MapReduce Fest - Desafío 3
Generación de un Índice Inverso

Lorena Etcheverry
Instituto de Computación,
Facultad de Ingeniería,
lorenae@fing.edu.uy

24 de mayo de 2013

Índice

1. Descripción del Problema

Supongamos que se quiere agregar links a StackOverflow a las páginas de Wikipedia que son referenciadas desde respuestas a preguntas en StackOverflow. Para esto es necesario analizar cada respuesta en StackOverflow en busca de links a la Wikipedia. Si se encuentra un link se genera como salida la pareja (link,ID del comentario). En la fase de reducción todos los IDs de comentarios que referencian a cierto link se agrupan, pudiéndose generar un archivo que permita luego actualizar páginas de la Wikipedia.

2. Descripción de la Solución

Este problema es una aplicación directa del patrón de índice inverso, descripto en la página 35 de [1]. Para aplicar dicho patrón es necesario construir, a partir de la entrada, parejas (clave,valor) donde la clave es el término a indexar y el valor es un identificador del documento donde se encuentra el término indexado. Luego, en la fase de reducción, se agrupan todos las parejas con igual clave (término a indexar) y se construye el conjunto de identificadores de documentos en los cuales el término a indezar fue encontrado. La Figura 1 representa gráficamente este proceso. Se implementó un mapper y un reducer (que también se utilizó como combiner) siguiendo las sugerencias encontradas en[1].

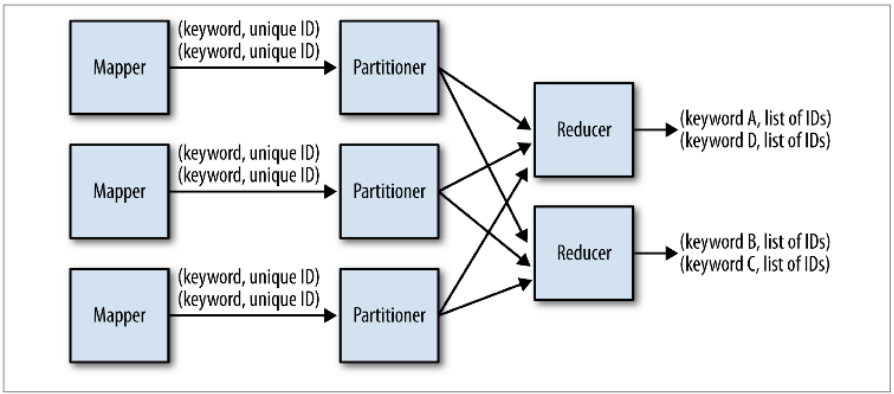

Figura 1: Patrón de diseño índice invertido[1]

2.1. Pruebas Realizadas

Se realizaron pruebas de la solución generada buscando analizar cómo varían los tiempos de ejecución y el uso de recursos a medida que el volumen de datos a procesar crece. También se evaluó los efectos que el uso de un combiner tiene sobre el desempeño de la solución. En esta sección se comienza describiendo los datos que fueron utilizados en las pruebas, para luego presentar los escenarios de prueba y los resultados obtenidos.

2.2. Preparación de los Datos de Prueba

Para probar la solución se procedió a obtener un volcado de datos de SO. Este sitio publica desde 2009 bajo la licencia Creative Commons todos los posts y comentarios que los usuarios realizan[1]. Para este desafío se seleccionó el archivo de datos correspondiente a Noviembre de 2010, el cual se obtuvo desde el sitio ClearBits[2]. De todos los archivos descargados (los cuales no sólo corresponden al sitio SO sino que también se incluyen datos de sitios hermanos como Meta StackOverflow) el archivo utilizado fue el que inlcuye a las publicaciones de los usuarios, llamado *posts.xml*. El tamaño de dicho archivo una vez descomprimido es de 4.09 GB.

Dicho archivo xml se estructura como una colección de *posts* representados por elementos del tipo `row`. La Figura 4 presenta un par de *posts* de ejemplo. El atributo `PostIdType` indica si el post corresponde a una pregunta (`PostTypeId=1`) o a una respuesta (`PostTypeId=2`).

A fin de realizar diferentes ejecuciones y evaluar el comportamiento de la solución en función del volúmen de datos a procesar, se decidió partir el archivo de entrada en archivos más pequeños. Para esto se utilizó el comando *split* indicando el tamaño deseado para cada archivo (40 MB). Dado que al restringir el tamaño del archivo podían quedar entradas mal formadas (elementos del tipo `row` truncados) se procedió a eliminar la primera y la segunda linea de cada

[1] `http://blog.stackoverflow.com/2009/06/stack-overflow-creative-commons-data-dump/`
[2] `http://www.clearbits.net/creators/146-stack-exchange-data-dump/contents`

```
<row Id="6939296" PostTypeId="2" ParentId="6939137"
CreationDate="2011-08-04T09:50:25.043" Score="4" ViewCount=""
Body="&lt;p&gt;You should have imported Poll with &lt;code&gt;
from polls.models import Poll&lt;/code&gt;&lt;/p&gt;&#xA;"
OwnerUserId="634150" LastActivityDate="2011-08-04T09:50:25.043"
CommentCount="1" />

<row Id="6939304" PostTypeId="1" AcceptedAnswerId="6939433"
CreationDate="2011-08-04T09:50:58.910" Score="1" ViewCount="26"
Body="&lt;p&gt;Is it possible to gzip a single asp.net 3.5 page? my
site is hosted on IIS7 and for technical reasons I cannot enable gzip
compression site wide. does IIS7 have an option to gzip individual pages or
will I have to override OnPreRender and write some code to compress the
output?&lt;/p&gt;&#xA;" OwnerUserId="743184"
LastActivityDate="2011-08-04T10:19:04.107" Title="gzip a single asp.net page"
Tags="&lt;asp.net&gt;&lt;iis7&gt;&lt;gzip&gt;"
AnswerCount="2" />
```

Figura 2: Ejemplo de contenido del archivo posts.xml[1]

uno de los archivos generados. La Figura 3 presenta los comandos ejecutados. Se obtuvieron de esta forma más de 90 archivos de 40 MB cada uno.

```
split -b 40m posts.xml splitted/posts_
for file in posts_*; do sed '1d;$d' $file > alt.$file.xml; done
```

Figura 3: Comandos ejecutados para preparar los datos de entrada.

2.3. Escenarios de Prueba y Resultados Obtenidos

Los archivos obtenidos luego del procesamiento descripto en la Sección 2.2 se organizaron en 10 conjuntos de archivos de 40, 400, 800, 1200, 1600, 2000, 2400, 2800, 3200 y 3600 MB respectivamente llamados DS0 a DS9. Para cada uno de estos conjuntos de archivos se ejecutó la solución desarrollada en dos modalidades: (a) sin *combiner* y (b) utilizando el *reducer* como *combiner*. Se armaron de esta manera 20 escenarios de prueba.

Se registraron los tiempos de ejecución de cada escenario y los valores de los contadores provistos por Hadoop. Se constató que el tiempo de ejecución puede ser utilizado como una primer aproximación al desempeño de la solución, pero la variabilidad del mismo es muy alta. Por ejemplo, para la ejecución sobre el DS5 con combiner se obtuvieron tiempos que oscilan entre los 8 y los 12 minutos. En la Tabla 2.3 se presentan los valores mínimos de tiempo de ejecución (expresados en minutos), para cada uno de los 20 escenarios, luego de realizar 5 ejecuciones de cada uno, mientras que la Figura presenta dichos valores en forma gráfica. Se puede apreciar que los tiempos de ejecución con combiner son ligeramente superiores a los que no lo utilizan. El análisis de los demás contadores no brinda información suficiente como para extraer una conclusión acerca de cual de las dos estrategias, en este caso, es más eficiente.

Sería deseable poder comparar esta solución (usando mapReduce) con una

	Con combiner	Sin combiner
DS0	00:43.42	00:37.30
DS1	01:40.21	01:39.03
DS2	02:50.31	02:49.37
DS3	04:43.76	04:38.91
DS4	06:55.59	06:00.26
DS5	08:01.05	07:13.31
DS6	08:41.37	08:18.12
DS7	10:13.36	09:16.02
DS8	14:03.57	12:50.61
DS9	13:56.76	12:40.92

Tabla 1: Mínimo tiempo de ejecución par cada escenario luego de 5 ejecuciones.

solución no distribuída para evaluar si los tiempos obtenidos son significativamente mejores que los que se pueden obtener con una solución no distribuída.

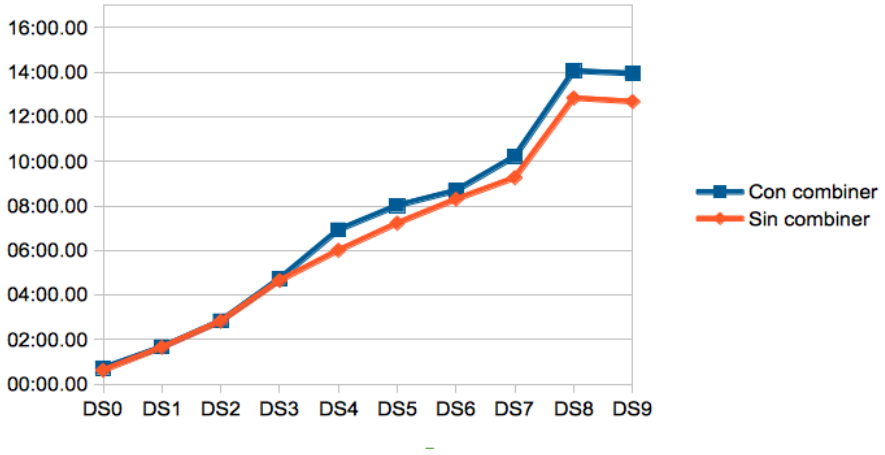

Figura 4: Mínimo tiempo de ejecución par cada escenario luego de 5 ejecuciones.

Referencias

[1] Tom White. *Hadoop: The definitive guide*. O'Reilly Media, Inc., 2012.