

Kubernetes, Hadoop y Spark

Integrantes:

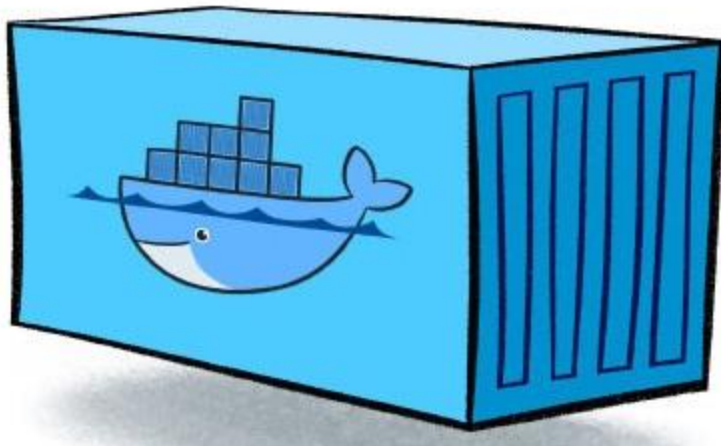
- Eduardo Estévez
- Jorge Gualpa



Contenedores

Los contenedores son una tecnología de virtualización de sistemas operativos que permite empaquetar y distribuir aplicaciones junto con sus dependencias en un entorno aislado y portátil.

Contenedores



Los contenedores se han vuelto cada vez más populares en el desarrollo de aplicaciones debido a su capacidad

Ventajas de los contenedores



Portabilidad

son portátiles y pueden ser ejecutados en diferentes plataformas



Escalabilidad

pueden ser duplicados rápidamente



Eficiencia

reduce la sobrecarga de recursos



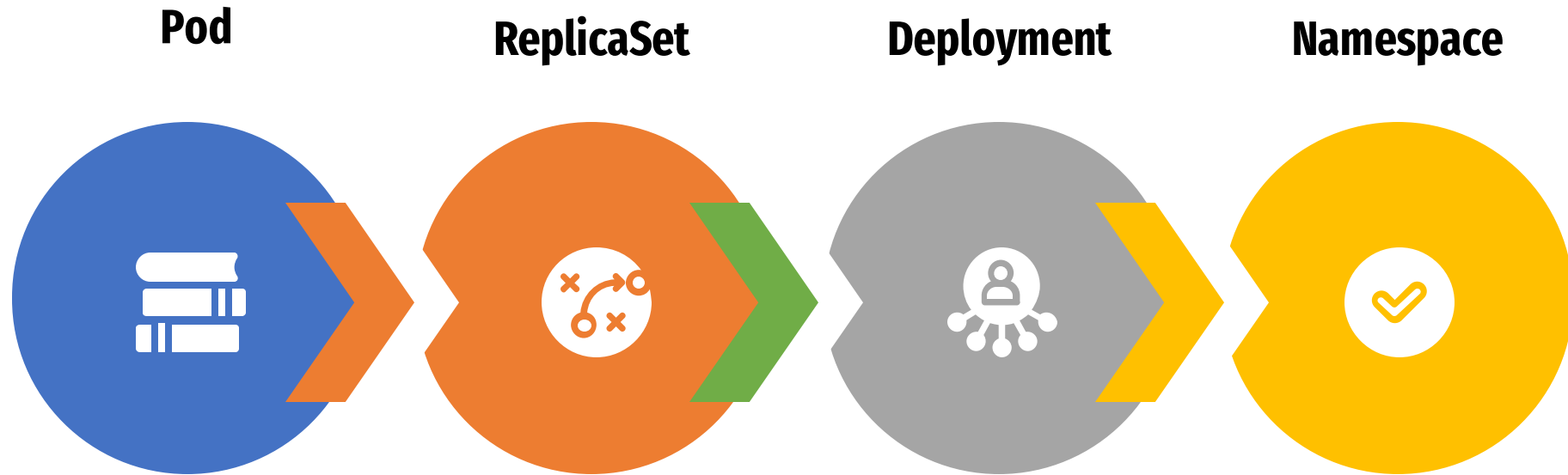
Aislamiento

Ofrecen un alto nivel de aislamiento

Kubernetes



Componentes de Kubernetes



Historia y evolución

2014

Creado por Google como
proyecto de código
abierto

contribución

Comunidad grande de
desarrolladores



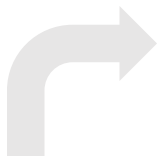
Actualidad

Más popular y de mayor
crecimiento

Como Funciona

funciona mediante la
creación de un clúster
de servidores

01



02



se basa en el concepto
de "pods"



03



Cada pod puede
contener uno o varios
contenedores



04



utiliza un sistema de
programación de
recursos



Arquitectura de Kubernetes

Plano de control

API Server

Etc

Scheduler

Controller manager

Nodos de trabajo

- Kubelet
- Kube-proxy
- Container Runtime

Ventajas de Kubernetes

- Alta disponibilidad
- Portabilidad
- Automatización
- Control y gestión centralizados
- Seguridad



Kubernetes

Qué Es y Para Qué Sirve?



Desventajas de Kubernetes

- **Complejidad**
- **Requisitos de recursos**
- **Curva de aprendizaje**
- **Posibles problemas de compatibilidad**
- **Dependencia de la infraestructura**



Como Realizar la instalación de Kubernetes en Windows

Para instalar Kubernetes primero instalamos kubectl ubicando en el cmd el siguiente comando:

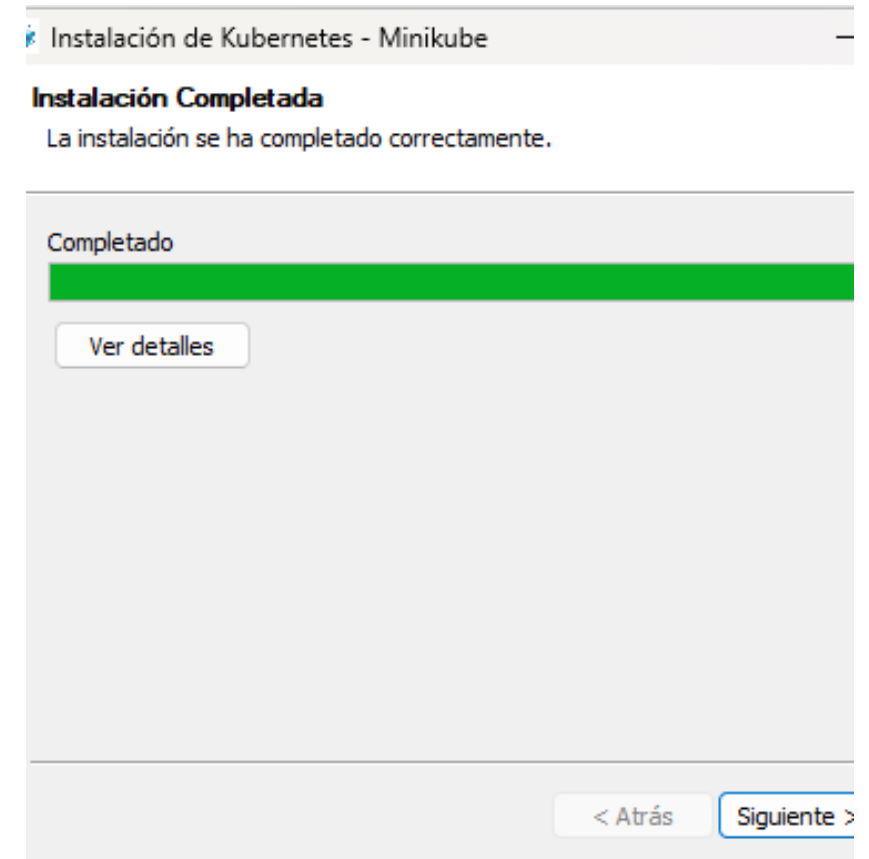
- `curl -LO https://dl.k8s.io/release/v1.27.0/bin/windows/amd64/kubectl.exe`

para verificar la versión utilizamos el siguiente comando:

- `kubectl version --client=true`

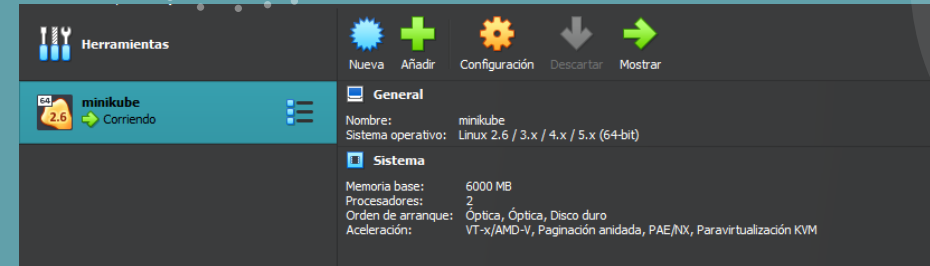
Como Realizar la instalación de Kubernetes en Windows

- Para instalar el custle de Kubernetes descargamos y instalamos minikube en el siguiente enlace:
 - <https://storage.googleapis.com/minikube/releases/latest/minikube-installer.exe>



Como Realizar la instalación de Kubernetes en Windows

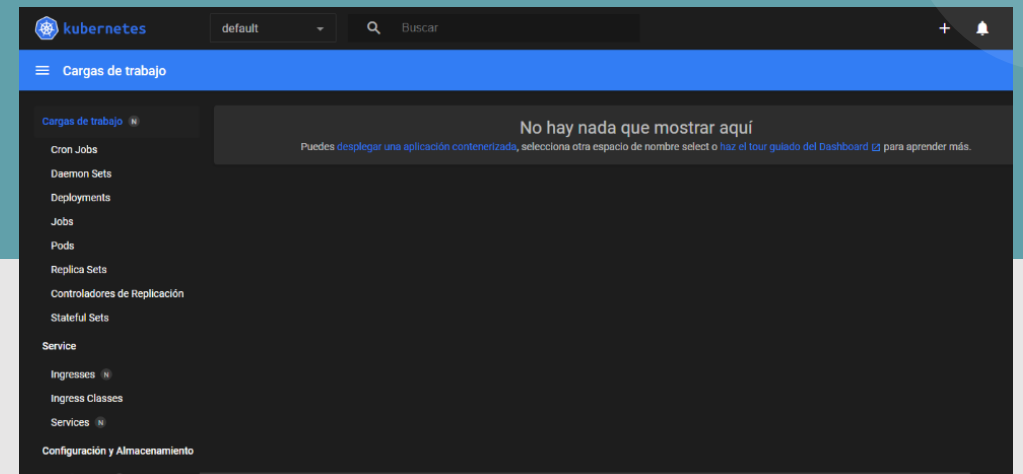
- Para arrancar minikube ubicamos:
 - minikube start
- Esto nos crea una máquina virtual de minikube:



```
C:\Windows\System32>minikube start
* minikube v1.30.1 en Microsoft Windows 11 Pro 10.0.22623.1325 Build 22623.1325
* Controlador virtualbox seleccionado automáticamente
* Descargando la imagen de arranque de la VM
  > minikube-v1.30.1-amd64.iso....: 65 B / 65 B [-----] 100.00% ? p/s 0s
  > minikube-v1.30.1-amd64.iso: 19.03 MiB / 282.84 MiB 6.73% 5.74 MiB p/s E_
```

Para entrar a la herramienta grafica utilizamos el comando: minikube dashboard

- Para pararlo utilizamos el comando: minikube stop El cual tambien parara la máquina virtual.



Como podemos ver namespaces en Kubernetes donde podemos ver la información como despliegues, etc.

Nombre	Etiquetas	Fase	Fecha de creación
kubernetee-dashboard	addonmanager.kubernetes.io/mode: Reconcile kubernetes.io/metadata.name: kubernetee-dashboard	Active	36 minutes ago
default	kubernetes.io/metadata.name: default	Active	36 minutes ago
kube-node-lease	kubernetes.io/metadata.name: kube-node-lease	Active	36 minutes ago
kube-public	kubernetes.io/metadata.name: kube-public	Active	36 minutes ago
kube-system	kubernetes.io/metadata.name: kube-system	Active	36 minutes ago

- Y para saber los namespaces por medio de la línea de comando es la siguiente:
 - `kubectl get namespaces`

```
C:\Windows\System32>kubectl get namespaces
NAME                STATUS    AGE
default             Active    42m
kube-node-lease     Active    42m
kube-public         Active    42m
kube-system         Active    42m
kubernetee-dashboard Active    37m

C:\Windows\System32>
```

```
C:\Windows\System32>kubectl create namespace eduardo-estevez
namespace/eduardo-estevez created
```

```
C:\Windows\System32>kubectl get namespaces
NAME                STATUS    AGE
default             Active    45m
eduardo-estevez     Active    2s
kube-node-lease     Active    45m
kube-public         Active    45m
kube-system         Active    45m
kubernetes-dashboard Active    40m
```

Espacios de nombres				
Nombre	Etiquetas	Fase	Fecha de creación ↑	
eduardo-estevez	kubernetes.io/metadata.name: eduardo-estevez	Active	a minute ago	⋮
kubernetes-dashboard	addonmanager.kubernetes.io/mode: Reconcile kubernetes.io/metadata.name: kubernetes-dashboard kubernetes.io/minikube-addons: dashboard	Active	42 minutes ago	⋮
default	kubernetes.io/metadata.name: default	Active	46 minutes ago	⋮
kube-node-lease	kubernetes.io/metadata.name: kube-node-lease	Active	46 minutes ago	⋮
kube-public	kubernetes.io/metadata.name: kube-public	Active	46 minutes ago	⋮
kube-system	kubernetes.io/metadata.name: kube-system	Active	46 minutes ago	⋮

Como crear un namespaces: Lo tenemos que realizar por medio de comando.

* kubectl create namespace eduardo-estevez

- Y como vemos el namespaces fue creado con éxito

- Para eliminar el namespaces ubicamos el siguiente comando:

- `kubectl delete namespace eduardo-estevez`

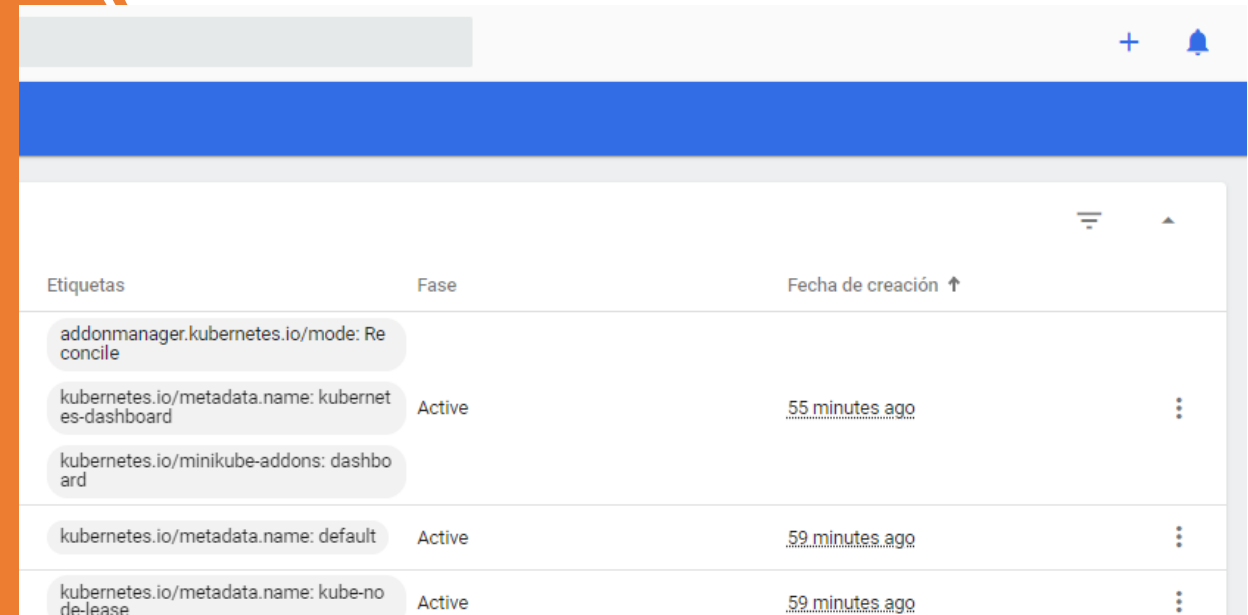
```
C:\Windows\System32>kubectl delete namespace eduardo-estevez  
namespace "eduardo-estevez" deleted
```

```
C:\Windows\System32>kubectl get namespaces
```

NAME	STATUS	AGE
default	Active	49m
kube-node-lease	Active	49m
kube-public	Active	49m
kube-system	Active	49m
kubernetes-dashboard	Active	44m

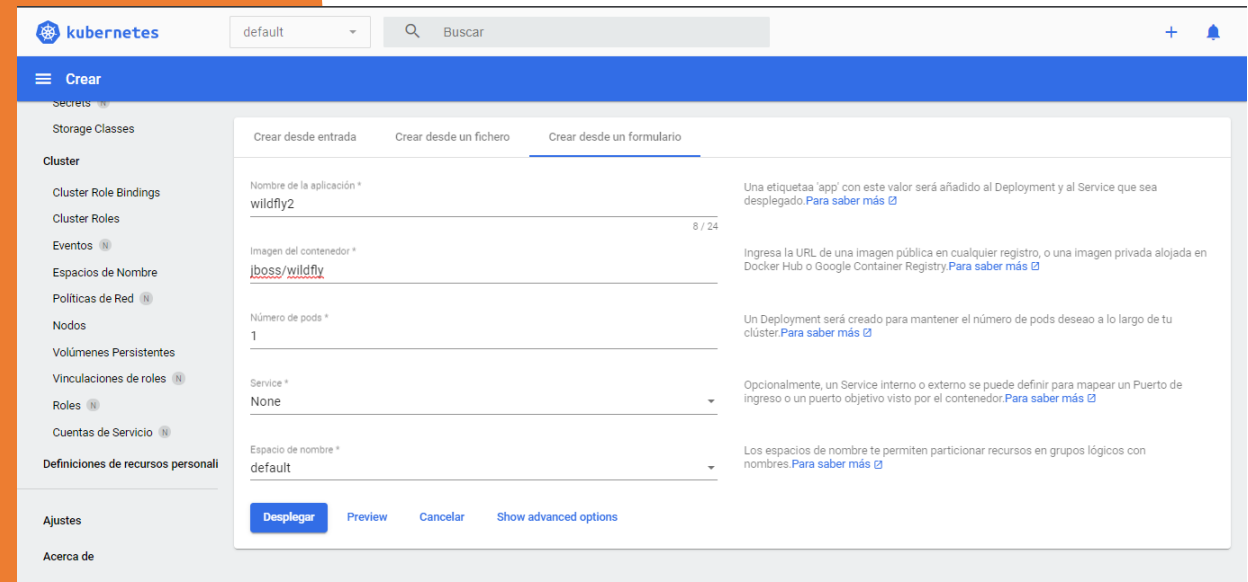
- Crear pod le damos en el mas que esta en la esquina parte superior

- Llenamos los campos de los formularios



The screenshot shows the Kubernetes dashboard with a table of pods. The table has columns for 'Etiquetas' (Labels), 'Fase' (Phase), and 'Fecha de creación' (Creation Time). There are four rows of pods, all in the 'Active' phase. The first row has labels 'addonmanager.kubernetes.io/mode: Reconcile', 'kubernetes.io/metadata.name: kubernetes-dashboard', and 'kubernetes.io/minikube-addons: dashboard'. The second row has label 'kubernetes.io/metadata.name: default'. The third row has label 'kubernetes.io/metadata.name: kube-node-lease'.

Etiquetas	Fase	Fecha de creación ↑
addonmanager.kubernetes.io/mode: Reconcile kubernetes.io/metadata.name: kubernetes-dashboard kubernetes.io/minikube-addons: dashboard	Active	55 minutes ago
kubernetes.io/metadata.name: default	Active	59 minutes ago
kubernetes.io/metadata.name: kube-node-lease	Active	59 minutes ago



The screenshot shows the 'Crear' (Create) form in the Kubernetes dashboard. The form is titled 'Crear desde un formulario' and has several fields for creating a new pod. The fields are: 'Nombre de la aplicación' (Application Name) with value 'wildfly2', 'Imagen del contenedor' (Container Image) with value 'jboss/wildfly', 'Número de pods' (Number of Pods) with value '1', 'Service' with value 'None', and 'Espacio de nombre' (Namespace) with value 'default'. There are also buttons for 'Desplegar' (Deploy), 'Preview', 'Cancelar' (Cancel), and 'Show advanced options'.

Crear desde entrada Crear desde un fichero Crear desde un formulario

Nombre de la aplicación *
wildfly2

Imagen del contenedor *
jboss/wildfly

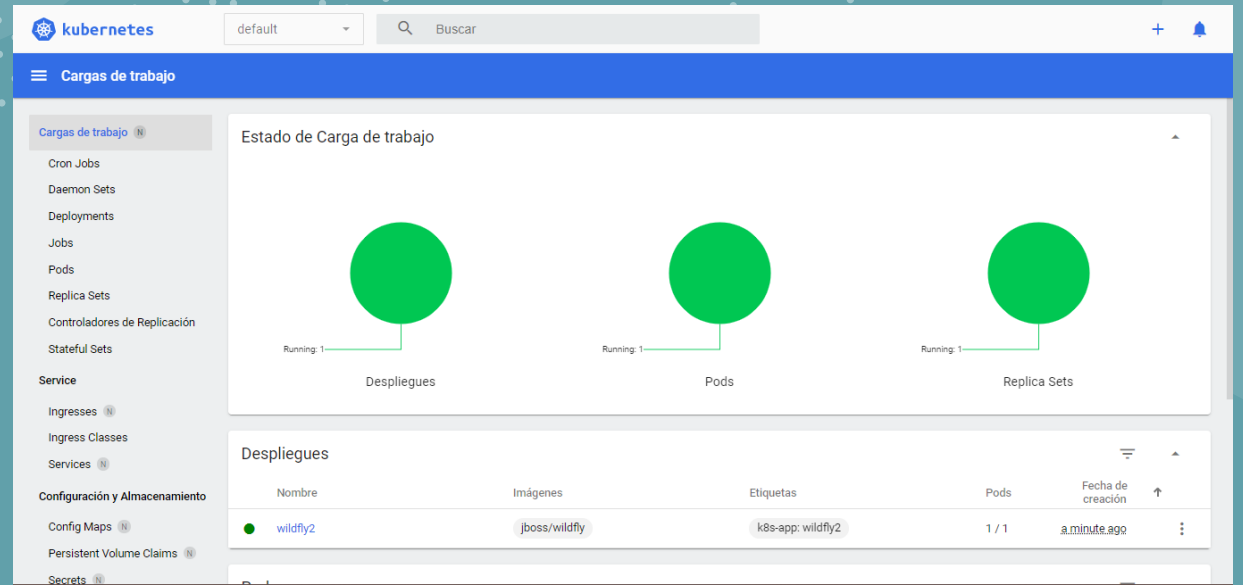
Número de pods *
1

Service *
None

Espacio de nombre *
default

Desplegar Preview Cancelar Show advanced options

- Y vemos que ya esta cargado el pod



- Para ver por la línea de comando es:
- `kubectl get pods`

```
C:\Windows\System32>kubectl get pods
NAME                                READY   STATUS    RESTARTS   AGE
wildfly2-76ff9769bd-dhvhr          1/1     Running   0           2m47s

C:\Windows\System32>
```

- Y con el siguiente comando podemos ver la descripción del pod que hemos creado:

- `kubectl describe pod wildfly2`

- Para acceder a un pod podemos ir a la siguiente parte:

```
C:\Windows\System32>kubectl describe pod wildfly2
Error from server (NotFound): pods "wildfly2" not found

C:\Windows\System32>kubectl describe pod wildfly2
Name:          wildfly2-76ff9769bd-dhvr
Namespace:     default
Priority:       0
Service Account: default
Node:          minikube/192.168.59.100
Start Time:    Sun, 23 Apr 2023 18:48:33 -0500
Labels:        k8s-app=wildfly2
               pod-template-hash=76ff9769bd
Annotations:   <none>
Status:        Running
IP:            10.244.0.5
IPs:           IP: 10.244.0.5
Controlled By: ReplicaSet/wildfly2-76ff9769bd
Containers:
  wildfly2:
    Container ID:  docker://6515b71221ff5c1c4f1c1bfec4d42e412c68be53fb99f45b079bcb5562d40eb
    Image:          jboss/wildfly
    Image ID:       docker-pullable://jboss/wildfly@sha256:35320abafdec6d360559b411aff466514d5741c3c527221445f48246350fdfe5
    Port:          <none>
    Host Port:     <none>
    State:         Running
      Started:     Sun, 23 Apr 2023 18:49:45 -0500
    Ready:         True
    Restart Count: 0
    Environment:   <none>
    Mounts:
      /var/run/secrets/kubernetes.io/serviceaccount from kube-api-access-58x5v (ro)
Conditions:
  Type              Status
  Initialized       True
  Ready             True
  ContainersReady   True
  PodScheduled      True
Volumes:
  kube-api-access-58x5v:
```

The screenshot shows the Kubernetes dashboard interface. The left sidebar contains a navigation menu with options like 'Cargas de trabajo', 'Pods', 'Replica Sets', etc. The main content area displays the details for the pod 'wildfly2-76ff9769bd-dhvr' in the 'default' namespace. It includes a 'Metadatos' section with fields for Name, Namespace, Creation Time, Age, and UID. Below that is the 'Etiquetas' (Labels) section showing 'k8s-app: wildfly2' and 'pod-template-hash: 76ff9769bd'. The 'Información del recurso' (Resource Information) section shows the pod is running on the 'minikube' node with IP '10.244.0.5'. The 'Condiciones' (Conditions) table at the bottom shows the pod is in a 'Ready' state.

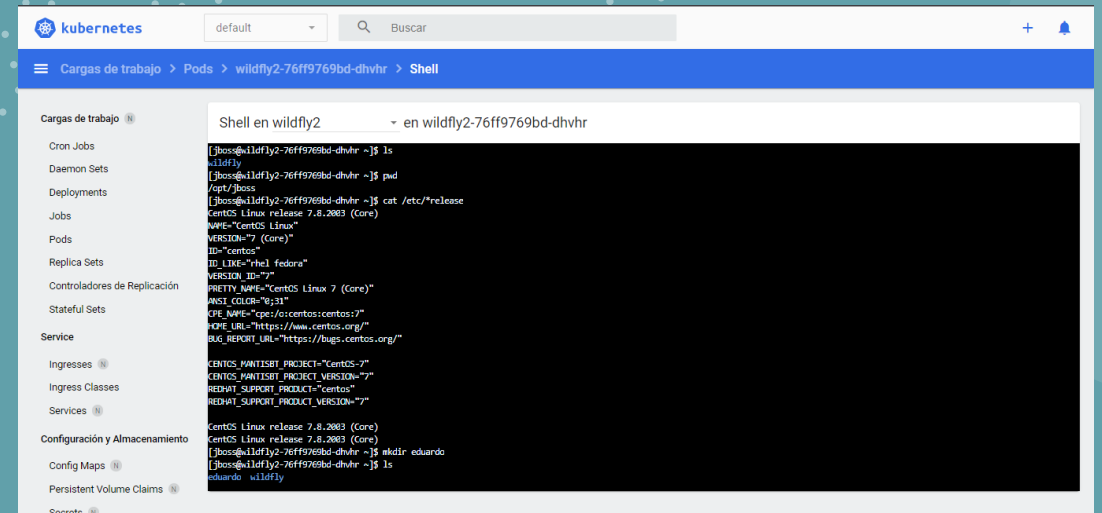
Metadatos					
Nombre	Espacio de nombre	Fecha de creación	Edad	UID	
wildfly2-76ff9769bd-dhvr	default	23.abr.2023	18.minutes.ago	3bb3eba2-c721-4d5b-8d2b-8202a64e808e	

Etiquetas	
Etiqueta	Valor
k8s-app: wildfly2	
pod-template-hash: 76ff9769bd	

Información del recurso					
Nodo	Estado	IP	QoS Class	Reinicios	Service Account
minikube	Running	10.244.0.5	BestEffort	0	default

Condiciones					
Tipo	Estado	Último sondeo	Última transición	Motivo	Mensaje
Initialized	True	-	18.minutes.ago	-	-
Ready	True	-	17.minutes.ago	-	-

- Nos presentará lo siguiente donde podemos ver que es una terminal igual a Linux



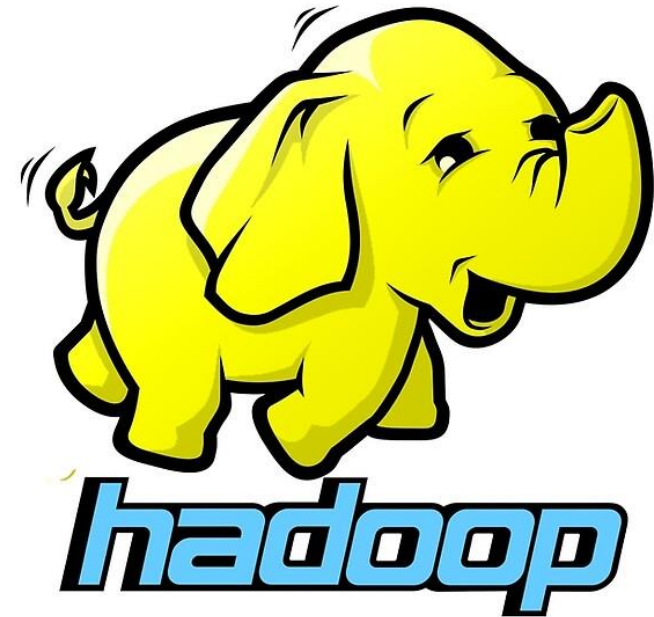
- Otra forma seria por el siguiente comando:

- `exec -it wildfly2-76ff9769bd-dhvhr -- /bin/bash`

```
C:\Windows\System32>kubectl exec -it wildfly2-76ff9769bd-dhvhr -- /bin/bash
[jboss@wildfly2-76ff9769bd-dhvhr ~]$ ls
eduardo wildfly
[jboss@wildfly2-76ff9769bd-dhvhr ~]$ pwd
/opt/jboss
[jboss@wildfly2-76ff9769bd-dhvhr ~]$
```

Hadoop

- Es un framework de software de código abierto para procesamiento distribuido de grandes conjuntos de datos en clusters de computadoras. Es capaz de procesar grandes volúmenes de datos en paralelo y puede escalar horizontalmente para manejar conjuntos de datos cada vez más grandes.
- Hadoop se compone de dos componentes principales: el sistema de almacenamiento distribuido Hadoop Distributed File System (HDFS) y el sistema de procesamiento de datos MapReduce. HDFS distribuye los datos a través del clúster de computadoras para su procesamiento y MapReduce proporciona un modelo de programación para procesar los datos distribuidos en paralelo.



Componentes Principales

Para comprender completamente cómo funciona Hadoop debemos entender sus tres componentes principales. Hemos dedicado una entrada a cada uno de ellos:

- **MapReduce:** Hadoop MapReduce es un paradigma de procesamiento de datos caracterizado por dividirse en dos fases o pasos diferenciados: **Map y Reduce**.
- **Yarn (Yet Another Resource Negotiator):** Yarn (Yet Another Resource Negotiator) es una pieza fundamental en el ecosistema Hadoop. Es el framework que permite a Hadoop soportar varios motores de ejecución incluyendo MapReduce,
- **HDFS (Hadoop Distributed File System):** HDFS (Hadoop Distributed File System) es el componente principal del ecosistema Hadoop. Esta pieza hace posible almacenar data sets masivos con tipos de datos estructurados, semi-estructurados y no estructurados como imágenes, vídeo, datos de sensores, etc.

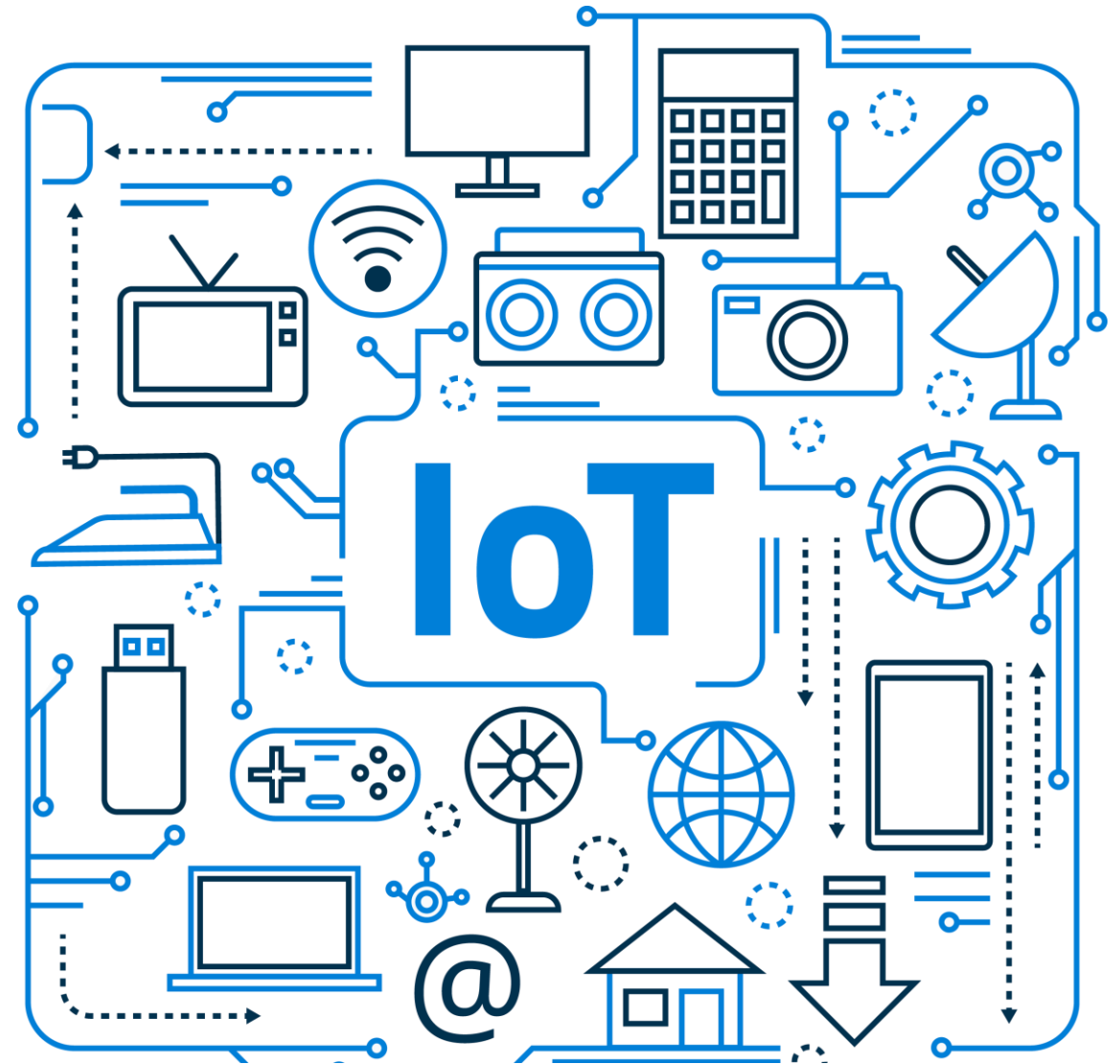


Tecnologías relacionadas

- Los proyectos open source más populares del ecosistema formado alrededor de Apache Hadoop seguramente te sonarán. Se trata de los siguientes:
 - Spark:
 - Ambari:.
 - Oozie:
 - Pig:
 - Tez:
 - Zookeeper:

La función de Hadoop en la IoT

- Una solución que ofrece Hadoop es la capacidad de almacenar y analizar cantidades masivas de datos. El motivo principal que dio lugar a este aumento masivo en la generación de datos es la ola tecnológica actual llamada la “Internet de las cosas” (o IoT, por su nombre en inglés). Esto es cuando los objetos físicos comunes se conectan a Internet y se controlan a través de dicha red.



Ventajas de Hadoop



Escalabilidad



Eficiencia



Reducción de
costes

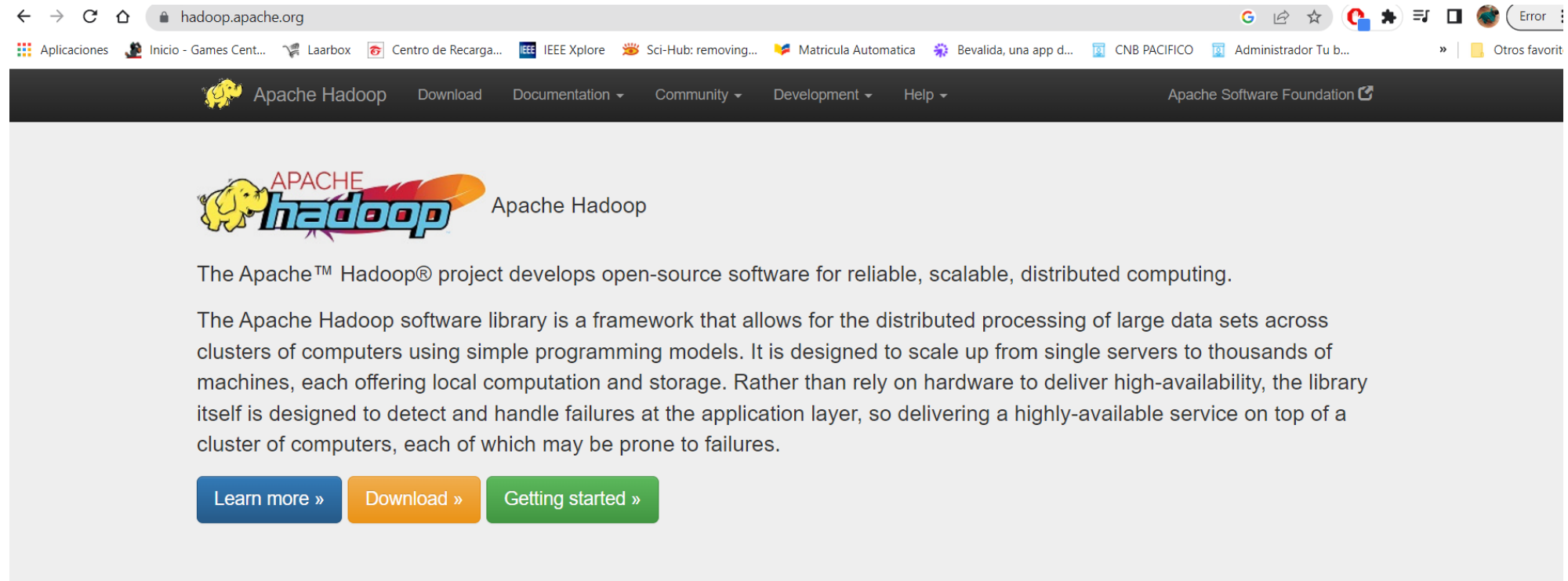


Flexibilidad



Gestión de
errores

Hadoop - Configuración Entorno




The screenshot shows the Apache Hadoop website. The browser's address bar displays 'hadoop.apache.org'. The website's navigation bar includes links for 'Download', 'Documentation', 'Community', 'Development', and 'Help', along with the 'Apache Software Foundation' logo. The main content area features the Apache Hadoop logo, which consists of a yellow elephant and the text 'APACHE hadoop'. Below the logo, the text reads: 'The Apache™ Hadoop® project develops open-source software for reliable, scalable, distributed computing.' This is followed by a paragraph describing the software library as a framework for distributed processing of large data sets across clusters of computers. At the bottom of the main content area, there are three buttons: 'Learn more »' (blue), 'Download »' (orange), and 'Getting started »' (green).

hadoop.apache.org

Aplicaciones Inicio - Games Cent... Laarbox Centro de Recarga... IEEE Xplore Sci-Hub: removing... Matricula Automatica Bevalida, una app d... CNB PACIFICO Administrador Tu b... Otros favorit

Apache Hadoop Download Documentation Community Development Help Apache Software Foundation

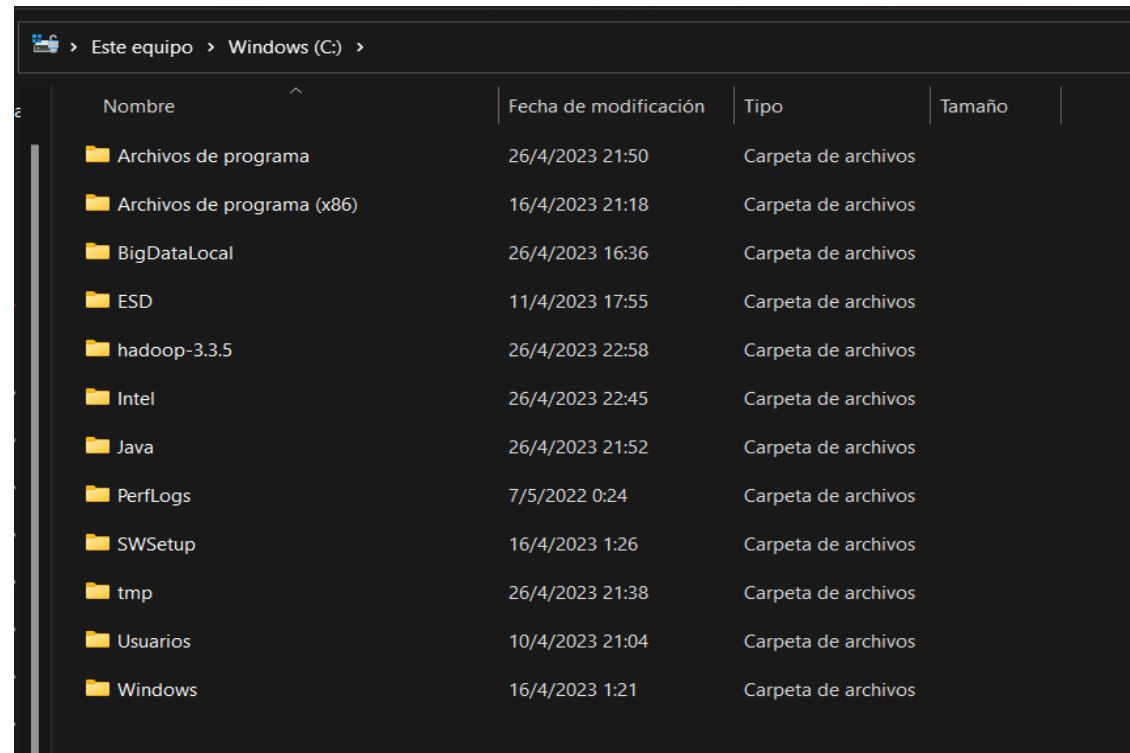
 Apache Hadoop

The Apache™ Hadoop® project develops open-source software for reliable, scalable, distributed computing.

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.

[Learn more »](#) [Download »](#) [Getting started »](#)

Hadoop - Configuración Entorno

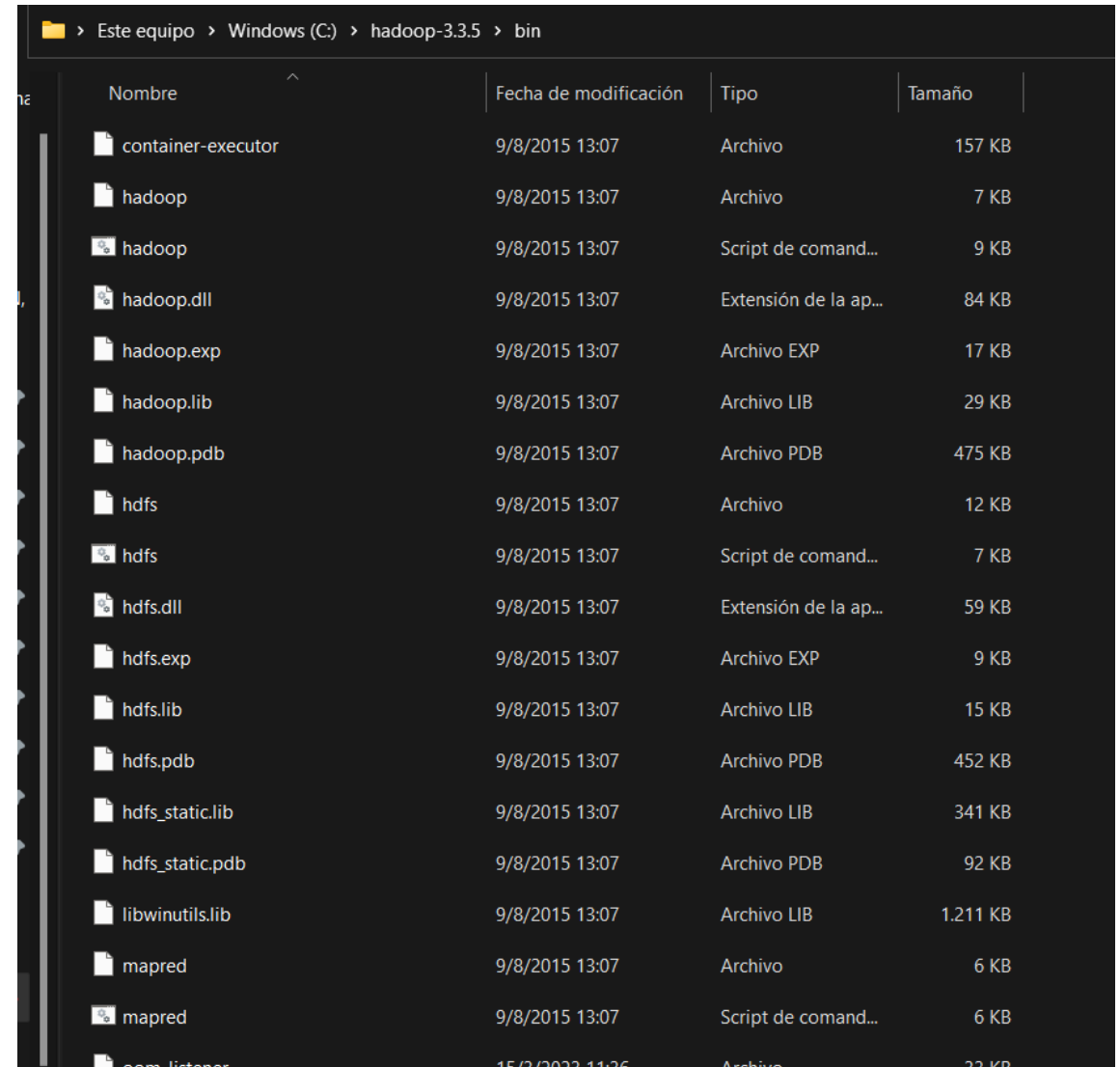


A screenshot of a Windows File Explorer window showing the contents of the C: drive. The address bar indicates the path is 'Este equipo > Windows (C:) >'. The file list is displayed in a table with columns for 'Nombre', 'Fecha de modificación', 'Tipo', and 'Tamaño'. The files listed are all folders, including 'Archivos de programa', 'Archivos de programa (x86)', 'BigDataLocal', 'ESD', 'hadoop-3.3.5', 'Intel', 'Java', 'PerfLogs', 'SWSetup', 'tmp', 'Usuarios', and 'Windows'.

Nombre	Fecha de modificación	Tipo	Tamaño
Archivos de programa	26/4/2023 21:50	Carpeta de archivos	
Archivos de programa (x86)	16/4/2023 21:18	Carpeta de archivos	
BigDataLocal	26/4/2023 16:36	Carpeta de archivos	
ESD	11/4/2023 17:55	Carpeta de archivos	
hadoop-3.3.5	26/4/2023 22:58	Carpeta de archivos	
Intel	26/4/2023 22:45	Carpeta de archivos	
Java	26/4/2023 21:52	Carpeta de archivos	
PerfLogs	7/5/2022 0:24	Carpeta de archivos	
SWSetup	16/4/2023 1:26	Carpeta de archivos	
tmp	26/4/2023 21:38	Carpeta de archivos	
Usuarios	10/4/2023 21:04	Carpeta de archivos	
Windows	16/4/2023 1:21	Carpeta de archivos	

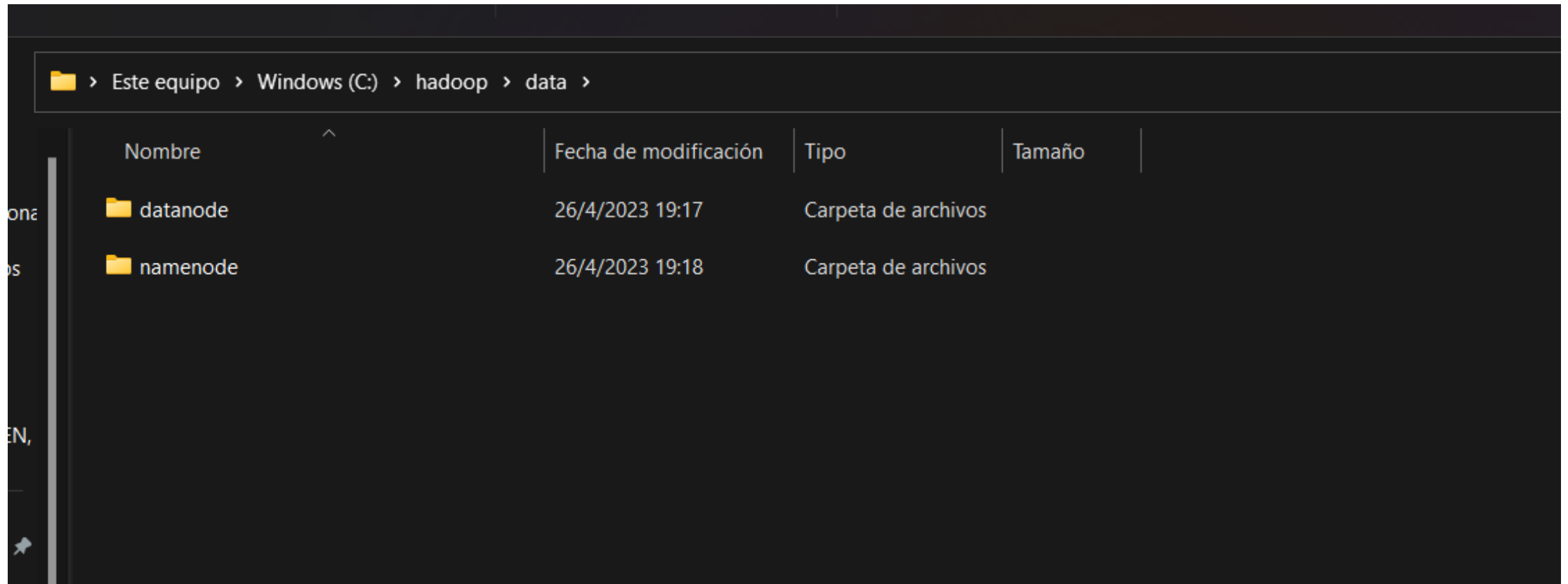
Descargar binarios compatibles con Windows

- Para descargar los archivos necesarios, se descargara del siguiente repositorio “<https://drive.google.com/drive/folders/1Q8bOFv1jVNreTuE34ISGhN7wzJ5b5w1P>”
- Descomprima el zip y copie todos los archivos presentes en la carpeta bin en C:\hadoop-3.3.5\bin También reemplace los archivos existentes.

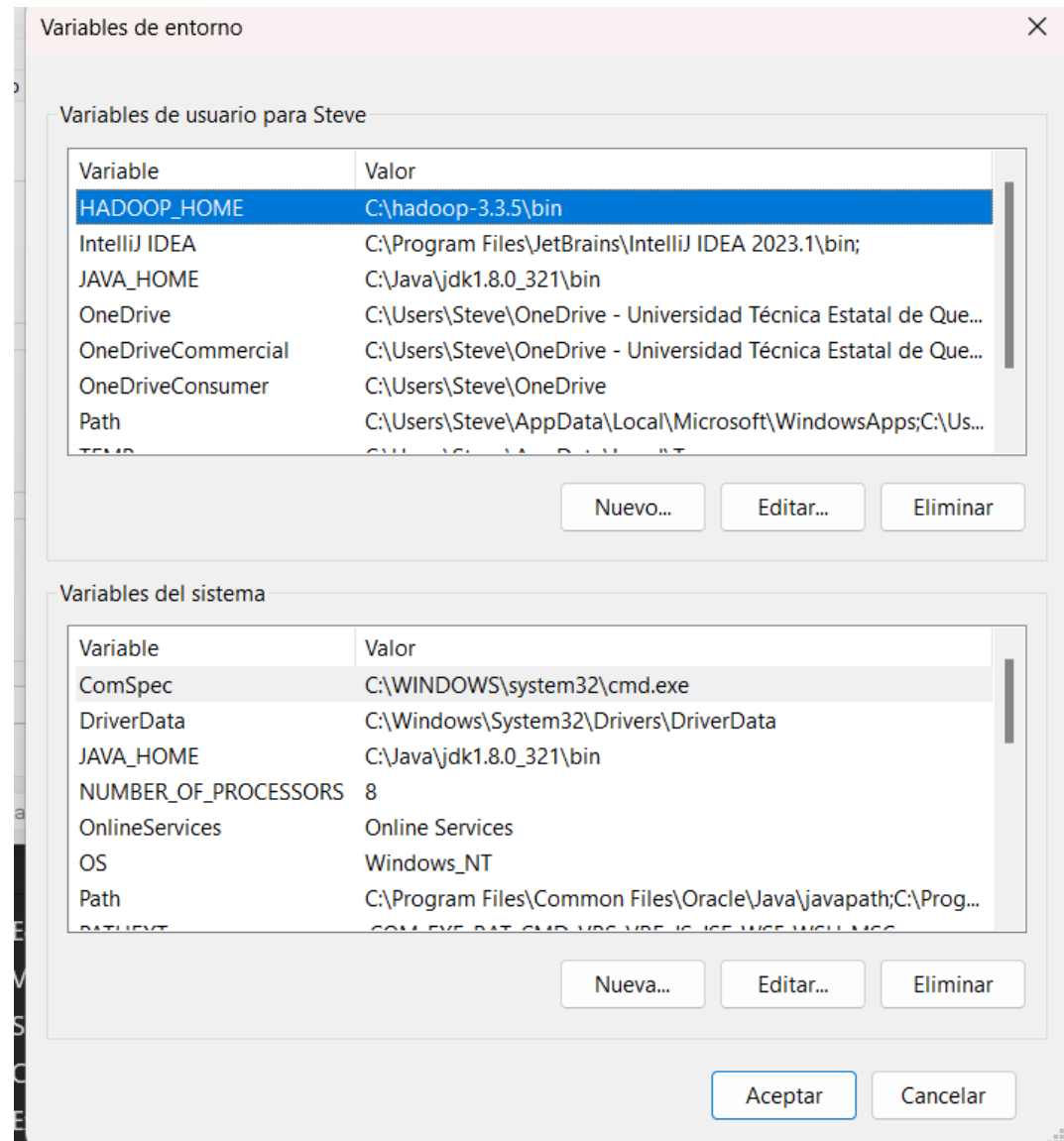


Nombre	Fecha de modificación	Tipo	Tamaño
container-executor	9/8/2015 13:07	Archivo	157 KB
hadoop	9/8/2015 13:07	Archivo	7 KB
hadoop	9/8/2015 13:07	Script de comand...	9 KB
hadoop.dll	9/8/2015 13:07	Extensión de la ap...	84 KB
hadoop.exp	9/8/2015 13:07	Archivo EXP	17 KB
hadoop.lib	9/8/2015 13:07	Archivo LIB	29 KB
hadoop.pdb	9/8/2015 13:07	Archivo PDB	475 KB
hdfs	9/8/2015 13:07	Archivo	12 KB
hdfs	9/8/2015 13:07	Script de comand...	7 KB
hdfs.dll	9/8/2015 13:07	Extensión de la ap...	59 KB
hdfs.exp	9/8/2015 13:07	Archivo EXP	9 KB
hdfs.lib	9/8/2015 13:07	Archivo LIB	15 KB
hdfs.pdb	9/8/2015 13:07	Archivo PDB	452 KB
hdfs_static.lib	9/8/2015 13:07	Archivo LIB	341 KB
hdfs_static.pdb	9/8/2015 13:07	Archivo PDB	92 KB
libwinutils.lib	9/8/2015 13:07	Archivo LIB	1.211 KB
mapred	9/8/2015 13:07	Archivo	6 KB
mapred	9/8/2015 13:07	Script de comand...	6 KB
com.listener	15/3/2023 11:36	Archivo	23 KB

Crear carpetas para datanode y namenode



Establecer Variables de entorno de Hadoop



Establecer Variables de entorno de PATH

C:\Program Files\Common Files\Oracle\Java\javapath
C:\Program Files (x86)\Common Files\Oracle\Java\javapath
C:\Program Files (x86)\VMware\VMware Player\bin\
%SystemRoot%\system32
%SystemRoot%
%SystemRoot%\System32\Wbem
%SYSTEMROOT%\System32\WindowsPowerShell\v1.0\
%SYSTEMROOT%\System32\OpenSSH\
C:\Program Files\Microsoft SQL Server\110\DTS\Binn\
C:\Program Files (x86)\Microsoft SQL Server\110\Tools\Binn\
C:\Program Files\Microsoft SQL Server\110\Tools\Binn\
C:\Program Files (x86)\Microsoft SQL Server\110\Tools\Binn\Ma...
C:\Program Files (x86)\Microsoft Visual Studio 10.0\Common7\I...
C:\Program Files (x86)\Microsoft SQL Server\110\DTS\Binn\
C:\Java\jdk1.8.0_321\bin
C:\hadoop-3.3.5\bin
C:\hadoop-3.3.5\sbin

Nuevo

Editar

Examinar...

Eliminar

Subir

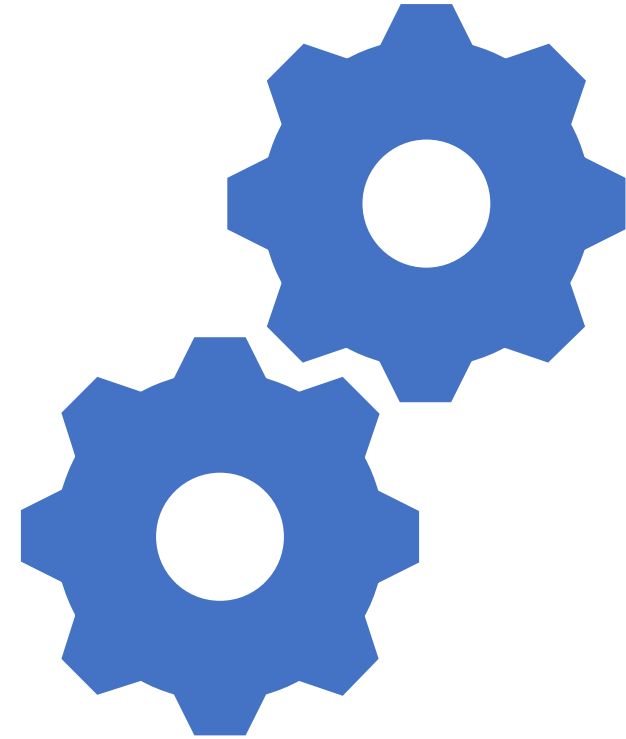
Bajar

Editar texto...

Configurar Hadoop

Después de configurar las variables de entorno, necesitamos configurar Hadoop editando el siguiente archivo de configuración.

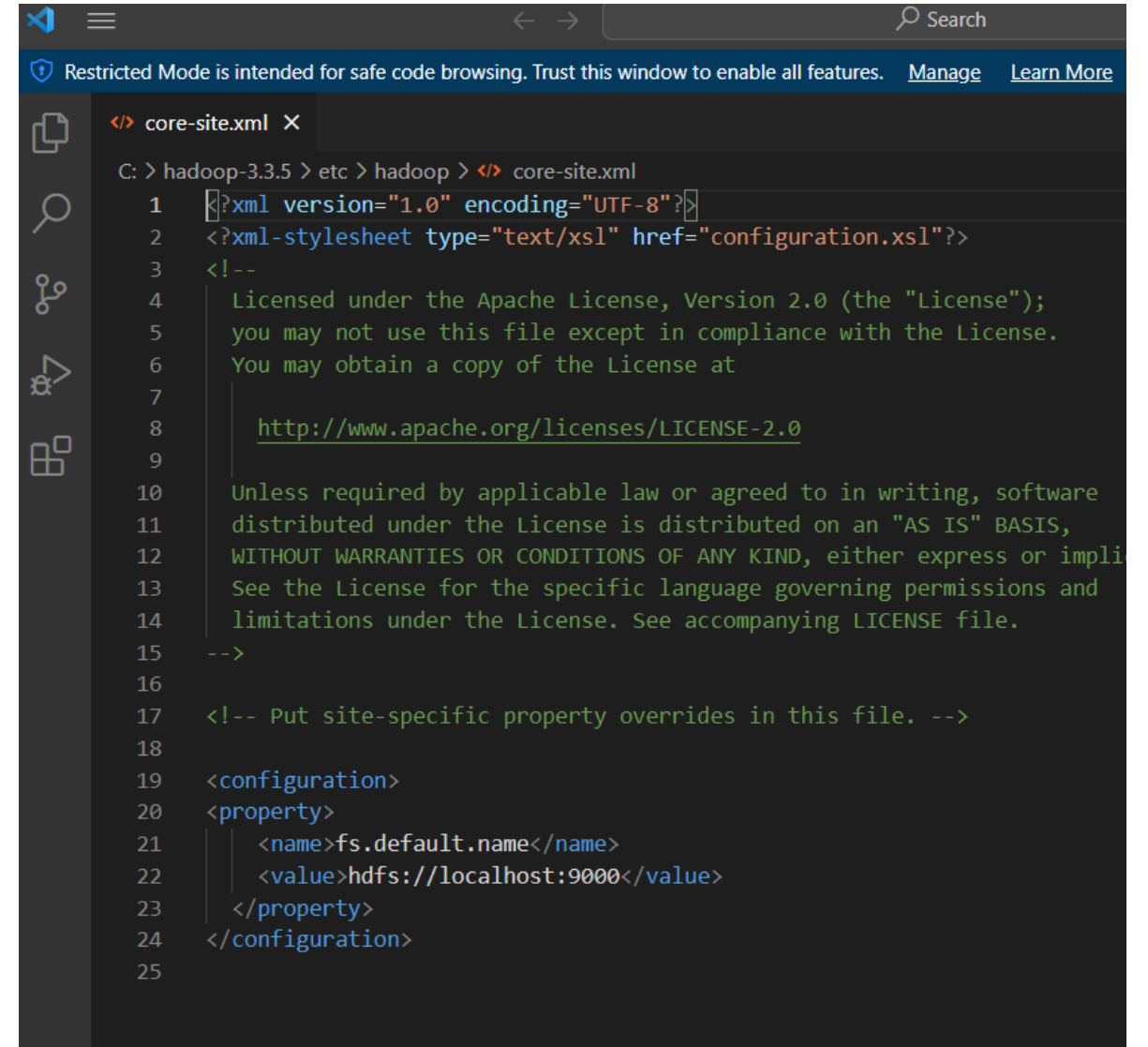
- `hadoop-env.cmd`
- `core-site.xml`
- `hdfs-site.xml`
- `mapred-site.xml`
- `yarn-site`



Configurar Hadoop “core-site.xml”

Ahora, configure los ajustes de Hadoop Core.

Abra C:\hadoop-3.3.5\etc\hadoop\core-site.xml
y ábralo debajo del contenido dentro de la
etiqueta </ configuration>.

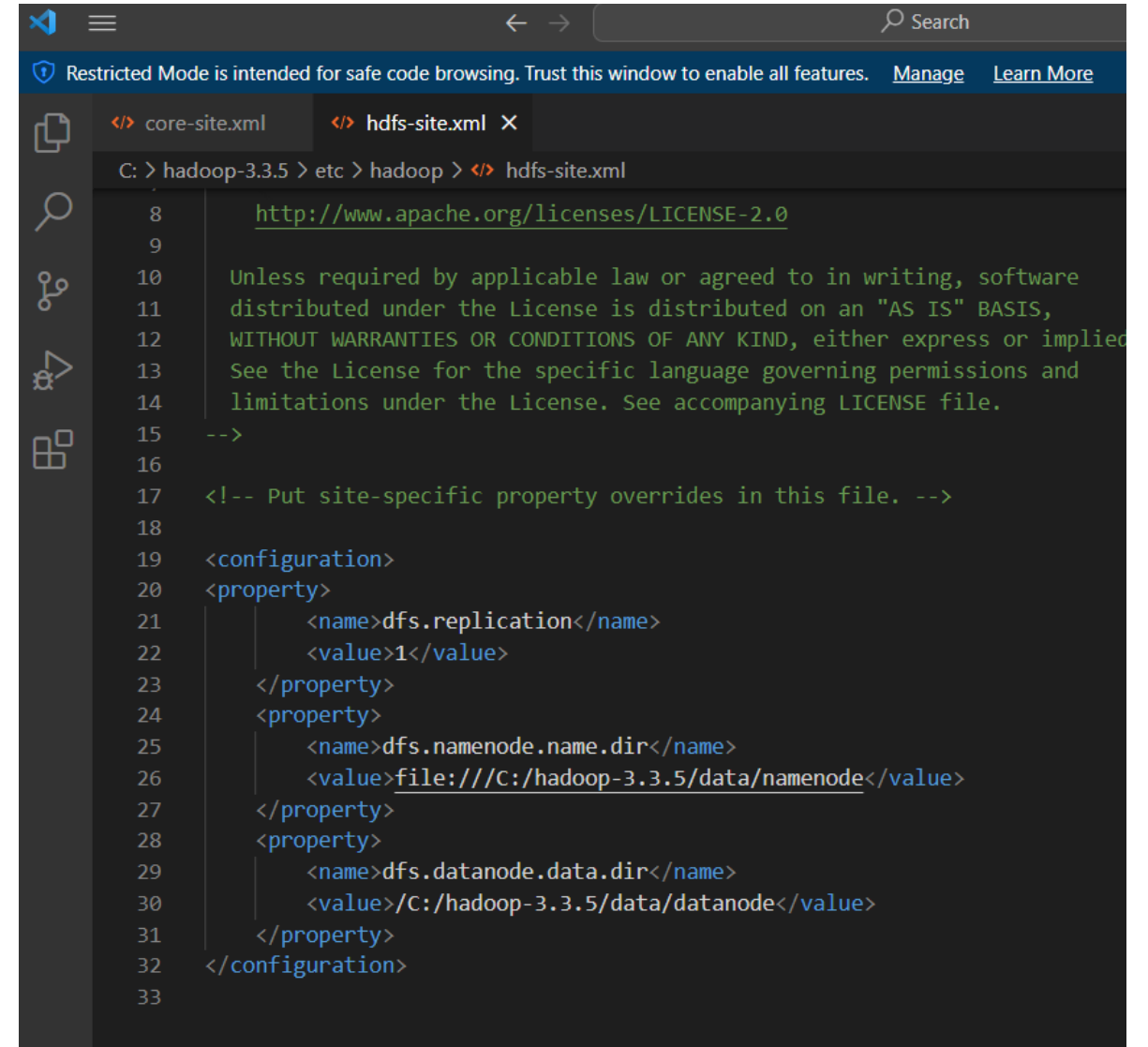


```
C: > hadoop-3.3.5 > etc > hadoop > </> core-site.xml
1  <?xml version="1.0" encoding="UTF-8"?>
2  <?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
3  <!--
4      Licensed under the Apache License, Version 2.0 (the "License");
5      you may not use this file except in compliance with the License.
6      You may obtain a copy of the License at
7
8          http://www.apache.org/licenses/LICENSE-2.0
9
10     Unless required by applicable law or agreed to in writing, software
11     distributed under the License is distributed on an "AS IS" BASIS,
12     WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or impli
13     See the License for the specific language governing permissions and
14     limitations under the License. See accompanying LICENSE file.
15 -->
16
17 <!-- Put site-specific property overrides in this file. -->
18
19 <configuration>
20 <property>
21     <name>fs.default.name</name>
22     <value>hdfs://localhost:9000</value>
23 </property>
24 </configuration>
25
```

Configurar Hadoop “hdfs-site.xml”

Después de editar core-site.xml, debe establecer el factor de replicación y la ubicación del namenode y datanode.

Abra C:\hadoop-3.3.5\etc\hadoop\hdfs-site.xml y ábralo bajo el contenido dentro de la etiqueta </ configuration>.

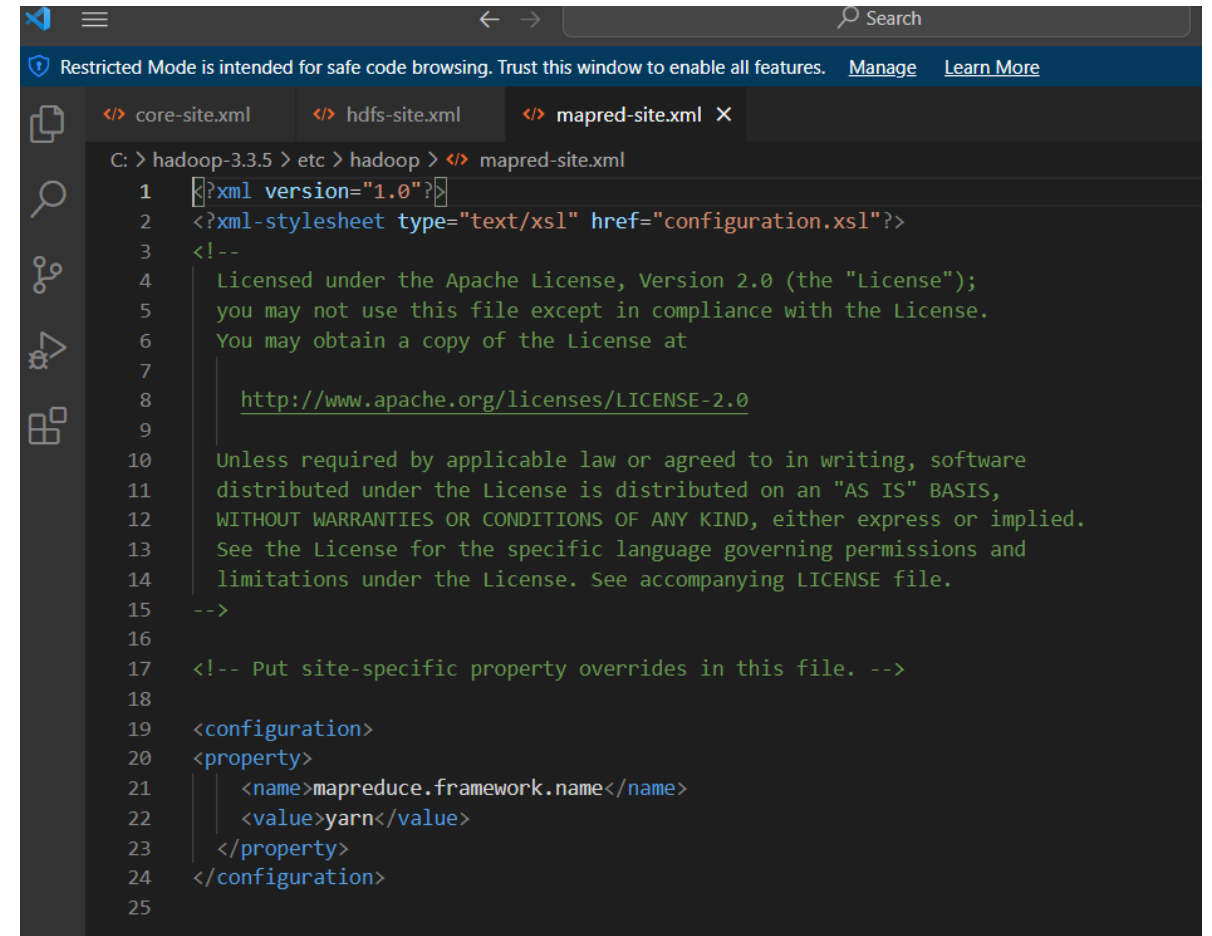


```
8      http://www.apache.org/licenses/LICENSE-2.0
9
10     Unless required by applicable law or agreed to in writing, software
11     distributed under the License is distributed on an "AS IS" BASIS,
12     WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
13     See the License for the specific language governing permissions and
14     limitations under the License. See accompanying LICENSE file.
15     -->
16
17     <!-- Put site-specific property overrides in this file. -->
18
19     <configuration>
20     <property>
21         <name>dfs.replication</name>
22         <value>1</value>
23     </property>
24     <property>
25         <name>dfs.namenode.name.dir</name>
26         <value>file:///C:/hadoop-3.3.5/data/namenode</value>
27     </property>
28     <property>
29         <name>dfs.datanode.data.dir</name>
30         <value>/C:/hadoop-3.3.5/data/datanode</value>
31     </property>
32 </configuration>
33
```

Configurar Hadoop “mapred-site.xml”

Configuremos las propiedades para el marco Map-Reduce.

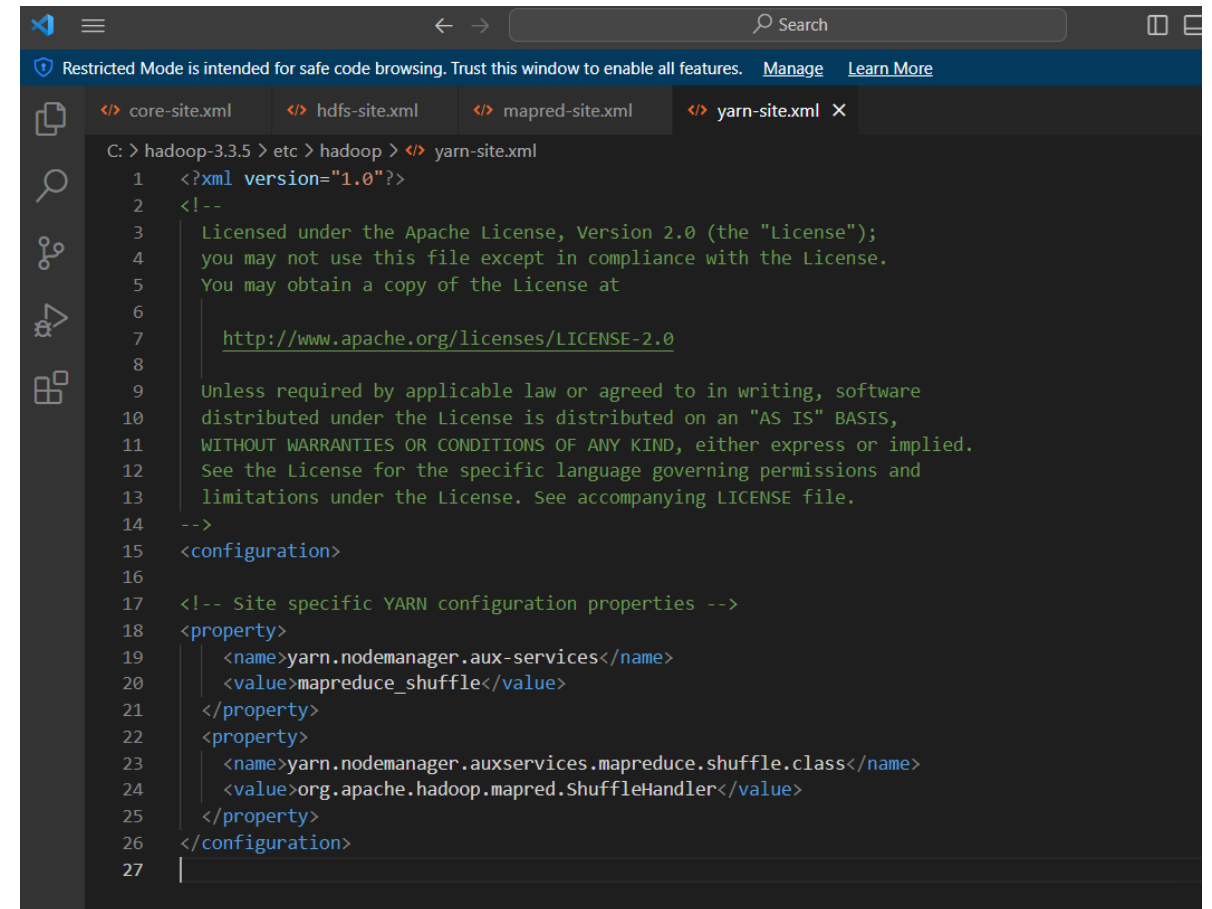
Abra C:\hadoop-3.3.5\etc\hadoop\mapred-site.xml y abra debajo del contenido dentro de la etiqueta </ configuration>. Si no ve mapred-site.xml, abra el archivo mapred-site.xml.template y cámbiele el nombre a mapred-site.xml



```
C: > hadoop-3.3.5 > etc > hadoop > </> mapred-site.xml
1  <?xml version="1.0"?>
2  <?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
3  <!--
4  Licensed under the Apache License, Version 2.0 (the "License");
5  you may not use this file except in compliance with the License.
6  You may obtain a copy of the License at
7
8  http://www.apache.org/licenses/LICENSE-2.0
9
10 Unless required by applicable law or agreed to in writing, software
11 distributed under the License is distributed on an "AS IS" BASIS,
12 WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
13 See the License for the specific language governing permissions and
14 limitations under the License. See accompanying LICENSE file.
15 -->
16
17 <!-- Put site-specific property overrides in this file. -->
18
19 <configuration>
20 <property>
21   <name>mapreduce.framework.name</name>
22   <value>yarn</value>
23 </property>
24 </configuration>
25
```

Configurar Hadoop “yarn-site.xml”

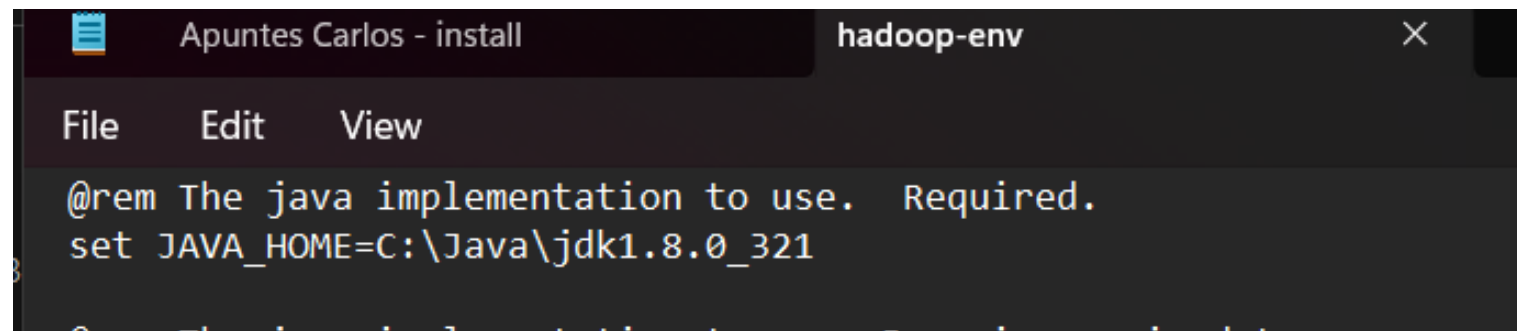
Ahora, configure los ajustes de Hadoop yarn-site
Abra C:\hadoop-3.3.5\etc\hadoop\ yarn-site.xml
y ábralo debajo del contenido dentro de la
etiqueta </ configuration



```
C: > hadoop-3.3.5 > etc > hadoop > yarn-site.xml
1  <?xml version="1.0"?>
2  <!--
3    Licensed under the Apache License, Version 2.0 (the "License");
4    you may not use this file except in compliance with the License.
5    You may obtain a copy of the License at
6
7    http://www.apache.org/licenses/LICENSE-2.0
8
9    Unless required by applicable law or agreed to in writing, software
10   distributed under the License is distributed on an "AS IS" BASIS,
11   WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
12   See the License for the specific language governing permissions and
13   limitations under the License. See accompanying LICENSE file.
14 -->
15 <configuration>
16
17 <!-- Site specific YARN configuration properties -->
18 <property>
19   <name>yarn.nodemanager.aux-services</name>
20   <value>mapreduce_shuffle</value>
21 </property>
22 <property>
23   <name>yarn.nodemanager.auxservices.mapreduce.shuffle.class</name>
24   <value>org.apache.hadoop.mapred.ShuffleHandler</value>
25 </property>
26 </configuration>
27
```

Configurar Hadoop “hadoop-env”

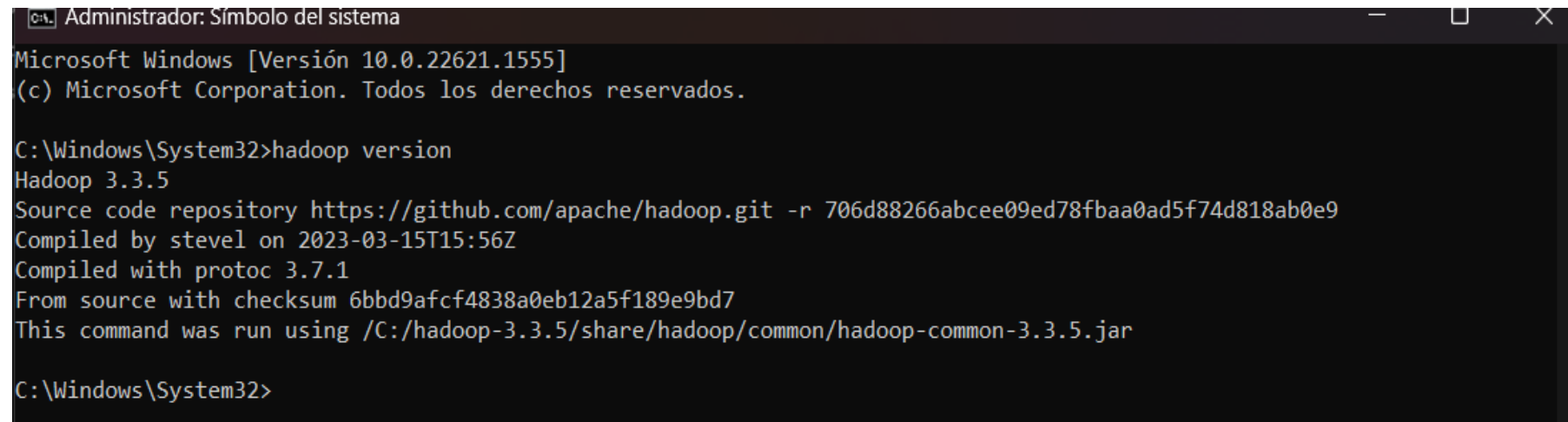
Agregamos la ruta del Jdk que tenemos instalado



```
File Edit View
@rem The java implementation to use. Required.
set JAVA_HOME=C:\Java\jdk1.8.0_321
```

Inicia Hadoop

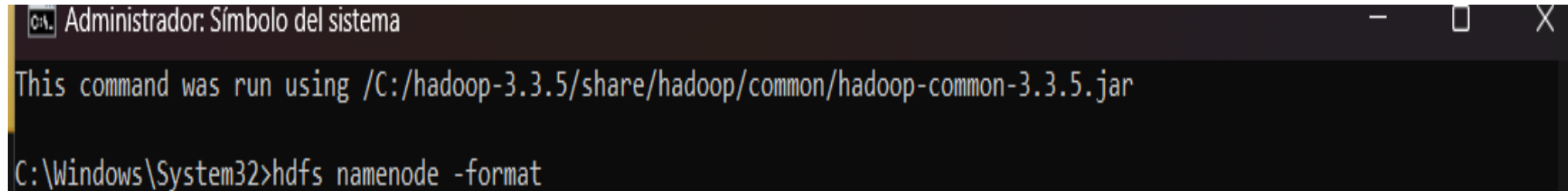
- Primero abrimos un nuevo símbolo del sistema de Windows como administrador y comprobamos la versión de Hadoop version



```
C:\> Administrador: Símbolo del sistema
Microsoft Windows [Versión 10.0.22621.1555]
(c) Microsoft Corporation. Todos los derechos reservados.

C:\Windows\System32>hadoop version
Hadoop 3.3.5
Source code repository https://github.com/apache/hadoop.git -r 706d88266abcee09ed78fbaa0ad5f74d818ab0e9
Compiled by stevel on 2023-03-15T15:56Z
Compiled with protoc 3.7.1
From source with checksum 6bbd9afcf4838a0eb12a5f189e9bd7
This command was run using /C:/hadoop-3.3.5/share/hadoop/common/hadoop-common-3.3.5.jar

C:\Windows\System32>
```



```
Administrador: Símbolo del sistema
This command was run using /C:/hadoop-3.3.5/share/hadoop/common/hadoop-common-3.3.5.jar
C:\Windows\System32>hdfs namenode -format
```

Solo por primera vez para instalar usaremos el siguiente comando

`hdfs namenode -format`


```
C:\%_ Administrador: Símbolo del sistema
mage.ckpt_00000000000000000000 using no compression
2023-04-26 22:59:11,989 INFO namenode.FSImageFormatProtobuf: Image file C:\hadoop-3.3.5\data\namenode\current\fsimage.ck
pt_00000000000000000000 of size 400 bytes saved in 0 seconds .
2023-04-26 22:59:12,009 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
2023-04-26 22:59:12,024 INFO namenode.FSNamesystem: Stopping services started for active state
2023-04-26 22:59:12,024 INFO namenode.FSNamesystem: Stopping services started for standby state
2023-04-26 22:59:12,036 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid=0 when meet shutdown.
2023-04-26 22:59:12,036 INFO namenode.NameNode: SHUTDOWN_MSG:
/*****
SHUTDOWN_MSG: Shutting down NameNode at BRYAN/192.168.154.1
*****/
```

Una vez ejecutado esperamos a que termine el proceso.

Luego iniciamos Hadoop

```
Administrador: Símbolo del sistema
***** /

C:\Windows\System32>cd C:\hadoop-3.3.5\sbin

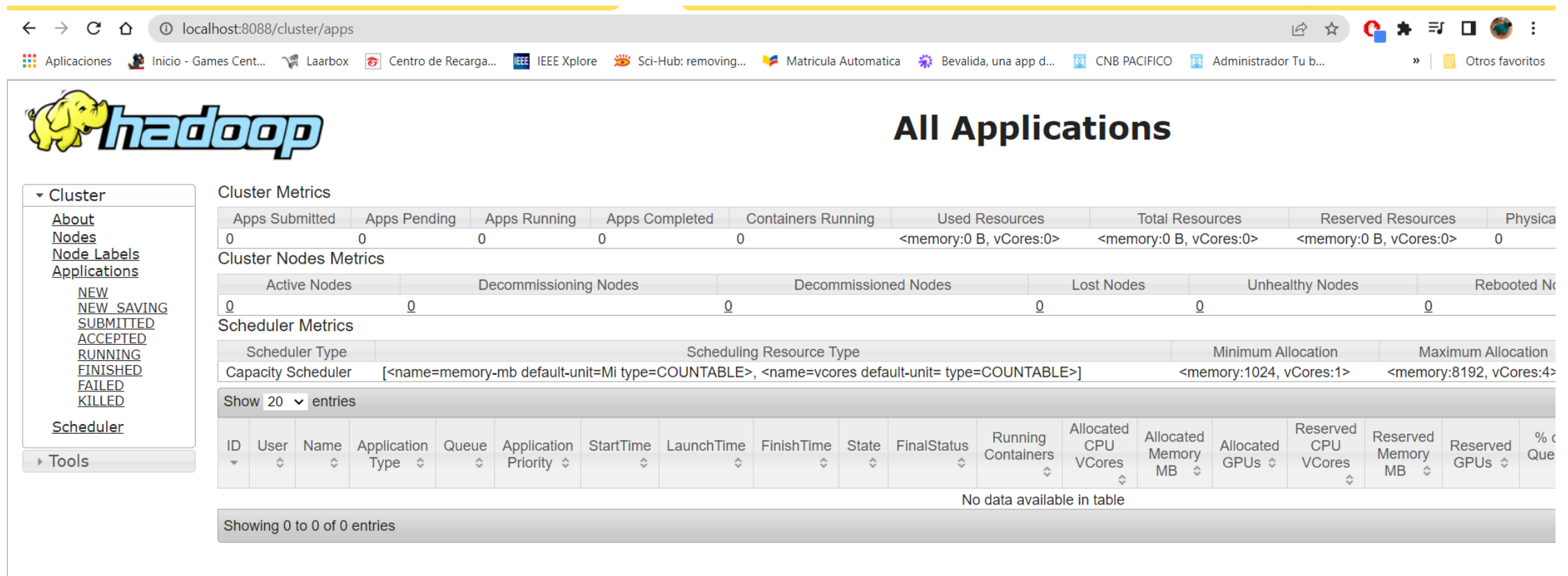
C:\hadoop-3.3.5\sbin>start-all
This script is Deprecated. Instead use start-dfs.cmd and start-yarn.cmd
starting yarn daemons
```

anager y



Interfaz de usuario web de Hadoop

- Finalmente, controlemos el funcionamiento del demonio Hadoop. Sin mencionar que puede usar la interfaz de usuario web para diversas actividades de administración y monitoreo. Abre tu navegador y comienza a usarlo
- Abrir <http://localhost:8088/cluster/apps> para abrir el administrador de recursos



The screenshot displays the Hadoop web interface at localhost:8088/cluster/apps. The interface features a sidebar on the left with navigation links: Cluster, About, Nodes, Node Labels, Applications, NEW, NEW SAVING, SUBMITTED, ACCEPTED, RUNNING, FINISHED, FAILED, KILLED, Scheduler, and Tools. The main content area is titled "All Applications" and contains several data tables.

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Used Resources	Total Resources	Reserved Resources	Physical Memory
0	0	0	0	0	<memory:0 B, vCores:0>	<memory:0 B, vCores:0>	<memory:0 B, vCores:0>	0

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes
0	0	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:4>

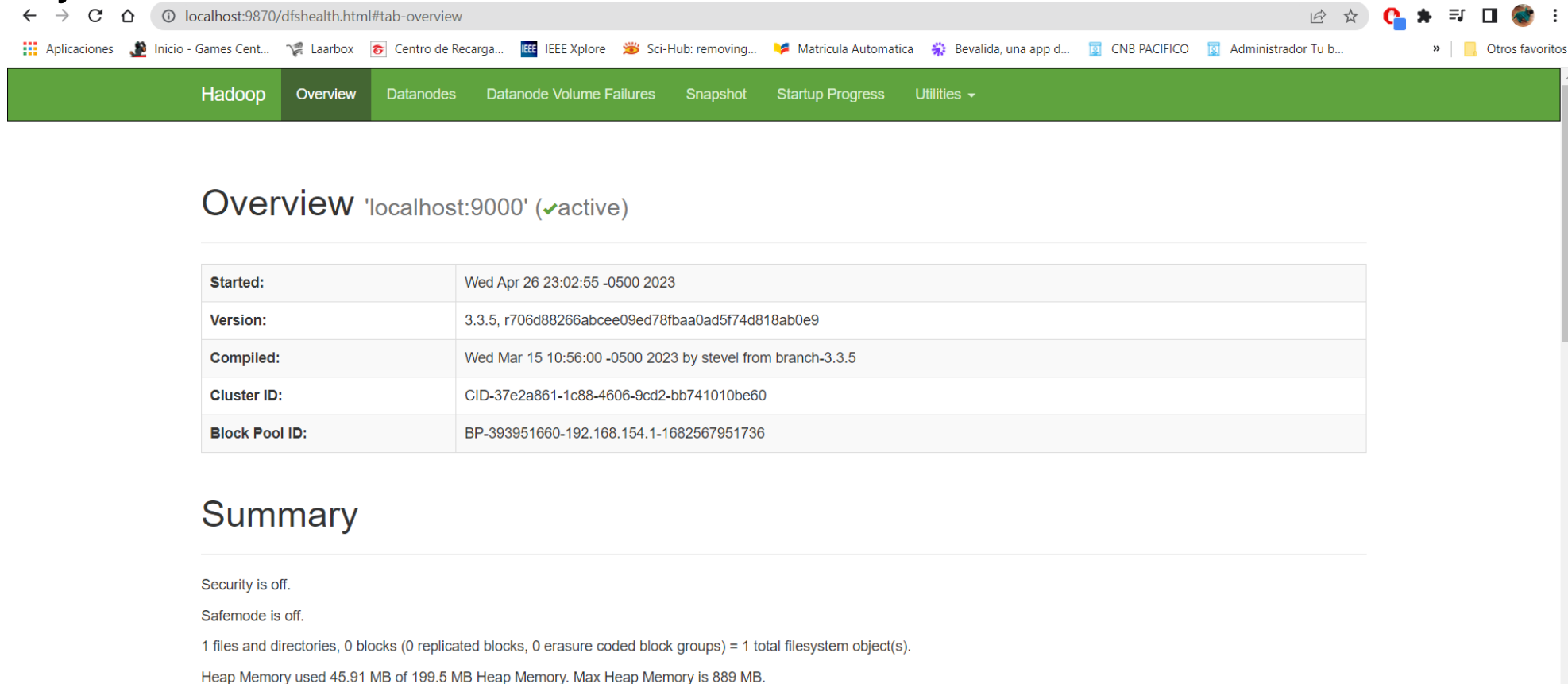
Applications Table

Show 20 entries

ID	User	Name	Application Type	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	Allocated GPUs	Reserved CPU VCores	Reserved Memory MB	Reserved GPUs	% of Queue
No data available in table																		

Showing 0 to 0 of 0 entries

Abra <http://localhost:9870/dfshealth.html#tab-overview> para verificar el estado de ejecución del nodo de nombre



The screenshot shows a web browser window with the address bar displaying `localhost:9870/dfshealth.html#tab-overview`. The browser's address bar and tabs are visible at the top. Below the browser window, the Hadoop DFS Health Overview page is shown. The page has a green header bar with the following navigation items: Hadoop, Overview (selected), Datanodes, Datanode Volume Failures, Snapshot, Startup Progress, and Utilities. The main content area is titled "Overview 'localhost:9000' (✓active)". Below the title is a table with the following information:

Started:	Wed Apr 26 23:02:55 -0500 2023
Version:	3.3.5, r706d88266abcee09ed78fbaa0ad5f74d818ab0e9
Compiled:	Wed Mar 15 10:56:00 -0500 2023 by stevel from branch-3.3.5
Cluster ID:	CID-37e2a861-1c88-4606-9cd2-bb741010be60
Block Pool ID:	BP-393951660-192.168.154.1-1682567951736

Below the table is a section titled "Summary". The summary text is as follows:

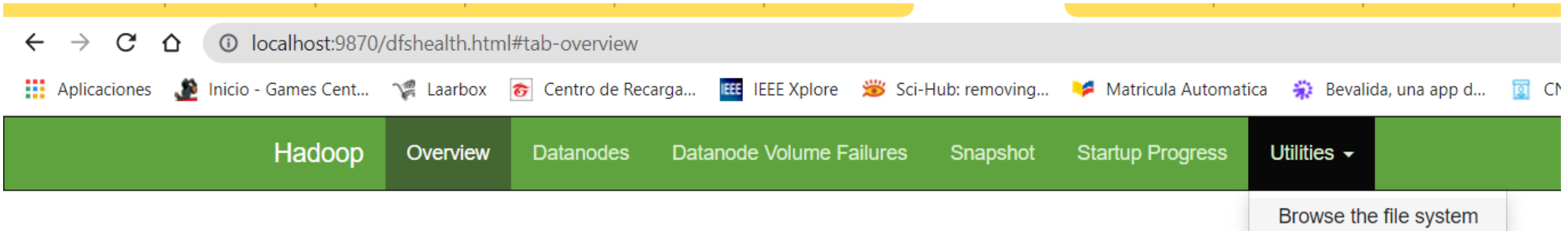
Security is off.
Safemode is off.
1 files and directories, 0 blocks (0 replicated blocks, 0 erasure coded block groups) = 1 total filesystem object(s).
Heap Memory used 45.91 MB of 199.5 MB Heap Memory. Max Heap Memory is 889 MB.

Probando Hadoop

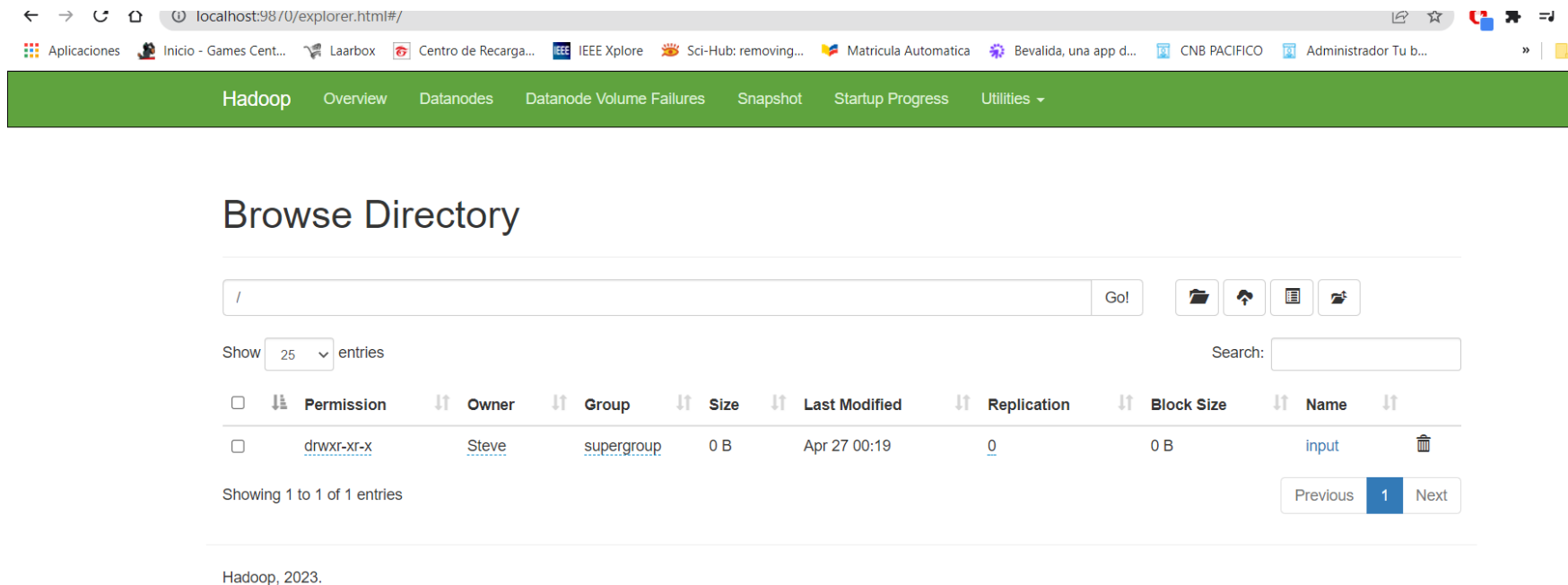
Creación de una carpeta para entrada

```
Administrador: Símbolo del sistema
***** /
C:\Windows\System32>cd C:\hadoop-3.3.5\sbin
C:\hadoop-3.3.5\sbin>start-all
This script is Deprecated. Instead use start-dfs.cmd and start-yarn.cmd
starting yarn daemons
C:\hadoop-3.3.5\sbin>hadoopp fs -mkdir /input
"hadoopp" no se reconoce como un comando interno o externo,
programa o archivo por lotes ejecutable.
C:\hadoop-3.3.5\sbin>hadoop fs -mkdir /input
C:\hadoop-3.3.5\sbin>
```

Nos dirigimos a Utilidades



Y comprobamos que nuestra carpeta se creo







The screenshot shows the Hadoop web interface in a browser. The address bar indicates the URL is `localhost:9870/explorer.html#`. The interface has a green navigation bar with links: Hadoop, Overview, Datanodes, Datanode Volume Failures, Snapshot, Startup Progress, and Utilities. Below this is the "Browse Directory" section. A search bar contains the path `/` and a "Go!" button. To the right are icons for folder, upload, list, and refresh. Below the search bar, it says "Show 25 entries" and a "Search:" input field. A table lists the directory contents with columns: Permission, Owner, Group, Size, Last Modified, Replication, Block Size, Name, and an action icon. One entry is shown: a folder named "input" with permissions "drwxr-xr-x", owner "Steve", group "supergroup", size "0 B", last modified "Apr 27 00:19", and replication "0". At the bottom, it says "Showing 1 to 1 of 1 entries" and a pagination bar with "Previous", "1", and "Next". The footer text reads "Hadoop, 2023."

localhost:9870/explorer.html#


Aplicaciones Inicio - Games Cent... Laarbox Centro de Recarga... IEEE Xplore Sci-Hub: removing... Matricula Automatica Bevalida, una app d... CNB PACIFICO Administrador Tu b...

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities

Browse Directory

/ Go!    

Show 25 entries Search:

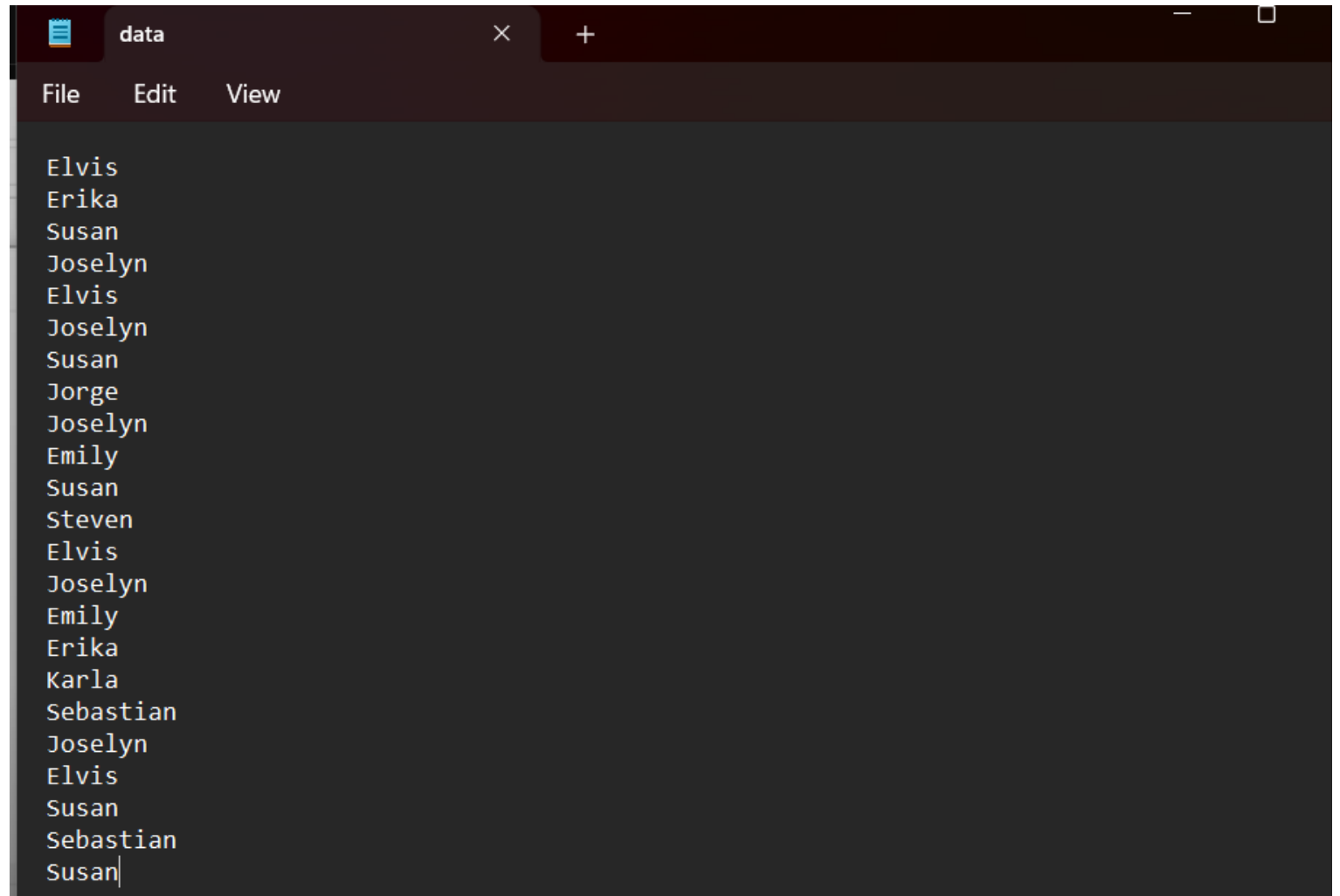
<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
<input type="checkbox"/>	drwxr-xr-x	Steve	supergroup	0 B	Apr 27 00:19	0	0 B	input	

Showing 1 to 1 of 1 entries

Previous 1 Next

Hadoop, 2023.

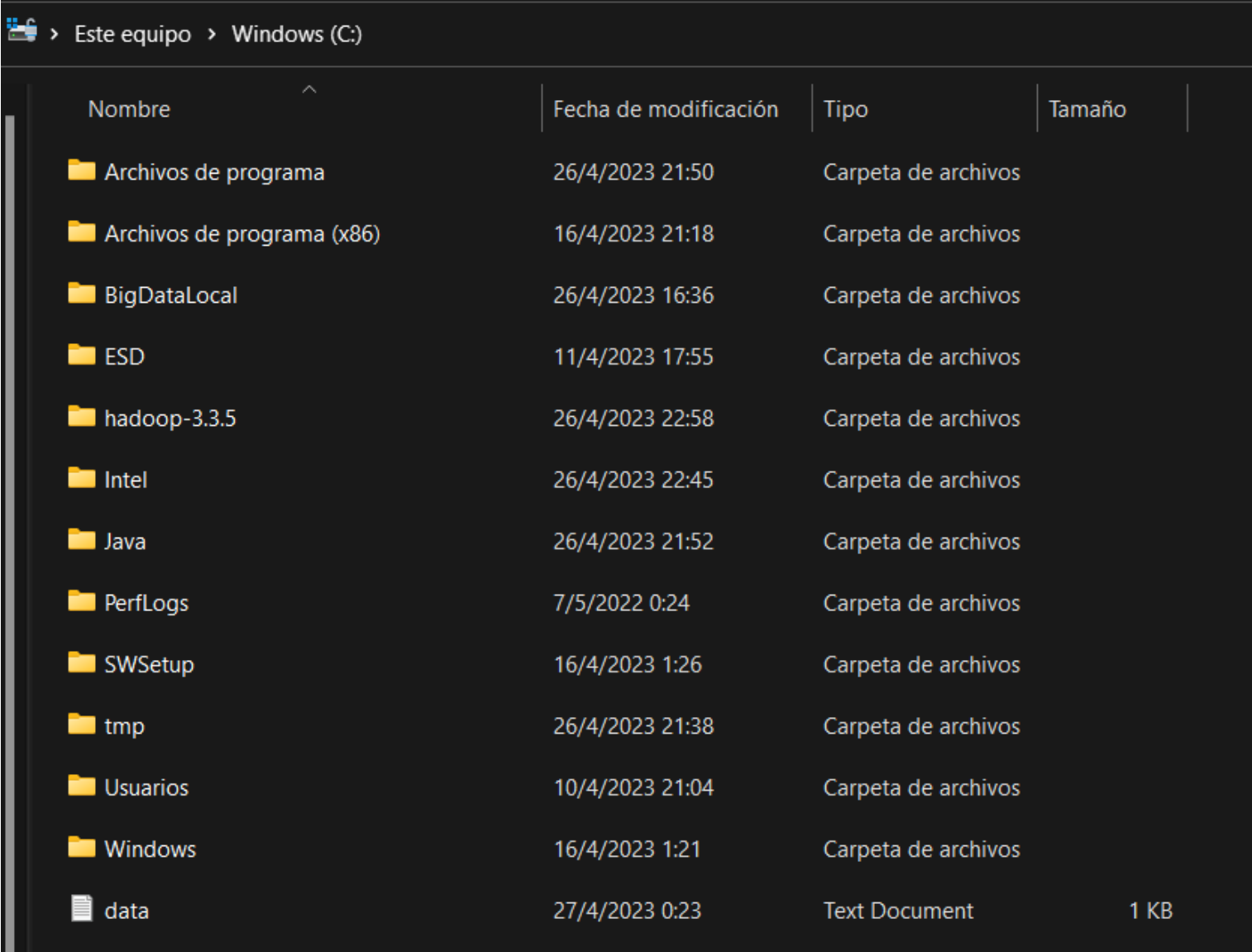
Luego creamos
un Arhivo con el
nombre de Data
con los
siguientes
datos.



A screenshot of a text editor window with a dark theme. The window has a title bar with the name 'data' and standard window controls (close, maximize, and a plus sign for a new tab). Below the title bar is a menu bar with 'File', 'Edit', and 'View'. The main text area contains a list of names, one per line, with a cursor at the end of the last line. The names are: Elvis, Erika, Susan, Joselyn, Elvis, Joselyn, Susan, Jorge, Joselyn, Emily, Susan, Steven, Elvis, Joselyn, Emily, Erika, Karla, Sebastian, Joselyn, Elvis, Susan, Sebastian, and Susan.

```
data
File Edit View
Elvis
Erika
Susan
Joselyn
Elvis
Joselyn
Susan
Jorge
Joselyn
Emily
Susan
Steven
Elvis
Joselyn
Emily
Erika
Karla
Sebastian
Joselyn
Elvis
Susan
Sebastian
Susan|
```

Y lo
colocamos en
el Disco local
C



Nombre	Fecha de modificación	Tipo	Tamaño
Archivos de programa	26/4/2023 21:50	Carpeta de archivos	
Archivos de programa (x86)	16/4/2023 21:18	Carpeta de archivos	
BigDataLocal	26/4/2023 16:36	Carpeta de archivos	
ESD	11/4/2023 17:55	Carpeta de archivos	
hadoop-3.3.5	26/4/2023 22:58	Carpeta de archivos	
Intel	26/4/2023 22:45	Carpeta de archivos	
Java	26/4/2023 21:52	Carpeta de archivos	
PerfLogs	7/5/2022 0:24	Carpeta de archivos	
SWSetup	16/4/2023 1:26	Carpeta de archivos	
tmp	26/4/2023 21:38	Carpeta de archivos	
Usuarios	10/4/2023 21:04	Carpeta de archivos	
Windows	16/4/2023 1:21	Carpeta de archivos	
data	27/4/2023 0:23	Text Document	1 KB

Y lo colocamos
en la carpeta
Input

```
Administrador: Símbolo del sistema
*****/

C:\Windows\System32>cd C:\hadoop-3.3.5\sbin

C:\hadoop-3.3.5\sbin>start-all
This script is Deprecated. Instead use start-dfs.cmd and start-yarn.cmd
starting yarn daemons

C:\hadoop-3.3.5\sbin>hadoopp fs -mkdir /input
"hadoopp" no se reconoce como un comando interno o externo,
programa o archivo por lotes ejecutable.

C:\hadoop-3.3.5\sbin>hadoop fs -mkdir /input

C:\hadoop-3.3.5\sbin>hadoop fs -put C:\data.txt /input_
```

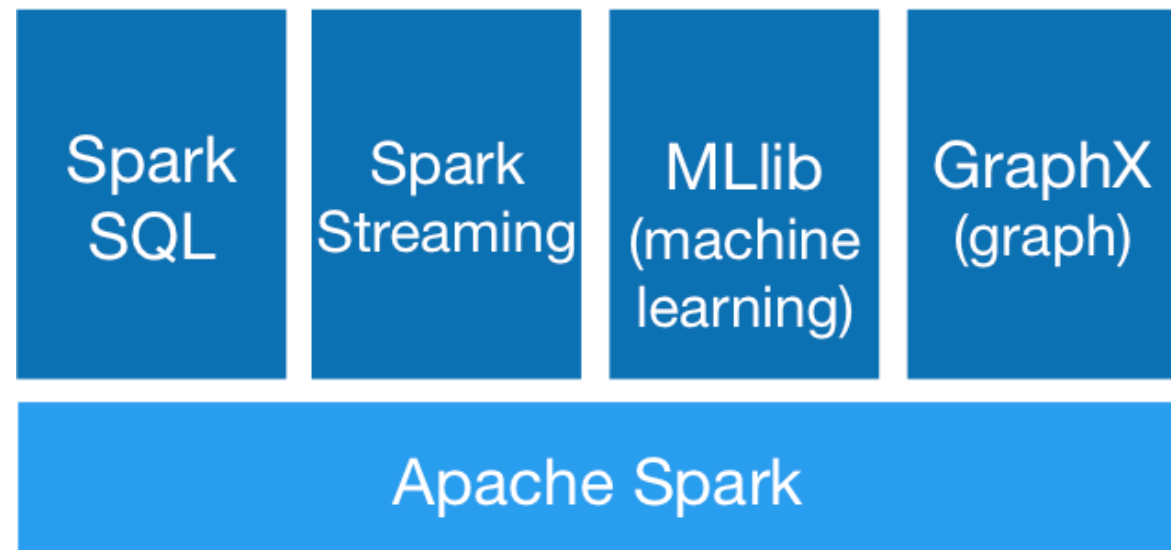
Spark

- Es un sistema de procesamiento de datos distribuido de código abierto que permite procesar grandes cantidades de datos de manera rápida y eficiente. Fue desarrollado por Apache Software Foundation y se basa en el lenguaje de programación Scala.
- Spark ofrece un conjunto de herramientas para el procesamiento de datos en diferentes formatos, incluyendo CSV, JSON, Parquet y Avro, y es compatible con varios sistemas de almacenamiento de datos, como Hadoop Distributed File System (HDFS), Apache Cassandra, Apache HBase y Amazon S3





Componentes de Spark

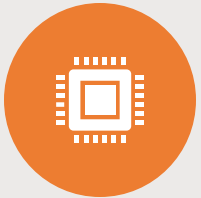


Como funciona Spark

El funcionamiento de Spark se puede resumir en los siguientes pasos:

- Spark distribuye los datos:
- Spark procesa los datos
- Spark almacena los resultados
- Spark optimiza el procesamiento
- Spark gestiona la tolerancia a fallos

Ventajas de Spark



**RENDIMIENTO
SUPERIOR:**



FLEXIBILIDAD:



TOLERANCIA A FALLOS



**INTEGRACIÓN CON
OTRAS TECNOLOGÍAS
DE BIG DATA**

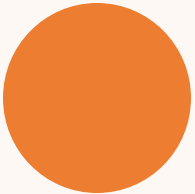


**BIBLIOTECAS DE
MACHINE LEARNING Y
ANÁLISIS DE DATOS**

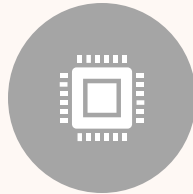


COMUNIDAD ACTIVA

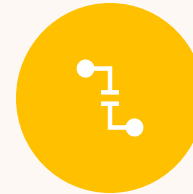
Desventajas de Spark



Requiere una curva de aprendizaje: Spark es una herramienta compleja y requiere una curva de aprendizaje para entender su funcionamiento y utilizarlo de manera efectiva.



Consumo de memoria: Aunque Spark utiliza una arquitectura de procesamiento en memoria.



Dificultades de depuración: El procesamiento distribuido en Spark puede dificultar la depuración de errores en el código,.










Configuración del clúster: La configuración y el mantenimiento de un clúster de Spark pueden ser complicados y requerir habilidades de administración de sistemas




Limitaciones en operaciones complejas: Aunque Spark es capaz de procesar grandes conjuntos de datos, algunas operaciones complejas.

Como Realizar la instalación de Spark en Windows

Descargamos e instalamos primero el JDK8

Mac OS X x64	249.15 MB	 jdk-8u202-macosx-x64.dmg
Solaris SPARC 64-bit (SVR4 package)	125.09 MB	 jdk-8u202-solaris-sparcv9.tar.Z
Solaris SPARC 64-bit	88.1 MB	 jdk-8u202-solaris-sparcv9.tar.gz
Solaris x64 (SVR4 package)	124.37 MB	 jdk-8u202-solaris-x64.tar.Z
Solaris x64	85.38 MB	 jdk-8u202-solaris-x64.tar.gz
Windows x86	201.64 MB	 jdk-8u202-windows-i586.exe
Windows x64	211.58 MB	 jdk-8u202-windows-x64.exe

Luego nos dirigimos a la página oficial de Apache Spark

[Download](#) [Libraries](#) [Documentation](#) [Examples](#) [Community](#) [Developers](#) [Apache Software Foundation](#)

Download Apache Spark™

- Choose a Spark release:
- Choose a package type:
- Download Spark: [spark-3.4.0-bin-hadoop3.tgz](#)
- Verify this release using the 3.4.0 [signatures](#), [checksums](#) and [project release KEYS](#) by following these [procedures](#).

Note that Spark 3 is pre-built with Scala 2.12 in general and Spark 3.2+ provides additional pre-built distribution with Scala 2.13.

Link with Spark

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

```
groupId: org.apache.spark
artifactId: spark-core_2.12
version: 3.4.0
```

Installing with PyPi

[PySpark](#) is now available in pypi. To install just run `pip install pyspark`.

Convenience Docker Container Images

[Spark Docker Container images](#) are available from [DockerHub](#), these images contain non-ASF software and may be subject to

Latest News


[Spark 3.4.0 released](#) (Apr 13, 2023)

[Spark 3.2.4 released](#) (Apr 13, 2023)

[Spark 3.3.2 released](#) (Feb 17, 2023)

[Spark 3.2.3 released](#) (Nov 28, 2022)

[Archive](#)

[LEARN MORE](#)

[DOWNLOAD SPARK](#)

Built-in Libraries:

[SQL and DataFrames](#)

[Spark Streaming](#)

[MLlib \(machine learning\)](#)

[GraphX \(graph\)](#)

[Third-Party Projects](#)

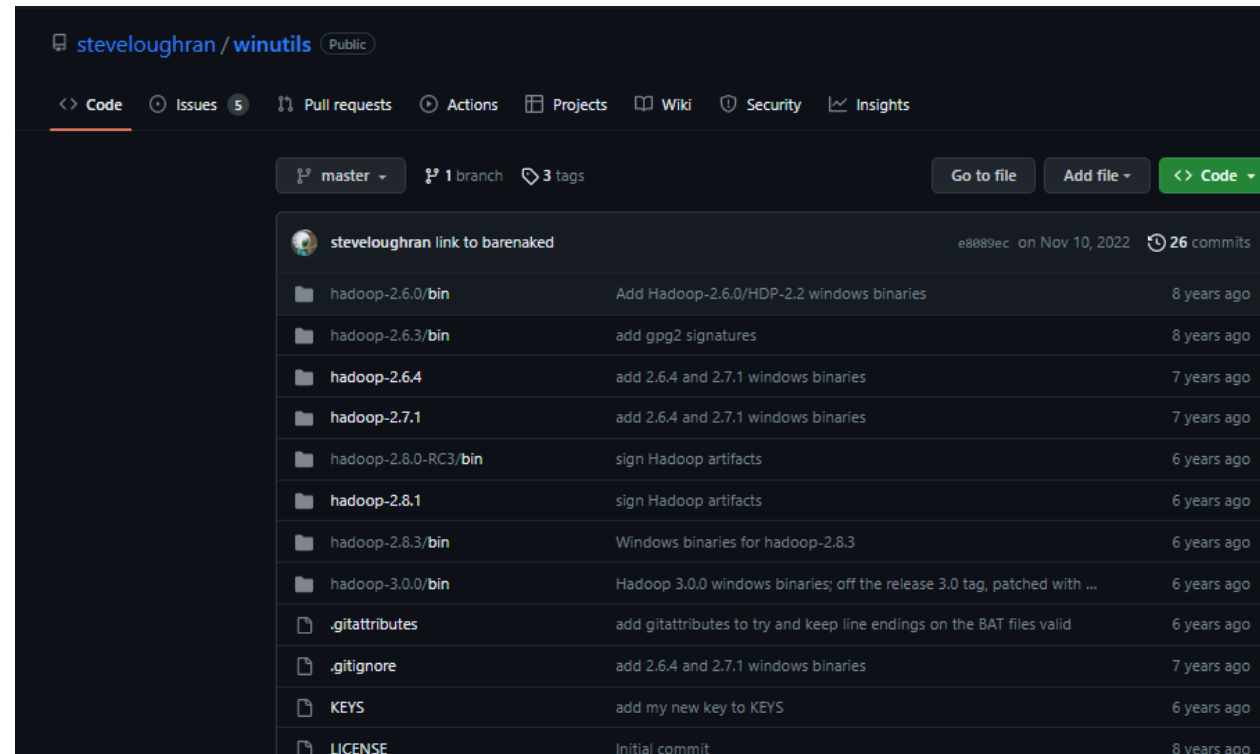
Descargamos Spark

Download Apache Spark™

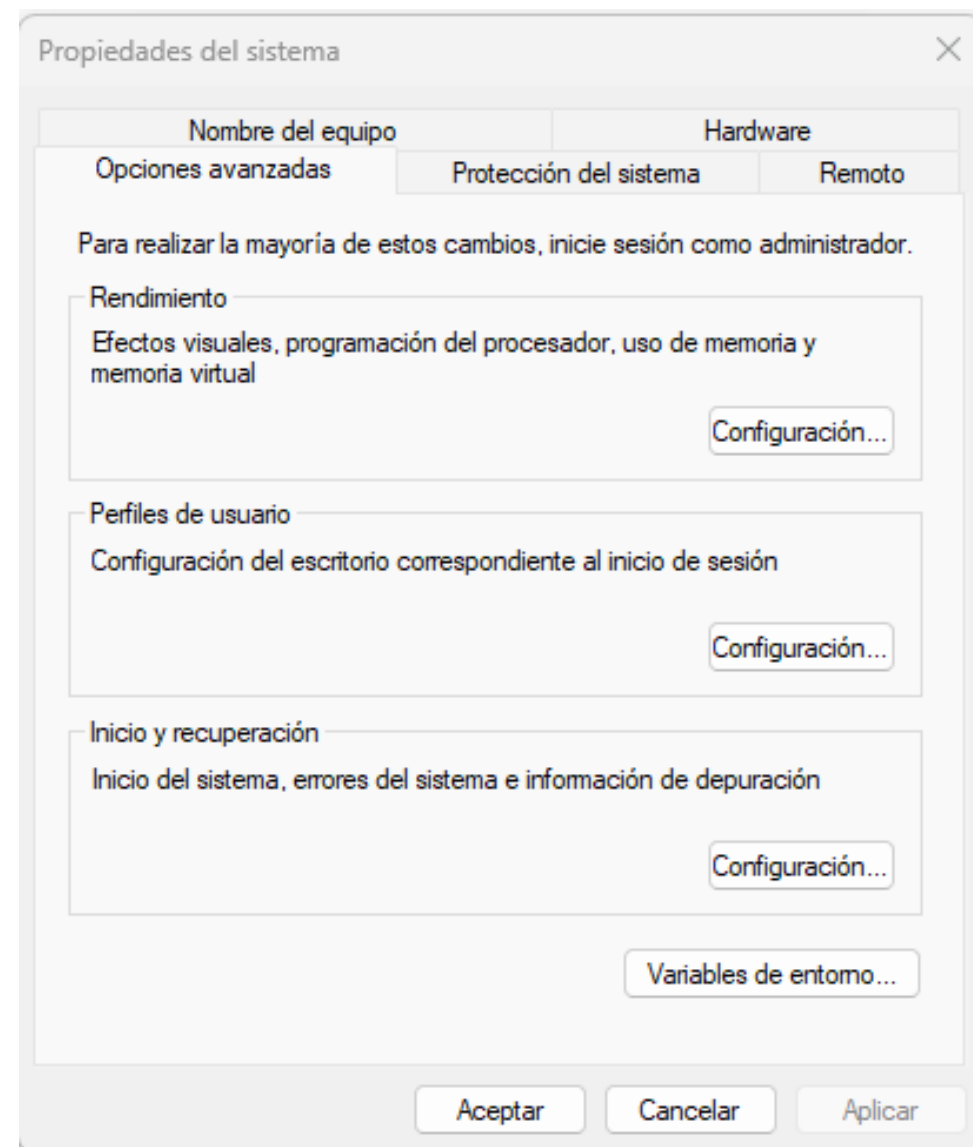
1. Choose a Spark release:
2. Choose a package type:
3. Download Spark: [spark-3.4.0-bin-hadoop3.tgz](#)
4. Verify this release using the 3.4.0 [signatures](#), [checksums](#) and [project release KEYS](#) by following these [procedures](#).

Note that Spark 3 is pre-built with Scala 2.12 in general and Spark 3.2+ provides additional pre-built distribution with Scala 2.13.

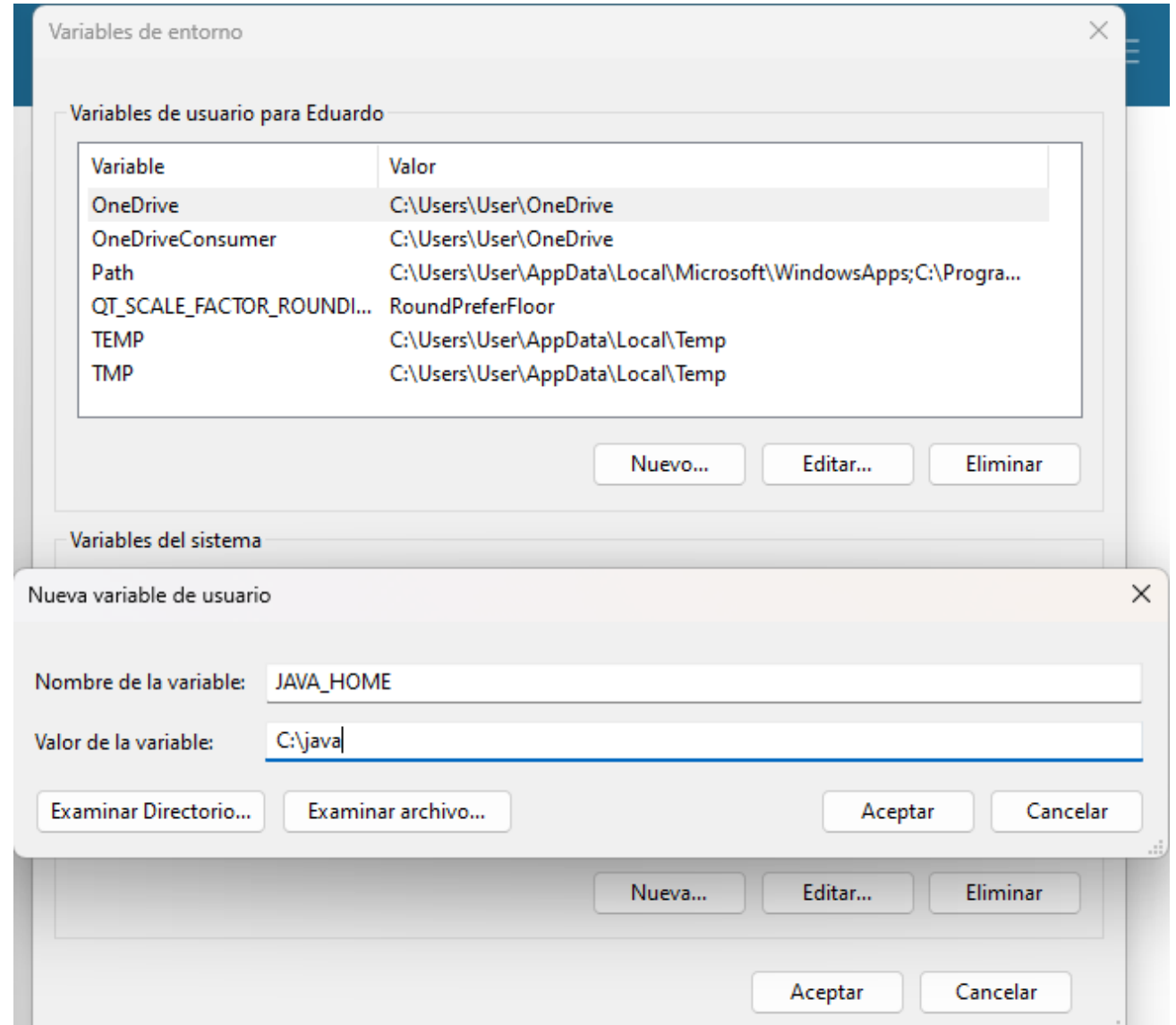
Por último descargamos el repositorio winutils



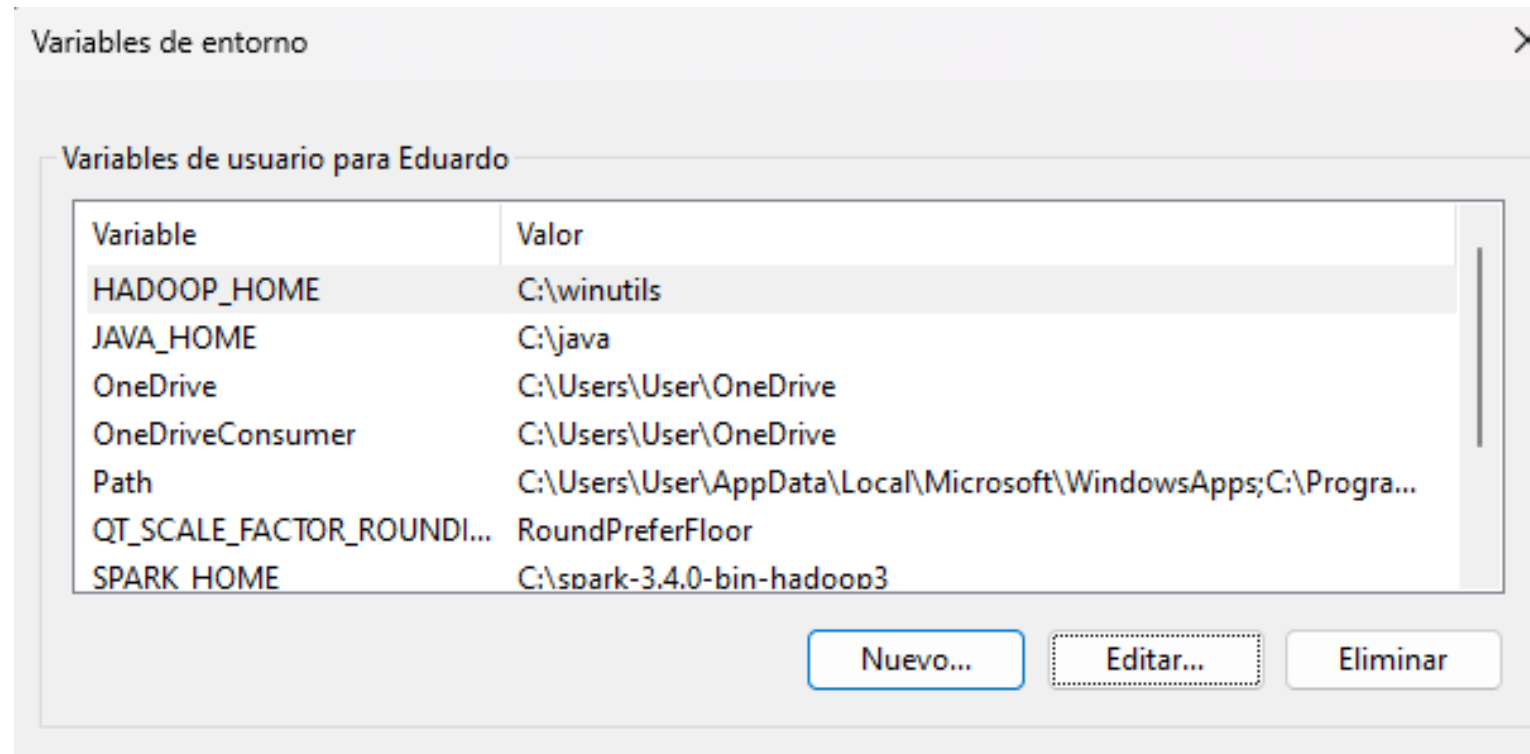
Buscamos la ventana propiedades del sistema



Creamos una variable de entorno con la dirección del java



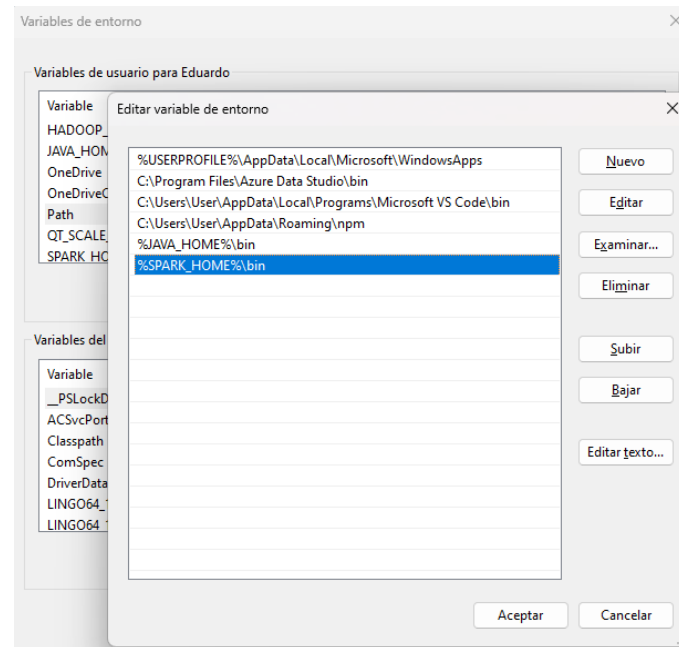
Y lo repetimos hasta que las tres cosas que descargamos tengan su variable de entorno



Editamos el Path y agregamos esas dos variables de entorno:

%JAVA_HOME%\bin

%SPARK_HOME%\bin



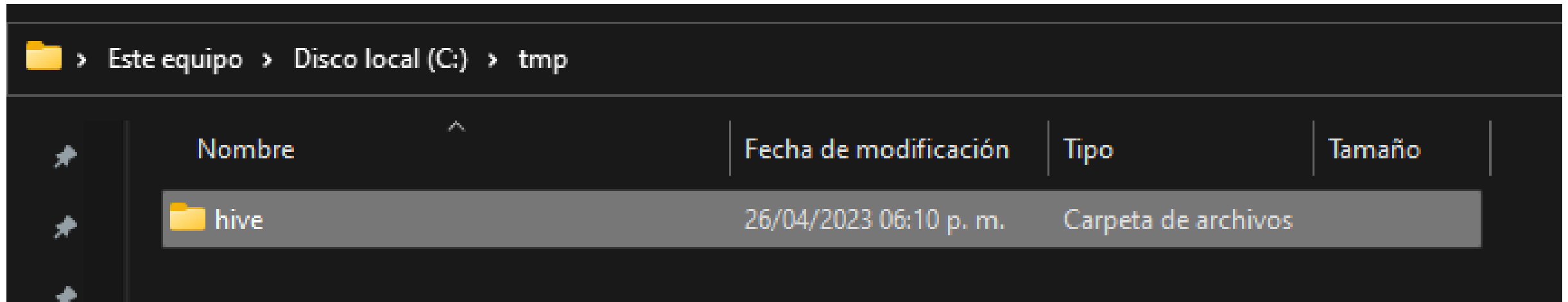
Verificamos que tengamos la versión de java

```
CA: Administrador: Símbolo del sistema
Microsoft Windows [Versión 10.0.22623.1325]
(c) Microsoft Corporation. Todos los derechos reservados.

C:\Windows\System32>java -version
java version "19.0.2" 2023-01-17
Java(TM) SE Runtime Environment (build 19.0.2+7-44)
Java HotSpot(TM) 64-Bit Server VM (build 19.0.2+7-44, mixed mode, sharing)

C:\Windows\System32>D_
```

Creamos dos carpetas en el disco local C
Tmp y dentro hive



```
C:\Windows\System32\cmd.e X + v
Microsoft Windows [Versión 10.0.22623.1325]
(c) Microsoft Corporation. Todos los derechos reservados.

C:\winutils\bin>dir
El volumen de la unidad C no tiene etiqueta.
El número de serie del volumen es: 14CE-07E1

Directorio de C:\winutils\bin

26/04/2023  05:56 p. m.    <DIR>          .
26/04/2023  05:56 p. m.    <DIR>          ..
26/04/2023  05:56 p. m.             112,640 winutils.exe
                1 archivos             112,640 bytes
                2 dirs  196,320,575,488 bytes libres

C:\winutils\bin>winutils chomd 777 C:\tmp\hive|
```

Para terminar de instalar en un CMD nos ubicamos en la dirección del winutils y procedemos a poner el siguiente comando

```
winutils chomd
777 C:\tmp\hive
```

Por último, instalamos las librerías que vamos a utilizar

```
Administrador: Símbolo del sistema - pip install pyspark findspark
Microsoft Windows [Versión 10.0.22623.1325]
(c) Microsoft Corporation. Todos los derechos reservados.

C:\Windows\System32>pip install pyspark findspark
Collecting pyspark
  Downloading pyspark-3.4.0.tar.gz (310.8 MB)
    ----- 310.8/310.8 MB 3.4 MB/s eta 0:00:00
    Preparing metadata (setup.py) ... done
Collecting findspark
  Downloading findspark-2.0.1-py2.py3-none-any.whl (4.4 kB)
Collecting py4j==0.10.9.7
  Downloading py4j-0.10.9.7-py2.py3-none-any.whl (200 kB)
    ----- 200.5/200.5 kB 12.7 MB/s eta 0:00:00
Installing collected packages: py4j, findspark, pyspark
  DEPRECATION: pyspark is being installed using the legacy 'setup.py install' method, because it does not have a 'pyproject.toml' and the 'wheel' package is not installed. pip 23.1 will enforce this behaviour change. A possible replacement is to enable the '--use-pep517' option. Discussion can be found at https://github.com/pypa/pip/issues/8559
  Running setup.py install for pyspark ... \
```

```
Selecionar Administrador: Símbolo del sistema
Microsoft Windows [Versión 10.0.22623.1325]
(c) Microsoft Corporation. Todos los derechos reservados.

C:\Windows\System32>pip install pyspark findspark_
```

```
C:\Windows\System32>  
C:\Windows\System32>  
C:\Windows\System32>jupyter notebook_
```

Abrimos un jupyter
notebook para hacer
una prueba

Dentro del jupyter realizamos una prueba para verificar si está bien instalado creamos una nueva SparkSession por medio de los siguientes comandos.

```
In [5]: import findspark
        findspark.init()

In [6]: from pyspark.sql import SparkSession

In [7]: spark = SparkSession.builder.getOrCreate()

In [8]: spark
Out[8]: SparkSession - in-memory
        SparkContext

        Spark UI
        Version
        v3.4.0
        Master
        local[*]
        AppName
        pyspark-shell

In [ ]:
```

- conda install -c conda-forge findspark
- import findspark
- findspark.init()
- from pyspark.sql import SparkSession
- spark = SparkSession.builder.getOrCreate()
- y para mirar si fue creada escribimos: spark

Ejemplo de cómo cargar y transformar datos utilizando PySpark en un entorno de Spark configurado con findspark .

1. Importar la librería findspark e inicializarla

```
import findspark  
findspark.init()
```

2. Importar la clase SparkSession de PySpark

```
from pyspark.sql import SparkSession
```

```
In [1]: import findspark  
findspark.init()
```

```
In [2]: from pyspark.sql import SparkSession
```

Ejemplo de cómo cargar y transformar datos utilizando PySpark en un entorno de Spark configurado con findspark .

3. Crear una instancia de SparkSession

```
spark = SparkSession.builder \  
    .appName("MiApp") \  
    .getOrCreate()
```

```
In [3]: spark = SparkSession.builder \  
        .appName("MiApp") \  
        .getOrCreate()
```

4. Cargar un archivo CSV como un DataFrame

```
df = spark.read.csv("datos.csv", header=True, \  
inferSchema=True)
```

	A	B	C	
1	nombre	edad	ciudad	
2	Juan	25	Madrid	
3	María	35	Barcelona	
4	Pedro	42	Madrid	
5	Ana	18	Valencia	
6	Luis	30	Barcelona	
7				

```
In [15]: df = spark.read.csv("datos.csv", header=True, inferSchema=True)
```


Ejemplo de cómo cargar y transformar datos utilizando PySpark en un entorno de Spark configurado con findspark .

5. Mostrar el esquema del DataFrame

```
df.printSchema()
```

```
In [16]: df.printSchema()
```

```
root
|-- nombre: string (nullable = true)
|-- edad: integer (nullable = true)
|-- ciudad: string (nullable = true)
```

6. Realizar una operación de transformación en el DataFrame

```
df2 = df.filter(df['edad'] >= 18).groupBy('ciudad').count()
```

```
In [17]: df2 = df.filter(df["edad"] >= 18).groupBy("ciudad").count()
```

Ejemplo de cómo cargar y transformar datos utilizando PySpark en un entorno de Spark configurado con findspark .

7. Mostrar los resultados en la consola
df2.show()

```
In [18]: df2.show()
```

```
+-----+-----+
| ciudad|count|
+-----+-----+
| Madrid|    2|
|Barcelona|    2|
| Valencia|    1|
+-----+-----+
```

Conclusiones

En un mundo cada vez más centrado en la tecnología, las empresas necesitan soluciones innovadoras para manejar grandes cantidades de datos de manera efectiva.