

# MapReduce Program + Full Inverted Index

Niyat Habtom Seghid -19967  
06/05/2025

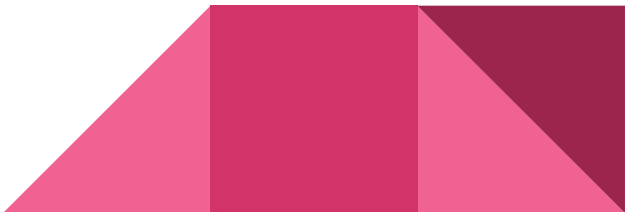
Github Link:

# Contents:

1. Introduction
2. Implementation
3. Test
4. Conclusion
5. Reference

---

# Introduction:

- This project focuses on implementing a **MapReduce program** to create a Full Inverted Index. The task involves several steps, starting with visualizing the operations of the mapper, combiner, and reducer using tables to process the contents of three text files.
  - We will then convert a WordCount MapReduce program into a **Partial Inverted Index** program, followed by modifying it to achieve a Full Inverted Index.
  - Finally, we will compile the program into a JAR file and execute it in a Hadoop environment to generate the **Full Inverted Index**, demonstrating the practical application of the MapReduce framework in text processing.
- 

# Implementation:

- Step 1: Please draw three tables to show the processes done by mapper, combiner, and reducer to show the Full Inverted Index of these three files:
  - file 0's content "it is what it is"
  - file 1's content "what is it"
  - file 2's content "it is a banana"

Mapper				Reducer			
Input Key	Input Value	Output Key	Output Value	Input Key	Input Key	Output Key	Output Value
0	it is what it is	it		0	a	{2}	{2}
		is		0	banana	{2}	{2}
		what		0	is	{0,0,1,2}	{0,1,2}
		it		0	it	{0,0,1,2}	{0,1,2}
		is		0	what	{0,1}	{0,1}
1	what is it	what		1			
		is		1			
		it		1			
2	it is a banana	it		2			
		is		2			
		a		2			
		banana		2			

Partial Inverted Index

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)				

Fully Inverted Index

# Implementation:

- Created 3 files – file0, file1, file2 with given lines in each file in InvertedIndex directory.

```
nseghid8444@big-data-week3-hw1:~$ mkdir InvertedIndex
nseghid8444@big-data-week3-hw1:~$ cd InvertedIndex
nseghid8444@big-data-week3-hw1:~/InvertedIndex$ vi file0.txt
nseghid8444@big-data-week3-hw1:~/InvertedIndex$ vi file1.txt
nseghid8444@big-data-week3-hw1:~/InvertedIndex$ vi file2.txt
nseghid8444@big-data-week3-hw1:~/InvertedIndex$
```



SSH-in-browser

```
it is a banana
```

```
~
~
~
~
~
```



SSH-in-browser

```
what is it
```

```
~
~
~
~
~
```



SSH-in-browser

```
it is what it is
```

```
~
~
~
~
~
```

# Implementation:

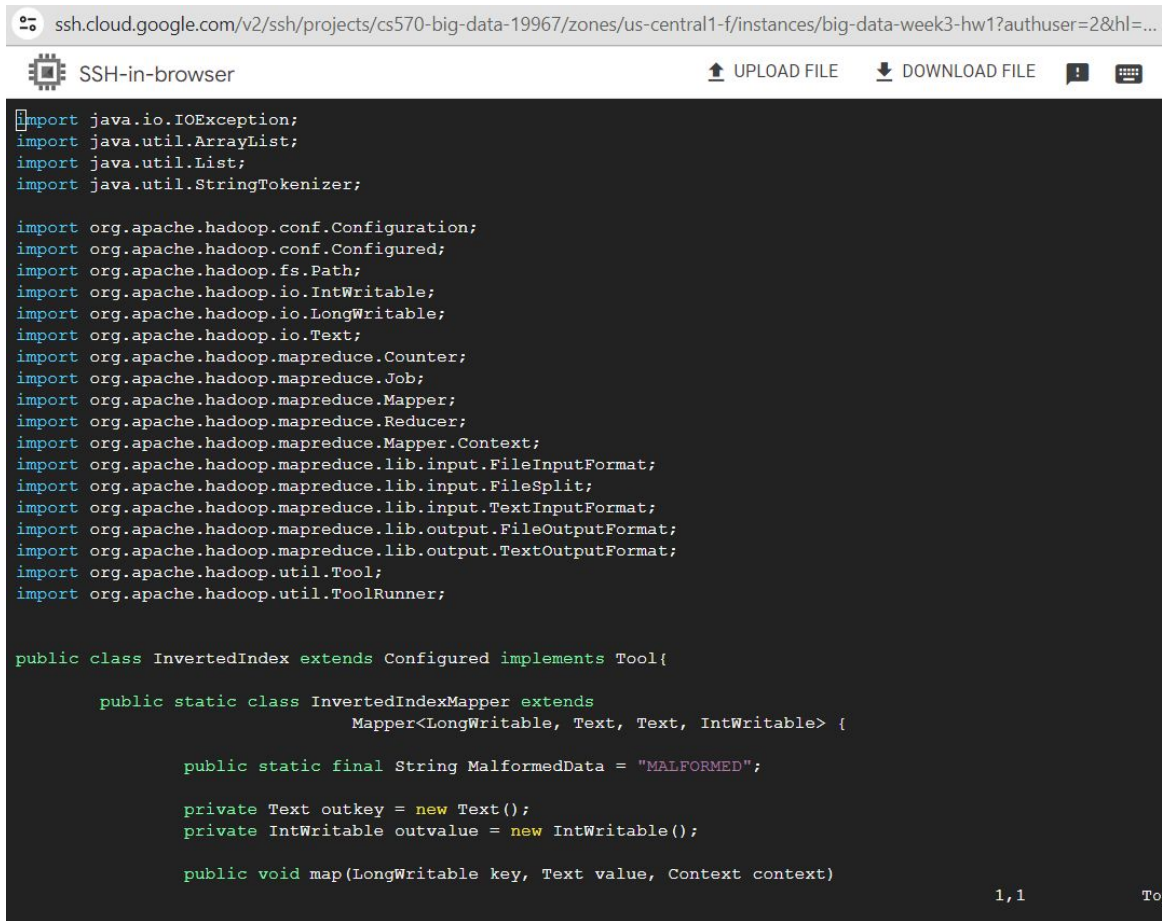
- Create HDFS directories and and copy the three files to HDFS to be used as input.

```
nseghid8444@big-data-week3-hw1:~/InvertedIndex$ ls
file0.txt  file1.txt  file2.txt
nseghid8444@big-data-week3-hw1:~/InvertedIndex$
```

```
nseghid8444@big-data-week3-hw1:~/hadoop-3.4.0$ bin/hdfs dfs -mkdir /user/nseghid8444/invertedIndex
nseghid8444@big-data-week3-hw1:~/hadoop-3.4.0$ bin/hdfs dfs -mkdir /user/nseghid8444/invertedIndex/input
nseghid8444@big-data-week3-hw1:~/hadoop-3.4.0$ bin/hadoop fs -put ../InvertedIndex/* /user/nseghid8444/invertedIndex/input
nseghid8444@big-data-week3-hw1:~/hadoop-3.4.0$ vi InvertedIndex.java
nseghid8444@big-data-week3-hw1:~/hadoop-3.4.0$ vi Tester.java
nseghid8444@big-data-week3-hw1:~/hadoop-3.4.0$ bin/hadoop com.sun.tools.javac.Main InvertedIndex.java
Note: InvertedIndex.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
nseghid8444@big-data-week3-hw1:~/hadoop-3.4.0$ ^C
```

# Implementation:

- Step 2: Convert a WordCount MapReduce program into a Partial Inverted Index MapReduce program with the three input files and expected output.
- ❑ This is InvertedIndex.java file, which is a mapreduce program for partial inverted Index.



The screenshot shows a web browser window with the address bar displaying a Google Cloud SSH session URL. The browser tab is titled "SSH-in-browser". The main content area shows a code editor with Java code for a MapReduce program. The code includes imports for various Java and Hadoop classes, followed by the definition of the `InvertedIndex` class and its `map` method.

```
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.LongWritable;
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.Mapper.Context;
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<LongWritable, Text, Text, IntWritable> {

        public static final String MalformedData = "MALFORMED";

        private Text outkey = new Text();
        private IntWritable outvalue = new IntWritable();

        public void map(LongWritable key, Text value, Context context)
```

1,1 To

# Implementation:

Compile InvertedIndex.java, create a jar, and Run the program on GCP.

```
nseghid8444@big-data-week3-hw1:~/hadoop-3.4.0$ vi InvertedIndex.java
nseghid8444@big-data-week3-hw1:~/hadoop-3.4.0$ vi Tester.java
nseghid8444@big-data-week3-hw1:~/hadoop-3.4.0$ bin/hadoop com.sun.tools.javac.Main InvertedIndex.java
Note: InvertedIndex.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
nseghid8444@big-data-week3-hw1:~/hadoop-3.4.0$ ^C
nseghid8444@big-data-week3-hw1:~/hadoop-3.4.0$ bin/hadoop com.sun.tools.javac.Main InvertedIndex.java
Note: InvertedIndex.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
nseghid8444@big-data-week3-hw1:~/hadoop-3.4.0$ jar cf invertedindex.jar InvertedIndex*.class
nseghid8444@big-data-week3-hw1:~/hadoop-3.4.0$ bin/hadoop jar invertedindex.jar InvertedIndex /user/nseghid8444/invertedIndex/input /user/nseghid8444/invertedIndex/output
2024-06-05 21:45:53,373 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2024-06-05 21:45:53,625 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2024-06-05 21:45:53,627 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2024-06-05 21:45:54,128 INFO input.FileInputFormat: Total input files to process : 3
2024-06-05 21:45:54,180 INFO mapreduce.JobSubmitter: number of splits:3
2024-06-05 21:45:54,403 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local940059751_0001
2024-06-05 21:45:54,403 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-06-05 21:45:54,693 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2024-06-05 21:45:54,696 INFO mapreduce.Job: Running job: job_local940059751_0001
2024-06-05 21:45:54,705 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2024-06-05 21:45:54,721 INFO output.PathOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommitterFactory
2024-06-05 21:45:54,723 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2024-06-05 21:45:54,723 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output
directory:false, ignore cleanup failures: false
2024-06-05 21:45:54,725 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
2024-06-05 21:45:54,814 INFO mapred.LocalJobRunner: Waiting for map tasks
2024-06-05 21:45:54,816 INFO mapred.LocalJobRunner: Starting task: attempt_local940059751_0001_m_000000_0
2024-06-05 21:45:54,867 INFO output.PathOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommitterFactory
2024-06-05 21:45:54,867 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2024-06-05 21:45:54,867 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output
directory:false, ignore cleanup failures: false
```



# Implementation: Format the File System

```
nseghid8444@big-data-week2-hw1:~/hadoop-3.4.0$ bin/hdfs namenode -format
WARNING: /home/nseghid8444/hadoop-3.4.0/logs does not exist. Creating.
2024-05-30 02:59:51,980 INFO namenode.NameNode: STARTUP_MSG:
/*****
STARTUP_MSG: Starting NameNode
STARTUP_MSG: host = big-data-week2-hw1.us-central1-f.c.cs570-big-data-19967.internal/10.128.0.2
STARTUP_MSG: args = [-format]
STARTUP_MSG: version = 3.4.0
STARTUP_MSG: classpath = /home/nseghid8444/hadoop-3.4.0/etc/hadoop:/home/nseghid8444/hadoop-3.4.0/share/h
adoop/common/lib/kerb-client-2.0.3.jar:/home/nseghid8444/hadoop-3.4.0/share/hadoop/common/lib/curator-clien
t-5.2.0.jar:/home/nseghid8444/hadoop-3.4.0/share/hadoop/common/lib/netty-resolver-dns-native-macos-4.1.100.
Final-osx-aarch_64.jar:/home/nseghid8444/hadoop-3.4.0/share/hadoop/common/lib/netty-codec-http-4.1.100.Fina
l.jar:/home/nseghid8444/hadoop-3.4.0/share/hadoop/common/lib/netty-codec-4.1.100.Final.jar:/home/nseghid844
4/hadoop-3.4.0/share/hadoop/common/lib/kerb-util-2.0.3.jar:/home/nseghid8444/hadoop-3.4.0/share/hadoop/comm
on/lib/jetty-security-9.4.53.v20231009.jar:/home/nseghid8444/hadoop-3.4.0/share/hadoop/common/lib/jakarta.a
ctivation-api-1.2.1.jar:/home/nseghid8444/hadoop-3.4.0/share/hadoop/common/lib/netty-codec-stomp-4.1.100.Fi
nal.jar:/home/nseghid8444/hadoop-3.4.0/share/hadoop/common/lib/jetty-server-9.4.53.v20231009.jar:/home/nse
ghid8444/hadoop-3.4.0/share/hadoop/common/lib/kerby-xdr-2.0.3.jar:/home/nseghid8444/hadoop-3.4.0/share/hadoo
p/common/lib/netty-transport-native-unix-common-4.1.100.Final.jar:/home/nseghid8444/hadoop-3.4.0/share/hado
op/common/lib/netty-transport-classes-epoll-4.1.100.Final.jar:/home/nseghid8444/hadoop-3.4.0/share/hadoop/c
ommon/lib/netty-handler-proxy-4.1.100.Final.jar:/home/nseghid8444/hadoop-3.4.0/share/hadoop/common/lib/zook
eeper-jute-3.8.3.jar:/home/nseghid8444/hadoop-3.4.0/share/hadoop/common/lib/commons-text-1.10.0.jar:/home/n
seghid8444/hadoop-3.4.0/share/hadoop/common/lib/jaxb-api-2.2.11.jar:/home/nseghid8444/hadoop-3.4.0/share/ha
dooop/common/lib/netty-handler-ssl-ocsp-4.1.100.Final.jar:/home/nseghid8444/hadoop-3.4.0/share/hadoop/common
/lib/netty-codec-socks-4.1.100.Final.jar:/home/nseghid8444/hadoop-3.4.0/share/hadoop/common/lib/jetty-http-
9.4.53.v20231009.jar:/home/nseghid8444/hadoop-3.4.0/share/hadoop/common/lib/netty-resolver-4.1.100.Final.ja
r:/home/nseghid8444/hadoop-3.4.0/share/hadoop/common/lib/hadoop-shaded-guava-1.2.0.jar:/home/nseghid8444/ha
dooop-3.4.0/share/hadoop/common/lib/netty-transport-native-epoll-4.1.100.Final.jar:/home/nseghid8444/hadoop-
3.4.0/share/hadoop/common/lib/jetty-util-9.4.53.v20231009.jar:/home/nseghid8444/hadoop-3.4.0/share/hadoop/c
ommon/lib/jackson-annotations-2.12.7.jar:/home/nseghid8444/hadoop-3.4.0/share/hadoop/common/lib/nimbus-jose
-jwt-9.3.1.jar:/home/nseghid8444/hadoop-3.4.0/share/hadoop/common/lib/kerby-pkix-2.0.3.jar:/home/nseghid8444
```

# Implementation: Start NameNode daemon and DataNode daemon, and Check localhost connection

```
nseghid8444@big-data-week3-hw1:~/hadoop-3.4.0$ sbin/start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [big-data-week3-hw1]
nseghid8444@big-data-week3-hw1:~/hadoop-3.4.0$ wget http://localhost:9870/
--2024-06-05 18:26:04-- http://localhost:9870/
Resolving localhost (localhost)... 127.0.0.1
Connecting to localhost (localhost)|127.0.0.1|:9870... connected.
HTTP request sent, awaiting response... 302 Found
Location: http://localhost:9870/index.html [following]
--2024-06-05 18:26:04-- 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'

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

2024-06-05 18:26:04 (97.3 MB/s) - 'index.html' saved [1079/1079]

nseghid8444@big-data-week3-hw1:~/hadoop-3.4.0$ □
```

# Output:

## Step 4: Partial Inverted Index

**Note:** we can observe that the output shows the inverted index for each word, indicating the file names where the word appears.

```
Map output bytes=91
Map output materialized bytes=133
Input split bytes=399
Combine input records=0
Combine output records=0
Reduce input groups=5
Reduce shuffle bytes=133
Reduce input records=12
Reduce output records=5
Spilled Records=24
Shuffled Maps =3
Failed Shuffles=0
Merged Map outputs=3
GC time elapsed (ms)=46
Total committed heap usage (bytes)=1179123712

Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0

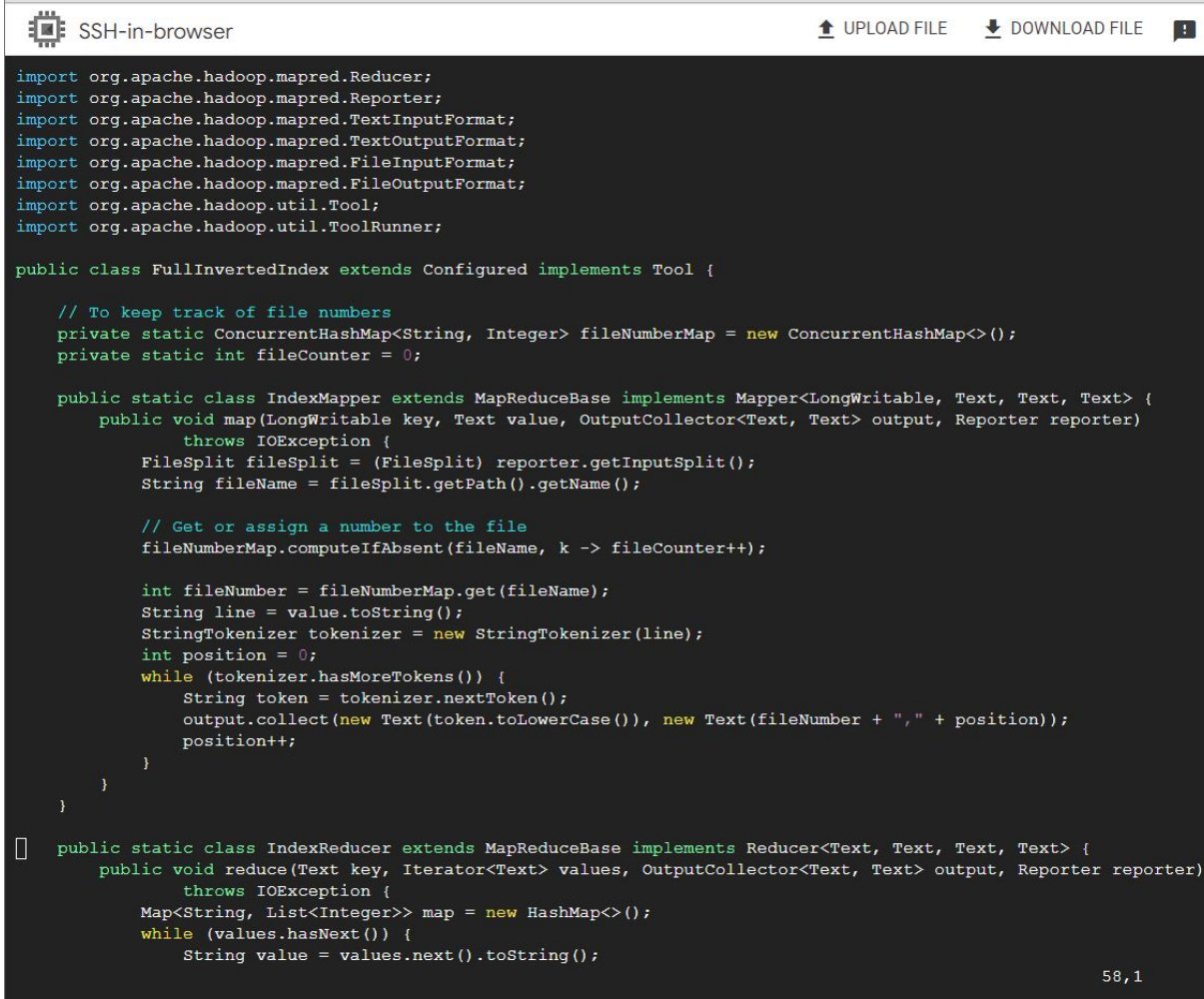
File Input Format Counters
Bytes Read=43
File Output Format Counters
Bytes Written=55

nseghid8444@big-data-week3-hw1:~/hadoop-3.4.0$ bin/hdfs dfs -ls /user/nseghid8444/invertedIndex/output
Found 2 items
-rw-r--r--  1 nseghid8444 supergroup          0 2024-06-05 21:45 /user/nseghid8444/invertedIndex/output/_SUCCESS
-rw-r--r--  1 nseghid8444 supergroup        55 2024-06-05 21:45 /user/nseghid8444/invertedIndex/output/part-r-00000
nseghid8444@big-data-week3-hw1:~/hadoop-3.4.0$ bin/hdfs dfs -cat /user/nseghid8444/invertedIndex/output/part-r-00000
a      [2]
banana [2]
is     [2, 1, 0]
it     [0, 2, 1]
what   [1, 0]
nseghid8444@big-data-week3-hw1:~/hadoop-3.4.0$
```

# Implementation:

- Step 3: Convert a [Partial Inverted Index](#) MapReduce program into a [Full Inverted Index](#) MapReduce program with the three input files and expected output.

□ This is FullInvertedIndex.java file, which is a mapreduce program for fully inverted Index.



```
import org.apache.hadoop.mapred.Reducer;
import org.apache.hadoop.mapred.Reporter;
import org.apache.hadoop.mapred.TextInputFormat;
import org.apache.hadoop.mapred.TextOutputFormat;
import org.apache.hadoop.mapred.FileInputFormat;
import org.apache.hadoop.mapred.FileOutputFormat;
import org.apache.hadoop.util.Tool;
import org.apache.hadoop.util.ToolRunner;

public class FullInvertedIndex extends Configured implements Tool {

    // To keep track of file numbers
    private static ConcurrentHashMap<String, Integer> fileNumberMap = new ConcurrentHashMap<>();
    private static int fileCounter = 0;

    public static class IndexMapper extends MapReduceBase implements Mapper<LongWritable, Text, Text, Text> {
        public void map(LongWritable key, Text value, OutputCollector<Text, Text> output, Reporter reporter)
            throws IOException {
            FileSplit fileSplit = (FileSplit) reporter.getInputSplit();
            String fileName = fileSplit.getPath().getName();

            // Get or assign a number to the file
            fileNumberMap.computeIfAbsent(fileName, k -> fileCounter++);

            int fileNumber = fileNumberMap.get(fileName);
            String line = value.toString();
            StringTokenizer tokenizer = new StringTokenizer(line);
            int position = 0;
            while (tokenizer.hasMoreTokens()) {
                String token = tokenizer.nextToken();
                output.collect(new Text(token.toLowerCase()), new Text(fileNumber + "," + position));
                position++;
            }
        }
    }

    public static class IndexReducer extends MapReduceBase implements Reducer<Text, Text, Text, Text> {
        public void reduce(Text key, Iterator<Text> values, OutputCollector<Text, Text> output, Reporter reporter)
            throws IOException {
            Map<String, List<Integer>> map = new HashMap<>();
            while (values.hasNext()) {
                String value = values.next().toString();
            }
        }
    }
}
```



# Implementation:

Compile FullInvertedIndex.java, create a jar, and Run the program on GCP.

```
nseghid8444@big-data-week3-hw1:~/hadoop-3.4.0$ vi FullInvertedIndex.java
nseghid8444@big-data-week3-hw1:~/hadoop-3.4.0$ bin/hadoop com.sun.tools.javac.Main FullInvertedIndex.java
nseghid8444@big-data-week3-hw1:~/hadoop-3.4.0$ jar cf fullinvertedindex.jar FullInvertedIndex*.class
nseghid8444@big-data-week3-hw1:~/hadoop-3.4.0$ bin/hadoop jar fullinvertedindex.jar FullInvertedIndex /user/nseghid8444/in
vertedIndex/input /user/nseghid8444/invertedIndex/output5
2024-06-05 23:49:07,141 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2024-06-05 23:49:07,316 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2024-06-05 23:49:07,316 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2024-06-05 23:49:07,337 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2024-06-05 23:49:07,760 INFO mapred.FileInputFormat: Total input files to process : 3
2024-06-05 23:49:07,792 INFO mapreduce.JobSubmitter: number of splits:3
2024-06-05 23:49:08,012 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1351526280_0001
2024-06-05 23:49:08,013 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-06-05 23:49:08,243 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2024-06-05 23:49:08,245 INFO mapreduce.Job: Running job: job_local1351526280_0001
2024-06-05 23:49:08,251 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2024-06-05 23:49:08,256 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
2024-06-05 23:49:08,262 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2024-06-05 23:49:08,262 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output
directory:false, ignore cleanup failures: false
2024-06-05 23:49:08,335 INFO mapred.LocalJobRunner: Waiting for map tasks
2024-06-05 23:49:08,340 INFO mapred.LocalJobRunner: Starting task: attempt_local1351526280_0001_m_000000_0
2024-06-05 23:49:08,382 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2024-06-05 23:49:08,383 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output
directory:false, ignore cleanup failures: false
2024-06-05 23:49:08,437 INFO mapred.Task: Using ResourceCalculatorProcessTree : [ ]
2024-06-05 23:49:08,447 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/nseghid8444/invertedIndex/input/
file0.txt:0+17
2024-06-05 23:49:08,486 INFO mapred.MapTask: numReduceTasks: 1
2024-06-05 23:49:08,515 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2024-06-05 23:49:08,515 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2024-06-05 23:49:08,515 INFO mapred.MapTask: soft limit at 83886080
2024-06-05 23:49:08,515 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2024-06-05 23:49:08,515 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
```

# Output:

## Step 4: Fully Inverted Index

**Note:** we can observe that the output shows the inverted index for each word, indicating the file names and line numbers where the word appears.

```
Map output materialized bytes=229
Input split bytes=360
Combine input records=0
Combine output records=0
Reduce input groups=5
Reduce shuffle bytes=229
Reduce input records=12
Reduce output records=5
Spilled Records=24
Shuffled Maps =3
Failed Shuffles=0
Merged Map outputs=3
GC time elapsed (ms)=23
Total committed heap usage (bytes)=1510473728

Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0

File Input Format Counters
Bytes Read=43

File Output Format Counters
Bytes Written=179

nseghid8444@big-data-week3-hw1:~/hadoop-3.4.0$ bin/hdfs dfs -ls /user/nseghid8444/invertedIndex/output5
Found 2 items
-rw-r--r-- 1 nseghid8444 supergroup 0 2024-06-05 23:49 /user/nseghid8444/invertedIndex/output5/_SUCCESS
-rw-r--r-- 1 nseghid8444 supergroup 179 2024-06-05 23:49 /user/nseghid8444/invertedIndex/output5/part-00000
nseghid8444@big-data-week3-hw1:~/hadoop-3.4.0$ bin/hdfs dfs -cat /user/nseghid8444/invertedIndex/output5/part-00000
a      [{file2.txt,2}]
banana [{file2.txt,3}]
is     [{file0.txt,4,1}, {file1.txt,1}, {file2.txt,1}]
it     [{file0.txt,3,0}, {file1.txt,2}, {file2.txt,0}]
what   [{file0.txt,2}, {file1.txt,0}]
nseghid8444@big-data-week3-hw1:~/hadoop-3.4.0$
```

# Output:

## Step 4: Fully Inverted Index

```
Bytes Written=164
nseghid8444@big-data-week3-hw1:~/hadoop-3.4.0$ bin/hdfs dfs -ls /user/nseghid8444/invertedIndex/output3
Found 2 items
-rw-r--r--  1 nseghid8444 supergroup          0 2024-06-05 23:14 /user/nseghid8444/invertedIndex/output3/_SUCCESS
-rw-r--r--  1 nseghid8444 supergroup        164 2024-06-05 23:14 /user/nseghid8444/invertedIndex/output3/part-00000
nseghid8444@big-data-week3-hw1:~/hadoop-3.4.0$ bin/hdfs dfs -ls /user/nseghid8444/invertedIndex/output3
Found 2 items
-rw-r--r--  1 nseghid8444 supergroup          0 2024-06-05 23:14 /user/nseghid8444/invertedIndex/output3/_SUCCESS
-rw-r--r--  1 nseghid8444 supergroup        164 2024-06-05 23:14 /user/nseghid8444/invertedIndex/output3/part-00000
nseghid8444@big-data-week3-hw1:~/hadoop-3.4.0$ bin/hdfs dfs -cat /user/nseghid8444/invertedIndex/output3/part-00000
a      [file2.txt,2]
banana [file2.txt,3]
is     [file0.txt,4,1] [file1.txt,1] [file2.txt,1]
it     [file0.txt,3,0] [file1.txt,2] [file2.txt,0]
what   [file0.txt,2] [file1.txt,0]
nseghid8444@big-data-week3-hw1:~/hadoop-3.4.0$
```

# References:

Chang, H. (2022, 10 09). *MapReduce* Inverted Index.

[https://hc.labnet.sfbu.edu/~henry/npu/classes/mapreduce/inverted\\_index/slide/overview.html](https://hc.labnet.sfbu.edu/~henry/npu/classes/mapreduce/inverted_index/slide/overview.html)

