**PREWORK**

Sesión 03

**YARN y Sistema de Ficheros HDFS en Hadoop**

Instala en tu computadora Apache Hadoop

**🎯 Objetivos**

* Descargar y usar Hadoop en tu distribución Linux
* Correr en la línea de comandos de Linux la descompresión de Hadoop para poder empezar a utilizarlo

### 👨‍💻 **Requisitos**

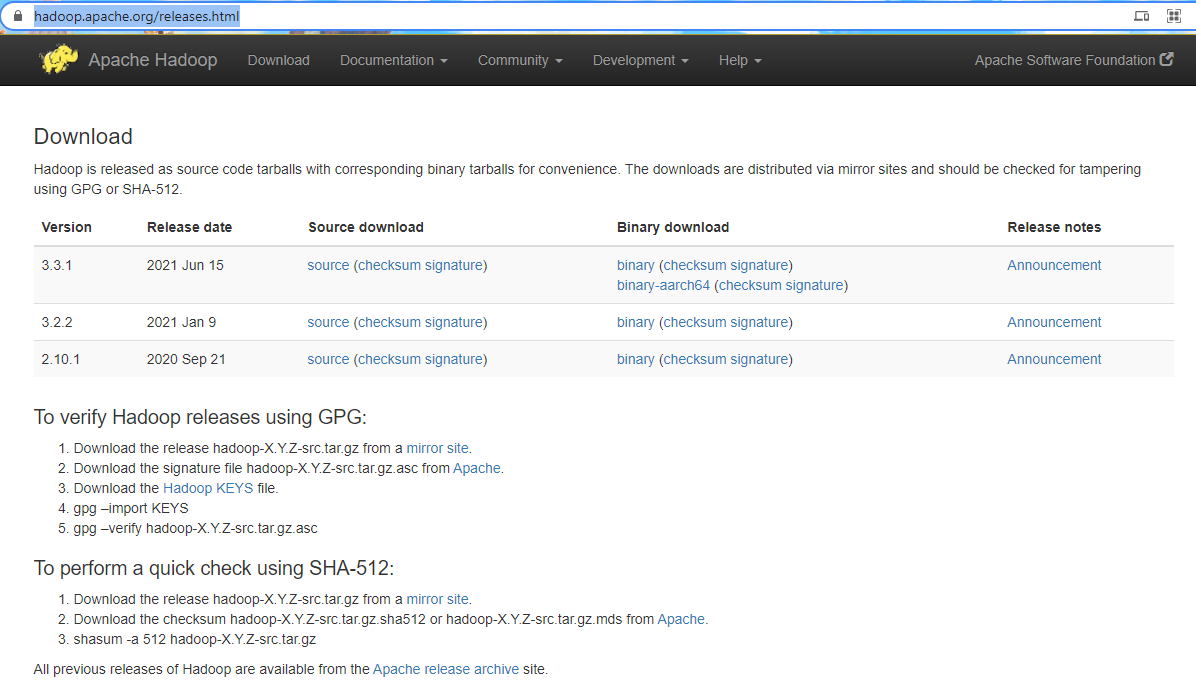
* Contar con una distribución Linux. También es posible poder correr una máquina virtual dentro de tu Windows.

### 

### 🚀 **Desarrollo**

### **Parte 1: Descarga Hadoop**

Ya hemos implementado Hadoop en Windows, pero puedes descargarlo de nuevo desde la siguiente [liga](https://hadoop.apache.org/releases.html).



Descomprima la distribución de Hadoop usando el siguiente comando. Tendrá que cambiar el x.x en el nombre del archivo con la versión real que ha descargado. Si está utilizando Windows, debería utilizar su programa de descompresión favorito, como WinZip o WinRAR, para extraer Hadoop. A partir de este punto, llamaremos al directorio Hadoop descomprimido HADOOP\_HOME.

**>tar -zxvf hadoop-1.x.x.tar.gz**

Puede usar el modo local de Hadoop después de descomprimir la distribución. Tu instalación está lista. Ahora, puede ejecutar trabajos de Hadoop a través del comando bin / hadoop, pero ello será explicado en el Work.

¿Y cómo funciona? El modo local de Hadoop no inicia ningún servidor, pero hace todo el trabajo dentro de la misma JVM (Java Virtual Machine, un entorno simulado y ejecutado en el lenguaje Java). Cuando envía un trabajo a Hadoop en el modo local, ese proceso inicia una JVM para ejecutar el trabajo y esa JVM realiza el trabajo. El resultado y el comportamiento del trabajo son los mismos que los de un trabajo Hadoop distribuido, excepto por el hecho de que el trabajo solo puede usar el nodo actual para ejecutar tareas. En el work, descubriremos cómo ejecutar un programa MapReduce utilizando la distribución Hadoop descomprimida.

**¿Qué es HDFS y cómo funciona?**



El sistema de archivos distribuido de Hadoop (en inglés, Hadoop Distributed File System o HDFS) es el sistema de almacenamiento de datos principal utilizado por las aplicaciones de Hadoop. HDFS emplea una arquitectura NameNode y DataNode para implementar un sistema de archivos distribuido que proporciona acceso de alto rendimiento a los datos en clústeres Hadoop altamente escalables.

Hadoop en sí es un framework basado en Java de procesamiento distribuido de código abierto que administra el procesamiento y almacenamiento de datos para aplicaciones de datos de enorme dimensión

HDFS permite la transferencia rápida de datos entre nodos. Al principio, estaba estrechamente acoplado con MapReduce, un marco para el procesamiento de datos que filtra y divide el trabajo entre los nodos de un clúster, y organiza y condensa los resultados en una respuesta coherente a una consulta. De manera similar, cuando HDFS toma datos, divide la información en bloques separados y los distribuye a diferentes nodos en un clúster.

Con HDFS, los datos se escriben en el servidor una vez y se leen y reutilizan varias veces después de eso. HDFS tiene un NameNode principal, que realiza un seguimiento de dónde se guardan los datos de archivo en el clúster.

HDFS también tiene varios DataNodes en un clúster de hardware básico, generalmente uno por nodo en un clúster. Los DataNodes generalmente se organizan dentro del mismo rack en el centro de datos. Los datos se dividen en bloques separados y se distribuyen entre los distintos DataNodes para su almacenamiento. Los bloques también se replican en todos los nodos, lo que permite un procesamiento paralelo altamente eficiente.

El NameNode sabe qué DataNode contiene qué bloques y dónde residen los DataNodes dentro del clúster de la máquina. NameNode también administra el acceso a los archivos, incluidas las lecturas, escrituras, creaciones, eliminaciones y la replicación del bloque de datos en los DataNodes. Este opera junto con DataNodes y como resultado, el clúster puede adaptarse dinámicamente a la demanda de capacidad del servidor en tiempo real agregando o restando nodos según sea necesario.

Los DataNodes están en comunicación constante con NameNode para determinar si los DataNodes necesitan completar tareas específicas: NameNode siempre conoce el estado de cada DataNode, donde si cada NameNode se da cuenta de que un DataNode no está funcionando correctamente, puede reasignar inmediatamente la tarea de ese DataNode a un nodo diferente que contenga el mismo bloque de datos. Los DataNodes también se comunican entre sí, lo que les permite cooperar durante las operaciones normales de archivos.

Además, el HDFS está diseñado para ser altamente tolerante a fallas. El sistema de archivos replica, o copia, cada dato varias veces y distribuye las copias a los nodos individuales, colocando al menos una copia en un bastidor de servidor diferente al de las otras copias.

❓ **Actividad.** Cuestionario Hadoop en Linux

1. ¿Al haber instalado Hadoop en nuestro Linux, ya podemos hacer cómputo paralelo?

1. **No, aún no configuramos nodos esclavos**
2. Si, ya tenemos todo listo
3. Si, solo que necesitamos calibrar los sub-nodos de nuestra pc

2. ¿Qué es una JVM?

1. **Java Virtual Machine**
2. No tiene significado
3. Una nomenclatura de Javascript

3. ¿Al ejecutar un nodo unidistribuido, podemos emular lo que genera una serie de nodos efectivamente distribuidos?

1. Es un switch que altera el flujo de datos
2. Es un conjunto a agrupación de clústeres
3. **Si, pero el poder de procesamiento no se ve reflejado ya que es una mera emulación.**