Full Inverted Index Using MapReduce

By Shruti Kavishwar
San Francisco Bay University

GuidedBy: Prof. Henry Chang

Agenda

- Introduction
- Problem Statement
- Manually Solved Solution
- Procedure
- Result

Introduction

An inverted index is a data structure that maps terms (words) to the documents they appear in. It allows for efficient retrieval of documents containing specific terms. In the context of MapReduce, constructing an inverted index involves the following steps:

1. **Map Phase:**
    - The map function parses each line in an input file and emits a sequence of <word, line number> pairs.
    - For each word encountered in a line, the map function emits the word along with the line number where it appears.
2. **Reduce Phase:**
    - The reduce function accepts all pairs for a given word.
    - It sorts the corresponding line numbers and emits a <word, list(line numbers)> pair.

Problem Statement

- Create a Full Inverted Index MapReduce Program using the input files below

  file 0's content **"it is what it is"**

  file 1's content **"what is it"**

  file 2's content **"it is a banana”**

- Use a VM to create Jar file and then copy the files to the hadoop environment.

Manually Solution

Mapper								Reducer			
Map()				*Combine()*				*Reduce()*			
Input		**Output**		**Input**		**Output**		**Input**		**Output**	
Key	*Value*	*Key*	*Value*	*Key*	*Value*	*Key*	*Value*	*Key*	*Value*	*Key*	*Value*
file 0	it is what it is	it	(0,0)	is	{(0,1),(0,4)}	is	{(0,1),(0,4)}	a	{(2,2)}	a	{(2,2)}
		is	(0,1)	it	((0,0),(0,3)}	it	((0,0),(0,3)}	banana	{(2,3)}	banana	{(2,3)}
		what	(0,2)	what	({0,2)}	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	{(1,0)}	what	{(1,0)}
file 1	what is it	what	(1,0)	is	{(1,1)}	is	{(1,1)}				
		is	(1,1)	it	{(1,2)}	it	{(1,2)}				
		it	(1,2)	what	{(1,0)}	what	{(1,0)}				
file 2	it is a banana	it	(2,0)	a	{(2,2)}	a	{(2,2)}				
		is	(2,1)	banana	{(2,3)}	banana	{(2,3)}				
		a	(2,2)	is	{(2,1)}	is	{(2,1)}				
		banana	(2,3)	it	{(2,0)}	it	{(2,0)}				

Procedure

- Create input files with Data

```
skavishw276@mapreduce-week2-hw1:~/FullInvertedIndex$ ls
DocSumWritable.java  FullInvertedIndex.java  Input
skavishw276@mapreduce-week2-hw1:~/FullInvertedIndex$ cd Input/
skavishw276@mapreduce-week2-hw1:~/FullInvertedIndex/Input$ ls
file0  file1  file2
skavishw276@mapreduce-week2-hw1:~/FullInvertedIndex/Input$ cat file0
it is what it is
skavishw276@mapreduce-week2-hw1:~/FullInvertedIndex/Input$ cat file1
what is it
skavishw276@mapreduce-week2-hw1:~/FullInvertedIndex/Input$ cat file2
it is a banana
skavishw276@mapreduce-week2-hw1:~/FullInvertedIndex/Input$
```

Procedure

- Create the following directories

```
skavishw276@mapreduce-week2-hw1:~/hadoop-3.4.0$ bin/hdfs dfs -mkdir /user
skavishw276@mapreduce-week2-hw1:~/hadoop-3.4.0$ bin/hdfs dfs -mkdir /user/skavishw276
skavishw276@mapreduce-week2-hw1:~/hadoop-3.4.0$ bin/hdfs dfs -mkdir /user/skavishw276/FullInvertedIndex
skavishw276@mapreduce-week2-hw1:~/hadoop-3.4.0$ bin/hdfs dfs -mkdir /user/skavishw276/FullInvertedIndex/input
skavishw276@mapreduce-week2-hw1:~/hadoop-3.4.0$ bin/hdfs dfs -put ../FullInvertedIndex/Input/* /user/skavishw276/FullInvertedIndex/input
```

- Create the Java Code file for Full Inverted Index along with MapReduce functions.
- Compile the code and create the jar file

```
        at org.apache.hadoop.util.RunJar.main(RunJar.java:245)
skavishw276@mapreduce-week2-hw1:~/hadoop-3.4.0$ vi FullInvertedIndex.java
skavishw276@mapreduce-week2-hw1:~/hadoop-3.4.0$ vi FullInvertedIndex.java
skavishw276@mapreduce-week2-hw1:~/hadoop-3.4.0$ bin/hadoop com.sun.tools.javac.Main ./FullInvertedIndex.java
skavishw276@mapreduce-week2-hw1:~/hadoop-3.4.0$ jar cf fii.jar FullInvertedIndex*.class
skavishw276@mapreduce-week2-hw1:~/hadoop-3.4.0$ ls
 CalculatePi.class                  WordCount.class
 CalculatePi.java                   WordCount.java
'CalculatePiMR$Map.class'           bin
'CalculatePiMR$Reduce.class'        etc
 CalculatePiMR.class                fii.jar
```

```
skavishw276@mapreduce-week2-hw1:~/hadoop-3.4.0$ bin/hadoop jar fii.jar FullInvertedIndex /user/skavishw276/FullInvertedIndex/input /user/skavishw276/FullInverte
dIndex/Output_i
2024-06-06 04:53:01,145 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
```

Result

- Run the application and match the result with the manually solved solution.

```
skavishw276@mapreduce-week2-hw1:~/hadoop-3.4.0$ bin/hdfs dfs -get FullInvertedIndex/Output_i
skavishw276@mapreduce-week2-hw1:~/hadoop-3.4.0$ cat Output_i/*
a       file2:2
banana  file2:3
is      file2:1 file0:4,1 file1:1
it      file2:0 file0:3,0 file1:2
what    file0:2 file1:0
```