Tema 4
Actividad 2

Autor
Juan José Méndez Torrero

Cuestiones sobre el Ecosistema Hadoop

1. ¿Qué es un TaskTracker?

TaskTracker es un nodo dentro de nuestro cluster que irá aceptando las tareas recibidas de un JobTracker. Estas tareas podrían ser Map, Reduce o Shuffle, por ejemplo. Además, cada TaskTracker cuenta con un conjunto de ranuras que indican el número de tareas que pueden ser aceptadas en ese nodo.

2. ¿Por qué se utiliza HDFS para aplicaciones con grandes conjuntos de datos y no para aplicaciones que tienen una gran cantidad de archivos pequeños?

Esto es debido a la forma en la que trabaja HDFS. Este sistema utiliza la transmisión de datos por lotes, con lo que una gran cantidad de archivos pequeños haría que el número de lotes incrementase bastante. Esto no pasa con archivos grandes, al ser por lotes, el número de lotes que habría que crear es mucho inferior si tenemos los datos almacenados en un mismo conjunto que en diferentes conjuntos más pequeños.

3. ¿Qué es una señal Heartbeat en HDFS?

En HDFS un Heartbeat se refiere a la señal que es enviada desde un DataNode a un NameNode para indicar que el primero está activo. En el caso de no haber ninguna señal, será indicador de que algo ha ido mal y que DataNode y NameNode no han podido realizar ningún cálculo.

4. ¿Qué es un NameNode secundario? ¿Es el NameNode secundario un sustituto de NameNode?

Un NameNode secundario es un servidor que se encuentra separado del resto. Este nodo es el único en el cluster capaz de copiar el registro de transacciones de la imagen HDFS y del bloque de archivos a una carpeta temporal, aplicando los cambios acumulados en el registro de transacciones a la imagen HDFS.

El NameNode secundario se usa como sustituto de NameNode en el caso de necesitar una rápida recuperación manual de éste cuando ha fallado.

5. ¿Qué es un rack?

El término *rack* o bastidor, es una colección física de nodos en nuestro cluster. Un gran cluster se compone de muchos *racks,* y con su ayuda, el Namenode elige el DataNode más cercano para lograr el máximo rendimiento mientras se realiza la información de lectura o escritura que reduce el tráfico de la red.

6. ¿Qué es un combinador?

Un combinador o *combiner*, siempre actúa entre Mapper y Reducer. Este combinador es el encargado de minimizar la congestión de la red. Además, ayuda a producir detalles abstractos o en resumen de conjuntos de datos muy grandes.

7. Si el tamaño de un archivo es de 500 MB, el tamaño de bloque es de 128 MB y el factor de replicación es 2, ¿cuál es la cantidad total de bloques que ocupa?

Para calcular el número de bloques total, habría que dividir los 500MB entre distintos bloques que ocupan 128MB. Esto quiere decir:

128 +128 +128 + 116 = 500

Esto quiere decir que tienen que crear un total de 4 bloques en total para almacenar el archivo de 500MB. Finalmente, al tener un factor de realización de 2, el número total de bloques que ocupa es de 4 bloques * 2 replicas = 8 bloques

8. Si el tamaño de un archivo es de 800 MB, el tamaño de bloque es de 128 MB y el factor de replicación es 3, ¿cuál es el número total de bloques que ocupa? ¿Cuál es el tamaño de cada bloque?

Para calcular el número de bloques total, habría que dividir los 800MB entre distintos bloques que ocupan 128MB. Esto quiere decir:

128 +128 + 128 + 128 + 128 + 128 + 32 = 800

Esto quiere decir que tienen que crear un total de 7 bloques en total para almacenar el archivo de 800MB. Finalmente, al tener un factor de realización de 3, el número total de bloques que ocupa es de 7 bloques * 3 replicas = 21 bloques

En este caso, todos los bloques tendrían el mismo tamaño excepto el último, que tiene un tamaño de 32MB,.

9. ¿Por qué se realizan las replicaciones de datos en diferentes racks?

La razón por la cual se realizan replicaciones de datos en diferentes racks es porque, en caso de que un rack falle, el sistema pueda mantener la integridad de los datos y que el sistema no se caiga por completo.

10. ¿Hadoop es adecuado para manejar datos de transmisión?

No, ya que Hadoop está pensado para el procesamiento por lotes, poniendo el énfasis en el diseño para altas tasas de rendimiento de datos, que se adaptarán al acceso de transmisión a conjuntos de datos, lo que el uso interactivo de Hadoop no es adecuado.