



UNIVERSIDAD DE EXTREMADURA

CENTRO UNIVERSITARIO DE MÉRIDA

**INGENIERO EN INFORMÁTICA EN TECNOLOGÍAS DE LA
INFORMACIÓN**

PROYECTO FIN DE GRADO

**INTEGRACIÓN DE UNA
HERRAMIENTA DE CÓMPUTO
EVOLUTIVO CON UNA DE
PROCESAMIENTO MASIVO DE
INFORMACIÓN**

Autor: Daniel Lanza García

Mérida - Junio 2015



UNIVERSIDAD DE EXTREMADURA

CENTRO UNIVERSITARIO DE MÉRIDA

**INGENIERO EN INFORMÁTICA EN TECNOLOGÍAS DE LA
INFORMACIÓN**

PROYECTO FIN DE GRADO

**INTEGRACIÓN DE UNA
HERRAMIENTA DE CÓMPUTO
EVOLUTIVO CON UNA DE
PROCESAMIENTO MASIVO DE
INFORMACIÓN**

Autor: Daniel Lanza García

Director: Francisco Fernández de Vega

Codirector: Francisco Chávez de la O

Mérida - Junio 2015

Prólogo

En las primeras líneas que describen este proyecto fin de carrera, ...

Agradecimientos

Quiero dar mi mas sincero agradecimiento a A todos muchas gracias.

Índice general

1. Introducción	1
1.1. Motivaciones	1
1.2. Objetivos	1
1.3. Recursos empleados	1
1.4. Organización del documento	1
2. Análisis del sistema	2
2.1. Algoritmos evolutivos	2
2.1.1. Paralelización	4
2.2. Procesamiento masivo de información	4
2.2.1. Modelo computacional: Map/Reduce	5
3. Desarrollo del proyecto	6
3.1. Estudio de ECJ, herramienta de cómputo evolutivo	6
3.1.1. Proceso evolutivo	7
3.2. Estudio de Hadoop, herramienta de procesamiento masivo de información	8
3.2.1. El sistema de ficheros	8
3.2.2. La capa MapReduce	9
3.3. Implementación	10
4. Manual de usuario	11
4.1. Obtención	11
4.2. Importar a entorno de desarrollo	12
4.3. Compilación	12
5. Trabajos futuros	14
6. Conclusiones	15
7. Anexo I. Código fuente	16
Bibliografía	17

Índice de figuras

2.1. Fases del proceso evolutivo	3
3.1. Clases que representan el proceso evolutivo en ECJ	7
3.2. Procesos en los diferentes nodos de un cluster Hadoop	8
3.3. Fases de un trabajo MapReduce	10

Índice de tablas

1. INTRODUCCIÓN

1.1 Motivaciones

La computación evolutiva está adquiriendo cada vez mas importancia a lo largo de los años, sus cada vez mas aplicaciones en sistemas de diferente naturaleza la hacen imprescindible para la resolución de problemas, que haciendo uso de otro modelo computacional, no podrían ser resueltos.

Con el paso de los años y la evolución de los sistemas computacionales, la computación evolutiva se ha intentado aplicar para la resolución de problemas cada vez mas complejos, lo cual en la mayoría de los casos, suele conllevar el uso de más recursos como capacidad de cómputo, memoria y tiempo. Esto hace que se busquen soluciones para mejorar el uso de estos recursos.

Numerosas investigaciones se han llevado a cabo con el fin de minimizar el uso de uno de los recursos mas importantes mencionados anteriormente, el tiempo, se han diseñado algoritmos en este modelo cuya ejecución puede tomar años y que por lo tanto su tiempo de ejecución es impracticable. Varias metodologías se han aplicado para reducir el tiempo de ejecución, una de ellas es la paralelización de parte del proceso, esto puede llevarse a cabo con los procesadores de última generación los cuales poseen varios núcleos de procesamiento o también se puede conseguir con el uso de varios computadores conectados en red.

Este trabajo plantea una solución para utilizar los recursos disponibles de forma eficiente y así reducir los tiempos de ejecución, la solución que se describirá hace uso de una herramienta que explota ambas metodologías, ejecución paralela en procesadores multi-núcleo y uso de numerosas computadoras.

1.2 Objetivos

1.3 Recursos empleados

1.4 Organización del documento

2. ANÁLISIS DEL SISTEMA

En este capítulo nos sumergiremos en los dos campos de conocimiento que este proyecto contempla, la computación evolutiva y el procesamiento masivo de información. Ambos están adquiriendo cada vez más importancia en el mundo de la computación y la resolución de problemas no convencionales.

Se cubrirán temas como porque surgen estas metodologías, sus propósitos y de que manera intentan resolver el problema para el cual han sido desarrolladas.

2.1 Algoritmos evolutivos

Existen problemas computacionales que no pueden ser resueltos con metodologías tradicionales, o porque no existe una solución que pueda proporcionar un resultado aceptable o porque la solución desarrollada necesita de un tiempo y recursos que no son manejables. Para este tipo de problemas se buscan implementaciones que no proporcionan siempre la mejor solución pero que intentar acercarse lo máximo posible, a estos problemas se les conoce como problemas de optimización.

Se han desarrollado diferentes formas de afrontar estos problemas de optimización y una de ellas es la computación evolutiva, este modelo se basa en la teoría de la evolución que Charles Darwin postuló. Esta idea de aplicar la teoría Darwiniana de la evolución surgió en los años 50 y desde entonces han surgido diferentes corrientes de investigación:

- Algoritmos genéticos, desarrolla programas informáticos, tradicionalmente representados en la memoria como estructuras de árboles. Los árboles pueden ser fácilmente evaluados de forma recursiva. Cada nodo del árbol tiene una función como operador y cada nodo terminal tiene un operando.
- Programación evolutiva, una variación de los algoritmos genéticos, donde lo que cambia es la representación de los individuos. En el caso de la Programación evolutiva los individuos son ternas cuyos valores representan estados de un autómata finito.
- Estrategias evolutivas, se diferencia de las demás en que la representación de cada individuo de la población consta de dos tipos de variables: las variables objeto, posibles valores que hacen que la función objetivo alcance el óptimo, y las variables estratégicas, las cuales indican de qué manera las variables objeto son afectadas por la mutación..

Este modelo por tanto, se basa generalmente en la evolución de una población y la lucha por la supervivencia. En su teoría, Darwin dictaminó que durante muchas generaciones, la variación, la selección natural y la herencia dan forma a las especies con el fin de satisfacer las demandas del entorno, de la misma manera pero con el fin de satisfacer una buena solución se basa la computación evolutiva. Podemos observar entonces, algunos elementos importantes como son:

- La población de individuos, donde cada una de ellos representa directa o indirectamente una solución al problema.
- Aptitud de los individuos, atributo que describe cuanto de cerca esta este individuo (solución) de la solución óptima.
- Procedimientos de selección, es la estrategia a seguir para elegir los progenitores de la siguiente

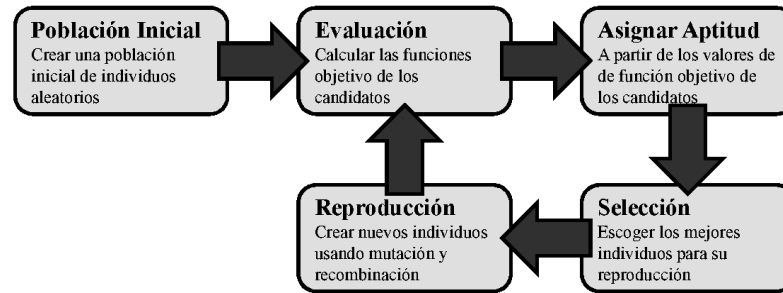


Figura 2.1: Fases del proceso evolutivo

generación. Esta normalmente elige a los individuos mas apto pero existen otras muchas técnicas.

- Procedimiento de transformación, se lleva a cabo sobre los individuos seleccionados y puede consistir en la combinación de varios individuos o en la mutación (cambios normalmente aleatorios en el individuo).

Para llevar a cabo la implementación en computadoras, se ha dividido el problema en diferentes fases y procedimientos, los cuales se ejecutan con un orden determinado, describimos a continuación de forma general como se lleva a cabo la resolución de problemas utilizando este modelo.

1. Inicialización, en esta primera fase se genera la población inicial, normalmente se genera una cantidad de individuos que es configurada y cada uno de ellos se genera de manera aleatoria, siempre respetando las restricciones que el problema imponga a la solución.
2. Evaluación, se calcula la aptitud de cada uno de los individuos de la población para poder determinar posteriormente cuales son más aptos.
3. Las fases que siguen a continuación se repiten hasta que se cumpla una de las siguientes dos condiciones, o que se encuentre la solución óptima o que se alcance un limite impuesto por el programador como un número de generaciones máximo o un tiempo máximo.
 - a) Selección, siguiendo la estrategia de selección de progenitores elegida, se eligen individuos de la población. Normalmente los que sean mas aptos.
 - b) Procreación, utilizando los individuos seleccionados, se combinan para generar nuevos individuos, y con ellos una nueva población (generación).
 - c) Mutación, se eligen normalmente de forma aleatoria individuos a los que se les aplica una modificación también aleatoria, estos nuevos individuos se incluyen también en la nueva población.
 - d) Evaluación, todos los individuos de la nueva población son evaluados.

Con el fin de entender mejor este proceso, se muestra una imagen (ver figura 2.1) donde se observan cada una de las etapas descritas anteriormente.

Varias herramientas han surgido en la comunidad para ayudar a la investigación de este modelo, en diferentes lenguajes de programación y plataformas. En nuestro caso hemos elegido ECJ que es un framework bien conocido por la comunidad, implementado en el lenguaje de programación Java y que posee una flexibilidad importante para la ejecución de problemas de muy distinta naturaleza. Más adelante (ver apartado 3.1) se describe con más detalle esta herramienta, explicando su funcionamiento e implementación.

2.1.1 Paralelización

Como hemos comentado anteriormente (ver apartado 1.1) este proyecto surge de la necesidad de optimizar el uso de recursos cuando se intentan resolver problemas complejos con este modelo. Con este fin, y con la posibilidad de paralizar el procesamiento de algunas de las partes del proceso, surge la viabilidad de este proyecto.

Varias partes del proceso evolutivo pueden ser paralizadas, pero no todas merecen el esfuerzo ya que el coste en algunas de ellas es mínimo. Una de las fases que suele conllevar un coste computacional alto y que su paralelización en la mayoría de problemas es sencilla, es la fase de evaluación de individuos.

Se han utilizado diferentes técnicas para este propósito, una de ellas es la ejecución de la evaluación de individuos haciendo uso procesadores multinúcleo/multihilo. Esta técnica consigue buenos resultados pero se limita a las capacidades del procesador que contenga esa computadora. Otro intento para llevar a cabo la paralelización del proceso ha sido la ejecución en diferentes máquinas, las cuales se conectan haciendo uso de una red. Este planteamiento requiere de una implementación más compleja y no suele explotar todos los recursos, además de que algunas soluciones planteadas carecen de la escalabilidad deseada. También han surgido implementaciones que hacen uso de ambas técnicas, esta solución suele ser la más apropiada ya que hace un uso más eficiente del hardware proporcionado, aunque requiere de una implementación aún más costosa.

2.2 Procesamiento masivo de información

Vivimos en el momento más álgido en la generación de información, nunca antes había existido plataformas que generaran la cantidad de información que se genera hoy. El hecho de que del análisis de grandes cantidades de información se puedan extraer valiosos datos como estrategias de negocio, hacen que numerosas empresas y organizaciones almacenen cantidades ingentes de información para poder sacarle el máximo partido.

La generación de información se está produciendo en ámbitos muy dispares, estos pueden ser redes sociales que mantienen millones de usuarios, grandes empresas con muchos clientes, laboratorios de física con redes de millones sensores y muchos otros ejemplos que podríamos mencionar. Todos ellos queriéndole sacan el máximo valor a la información que recaban.

■ Computación distribuida

Cuando hablamos de cantidades de información del orden de terabytes o petabytes no podemos pensar en otra cosa que no sea computación distribuida. Para una sola máquina, el tiempo que conllevaría procesar esas cantidades de datos podrían ser del orden de años.

Por tanto, la computación distribuida es la única solución con el hardware que hoy en día manejamos. Esta solución implica el uso de múltiples computadores conectados a la red, cada una de las cuales tiene su propio procesador, arquitectura, etc, con lo que pueden ser totalmente heterogéneas, en contraposición a la computación paralela, que consiste en utilizar más de un hilo de procesamiento simultáneamente para ejecutar un único programa. Idealmente, el procesamiento paralelo permite que un programa se ejecute más rápido, en la práctica, suele ser difícil dividir un programa de forma que CPU separadas ejecuten diferentes porciones del programa sin ninguna interacción.

2.2.1 Modelo computacional: Map/Reduce

En este ámbito, surge la necesidad de diseñar herramientas que puedan no solo mantener esta información, si no también que tengan la capacidad de analizarla y extraer el valor que se desea de una forma distribuida y con un modelo sencillo.

Google, el buscador de internet más utilizado en el mundo, ha hecho frente a este problema antes que nadie ya que desde hace años maneja cantidades de información realmente grandes, es por esto que sus investigaciones y experiencia son avanzadas. Varios años atrás hicieron pública [1] una solución para el análisis de grandes cantidad de datos de forma distribuida y con un modelo que da solución a la complejidad de dividir el problema para poder paralelizarlo, ha este modelo se le conoce como Map/Reduce. Esta publicación ha dado pie a que se implemente una herramienta que se conoce con el nombre de Hadoop [4], la cual se describe con más detalle en un capítulo posterior (ver apartado 3.2). La creación de esta herramienta y el hecho de que se hayan obtenidos buenos resultados de ella, ha provocado que surjan otras muchas herramientas a su alrededor, las cuales ayudan a diferentes tareas como el volcado de información (Sqoop), bases de datos para consulta (Impala), gestion de flujos de datos (Flume) y otras muchas.

La propuesta de computación distribuida que se hace en esta publicación, hace uso de un modelo computacional anteriormente conocido en la programación funcional. Este modelo se basa en aplicar dos funciones básicas conocidas como Map (mapeo) y Reduce (reducción). La función de mapeo consiste en aplicar una transformación o procedimiento a todos los datos, obteniendo así la entrada de la siguiente fase, reducción, la cual consiste en "resumir", aplicando la misma función a diferentes partes de la salida de la fase de mapeo, obteniendo finalmente la salida deseada. Numerosas fases de mapeo y reducción pueden ser concatenadas en el orden que se quiera con el fin de producir el resultado esperado. Este modelo es el implementado en Hadoop pero con algunas peculiaridades(ver apartado ??).

3. DESARROLLO DEL PROYECTO

En este capítulo se describirá el desarrollo del proyecto, el cual ha consistido en primer lugar en analizar las dos herramientas que se integraran, ECJ y Hadoop, con el objetivo de conocerlas en profundidad y así poder integrarlas de la mejor manera posible. Una vez analizadas ambas, se explicará el procedimiento llevado a cabo para su integración.

3.1 Estudio de ECJ, herramienta de cómputo evolutivo

Surgen numerosas herramientas en este campo que ayudan a su investigación, una bien conocida por la comunidad es ECJ. Esta herramienta ha sido llevada a cabo por el departamento de ciencias de la computación de la universidad de George Mason, US y ahora mismo se encuentra en su version 22.

Esta herramienta ha sido concedida para proporcionar una amplia flexibilidad para poder albergar numerosos tipos de problema, además, persiguiendo este mismo objetivo, ha sido desarrollada en el lenguaje de programación Java, lo que permite que pueda ser ejecutada en cualquier sistema operativo tanto Windows como Linux.

Algunas de sus características mas importantes y por las cuales ha adquirido la popularidad que posee son las siguientes:

- Facilidad de analizar la ejecución con una util implementación de logging.
- Posibilidad de generar puntos de restauración, los cuales permiten detener la ejecución y reiniciarla en otro momento.
- Ficheros de parámetros anidados, lo cual permite jerarquizar la configuración y mantener mas clara su configuración.
- Ejecución multihilo de diferentes partes del proceso, esto da la posibilidad de paralizar el proceso en procesadores que posean varias unidades de procesamiento.
- Generación de numeros pseudoaleatorios de manera que se permite reproducir los resultados generados en anteriores ejecuciones.
- Soporte para diferentes técnicas evolutivas.
- Proceso de reproducción muy flexible representado por una jerarquía de operaciones.

Estas y otras características han hecho que ECJ se posicione como una de las herramientas favoritas para la investigación de la computación evolutiva. El hecho de su amplia utilización en la comunidad ha motivado el uso de ECJ en este trabajo ya que de este modo el publico al que puede ir dirigido es más amplio y mas grupos de investigación puedan beneficiarse de los resultados que de este se obtengan.

La ejecución de cualquier algoritmo en ECJ esta guiada a través de los ficheros de configuración que deben ser anteriormente establecidos. Estos ficheros de configuración poseen una estructura jerárquica lo que permite que un fichero de configuración pueda incluir a otro/s y sobrescribir parámetros que hayan sido establecidos. Esta estructura permite que se pueda describir un problema con pocos parámetros ya que los que se utilicen de manera general estarán contenidos en

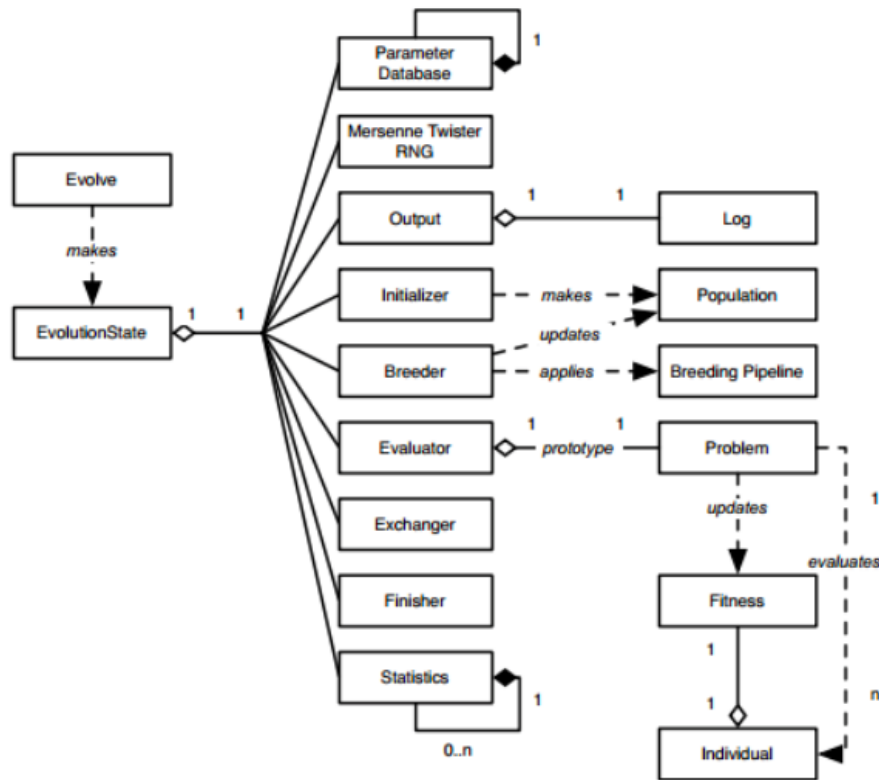


Figura 3.1: Clases que representan el proceso evolutivo en ECJ

otros que serán simplemente incluidos. En estos ficheros de configuración se incluyen numerosos parámetros que describen el comportamiento del proceso evolutivo y existen algunos de ellos que son obligatorios especificar, como por ejemplo los que determinan que clase implementa cada una de las partes de la evolución.

3.1.1 Proceso evolutivo

En la documentación de ECJ se facilita un diagrama (ver figura 3.1) que describe de forma clara como se implementa el proceso evolutivo en esta herramienta. Podemos observar que la clase que inicia el proceso evolutivo es *Evolve*, la cual genera un *EvolutionState* que es el contiene cada una de las clases/etapas del proceso. La mayor parte de estas clases deben ser especificadas en los ficheros de configuración, de esta manera ECJ sabe que clase implementa cada una de las partes.

Como se ha comentado anteriormente (ver apartado 2.1.1), la parte del proceso que suele ser más costosa es la evaluación de individuos, representada (ver figura 3.1) en ECJ por la clase *Evaluator*. En esta herramienta esta es la clase encargada de la evaluación de cada uno de los individuos de la población, lo cual suele ser lo más costoso computacionalmente y por consiguiente será la parte de ECJ donde se centre la implementación de la integración con Hadoop.

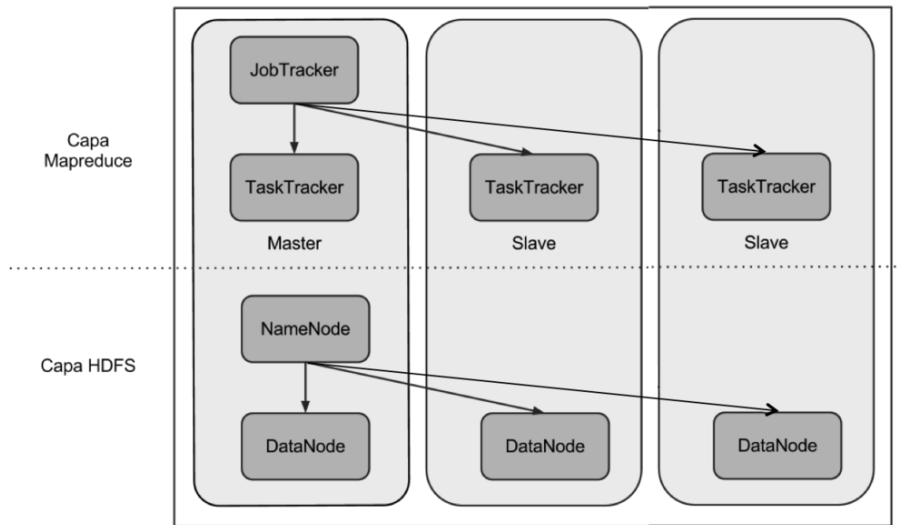


Figura 3.2: Procesos en los diferentes nodos de un cluster Hadoop

3.2 Estudio de Hadoop, herramienta de procesamiento masivo de información

Apache Hadoop es un framework de software que soporta aplicaciones distribuidas con capacidades de procesamiento de cantidades masivas de información. Permite a las aplicaciones trabajar con miles de nodos y petabytes de datos. Hadoop se inspiró en los documentos de Google para MapReduce [1] y Google File System (GFS) [3].

Hadoop es un proyecto de alto nivel Apache que está siendo construido y usado por una comunidad global de contribuyentes bastante amplia y activa, mediante el lenguaje de programación Java. Yahoo! ha sido el mayor contribuyente al proyecto, y usa Hadoop extensivamente en su negocio.

Hadoop requiere tener instalados entre nodos en el clúster JRE 1.6 o superior, y SSH. Un clúster típico incluye un nodo maestro y múltiples nodos esclavo. El nodo maestro consiste en un proceso jobtracker (rastreador de trabajo), tasktracker (rastreador de tareas), namenode (nodo de nombres), y datanode (nodo de datos). Un esclavo o compute node (nodo de cómputo) consisten en un nodo de datos y un rastreador de tareas (ver figura 3.2).

Como se puede apreciar, Hadoop puede ser claramente dividido en dos partes, el sistema de ficheros, HDFS, y la parte que implementa la parte de procesamiento de información, la capa MapReduce.

3.2.1 El sistema de ficheros

El Hadoop Distributed File System (HDFS) es un sistema de archivos distribuido, escalable y portátil escrito en Java para el framework Hadoop. Cada nodo en una instancia Hadoop típicamente tiene un único nodo de datos; un clúster de datos forma el clúster HDFS. La situación es típica porque cada nodo no requiere un nodo de datos para estar presente.

Cada nodo sirve bloques de datos sobre la red usando un protocolo de bloqueo específico para HDFS. El sistema de archivos usa la capa TCP/IP para la comunicación; los clientes usan RPC para comunicarse entre ellos. El HDFS almacena archivos grandes, a través de múltiples máquinas. Consigue fiabilidad mediante replicado de datos a través de múltiples hosts, y no requiere

almacenamiento RAID en ellos. Con el valor de replicación por defecto, 3, los datos se almacenan en 3 nodos: dos en el mismo rack, y otro en un rack distinto. Los nodos de datos pueden hablar entre ellos para reequilibrar datos, mover copias, y conservar alta la replicación de datos. HDFS no cumple totalmente con POSIX porque los requerimientos de un sistema de archivos POSIX difieren de los objetivos de una aplicación Hadoop, porque el objetivo no es tanto cumplir los estándares POSIX sino la máxima eficacia y rendimiento de datos. HDFS fue diseñado para gestionar archivos muy grandes. Algo que debe ser tenido en cuenta es que no proporciona alta disponibilidad, ya que la caída de su nodo maestro supone la caída del sistema de ficheros.

Su escalabilidad y su rendimiento se pueden llevar a cabo gracias a la falta de una de las características más comunes en un sistema de ficheros, la capacidad de modificar el contenido de los ficheros que contiene, una característica no necesaria para el propósito para el que está diseñado.

Aunque Hadoop tiene su propio sistema de ficheros, HDFS, no está cerrado al uso de tan solo ese sistema, otros sistemas de ficheros son también compatibles como Amazon S3, CloudStore o FTP.

3.2.2 La capa MapReduce

Aparte del sistema de archivos, está el motor MapReduce, que consiste en un Job Tracker (rastreador de trabajos), para el cual las aplicaciones cliente envían trabajos MapReduce.

El rastreador de trabajos (Job Tracker) impulsa el trabajo fuera a los nodos Task Tracker disponibles en el clúster, intentando mantener el trabajo tan cerca de los datos como sea posible. Con un sistema de archivos consciente del rack en el que se encuentran los datos, el Job Tracker sabe qué nodo contiene la información, y cuáles otras máquinas están cerca. Si el trabajo no puede ser almacenado en el nodo actual donde residen los datos, se da la prioridad a los nodos del mismo rack. Esto reduce el tráfico de red en la red principal backbone. Si un Task Tracker (rastreador de tareas) falla o no llega a tiempo, la parte de trabajo se reprograma. El TaskTracker en cada nodo genera un proceso separado JVM para evitar que el propio TaskTracker mismo falle si el trabajo en cuestión tiene problemas. Se envía información desde el TaskTracker al JobTracker cada pocos minutos para comprobar su estado. El estado del Job Tracker y el TaskTracker y la información obtenida se pueden ver desde un navegador web proporcionado por Jetty.

■ Un trabajo MapReduce

Un trabajo MapReduce aplica dos funciones, como su nombre indica, una es una función Map (una operación a cada uno de los registros de entrada) y una función Reduce (una operación que resume la salida de la función Map) para aplicar este modelo, Hadoop divide el proceso en diferentes fases, podemos resumirlas de la siguiente manera:

1. Se divide los ficheros de entrada en tantas partes como número de tareas Map vayan a componer el trabajo.
2. Cada tarea Map lee una de las partes y aplica a cada registro de entrada una función que define el usuario. La salida de esta fase son tuplas de clave-valor.
3. La salida de cada tarea Map se organiza de forma local en el nodo que la ha ejecutado formando grupos, cada uno de ellos contiene todas las tuplas con la misma clave.
4. Se transfieren todos los grupos con la misma clave al nodo que vaya a ejecutar la tarea de Reduce, de forma que las claves se entregan de manera ordenada.
5. Se unen todos los grupos recibidos con la misma clave creando así la entrada de la tarea de Reduce.

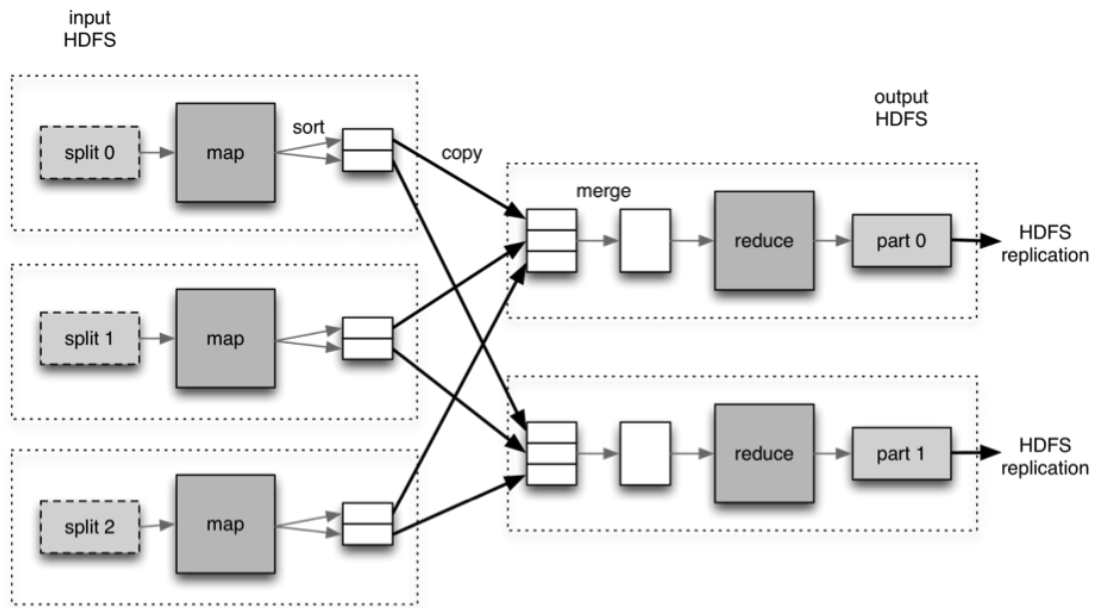


Figura 3.3: Fases de un trabajo MapReduce

6. Se aplica la función de Reduce la cual recibe todas las tuplas con la misma clave y produce también tuplas de clave-valor.
7. La salida de la operación de Reduce se almacena en el sistema de ficheros.

Todo este proceso se ve resumido el diagram (ver figura 3.3), donde se muestran cada una de estas fases.

La idea de la integración es que la entrada del trabajo de MapReduce este compuesta por todos los individuos a evaluar, de manera que la función Map sea la que evalúe cada individuo, prescindiendo de la etapa de Reduce, de manera que se escriba en el sistema de ficheros directamente la salida de la fase de Map, el resultado de la evaluación de los individuos.

3.3 Implementación

4. MANUAL DE USUARIO

4.1 Obtención

El primer paso para la utilización de la solución implementada, no puede ser otro que obtenerla. Con el fin de que la comunidad pueda conseguirla sin problema, se ha creado un repositorio publico desde donde se puede descargar.

El repositorio está localizado en la conocida página por los desarrolladores GitHub (<https://github.com>), en ella se encuentran numerosos proyectos de código abierto y en la cual se ha almacenado este.

Todos los proyectos almacenados en esta página hacen uso de una herramienta de control de versionado llamada Git, la cual hemos también utilizado en este proyecto.

¿Qué es Git?

Es un software de control de versiones diseñado por Linus Torvalds, pensando en la eficiencia y la confiabilidad del mantenimiento de versiones de aplicaciones cuando estas tienen un gran número de archivos de código fuente. Hay algunos proyectos de mucha relevancia que ya usan Git, en particular, el grupo de programación del núcleo Linux. Algunas de sus características mas destacadas son:

- Fuerte apoyo al desarrollo no lineal, por ende rapidez en la gestión de ramas y mezclado de diferentes versiones.
- Gestión distribuida. Git le da a cada programador una copia local del historial del desarrollo entero, y los cambios se propagan entre los repositorios locales.
- Los almacenes de información pueden publicarse por HTTP, FTP, rsync o mediante un protocolo nativo, ya sea a través de una conexión TCP/IP simple o a través de cifrado SSH.
- Los repositorios Subversion y svk se pueden usar directamente con git-svn.
- Gestión eficiente de proyectos grandes.

El proyecto se puede encontrar a través del buscador del sitio escribiendo "ecj_hadoop", o accediendo directamente al repositorio siguiendo esta dirección: https://github.com/dlanza1/ecj_hadoop . Una vez en el repositorio, el proyecto se puede obtener de dos formas, una es pulsando en el botón "Download ZIP" (situado en la columna de la derecha), y una vez descargado descomprimir el fichero, y la otra es clonar el repositorio con Git (crear un repositorio igual de forma local).

Para clonar el repositorio con Git se debe tener instalada esta herramienta en el sistema [5] o utilizar un entorno de desarrollo como Eclipse, el cual la trae incluida.

Para clonarlo desde un entorno de desarrollo como Eclipse [2] el procedimiento suele ser el de importar un proyecto pero con la diferencia de que se debe indicar que la fuente es un repositorio Git, nos solicitará el repositorio que queremos clonar y es ahí donde debemos escribir la URL del repositorio. Algo que debe ser aclarado es que esta operación creará un repositorio local, no genera el proyecto en el entorno de desarrollo, lo cual se explica en la sección siguiente.

Si de otro modo, tenemos la herramienta instalada en el sistema, debemos en primer lugar abrir una consola, posteriormente dirigimos al directorio donde queremos crear el repositorio local y por último clonar el repositorio con el siguiente comando:

```
[usu@host repo]$ git clone https://github.com/dlanza1/ecj_hadoop
```

Una vez clonado tendremos una copia idéntica del repositorio la cual contiene la última versión del proyecto además de toda la información necesaria para que funcione el sistema de versionado utilizada por Git.

4.2 Importar a entorno de desarrollo

Esta sección tiene como objetivo explicar el procedimiento para importar el proyecto en un entorno de desarrollo, esto no es necesario para su ejecución por lo que no estén interesados en modificar/ampliar el proyecto pueden continuar la lectura en la sección siguiente (Compilación).

En el apartado anterior se ha explicado como obtener el proyecto, pero lo obtenido es básicamente los ficheros de código fuente, no se incluyen proyectos de entornos de desarrollo ni ejecutables.

Si acabamos de descargar/clonar el proyecto, no lo tendremos importado en nuestro entorno, para llevar a cabo esta operación tan solo debemos dirigirnos al menu de importación de proyectos, solucionar que es un proyecto Maven y seleccionar el directorio que contiene el proyecto.

Otra opción, teniendo la herramienta Maven instalada en el sistema [6], es generar primero el proyecto del entorno de desarrollo y posteriormente importarlo como un proyecto normal. Para generar el proyecto del entorno de desarrollo debemos abrir una consola, dirigimos al directorio donde se sitúa el proyecto y ejecutar el comando Maven correspondiente a nuestro entorno, por ejemplo para Eclipse debemos ejecutar el siguiente comando:

```
[usu@host repo]$ mvn eclipse:eclipse
```

Esto nos generará el proyecto Eclipse y podremos importarlo como cualquier otro proyecto.

4.3 Compilación

Como se mencionaba anteriormente, los ficheros ejecutables no son proporcionados por lo que se deben generar. Con el uso de la herramienta de construcción de proyectos, Maven, explicaremos como construir el proyecto para que pueda ser ejecutado, no obstante, si tenemos el proyecto importado en un IDE, no es necesario utilizar Maven para compilarlo, podemos utilizar los usuales procedimientos para compilarlo.

¿Qué es Maven?

Maven es una herramienta de software para la gestión y construcción de proyectos Java creada por Jason van Zyl, de Sonatype, en 2002. Tiene un modelo de configuración de construcción simple, basado en un formato XML. Es un proyecto de nivel superior de la Apache Software Foundation.

Maven utiliza un Project Object Model (POM) para describir el proyecto de software a construir, sus dependencias de otros módulos y componentes externos, y el orden de construcción de los elementos. Viene con objetivos predefinidos para realizar ciertas tareas claramente definidas, como la compilación del código y su empaquetado.

El motor incluido en su núcleo puede dinámicamente descargar plugins de un repositorio, el mismo repositorio que provee acceso a muchas versiones de diferentes proyectos. Este repositorio pugnan por ser el mecanismo de facto de distribución de aplicaciones en Java. Una caché local de artefactos actúa como la primera fuente para sincronizar la salida de los proyectos a un sistema local.

Maven se ejecuta en unas circunstancias similares a las que explicábamos con Git. Existen dos opciones, o tener Maven instalado en el sistema [6], o utilizar un entorno de desarrollo como Eclipse [2], el cual lo trae incluido.

Si tenemos importado el proyecto en un IDE, para compilarlo (o en nomenclatura de Maven, construirlo) es necesario hacer clic derecho en el fichero incluido en el proyecto, pom.xml, y seleccionar la opción: Run as (Ejecutar como) y en el submenú, Maven build, si hacemos esto por primera vez nos aparecerá un cuadro de dialogo donde se nos solicita los objetivos (goals), ahí debemos indicar "install" (sin las comillas) y aceptar/ejecutar.

En caso de que no estemos utilizando un IDE, necesitaremos tener instalado Maven. Para compilarlo de este modo debemos abrir una consola, dirigirnos al directorio donde se encuentra el proyecto y ejecutar el siguiente comando:

```
[usu@host repo]$ mvn install
```

Al compilarlo con Maven, tanto desde el IDE como desde la consola, se generará un fichero .jar el cual contiene todas las clases compiladas.

5. TRABAJOS FUTUROS

6. CONCLUSIONES

7. ANEXO I. CÓDIGO FUENTE

BIBLIOGRAFÍA

- [1] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: Simplified data processing on large clusters. *Google, Inc.*, 2004.
- [2] Entorno de desarrollo Eclipse [online]. URL: <https://eclipse.org/>.
- [3] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. The google file system. *Google, Inc.*, 2003.
- [4] Apache Hadoop [online]. URL: <http://hadoop.apache.org/>.
- [5] Procedimientos para la instalación de Git [online]. URL: <http://symfony.es/documentacion/guia-de-instalacion-de-git/>.
- [6] Procedimientos para la instalación de Maven [online]. URL: <https://eljaviador.wordpress.com/2013/05/21/guia-de-instalacion-de-maven/>.
- [7] Artículo de la Wikipedia sobre Git [online]. URL: <http://es.wikipedia.org/wiki/Git>.
- [8] Artículo de la Wikipedia sobre Maven [online]. URL: <http://es.wikipedia.org/wiki/Maven>.