MBIT School
Madrid Business Intelligence Technology

Taller MapReduce

Diego J. Bodas Sagi

Índice

- Presentación del ejercicio
- Técnicas útiles para resolver el ejercicio
  - Paso de variables a la configuración
  - Memoria cache
- Conclusiones

Presentación del ejercicio

MBIT School
Madrid Business Intelligence Technology

```
conf/performance/MayrL76:::Heinrich C. Mayr::Peter C. Lockemann:::Formal Modelling of Discrete Dynamic Systems.
conf/performance/HsiaoL87:::Man-tung T. Hsiao::Aurel A. Lazar:::A Game Theoretic Approach to Decentralized Flow Control of Markovian Queueing Networks.
conf/performance/Epema90:::Dick H. J. Epema:::Mean Waiting Times in a General Feedback Queue with Priorities.
conf/performance/DowdyPS83:::Lawrence W. Dowdy::Alfredo de J. Perez-Davila::Lindsey E. Stephens:::Performance Bounds Based upon Throughput Curve Properties.
conf/performance/KrzesinskiT77:::Anthony E. Krzesinski::Peter Teunissen:::Analysis of the Page Size Problem Using a Network Analyizer.
conf/performance/PinskyY84:::Eugene Pinsky::Yechiam Yemini:::A Statistical Mechanics of Some Interconnection Networks.
conf/performance/Minetti76:::Vito Minetti:::Performance Evaluation of a Batch-time Sharing Computer System Using a Trace Driven Model.
conf/performance/MitraM84a:::Debasis Mitra::J. McKenna:::Some Results on Asymptotic Expansions for Closed Markovian Networks with State Dependent Service Rates.
journals/lncs/Kiss91:::Tibor Kiss:::The Grammars of LILOG.
journals/lncs/Wilms91:::Jan Wilms:::The Text Understanding System LEU/2.
```

Enunciado

MBIT School
Madrid Business Intelligence Technology

- Esta práctica se puede hacer de forma individual o en grupo. En caso de hacerse en grupo el número máximo de alumnos permitidos es 3.
- Objetivo: los alumnos han recibido un archivo comprimido (allPapers.zip) que contiene un amplio listado de publicaciones técnicas y científicas (se ha comentado en clase). Con ese material se pueden calcular cosas como:
- Palabras (no stopwords) más frecuentes en los títulos. Se valorará de forma especial que las stopwords provengan de un fichero de texto y no se introduzcan manualmente en el código
  - StopWords: “the”, “a”, “an”, “in”…
  - No aportan información al análisis
  - Multitud de ficheros disponibles en internet para los distintos idiomas

Enunciado

- Número de publicaciones para cada autor.
- Para cada autor, número de publicaciones en equipo.
- Para cada autor, número de publicaciones en solitario.
- Dadas las 5 palabras más frecuentes en los títulos, ¿quienes son los autores que más usan esas palabras en sus publicaciones (ya sean autores únicos o co-autores)?

A considerar

- Se debe entregar:
  - Una pequeña memoria en formato pdf explicando de forma básica cómo se ha resuelto el caso de uso.
  - Código del proyecto (anexar el proyecto eclipse sin los artículos).
  - Archivo LEEME.txt con los nombres y email de los integrantes del proyecto.
- Valoración de la práctica
  - Se valorará la limpieza, precisión y organización del código.
  - Los comentarios son obligatorios en el código.
  - Usar nombres descriptivos en variables y funciones.

Ayuda 1: conjuntos en Java

```
private boolean stopWord(String word) {
    Set<String> stopWords = new HashSet<String>();
    stopWords.add("a");
    stopWords.add("the");
    stopWords.add("for");
    stopWords.add("in");
    stopWords.add("of");
    stopWords.add("or");
    stopWords.add("why");
    stopWords.add("not");
    stopWords.add("no");
    stopWords.add("to");
    return stopWords.contains(word);
  }
```

Pero… ¿vamos a añadir a mano todas las stopwords?

Ayuda 2: paso de variables a la configuración

MBIT School
Madrid Business Intelligence Technology

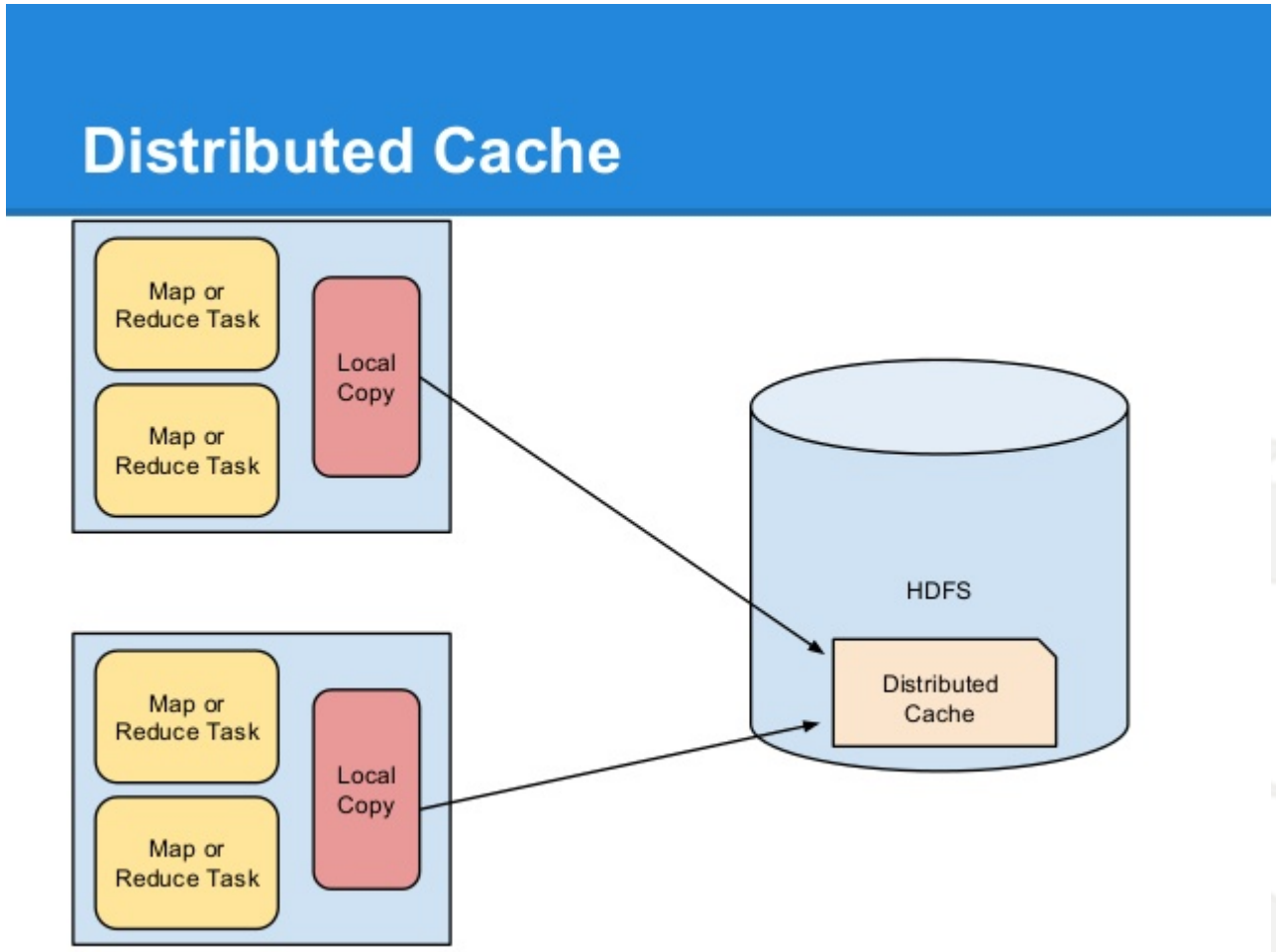
- En el driver, usando la configuración bajo la que se lanza el job, se puede configurar la variable que se desee pasando el valor que queremos compartir con mappers y reducers
  - Por ejemplo, si quiero compartir la ruta de un fichero:
    - conf.setStrings("filePath", "./StopWords/stopWords.txt");
  - Recordar la clase "Path" de java para todo lo que tenga que ver con rutas de directorios
  - En el mapper:
    - context.getConfiguration().get("filePath")
    - Analizar lo que devuelve el método anterior y cómo recogerlo

Ayuda 2: usando valores pasados como parámetros

MBIT School
Madrid Business Intelligence Technology

- En la orden de ejecución, el usuario, puede pasar tantos argumentos como quiera, por ejemplo:
  - … input_dir output_dir stopwords_dir

Ayuda 3: uso de la memoria caché



Ayuda 3: procedimiento

MBIT School
Madrid Business Intelligence Technology

```
 //añadimos el fichero de stopwords a la cache distribuida
job.addCacheFile(new Path(args[3]).toUri());

//confirmamos que el fichero se ha añadido
URI[] cacheFiles = job.getCacheFiles();
if (cacheFiles != null) {
        for (URI cacheFile : cacheFiles) {
                System.out.println("Cache file ->" + cacheFile);
        }
}
```

Ayuda 3: cargar las stopwords en el método setup del mapper

MBIT School
Madrid Business Intelligence Technology

```
//comprobamos que las Stopwords no estan cargadas
if (StopWordsList == null) {

    StopWordsList = new ArrayList<String>();
    if (context.getCacheFiles() != null && context.getCacheFiles().length > 0) {
      URI mappingFileUri = context.getCacheFiles()[0];
      String StopWordsStr = new String();
      //leemos el fichero de la cache distribuida.
      if (mappingFileUri != null) {
        BufferedReader fis =
            new BufferedReader(new FileReader(new File(
                mappingFileUri.getPath()).getName()));
        String pattern;
        while ((pattern = fis.readLine()) != null) {
          StopWordsStr += pattern;
        }
        fis.close();
```

Ayuda 3: ¿qué está pendiente?

MBIT School
Madrid Business Intelligence Technology

- Procesar las stop words para añadirlas a la lista o conjunto correspondiente

- EJERCICIO

Resumen

MBIT School
Madrid Business Intelligence Technology

- Hemos presentado un ejemplo de aplicación “curioso” donde se puede desarrollar la potencia de Hadoop
- Recordemos que trabajamos con Java, lo que implica que podemos usar todo lo que está disponible en Java, como los conjuntos
- Se pueden pasar variables a la configuración que son empleadas en mappers o reducers
- La caché distribuida es un excelente medio de cargar información que será compartida por todos
- El ejercicio queda pendiente de finalización…