



Proyectos

Implementación de Clúster Hadoop

Versión 2

Historial de Revisiones

Seguimiento de revisiones		
 Revisor	 Estado	 Notas
Ernesto Rios	Aprobada ▾	Versión 1 - Octubre 2023
Ernesto Rios	Aprobada ▾	Actualización de versión de Python, Spark y Jupyter Notebook

Introducción

Big Data es la convergencia de grandes cantidades de datos, tanto estructurados, semiestructurados y no estructurados. Big Data es una filosofía que podemos aplicar con productos como Hadoop.

Dado que las tecnologías tradicionales no pueden hacer frente a esta gran cantidad de información, es necesario utilizar nuevas estrategias, es necesario pasar del almacenamiento y procesamiento de datos en grandes servidores a su equivalente en entornos distribuidos.

Hadoop es un entorno distribuido de datos y procesos. Hadoop implementa procesamiento en paralelo a través de nodos de datos en un sistema de archivos distribuidos. Nodos maestros y nodos workers.

Uno de los puntos fuertes de Hadoop es que está diseñado para ejecutarse en servidores de bajo costo y que dispone de una gran tolerancia a fallos.

Hadoop es un entorno que suministra librerías open source para la computación distribuida. Está diseñado para escalar desde unos pocos nodos a miles de máquinas, cada una de ellas ofreciendo la lógica de negocio y el almacenamiento a nivel local.

El core de Hadoop está conformado por dos componentes básicos:

- Datos
- Procesamiento

Datos: HDFS (Hadoop Distributed File System) es un sistema de almacenamiento tolerante a fallos que puede almacenar gran cantidad de datos, escalar de forma incremental y sobrevivir a fallos de hardware sin perder datos. Si uno falla, el clúster puede continuar trabajando sin perder datos o sin

interrumpir el trabajo, sencillamente redistribuye el trabajo entre los nodos restantes del clúster.

Procesos: Map Reduce V1 y Map Reduce V2 - YARN. De forma general son algoritmos de procesamiento de datos que implementan procesos en paralelo. Es decir, distribuye las tareas a través de los nodos de un clúster.

Existen distintas empresas que ofrecen soluciones empaquetadas para Hadoop, conocidas como distribuciones Hadoop, las cuales están disponibles desde máquinas virtuales pre hechas o mediante soluciones en la nube. Entre las más importantes están: Cloudera, Hortonworks, IBM Open Platform, entre otros.

Sin embargo, mediante el presente proyecto vamos a implementar un muy pequeño entorno de Big Data con Hadoop. Vamos a utilizar la informática distribuida, vamos a distribuir datos y procesos.

Propósito del Proyecto

Este proyecto ***“Implementación de Clúster Hadoop”*** tiene como propósito de contar con un pequeño entorno de laboratorio personal de prácticas para el almacenamiento y el procesamiento distribuido de datos.

Al tratarse de un clúster muy pequeño, sólo se pretende aplicar y explorar los conceptos teóricos y tecnologías alrededor de Hadoop y Big Data, con fines prácticos y académicos, a fin de aprender y entender su funcionamiento y forma de trabajo.

Este clúster debe implementar, principalmente, el trabajo con el sistema de archivos distribuido de Hadoop (HDFS), Hive, Spark, entre otras tecnologías adyacentes.

Descripción de la Infraestructura

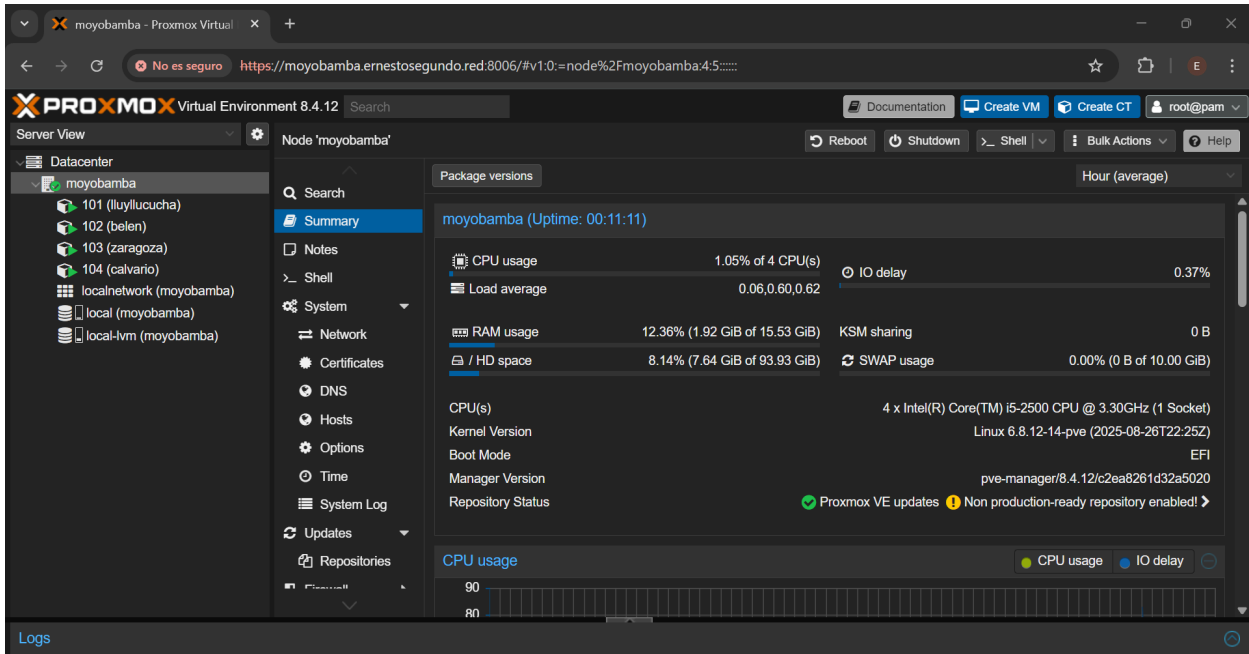
Como mencionamos, Hadoop está diseñado para ejecutarse en servidores de bajo costo y, dado el propósito del proyecto, se cuenta con el siguiente hardware para su implementación:

Hardware	
CPU	4 x Intel(R) Core(TM) i5-2500 CPU @ 3.30GHz (1 Socket)
RAM	16 GB
Disco	500 GB

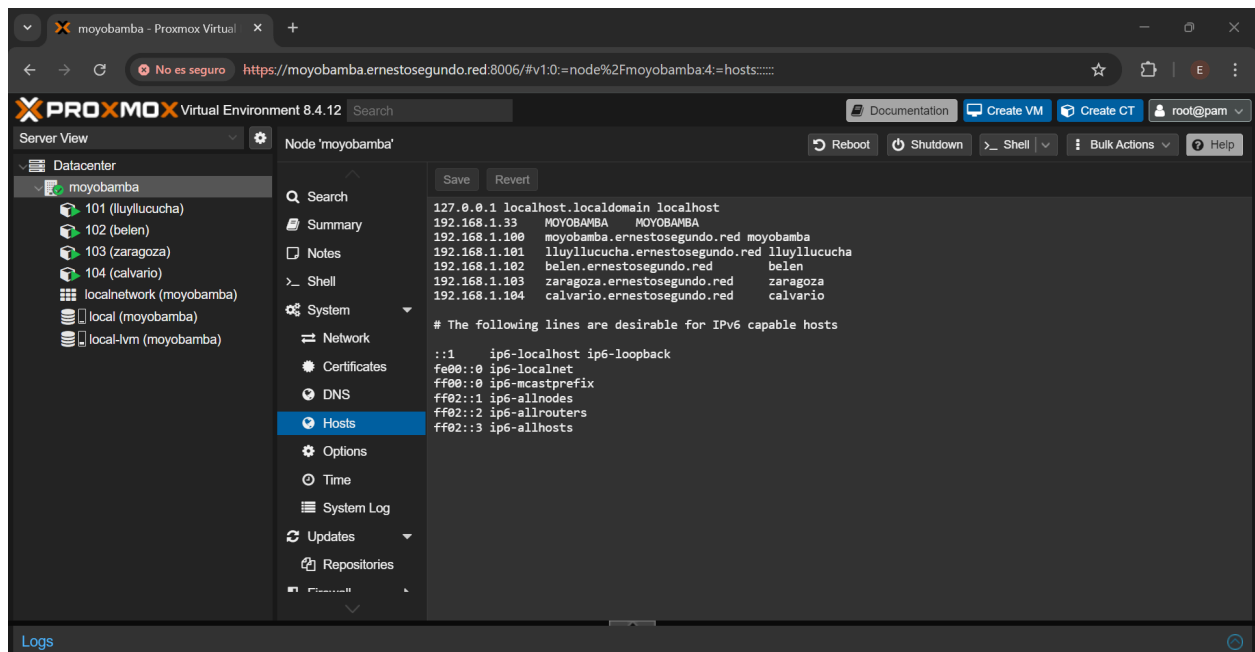
En cuanto al software utilizado, se implementaron los servidores usando la plataforma de virtualización Proxmox VE. Proxmox VE es una plataforma de código abierto, basada en Debian, para la administración de servidores virtuales.

Software	
Plataforma de Virtualización	Proxmox VE 8.4.12
Tecnología	Linux Containers
Kernel Version	Linux 6.8.12-14-pve

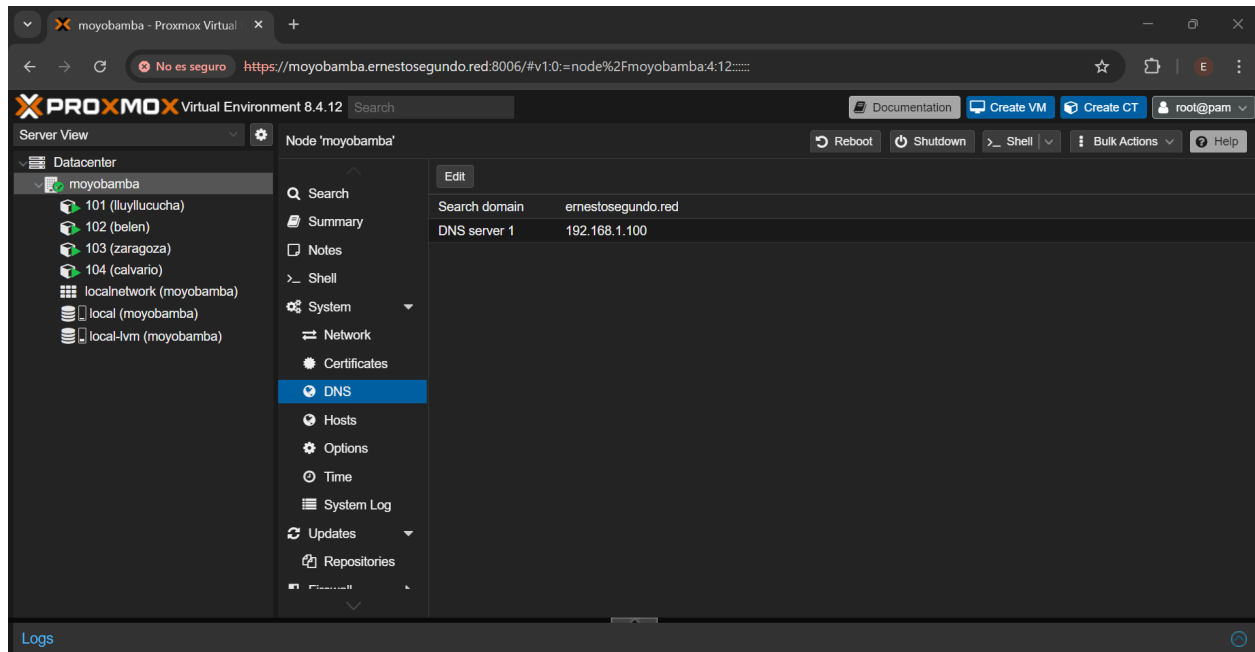
Proxmox VE integra estrechamente el hipervisor KVM, los Linux Containers (LXC), el almacenamiento definido por software y la funcionalidad de red en una sola plataforma, y gestiona fácilmente todos los componentes en la interfaz gráfica de gestión web.



La infraestructura de servidores se encuentra en el nodo Proxmox con nombre de host moyobamba, que alberga los LXC que conforman el clúster Hadoop: lluyllucucha, belen, zaragoza y calvario.



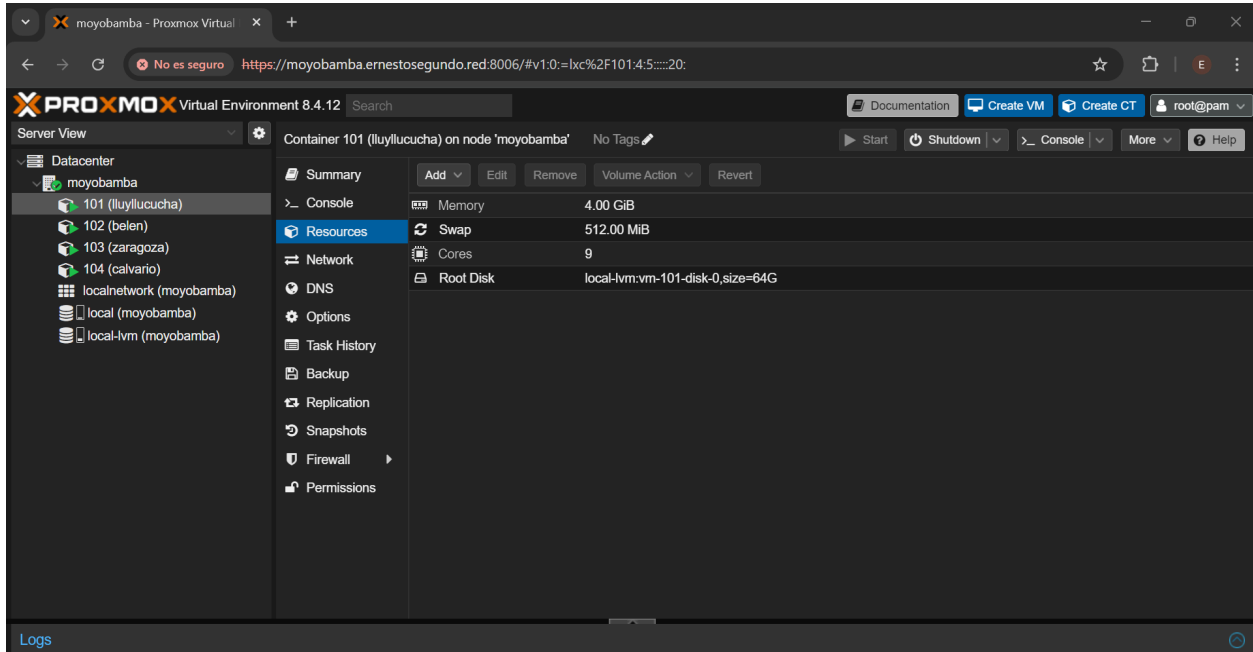
Este nodo Proxmox también es un servidor DNS para el dominio ernestosegundo.red (192.168.1.100/24).



Es decir, Proxmox alberga los cuatro servidores que conforman la infraestructura para la implementación del clúster Hadoop, los cuales se encuentran implementados mediante Linux Containers (LXC):

Nombre de Host	Dirección IP
lluyllucucha.ernestosegundo.red	192.168.1.101/24
belen.ernestosegundo.red	192.168.1.102/24
zaragoza.ernestosegundo.red	192.168.1.103/24
calvario.ernestosegundo.red	192.168.1.104/24

Servidor lluyllucucha.ernestosegundo.red



Proxmox Virtual Environment 8.4.12

Container 101 (lluyllucucha) on node 'moyobamba' No Tags

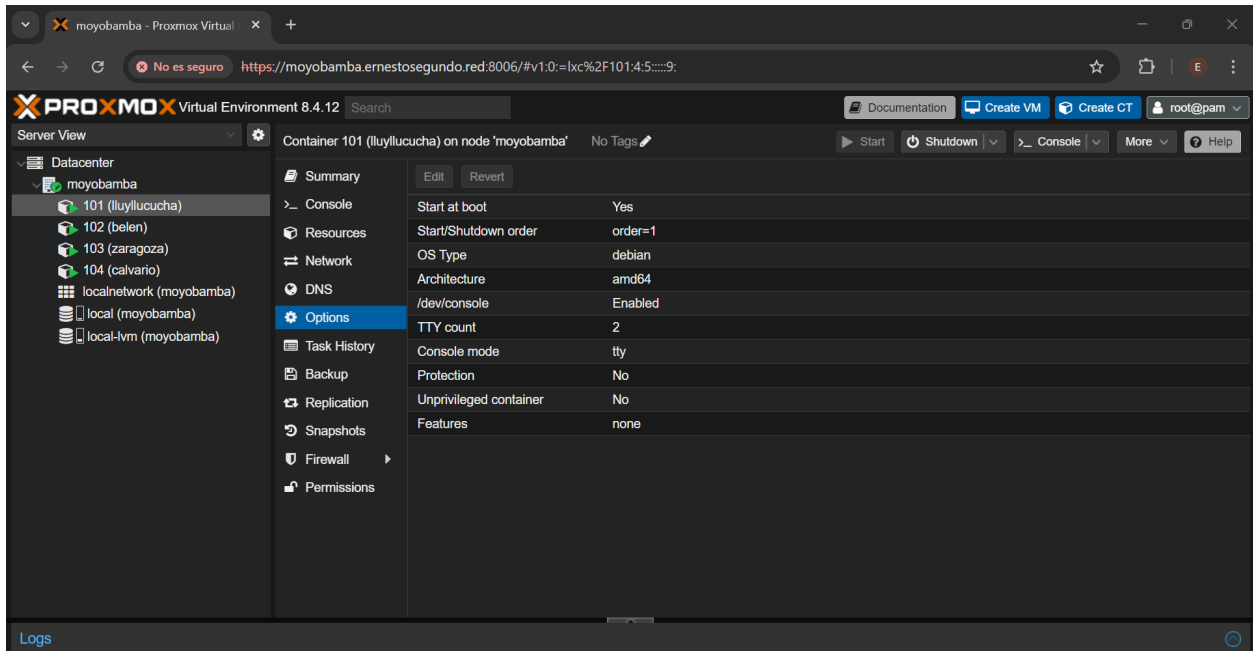
Start Shutdown Console More Help

Summary

Resources

Memory	4.00 GiB
Swap	512.00 MiB
Cores	9
Root Disk	local-lvm:vm-101-disk-0,size=64G

Logs



Proxmox Virtual Environment 8.4.12

Container 101 (lluyllucucha) on node 'moyobamba' No Tags

Start Shutdown Console More Help

Summary

Options

Start at boot	Yes
Start/Shutdown order	order=1
OS Type	debian
Architecture	amd64
/dev/console	Enabled
TTY count	2
Console mode	tty
Protection	No
Unprivileged container	No
Features	none

Logs

Servidor belen.ernestosegundo.red

The screenshot shows the Proxmox Virtual Environment 8.4.12 interface. The left sidebar displays the 'Server View' with a tree structure: Datacenter > moyobamba > 101 (luyllucucha) > 102 (belen). The main panel shows the configuration for 'Container 102 (belen) on node 'moyobamba''. The 'Resources' tab is selected, displaying the following details:

Resource	Value
Memory	4.00 GiB
Swap	512.00 MiB
Cores	2
Root Disk	local-lvm:vm-102-disk-0,size=64G

Other tabs visible include Summary, Console, Network, DNS, Options, Task History, Backup, Replication, Snapshots, Firewall, and Permissions. The top bar shows the user 'root@pam' and various action buttons like Start, Shutdown, Console, and Help.

The screenshot shows the same Proxmox Virtual Environment 8.4.12 interface, but with the 'Options' tab selected for 'Container 102 (belen) on node 'moyobamba''. The 'Options' tab displays the following configuration details:

Option	Value
Start at boot	Yes
Start/Shutdown order	order=2
OS Type	debian
Architecture	amd64
/dev/console	Enabled
TTY count	2
Console mode	tty
Protection	No
Unprivileged container	No
Features	none

The interface also shows the same sidebar and top bar as the previous screenshot.

Servidor zaragoza.ernestosegundo.red

The screenshot shows the Proxmox Virtual Environment 8.4.12 interface. The left sidebar displays the 'Server View' with a tree structure: Datacenter > moyobamba > 103 (zaragoza). The main panel shows the configuration for 'Container 103 (zaragoza) on node 'moyobamba''. The 'Resources' tab is selected, displaying the following details:

Resource	Value
Memory	4.00 GiB
Swap	512.00 MiB
Cores	2
Root Disk	local-lvm:vm-103-disk-0,size=64G

Other tabs visible include Summary, Console, Network, DNS, Options, Task History, Backup, Replication, Snapshots, Firewall, and Permissions. The top bar shows the user 'root@pam' and various action buttons like Start, Shutdown, and Console.

This screenshot shows the same Proxmox interface, but with the 'Options' tab selected for 'Container 103 (zaragoza)'. The configuration details are as follows:

Option	Value
Start at boot	Yes
Start/Shutdown order	order=4
OS Type	debian
Architecture	amd64
/dev/console	Enabled
TTY count	2
Console mode	tty
Protection	No
Unprivileged container	No
Features	none

The interface elements, including the sidebar and top bar, are consistent with the previous screenshot.

Servidor calvario.ernestosegundo.red

The screenshot shows the Proxmox Virtual Environment 8.4.12 interface. The left sidebar displays the 'Server View' with a tree structure under 'Datacenter' > 'moyobamba'. The selected container is '104 (calvario)'. The main panel shows the 'Resources' tab for 'Container 104 (calvario) on node 'moyobamba''. The resources table is as follows:

Resource	Value
Memory	4.00 GiB
Swap	512.00 MiB
Cores	2
Root Disk	local-lvm:vm-104-disk-0,size=64G

At the bottom of the interface, a 'Logs' tab is visible.

The screenshot shows the Proxmox Virtual Environment 8.4.12 interface, similar to the first one, but with the 'Options' tab selected for 'Container 104 (calvario) on node 'moyobamba''. The options table is as follows:

Option	Value
Start at boot	Yes
Start/Shutdown order	order=3
OS Type	debian
Architecture	amd64
/dev/console	Enabled
TTY count	2
Console mode	tty
Protection	No
Unprivileged container	No
Features	none

At the bottom of the interface, a 'Logs' tab is visible.

Descripción del Clúster Hadoop

El presente clúster Hadoop está formado por cuatro servidores, organizados de la siguiente manera:

lluyllucucha.ernestosegundo.red		
192.168.1.101/24	Nodo principal	Namenode
belen.ernestosegundo.red		
192.168.1.102/24	Nodo secundario	Namenode, Datanode
zaragoza.ernestosegundo.red		
192.168.1.103/24	Nodo secundario	Datanode
calvario.ernestosegundo.red		
192.168.1.104/24	Nodo secundario	Datanode

El usuario configurado para todos los nodos del cluster es ernestosegundo.

```
ernestosegundo@lluyllucucha: ~  
login as: ernestosegundo  
ernestosegundo@lluyllucucha.ernestosegundo.red's password:  
Linux lluyllucucha 6.8.12-14-pve #1 SMP PREEMPT_DYNAMIC PMX 6.8.12-14 (2025-08-26T22:25Z) 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: Fri Sep 5 23:36:30 2025 from 192.168.1.33  
ernestosegundo@lluyllucucha:~$
```

```
ernestosegundo@belen: ~  
login as: ernestosegundo  
ernestosegundo@belen.ernestosegundo.red's password:  
Linux belen 6.8.12-14-pve #1 SMP PREEMPT_DYNAMIC PMX 6.8.12-14 (2025-08-26T22:25Z) 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: Fri Sep 5 15:35:27 2025  
ernestosegundo@belen:~$
```

```
ernestosegundo@zaragoza: ~  
login as: ernestosegundo  
ernestosegundo@zaragoza.ernestosegundo.red's password:  
Linux zaragoza 6.8.12-14-pve #1 SMP PREEMPT_DYNAMIC PMX 6.8.12-14 (2025-08-26T22:25Z) 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: Fri Sep 5 15:36:04 2025  
ernestosegundo@zaragoza:~$
```

```
ernestosegundo@calvario: ~  
login as: ernestosegundo  
ernestosegundo@calvario.ernestosegundo.red's password:  
Linux calvario 6.8.12-14-pve #1 SMP PREEMPT_DYNAMIC PMX 6.8.12-14 (2025-08-26T22:25Z) 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: Fri Sep 5 15:36:35 2025  
ernestosegundo@calvario:~$
```

Del mismo modo, se realizó la configuración para la conectividad vía SSH entre los cuatro nodos.

```
ernestosegundo@lluyllucucha: ~  
ernestosegundo@lluyllucucha:~$ ls -lia .ssh/  
total 24  
2883587 drwx----- 2 ernestosegundo ernestosegundo 4096 Nov 28 2024 .  
2883586 drwxr-xr-x 19 ernestosegundo ernestosegundo 4096 Sep 7 08:30 ..  
2928008 -rw----- 1 ernestosegundo ernestosegundo 3063 Nov 28 2024 authorized_keys  
2883593 -rw----- 1 ernestosegundo ernestosegundo 2622 Dec 12 2022 id_rsa  
2883596 -rw-r--r-- 1 ernestosegundo ernestosegundo 581 Dec 12 2022 id_rsa.pub  
2883598 -rw-r--r-- 1 ernestosegundo ernestosegundo 3330 Nov 28 2024 known_hosts  
ernestosegundo@lluyllucucha:~$
```

```
ernestosegundo@belen: ~  
ernestosegundo@belen:~$ ls -lia .ssh/  
total 24  
1048583 drwxr-xr-x 2 ernestosegundo ernestosegundo 4096 Nov 28 2024 .  
1048578 drwxr-xr-x 10 ernestosegundo ernestosegundo 4096 Jan 8 2025 ..  
1048584 -rw----- 1 ernestosegundo ernestosegundo 3063 Nov 28 2024 authorized_keys  
1048585 -rw----- 1 ernestosegundo ernestosegundo 2610 Dec 15 2022 id_rsa  
1048586 -rw-r--r-- 1 ernestosegundo ernestosegundo 574 Dec 15 2022 id_rsa.pub  
1048587 -rw-r--r-- 1 ernestosegundo ernestosegundo 1776 Feb 8 2023 known_hosts  
ernestosegundo@belen:~$
```



```
ernestosegundo@zaragoza:~$ ls -lia .ssh/
total 24
3276807 drwxr-xr-x  2 ernestosegundo ernestosegundo 4096 Nov 28  2024 .
3276802 drwxr-xr-x 10 ernestosegundo ernestosegundo 4096 Jan  8  2025 ..
3276808 -rw-----  1 ernestosegundo ernestosegundo 3063 Nov 28  2024 authorized_keys
3276809 -rw-----  1 ernestosegundo ernestosegundo 2610 Dec 15  2022 id_rsa
3276810 -rw-r--r--  1 ernestosegundo ernestosegundo  577 Dec 15  2022 id_rsa.pub
3276811 -rw-r--r--  1 ernestosegundo ernestosegundo 1332 Dec 17  2022 known_hosts
ernestosegundo@zaragoza:~$ █
```

```
ernestosegundo@calvario:~$ ls -lia .ssh/
total 24
786439 drwxr-xr-x  2 ernestosegundo ernestosegundo 4096 Nov 28  2024 .
786434 drwxr-xr-x 10 ernestosegundo ernestosegundo 4096 Jan  8  2025 ..
786442 -rw-----  1 ernestosegundo ernestosegundo 3063 Nov 28  2024 authorized_keys
786440 -rw-----  1 ernestosegundo ernestosegundo 2610 Dec 15  2022 id_rsa
786441 -rw-r--r--  1 ernestosegundo ernestosegundo  577 Dec 15  2022 id_rsa.pub
786443 -rw-r--r--  1 ernestosegundo ernestosegundo 1332 Dec 15  2022 known_hosts
ernestosegundo@calvario:~$ █
```

Como se sabe, Hadoop necesita de Java para su funcionamiento por lo que se usa la versión 8 de este software.

El clúster implementado considera la inclusión de las siguientes tecnologías adyacentes a Hadoop:

- Hive: Para acceder a HDFS con comandos parecidos a SQL (HiveSQL)
- Hue: Interfaz web que permite interactuar con Hive
- HBase: Para el almacenamiento no relacional para Hadoop
- Sqoop: Para transferir grandes volúmenes de datos de manera eficiente entre Hadoop y los gestores de bases de datos relacionales

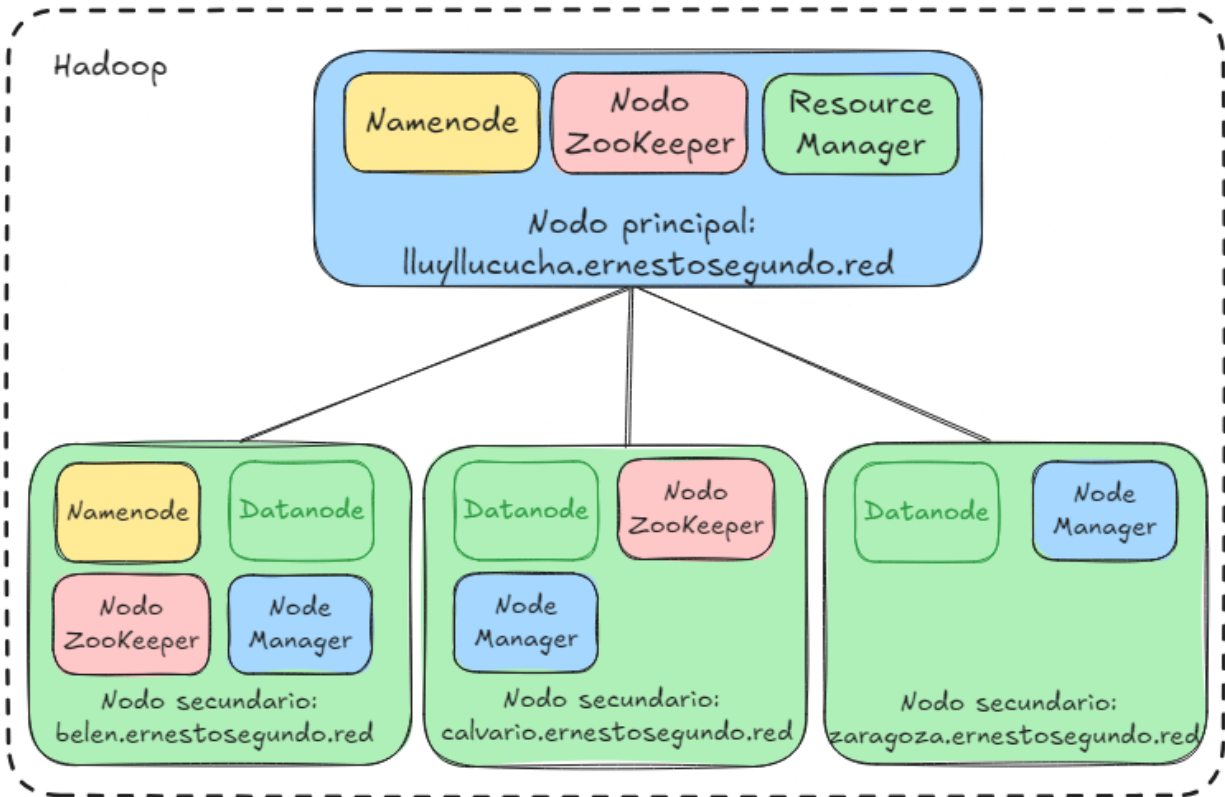
- ZooKeeper: Para mantener alta disponibilidad del clúster Hadoop
- Spark: Para el procesamiento de datos en memoria

Con este conjunto de tecnologías de tratamiento de datos, se tiene lo siguiente:

Tecnología	Versión
Apache Hadoop	3.3.4
Apache Spark	3.5.7
Apache Hive	3.1.3
Hue	4.11.0
Sqoop	1.4.7
HBase	2.5.3
ZooKeeper	3.7.1
Java Development Kit	1.8.0_361
Python	3.13.5

ZooKeeper es la herramienta, el componente con el cual vamos a poner el clúster en alta disponibilidad. Los nodos ZooKeeper están configurados de la siguiente manera:

llyllucucha.ernestosegundo.red		
192.168.1.101/24	Nodo principal Nodo ZooKeeper	Namenode
belen.ernestosegundo.red		
192.168.1.102/24	Nodo secundario Nodo ZooKeeper	Namenode, Datanode
zaragoza.ernestosegundo.red		
192.168.1.103/24	Nodo secundario	Datanode
calvario.ernestosegundo.red		
192.168.1.104/24	Nodo secundario Nodo ZooKeeper	Datanode



Verificación de ejecución de procesos en el nodo principal: lluyllucucha.ernestosegundo.red

```
ernestosegundo@lluyllucucha: ~  
ernestosegundo@lluyllucucha:~$ jps  
753 QuorumPeerMain  
1267 DFSZKFailoverController  
1525 JobHistoryServer  
919 NameNode  
1383 ResourceManager  
1128 JournalNode  
1561 RunJar  
5071 Jps  
ernestosegundo@lluyllucucha:~$
```

Verificación de ejecución de procesos en el nodo secundario: belen.ernestosegundo.red

```
ernestosegundo@belen: ~  
ernestosegundo@belen:~$ jps  
737 DataNode  
594 QuorumPeerMain  
792 JournalNode  
921 NodeManager  
681 NameNode  
1004 JobHistoryServer  
4846 Jps  
847 DFSZKFailoverController  
ernestosegundo@belen:~$
```

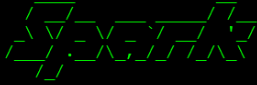
Verificación de ejecución de procesos en el nodo secundario: calvario.ernestosegundo.red

```
ernestosegundo@calvario: ~  
ernestosegundo@calvario:~$ jps  
1954 Jps  
724 JournalNode  
668 DataNode  
781 NodeManager  
861 JobHistoryServer  
574 QuorumPeerMain  
ernestosegundo@calvario:~$
```

Verificación de ejecución de procesos en el nodo secundario: zaragoza.ernestosegundo.red

```
ernestosegundo@zaragoza: ~  
ernestosegundo@zaragoza:~$ jps  
576 DataNode  
634 NodeManager  
715 JobHistoryServer  
1356 Jps  
ernestosegundo@zaragoza:~$
```

spark-shell

```
ernestosegundo@lluyllucucha: ~  
ernestosegundo@lluyllucucha:~$ spark-shell  
Setting default log level to "WARN".  
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).  
Setting Spark log level to "ERROR".  
Spark context Web UI available at http://lluyllucucha.ernestosegundo.red:4040  
Spark context available as 'sc' (master = local[*], app id = local-1759880195292).  
Spark session available as 'spark'.  
Welcome to  
 version 3.5.7  
  
Using Scala version 2.12.18 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_361)  
Type in expressions to have them evaluated.  
Type :help for more information.  
  
scala> █
```

pyspark

```
ernestosegundo@lluyllucucha: ~  
ernestosegundo@lluyllucucha:~$ pyspark  
Python 3.13.5 | packaged by Anaconda, Inc. | (main, Jun 12 2025, 16:09:02) [GCC 11.2.0] on linux  
Type "help", "copyright", "credits" or "license" for more information.  
Setting default log level to "WARN".  
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).  
25/10/07 18:38:19 WARN HiveConf: HiveConf of name hive.metastore.wm.default.pool.size does not exist  
25/10/07 18:38:19 WARN HiveConf: HiveConf of name hive.llap.task.scheduler.preempt.independent does not exist  
25/10/07 18:38:19 WARN HiveConf: HiveConf of name hive.llap.output.format.arrow does not exist  
25/10/07 18:38:19 WARN HiveConf: HiveConf of name hive.tez.llap.min.reducer.per.executor does not exist  
25/10/07 18:38:19 WARN HiveConf: HiveConf of name hive.arrow.root allocator.limit does not exist  
25/10/07 18:38:19 WARN HiveConf: HiveConf of name hive.vectorized.use.checked.expressions does not exist  
25/10/07 18:38:19 WARN HiveConf: HiveConf of name hive.tez.dynamic.semijoin.reduction.for.mapjoin does not exist  
25/10/07 18:38:19 WARN HiveConf: HiveConf of name hive.vectorized.complex.types.enabled does not exist  
25/10/07 18:38:19 WARN HiveConf: HiveConf of name hive.server2.wm.worker.threads does not exist  
25/10/07 18:38:19 WARN HiveConf: HiveConf of name hive.repl.partitions.dump.parallelism does not exist  
25/10/07 18:38:19 WARN HiveConf: HiveConf of name hive.metastore.uri.selection does not exist  
25/10/07 18:38:19 WARN HiveConf: HiveConf of name hive.strict.checks.no.partition.filter does not exist  
25/10/07 18:38:19 WARN HiveConf: HiveConf of name hive.tez.dynamic.semijoin.reduction.for.dpp.factor does not exist  
25/10/07 18:38:19 WARN HiveConf: HiveConf of name hive.stats.filter.in.min.ratio does not exist  
25/10/07 18:38:19 WARN HiveConf: HiveConf of name hive.metastore.client.cache.initial.capacity does not exist  
25/10/07 18:38:19 WARN HiveConf: HiveConf of name hive.stats.ndv.estimate.percent does not exist  
25/10/07 18:38:19 WARN HiveConf: HiveConf of name hive.server2.webui.cors.allowed.methods does not exist  
25/10/07 18:38:19 WARN HiveConf: HiveConf of name hive.optimize.joinreducededuplication does not exist  
25/10/07 18:38:19 WARN HiveConf: HiveConf of name hive.metastore.client.cache.enabled does not exist  
25/10/07 18:38:19 WARN HiveConf: HiveConf of name hive.stats.fetch.bitvector does not exist  
25/10/07 18:38:19 WARN HiveConf: HiveConf of name hive.disable.unsafe.external.table.operations does not exist  
25/10/07 18:38:19 WARN HiveConf: HiveConf of name hive.materializedview.rewriting.incremental does not exist  
25/10/07 18:38:19 WARN HiveConf: HiveConf of name hive.server2.materializedviews.registry.impl does not exist  
25/10/07 18:38:19 WARN HiveConf: HiveConf of name hive.metastore.event.db.notification.api.auth does not exist  
25/10/07 18:38:19 WARN HiveConf: HiveConf of name hive.exec.orc.delta.streaming.optimizations.enabled does not exist  
25/10/07 18:38:19 WARN HiveConf: HiveConf of name hive.stats.ndv.algo does not exist  
25/10/07 18:38:19 WARN HiveConf: HiveConf of name hive.spark.job.max.tasks does not exist  
25/10/07 18:38:19 WARN HiveConf: HiveConf of name hive.mack.repair.batch.max.retries does not exist
```

```
ernestosegundo@lluyllucucha: ~
25/10/07 18:38:19 WARN HiveConf: HiveConf of name hive.llap.io.vrb.queue.limit.min does not exist
25/10/07 18:38:19 WARN HiveConf: HiveConf of name hive.server2.wm.pool.metrics does not exist
25/10/07 18:38:19 WARN HiveConf: HiveConf of name hive.repl.add.raw.reserved.namespace does not exist
25/10/07 18:38:19 WARN HiveConf: HiveConf of name hive.resource.use.hdfs.location does not exist
25/10/07 18:38:19 WARN HiveConf: HiveConf of name hive.stats.num.nulls.estimate.percent does not exist
25/10/07 18:38:19 WARN HiveConf: HiveConf of name hive.llap.io.acid does not exist
25/10/07 18:38:19 WARN HiveConf: HiveConf of name hive.llap.zk.sm.session.timeout does not exist
25/10/07 18:38:19 WARN HiveConf: HiveConf of name hive.vectorized.ptf.max.memory.buffering.batch.count does not exist
25/10/07 18:38:19 WARN HiveConf: HiveConf of name hive.llap.task.scheduler.am.registry does not exist
25/10/07 18:38:19 WARN HiveConf: HiveConf of name hive.druid.overlord.address.default does not exist
25/10/07 18:38:19 WARN HiveConf: HiveConf of name hive.optimize.remove.sq_count_check does not exist
25/10/07 18:38:19 WARN HiveConf: HiveConf of name hive.server2.webui.enable.cors does not exist
25/10/07 18:38:19 WARN HiveConf: HiveConf of name hive.vectorized.row.serde.inputformat.excludes does not exist
25/10/07 18:38:19 WARN HiveConf: HiveConf of name hive.query.reexecution.stats.cache.size does not exist
25/10/07 18:38:19 WARN HiveConf: HiveConf of name hive.combine.equivalent.work.optimization does not exist
25/10/07 18:38:19 WARN HiveConf: HiveConf of name hive.lock.query.string.max.length does not exist
25/10/07 18:38:19 WARN HiveConf: HiveConf of name hive.llap.io.track.cache.usage does not exist
25/10/07 18:38:19 WARN HiveConf: HiveConf of name hive.use.orc.codec.pool does not exist
25/10/07 18:38:19 WARN HiveConf: HiveConf of name hive.query.results.cache.max.size does not exist
25/10/07 18:38:19 WARN HiveConf: HiveConf of name hive.repl.bootstrap.dump.open.txn.timeout does not exist
Setting Spark log level to "ERROR".
Welcome to

  ____  _
 / ___|| | | |
| |___| |_| |
|___ \  _  |
   __| | | |
  |___|_|_|_|

 version 3.5.7

Using Python version 3.13.5 (main, Jun 12 2025 16:09:02)
spark context Web UI available at http://lluyllucucha.ernestosegundo.red:4040
Spark context available as 'sc' (master = local[*], app id = local-1759880300048).
SparkSession available as 'spark'.
>>> []
```

hive

```
ernestosegundo@lluyllucucha: ~
ernestosegundo@lluyllucucha:~$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/opt/hadoop/hive/lib/log4j-slf4j-impl-2.17.1.jar!/org/slf4j/impl/StaticLogger
Binder.class]
SLF4J: Found binding in [jar:file:/opt/hadoop/share/hadoop/common/lib/slf4j-reload4j-1.7.36.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]
Hive Session ID = 7da46596-17d4-4394-b002-a3f0c304a25b

Logging initialized using configuration in file:/opt/hadoop/hive/conf/hive-log4j2.properties Async: true
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.
Hive Session ID = b275639c-1559-4c44-98b2-e4388b4fb765
hive> show databases;
OK
airtribu
default
ernxto
ernxtometorfano
hr_db
reTail_db
training
Time taken: 1.275 seconds, Fetched: 7 row(s)
hive> []
```

Web de administración de Hadoop MapReduce

<http://lluyllucucha.ernestosegundo.red:8088/>

moyobamba - Proxmox Virtual

Namenode information

Namenode information

No es seguro

lluyllucucha.ernestosegundo.red:9870/dfshealth.html#tab-overview

E

Hadoop

Overview

Datanodes

Datanode Volume Failures

Snapshot

Startup Progress

Utilities

Overview 'lluyllucucha:9000' (active)

Namespace:	ha-cluster
Namenode ID:	lluyllucucha
Started:	Sun Sep 07 11:04:19 -0500 2025
Version:	3.3.4, ra585a73c3e02ac62350c136643a5e7f6095a3dbb
Compiled:	Fri Jul 29 07:32:00 -0500 2022 by stevel from branch-3.3.4
Cluster ID:	CID-40723c7a-d4ce-49c1-82b8-67e7c12ce216
Block Pool ID:	BP-43467781-192.168.1.101-1676837218964

Summary

Security is off.

Safemode is off.

1459 files and directories, 74 blocks (74 replicated blocks, 0 erasure coded block groups) = 1533 total filesystem object(s).

Heap Memory used 339.96 MB of 595.5 MB Heap Memory. Max Heap Memory is 3.45 GB.

Non Heap Memory used 79.93 MB of 81.94 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.

moyobamba - Proxmox Virtual

Namenode information

Namenode information

No es seguro

belen.ernestosegundo.red:9870/dfshealth.html#tab-overview

E

Hadoop

Overview

Datanodes

Datanode Volume Failures

Snapshot

Startup Progress

Utilities

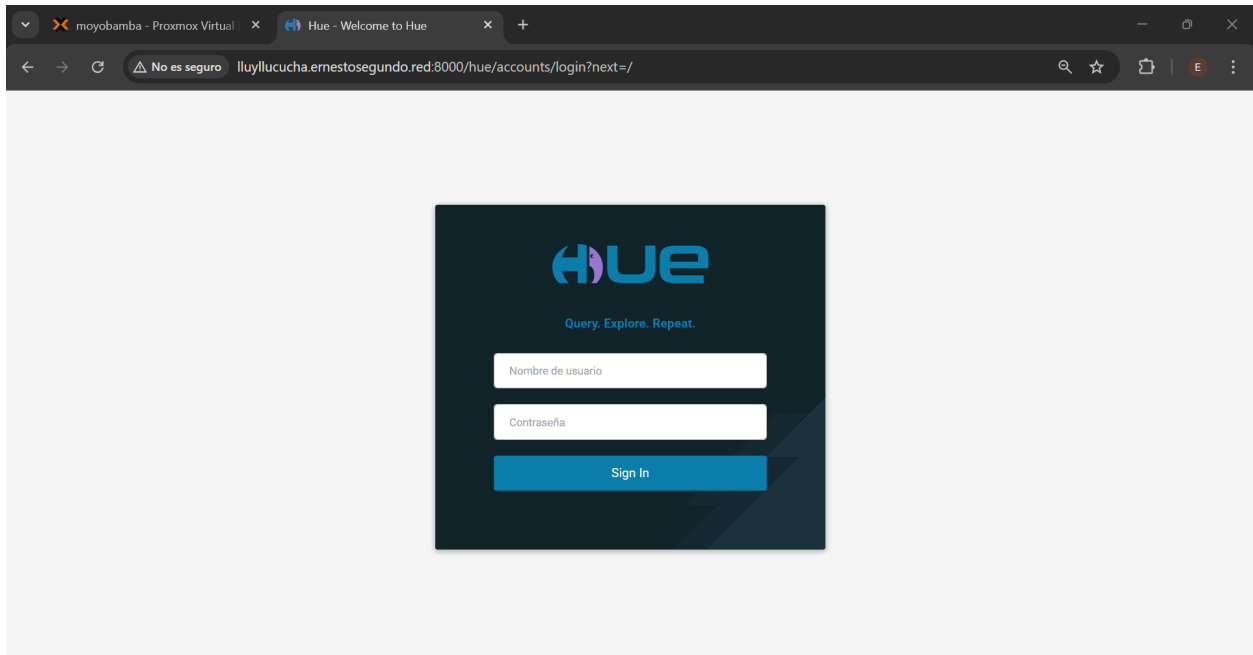
Overview 'belen:9000' (standby)

Namespace:	ha-cluster
Namenode ID:	belen
Started:	Sun Sep 07 11:04:43 -0500 2025
Version:	3.3.4, ra585a73c3e02ac62350c136643a5e7f6095a3dbb
Compiled:	Fri Jul 29 07:32:00 -0500 2022 by stevel from branch-3.3.4
Cluster ID:	CID-40723c7a-d4ce-49c1-82b8-67e7c12ce216
Block Pool ID:	BP-43467781-192.168.1.101-1676837218964

Summary

Security is off.

Hue: Interfaz web que permite interactuar con Hive
<http://lluyllucucha.ernestosegundo.red:8000/hue>



Search saved documents...

0.48s airtribu

```
SELECT *
FROM airtribu.origenes
LIMIT 100;
```

Query History

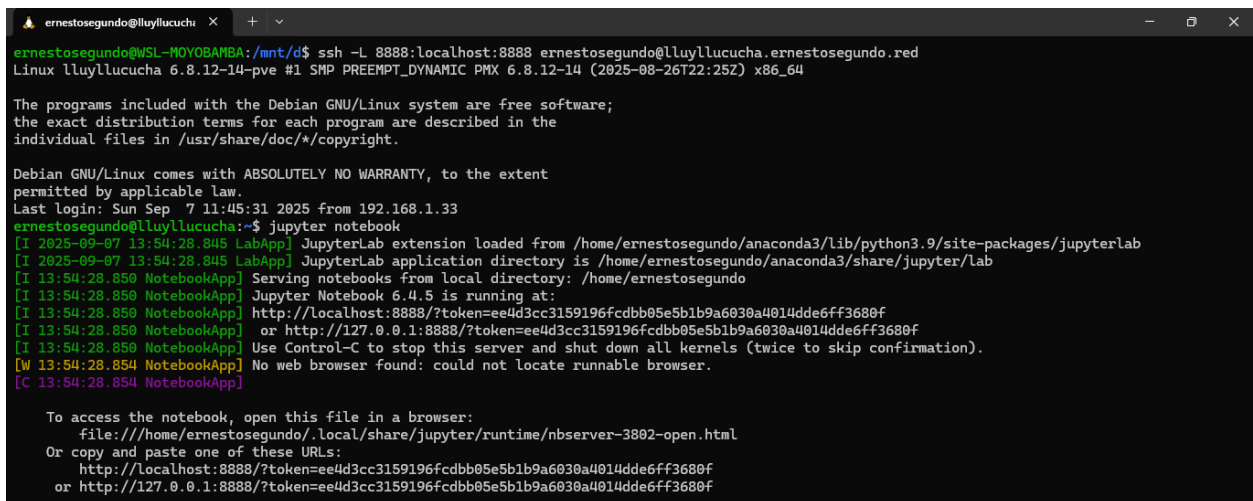
	origenes.origen	origenes.numero_vuelos
1	ARG	5
2	AUS	2
3	BRA	5
4	CAN	7
5	CHI	2
6	CHN	2
7	COL	5

Adicionalmente, se tiene configurado Jupyter Notebook para acceder al clúster Hadoop para interactuar con HDFS, Bash, Python y Spark.

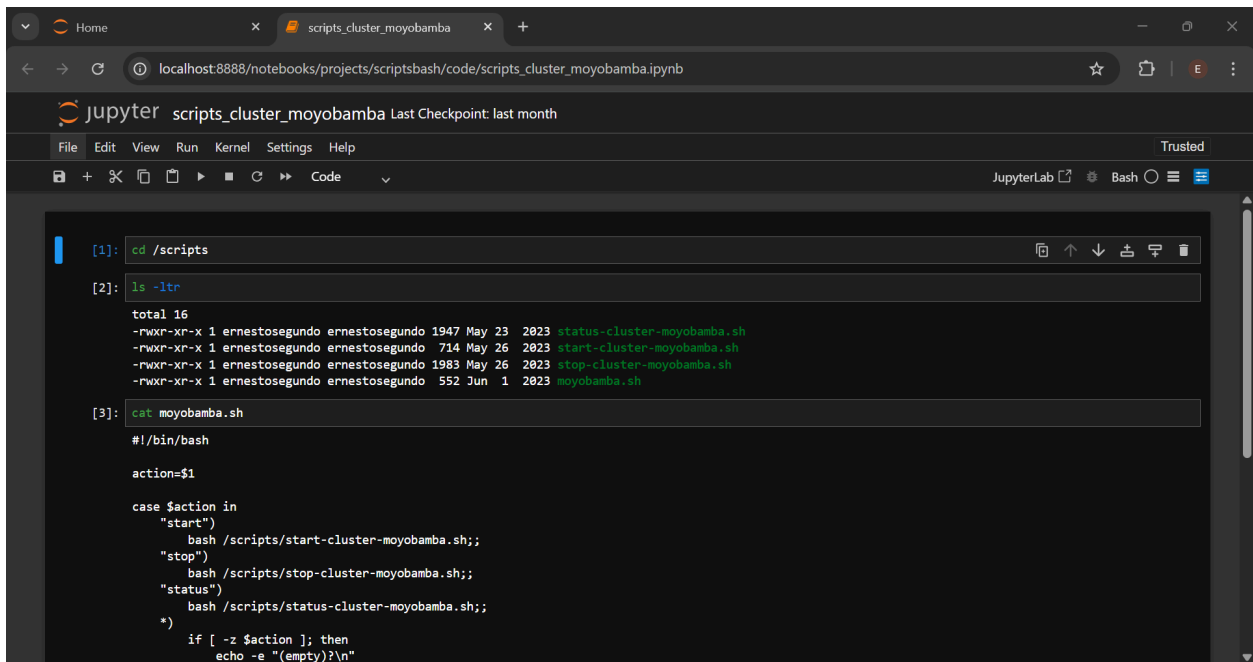
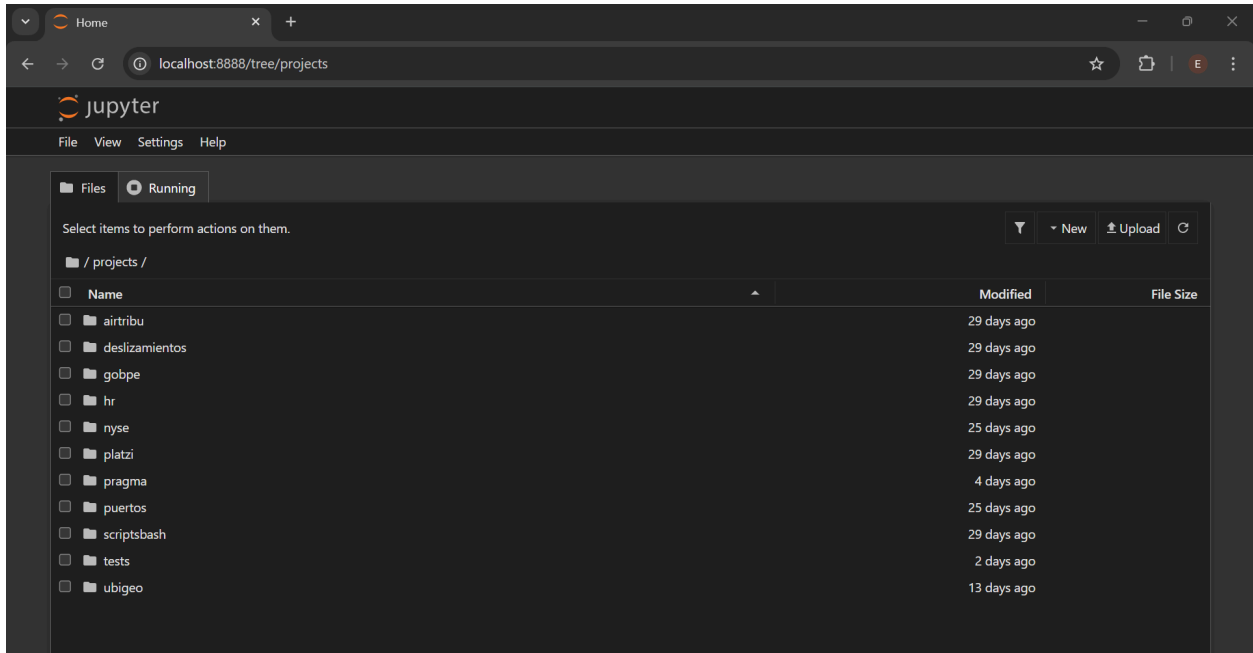
Para acceder a Jupyter Notebook desde una computadora en el mismo rango de red, usamos el siguiente comando desde WSL (Windows Subsystem Linux):

```
ssh -L 8888:localhost:8888
```

```
ernestosegundo@lluyllucucha.ernestosegundo.red
```



```
ernestosegundo@lluyllucucha: ~  
ernestosegundo@WSL-MOYOBAMBA: /mnt/d$ ssh -L 8888:localhost:8888 ernestosegundo@lluyllucucha.ernestosegundo.red  
Linux lluyllucucha 6.8.12-14-pve #1 SMP PREEMPT_DYNAMIC PMX 6.8.12-14 (2025-08-26T22:25Z) 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 Sep  7 11:45:31 2025 from 192.168.1.33  
ernestosegundo@lluyllucucha:~$ jupyter notebook  
[I 2025-09-07 13:54:28.845 LabApp] JupyterLab extension loaded from /home/ernestosegundo/anaconda3/lib/python3.9/site-packages/jupyterlab  
[I 2025-09-07 13:54:28.845 LabApp] JupyterLab application directory is /home/ernestosegundo/anaconda3/share/jupyter/lab  
[I 13:54:28.850 NotebookApp] Serving notebooks from local directory: /home/ernestosegundo  
[I 13:54:28.850 NotebookApp] Jupyter Notebook 6.4.5 is running at:  
[I 13:54:28.850 NotebookApp] http://localhost:8888/?token=ee4d3cc3159196fcd3bb05e5b1b9a6030a4014dde6ff3680f  
[I 13:54:28.850 NotebookApp] or http://127.0.0.1:8888/?token=ee4d3cc3159196fcd3bb05e5b1b9a6030a4014dde6ff3680f  
[I 13:54:28.850 NotebookApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).  
[W 13:54:28.854 NotebookApp] No web browser found: could not locate runnable browser.  
[C 13:54:28.854 NotebookApp]  
  
To access the notebook, open this file in a browser:  
file:///home/ernestosegundo/.local/share/jupyter/runtime/nbserver-3802-open.html  
Or copy and paste one of these URLs:  
http://localhost:8888/?token=ee4d3cc3159196fcd3bb05e5b1b9a6030a4014dde6ff3680f  
or http://127.0.0.1:8888/?token=ee4d3cc3159196fcd3bb05e5b1b9a6030a4014dde6ff3680f
```



```

[12]: from pyspark.sql import SparkSession
      from pyspark.sql.types import StructType, StructField, StringType, IntegerType

[13]: spark = SparkSession.builder.appName("createDataFrameOverview1").getOrCreate()
      spark.sparkContext.setLogLevel("ERROR")

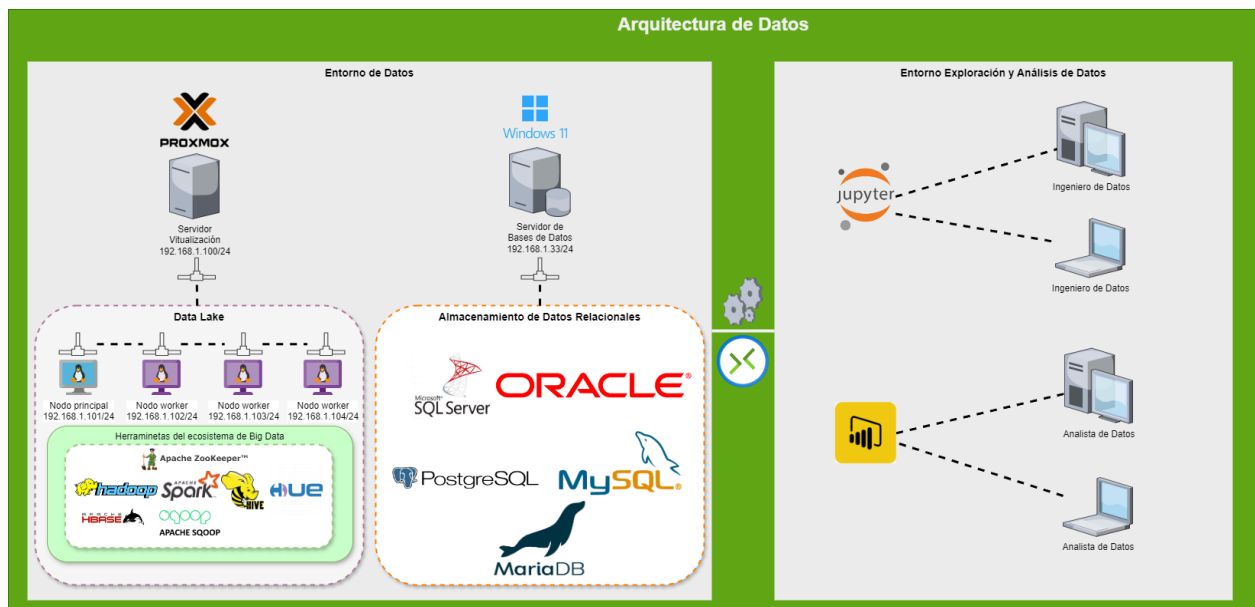
Crear un DataFrame a partir de una lista de tuplas

[14]: spark.createDataFrame([(('Moyobamba', 0),
                               ('Calzada', 1),
                               ('Habana', 2),
                               ('Jepelacio', 3),
                               ('Soritor', 4),
                               ('Yantalo', 5))]).show()

+-----+-----+
|_1|_2|
+-----+-----+
|Moyobamba| 0|
|Calzada| 1|

```

En resumen, la arquitectura de datos implementada se ve de la siguiente manera:



Anexos

Variables de entorno en los distintos servidores del clúster

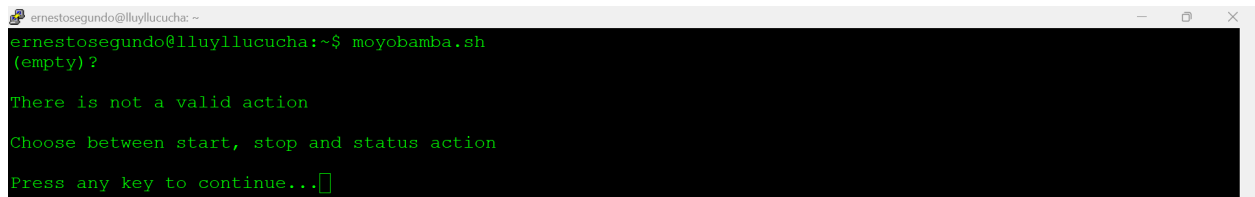
```
ernestosegundo@luyilucucha: ~
# Sources /etc/bash.bashrc).
if ! shopt -oq posix; then
  if [ -f /usr/share/bash-completion/bash_completion ]; then
    . /usr/share/bash-completion/bash_completion
  elif [ -f /etc/bash_completion ]; then
    . /etc/bash_completion
  fi
fi
export JAVA_HOME="/opt/java/jdk1.8.0_361"
export HADOOP_CLASSPATH="$JAVA_HOME/lib/tools.jar"
export HADOOP_HOME="/opt/hadoop"
export HIVE_HOME="/opt/hadoop/hive"
export SQOOP_HOME="/opt/hadoop/sqoop"
export ZOOKEEPER_HOME="/opt/hadoop/zookeeper"
export SPARK_HOME="/opt/hadoop/spark"
export HBASE_HOME="/opt/hadoop/hbase"
export CONDA3_HOME="/opt/anaconda3"
export HOME_SCRIPTS="/scripts"
export PATH="$PATH:$JAVA_HOME/bin:$HADOOP_HOME/bin:$HADOOP_HOME/sbin:$HIVE_HOME/bin:$SQOOP_HOME/bin:$ZOOKEEPER_HOME/bin:$SPARK_HOME/bin:$SPARK_HOME/sbin:$HBASE_HOME/bin:$CONDA3_HOME/bin:$HOME_SCRIPTS"
export SPARK_DIST_CLASSPATH=$(hadoop classpath)
export HADOOP_CONF_DIR="$HADOOP_HOME/etc/hadoop"
export HADOOP_HOME_WARN_SUPPRESS=1
export HADOOP_ROOT_LOGGER="ERROR,DREA"
export PYTHON_VER=python3.13
export PYTHONPATH="$SPARK_HOME/python:$SPARK_HOME/python/lib/py4j-0.10.9.7-src.zip:$PYTHONPATH"
export PYSARK_PYTHON=python
export PYSARK_DRIVER_PYTHON=python
export HADOOP_COMMON_LIB_NATIVE_DIR="$HADOOP_HOME/lib/native"
export HADOOP_OPTS="$HADOOP_OPTS -Djava.library.path=$HADOOP_HOME/lib"
export LD_LIBRARY_PATH="$HADOOP_HOME/lib/native"
export JAVA_LIBRARY_PATH="$HADOOP_HOME/lib/native:$JAVA_LIBRARY_PATH"
136,1 Bot
```

```
ernestosegundo@belen: ~
# Sources /etc/bash.bashrc).
if ! shopt -oq posix; then
  if [ -f /usr/share/bash-completion/bash_completion ]; then
    . /usr/share/bash-completion/bash_completion
  elif [ -f /etc/bash_completion ]; then
    . /etc/bash_completion
  fi
fi
export JAVA_HOME="/opt/java/jdk1.8.0_361"
export HADOOP_CLASSPATH="$JAVA_HOME/lib/tools.jar"
export HADOOP_HOME="/opt/hadoop"
export HIVE_HOME="/opt/hadoop/hive"
export SQOOP_HOME="/opt/hadoop/sqoop"
export ZOOKEEPER_HOME="/opt/hadoop/zookeeper"
export SPARK_HOME="/opt/hadoop/spark"
export HBASE_HOME="/opt/hadoop/hbase"
export CONDA3_HOME="/opt/anaconda3"
export HOME_SCRIPTS="/scripts"
export PATH="$PATH:$JAVA_HOME/bin:$HADOOP_HOME/bin:$HADOOP_HOME/sbin:$HIVE_HOME/bin:$SQOOP_HOME/bin:$ZOOKEEPER_HOME/bin:$SPARK_HOME/bin:$SPARK_HOME/sbin:$HBASE_HOME/bin:$CONDA3_HOME/bin:$HOME_SCRIPTS"
export SPARK_DIST_CLASSPATH=$(hadoop classpath)
export HADOOP_CONF_DIR="$HADOOP_HOME/etc/hadoop"
export HADOOP_HOME_WARN_SUPPRESS=1
export HADOOP_ROOT_LOGGER="ERROR,DREA"
export PYTHON_VER=python3.13
export PYTHONPATH="$SPARK_HOME/python:$SPARK_HOME/python/lib/py4j-0.10.9.7-src.zip:$PYTHONPATH"
export PYSARK_PYTHON=python
export PYSARK_DRIVER_PYTHON=python
export HADOOP_COMMON_LIB_NATIVE_DIR="$HADOOP_HOME/lib/native"
export HADOOP_OPTS="$HADOOP_OPTS -Djava.library.path=$HADOOP_HOME/lib"
export LD_LIBRARY_PATH="$HADOOP_HOME/lib/native"
export JAVA_LIBRARY_PATH="$HADOOP_HOME/lib/native:$JAVA_LIBRARY_PATH"
125,32 Bot
```

Inicio de servicios del clúster Hadoop

Teniendo en cuenta que es necesario ejecutar varios comandos para iniciar los diferentes servicios del clúster, se ha implementado un script bash que los ejecuta secuencialmente.

El script es [moyobamba.sh](#) y requiere un argumento, el cual puede ser uno de los siguientes valores: start, stop o status.



```
ernestosegundo@lluyllucucha: ~  
ernestosegundo@lluyllucucha:~$ moyobamba.sh  
(empty)?  
There is not a valid action  
Choose between start, stop and status action  
Press any key to continue...[]
```

Notas:

Instalar una nueva versión de Python en Linux Debian

Compilar desde el código fuente (para una versión específica o más nueva)

- Descarga el código fuente: Del sitio oficial de Python (python.org) copiar el enlace al archivo comprimido (tarball) de la versión deseada.

```
wget <enlace_del_tarball>  
tar -xvf Python-<nombre_de_la_version>.tgz  
cd Python-<nombre_de_la_version>
```

- Ejecuta el script de configuración y compila el código fuente:

```
./configure  
make  
sudo make altinstall
```


altinstall es importante para instalar la nueva versión sin reemplazar la que podría estar ya en tu sistema.

```
update-alternatives --list python
```

```
whereis python3.13
```

```
sudo update-alternatives --remove python  
/usr/bin/python3.9
```

```
sudo update-alternatives --install  
/usr/lib/python3.9 python3.9  
/usr/local/lib/python3.9 1
```

```
sudo update-alternatives --install  
/usr/lib/python3.13 python  
/usr/local/lib/python3.13 1
```

```
python --version
```

Actualizar conector para el metastore de Hive con MySQL

Desde el backup de la instalación anterior de Spark, copiar el archivo **mysql-connector-j-8.0.31.jar** a la nueva instalación en la ruta **/opt/hadoop/spark/jars/**