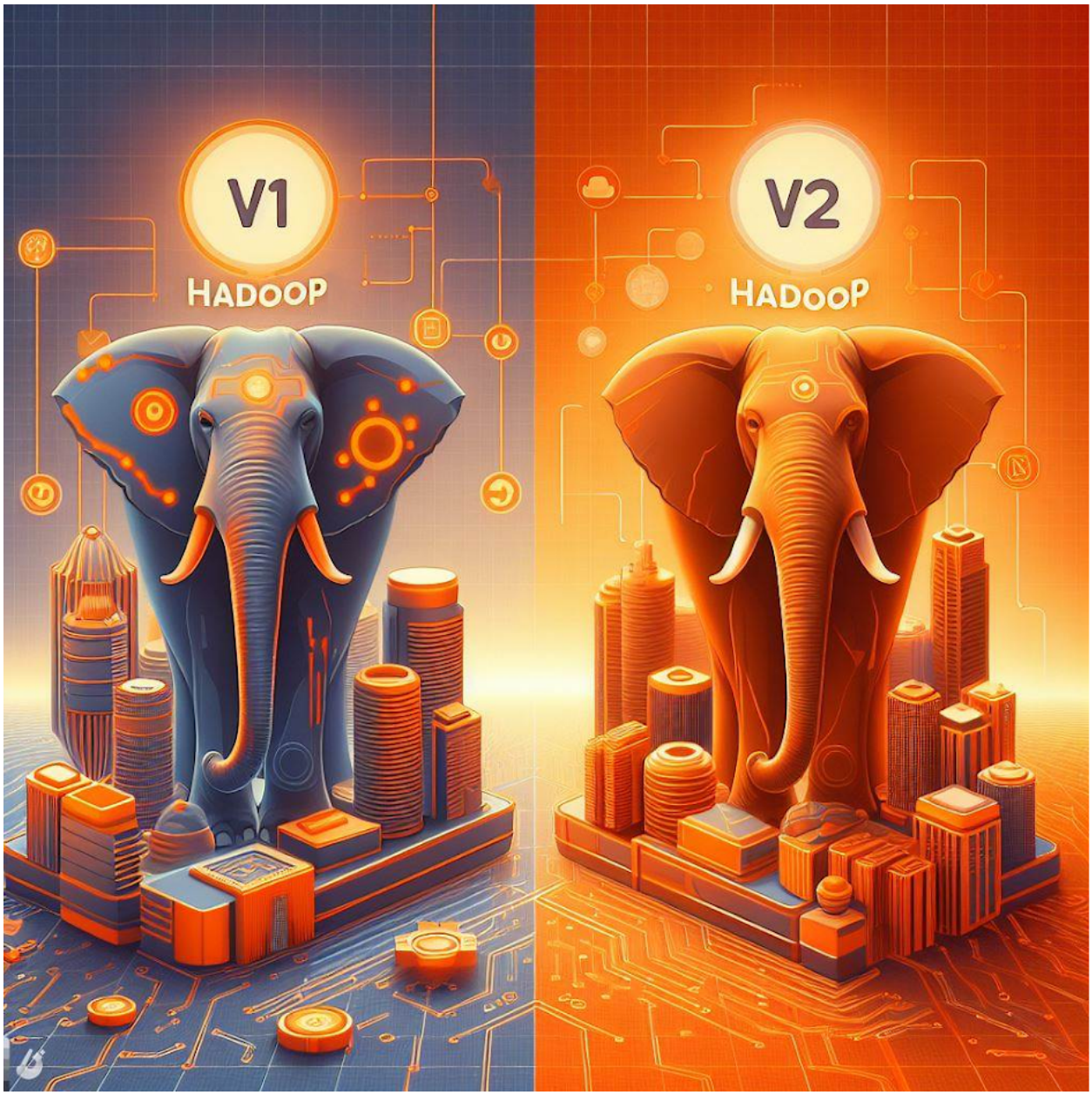
Néstor Batista Díaz

Hadoop v1
vs
Hadoop v2



DESARROLLO

1. **Investiga exhaustivamente la gestión de recursos en Hadoop v1. Examina cómo el JobTracker despliega recursos para los trabajos de MapReduce y señala las limitaciones asociadas con este enfoque.**

   TaskTracker es un proceso se ejecuta en los DataNodes. Los TaskTrackers están en constante comunicación con el **JobTracker**, señalando el progreso de la tarea en ejecución. Las tareas del Mapper y Reducer se ejecutan en los DataNodes administrados por los TaskTrackers. Los TaskTrackers serán asignados por el JobTracker para ejecutar tareas del Mapper y Reducer.

   Limitaciones:

   - **Punto único de fallo**: Tanto el JobTracker como el NameNode son puntos únicos de fallo en Hadoop v1. Si el JobTracker falla, todos los trabajos en ejecución se detienen.
   - **Escalabilidad**: El JobTracker puede convertirse en un cuello de botella de recursos cuando el clúster se amplía lo suficiente (generalmente alrededor de 4,000 clústers).
   - **Asignación de trabajos simple**: La asignación de trabajos por parte de los JobTrackers es muy sencilla. Cada TaskTracker tiene un número de plazas disponibles (ranuras o slots). El JobTracker asigna las tareas Map o Reduce al TaskTracker más cercano a los datos.

2. **Investiga la gestión de recursos en Hadoop v2, centrándote en el framework YARN. Explora cómo YARN asigna recursos a diversas aplicaciones, como MapReduce, Spark y Hive, y destaca sus características clave.**

   YARN (Yet Another Resource Negotiator) es un componente fundamental de Hadoop v2. Se encarga de la gestión de recursos y la planificación y monitorización de trabajos en el clúster de Hadoop. Aquí te dejo algunos detalles clave sobre cómo YARN asigna recursos a diversas aplicaciones y sus características principales:

- **Componentes principales de YARN**:
  - **Resource Manager (RM)**: Es el encargado de gestionar la distribución de los recursos del clúster de YARN. Tiene dos componentes: el Scheduler, que distribuye los recursos del clúster, y el Applications Manager, que acepta las peticiones de trabajos, negocia el contenedor en el que ejecutar la aplicación y proporciona reinicios de los trabajos en caso de errores.
  - **Node Manager (NM)**: Gestiona los trabajos con las instrucciones del Resource Manager y proporciona los recursos computacionales necesarios para las aplicaciones en forma de contenedores.
  - **Application Master (AM)**: Es el responsable de negociar los recursos apropiados con el Resource Manager y monitorizar su estado y su progreso.
- **Asignación de recursos a aplicaciones**:
  - **MapReduce**: Las aplicaciones de MapReduce son compatibles con YARN. YARN permite a Hadoop soportar varios motores de ejecución, incluyendo MapReduce.
  - **Spark**: Spark puede ejecutarse en YARN. YARN proporciona una mejor gestión de recursos en Hadoop, lo que resulta en una mayor eficiencia del clúster y un mejor rendimiento de las aplicaciones.
  - **Hive**: Hive también puede ejecutarse en YARN. Sin embargo, para ver mejoras significativas en el rendimiento, es posible que se necesite una versión de Hive que sea compatible con Hadoop v25.
- **Características clave de YARN**:
  - **Escalabilidad**: YARN permite escalar el procesamiento de datos horizontalmente, permitiendo agregar más nodos para procesar mayores cantidades de datos.
  - **Eficiencia**: Reduce los tiempos de procesamiento al planificar las tareas de forma automática.
  - **Flexibilidad**: Permite desarrollar diferentes aplicaciones a partir de distintos lenguajes de programación.

        - **Administración centralizada**: Permite gestionar y administrar recursos, tareas y aplicaciones.
        - **Procesamiento distribuido**: Realiza diferentes tareas en distintos nodos de forma simultánea.

3. **Realiza una comparación detallada entre los dos sistemas de gestión de recursos. Examina aspectos como la escalabilidad, la resiliencia y la flexibilidad, resaltando las diferencias significativas que definen la eficacia de cada versión.**

- **Modelo de procesamiento**:
    - **Hadoop v1** se limita a un modelo de procesamiento basado en MapReduce.
    - **Hadoop v2** ha evolucionado desde su modelo de procesamiento basado en MapReduce, ofreciendo mejoras significativas para el procesamiento de Big Data. Permite modelos de procesamiento efectivos que se prestan a muchos usos de Big Data, incluyendo consultas SQL interactivas sobre big data, análisis de gráficos a escala de Big Data y habilidades de aprendizaje automático escalables.
- **Gestión de recursos**:
    - En **Hadoop v1**, MapReduce se encarga tanto de la gestión de recursos como del procesamiento de datos.
    - En **Hadoop v2**, YARN (Yet Another Resource Negotiator) se encarga de la gestión de recursos, mientras que MapReduce se encarga de la gestión de aplicaciones. YARN ha introducido dos nuevos demonios con Hadoop 2: Resource Manager y Node Manager.
- **Escalabilidad y resiliencia**:
    - **Hadoop v2** introduce la federación HDFS y el modo de alta disponibilidad, lo que permite una mejor gestión de recursos y escalabilidad.
    - El modo de alta disponibilidad de Hadoop 2 mejora la fiabilidad de Namenode.
- **Flexibilidad**:
    - **Hadoop v1** trata todos los dispositivos de almacenamiento como un solo grupo uniforme en un DataNode.
    - **Hadoop v2** admite todo tipo de almacenamiento heterogéneo.

- **Tamaño de bloque**:
    - En **Hadoop v1**, el tamaño de bloque predeterminado es de 64MB.
    - En **Hadoop v2**, el tamaño de bloque predeterminado es de 128MB.

4. **Destaca específicamente las ventajas de Hadoop v2 en la gestión de recursos. Enfócate en aspectos como la asignación dinámica de recursos, el soporte para múltiples tipos de aplicaciones y la mejor escalabilidad que ofrece YARN en comparación con el JobTracker.**

- **Asignación dinámica de recursos**: YARN permite una asignación dinámica de recursos, lo que mejora la utilización y la eficiencia del clúster. A diferencia de Hadoop v1, donde los recursos se asignan estáticamente, YARN puede asignar recursos de manera flexible en función de las demandas de las aplicaciones.
- **Soporte para múltiples tipos de aplicaciones**: YARN no está limitado a MapReduce y puede soportar múltiples paradigmas de procesamiento de datos. Esto significa que Hadoop v2 puede manejar una variedad más amplia de cargas de trabajo, incluyendo procesamiento por lotes, procesamiento interactivo, transmisión en tiempo real y gráficos.
- **Mejor escalabilidad**: YARN mejora la escalabilidad del clúster al separar la gestión de recursos y la programación/monitorización de trabajos. Esto permite a Hadoop v2 escalar a un número mucho mayor de nodos y manejar una mayor cantidad de datos en comparación con Hadoop v11.
- **Tolerancia a fallos**: Hadoop v2 mejora la tolerancia a fallos al permitir que las tareas de un nodo que falla sean transferidas de manera transparente a otro nodo. Los datos son replicados automáticamente en múltiples máquinas, lo que aumenta la resiliencia del sistema.
- **Flexibilidad**: Hadoop v2 ofrece una mayor flexibilidad al permitir que los datos no sean procesados previamente a su almacenamiento. Se pueden almacenar tantos datos, estructurados o no estructurados, como se necesite y decidir posteriormente cómo se van a utilizar.

PREGUNTAS

1. **¿Cuáles son las disparidades más destacadas en la gestión de recursos entre Hadoop v1 y Hadoop v2?**
    - ➢ Hadoop v1 sólo admite un espacio de nombres para gestionar el sistema de archivos HDFS, mientras que Hadoop v2 admite varios espacios de nombres.
    - ➢ Hadoop v1 soporta uno y sólo un modelo de programación: MapReduce. Hadoop v2 admite varios modelos de programación con el componente YARN, como MapReduce, Interative, Streaming, Graph, Spark, Storm, etc.
    - ➢ Hadoop v1 tiene muchas limitaciones de escalabilidad. Hadoop v2 ha superado esa limitación con una nueva arquitectura.
    - ➢ Hadoop v2 tiene soporte multi-tenencia, pero Hadoop v1 no.
    - ➢ El HDFS de Hadoop v1 utiliza un mecanismo de ranuras de tamaño fijo para el almacenamiento, mientras que Hadoop v2 utiliza contenedores de tamaño variable.
    - ➢ Hadoop v1 admite un máximo de 4.000 nodos por clúster, mientras que Hadoop v2 admite más de 10.000 nodos por clúster.
    - ➢ Hadoop v1 trabaja sobre conceptos de slots - los slots pueden ejecutar una tarea Map o una tarea Reduce solamente mientras que Hadoop v2 trabaja sobre conceptos de contenedores. Utilizando contenedores se pueden ejecutar tareas genéricas.
    - ➢ Hadoop v1 no es compatible con Microsoft Windows, mientras que Hadoop v2 ha añadido compatibilidad con Microsoft Windows.

2. **¿Cuáles son las ventajas clave de Hadoop v2 en términos de gestión de recursos, y cómo estas contribuyen a una mejor eficiencia del sistema?**

    Las ventajas clave de Hadoop v2 en términos de gestión de recursos se centran principalmente en la introducción de YARN. Aquí hay algunas de las ventajas más importantes:

- **Mayor flexibilidad en el procesamiento de datos**: YARN permite la ejecución de aplicaciones más allá de MapReduce, lo que significa que puedes ejecutar diferentes tipos de cargas de trabajo, como Spark, Tez u otros, en el mismo clúster. Esto proporciona una mayor flexibilidad y eficiencia al adaptar los recursos a las necesidades específicas de cada aplicación.
- **Mejora en la utilización de recursos**: Con YARN, la gestión de recursos es más dinámica y eficiente. Puedes asignar y liberar recursos de manera más granular, lo que mejora la utilización general del clúster. Esto es crucial para maximizar el rendimiento y reducir los tiempos de espera de las aplicaciones.
- **Escalabilidad mejorada**: YARN escala mejor que el modelo de Hadoop v1, ya que distribuye la responsabilidad de la administración de recursos entre el ResourceManager y los NodeManagers. Esto facilita la administración de clústeres más grandes y proporciona una mayor capacidad de procesamiento sin sacrificar la estabilidad.
- **Tolerancia a fallos mejorada**: La arquitectura de YARN mejora la tolerancia a fallos al descentralizar la gestión de recursos. Si un ResourceManager falla, otro puede tomar su lugar. Además, los nodos individuales (NodeManagers) son más robustos, lo que reduce la probabilidad de que un fallo afecte a todo el clúster.
- **Administración de recursos basada en políticas**: YARN permite la implementación de políticas de administración de recursos más avanzadas. Puedes asignar prioridades a las aplicaciones, garantizar cuotas de recursos y ajustar dinámicamente la asignación de recursos según las necesidades cambiantes.

3. **¿En qué escenarios sería preferible utilizar Hadoop v1 en lugar de Hadoop v2, considerando las limitaciones específicas de cada versión?**

   Hay 4 casos en los que es mejor utilizar Hadoop v1 en lugar de Hadoop v2:

- **Simplicidad y compatibilidad**: Si estás trabajando en un entorno donde la simplicidad es crucial y ya has implementado con éxito Hadoop v1, cambiar a Hadoop v2 puede ser una tarea compleja. En algunos casos, mantener la versión existente puede ser más conveniente, especialmente si no necesitas las características adicionales de YARN.
- **Aplicaciones que solo utilizan MapReduce**: Si tus aplicaciones se basan exclusivamente en el modelo de programación MapReduce y no tienes planes inmediatos de adoptar otras tecnologías de procesamiento de datos compatibles con YARN, podrías optar por seguir utilizando Hadoop v1.
- **Entornos de prueba o desarrollo pequeños**: En situaciones donde tienes entornos más pequeños, como configuraciones de prueba o desarrollo, donde la complejidad adicional de YARN no aportaría beneficios significativos, podrías optar por Hadoop v1 por su simplicidad y menor sobrecarga.
- **Recursos limitados**: En entornos con recursos limitados, donde la sobrecarga adicional de YARN puede afectar negativamente al rendimiento general, Hadoop v1 podría ser una opción más viable, ya que tiene una huella más ligera en términos de recursos.

BIBLIOGRAFÍA

- https://stackoverflow.com/questions/46684091/in-hadoop-what-is-the-difference-and-relationship-between-jobtracker-tasktracker
- https://hdfstutorial.com/blog/hadoop-1-vs-hadoop-2-differences/?utm_content=cmp-true
- https://www.geeksforgeeks.org/difference-between-hadoop-1-and-hadoop-2/
- https://www.tutorialspoint.com/difference-between-hadoop-1-and-hadoop-2
- https://www.quora.com/What-are-the-differences-between-Hadoop-1-x-and-Hadoop-2-x