



Full Inverted Index using MapREduce

CS570: Big Data Processing & Analytics
Submitted By Imran Noor Saleh Student ID 19648

Introduction

The fully inverted index is a popular data structure used in information retrieval systems. It allows efficient searching and retrieval of documents based on the words they contain. This program utilizes the power of MapReduce to create an inverted index for a given set of text files.

docID		geo-scopeID
1		Europe
2		Europe
3		France
4		England
5		Portugal
6		Quebec
7		Europe
8		Spain

Forward Index

geo-scopeID		docID
Europe		1 2 7
France		3
Portugal		5
England		4
Quebec		6
Spain		8

Inverted Index

Design

Full Inverted Index

a: {(2, 2)} ==> file 2, location 2
banana: {(2, 3)} ==> file 2, location 3
is: {(0, 1), (0, 4), (1, 1), (2, 1)}
it: {(0, 0), (0, 3), (1, 2), (2, 0)}
what: {(0, 2), (1, 0)} ==> file 0, location 2; file 1, location 0

Mapper				Reducer			
Input Key	Input Value	Output Key	Output Value	Input Key	Input Key	Output Key	Output Value
file0	it is what it is	it	(0,0)	a	{{(2,2)}}	a	{{(2,2)}}
		is	(0,1)	banana	{{(2,3)}}	banana	{{(2,3)}}
		what	(0,2)	is	{{(0,1),(0,4),(1,1),(2,1)}}	is	{{(0,1),(0,4),(1,1),(2,1)}}
		it	(0,3)	it	{{(0,0),(0,3),(1,2),(2,0)}}	it	{{(0,0),(0,3),(1,2),(2,0)}}
		is	(0,4)	what	{{(0,2),(1,0)}}	what	{{(0,2),(1,0)}}
file1	what is it	what	(1,0)				
		is	(1,1)				
		it	(1,2)				
file2	it is a banana	it	(2,0)				
		is	(2,1)				
		a	(2,2)				
		banana	(2,3)				

Implementation: Input Files

```
isaleh@instance-1:~$ mkdir index
isaleh@instance-1:~$ cd index
isaleh@instance-1:~/index$ mkdir input
isaleh@instance-1:~/index$ cd input
isaleh@instance-1:~/index/input$ vi file1
isaleh@instance-1:~/index/input$ vi file2
isaleh@instance-1:~/index/input$ cd ..
isaleh@instance-1:~/index$ cd ..
isaleh@instance-1:~$ cd hadoop-3.3.4
isaleh@instance-1:~/hadoop-3.3.4$ bin/hdfs namenode -format
```

```
what is it
```

```
it is a banana
```



Implementation: MapReduce

```
isaleh@instance-1:~/hadoop-3.3.4$ bin/hdfs namenode -format
2023-08-20 11:56:34,417 INFO namenode.NameNode: STARTUP_MSG:
/*****
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   host = instance-1/10.182.0.2
STARTUP_MSG:   args = [-format]
STARTUP_MSG:   version = 3.3.4
STARTUP_MSG:   classpath = /home/isaleh/hadoop-3.3.4/etc/hadoop:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/slf4j-reload4j-1.7.36.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/json-smart-2.4.7.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/jul-to-slf4j-1.7.36.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/jackson-xml-1.9.13.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/paranamer-2.3.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/jsr305-3.0.2.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/commons-net-3.6.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/jcip-annotations-1.0-1.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/zookeeper-jute-3.5.6.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/jetty-webapp-9.4.43.v20210629.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/zookeeper-3.5.6.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/jetty-util-ajax-9.4.43.v20210629.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/kerb-identity-1.0.1.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/accessors-smart-2.4.7.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/jackson-core-asl-1.9.13.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/javax.servlet-api-3.1.0.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/commons-beanutils-1.9.4.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/asm-5.0.4.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/jackson-core-2.12.7.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/curator-client-4.2.0.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/httpclient-4.5.13.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/listenablefuture-9999.0-empty-to-avoid-conflict-with-guava.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/gson-2.8.9.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/commons-logging-1.1.3.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/jersey-servlet-1.19.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/avro-1.7.7.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/kerby-xdr-1.0.1.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/jaxb-api-2.2.11.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/commons-configuration2-2.11.1.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/jetty-security-9.4.43.v20210629.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/lib/failureaccess-1.0.jar:/home/isaleh/hadoop-3.3.4/share/hadoop/common/
```



Implementation: MapReduce

```
isaleh@instance-1:~/hadoop-3.3.4$ sbin/start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [instance-1]
isaleh@instance-1:~/hadoop-3.3.4$
```

```
isaleh@instance-1:~/hadoop-3.3.4$ wget http://localhost:9870/
--2023-08-20 12:02:46-- http://localhost:9870/
Resolving localhost (localhost)... ::1, 127.0.0.1
Connecting to localhost (localhost)|::1|:9870... failed: Connection refused.
Connecting to localhost (localhost)|127.0.0.1|:9870... connected.
HTTP request sent, awaiting response... 302 Found
Location: http://localhost:9870/index.html [following]
--2023-08-20 12:02:46-- http://localhost:9870/index.html
Reusing existing connection to localhost:9870.
HTTP request sent, awaiting response... 200 OK
Length: 1079 (1.1K) [text/html]
Saving to: 'index.html.3'

index.html.3          100%[=====>]  1.05K  --.-KB/s    in 0s

2023-08-20 12:02:46 (122 MB/s) - 'index.html.3' saved [1079/1079]
```

Implementation: InvertedIndex.java

```
import java.io.IOException;
import java.util.ArrayList;
import java.util.List;
import java.util.StringTokenizer;

import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.conf.Configured;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Counter;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.input.FileSplit;
import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import org.apache.hadoop.mapreduce.lib.output.TextOutputFormat;
import org.apache.hadoop.util.Tool;
import org.apache.hadoop.util.ToolRunner;

public class InvertedIndex extends Configured implements Tool {

    public static class InvertedIndexMapper extends Mapper<Object, Text, Text, Text> {

        private Text word = new Text();
        private Text location = new Text();

        public void map(Object key, Text value, Context context) throws IOException, InterruptedException {
            FileSplit fileSplit = (FileSplit) context.getInputSplit();
            String filename = fileSplit.getPath().getName();
            String line = value.toString();

            StringTokenizer tokenizer = new StringTokenizer(line);
            int index = 0;
            while (tokenizer.hasMoreTokens()) {
                word.set(tokenizer.nextToken().toLowerCase());
                int fileNumber = Integer.parseInt(filename.replaceAll("[0-9]", ""));
                String toReduce = "(" + fileNumber + "," + index + ")";
                index++;
            }
        }
    }

    1,1    Top
```

```
        context.write(word, new Text(toReduce));
    }
}

public static class InvertedIndexReducer extends Reducer<Text, Text, Text, Text> {

    private Text result = new Text();

    public void reduce(Text key, Iterable<Text> values, Context context) throws IOException, InterruptedException {
        List<String> locations = new ArrayList<>();

        for (Text value : values) {
            locations.add(value.toString());
        }

        result.set(locations.toString());
        context.write(key, result);
    }

    public int run(String[] args) throws Exception {
        Configuration conf = getConf();
        Job job = Job.getInstance(conf, "InvertedIndex");
        job.setJarByClass(InvertedIndex.class);

        job.setInputFormatClass(TextInputFormat.class);
        job.setOutputFormatClass(TextOutputFormat.class);

        job.setMapperClass(InvertedIndexMapper.class);
        job.setReducerClass(InvertedIndexReducer.class);

        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(Text.class);

        FileInputFormat.addInputPath(job, new Path(args[0]));
        FileOutputFormat.setOutputPath(job, new Path(args[1]));

        return job.waitForCompletion(true) ? 0 : 1;
    }

    public static void main(String[] args) throws Exception {

    77,1    918
```



Implementation: Copying File

```
isaleh@instance-1:~/hadoop-3.3.4$ bin/hadoop com.sun.tools.javac.Main ../index/InvertedIndex.java
isaleh@instance-1:~/hadoop-3.3.4$ cp ../index/*.class .
cp: cannot stat '../index/*.class': No such file or directory
isaleh@instance-1:~/hadoop-3.3.4$ cp ../index/*.class .
isaleh@instance-1:~/hadoop-3.3.4$ cp ../index/*.java .
isaleh@instance-1:~/hadoop-3.3.4$ jar cf in.jar InvertedIndex*.class
```




Implementation

```
isaleh@instance-1:~/hadoop-3.3.4$ bin/hadoop jar in.jar InvertedIndex /user/isaleh/index/input
/user/isaleh/index/output
2023-08-20 12:21:06,568 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2023-08-20 12:21:06,741 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 seconds
2023-08-20 12:21:06,741 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2023-08-20 12:21:07,119 INFO input.FileInputFormat: Total input files to process : 2
2023-08-20 12:21:07,147 INFO mapreduce.JobSubmitter: number of splits:2
2023-08-20 12:21:07,487 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local748548891_0001
2023-08-20 12:21:07,487 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-08-20 12:21:07,710 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2023-08-20 12:21:07,711 INFO mapreduce.Job: Running job: job_local748548891_0001
2023-08-20 12:21:07,718 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2023-08-20 12:21:07,725 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2023-08-20 12:21:07,725 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup temporary folders under output directory:false, ignore cleanup failures: false
2023-08-20 12:21:07,726 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
2023-08-20 12:21:07,822 INFO mapred.LocalJobRunner: Waiting for map tasks
2023-08-20 12:21:07,827 INFO mapred.LocalJobRunner: Starting task: attempt_local748548891_0001_m_000000_0
2023-08-20 12:21:07,860 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2023-08-20 12:21:07,863 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup temporary folders under output directory:false, ignore cleanup failures: false
2023-08-20 12:21:07,903 INFO mapred.Task: Using ResourceCalculatorProcessFree: []
2023-08-20 12:21:07,909 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/isaleh/index/input/file2:0x15
2023-08-20 12:21:07,979 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2023-08-20 12:21:07,979 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2023-08-20 12:21:07,979 INFO mapred.MapTask: soft limit at 83886080
2023-08-20 12:21:07,979 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2023-08-20 12:21:07,979 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2023-08-20 12:21:07,984 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
2023-08-20 12:21:08,232 INFO mapred.LocalJobRunner:
2023-08-20 12:21:08,239 INFO mapred.MapTask: Starting flush of map output
2023-08-20 12:21:08,240 INFO mapred.MapTask: Spilling map output
2023-08-20 12:21:08,240 INFO mapred.MapTask: bufstart = 0; bufend = 39; bufvoid = 104857600
```



Output

```
isaleh@instance-1:~/hadoop-3.3.4$ bin/hdfs dfs -ls /user/isaleh/index/output
Found 2 items
-rw-r--r--    1 isaleh supergroup          0 2023-08-20 12:21 /user/isaleh/index/output/_SUCCESS
-rw-r--r--    1 isaleh supergroup       74 2023-08-20 12:21 /user/isaleh/index/output/part-r-00000
isaleh@instance-1:~/hadoop-3.3.4$ bin/hdfs dfs -cat /user/isaleh/index/output/part-r-0000
cat: `/user/isaleh/index/output/part-r-0000': No such file or directory
isaleh@instance-1:~/hadoop-3.3.4$ bin/hdfs dfs -cat /user/isaleh/index/output/part-r-00000
a      [(2,2)]
banana [(2,3)]
is     [(2,1), (1,1)]
it     [(1,2), (2,0)]
what   [(1,0)]
isaleh@instance-1:~/hadoop-3.3.4$
```