INSTALAR APACHE SPARK

Los datos son el elemento vital de los procesos científicos y empresariales modernos. Los científicos e ingenieros trabajan habitualmente con enormes conjuntos de datos en genética, ingeniería y áreas relacionadas.

Apache Spark es un sistema informático distribuido de código abierto diseñado para el procesamiento de datos a gran escala. Sin embargo, necesita una base confiable para alojar su instalación de Apache Spark. Ahí es donde necesita Ubuntu (o distribuciones similares) que faciliten las operaciones de Spark.

Comencemos con una breve introducción a Apache Spark.

Una breve introducción a Apache Spark

Apache Spark es una plataforma de código abierto popular para configurar y acceder a una interfaz para programar clústeres con paralelismo de datos implícito y tolerancia a fallas. Spark es conocido por su velocidad, facilidad de uso y sofisticadas capacidades de análisis.

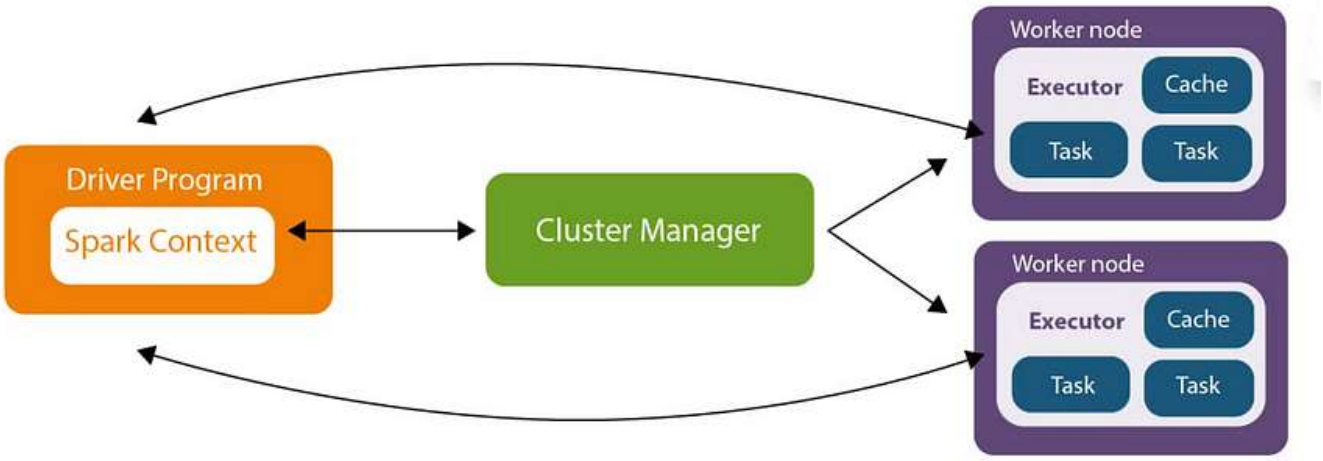
Ubuntu, al ser una distribución de Linux popular, ofrece un entorno estable y eficiente para ejecutar Spark. Al aprovechar Spark en Ubuntu, puede procesar grandes cantidades de datos rápidamente, realizar análisis avanzados y crear modelos de aprendizaje automático potentes.

La arquitectura de Apache Spark

La arquitectura de Apache Spark sigue un modelo maestro/esclavo con dos demonios principales y un administrador de clúster.

El **demonio maestro**, también conocido como **proceso maestro/controlador**, es responsable de coordinar y supervisar la ejecución de tareas dentro del clúster Spark. Mantiene información sobre los recursos disponibles en el clúster y distribuye las cargas de trabajo a **los demonios de trabajo** en consecuencia.

Por otro lado, **los Worker Daemons**, o **procesos esclavos**, son los encargados de ejecutar las tareas que les asigna el **Master Daemon** . Estos nodos trabajadores realizan cálculos y almacenan datos localmente para acelerar el procesamiento de datos.



Además de los **daemons maestro** y trabajador, un clúster Spark también incluye un **administrador de clúster** que administra los recursos en todo el clúster de manera eficiente.

El **administrador de clústeres** es responsable de asignar recursos, como núcleos de CPU y memoria, a las diferentes aplicaciones que se ejecutan en el clúster Spark. Ayuda a optimizar la utilización de los recursos programando tareas en función de los requisitos de carga de trabajo y los recursos disponibles en el clúster.

Gracias a esta arquitectura, Apache Spark puede lograr alta disponibilidad, tolerancia a fallas y escalabilidad al distribuir cargas de trabajo entre múltiples nodos en un entorno informático distribuido.

Instalar Spark en Ubuntu

Ahora que comprende bien cómo funciona Apache Spark, analizaremos en detalle cómo instalarlo y ejecutarlo en Ubuntu.

Paso n.° 1: Verifique los requisitos previos

Antes de instalar Apache Spark debes asegurarte de haber completado los dos pasos siguientes.

Actualizar paquetes del sistema

Antes de instalar Spark, es fundamental asegurarse de que el sistema esté actualizado.

 # sudo apt update

 # sudo apt upgrade

Comprobar la instalación y la versión de Java

Apache Spark requiere Java para ejecutarse. Por lo tanto, comprobar si Java está instalado en el sistema es otro requisito fundamental. Para ello, ejecute el siguiente comando que imprime la información de la versión de Java:

 # java -version

Si Java no está instalado o necesita actualizarlo, instale la última versión de OpenJDK con este comando:

 # sudo apt install openjdk-11-jdk

Paso n.° 2: descargue el paquete Spark

Recomendamos descargar Apache Spark desde el sitio web oficial de Apache Spark. Tenga en cuenta el nombre del archivo de la última versión.

A continuación, ejecute el siguiente comando **wget** para descargar el archivo Spark seleccionado:

 # wget https://archive.apache.org/dist/spark/spark-3.5.1/spark-3.5.1.tgz

Nota : Reemplace 3.5.1 con el último número de versión disponible en el sitio web de Spark.

Paso n.° 3: Instalar Spark en Ubuntu

Ahora que ha descargado el archivo, siga estos pasos para instalar Spark en Ubuntu.

Configurar el directorio de instalación

Elige un directorio donde quieras instalar Spark. Por ejemplo, puedes crear un directorio llamado **spark** en el directorio **de inicio** .

Recomendamos la siguiente serie de comandos que utilizan el comando **mkdir** para crear el directorio de destino, mover el archivo descargado a este directorio y cambiar el directorio de trabajo actual a este directorio recién creado:

 # mkdir ~/spark
 # mv spark-3.5.1.tgz chispa/
 # cd ~/spark

Extraer el archivo tar de Spark

Extraiga el archivo tar descargado en el directorio **spark** con el siguiente comando **tar** :

 # tar -xvzf spark-3.5.1.tgz

Configurar variables de entorno

Antes de poder ejecutar comandos Spark desde cualquier directorio, debe configurar las variables de entorno.

Para ello, comience abriendo el archivo de configuración de Bash (**.bashrc**) en su editor de texto preferido. Abriremos el archivo en Nano:

 # nano ~/.bashrc

Ahora, agregue las siguientes líneas al final del archivo:

export SPARK_HOME=~/spark/spark-3.5.1

export PATH=$SPARK_HOME/bin:$PATH

export PATH=$SPARK_HOME/sbin:$PATH

Guarde el archivo y ejecute este comando para recargar las variables de entorno:

 # source ~/.bashrc

Paso n.° 4: Verificar la instalación

Siempre es una buena idea verificar una instalación de Spark ejecutando el siguiente comando:

 # spark-shell

Si Spark está instalado correctamente, deberías ver el mensaje de aviso del shell de Spark indicando que Spark está listo para usarse. Puedes ver que el mensaje de aviso cambió a **scala>** para indicar que el shell de Spark está activo.

Si alguna vez te encuentras con el siguiente error,

No se pudo encontrar el JAR del ensamblado de Spark. Debe compilar Spark antes de ejecutar este programa

Intente ejecutar este comando para compilar Spark:

 # ./build/mvn -DskipTests clean package

En este punto, Spark está listo para su lanzamiento.

Iniciar Apache Spark para su uso

Ahora que ha instalado Apache Spark, es momento de iniciar los distintos componentes de la arquitectura Spark.

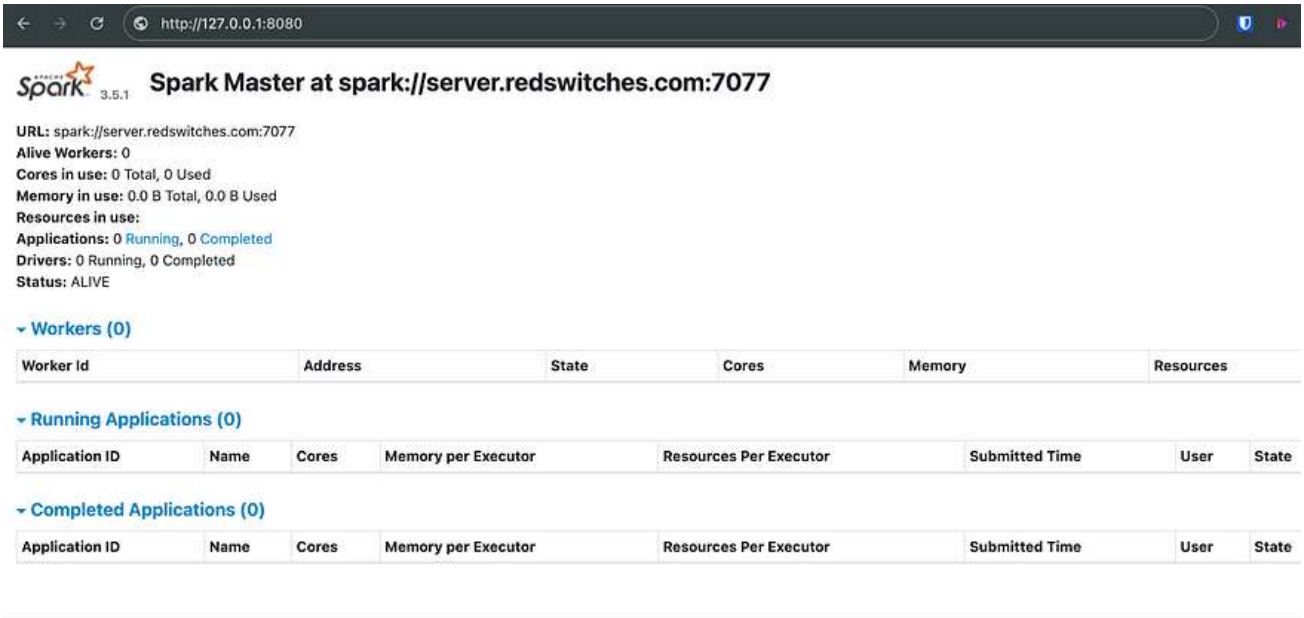
Iniciar el servidor maestro independiente Spark

Ejecute el siguiente comando para iniciar el servidor maestro independiente:

$start-master.sh

```
root@server:~# start-master.sh
starting org.apache.spark.deploy.master.Master, logging to /root/spark/spark-3.5.1/logs/spark-root-org.apache.spark.deploy.master.Master-1-server.out
root@server:~#
```

A continuación, acceda a la interfaz de usuario de Spark Web. Para ello, abra una pestaña del navegador web y navegue hasta la dirección IP del servidor en el puerto 8080.



Iniciar el servidor esclavo Spark (iniciar un proceso de trabajo)

Junto con el Master, también debe iniciar **Worker(s)** para configurar e iniciar las operaciones.

$start-worker.sh

Interfaz de usuario web de Spark

Explora la interfaz de usuario de Spark para obtener información sobre los nodos de trabajo, las aplicaciones en ejecución y los recursos del clúster. Recomendamos iniciar la interfaz pública en http://server public IP:8080.

Conclusión

Instalar Apache Spark en Ubuntu es un proceso sencillo que implica actualizar el sistema, asegurarse de que Java esté instalado, descargar el archivo tar de Spark, extraerlo y configurar las variables de entorno. Una vez configurado, puede aprovechar las potentes capacidades de procesamiento de datos de Spark para manejar grandes conjuntos de datos de manera eficiente.

Preguntas frecuentes

P. ¿Qué es Apache Spark?

R: Apache Spark es un sistema informático distribuido de código abierto que se utiliza para el procesamiento y análisis de grandes cantidades de datos.

P. ¿Por qué necesito Java para ejecutar Spark?

Spark está construido sobre Java, por lo que requiere un entorno de ejecución Java para ejecutar sus programas.

P. ¿Puedo instalar Spark en otros sistemas operativos?

Sí, Apache Spark se puede instalar en varios sistemas operativos, incluidos Windows y macOS.

P. ¿Cómo puedo comprobar si Spark está instalado correctamente?

Puede verificar la instalación ejecutando el comando spark-shell en su terminal. Si Spark se inicia, la instalación se realizó correctamente.

P. ¿Qué pasa si encuentro problemas durante la instalación?

Asegúrese de haber seguido todos los pasos correctamente. Compruebe si hay mensajes de error y consulte la documentación de Apache Spark o los foros de la comunidad para obtener ayuda.

P. ¿Qué es Spark UI?

Spark UI es una interfaz gráfica de usuario para aplicaciones Spark. Proporciona una descripción detallada de los distintos componentes de su aplicación Spark, lo que le ayuda a supervisar y depurar sus trabajos.

P. ¿Qué información puedo encontrar en la interfaz de usuario de Spark?

La interfaz de usuario de Spark proporciona información sobre:

• Etapas: Detalles sobre las diferentes etapas de su trabajo, incluidas las tareas dentro de cada etapa y su estado de finalización.

• Almacenamiento: información sobre RDD (conjuntos de datos distribuidos resilientes) almacenados en caché, incluidos sus tamaños y ubicaciones en la memoria.

• Variables de entorno: muestra la configuración y las variables de entorno utilizadas por su aplicación Spark.

• Ejecutores: detalles sobre los ejecutores, incluido su uso de memoria y CPU, tareas en ejecución y registros.