   /
    \ /
     \
      version 2.4.5

Using Python version 2.7.0 (r27:82500, May  9 2020 21:17:37)
SparkSession available as 'spark'.
>>> exit
>>> exit
Use exit() or Ctrl-D (i.e. EOF) to exit
>>>
[hadoop@node1 spark]$
```

☐ Farem un exemple de prova (el mateix que hem fet anteriorment amb Scala)

```
[hadoop@node1 bin]$ pyspark
>>> fitxer=spark.read.text("file:///opt/hadoop/spark/REARME.md")
>>> fitxer.count()
```

```
[hadoop@node1:~/opt/hadoop/spark]$ pyspark
[hadoop@node1 spark]$ pyspark
Python 2.7 (r27:82500, May  9 2020, 21:17:37)
[GCC 4.4.7 20120313 (Red Hat 4.4.7-23)] on linux2
Type "help", "copyright", "credits" or "license" for more information.
20/05/09 21:30:23 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
20/05/09 21:30:24 WARN util.Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
Welcome to

    / \ \
   /   \
  /     \
 /       \
/         \
 \       /
  \     /
   \   /
    \ /
     \
      version 2.4.5

Using Python version 2.7.0 (r27:82500, May  9 2020 21:17:37)
SparkSession available as 'spark'.
>>> fitxer=spark.read.text("file:///opt/hadoop/spark/REARME.md")
>>> fitxer
>>> DataFrame[Value: string]
>>> fitxer.count()
104
>>>
```



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

Exemple amb Spark Shell interactuant amb HDFS

- Per llençar comandes de Spark contra el clúster haurem d'utilitzar la comanda **spark-submit**
- A continuació mostrem els arguments utilitzats:
 - **-class** : utilitzarem un exemple que serveix per calcular el numero Pi
 - **-master yarn** : Estem dient que executarem aquest comando spark contra el nostre clúster Hadoop (de tipus yarn)
 - **--deploy-mode cluster** : S'ha d'especificar ja que sino treballaria en mode local
 - **--name** : Per posar un nom a l'aplicació
 - després haurem d'especificar el path del fitxer + 1 un número (com més gran sigui, més operacions farà i més precís serà)
- Com que ens dóna error, ja que no troba el context, tornem a carregar les variables d'entorn. I executarem **jps** en tots els nodes per observar els processos que 'estan executant.

```
[hadoop@node1 spark]$: source ~/.bashrc
[hadoop@node1 spark]$: spark-submit --class org.apache.spark.examples.SparkPi --master yarn --deploy-mode cluster --name "aplicacio_Pi01" /opt/hadoop/spark/examples/jars/spark-examples_2.11-2.4.5.jar 10
```

```
[hadoop@node1 spark]$: jps
[hadoop@node2 ~]$: jps
[hadoop@node3 ~]$: jps
```

```
hadoop@node3:~$ jps
12634 Jps
8270 NodeManager
3326 DataNode
[hadoop@node3 ~]$ 

hadoop@node2:~$ jps
13377 ApplicationMaster
13414 Jps
8277 NodeManager
3342 DataNode
[hadoop@node2 ~]$ 

hadoop@node1:~$ jps
[hadoop@node1 ~]$ jps
13377 ApplicationMaster
13414 Jps
8277 NodeManager
3342 DataNode
[hadoop@node1 ~]$ 

hadoop@node1:~/opt/hadoop/spark
[hadoop@node1:~/opt/hadoop/spark]$ spark-submit --class org.apache.spark.examples.SparkPi --master yarn --deploy-mode cluster --name "aplicacio_Pi01" /opt/hadoop/spark/examples/jars/spark-examples_2.11-2.4.5.jar 10
Exception in thread "main" org.apache.spark.SparkException: When running with master [yarn] either HADOOP_CONF_DIR or YARN_CONF_DIR must be set in the environment.

at org.apache.spark.deploy.SparkSubmit$Arguments.error(SparkSubmit$Arguments.scala:65)
at org.apache.spark.deploy.SparkSubmit$Arguments.validateSubmitArguments(SparkSubmit$Arguments.scala:290)
at org.apache.spark.deploy.SparkSubmit$Arguments.<init>(SparkSubmit$Arguments.scala:128)
at org.apache.spark.deploy.SparkSubmit$.main(SparkSubmit$.scala:907)
at org.apache.spark.deploy.SparkSubmit$$anon$2.parseArguments(SparkSubmit$.scala:907)
at org.apache.spark.deploy.SparkSubmit$.doSubmit$(SparkSubmit$.scala:81)
at org.apache.spark.deploy.SparkSubmit$.doSubmit(SparkSubmit$.scala:920)
at org.apache.spark.deploy.SparkSubmit$.main(SparkSubmit$.scala:929)
at org.apache.spark.deploy.SparkSubmit$.main$(SparkSubmit$.scala)

[hadoop@node1:~/opt/hadoop/spark]$ source ~/.bashrc
[hadoop@node1:~/opt/hadoop/spark]$ spark-submit --class org.apache.spark.examples.SparkPi --master yarn --deploy-mode cluster --name "aplicacio_Pi01" /opt/hadoop/spark/examples/jars/spark-examples_2.11-2.4.5.jar 10
20/05/10 11:22:30 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
20/05/10 11:22:30 INFO client.RMProxy: Connecting to ResourceManager at node1/192.168.0.101:8082
20/05/10 11:22:31 INFO yarn.Client: Requesting a new application from cluster with 2 NodeManagers
20/05/10 11:22:31 INFO resource.ResourceUtils: Adding resource type - name = memory_mb, units = Mi, type = COUNTABLE
20/05/10 11:22:31 INFO resource.ResourceUtils: Adding resource type - name = vcores, units = , type = COUNTABLE
20/05/10 11:22:31 INFO yarn.Client: Verifying our application has not requested more than the maximum memory capability of the cluster (8192 MB per container)
20/05/10 11:22:31 INFO yarn.Client: Will allocate AM container, with 1408 MB memory including 384 MB overhead
20/05/10 11:22:31 INFO yarn.Client: Setting up container launch context for our AM
20/05/10 11:22:31 INFO yarn.Client: Setting up the launch environment for our AM container
20/05/10 11:22:31 INFO yarn.Client: Preparing resources for our AM container
20/05/10 11:22:31 INFO yarn.yarn: spark.archive is set to fall back to uploading libraries under SPARK HOME.
20/05/10 11:22:30 INFO yarn.Client: Uploading resource file:/tmp/spark-1fb0b4c1-46d2-4cfc-9c4c-3e7ca39194/_spark_libs_004529931872138312.zip
20/05/10 11:22:38 INFO yarn.Client: Uploading resource file:/opt/hadoop/spark/examples/jars/spark-examples_2.11-2.4.5.jar -> hdfs://node1:9000/user/hadoop/Staging/application_1588925284107_0003/spark-examples_2.11-2.4.5.jar
20/05/10 11:22:38 INFO hdfs.DataStreamer: Exception in createBlockOutputStream
java.io.EOFException
    at org.apache.hadoop.hdfs.protocolPB.HdfsProtocolClient.writePrefix(HdfsProtocolClient.java:456)
    at org.apache.hadoop.hdfs.DataStreamer.createBlockOutputStream(DataStreamer.java:1734)
    at org.apache.hadoop.hdfs.DataStreamer.nextBlockOutputStream(DataStreamer.java:1655)
    at org.apache.hadoop.hdfs.DataStreamer.run(DataStreamer.java:710)
20/05/10 11:22:38 WARN hdfs.DataStreamer: Abandoning BP-454387204-192.168.0.101-1588789389960:blk_1073741951_1127
20/05/10 11:22:38 WARN hdfs.DataStreamer: Excluding datanode from storage[192.168.0.103:58010,05-bf4953e9-f6b1-434b-a1af-d45b67c1a193,DISK]
```

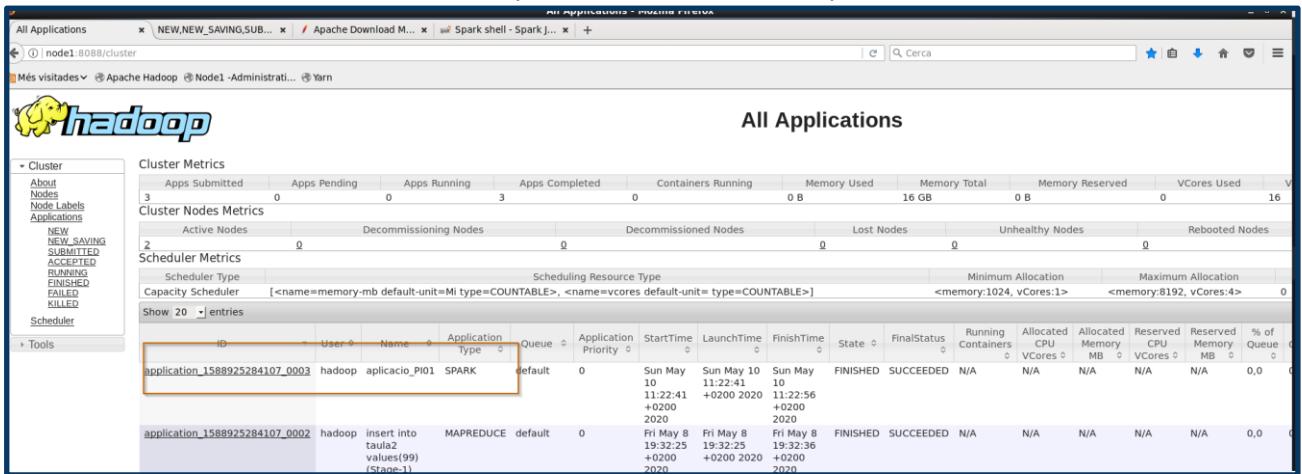
Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

```

hadoop@node1:/opt/hadoop/spark$ fitxer Edita Visualitza Cerca Terminal Pestanyes Ajuda
hadoop@node1:/opt/hadoop/spark$ hadoop@node1:/opt/hadoop/spark$ hadoop@node1:/opt/hadoop/spark$ hadoop@node1:/opt/hadoop/spark$ hadoop@node1:~/
diagnostics: AM container is launched, waiting for AM container to Register with RM
ApplicationMaster host: N/A
ApplicationMaster RPC port: -1
queue: default
start time: 1589102561207
final status: UNDEFINED
tracking URL: http://node1:8088/proxy/application_1588925284107_0003/
user: hadoop
20/05/10 11:22:43 INFO yarn.Client: Application report for application_1588925284107_0003 (state: ACCEPTED)
20/05/10 11:22:44 INFO yarn.Client: Application report for application_1588925284107_0003 (state: ACCEPTED)
20/05/10 11:22:45 INFO yarn.Client: Application report for application_1588925284107_0003 (state: ACCEPTED)
20/05/10 11:22:46 INFO yarn.Client: Application report for application_1588925284107_0003 (state: ACCEPTED)
20/05/10 11:22:47 INFO yarn.Client: Application report for application_1588925284107_0003 (state: ACCEPTED)
20/05/10 11:22:47 INFO yarn.Client: Application report for application_1588925284107_0003 (state: RUNNING)
20/05/10 11:22:47 INFO yarn.Client:
client token: N/A
diagnostics: N/A
ApplicationMaster host: node2
ApplicationMaster RPC port: 38514
queue: default
start time: 1589102561207
final status: UNDEFINED
tracking URL: http://node1:8088/proxy/application_1588925284107_0003/
user: hadoop
20/05/10 11:22:48 INFO yarn.Client: Application report for application_1588925284107_0003 (state: RUNNING)
20/05/10 11:22:49 INFO yarn.Client: Application report for application_1588925284107_0003 (state: RUNNING)
20/05/10 11:22:50 INFO yarn.Client: Application report for application_1588925284107_0003 (state: RUNNING)
20/05/10 11:22:51 INFO yarn.Client: Application report for application_1588925284107_0003 (state: RUNNING)
20/05/10 11:22:52 INFO yarn.Client: Application report for application_1588925284107_0003 (state: RUNNING)
20/05/10 11:22:53 INFO yarn.Client: Application report for application_1588925284107_0003 (state: RUNNING)
20/05/10 11:22:54 INFO yarn.Client: Application report for application_1588925284107_0003 (state: RUNNING)
20/05/10 11:22:55 INFO yarn.Client: Application report for application_1588925284107_0003 (state: RUNNING)
20/05/10 11:22:56 INFO yarn.Client: Application report for application_1588925284107_0003 (state: RUNNING)
20/05/10 11:22:57 INFO yarn.Client: Application report for application_1588925284107_0003 (state: FINISHED)
20/05/10 11:22:57 INFO yarn.Client:
client token: N/A
diagnostics: N/A
ApplicationMaster host: node2
ApplicationMaster RPC port: 38514
queue: default
start time: 1589102561207
final status: SUCCEEDED
tracking URL: http://node1:8088/proxy/application_1588925284107_0003/
user: hadoop
20/05/10 11:22:57 INFO util.ShutdownHookManager: Shutdown hook called
20/05/10 11:22:57 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-1fb9bbfd-c186-40d2-84ca-35c7e3339194
20/05/10 11:22:57 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-617d75a2-f848-4c2d-9a54-9a295fdb637e
[hadoop@node1 spark]$ 

```

□ Accedim a l'interfície web del Yarn per veure informació de l'aplicació



The screenshot shows the Hadoop YARN Web UI with the title "All Applications". The left sidebar has sections for Cluster (About, Nodes, Node Labels, Applications: NEW, NEW_SAVING, SUBMITTED, ACCEPTED, RUNNING, FINISHED, FAILED, KILLED, Scheduler), Tools (Hadoop, Spark shell, Yarn), and a search bar. The main content area displays "Cluster Metrics" and "Cluster Nodes Metrics" tables. Below these are "Scheduler Metrics" and "Capacity Scheduler" settings. The "Applications" section lists two entries:

ID	User	Name	Application Type	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU	Allocated Memory	Reserved CPU	Reserved Memory	% of Queue
application_1588925284107_0003	hadoop	aplicacio_PI01	SPARK	Default	0	Sun May 10 11:22:41 +0200 2020	Sun May 10 11:22:56 +0200 2020	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0,0	
application_1588925284107_0002	hadoop	insert into taula2 values(99) (Stage=1)	MAPREDUCE	default	0	Fri May 8 19:32:25 +0200 2020	Fri May 8 19:32:36 +0200 2020	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0,0	

Nom i Cognoms

Arnau Subirós Puigarnau

Data

02-06-2020

Application application_1588925284107_0003 - Mozilla Firefox

User: hadoop
 Name: aplicacio_Pi01
 Application Type: SPARK

Application tags:
 Application Priority: 0 (Higher integer value indicates higher priority)
 YarnApplicationState: FINISHED
 Queue: default

FinalStatus Reported by AM: SUCCEEDED

Started: dg. de maig 10 11:22:41 +0200 2020
 Launched: dg. de maig 10 11:22:41 +0200 2020
 Finished: dg. de maig 10 11:22:56 +0200 2020
 Elapsed: 15sec

Tracking URL: History
 Log Aggregation Status: DISABLED
 Application Timeout (Remaining Time): Unlimited

Diagnostics:
 Unmanaged Application: false
 Application Node Label expression: <Not set>
 AM Container Node Label expression: <DEFAULT_PARTITION>

Application Metrics

Total Resource Preempted: <memory:0, vCores:0>
 Total Number of Non-AM Containers Preempted: 0
 Total Number of AM Containers Preempted: 0
 Resource Preempted from Current Attempt: <memory:0, vCores:0>
 Number of Non-AM Containers Preempted from Current Attempt: 0
 Aggregate Resource Allocation: 71150 MB-seconds, 34 vcore-seconds
 Aggregate Preempted Resource Allocation: 0 MB-seconds, 0 vcore-seconds

Show 20 entries

Attempt ID	Started	Node	Logs	Nodes blacklisted by the app	Nodes blacklisted by the system
appattempt_1588925284107_0003_000001	Sun May 10 11:22:41 +0200 2020	http://node2:8042	Logs	0	0

Logs for container_1588925284107_0003_01_000001 - Mozilla Firefox

Logs for container_1588925284107_0003_01_000001

prelaunch.err : Total file length is 0 bytes.
 prelaunch.out : Total file length is 70 bytes.
 stderr : Total file length is 20772 bytes.
 stdout : Total file length is 31 bytes.

Logs for container_1588925284107_0003_01_000001 - Mozilla Firefox

Logs for container_1588925284107_0003_01_000001

Pi is roughly 3.14039114039114



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

3.3. Clúster Hadoop amb Ambari

- En aquest clúster tornarem a fer servir la màquina virtual node1 on només té el sistema operatiu (CentOS 6 i les Guest Additions)
- Haurem de canviar els noms del host, configurar les xarxes i configura el SSH com ja havíem fet anteriorment
- Perquè funcioni, hem d'utilitzar l'usuari “root”.
- Hem de tenir en compte que un cop estigui operatiu, si parem les màquines haurem d'encendre el servei (1 per un, tenint en compte les prioritats, he fet la prova i demora uns 20 min aprox.)

3.3.1. Connexions entre el clúster: Xarxa interna

Node1-ambari

- Modifiquem el nom del host
 - [root@node1 Escriptori]# hostname node1-ambari
 - [root@node1-ambari Escriptori]# nano /etc/sysconfig/network
- Configurant l'arxiu /etc/hosts
 - [root@node1 Escriptori]# nano /etc/hosts
 - 192.168.30.201 node1-ambari
 - 192.168.30.202 node2-ambari
 - 192.168.30.203 node3-ambari
 - [root@node1 Escriptori]# ip a

Node2-ambari

- Modifiquem el nom del host
 - [root@node1 Escriptori]# hostname node1-ambari
 - [root@node2-ambari Escriptori]# nano /etc/sysconfig/network
- Configurant l'arxiu /etc/hosts
 - [root@node2 Escriptori]# nano /etc/hosts
 - 192.168.30.201 node1-ambari
 - 192.168.30.202 node2-ambari
 - 192.168.30.203 node3-ambari
 - [root@node2 Escriptori]# ip a

Node3-ambari

- Modifiquem el nom del host
 - [root@node1 Escriptori]# hostname node1-ambari
 - [root@node3-ambari Escriptori]# nano /etc/sysconfig/network
- Configurant l'arxiu /etc/hosts
 - [root@node3 Escriptori]# nano /etc/hosts
 - 192.168.30.201 node1-ambari
 - 192.168.30.202 node2-ambari
 - 192.168.30.203 node3-ambari
 - [root@node3 Escriptori]# ip a

- Fem ping entre les màquines per confirmar que es poden veure entre elles.



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

3.3.2. SSH :Claus públiques entre els nodes

- No mostrarem les captures, ja que es el mateix que abans. Així i tot mostrarem les comandes.

Node1-ambari

- [root@node1-ambari .ssh]# **ssh-keygen**
- [root@node1-ambari .ssh]# **cp id_rsa.pub authorized_keys**
- [root@node1-ambari .ssh]# **cat authorized_keys**
- [root@node1-ambari .ssh]# **scp authorized_keys node2-ambari:/root/.ssh**
-

Node2-ambari

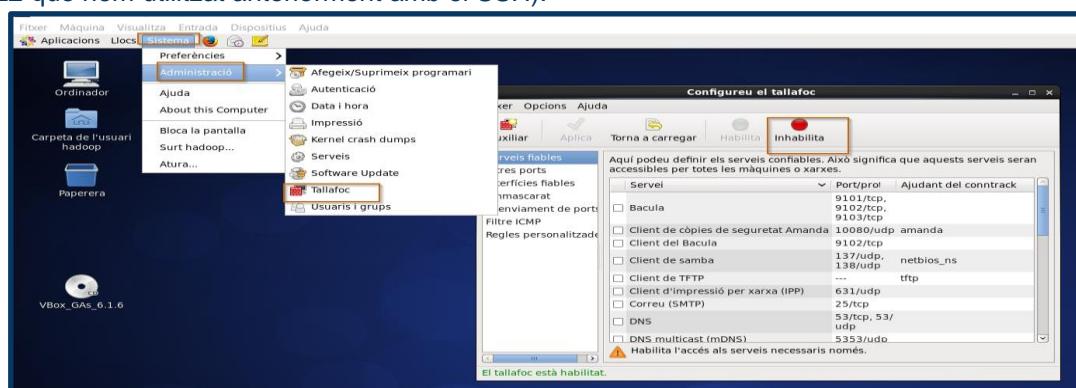
- [root@node1-ambari Escriptori]# **ssh node2-ambari**
- [root@node1-ambari Escriptori]# **cd .ssh**
- [root@node2-ambari .ssh]# **ssh-keygen**
- [root@node2-ambari .ssh]# **cat id_rsa.pub >> authorized_keys**
- [root@node2-ambari .ssh]# **scp authorized_keys node3-ambari:/root/.ssh**
-

Node3-ambari

- [root@node1-ambari Escriptori]# **ssh node3-ambari**
- [root@node3-ambari Escriptori]# **cd .ssh**
- [root@node3-ambari .ssh]# **ssh-keygen**
- [root@node3-ambari .ssh]# **cat id_rsa.pub >> authorized_keys**
- [root@node3-ambari .ssh]# **scp authorized_keys node2-ambari:/root/.ssh**
- [root@node3-ambari .ssh]# **scp authorized_keys node1-ambari:/root/.ssh**

3.3.3. Instal.lació i configuració del servidor Ambari

- Abans de continuar ens hem d'assegurar que el firewall estigui desactivat (o crear regles en les iptables de tots els ports que ens interessi que estiguin oberts, ja que per defecte només està obert el port 22 que hem utilitzat anteriorment amb el SSH).



- Com que el node1-ambari farà de master instal·larem Ambari 2.2.1 des de repositoris públics seguint les pautes de la url oficial :
 - <https://cwiki.apache.org/confluence/display/AMBARI/Install+Ambari+2.2.1+from+Public+Repositories>

Node1-ambari



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

```
[root@node1-ambari ~]# cd /etc/yum.repos.d/
[root@node1-ambari ~]# wget http://public-repo-1.hortonworks.com/ambari/centos6/2.x/updates/2.2.1.0/ambari.repo
[root@node1-ambari ~]# yum install ambari-server
[root@node1-ambari ~]# yum ambari-server setup
```

- Hem de desactivar SELinux¹⁷. Si no tenim JAVA i JCE ens ho instal.lara

- Si no tenim cap base de dades, una opció ràpida és la número 1 PostgreSQL ja està integrat

¹⁷ És un mòdul de seguretat del kernel de Linux

Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

```
hadoop@node1_ambari:/home/hadoop/Escriptori
Fitxer Edita Visualitza Cerca Terminal Ajuda
Successfully installed JDK to /usr/jdk64/
Downloading JCE Policy archive from http://public-repo-1.hortonworks.com/ARTIFAC
TS/jce_policy-8.zip to /var/lib/ambari-server/resources/jce_policy-8.zip

Successfully downloaded JCE Policy archive to /var/lib/ambari-server/resources/jce_policy-8.zip
Installing JCE policy...
Completing setup...
ver.u Configuring database...
Enter advanced database configuration [y/n] (n)? y
Configuring database...
=====
Choose one of the following options:
Server [1] - PostgreSQL (Embedded)
[2] - Oracle
[3] - MySQL
[4] - PostgreSQL
[5] - Microsoft SQL Server (Tech Preview)
[6] - SQL Anywhere
=====
Enter choice (1): 1
Database name (ambari):
Postgres schema (ambari):
Username (ambari):
Enter Database Password (bigdata):
```

```
hadoop@node1_ambari:/home/hadoop/Escriptori
Fitxer Edita Visualitza Cerca Terminal Ajuda
[6] - SQL Anywhere
=====
Enter choice (1): 1
Database name (ambari):
Postgres schema (ambari):
Username (ambari):
Enter Database Password (bigdata):
Default properties detected. Using built-in database.
Configuring ambari database...
Checking PostgreSQL...
Running initdb: This may take upto a minute.
S'està inicialitzant la base de dades: [ FET ]

int About to start PostgreSQL
Configuring local database...
Connecting to local database...connection timed out...retrying (1)
Connecting to local database...done.
Configuring PostgreSQL...
Extracting system views...
.....ambari-admin-2.2.1.0.161.jar

Adjusting ambari-server permissions and ownership...
Ambari Server 'setup' completed successfully.
[root@node1_ambari Escriptori]#
```

- Arrenquem el servei i revisem els processos

Node1-ambari

```
[root@node1-ambari ~]# yum ambari-server start
[root@node1-ambari ~]# ps -ef | grep ambari
```

Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

```

hadoop@node1_ambari:/home/hadoop/Escriptori
Fitxer Edita Visualitza Cerca Terminal Ajuda
[root@node1_ambari Escriptori]# ambari-server start
Using python /usr/bin/python2
Starting ambari-server
Ambari Server running with administrator privileges.
Running initdb: This may take upto a minute.
About to start PostgreSQL
Organizing resource files at /var/lib/ambari-server/resources...
Server PID at: /var/run/ambari-server/ambari-server.pid
Server out at: /var/log/ambari-server/ambari-server.out
Server log at: /var/log/ambari-server/ambari-server.log
Waiting for server start.....
Ambari Server 'start' completed successfully.
[root@node1_ambari Escriptori]#

```



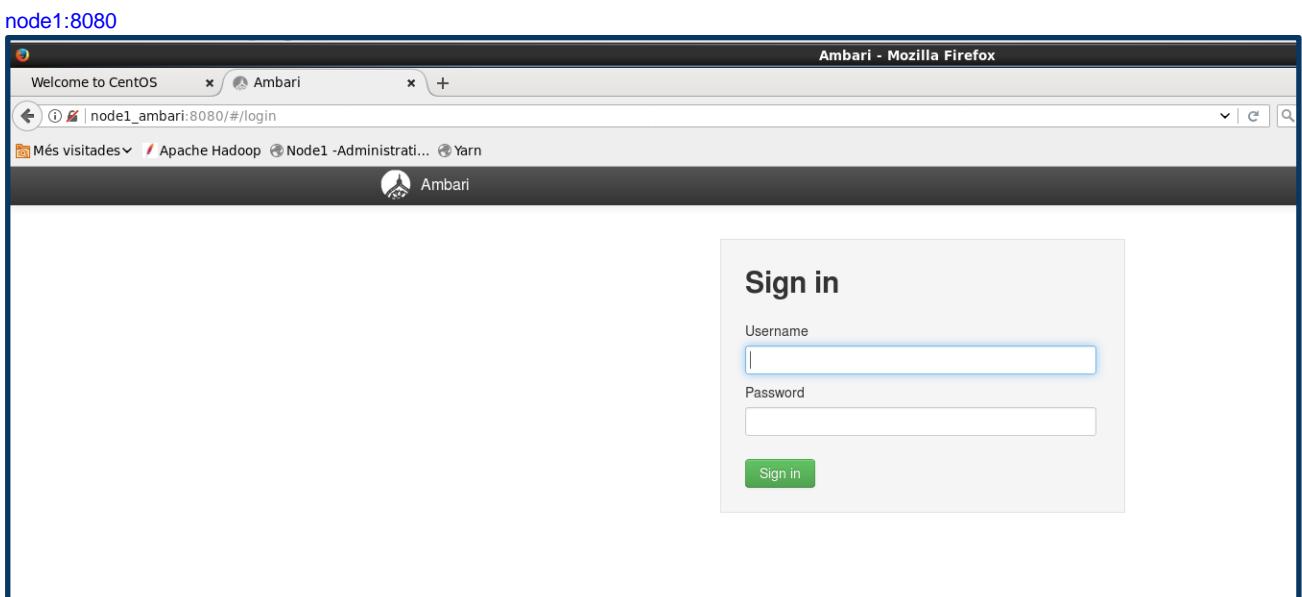
```

hadoop@node1_ambari:/home/hadoop/Escriptori
Fitxer Edita Visualitza Cerca Terminal Ajuda
[root@node1_ambari Escriptori]# ps -ef | grep ambari
root      4529     1 59 11:35 pts/0    00:00:42 /usr/jdk64/jdk1.8.0_60/bin/java -server -XX:NewRatio=3 -XX:+UseConcMarkSweepGC -XX:-UseGCOverheadLimit -XX:CMSInitiatingOccupancyFraction=60 -Dsun.zip.disableMemoryMapping=true -Xms512m -Xmx2048m -Djava.security.auth.login.config=/etc/ambari-server/conf/krb5JAASLogin.conf -Djava.security.krb5.conf=/etc/krb5.conf -Djavax.security.auth.useSubjectCredsOnly=false -cp /etc/ambari-server/conf:/usr/lib/ambari-server/*:/usr/share/java/postgresql-jdbc.jar org.apache.ambari.server.controller.AmbariServer
postgres   4544   3734  0 11:35 ?        00:00:00 postgres: ambari ambari 127.0.0.1(42172) idle
postgres   4547   3734  0 11:35 ?        00:00:00 postgres: ambari ambari 127.0.0.1(42176) idle
postgres   4591   3734  0 11:36 ?        00:00:00 postgres: ambari ambari 127.0.0.1(42182) idle
root      4601   3442  0 11:36 pts/0    00:00:00 grep ambari
[root@node1_ambari Escriptori]#

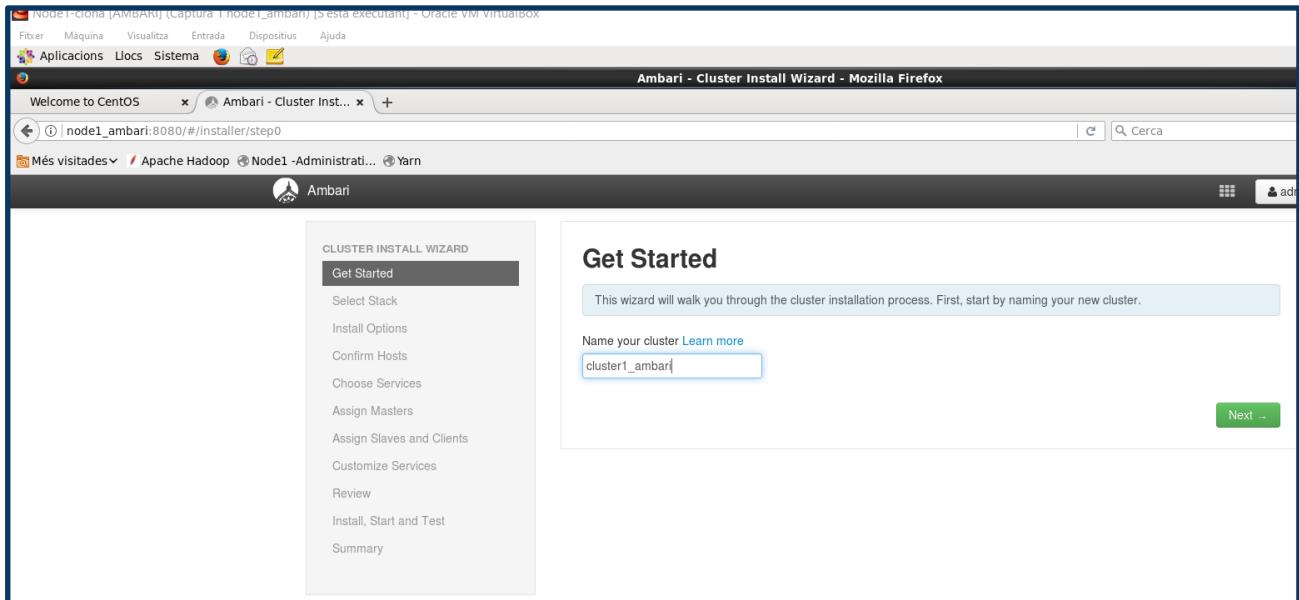
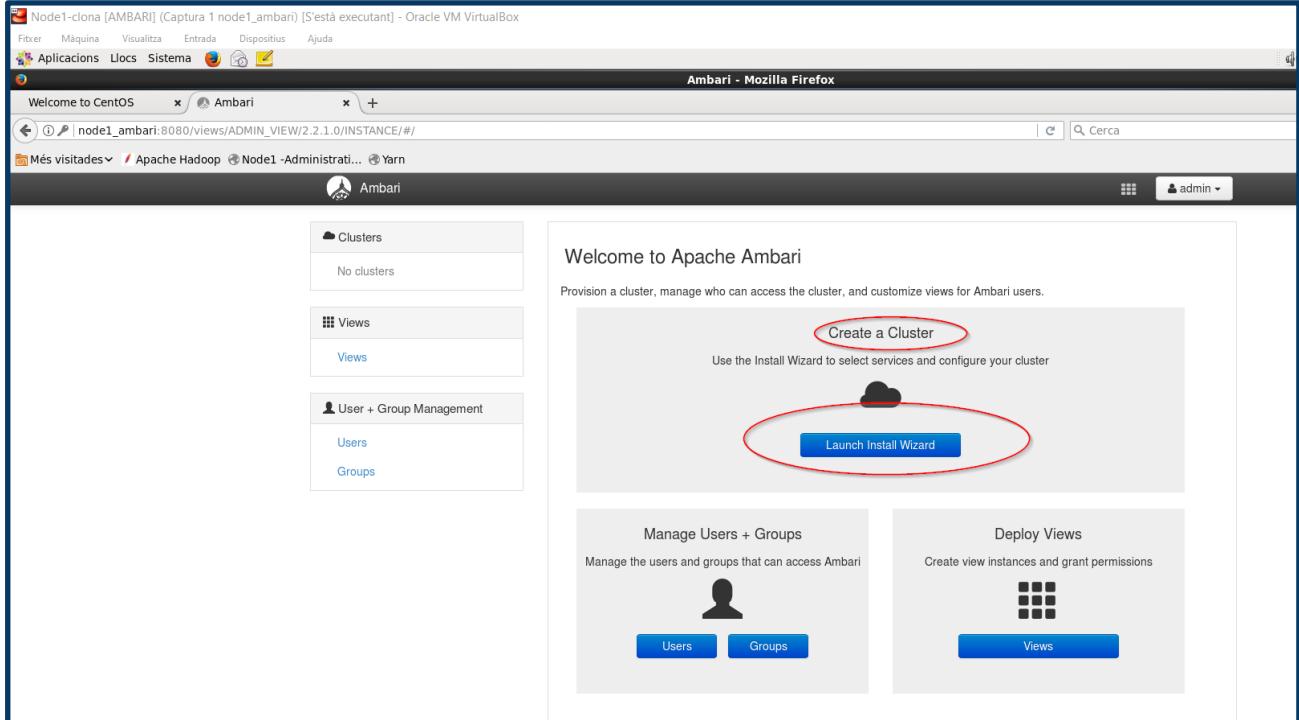
```

- Accedim a la interfície web del servidor Ambari (utilitza el port 8080) on haurem de fer la configuració del clúster.

Nota: per defecte el username i password és **admin**



Nom i Cognoms	Data
Arnaud Subirós Puigarnau	02-06-2020





Nom i Cognoms

Arnaud Subirós Puigarnau

Data

02-06-2020

Seleccionem l'última versió dels repositoris de Hortonworks

The screenshot shows the 'Select Stack' step of the Ambari Cluster Install Wizard. On the left, a sidebar lists steps: Get Started, Select Stack (which is highlighted), Install Options, Confirm Hosts, Choose Services, Assign Masters, Assign Slaves and Clients, Customize Services, Review, Install, Start and Test, and Summary. The main panel title is 'Select Stack' with the sub-instruction 'Please select the service stack that you want to use to install your Hadoop cluster.' Below this, under 'Stacks', the radio button for 'HDP 2.4' is selected. There is also a link 'Advanced Repository Options'. At the bottom right is a green 'Next >' button.

- La següent opció hem d'afegir els noms del host i la clau privada del host([id_rsa del node1-ambari](#)). Com que la clau privada està en una carpeta oculta no la podrem insertar, per això l'haurem a una altre carpeta([/tmp](#)) que tingui visibilitat i permisos.

The screenshot shows the 'Install Options' step of the Ambari Cluster Install Wizard. The sidebar shows steps: Get Started, Select Stack, Install Options (which is highlighted), Confirm Hosts, Choose Services, Assign Masters, Assign Slaves and Clients, Customize Services, Review, Install, Start and Test, and Summary. The main panel has sections for 'Target Hosts' (listing node1_ambari, node2_ambari, node3_ambari) and 'Host Registration Information' (with a checked radio button for 'Provide your SSH Private Key to automatically register hosts'). A file dialog window titled 'Puja un fitxer' is overlaid, showing the contents of the '/tmp' directory. Inside the dialog, the 'id_rsa' file is selected and highlighted with a red box. The dialog includes buttons 'Cancel·la' and 'Obre' at the bottom right.



Nom i Cognoms

Arnaud Subirós Puigarnau

Data

02-06-2020

The screenshot shows the 'Install Options' step of the Ambari Cluster Install Wizard. On the left, a sidebar lists steps: Get Started, Select Stack, Install Options (selected), Confirm Hosts, Choose Services, Assign Masters, Assign Slaves and Clients, Customize Services, Review, Install, Start and Test, and Summary. The main area is titled 'Install Options' with a sub-section 'Target Hosts'. It contains a text input field for host names and a note about using Fully Qualified Domain Names (FQDN). Below it is a 'Host Registration Information' section with a radio button for 'Provide your SSH Private Key' and a terminal window showing the exchange of an RSA private key between 'node1_ambari' and 'hadoop@node1_ambari'. The terminal also shows the command 'id_rsa' being checked.

The screenshot shows the 'Confirm Hosts' step of the Ambari Cluster Install Wizard. The sidebar shows the 'Confirm Hosts' step is selected. The main area displays a table of hosts: node1_ambari, node2_ambari, and node3_ambari, all in the 'Installing' status. A progress bar indicates the installation process. The table includes columns for Host, Progress, Status, and Action (with a 'Remove' link). Buttons for 'Back' and 'Next' are at the bottom.

The screenshot shows the 'Confirm Hosts' step of the Ambari Cluster Install Wizard. The sidebar shows the 'Confirm Hosts' step is selected. The main area displays a table of hosts: node1_ambari, node2_ambari, and node3_ambari, all in the 'Success' status. A progress bar indicates the success of the installation. The table includes columns for Host, Progress, Status, and Action (with a 'Remove' link). A message at the bottom says 'Please wait while the hosts are being checked for potential problems...' with a loading icon. Buttons for 'Back' and 'Next' are at the bottom.



Nom i Cognoms

Arnaud Subirós Puigarnau

Data

02-06-2020

- Un cop instal·lat hem de seleccionar quins serveis volem afegir, tot i que n'hi ha alguns com "Zookeeper" t'obliga a instal·lar-los per poder continuar.

CLUSTER INSTALL WIZARD
Get Started
Select Stack
Install Options
Confirm Hosts
Choose Services
Assign Masters
Assign Slaves and Clients
Customize Services
Review
Install, Start and Test
Summary

Choose Services

Choose which services you want to install on your cluster.

Service	Version	Description
<input type="checkbox"/> HDFS	2.7.1.2.4	Apache Hadoop Distributed File System
<input checked="" type="checkbox"/> YARN + MapReduce2	2.7.1.2.4	Apache Hadoop NextGen MapReduce (YARN)
<input type="checkbox"/> Tez	0.7.0.2.4	Tez is the next generation Hadoop Query Processing framework written on top of YARN.
<input checked="" type="checkbox"/> Hive	1.2.1.2.4	Data warehouse system for ad-hoc queries & analysis of large datasets and table & storage management service
<input type="checkbox"/> HBase	1.1.2.2.4	A Non-relational distributed database, plus Phoenix, a high performance SQL layer for low latency applications.
<input checked="" type="checkbox"/> Pig	0.15.0.2.4	Scripting platform for analyzing large datasets
<input checked="" type="checkbox"/> Sqoop	1.4.6.2.4	Tool for transferring bulk data between Apache Hadoop and structured data stores such as relational databases
<input type="checkbox"/> Oozie	4.2.0.2.4	System for workflow coordination and execution of Apache Hadoop jobs. This also includes the installation of the optional Oozie Web Console which relies on and will install the ExtJS Library.
<input checked="" type="checkbox"/> ZooKeeper	3.4.6.2.4	Centralized service which provides highly reliable distributed coordination
<input type="checkbox"/> Falcon	0.6.1.2.4	Data management and processing platform

Ambari - Cluster Install Wizard - Mozilla Firefox

Welcome to CentOS | Ambari - Cluster Inst... | node1_ambari:8080/#/installer/step4

Més visitades | Apache Hadoop | Node1 -Administrati... | Yarn

Install the ExtJS Library.

Choose Services

Choose which services you want to install on your cluster.

Service	Version	Description
<input checked="" type="checkbox"/> ZooKeeper	3.4.6.2.4	Centralized service which provides highly reliable distributed coordination
<input type="checkbox"/> Falcon	0.6.1.2.4	Data management and processing platform
<input type="checkbox"/> Storm	0.10.0.2.4	Apache Hadoop Stream processing framework
<input type="checkbox"/> Flume	1.5.2.2.4	A distributed service for collecting, aggregating, and moving large amounts of streaming data into HDFS
<input type="checkbox"/> Accumulo	1.7.0.2.4	Robust, scalable, high performance distributed key/value store.
<input checked="" type="checkbox"/> Ambari Metrics	0.1.0	A system for metrics collection that provides storage and retrieval capability for metrics collected from the cluster
<input type="checkbox"/> Atlas	0.5.0.2.4	Atlas Metadata and Governance platform
<input type="checkbox"/> Kafka	0.9.0.2.4	A high-throughput distributed messaging system
<input type="checkbox"/> Knox	0.6.0.2.4	Provides a single point of authentication and access for Apache Hadoop services in a cluster
<input type="checkbox"/> Mahout	0.9.0.2.4	Project of the Apache Software Foundation to produce free implementations of distributed or otherwise scalable machine learning algorithms focused primarily in the areas of collaborative filtering, clustering and classification
<input type="checkbox"/> Slider	0.80.0.2.4	A framework for deploying, managing and monitoring existing distributed applications on YARN.
<input type="checkbox"/> SmartSense	1.2.1.0-161	SmartSense - Hortonworks SmartSense Tool (HST) helps quickly gather configuration, metrics, logs from common HDP services that aids to quickly troubleshoot support cases and receive cluster-specific recommendations.
<input checked="" type="checkbox"/> Spark	1.6.0.2.4	Apache Spark is a fast and general engine for large-scale data processing.

[Back](#) [Next →](#)

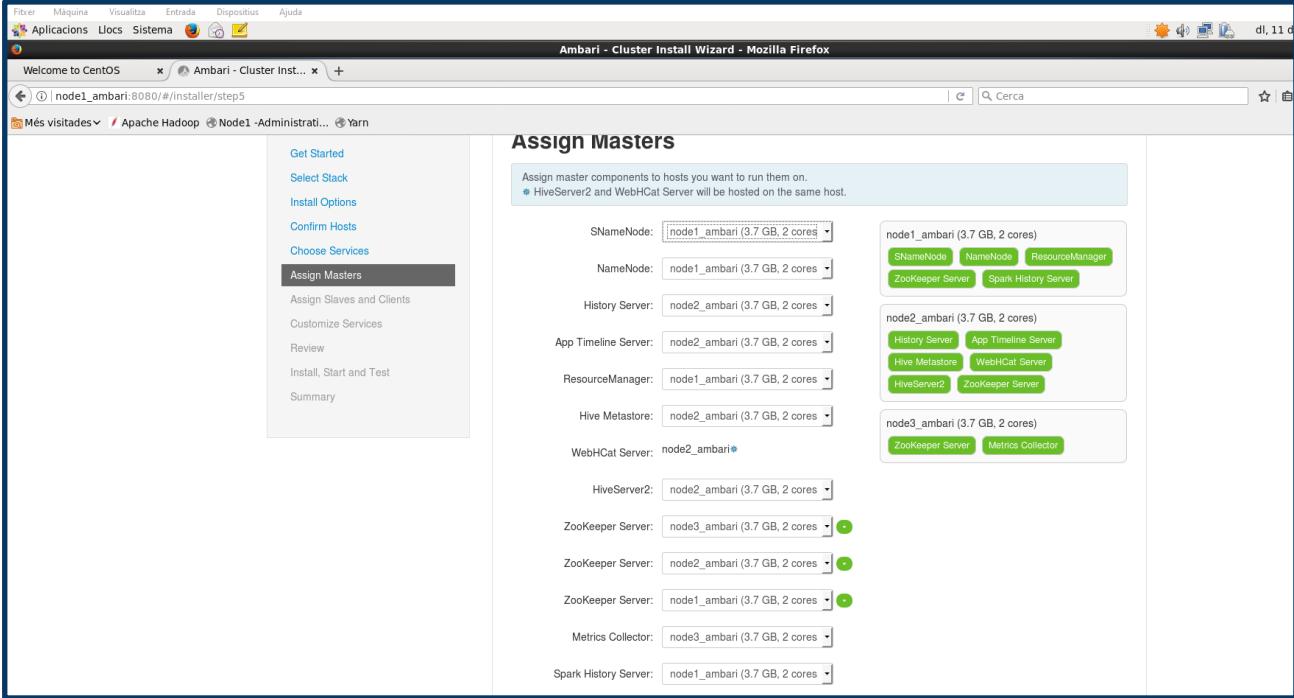
Nom i Cognoms

Arnau Subirós Puigarnau

Data

02-06-2020

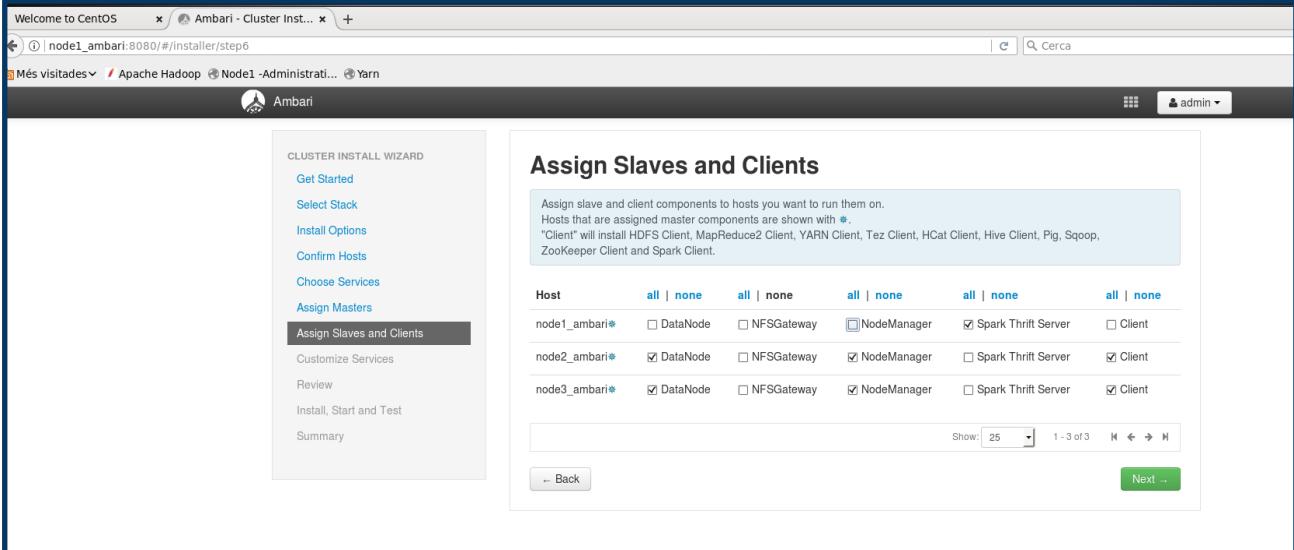
□ Després hem d'especificar a quin node volem instal·lar els servidors



The screenshot shows the 'Assign Masters' step of the Ambari Cluster Install Wizard. The left sidebar lists steps from 'Get Started' to 'Summary'. The main area is titled 'Assign Masters' with a sub-instruction: 'Assign master components to hosts you want to run them on.' Below this, a note says 'HiveServer2 and WebHCat Server will be hosted on the same host.' A table lists master components and their assigned hosts:

	SNameNode	NameNode	ResourceManager	ZooKeeper Server	Spark History Server
node1_ambari (3.7 GB, 2 cores)	node1_ambari (3.7 GB, 2 cores)				
node2_ambari (3.7 GB, 2 cores)					
node3_ambari (3.7 GB, 2 cores)					

Other master components listed but not assigned to specific hosts are: History Server, App Timeline Server, Hive Metastore, WebHCat Server, HiveServer2, ZooKeeper Server, Metrics Collector, and Spark History Server.



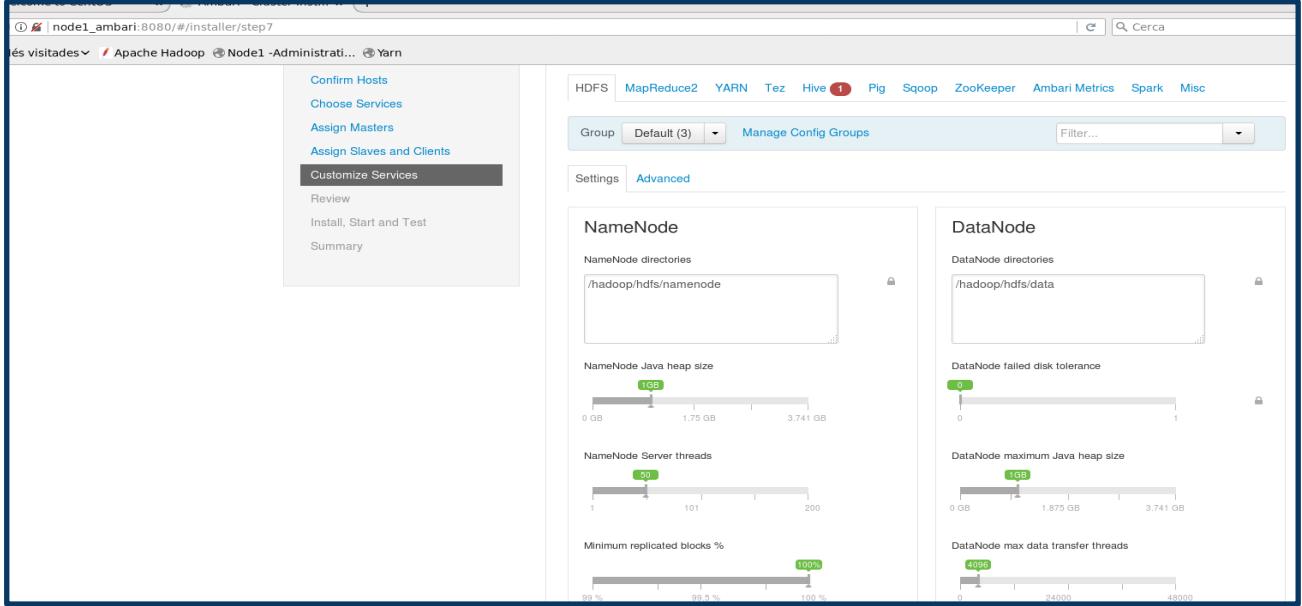
The screenshot shows the 'Assign Slaves and Clients' step of the Ambari Cluster Install Wizard. The left sidebar lists steps from 'Get Started' to 'Summary'. The main area is titled 'Assign Slaves and Clients' with a sub-instruction: 'Assign slave and client components to hosts you want to run them on.' It notes that hosts assigned master components are shown with *. Components assigned to each host are:

Host	all none	all none	all none	all none	all none
node1_ambari*	<input type="checkbox"/> DataNode	<input type="checkbox"/> NFSGateway	<input checked="" type="checkbox"/> NodeManager	<input checked="" type="checkbox"/> Spark Thrift Server	<input type="checkbox"/> Client
node2_ambari*	<input checked="" type="checkbox"/> DataNode	<input type="checkbox"/> NFSGateway	<input checked="" type="checkbox"/> NodeManager	<input type="checkbox"/> Spark Thrift Server	<input checked="" type="checkbox"/> Client
node3_ambari*	<input checked="" type="checkbox"/> DataNode	<input type="checkbox"/> NFSGateway	<input checked="" type="checkbox"/> NodeManager	<input type="checkbox"/> Spark Thrift Server	<input checked="" type="checkbox"/> Client

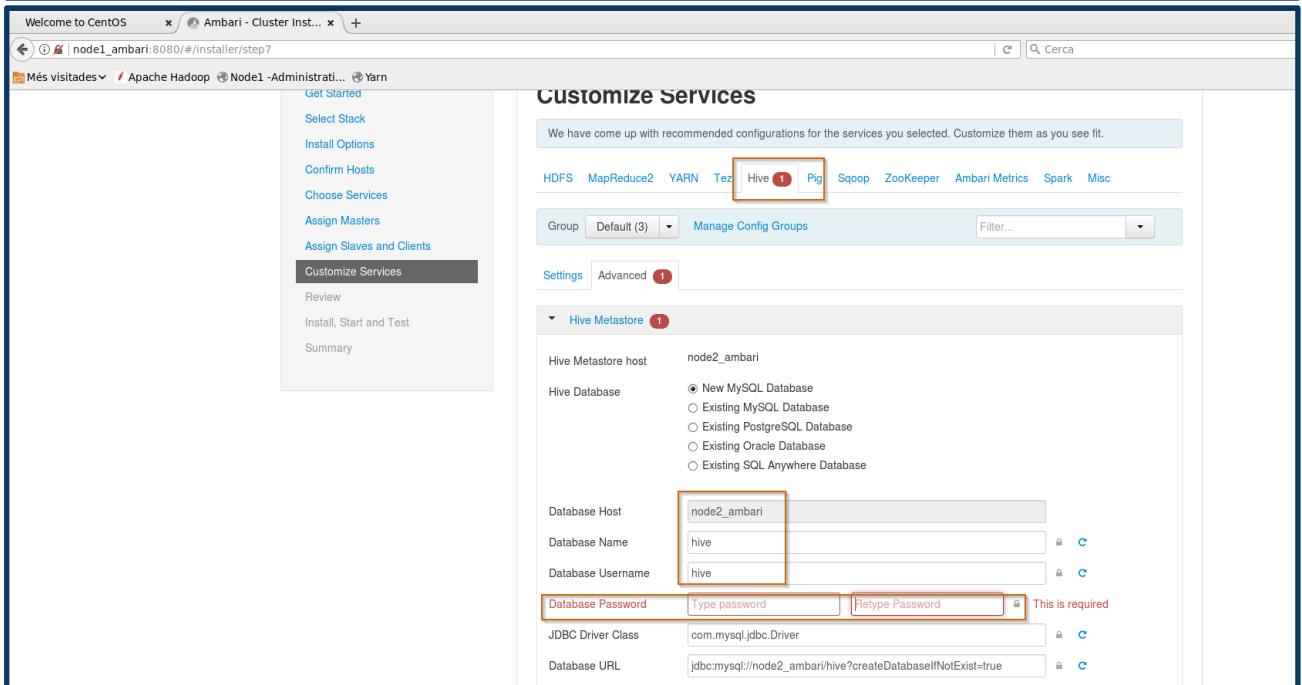
Slave components assigned are DataNode, NFSGateway, and NodeManager. Client components assigned are Spark Thrift Server and Client. Other components listed but not assigned are: MapReduce2 Client, YARN Client, Tez Client, HCat Client, Hive Client, Pig, Sqoop, ZooKeeper Client, and Spark Client.

Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

- Indica un error en el Hive, ja que la base de dades i usuari hive que s'ha generat hem de escriure el password



The screenshot shows the Ambari web interface for a Hadoop cluster. On the left, there's a sidebar with links like 'Customize Services', 'Review', 'Install, Start and Test', and 'Summary'. The main area has tabs for 'HDFS', 'MapReduce2', 'YARN', 'Tez', 'Hive' (which has a red notification dot), 'Pig', 'Sqoop', 'ZooKeeper', 'Ambari Metrics', 'Spark', and 'Misc'. Under 'HDFS', there are sections for 'NameNode' and 'DataNode', each with various configuration parameters and their current values.



This screenshot shows the 'Customize Services' step in the Ambari cluster installation wizard. It's specifically configured for the 'Hive Metastore' service. The 'Database Host' field is set to 'node2_ambari', 'Database Name' to 'hive', and 'Database Username' to 'hive'. The 'Database Password' and 'Retype Password' fields are highlighted with a red box, indicating they are required fields. Other settings include 'JDBC Driver Class' (set to 'com.mysql.jdbc.Driver') and 'Database URL' (set to 'jdbc:mysql://node2_ambari/hive?createDatabaseIfNotExist=true').



Nom i Cognoms

Arnaud Subirós Puigarnau

Data

02-06-2020

○ Existing SQL Anywhere Database

Database Host	node2_ambari
Database Name	hive
Database Username	hive
Database Password	*****
JDBC Driver Class	com.mysql.jdbc.Driver
Database URL	jdbc:mysql://node2_ambari/hive?createDatabaseIfNotExist=true
Hive Database Type	mysql

Welcome to CentOS x / Ambari - Cluster Inst... x +

node1_ambari:8080/#/installer/step8

Més visitades / Apache Hadoop / Node1 - Administrati... / Yarn

Ambari

CLUSTER INSTALL WIZARD

- Get Started
- Select Stack
- Install Options
- Confirm Hosts
- Choose Services
- Assign Masters
- Assign Slaves and Clients
- Customize Services
- Review**
- Install, Start and Test
- Summary

Review

Please review the configuration before installation

http://public-repo-1.hortonworks.com/HDP-UTILS-1.1.0.20/repos/ubuntu12

Services:

HDFS

- DataNode : 2 hosts
- NameNode : node1_ambari
- NFSGateway : 0 host
- SNameNode : node1_ambari

YARN + MapReduce2

- App Timeline Server : node2_ambari
- NodeManager : 2 hosts
- ResourceManager : node1_ambari

Tez

- Clients : 2 hosts

Hive

- Metastore : node2_ambari
- HiveServer2 : node2_ambari
- WebHCat Server : node2_ambari
- Database : MySQL (New MySQL Database)

Pig

- Clients : 2 hosts

Sqoop

- Clients : 2 hosts

ZooKeeper

... Back Print Deploy ...

Welcome to CentOS x / Ambari - Cluster Inst... x +

node1_ambari:8080/#/installer/step9

Més visitades / Apache Hadoop / Node1 - Administrati... / Yarn

Ambari

CLUSTER INSTALL WIZARD

- Get Started
- Select Stack
- Install Options
- Confirm Hosts
- Choose Services
- Assign Masters
- Assign Slaves and Clients
- Customize Services
- Review
- Install, Start and Test**
- Summary

Install, Start and Test

Please wait while the selected services are installed and started.

3 % overall

Host	Status	Message
node1_ambari	3%	Waiting to install HDFS Client
node2_ambari	3%	Waiting to install App Timeline Server
node3_ambari	3%	Waiting to install DataNode

Show: All (3) In Progress (3) | Warning (0) | Success (0) | Fail (0)

3 of 3 hosts showing - Show All

Show: 25 1 - 3 of 3 Next ...

Nom i Cognoms

Arnau Subirós Puigarnau

Data

02-06-2020

Ambari - Cluster Install Wizard - Mozilla Firefox

node1:8080/#/installer/step9

Més visitades ▾ / Apache Hadoop @ Node1 -Administrati... @ Yarn

Ambari admin

CLUSTER INSTALL WIZARD

- Get Started
- Select Stack
- Install Options
- Confirm Hosts
- Choose Services
- Assign Masters
- Assign Slaves and Clients
- Customize Services
- Review
- Install, Start and Test**
- Summary

Install, Start and Test

Please wait while the selected services are installed and started.

100 % overall

Show: All (3) In Progress (0) Warning (0) Success (3) Fail (0)		
Host	Status	Message
node1	100%	Success
node2	100%	Success
node3	100%	Success

3 of 3 hosts showing - Show All

Show: 25 1 - 3 of 3

Successfully installed and started the services.

Next →

Ambari - Cluster Inst... - Mozilla Firefox

node1:8080/#/installer/step10

Més visitades ▾ / Apache Hadoop @ Node1 -Administrati... @ Yarn

Ambari admin

CLUSTER INSTALL WIZARD

- Get Started
- Select Stack
- Install Options
- Confirm Hosts
- Choose Services
- Assign Masters
- Assign Slaves and Clients
- Customize Services
- Review
- Install, Start and Test
- Summary**

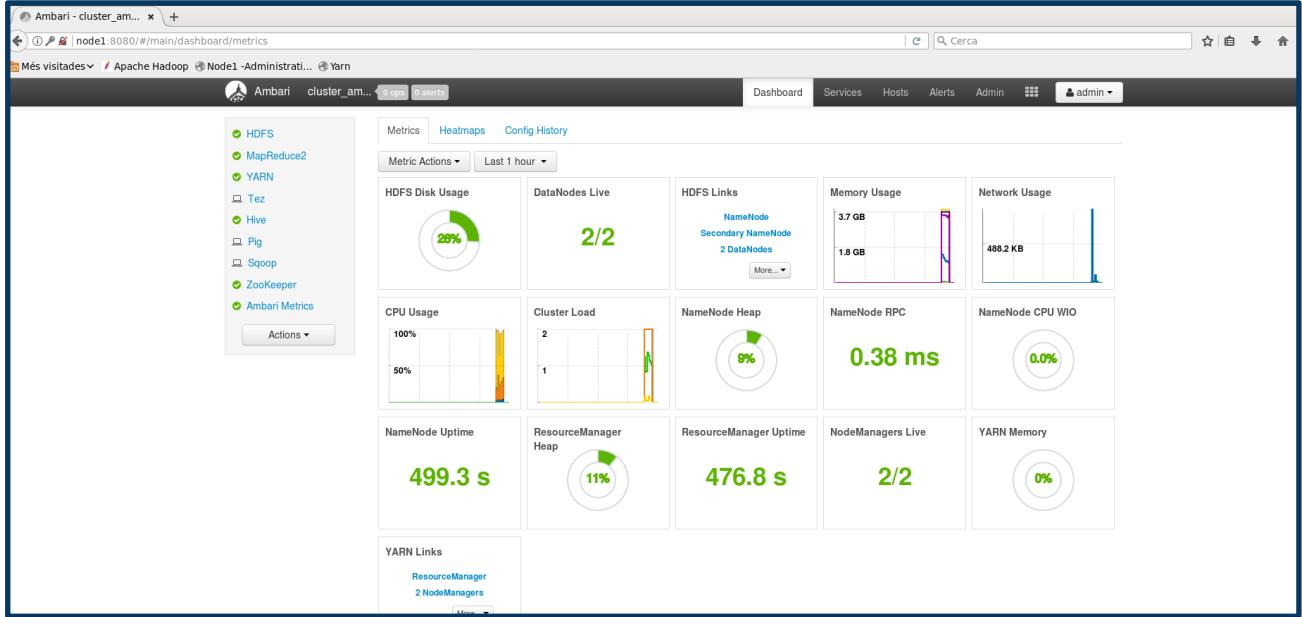
Summary

Here is the summary of the install process.

The cluster consists of 3 hosts
 Installed and started services successfully on 3 new hosts
 Master services installed
 NameNode installed on node1
 SNameNode installed on node1
 History Server installed on node2
 ResourceManager installed on node1
 HiveServer2 installed on node2
 All services started
 All tests passed
 Install and start completed in 14 minutes and 58 seconds

Complete ..

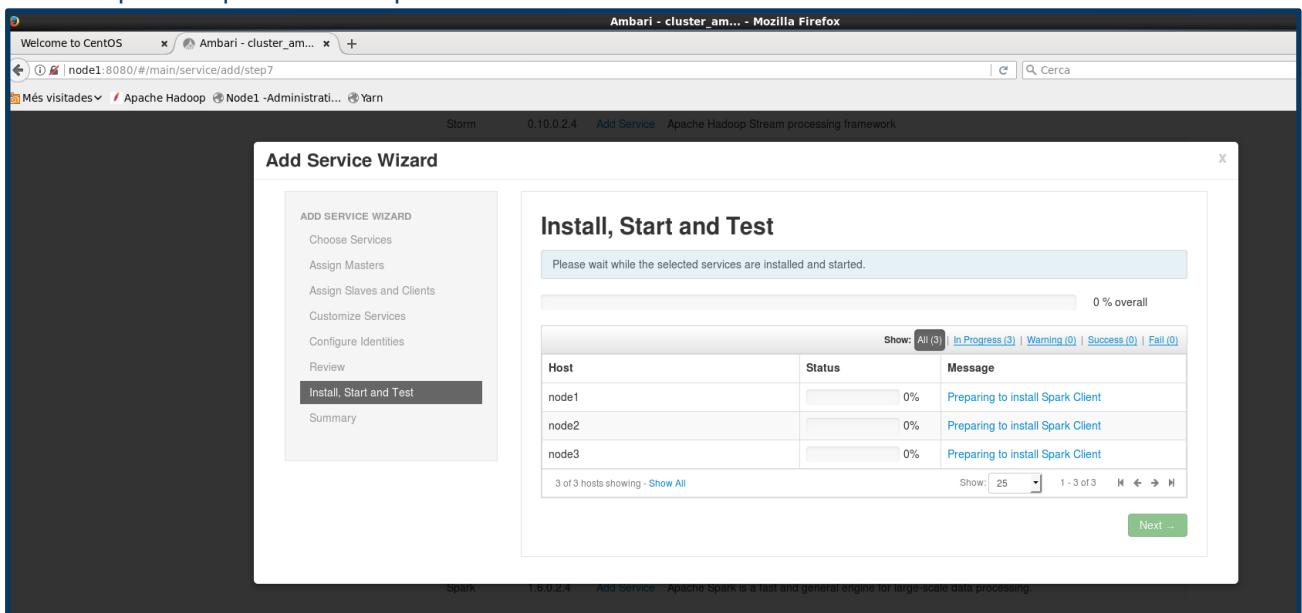
Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020



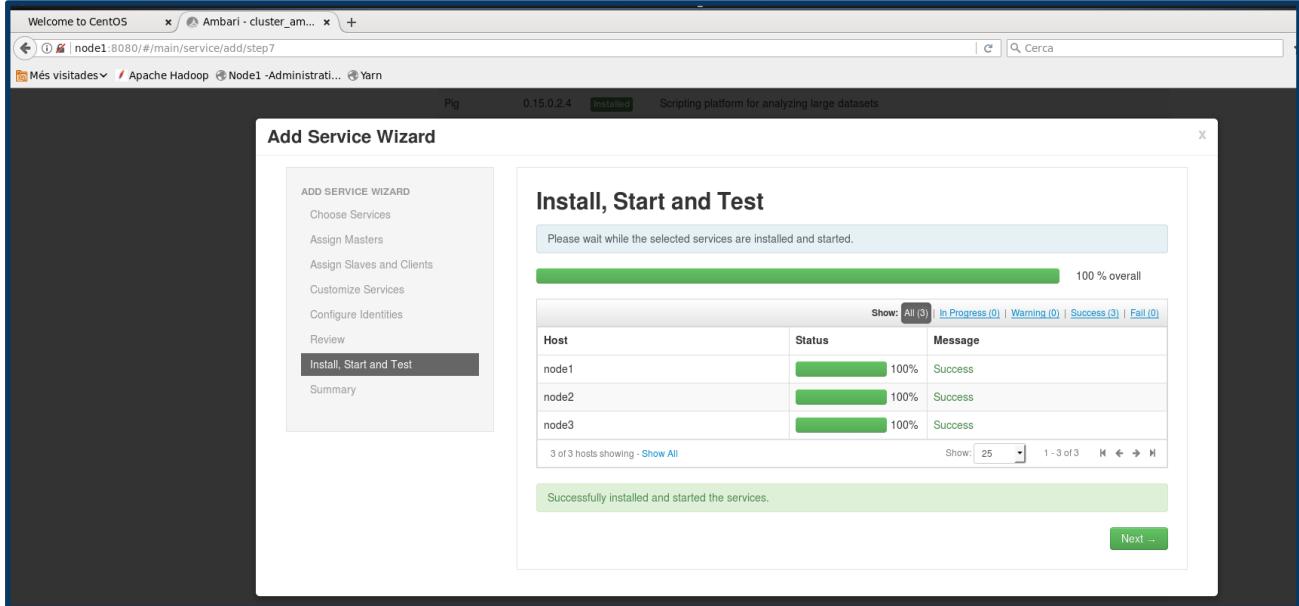
- Després de fer un backup de les màquines a un disc extern. En iniciar el servidor Ambari he tingut problemes, ja que no s'iniciaven els serveis. Per solucionar-ho desde el terminal com usuari hdfs hem tingut que desactivar el safe mode

```
[root@node1-ambari ~]#su - hdfs
[hdfs@node1-ambari ~]# hdfs dfsadmin -safemode leave
```

- Aprofitem per instal·lar Spark mode client

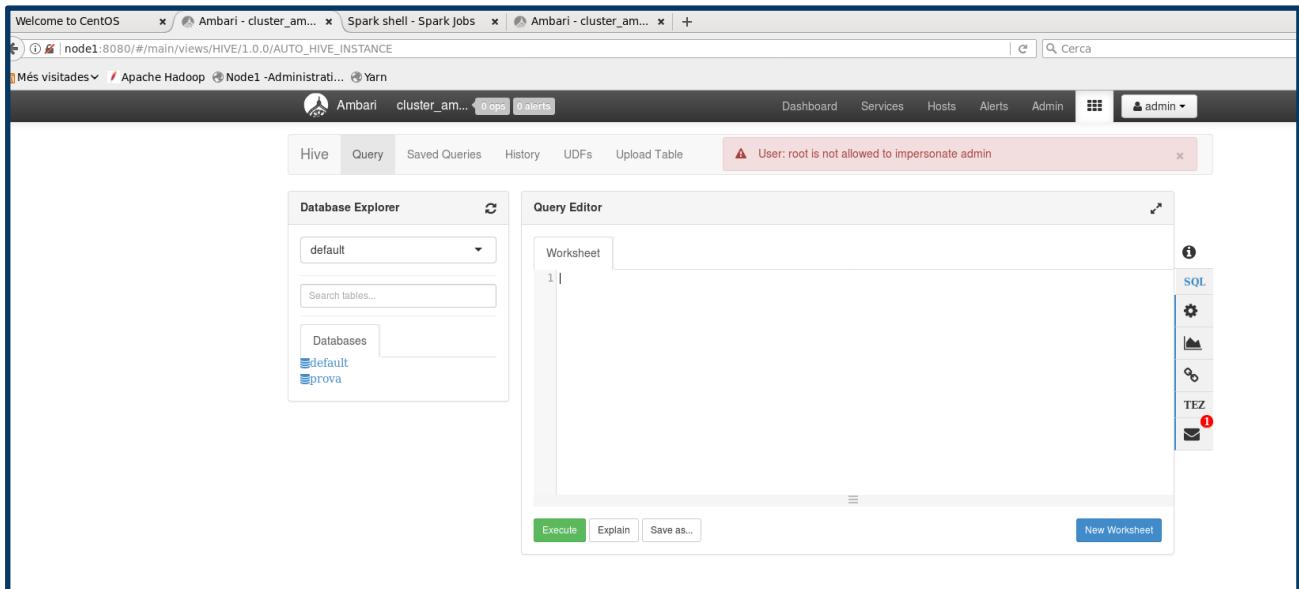


Nom i Cognoms	Data
Arnaud Subirós Puigarnau	02-06-2020



3.3.3.1. Configuració de Hive

- En accedir a l'editor de Hive, tinc problemes amb els permisos, ja que per defecte l'usuari és hive i no admin. Des del terminal fem la prova de crear una base de dades anomenada prova i l'intentem obrir sense èxit





Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

- ❑ Hem d'anar a les opcions avançades de HDFS i afegir unes propietats a [custom core-site](#).
- ❑ Totes les modificacions s'han de fer des de la interfície web, ja que de l'altra manera podria haver-hi errors
 - [hadoop.proxyuser.root.groups=*](#)
 - [hadoop.proxyuser.root.hosts=*](#)

The image contains two screenshots of the Ambari web interface, both showing the 'Add Property' dialog box. In the top screenshot, the 'Key' field is set to 'hadoop.proxyuser.root.groups' and the 'Value' field is empty. In the bottom screenshot, the 'Key' field is set to 'hadoop.proxyuser.root.hosts' and the 'Value' field is also empty. Both dialogs have 'Type' set to 'core-site.xml'.

- ❑ Després de fer les modificacions, podré accedir però no podré crear res. Al intentar fer un select de la taula01 de la bse de dades prova, no troba l'usuari admin

The image shows a screenshot of the Hive Query Editor in the Ambari interface. The 'Database Explorer' shows databases 'default' and 'prova'. The 'Query Editor' contains the following SQL code:

```
Worksheet
1 select * from test01;
```

A red error message at the top right of the editor window states: "E090 HDFS020 Could not write file /user/admin/hive/jobs/hive-job-4-2020-05-12_08-28/query.hql [HdfsApiException]".



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

- Creem l'usuari “arsupu” i el grup : “group_arsupu” i una instancia de HIVE anomenada “hive_arsupu_view” per no només tenir l'usuari admin.

The screenshot shows the Ambari interface for managing Apache Hadoop. A new instance is being created with the following details:

- Instance Name:** hive_arsupu_view
- Display Name:** hive_arsupu_view
- Description:** hive_arsupu_view
- Visible:** checked
- Permissions:** Grant permission to these users: admin, group_arsupu

- IMPORTANT:** Creem un directori HDFS pels 2 usuaris i li donem permisos.

```
[root@node1-ambari ~]# hdfs dfs -ls /user/hive
[root@node1-ambari ~]# hadoop fs -mkdir /user/admin
[root@node1-ambari ~]# hadoop fs -mkdir /user/arsupu
[root@node1-ambari ~]# hadoop fs -chown admin:hadoop /user/admin
[root@node1-ambari ~]# hadoop fs -chown arsupu:hadoop /user/arsupu
```

The terminal window shows the following commands being run:

```
Node1-Master clona_ambari (Captura 2 reiniciant ambari finalment ok) [S'està executant] - Oracle VM VirtualBox
Fitxer MÀquina Visualitzar Entrada Dispositius Ajuda
Aplicacions Llocs Sistema
hdbs@node1:~#
[root@node1 ambari-agent]# hdfs dfs -ls /user
Found 4 items
drwxrwx---  ambari-qa hdfs      0 2020-05-12 12:42 /user/ambari-qa
drwxr-xr-x   hcat     hdfs      0 2020-05-12 12:39 /user/hcat
drwxr-xr-x   hive     hdfs      0 2020-05-12 19:25 /user/hive
drwxrwxr-x   spark    hdfs      0 2020-05-12 18:39 /user/spark
[root@node1 ambari-agent]#
[root@node1 ambari-agent]# hdfs dfs -ls /user/hive
Found 1 items
drwxr-xr-x   - hive     hdfs      0 2020-05-12 19:25 /user/hive/.hiveJars
[root@node1 ambari-agent]# hdfs dfs -ls /user
Found 4 items
drwxrwx---  ambari-qa hdfs      0 2020-05-12 12:42 /user/ambari-qa
drwxr-xr-x   - hcat     hdfs      0 2020-05-12 12:39 /user/hcat
drwxr-xr-x   - hive     hdfs      0 2020-05-12 19:25 /user/hive
drwxrwxr-x   - spark    hdfs      0 2020-05-12 18:39 /user/spark
[root@node1 ambari-agent]#
[root@node1 ambari-agent]# su - hdfs
[hdbs@node1 ~]$ hadoop fs -mkdir /user/admin
[hdbs@node1 ~]$ hadoop fs -mkdir /user/arsupu
[hdbs@node1 ~]$ hadoop fs -chown admin:hadoop /user/admin
[hdbs@node1 ~]$ hadoop fs -chown arsupu:hadoop /user/arsupu
```



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

- ☐ Fem un select, per confirmar que funciona correctament, ja no tenim problema de permisos

```
use prova;
select * from test01;
```

Query Process Results (Status: Succeeded)

test01.nom	Albert

3.3.3.2. Visualització des de Windows 10

- ☐ Hem canviat el tipus d'adaptador pont de les màquines virtuals , no farem servir xarxa interna, sinó adaptador pont (Ethernet) ja que ens interessa fer servir la mateixa xarxa que utilitza el meu host amfitrió (192.168.1.0).
- ☐ Hem configurat les IP estàtiques en un rang que no doni problemes
- ☐ S'ha afegit les IP i el nom de les màquines en l'arxiu host (ja que no tenim servidor DNS)

Nom	Data
hosts	19/2/
Imhosts.sam	19/3/
networks	12/4/
protocol	12/4/
services	12/4/



Nom i Cognoms

Arnaud Subirós Puigarnau

Data

02-06-2020

A screenshot of a Windows hosts file editor window titled "hosts". The file contains sample host mappings. A new entry at the bottom is highlighted with an orange box:

```
1 # Copyright (c) 1993-2009 Microsoft Corp.  
2 #  
3 # This is a sample HOSTS file used by Microsoft TCP/IP for Windows.  
4 #  
5 # This file contains the mappings of IP addresses to host names. Each  
6 # entry should be kept on an individual line. The IP address should  
7 # be placed in the first column followed by the corresponding host name.  
8 # The IP address and the host name should be separated by at least one  
9 # space.  
10 #  
11 # Additionally, comments (such as these) may be inserted on individual  
12 # lines or following the machine name denoted by a '#' symbol.  
13 #  
14 # For example:  
15 #  
16 # 102.54.94.97      rhino.acme.com          # source server  
17 #       38.25.63.10      x.acme.com            # x client host  
18 #  
19 # localhost name resolution is handled within DNS itself.  
20 # 127.0.0.1          localhost  
21 # ::1                localhost  
22  
23  
24  
25 #79.158.191.62    rasp-asp.ddns.net    rasp-asp  
26  
27 #79.158.191.62    w10-asp2019.ddns.net 192.168.56.1  
28  
29  
30 192.168.1.211 ambari-node1 node1 master
```

A screenshot of a Mozilla Firefox browser window titled "Ambari". The address bar shows the URL "192.168.1.211:8080/#/login". The main content area displays the Ambari "Sign in" form with fields for "Username" and "Password" and a "Sign in" button.

A second screenshot of the same Mozilla Firefox browser window, showing the Ambari "Sign in" form again. The address bar is identical to the previous screenshot.

Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

NOTA: Però si volguéssim configurar-ho el clúster des de Windows 10 a més a més hauríem descarregar la clau privada (id_rsa) del server Ambari utilitzant Filezilla, ja que desde el navegador no podriem.

The screenshot shows the Ambari Cluster Install Wizard at step 2, titled "Install Options". The left sidebar lists steps: Get Started, Select Stack, **Install Options**, Confirm Hosts, Choose Services, Assign Masters, Assign Slaves and Clients, Customize Services, Review, Install, Start and Test, and Summary. The main area has a title "Install Options" and a sub-section "Target Hosts" with a text input field containing "ambari-node1", "rasp-node2", and "rasp-node3". Below it is a section "Host Registration Information" with two radio button options: "Provide your SSH Private Key to automatically register hosts" (selected) and "Perform manual registration on hosts and do not use SSH". A text input field for "SSH User Account" contains "root". At the bottom are "Back" and "Register and Confirm" buttons.

Ambari - Cluster Install Wizard

No segr | ambari-node1:8080/#/installer/step2

Install Options

Enter the list of hosts to be included in the cluster and provide your SSH key.

Target Hosts

Enter a list of hosts using the Fully Qualified Domain Name (FQDN), one per line. Or use Pattern Expressions

```
ambari-node1
rasp-node2
rasp-node3
```

Host Registration Information

Provide your **SSH Private Key** to automatically register hosts

Perform manual registration on hosts and do not use SSH

Tria un fitxer | No s'ha triat cap fitxer

```
-----BEGIN RSA PRIVATE KEY-----
MIIEowIBAAKCAQEAgA1+gryDAjHhnwvZ3e4ezpcTVzLcb5kmBLhKD5XxXG0tOfUx
P
```

SSH User Account

← Back Register and Confirm →

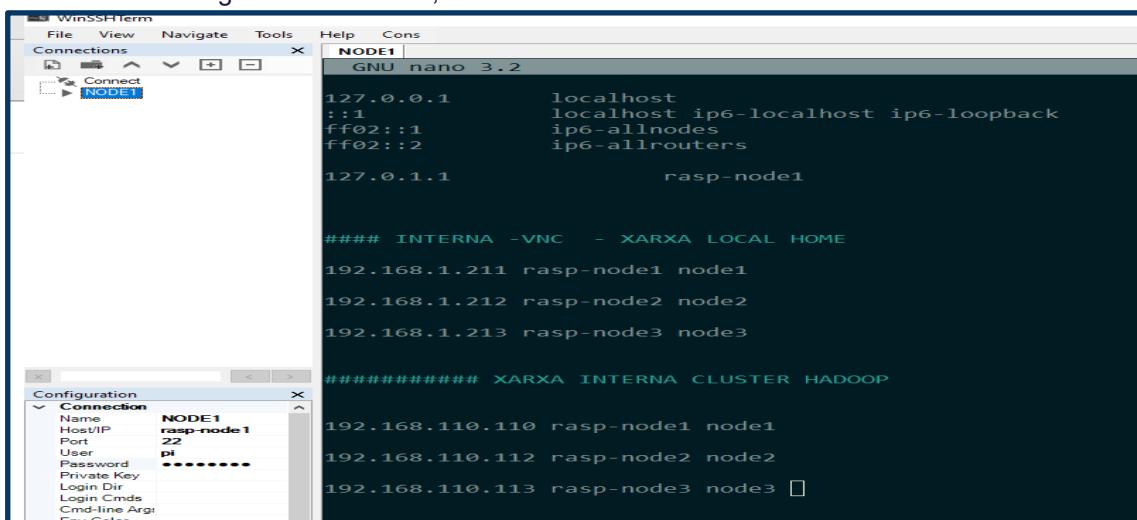


Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

4. Annex: Raspberry Pi Desktop (testing)

4.1. Intent amb Raspberri Pi Desktop (Debian 9 i Debian 10)

- ❑ Com que la idea original era fer un clúster d' Hadoop amb 3 Raspberry Pi, però a causa de la COVID-19 no s'ha pogut fer per falta de dispositius, s'ha intentat fer-ho al VirtualBox. S'ha de dir que VirtualBox no permet l'arquitectura ARM , però hi han isos de [Raspberry Pi Desktop\(Debian 9\)](#) i [Raspberry Pi Desktop\(Debian 10 \)](#) que són un Debian amb la interfície gràfica de la Raspberry Pi (amb algunes limitacions).
 - Es volia fer un clúster de Hadoop utilitzant Ambari i un nou sistema operatiu.
- ❑ Com s'ha fet anteriorment. Després d'instal·lar una màquina, s'ha clonat diversos cops.
- ❑ S'ha configurat el hostname, la xarxa i el SSH.





Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

- Hem afegit les 3 màquines a WinSSHTerm per podent-se connectar via SSH

```

WinSSHTerm
File View Navigate Tools Help Cons
Connections X NODE1 | NODE2 | NODE3
  Connect
    NODE1
    NODE2
    NODE3
Using username "pi".
Linux rasp-node3 4.19.0-9-amd64 #1 SMP Debian 4.19.118-2 (2020-04-29) 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.
Last login: Sun May 17 09:51:59 2020
pi@rasp-node3:~ $ 

```

- Respecte a SSH, en les distribucions Debian per defecte si volem accedir SSH amb l'usuari "root" està deshabilitat. Després de la modificació, s'ha de reiniciar el servei SSH.

root@rasp-node1 : ~/ssh# nano /etc/ssh/sshd_config

```

File View Navigate Tools Help Cons
Connections X NODE1 | NODE2 | NODE3
  Connect
    NODE1
    NODE2
    NODE3
GNU nano 3.2
/etc/ssh/sshd_config
#       $OpenBSD: sshd_config,v 1.103 2018/04/09 20:41:22 tj Exp $
# This is the sshd server system-wide configuration file. See
# sshd_config(5) for more information.

# This sshd was compiled with PATH=/usr/bin:/bin:/usr/sbin:/sbin

# The strategy used for options in the default sshd_config shipped with
# OpenSSH is to specify options with their default value where
# possible, but leave them commented. Uncommented options override the
# default value.

#Port 22
#AddressFamily any
#ListenAddress 0.0.0.0
#ListenAddress ::

#HostKey /etc/ssh/ssh_host_rsa_key
#HostKey /etc/ssh/ssh_host_ecdsa_key
#HostKey /etc/ssh/ssh_host_ed25519_key

# Ciphers and keying
#RekeyLimit default none

# Logging
#SyslogFacility AUTH
#LogLevel INFO

# Authentication:

#LoginGraceTime 2m
PermitRootLogin yes
#StrictModes yes
#MaxAuthTries 6
#MaxSessions 10

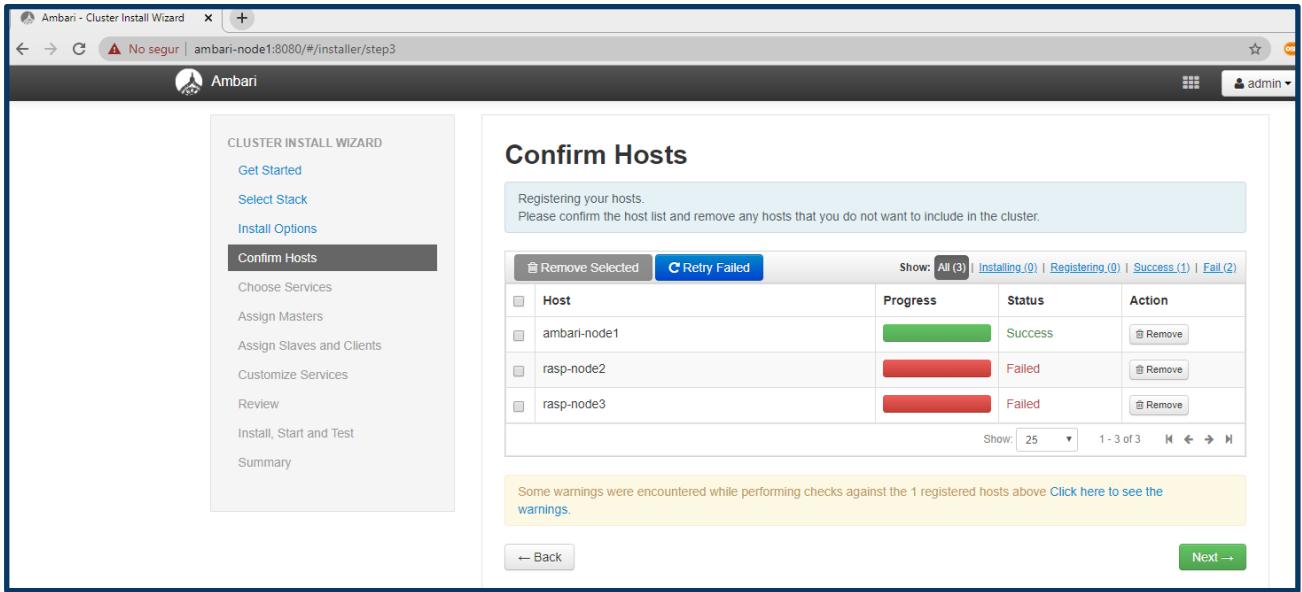
```

- Un cop ja hem creat les claus públiques de cada node com ja hem explicat anteriorment. Ens hem de descarregar els repositoris de Ambari 2.2.1
 - Només està disponible per Debian 7 (s'ha intentat amb aquesta i versions d'Ambari superior sense èxit, donava error.)

Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

4.2. Intent de S.O. hibrid (master :CentOS 6 i 2 slaves: Raspberri Pi Desktop)

- ❑ S'ha provat una altra alternativa, utilitzar el node master on hi ha el servidor d' Ambari que anteriorment s'havia configurat amb un CentOS 6 i fer que els slaves siguin Raspberri Pi Desktop. Però al moment de confirmar els hosts, ha donat error, ja que troba sistemes operatius diferents.
- ❑ S'ha intentat sense èxit instal·lar els agents d'Ambari de forma manual



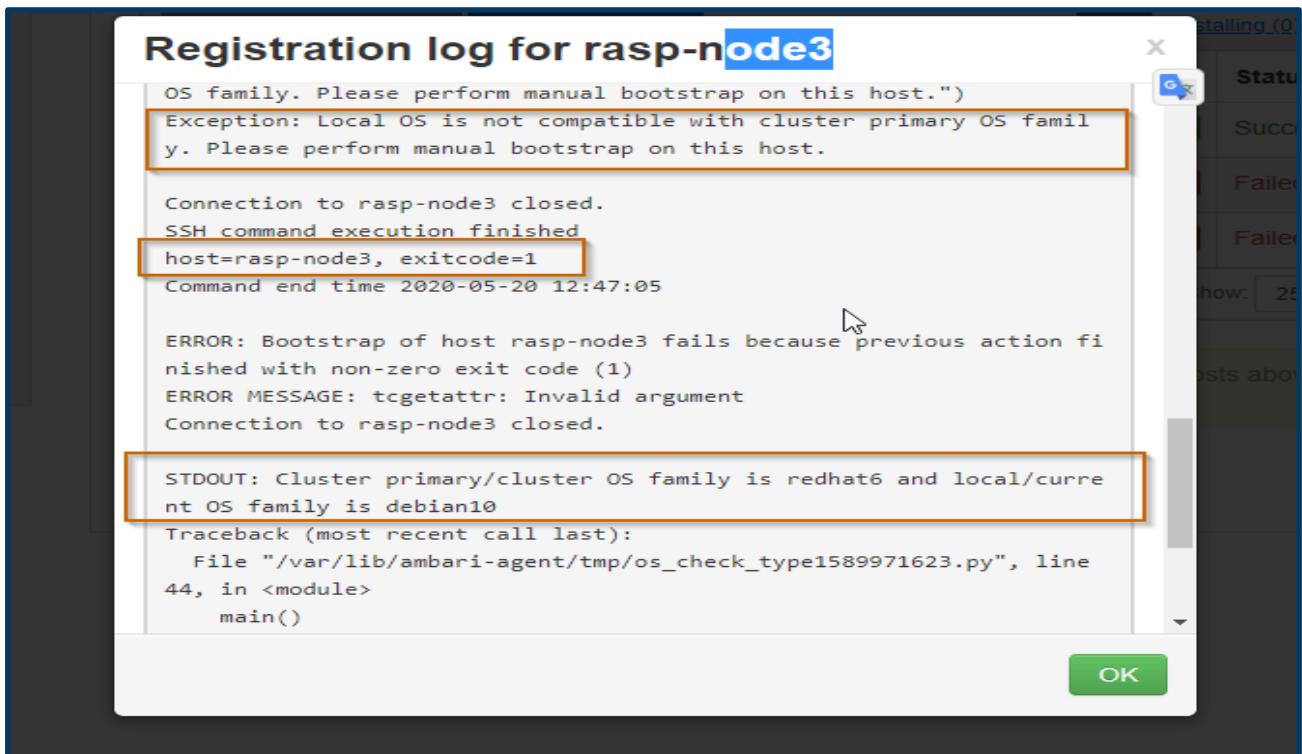
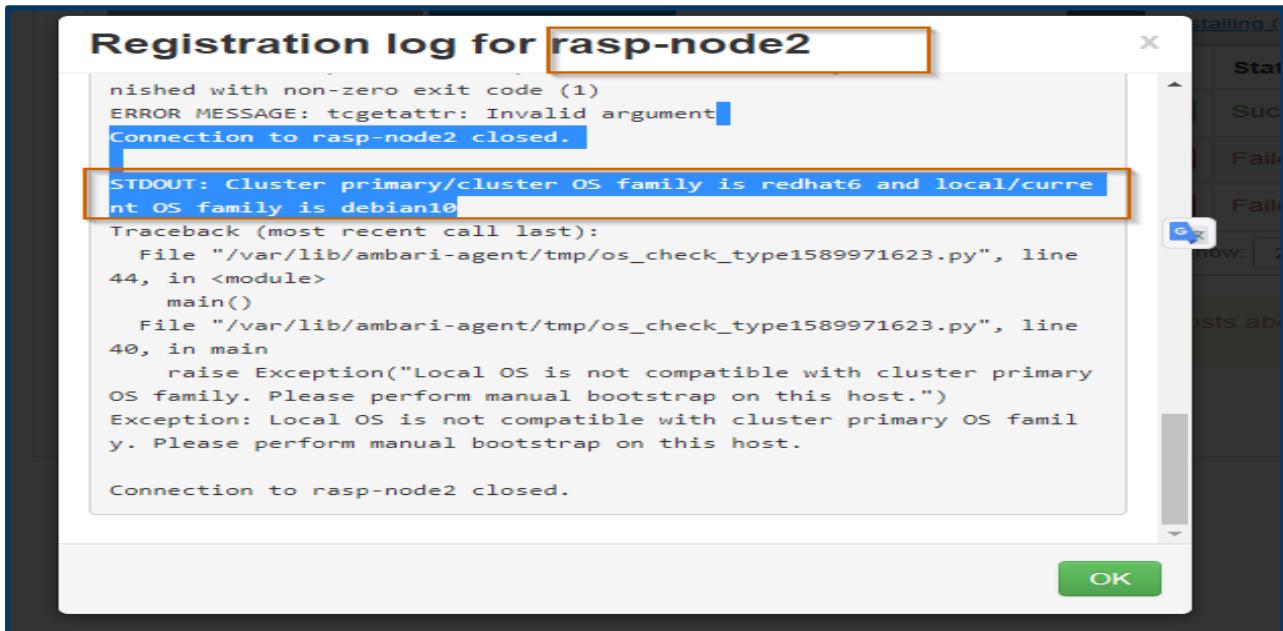
The screenshot shows the Ambari Cluster Install Wizard interface. The left sidebar lists steps: Get Started, Select Stack, Install Options, Confirm Hosts (selected), Choose Services, Assign Masters, Assign Slaves and Clients, Customize Services, Review, Install, Start and Test, and Summary. The main panel title is "Confirm Hosts". It says "Registering your hosts. Please confirm the host list and remove any hosts that you do not want to include in the cluster." Below is a table:

Host	Progress	Status	Action
ambari-node1	<div style="width: 100%;"> </div>	Success	Remove
rasp-node2	<div style="width: 0%; background-color: red;"> </div>	Failed	Remove
rasp-node3	<div style="width: 0%; background-color: red;"> </div>	Failed	Remove

At the bottom, a yellow box says: "Some warnings were encountered while performing checks against the 1 registered hosts above [Click here to see the warnings.](#)". Navigation buttons "Back" and "Next →" are at the bottom right.



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020





Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

5. Conclusions finals

Per finalitzar aquest projecte, podem confirmar que tot i les alternatives del confinament del Covid-19 per poder fer viable el projecte s'han pogut assolir els objectius de crear un clúster Hadoop de diversos nodes. S'ha de reconèixer que ha sigut un tema molt apassionant i enriquidor.

- Hem començat amb el clúster pseudodistribuit (1 node) per començar a practicar i tenir una mínima configuració, ja que després faríem clonacions. S'han fet diversos exemples explicatius:
 - HDFS
 - HDFS Snapshot
 - Diversos processos Mapreduce
- S'ha creat un cluster Hadoop (1 master i 2 slaves) .S'han fet diversos exemples explicatius:
 - Llençar diversos processos Mapreduce contra el clúster
 - Llençar un job amb una nova configuració de Yarn Schelduler
 - Un exemple de Spark-Shell
 - Un exemple de Pyspark
 - Un exemple de Spark-Shell interactuant amb HDFS
- S'ha creat de 0 un altre cluster utilitzant Ambari
 - S'ha configurat Hive pel seu funcionament
 - Visualització de Ambari sobre Windows 10
- Finalment s'ha volgut fer proves amb Raspberri Pi Desktop amb la finalitat de crear un clúster Hadoop amb Ambari però ha donat molts problemes de repositoris, versions i falta de dependències que per falta de temps s'ha volgut deixar-ho.



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

6. Bibliografia

- **Clúster Hadoop en Centos**
 - <https://www.tecmint.com/install-hadoop-multinode-cluster-in-centos/>
 - <https://tecadmin.net/set-up-hadoop-multi-node-cluster-on-centos-redhat/>
 - <https://jmchung.github.io/post/running-hadoop-on-centos-6-multi-node-cluster/>
 - <https://medium.com/@tiago.slucas/set-up-hadoop-cluster-on-virtualbox-machines-running-centos-7-df9121b1eab0>
- **Clúster Hadoop utilitzant Apache Ambari**
 - https://www.alibabacloud.com/blog/set-up-a-hadoop-cluster-with-apache-ambari_595722
 - <https://hadoopjournal.com/2015/08/09/hortonworks-hadoop-installation-using-apache-ambari-on-centos6/>
 - <https://community.cloudera.com/t5/Support-Questions/App-Timeline-Server-not-start/td-p/230460>
 - <https://community.cloudera.com/t5/Support-Questions/HDFS020-Could-not-write-file-user-admin-hive-jobs-hive-job/td-p/153109>
 - https://docs.cloudera.com/HDPDocuments/Ambari-2.2.0.0/bk_ambari_views_guide/content/_configuring_your_cluster_for_files_view.html
- **Instal.lar Hadoop a Ubuntu 18.04 Bionic Beaver**
 - <https://linuxconfig.org/how-to-install-hadoop-on-ubuntu-18-04-bionic-beaver-linux>
- **Editor de Text (Vi, Gedit)**
 - <https://docs.oracle.com/cd/E19620-01/805-7644/x-5le2/index.html>
 - <https://askubuntu.com/questions/52523/errors-shown-in-terminal-when-editing-files-with-gedit>
 - <https://ajpdsoft.com/modules.php?name=Foros&file=viewtopic&t=666>
- **CentOs6- canvi de versió de Python**
 - <http://www.ghanshammahajan.com/how-to-upgrade-python-2-6-to-2-7-on-centos-6/>
- **Raspberry Pi Desktop**
 - <http://www.aoakley.com/articles/2017-07-04-raspbian-x86-virtualbox.php>
 - <https://github.com/superjamie/lazyweb/wiki/Raspberry-Pi-Debian-Backports>
 - <https://www.raspberrypi.org/documentation/remote-access/vnc/>
 - <https://www.osboxes.org/raspbian/>



Nom i Cognoms	Data
Arnau Subirós Puigarnau	02-06-2020

Agraïments:

M'agradaria expressar tot el meu agraïment al meu professor responsable, el Sr. Sergi Grau per tot l'acompanyament i suport rebut durant el transcurs del projecte.