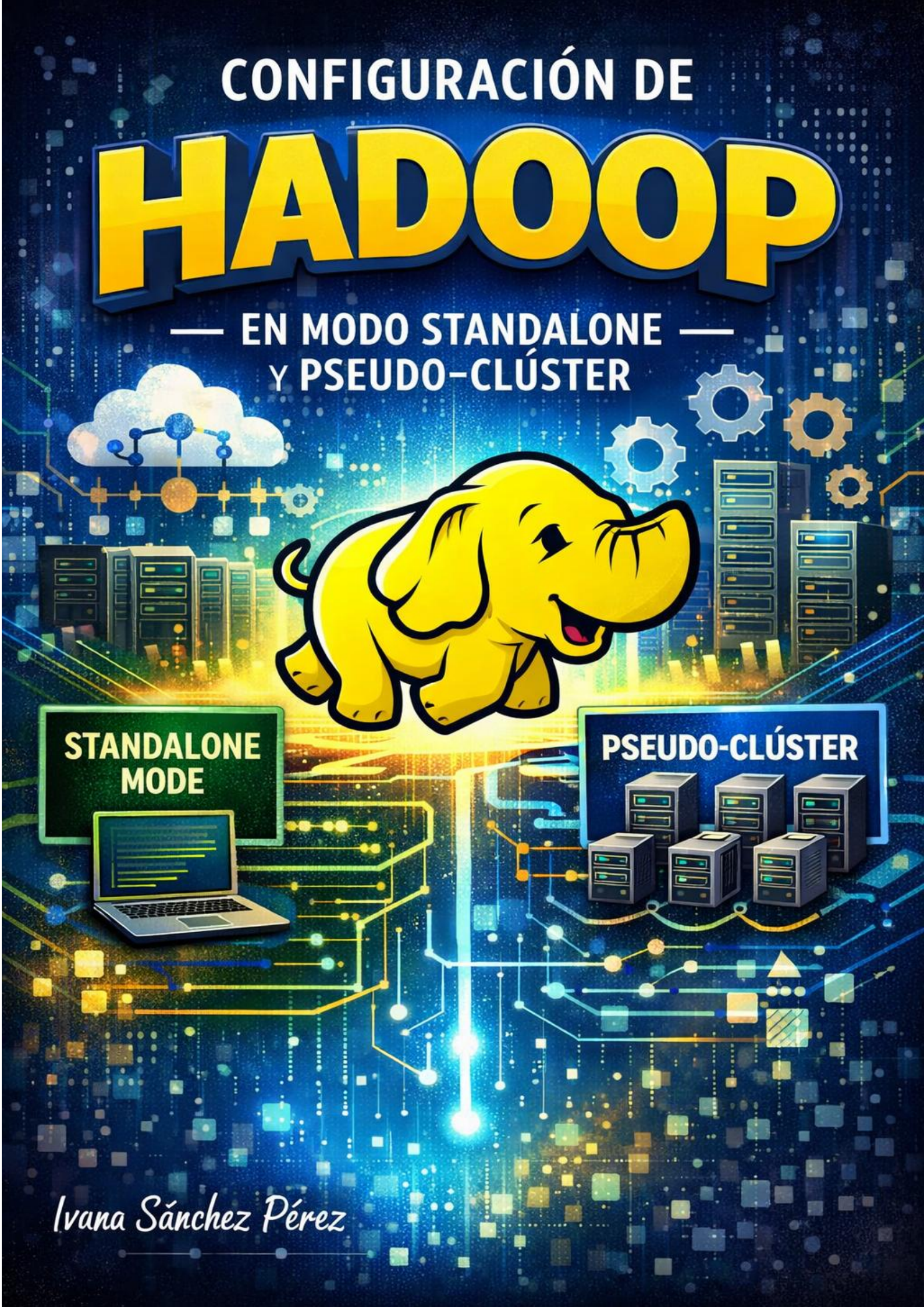
CONFIGURACIÓN DE HADOOP

— EN MODO STANDALONE — Y PSEUDO-CLÚSTER

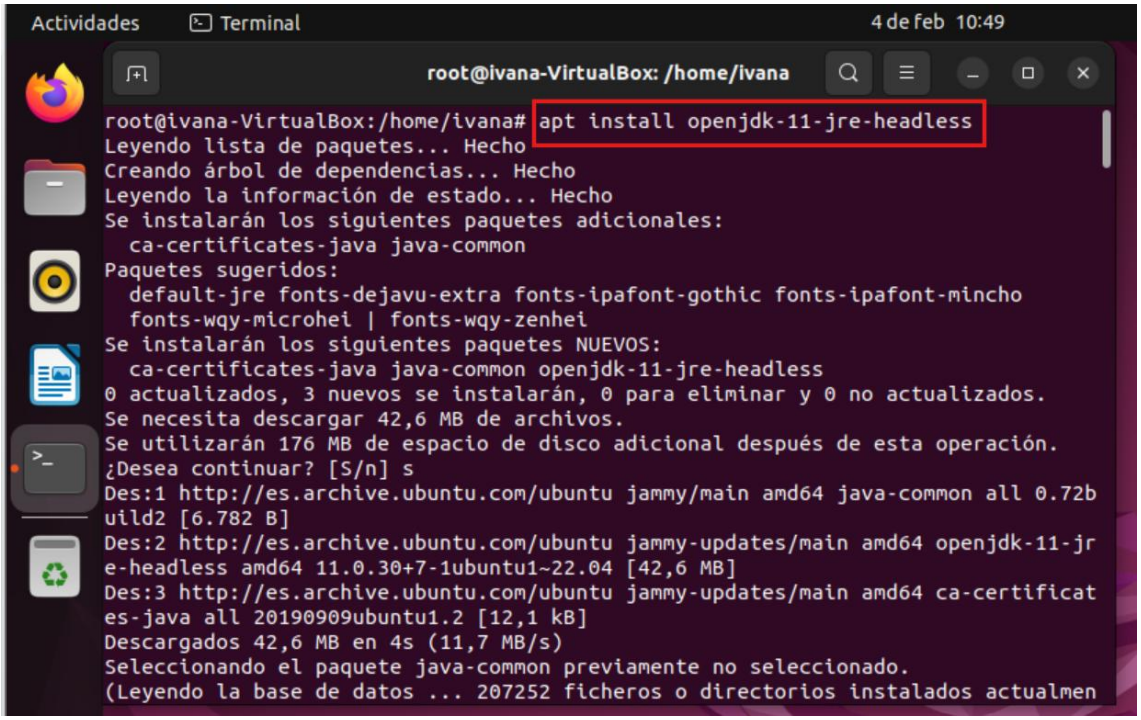


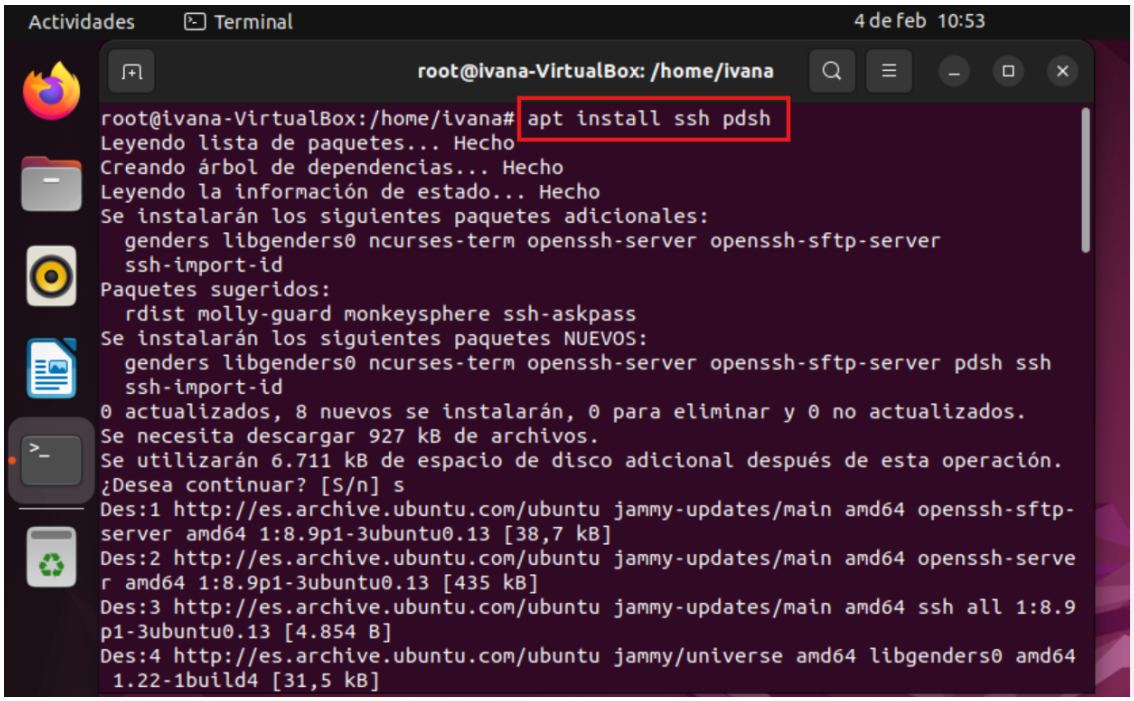
Ivana Sánchez Pérez

CONFIGURACIÓN DE HADOOP EN MODO STANDALONE Y PSEUDO-CLÚSTER

CONFIGURACIÓN DEL ENTORNO (MODO STANDALONE)

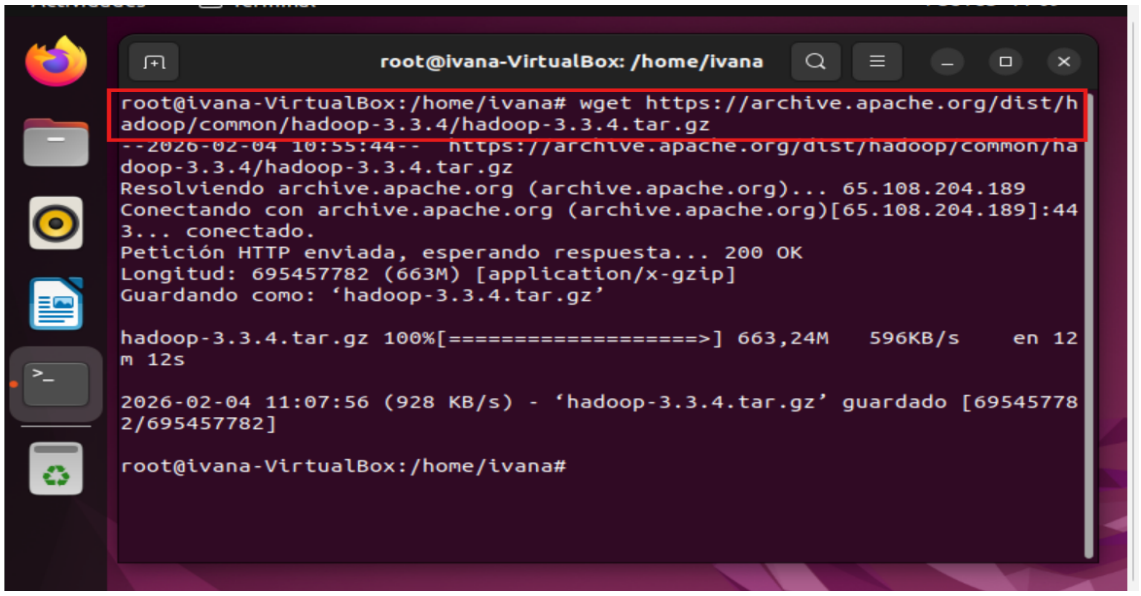
1- Instalación de Java 11, SSH y PDSH

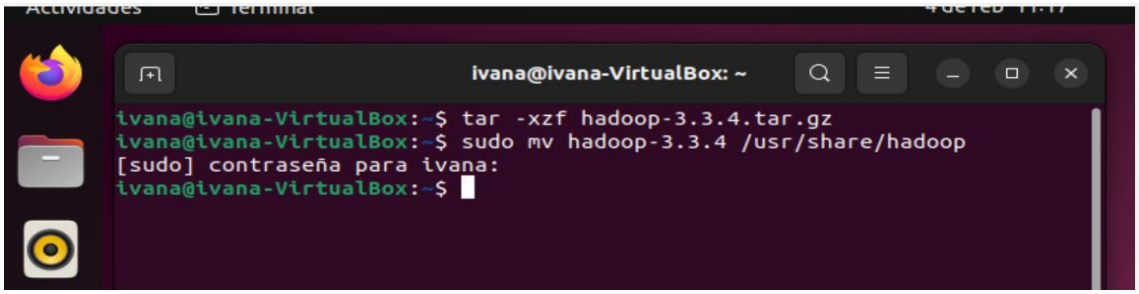




2- Instalación de Hadoop

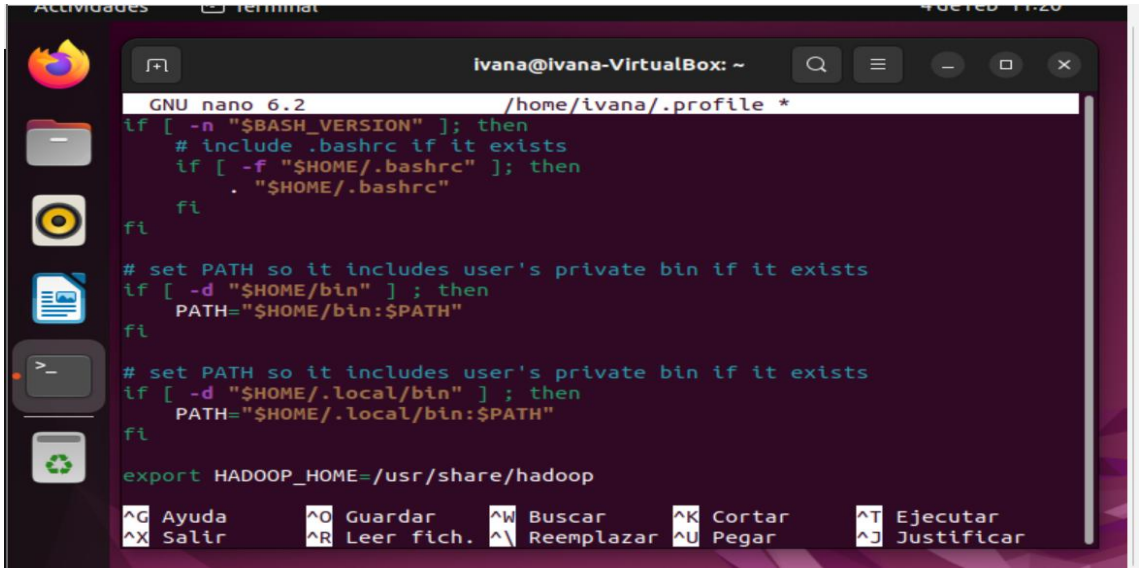
Descarga y comprensión de la versión 3.3.4 en la ruta ***/usr/share/hadoop.***



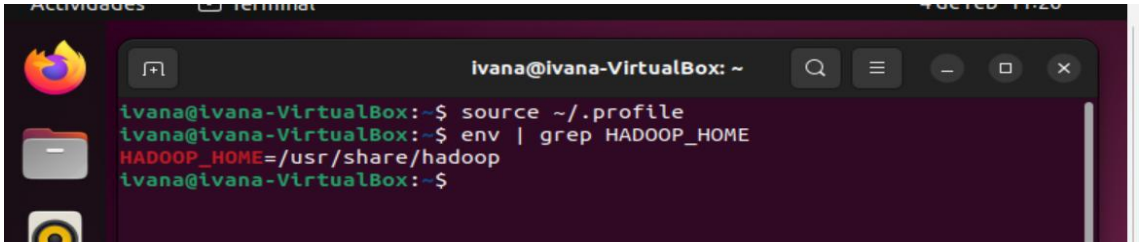


3- Variables del entorno

Configuración del archivo ***.profile*** para definir ***HADOOP_HOME*** y asegurar que el sistema reconozca los binarios.

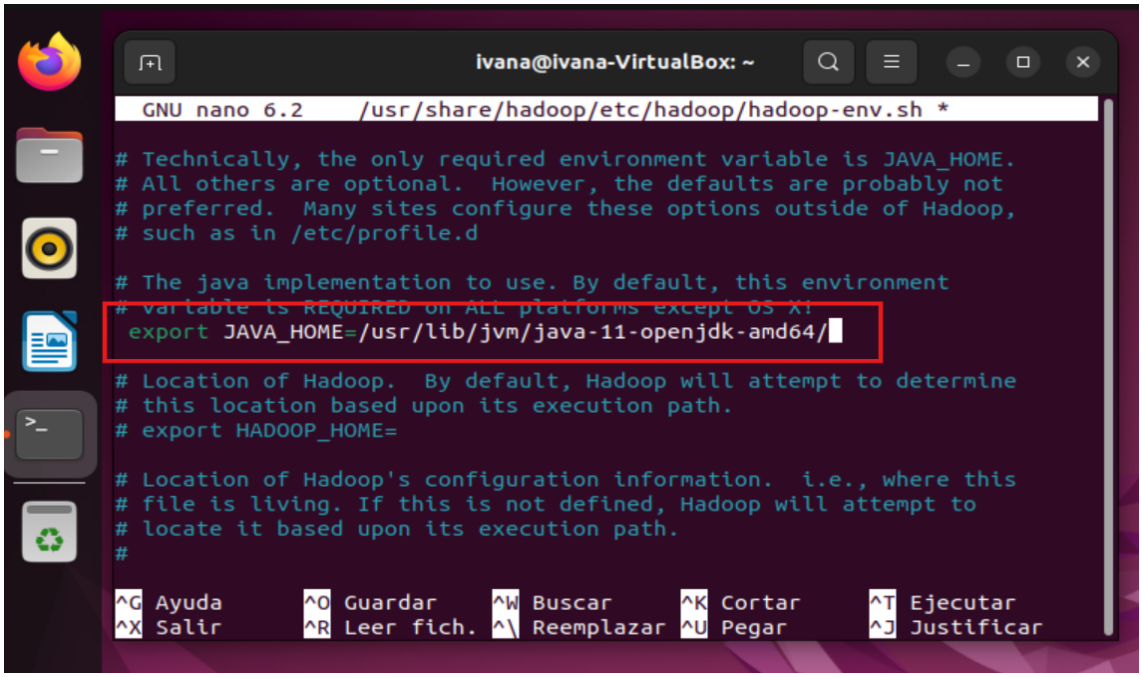


Cargamos los cambios y verificamos



4- Configuración de Java en Hadoop

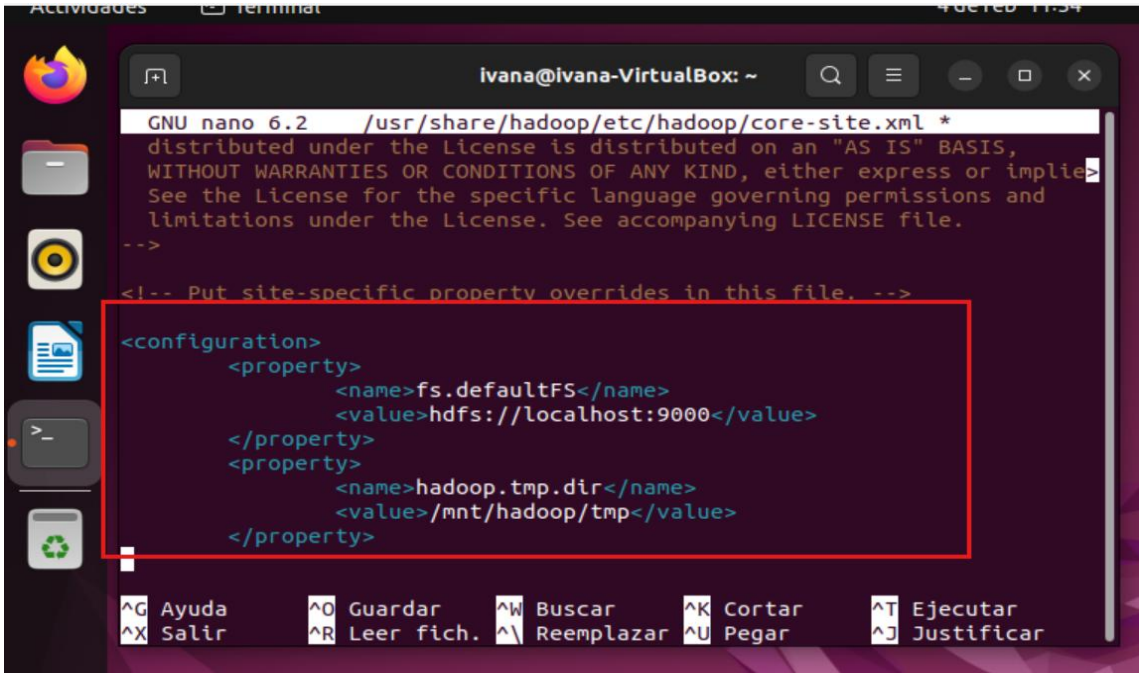
Hadoop necesita saber dónde está Java para arrancar sus servicios. Para ello definimos la ruta de Java dentro de hadoop-env.sh.

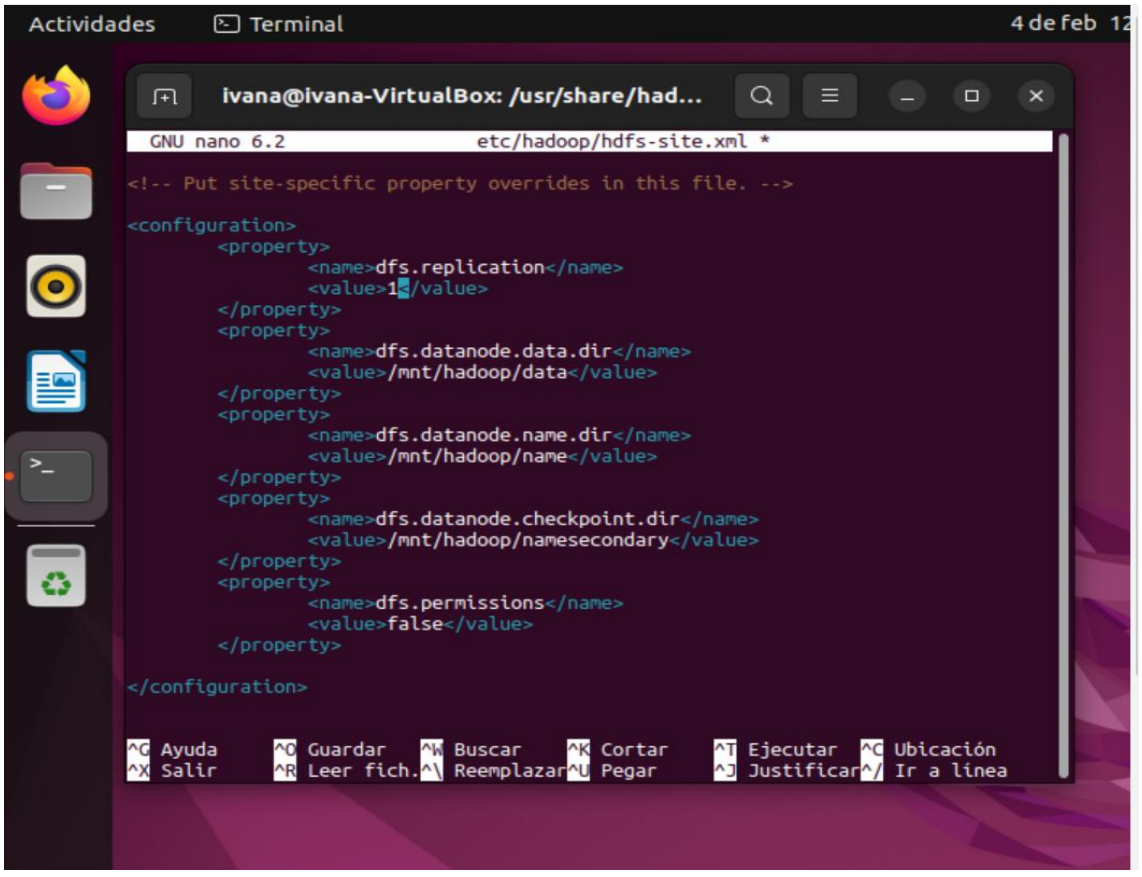


CONFIGURACIÓN EN MODO PSEUDO-CLÚSTER

Este modo simula un clúster en nuestra propia máquina.

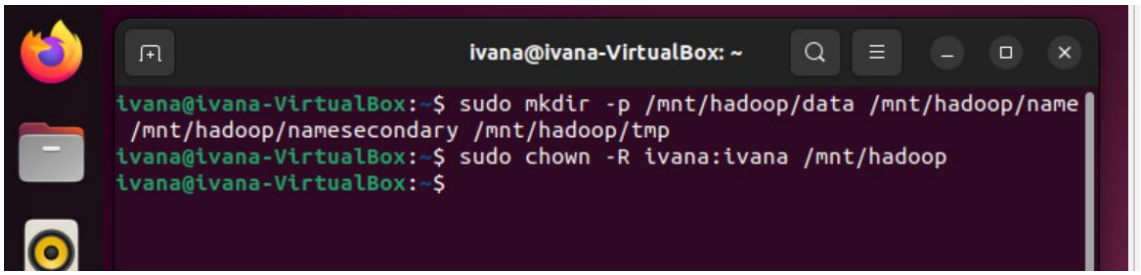
1- **Configuración de los archivos XML**
   Edición de ***core-site.xml*** y ***hdfs-site.xml*** para definir el sistema de archivos por defecto y factor de replicación.





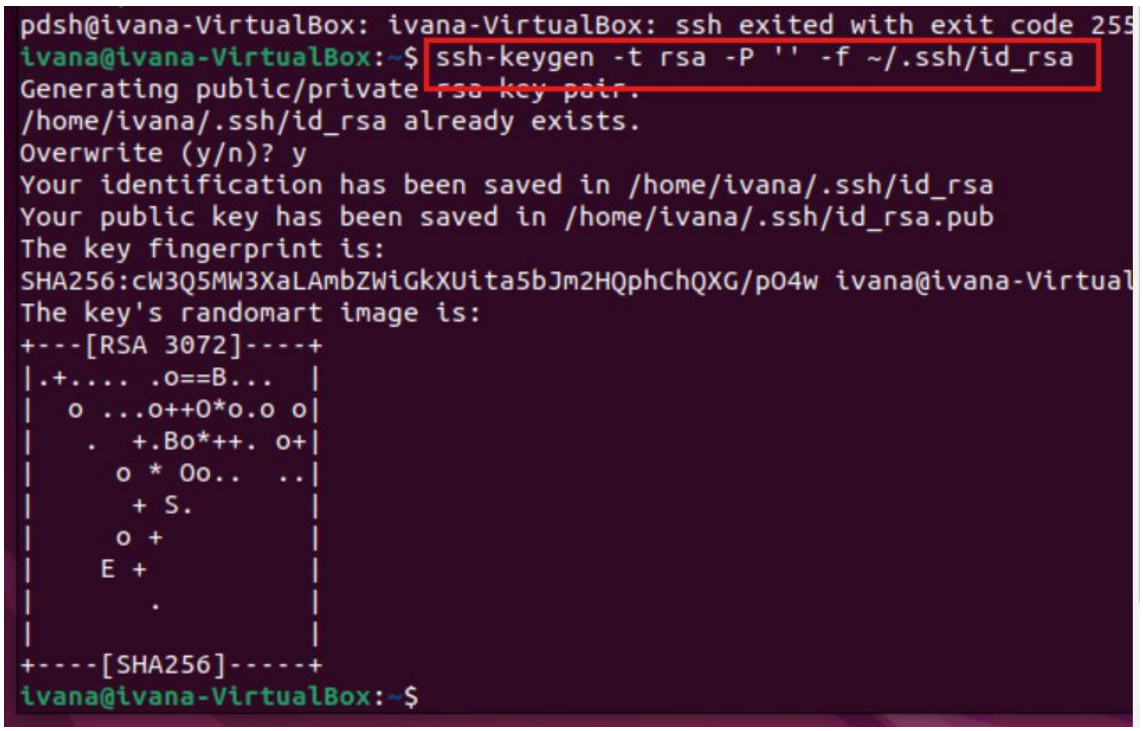
2- Preparación de directorios

Creación de las carpetas físicas para datos, nombres y temporales en ***/mnt/hadoop***, asignando permisos al usuario actual.



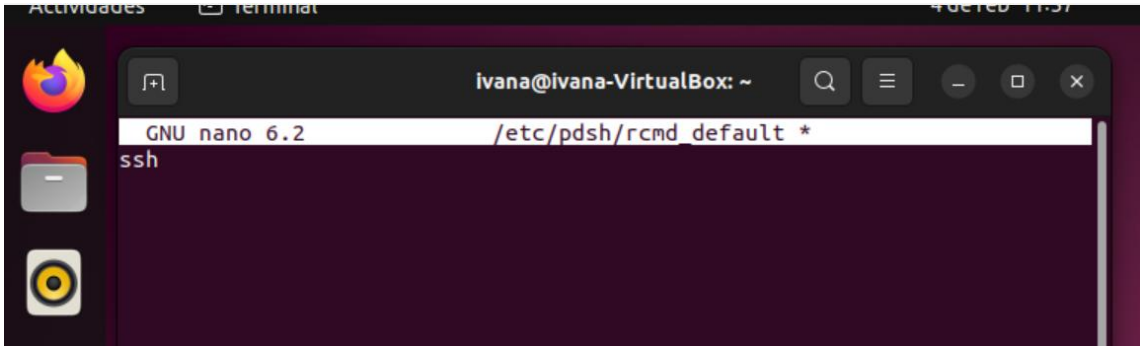
3- Acceso a SSH sin clave

Generación de llaves RSA para permitir que Hadoop se comunique internamente consigo mismo, sin solicitar contraseñas.

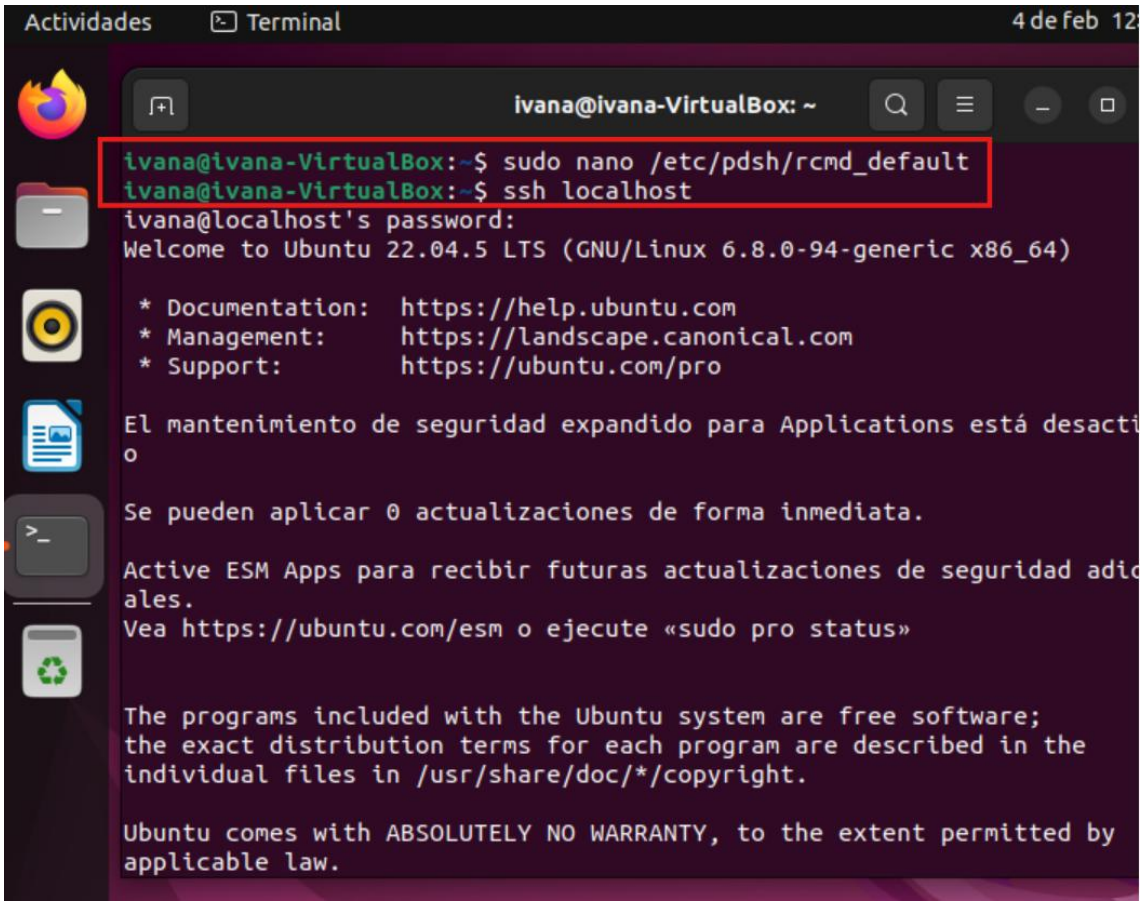


4- Configuramos el Shell remoto por defecto para PDSH

Obligatorio en Ubuntu para que Hadoop pueda arrancar los servicios sin errores. Creamos el archivo de configuración ***rcmd_default***



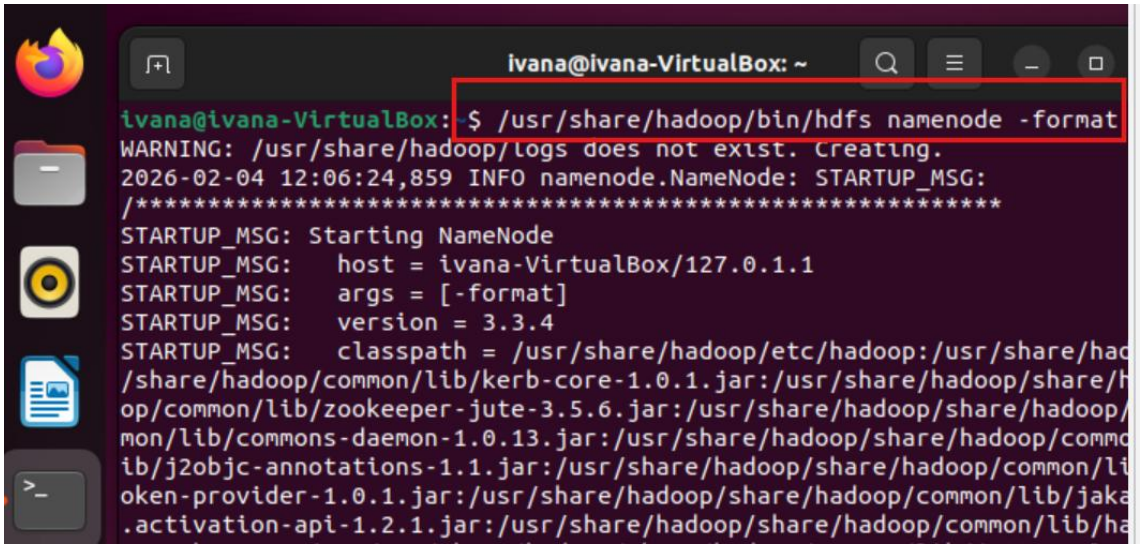
Verificamos la conexión SSH



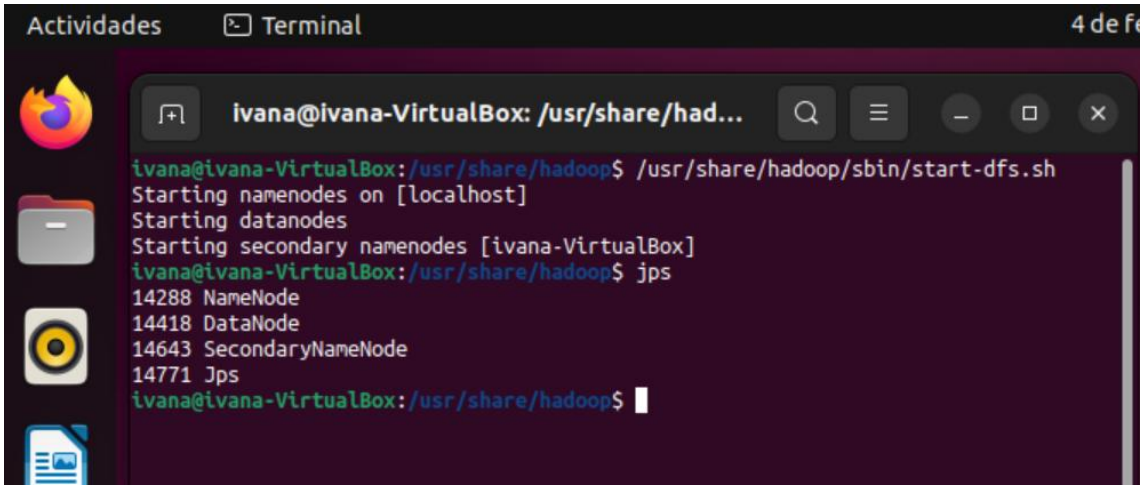
5- Formateo y arranque

Formateo del NameNode y primer arranque de los servicios de almacenamiento mediante start-dfs.sh

a) Formatear el NameNode



b) Arrancar los servicios de almacenamiento

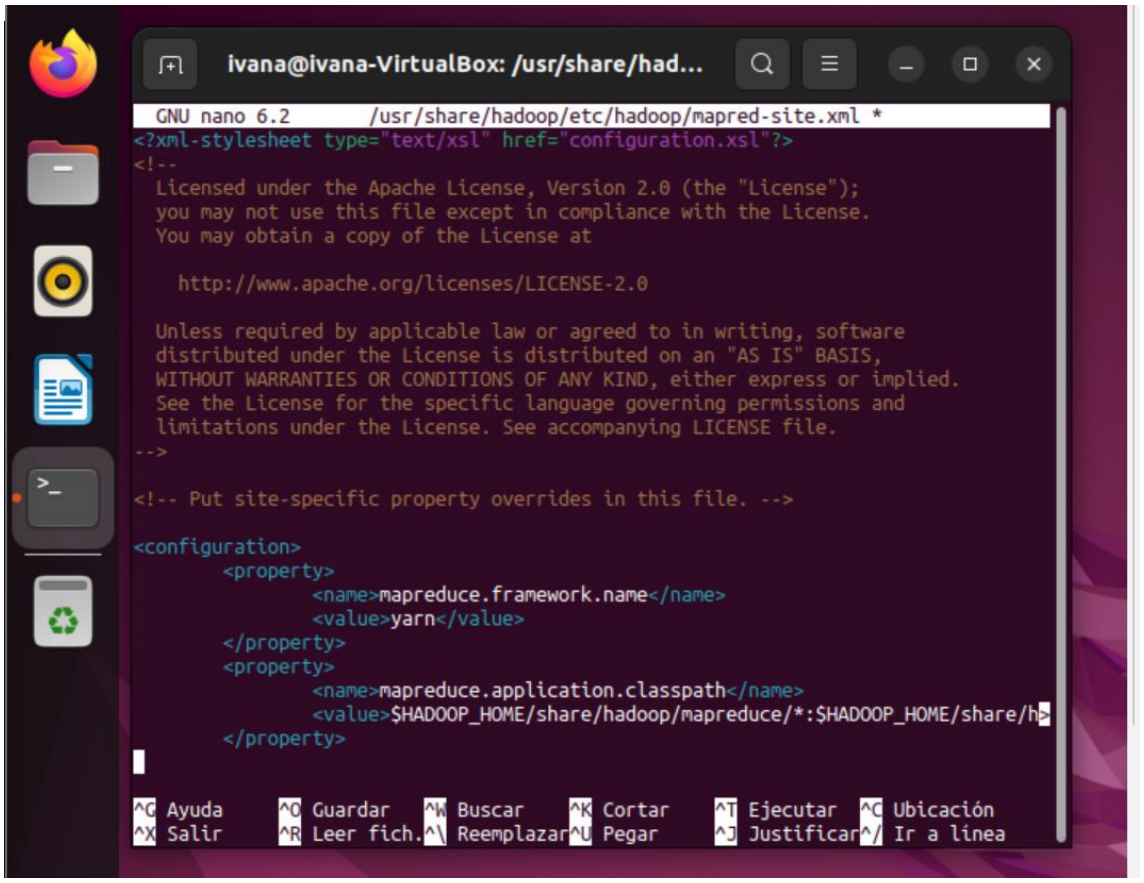


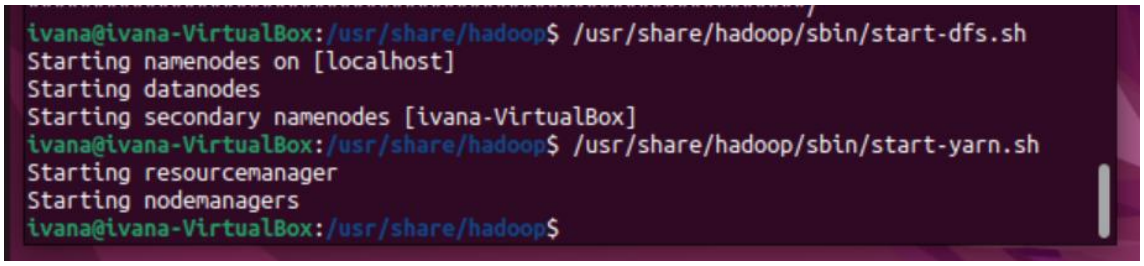
CONFIGURACIÓN DE YARN Y RESOLUCIÓN DE ERRORES

Necesario para gestionar los recursos del sistema.

1- Activación de YARN

Configuración de mapred.site.xml y arranque del gestor de recursos para ejecutar tareas MapReduce

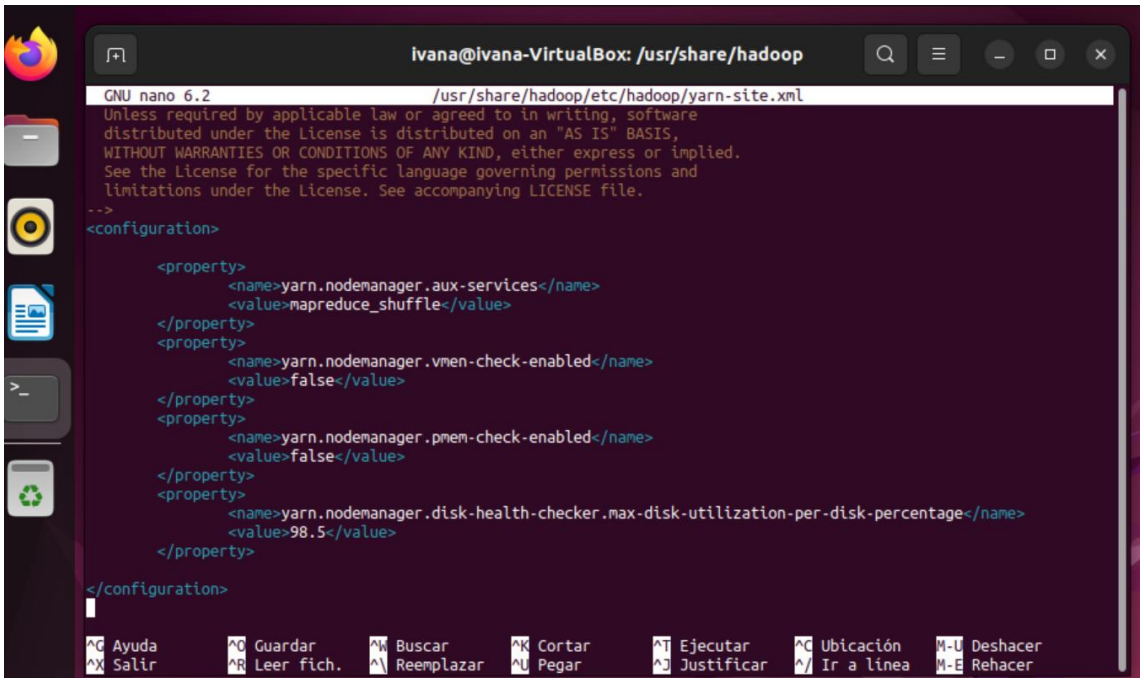




2- Depuración de errores

Debido al entorno de ejecución en una VM con recursos limitados, la tarea de WordCount fallaba inicialmente al superar los límites de memoria virtual permitidos por el sistema.

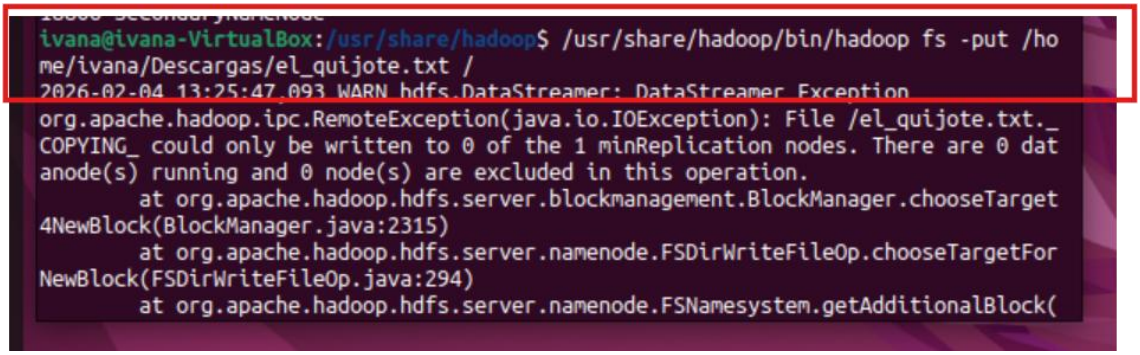
Para solucionar este inconveniente se ha procedido a modificar el archivo de configuración yarn-site.xml, para que desactivar la comprobación de la memorias tanto virtual como física y se ajustó al 98,5% para permitir que Hadoop siga trabajando incluso si el disco de la VM está casi lleno



EJECUCIÓN DE LA TAREA Y RESULTADOS FINALES

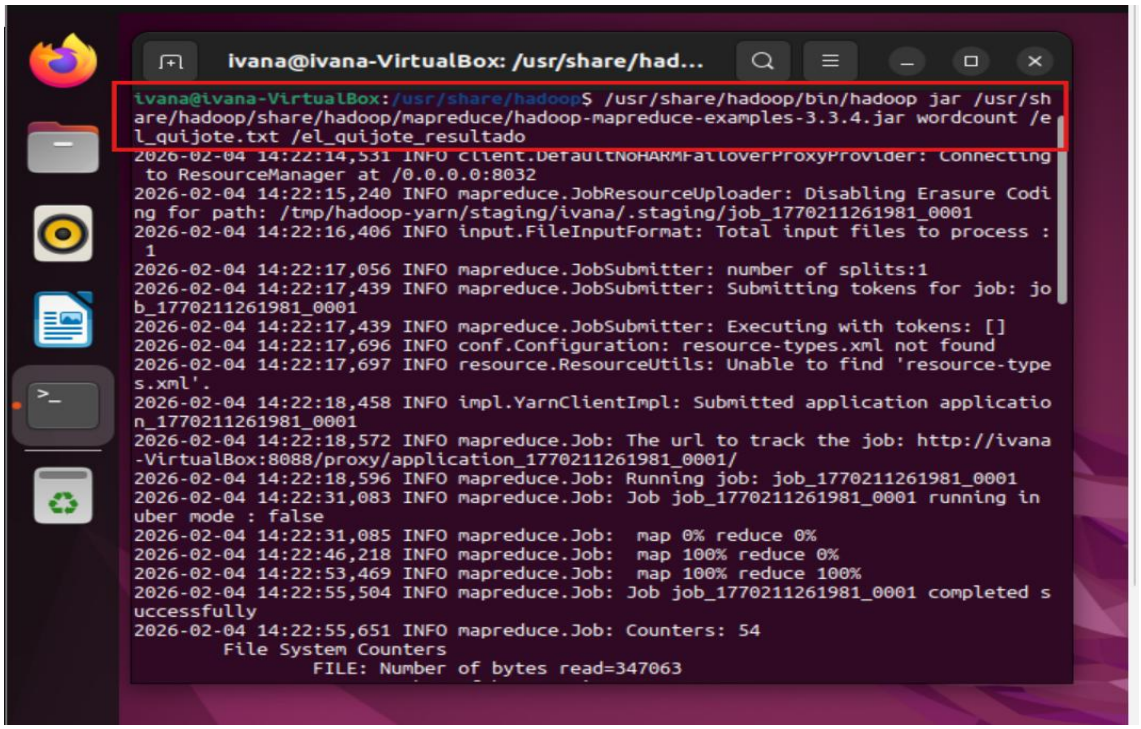
1- Carga de Datos

Subida del archivo el_quijote.txt (texto plano) al sistema HDFS



2- Procesamiento WordCount

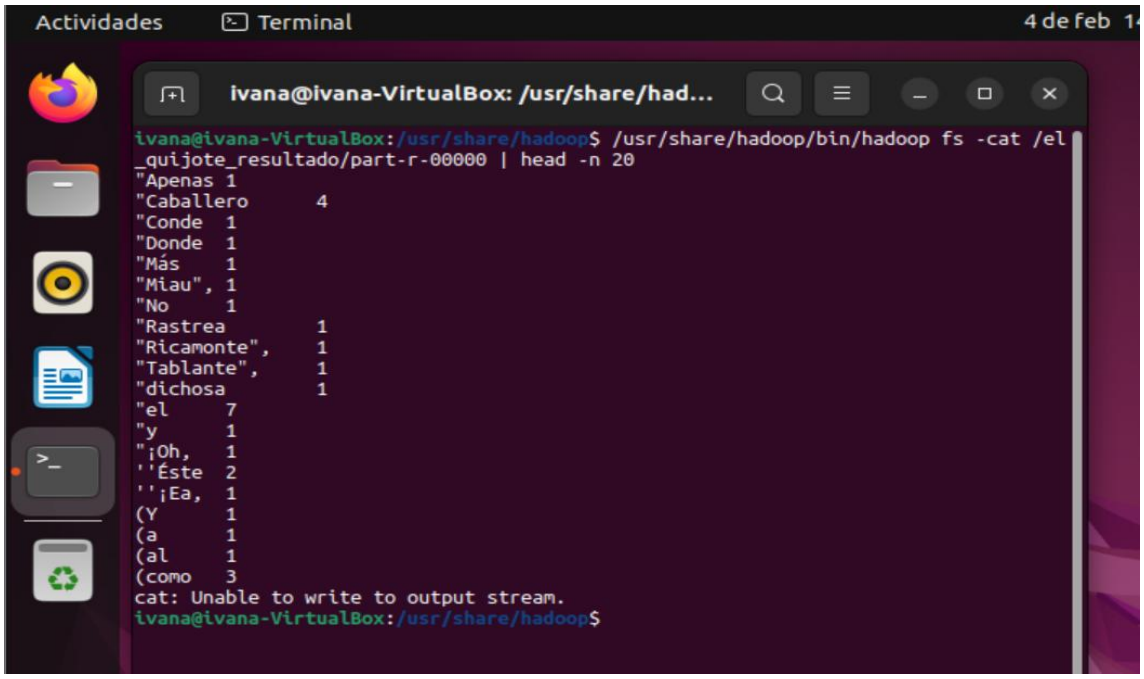
Ejecución del programa de ejemplo para contar palabras, alcanzando el estado de SUCCEEDED tras las correcciones de memoria.



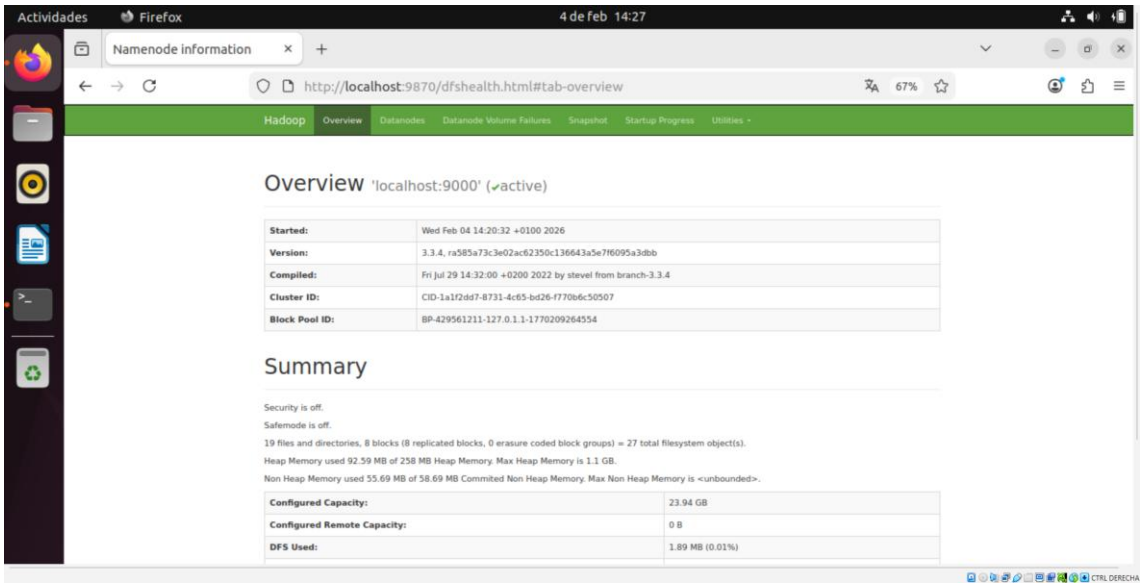

```
                Combine input records=187018
                Combine output records=22938
                Reduce input groups=22938
                Reduce shuffle bytes=347063
                Reduce input records=22938
                Reduce output records=22938
                Spilled Records=45876
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=159
                CPU time spent (ms)=9080
                Physical memory (bytes) snapshot=555212800
                Virtual memory (bytes) snapshot=5475381248
                Total committed heap usage (bytes)=416284672
                Peak Map Physical memory (bytes)=328871936
                Peak Map Virtual memory (bytes)=2733268992
                Peak Reduce Physical memory (bytes)=226340864
                Peak Reduce Virtual memory (bytes)=2742112256
        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=1060259
        File Output Format Counters
                Bytes Written=257233
ivana@ivana-VirtualBox:/usr/share/hadoop$
```

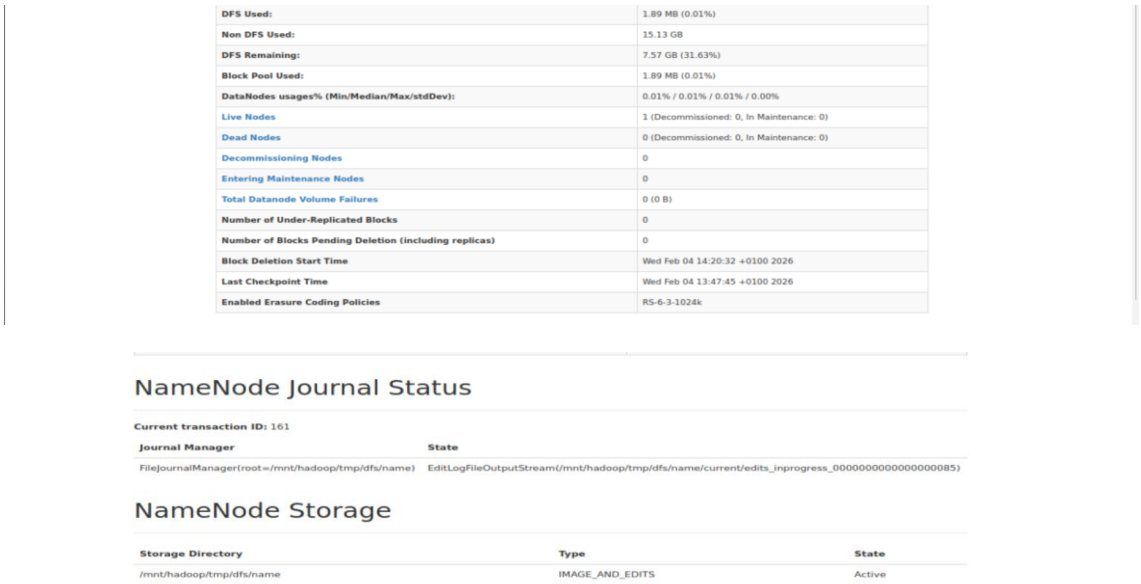
3- Verificaciones

a. **Terminal:** Visualización de las primeras 20 palabras procesadas.



b. **Interfaz Web:** Comprobación en https://localhost:9870 del estado Active del NameNode y la presencia de los datos en el clúster.





DFS Used:	1.89 MB (0.01%)
Non DFS Used:	15.13 GB
DFS Remaining:	7.57 GB (31.63%)
Block Pool Used:	1.89 MB (0.01%)
DataNodes usages% (Min/Median/Max/stdDev):	0.01% / 0.01% / 0.01% / 0.00%
Live Nodes	1 (Decommissioned: 0, In Maintenance: 0)
Dead Nodes	0 (Decommissioned: 0, In Maintenance: 0)
Decommissioning Nodes	0
Entering Maintenance Nodes	0
Total Datanode Volume Failures	0 (0 B)
Number of Under-Replicated Blocks	0
Number of Blocks Pending Deletion (including replicas)	0
Block Deletion Start Time	Wed Feb 04 14:20:32 +0100 2026
Last Checkpoint Time	Wed Feb 04 13:47:45 +0100 2026
Enabled Erasure Coding Policies	RS-6-3-1024k

NameNode Journal Status

Current transaction ID: 161

Journal Manager	State
FileJournalManager(root=/mnt/hadoop/tmp/dfs/name)	EditLogFileOutputStream(/mnt/hadoop/tmp/dfs/name/current/edits_inprogress_0000000000000000085)

NameNode Storage

Storage Directory	Type	State
/mnt/hadoop/tmp/dfs/name	IMAGE_AND_EDITS	Active

En la imagen se observan los siguientes puntos clave que demuestran que el sistema está bien configurado:

- **Estado Activo**: El NameNode aparece como 'localhost:9000' (active), lo que indica que el formateo y el arranque fueron exitosos.
- **Seguridad y Modo Seguro**: Ambos parámetros, Security is off y Safemode is off, están desactivados, tal como se pedía en la guía para facilitar las pruebas.
- **Nodos Vivos**: Se muestra Live Nodes: 1, confirmando que el DataNode está conectado y listo para recibir datos.
- **Uso de Disco**: El sistema reconoce la capacidad total y muestra que hay archivos almacenados (1.89 MB de uso de DFS), que corresponden al Quijote y sus resultados.

ÚLTIMO PASO: APAGADO CORRECTO

Para asegurar que no se dan los archivos de datos que acabamos de ver en el navegador, ejecutaremos el comando de apagado antes de cerrar nuestra VM.