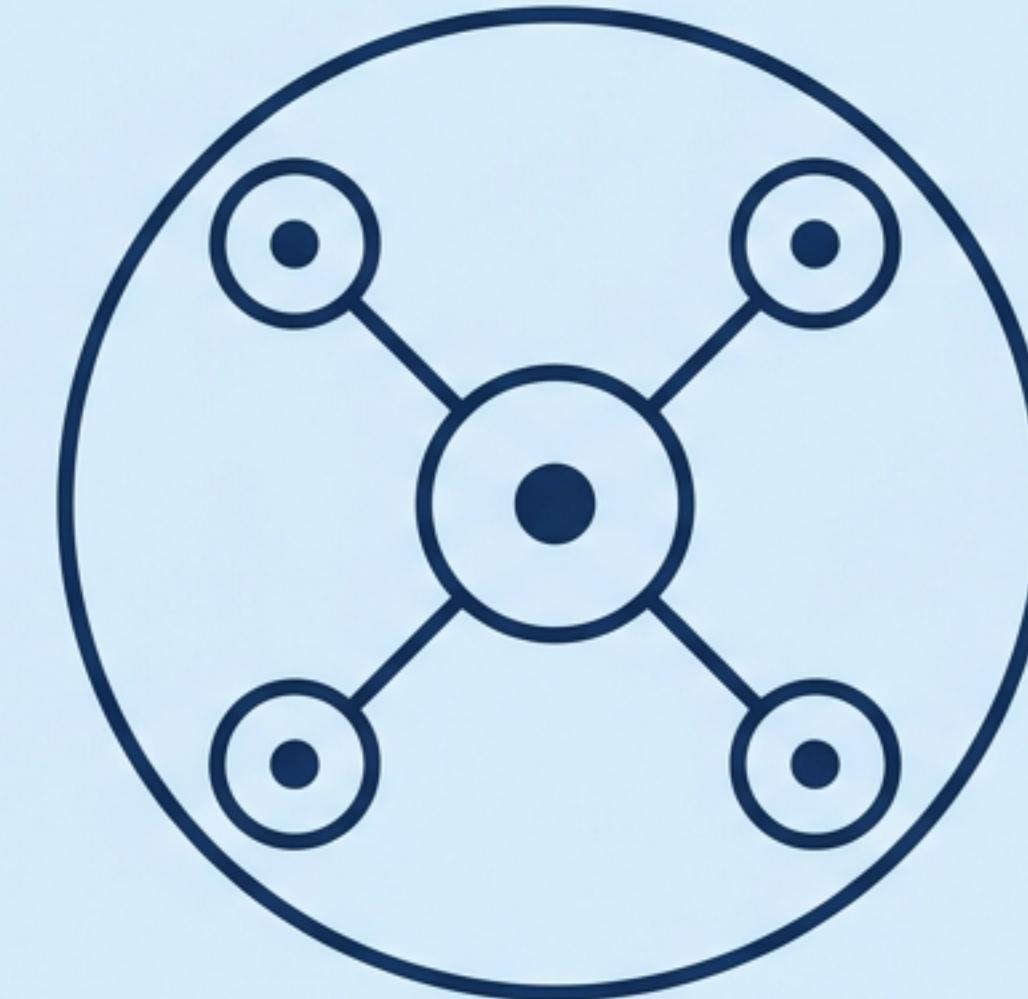


Construyendo tu Primer Pseudoclúster Hadoop



Una guía paso a paso para configurar un entorno de aprendizaje en tu propia máquina.

El Objetivo: Un Laboratorio Hadoop en tu Propia Máquina

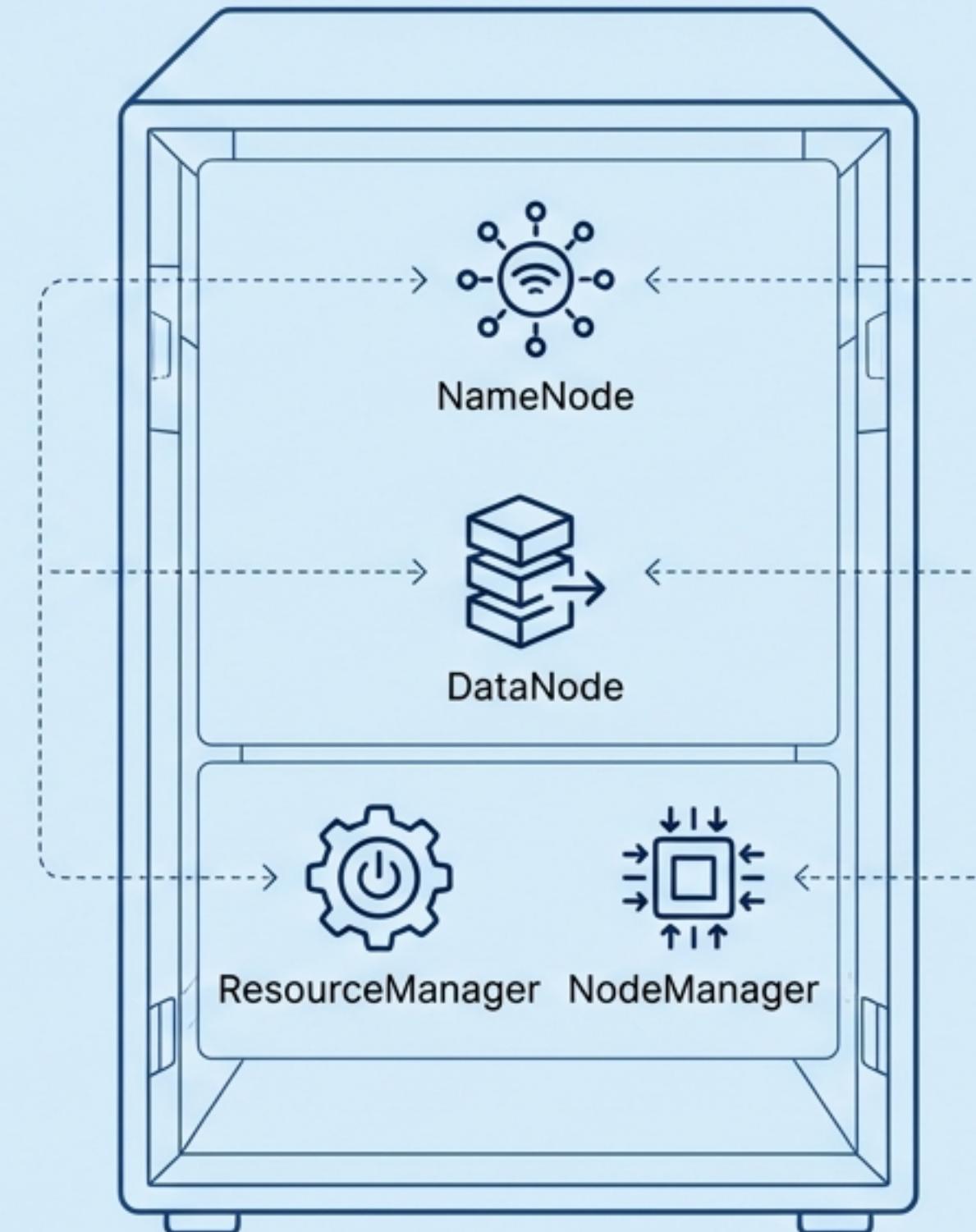
Un pseudoclúster ejecuta todos los componentes de Hadoop (NameNode, DataNode, ResourceManager, NodeManager) en una única máquina.



Ideal para Aprender: Permite familiarizarse con la arquitectura y comandos de Hadoop sin la necesidad de hardware múltiple.



Perfecto para Pruebas: Facilita el desarrollo y la depuración de aplicaciones Big Data en un entorno controlado.



Nuestro Viaje: De Cero a un Clúster Funcional

1



Preparación

Instalar Hadoop y configurar el entorno.

2



Configuración

Editar los archivos XML clave que definen el clúster.

3



Lanzamiento

Formatear el sistema de archivos e iniciar los servicios.

4



Verificación

Comprobar que todo funciona correctamente.

1

Etapa 1: Preparación del Entorno

Paso 1: Descargar y Ubicar Hadoop

Descarga la última versión desde el sitio oficial de Apache Hadoop.

Extrae los archivos en un directorio de tu sistema (ejemplo recomendado: `/opt/hadoop`).

Paso 2: Configurar Variables de Entorno

Define las rutas esenciales para que tu sistema pueda encontrar y ejecutar Hadoop.

Ejemplo para `~/.bashrc` o similar

```
export HADOOP_HOME=/opt/hadoop
export HADOOP_CONF_DIR=$HADOOP_HOME/etc/
hadoop
export PATH=$PATH:$HADOOP_HOME/bin:
$HADOOP_HOME/sbin
```

2 Etapa 2: Configurando el Cerebro del Clúster

Archivo Clave: `core-site.xml`

Propósito

Este archivo contiene la configuración básica del sistema de archivos distribuido (HDFS). Su parámetro más importante, `fs.defaultFS`, indica al clúster la dirección del NameNode.

Ubicación

/opt/hadoop/etc/hadoop/core-site.xml

```
<configuration>
  <property>
    <name>fs.defaultFS</name>
    <value>hdfs://localhost:9000</value>
  </property>
</configuration>
```

2 Etapa 2: Definiendo Dónde Viven los Datos

Archivo Clave: `hdfs-site.xml`

Propósito

Aquí se configuran los directorios locales donde el NameNode almacenará sus metadatos y el DataNode guardará los bloques de datos. También se define el factor de replicación.

Ubicación

`/opt/hadoop/etc/hadoop/hdfs-site.xml`

```
<configuration>
  <property>
    <name>dfs.replication</name>
    <value>1</value>
  </property>
  <property>
    <name>dfs.namenode.name.dir</name>
    <value>file:/datos/namenode</value>
  </property>
  <property>
    <name>dfs.datanode.data.dir</name>
    <value>file:/datos/datanode</value>
  </property>
</configuration>
```

Nota

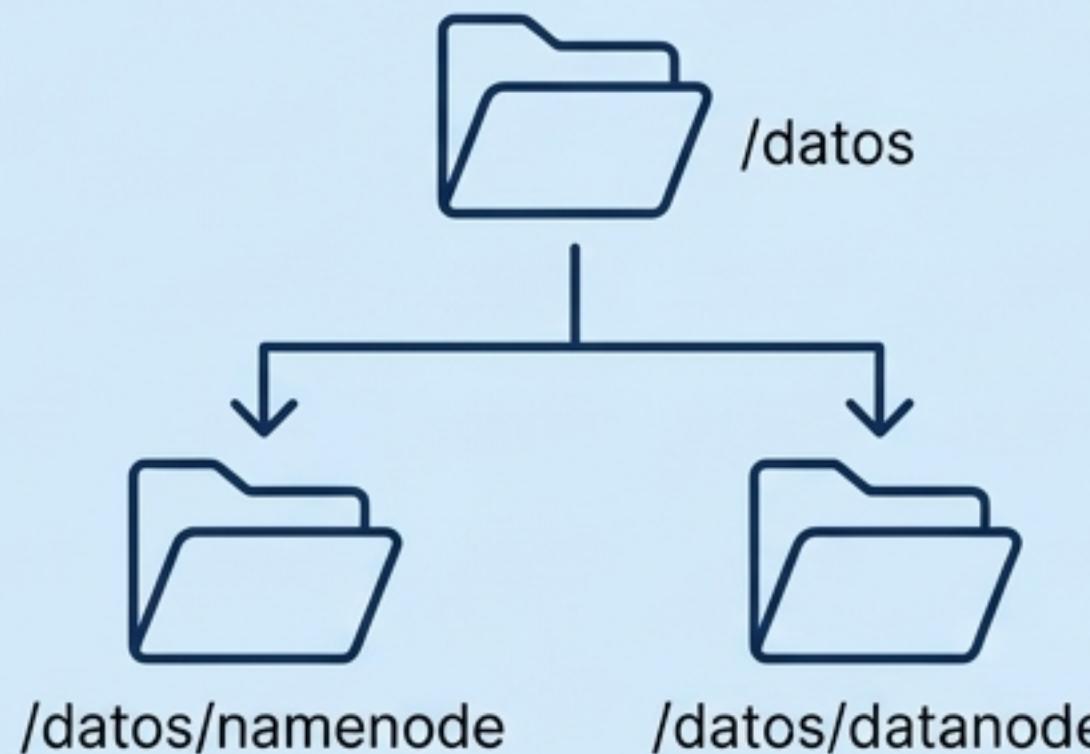
Un factor de replicación de `1` es adecuado para un pseudoclúster.

2 Etapa 2: Creando el Espacio Físico

Contexto

Los directorios que especificaste en `hdfs-site.xml` deben existir en tu sistema de archivos local antes de poder iniciar Hadoop.

Acción Requerida: Crea las carpetas usando la terminal.



```
# Crea una carpeta central para los datos de Hadoop  
sudo mkdir -p /datos
```

```
# Crea los directorios para el NameNode y DataNode  
sudo mkdir -p /datos/namenode  
sudo mkdir -p /datos/datanode
```

```
# Asegura los permisos para el usuario que ejecutará Hadoop  
sudo chown -R tu_usuario:tu_grupo /datos
```

Consejo: Asegúrate de que el usuario que ejecuta Hadoop tenga permisos de lectura y escritura sobre estos directorios.

3 Etapa 3: Inicialización del Sistema

Paso Clave: Formatear el NameNode



Antes de iniciar el clúster por primera vez, debes formatear el NameNode. Este paso inicializa la estructura de directorios y los metadatos del sistema de archivos distribuido de Hadoop (HDFS).



¡Atención!: Este comando se ejecuta **una sola vez**. Volver a ejecutarlo en un clúster existente borrará todos los datos de HDFS.

```
hdfs namenode -format
```

3 Etapa 3: ¡A Encender los Motores!

Iniciar los servicios de HDFS (NameNode y DataNode).



Hadoop proporciona scripts para iniciar y detener todos los servicios necesarios de forma conveniente.

start-dfs.sh

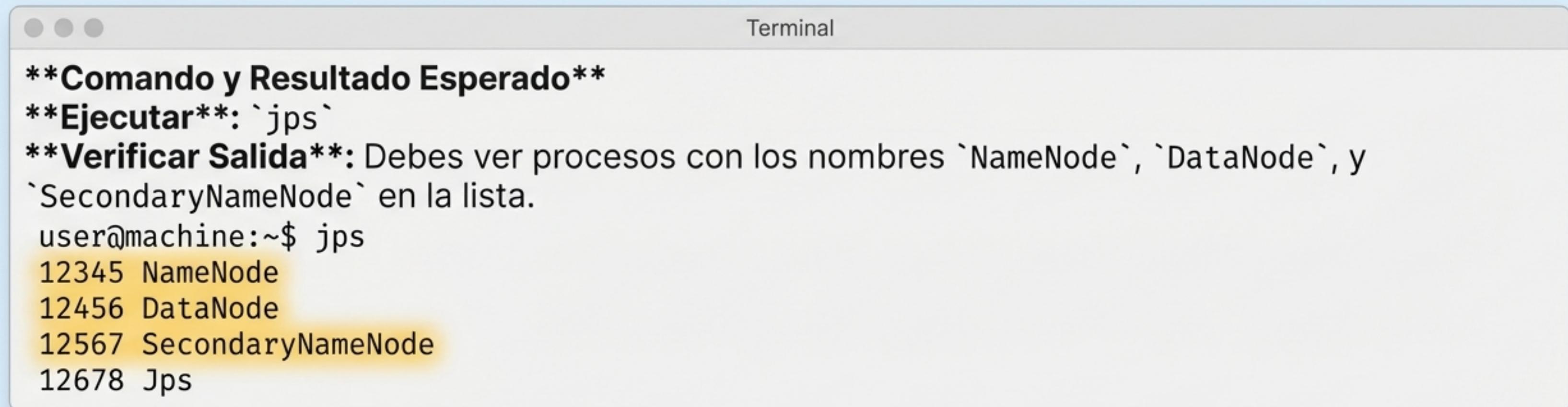
4 Etapa 4: Verificación por Línea de Comandos

****Herramienta****

`jps` (Java Virtual Machine Process Status Tool)

****Propósito****

Este comando te permite ver todos los procesos de Java que se están ejecutando en tu máquina, confirmando que los daemons de Hadoop han iniciado.



A screenshot of a terminal window titled "Terminal". The window shows the output of the `jps` command. The output lists several Java processes, including the NameNode, DataNode, and SecondaryNameNode daemons, along with other system processes like Jps and Jps.

```
user@machine:~$ jps
12345 NameNode
12456 DataNode
12567 SecondaryNameNode
12678 Jps
```

4 Etapa 4: Verificación a través de la Interfaz Web

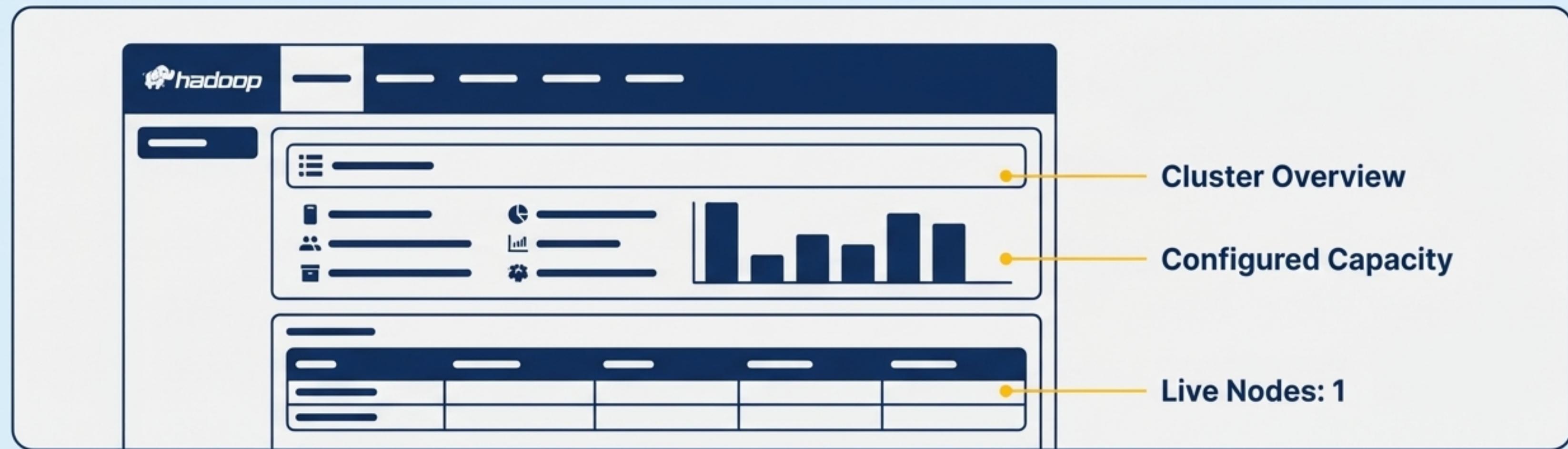
Acceso

Abre tu navegador web y dirígete a la siguiente dirección.

← → C <http://localhost:9870>

¿Qué Encontrarás?

Una página de resumen del estado de tu clúster HDFS. Podrás ver información general, la capacidad total, el espacio utilizado y el estado de los nodos de datos (DataNodes).



4 Etapa 4: La Prueba de Fuego, Cargar un Archivo

Objetivo: Realizar una prueba básica cargando un archivo a HDFS para verificar que el sistema de archivos está completamente operativo.

Paso 1: Crear un archivo de prueba

```
echo "Hola Hadoop" > prueba.txt
```

Paso 2: Subir el archivo a HDFS

```
hdfs dfs -put prueba.txt /
```

Paso 3: Verificar que el archivo está en HDFS

```
hdfs dfs -ls /
```

Resultado Esperado

Deberías ver `prueba.txt` en la lista de archivos.

```
Found 1 items  
-rw-r--r-- 1 user supergroup
```

```
12 2023-10-27 10:30 /prueba.txt
```

Consejos Clave para el Éxito



Ubicación Central

Todos los archivos de configuración clave residen en /opt/hadoop/etc/hadoop.



Permisos Correctos

Asegúrate de que los directorios locales (/datos/namenode, /datos/datanode) sean accesibles y tengan los permisos adecuados para el usuario que ejecuta Hadoop.



Consulta la Fuente

Si tienes dudas sobre alguna configuración, la documentación oficial de Hadoop es tu mejor recurso.



¡Misión Cumplida! Tu Pseudoclúster está Operativo.

Has configurado con éxito un entorno Hadoop funcional en tu propia máquina. Ahora tienes una poderosa herramienta para explorar el mundo del Big Data.

¿Y Ahora Qué? Próximos Pasos:

- Experimenta con más comandos de HDFS.
- Aprende a ejecutar tu primer trabajo MapReduce.
- Explora otros componentes del ecosistema Hadoop.