Tarea 3 Sistemas Distribuidos
Javier Guerrero - Darío Marmie

Para contar palabras dentro de un código de MapReduce se debe utilizar un trabajo de WordCount. Lo que principalmente realiza un trabajo de WordCount es tomar los datos, separarlos en bloques, y reordenar los outputs e inputs para reducir el número de tareas. En Hadoop, existe el Hadoop Distributed File System, que permite a todos los nodos pertenecientes al clúster acceder al mismo directorio de archivos. En dicho directorio se guardan los inputs y outputs de un trabajo.

En el caso de esta tarea, se utilizó un proceso de MapReduce que realizaba el conteo en base al archivo que se le entregaba. El Mapper entregaba una palabra, un 1 y el número del archivo del que se está leyendo (Dicho número de archivo se encontraba en la primera línea de los .txt a procesar). Luego, el Reducer toma todas las palabras y genera un diccionario donde asocia el archivo y el contador a la palabra respectiva.

La idea detrás del uso de MapReduce y de WordCount especialmente en Hadoop, es para poder dividir la carga que se genera cuando se trabaja con un gran volúmen de datos, pudiendo repartir los trabajos a través de los distintos nodos que pertenezcan al clúster de Hadoop. Esto se logra a través del uso de YARN (Yet Another Resource Negotiator), que permite la distribución de recursos en conjunto con el Resource Manager.