**PREWORK**

Sesión 02

**Introducción a Apache Hadoop**

Instala en tu computadora Windows Apache Hadoop

**🎯 Objetivos**

* Descargar Hadoop en tu computadora para que puedas tomar con satisfacción la sesión 02 (empezaremos a instalar y programar en hadoop). La instalación de Hadoop en Windows (el sistema operativo más ampliamente usado en el mundo y las empresas, aunque Linux sea más eficiente) es algo complicada, y es por ello que lo verás a detalle en el work junto con la ayuda del Experto.
* Descargar Java en tu computadora (Hadoop funciona gracias a este lenguaje de programación).
* Comprender y definir conceptos básicos para el cómputo paralelo en Hadoop

### 👨‍💻 **Requisitos**

* Instalar los prerrequisitos obligatorios para que Hadoop funcione adecuadamente: debes instalar la versión 8 de Java.
* Contar con una computadora con al menos 8 GB de memoria RAM, 60 GB libres en HDD o SSD, Windows 7 o superior

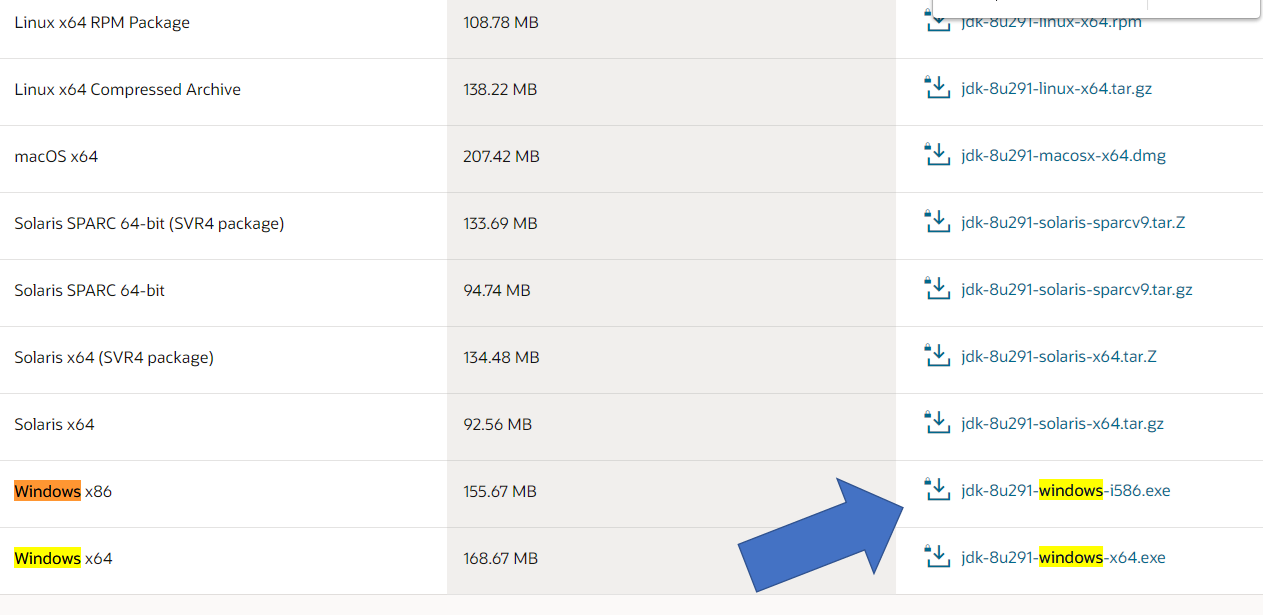
### 👨‍💻 **Notas**

* Si bien puedes desarrollar el módulo con Windows, la mayoría de cómputo paralelo en la industria se genera con Linux debido a la eficiencia y costos nulos de licencia. Es altamente recomendable que instales una distribución Linux para continuar con el módulo
* Si cuentas con una distribución Linux, INSTALA HADOOP DE UNA VEZ, esto debido a que es mucho más sencillo en dicho sistema (para el work hemos asignado una parte para la mayoría de los usuarios, así es, aquellos que usan Windows) te dejamos este video tutorial desarrollado por uno de los Expertos para que puedas hacerlo satisfactoriamente. En caso de que tengas MacOS, mira este otro videotutorial (también puedes pedir ayuda al Experto o un TAA durante la sesión)

### 🚀 **Desarrollo**

### **Parte 1: Descarga de Java y Hadoop**

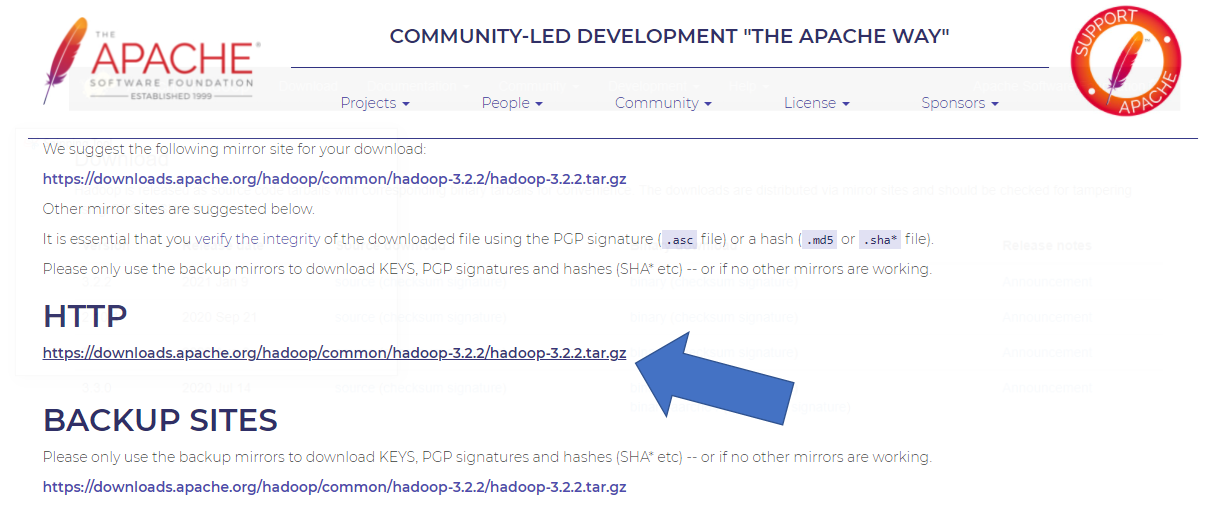
* Descarga Java 8 desde la siguiente [liga](https://www.oracle.com/java/technologies/javase/javase-jdk8-downloads.html#license-lightbox). Verás la siguiente pantalla (selecciona la versión de 32 o 64 bits, dependiendo de tu caso):



* Ahora bien, debes descargar Hadoop desde la siguiente [liga](https://hadoop.apache.org/releases.html). Verás la siguiente pantalla, donde deberás dar click en “binary” de la versión más reciente (en el momento en el que fue desarrollado el presente material, la más nueva era la 3.2.2 liberada el 9 de enero del 2021):



* Una vez hayas dado click en binary, te encontrarás en la siguiente página. Da click en la opción debajo de “HTTP”



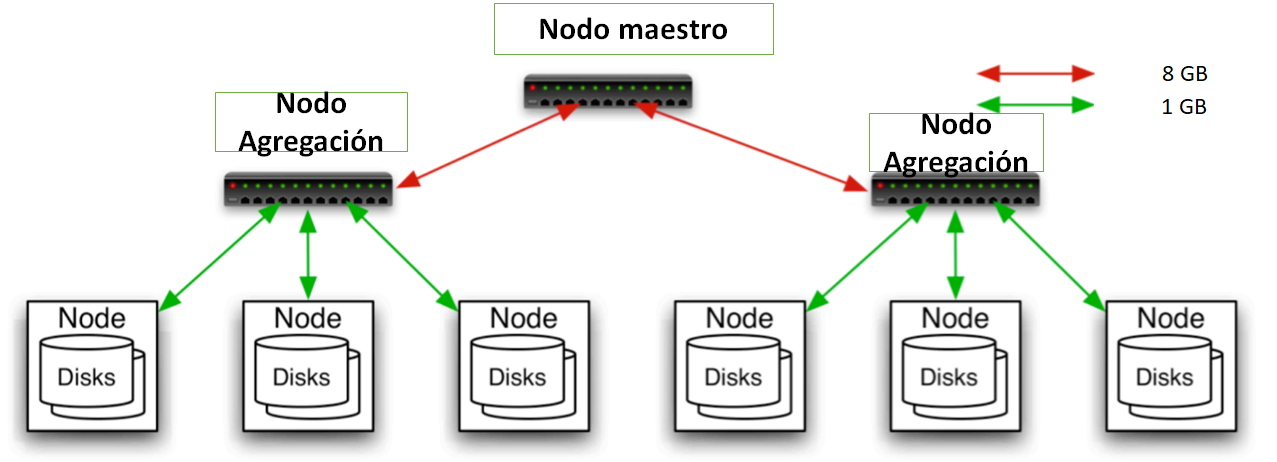
### **Parte 2: Algunos conceptos básicos de Hadoop**



Apache Hadoop es un ambiente de programación desarrollado con el objetivo de procesar y analizar grandes cantidades de datos en conjunto con tecnologías de cómputo paralelo. Recordemos algunos términos importantes:

* Clúster: Un clúster de Hadoop es un conjunto de computadoras, conocidas como nodos, que están conectadas en red para realizar este tipo de cálculos paralelos en grandes conjuntos de datos. A diferencia de otros clústeres de computadoras, los clústeres de Hadoop están diseñados específicamente para almacenar y analizar cantidades masivas de datos estructurados y no estructurados en un entorno informático distribuido. Además, lo que distingue al ecosistema Hadoop de otros en el mercado es justamente su estructura y arquitectura únicas.
* Nodo: Los clústeres de Hadoop consisten en una red de nodos maestros y esclavos conectados que utilizan hardware básico de alta disponibilidad y bajo costo. La capacidad de escalar linealmente y agregar o restar nodos rápidamente según las demandas de volumen que tú definas, es lo que hace sumamente competitivo y adecuado a Hadoop para tareas de análisis de big data con conjuntos de datos de tamaño que pueden variar dinámicamente.
* Rack: Un Rack es una colección de nodos (generalmente 10 o más) que se almacenan juntos y están conectados a un mismo Switch. Cuando un usuario solicita una lectura o escritura en un clúster grande de Hadoop, pueden existir problemas en el movimiento de los datos. Para mejorar el tráfico, se elige un nodo que esté más próximo a los datos solicitados (a esto se llama ‘rack awarness’).

En este orden de ideas, un cluster puede estar constituido por nodos simples que pertenecen a un nodo maestro, o también pueden estar agregados como racks. Aquí un ejemplo visual del primer caso, y que también viste a grandes rasgos en la sesión pasada:



A diferencia de otras herramientas, Hadoop permite generar cómputo paralelo a partir de clusters y nodos que incluso pueden ser no necesariamente servidores dedicados y costosos, sino que en su lugar, tus nodos de procesamiento pueden ser computadores personales (eso sí, con una capacidad de procesamiento y almacenamiento relativamente altas para poder desarrollar las tareas asignadas adecuadamente).

❓ **Actividad.** Cuestionario sobre Hadoop

1. ¿Qué es un clúster para Hadoop?

1. **Es un conjunto de nodos de procesamiento**
2. Es una forma de agrupar puntos geométricos
3. Es una metodología de programación

2. ¿Qué es un nodo para Hadoop?

1. **Se considera una célula de procesamiento. Es la unidad básica dentro de un cluster**
2. Es parte de una estructura de datos, en específico un Árbol de Decisión
3. Es la estructura de cómputo paralelo más general que existe para Hadoop

3. ¿Qué es un rack para Hadoop?

1. Es un switch que altera el flujo de datos
2. Es un conjunto a agrupación de clústeres
3. **Es un conjunto o agrupación de nodos, generalmente más de 10**

4. ¿Es posible generar cómputo paralelo con computadoras comerciales interconectadas?

1. No, solo servidores diseñados para el procesamiento de Grandes Datos
2. No, debido a que no tenemos asignaciones simétricas de RAM y de los procesadores.
3. **Si, Hadoop fue desarrollado pensado para procesar Big Data incluso tomando como nodos desde computadores personales, hasta servidores de alta calidad**