Step 1: First lets connect to the vm instance using ssh. Incase the publickey permission is denied just make sure to connect using the way mentioned below:

$ ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa

$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys

$ chmod 0600 ~/.ssh/authorized_keys

```
adagniew407@instance-20240603-073349:~$ ssh localhost
adagniew407@localhost: Permission denied (publickey).
adagniew407@instance-20240603-073349:~$ ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa
Generating public/private rsa key pair.
/home/adagniew407/.ssh/id_rsa already exists.
Overwrite (y/n)? y
Your identification has been saved in /home/adagniew407/.ssh/id_rsa
Your public key has been saved in /home/adagniew407/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:0CfcyYm8GDcp7xRzEnW+lO2Hemb3amG65tsOABshUmg adagniew407@instance-20240603-07334
9
The key's randomart image is:
+---[RSA 3072]----+
|    .oo o.. .    |
|    E. = B = o   |
|   .  + / B + .  |
|       B & . o . |
|      . S . . o .|
|       o   . .o. |
|        .   oo+..|
|            o*...|
|           o+=+..|
+----[SHA256]-----+
adagniew407@instance-20240603-073349:~$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_key
s
adagniew407@instance-20240603-073349:~$ chmod 0600 ~/.ssh/authorized_keys
adagniew407@instance-20240603-073349:~$ ssh localhost
Welcome to Ubuntu 20.04.6 LTS (GNU/Linux 5.15.0-1060-gcp x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/pro

 System information as of Wed Jun  5 20:57:13 PDT 2024
```

Step 2: Create the Input files and necessary java files for the task

$ vi file0

$ vi file1

$ vi file2

```
adagniew407@instance-20240603-073349:~/InvertedIndex$ ls
file0  file1  file2
adagniew407@instance-20240603-073349:~/InvertedIndex$ cat *
it is what it is
what is it
it is a banana
```

Go to the Hadoop-3.4.0/

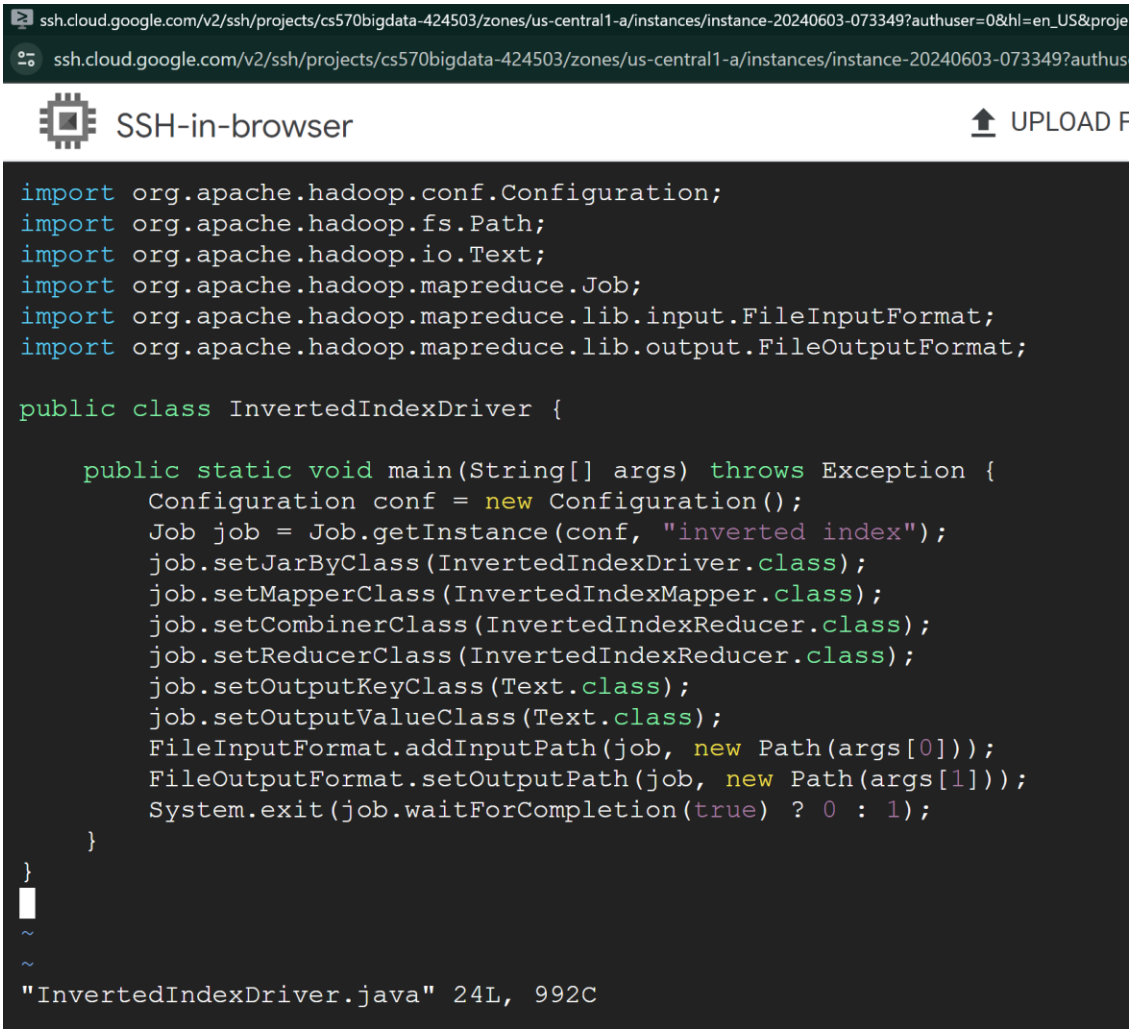
$ cd hadoop-3.4.0/

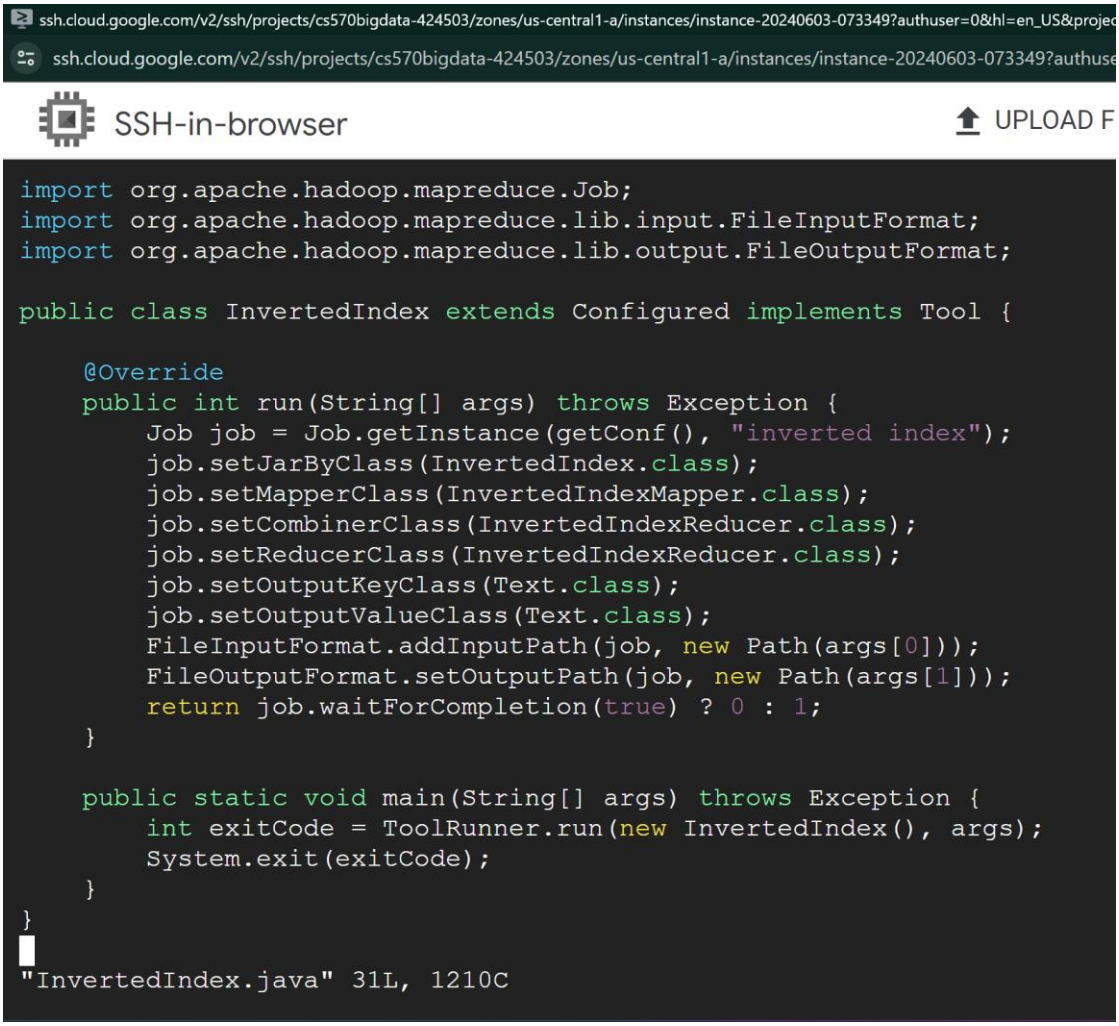
```
adagniew407@instance-20240603-073349:~$ cd hadoop-3.4.0/
adagniew407@instance-20240603-073349:~/hadoop-3.4.0$
```

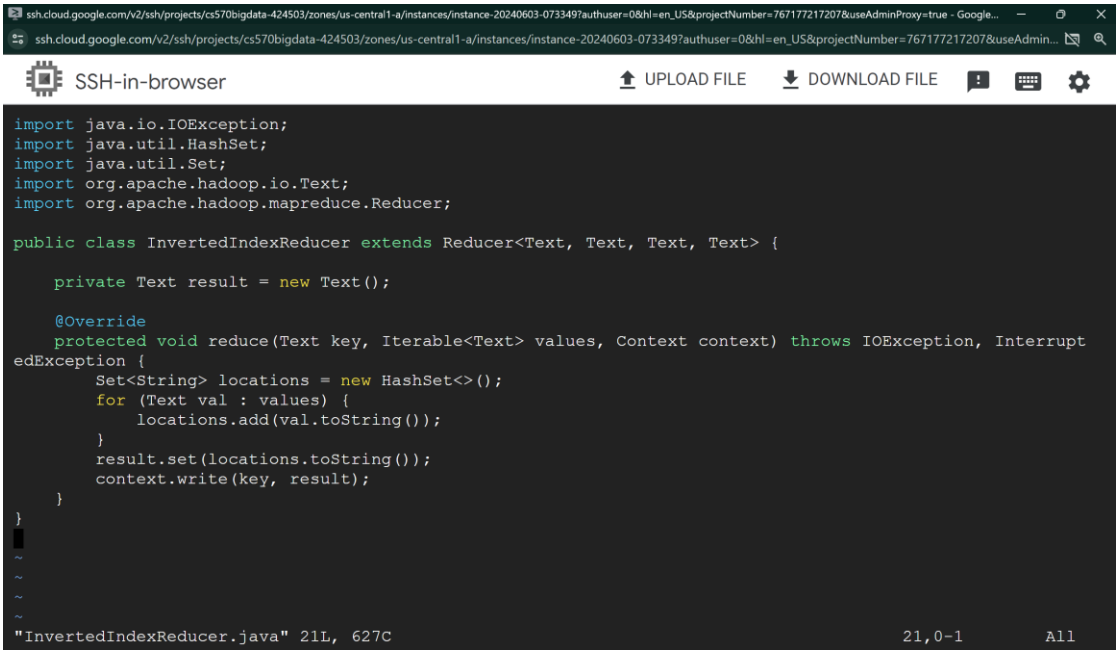
Process

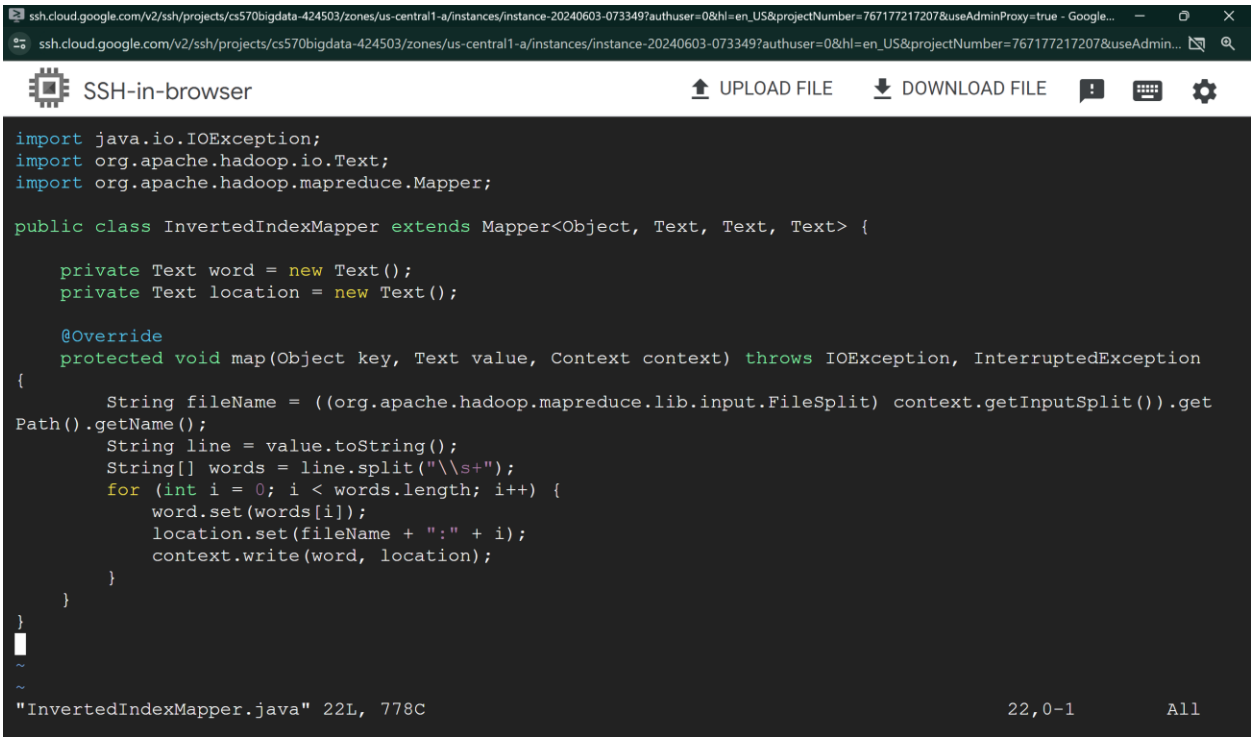
Lets create the files: java files for map-reduce

```
adagniew407@instance-20240603-073349:~/hadoop-3.4.0$ vi InvertedIndexDriver.java
adagniew407@instance-20240603-073349:~/hadoop-3.4.0$ vi InvertedIndex.java
adagniew407@instance-20240603-073349:~/hadoop-3.4.0$ vi InvertedIndexReducer.java
adagniew407@instance-20240603-073349:~/hadoop-3.4.0$ vi InvertedIndexMapper.java
```




ssh.cloud.google.com/v2/ssh/projects/cs570bigdata-424503/zones/us-central1-a/instances/instance-20240603-073349?authuser=0&hl=en_US&projec
ssh.cloud.google.com/v2/ssh/projects/cs570bigdata-424503/zones/us-central1-a/instances/instance-20240603-073349?authuse
SSH-in-browser
UPLOAD F
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
public class InvertedIndex extends Configured implements Tool {
@Override
public int run(String[] args) throws Exception {
Job job = Job.getInstance(getConf(), "inverted index");
job.setJarByClass(InvertedIndex.class);
job.setMapperClass(InvertedIndexMapper.class);
job.setCombinerClass(InvertedIndexReducer.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 {
int exitCode = ToolRunner.run(new InvertedIndex(), args);
System.exit(exitCode);
}
}
"InvertedIndex.java" 31L, 1210C


ssh.cloud.google.com/v2/ssh/projects/cs570bigdata-424503/zones/us-central1-a/instances/instance-20240603-073349?authuser=0&hl=en_US&projectNumber=767177217207&useAdminProxy=true - Google...
ssh.cloud.google.com/v2/ssh/projects/cs570bigdata-424503/zones/us-central1-a/instances/instance-20240603-073349?authuser=0&hl=en_US&projectNumber=767177217207&useAdmin...
SSH-in-browser
UPLOAD FILE
DOWNLOAD FILE
import java.io.IOException;
import java.util.HashSet;
import java.util.Set;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Reducer;
public class InvertedIndexReducer extends Reducer<Text, Text, Text, Text> {
private Text result = new Text();
@Override
protected void reduce(Text key, Iterable<Text> values, Context context) throws IOException, Interrupt
edException {
Set<String> locations = new HashSet<>();
for (Text val : values) {
locations.add(val.toString());
}
result.set(locations.toString());
context.write(key, result);
}
}
"InvertedIndexReducer.java" 21L, 627C
21,0-1
All




```
import java.io.IOException;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Mapper;

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

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

    @Override
    protected void map(Object key, Text value, Context context) throws IOException, InterruptedException
{
        String fileName = ((org.apache.hadoop.mapreduce.lib.input.FileSplit) context.getInputSplit()).get
Path().getName();
        String line = value.toString();
        String[] words = line.split("\\s+");
        for (int i = 0; i < words.length; i++) {
            word.set(words[i]);
            location.set(fileName + ":" + i);
            context.write(word, location);
        }
    }
}
```

Explanation:

1. Job Configuration (InvertedIndex.java)

- Main Class (**InvertedIndex**):
    - The class InvertedIndex extends Configured and implements Tool.
    - In the run method, a new Job is created with the job name "inverted index".
    - The job configuration specifies the jar file containing the code (InvertedIndex.class), the mapper class (InvertedIndexMapper.class), and the reducer class (InvertedIndexReducer.class).
    - Input and output paths are set using FileInputFormat.addInputPath and FileOutputFormat.setOutputPath.
    - The waitForCompletion method is called to execute the job and wait for its completion. It returns 0 if the job succeeds, otherwise 1.
    - The main method uses ToolRunner.run to parse command-line arguments and execute the job.

2. Mapper (InvertedIndexMapper.java)

- Mapper Class (**InvertedIndexMapper**):
    - This class extends Mapper<Object, Text, Text, Text>.

- The map method reads each line of the input file. The filename is obtained from the context.
- The line is split into words using whitespace as a delimiter.
- For each word, the filename and position within the line are combined into a single string (e.g., file0:0, file0:3, etc.).
- The word and its corresponding location are written to the context as key-value pairs (Text).

3. Reducer (InvertedIndexReducer.java)

- Reducer Class (**InvertedIndexReducer**):
  - This class extends Reducer<Text, Text, Text, Text>.
  - The reduce method receives a word (key) and a list of locations (values).
  - It collects all unique locations into a Set to remove duplicates.
  - The set of locations is converted to a string and written to the context with the word as the key and the locations as the value.

4. Execution

- Compiling and Running the Job:
  - Java source files are compiled using javac, and a JAR file is created using jar.
  - The job is run with the command bin/hadoop jar inverted-index.jar InvertedIndex /input /output.
  - Input files are read from HDFS, processed by the mapper and reducer, and the output is written back to HDFS**.**

5. Output Explanation

- The mapper emits key-value pairs where the key is a word and the value is a location (file and index).
- The reducer combines these values for each word into a set of unique locations and emits the word with its associated locations.

Step 4: Now lets compile these files in this step makes sure the java and hadoop configurations are taken care of:

Compile Java Code:

Navigate to the directory containing your Java source files

```
javac -classpath $(~/hadoop-3.4.0/bin/hadoop classpath) *.java
```

```
adagniew407@instance-20240603-073349:~/hadoop-3.4.0$ javac -classpath $(~/hadoop-3.4.0/bin/hadoop classpath) *.java
```

```
adagniew407@instance-20240603-073349:~/hadoop-3.4.0$ ls
InvertedIndex.class         InvertedIndexReducer.class  README.txt    index.html.2        logs
InvertedIndex.java          InvertedIndexReducer.java   bin           input               output
InvertedIndexDriver.class   LICENSE-binary              etc           inverted-index.jar  sbin
InvertedIndexDriver.java    LICENSE.txt                 include       lib                 share
InvertedIndexMapper.class   NOTICE-binary               index.html    libexec
InvertedIndexMapper.java    NOTICE.txt                  index.html.1  licenses-binary
```

Create JAR File:

```
jar cf inverted-index.jar *.class
```

```
adagniew407@instance-20240603-073349:~/hadoop-3.4.0$ javac -classpath $(~/hadoop-3.4.0/bin/hadoop classpath) *.java
adagniew407@instance-20240603-073349:~/hadoop-3.4.0$ jar cf inverted-index.jar *.class
```

Then lets make sure we are connected to the hadoop cluster

```
adagniew407@instance-20240603-073349:~/hadoop-3.4.0$ jps
2146 SecondaryNameNode
8502 Jps
1787 NameNode
1935 DataNode
```

Step 5: let's copy the files to the Hadoop cluster

```
adagniew407@instance-20240603-073349:~/hadoop-3.4.0$ sbin/start-dfs.sh
```

Create Directories in HDFS:

```
# Navigate to Hadoop home directory
cd ~/hadoop-3.4.0

# Create necessary directories in HDFS
bin/hdfs dfs -mkdir /user
bin/hdfs dfs -mkdir /user/adagniew407
bin/hdfs dfs -mkdir /user/adagniew407/fullinvertedindexcalculation
bin/hdfs dfs -mkdir/user/adagniew407/fullinvertedindexcalculation/input
```

```
adagniew407@instance-20240603-073349:~/hadoop-3.4.0$ bin/hdfs dfs -mkdir /user
mkdir: `/user': File exists
adagniew407@instance-20240603-073349:~/hadoop-3.4.0$ bin/hdfs dfs -mkdir /user/adagniew407
mkdir: `/user/adagniew407': File exists
adagniew407@instance-20240603-073349:~/hadoop-3.4.0$ bin/hdfs dfs -mkdir /user/adagniew407/fullinvertedindexcalculation
mkdir: `/user/adagniew407/fullinvertedindexcalculation': File exists
adagniew407@instance-20240603-073349:~/hadoop-3.4.0$ bin/hdfs dfs -mkdir /user/adagniew407/fullinvertedindexcalculation/input
mkdir: `/user/adagniew407/fullinvertedindexcalculation/input': File exists
```

Upload Input Files to HDFS:

```
bin/hdfs dfs -put ~/InvertedIndex/file0
/user/adagniew407/fullinvertedindexcalculation/input
bin/hdfs dfs -put ~/InvertedIndex/file1
/user/adagniew407/fullinvertedindexcalculation/input
bin/hdfs dfs -put ~/InvertedIndex/file2
/user/adagniew407/fullinvertedindexcalculation/input
```

```
adagniew407@instance-20240603-073349:~/hadoop-3.4.0$ bin/hdfs dfs -put ~/InvertedIndex/file0 /user/adagniew407/fullinvertedindexcalculation/input
put: `/user/adagniew407/fullinvertedindexcalculation/input/file0': File exists
adagniew407@instance-20240603-073349:~/hadoop-3.4.0$ bin/hdfs dfs -put ~/InvertedIndex/file1 /user/adagniew407/fullinvertedindexcalculation/input
put: `/user/adagniew407/fullinvertedindexcalculation/input/file1': File exists
adagniew407@instance-20240603-073349:~/hadoop-3.4.0$ bin/hdfs dfs -put ~/InvertedIndex/file2 /user/adagniew407/fullinvertedindexcalculation/input
put: `/user/adagniew407/fullinvertedindexcalculation/input/file2': File exists
adagniew407@instance-20240603-073349:~/hadoop-3.4.0$
```

```
adagniew407@instance-20240603-073349:~/hadoop-3.4.0$ bin/hdfs dfs -ls /user/adagniew407/fullinvertedindexcalculation/input
Found 3 items
-rw-r--r--   1 adagniew407 supergroup         17 2024-06-05 16:54 /user/adagniew407/fullinvertedindexcalculation/input/file0
-rw-r--r--   1 adagniew407 supergroup         11 2024-06-05 16:54 /user/adagniew407/fullinvertedindexcalculation/input/file1
-rw-r--r--   1 adagniew407 supergroup         15 2024-06-05 16:55 /user/adagniew407/fullinvertedindexcalculation/input/file2
```

Submit and Run - Full Inverted Index the Job:

```
# Navigate to Hadoop home directory
cd ~/hadoop-3.4.0

# Run the MapReduce job with the created JAR file
bin/hadoop jar inverted-index.jar InvertedIndex
/user/adagniew407/fullinvertedindexcalculation/input
/user/adagniew407/fullinvertedindexcalculation/output
```

7. **Verify Output:**

```
bin/hdfs dfs -ls /user/adagniew407/fullinvertedindexcalculation/output
bin/hdfs dfs -cat
/user/adagniew407/fullinvertedindexcalculation/output/part-*
```

```
adagniew407@instance-20240603-073349:~/hadoop-3.4.0$ bin/hadoop jar inverted-index.jar InvertedIndex /user/adagniew407/fullinvertedindexcalculation/input /user/
adagniew407/fullinvertedindexcalculation/output_new
2024-06-05 21:00:41,055 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2024-06-05 21:00:41,231 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2024-06-05 21:00:41,231 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2024-06-05 21:00:41,709 INFO input.FileInputFormat: Total input files to process : 3
2024-06-05 21:00:41,740 INFO mapreduce.JobSubmitter: number of splits:3
2024-06-05 21:00:41,934 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local363782931_0001
2024-06-05 21:00:41,934 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-06-05 21:00:42,189 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2024-06-05 21:00:42,191 INFO mapreduce.Job: Running job: job_local363782931_0001
2024-06-05 21:00:42,199 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2024-06-05 21:00:42,213 INFO output.PathOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommitterFactory
2024-06-05 21:00:42,216 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2024-06-05 21:00:42,216 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failur
es: false
2024-06-05 21:00:42,218 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
2024-06-05 21:00:42,291 INFO mapred.LocalJobRunner: Waiting for map tasks
2024-06-05 21:00:42,292 INFO mapred.LocalJobRunner: Starting task: attempt_local363782931_0001_m_000000_0
2024-06-05 21:00:42,328 INFO output.PathOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommitterFactory
2024-06-05 21:00:42,328 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2024-06-05 21:00:42,328 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failur
es: false
2024-06-05 21:00:42,370 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
2024-06-05 21:00:42,377 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/adagniew407/fullinvertedindexcalculation/input/file0:0+17
2024-06-05 21:00:42,414 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
```

```
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=135
                HDFS: Number of bytes written=153
                HDFS: Number of read operations=35
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=6
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Map input records=3
                Map output records=12
                Map output bytes=139
                Map output materialized bytes=193
                Input split bytes=432
                Combine input records=12
                Combine output records=10
                Reduce input groups=5
                Reduce shuffle bytes=193
                Reduce input records=10
                Reduce output records=5
                Spilled Records=20
                Shuffled Maps =3
                Failed Shuffles=0
                Merged Map outputs=3
                GC time elapsed (ms)=12
                Total committed heap usage (bytes)=1516765184
        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=153
adagniew407@instance-20240603-073349:~/hadoop-3.4.0$ # List output directory contents
```

Check the out put

```
adagniew407@instance-20240603-073349:~/hadoop-3.4.0$ # List output directory contents
adagniew407@instance-20240603-073349:~/hadoop-3.4.0$ bin/hdfs dfs -ls /user/adagniew407/fullinvertedindexcalculation/output_new
Found 2 items
-rw-r--r--   1 adagniew407 supergroup          0 2024-06-05 21:00 /user/adagniew407/fullinvertedindexcalculation/output_new/_SUCCESS
-rw-r--r--   1 adagniew407 supergroup        153 2024-06-05 21:00 /user/adagniew407/fullinvertedindexcalculation/output_new/part-r-00000
```

```
adagniew407@instance-20240603-073349:~/hadoop-3.4.0$ # View the output files
adagniew407@instance-20240603-073349:~/hadoop-3.4.0$ bin/hdfs dfs -cat /user/adagniew407/fullinvertedindexcalculation/output_new/part-*
a       [[file2:2]]
banana  [[file2:3]]
is      [[file2:1], [file1:1], [file0:1, file0:4]]
it      [[file1:2], [file0:0, file0:3], [file2:0]]
what    [[file1:0], [file0:2]]
```