Tema 4
Actividad 1

Autor
Juan José Méndez Torrero

Cuestiones sobre el Ecosistema Hadoop

1. ¿Qué es el framework Hadoop?

Hadoop es un framework de código abierto usado para almacenar datos y ejecutar aplicaciones en diferentes clusters, los cuales cuentan con un hardware muy básico (procesador, memoria, etc). Además, proporciona un almacenamiento de datos masivo de cualquier tipo, además de proporcionar un gran poder de procesamiento y la capacidad de realizar tareas prácticamente ilimitadas.

2. ¿Qué es la tolerancia a fallos?

La tolerancia a fallos en un framework indica la capacidad del sistema a protegerse contra fallos de hardware. Esto quiere decir que, si un nodo fallara por alguna razón, el flujo de la tarea se mandaría a un nodo diferente para así asegurar que la tarea es completada satisfactoriamente.

3. Nombre los cuatro componentes que forman el framework de Hadoop

Los cuatro componentes que forman Hadoop son:

- Capa de almacenamiento de datos: Ej. HDFS
- Capa de procesamiento de datos: Ej. YARN
- Capa de acceso a datos: Ej. SQOOP
- Capa de gestión de datos: Ej. Flume

4. Si la realización entre nodos en HDFS provoca que la redundancia de datos ocupe más memoria, ¿por qué se implementa?

Aunque esta redundancia de datos ocupe más memoria, este tipo de almacenamiento no requiere de un hardware costoso, con lo que el necesitar de más sistemas para el almacenamiento de los datos no implicaría un alto coste. Además, al realizar la lectura de datos de manera distribuida, esto reduce en bastante cantidad el tiempo necesario para la lectura de los datos, otro punto a favor del uso de HDFS.

5. ¿Qué es un nodo maestro y un nodo esclavo en Hadoop?

El nodo maestro se utiliza para almacenar toda la información relacionada con sus nodos esclavos. Además mantiene el estado de esos nodos esclavos que, a su vez, estos nodos esclavos son los que almacenan la información que será procesada posteriormente.

6. ¿Qué es un NameNode?

NameNode o Nodo de control, es un servidor único con código para administrar el espacio de nombre en el sistema de archivos. Además, almacena los metadatos de los archivos junto con sus directorios. Cabe señalar que es un componente obligatorio dentro de un cluster HDFS.

7. ¿El NameNode también es hardware básico?

Sí, ya que es el encargado de almacenar y administrar el sistema de archivos dentro del cluster creado.

8. ¿Qué es un DataNode?

DataNode, o también servidor de datos, es un componente obligatorio que se encarga de escribir y leer datos, ejecutar comandos o para enviar periódicamente mensajes de estado y procesar esas peticiones.

9. ¿Qué es MapReduce?

Es un paradigma de procesamiento de datos que se caracteriza por su división en dos fases: Map y Reduce. Normalmente, los datos son divididos y procesador por un mapper en paralelo. Posteriormente, el resultado es convertido en la entrada de los reducers (segunda fase).

10. ¿Qué es Jobtracker?

El JobTracker se ejecuta en un nodo separado y no suele estar dentro de un DataNode. Éste, recibe las solicitudes de ejecución de MapReduce desde el cliente, y es el encargado de hablar con el NameNode para determinar la ubicación de los datos. Finalmente, JobTracker encuentra los mejores nodos de TaskTracker para ejecutar las tareas basándose en la localización de los datos y las ranuras disponibles para ejecutar una tarea en un nodo determinado.