

Full Inverted Index

A full inverted index is an inverted index that additionally contains the positions of each word within a document.

Full Inverted Index

a: {(2, 2)}
banana: {(2, 3)}
is: {(0, 1), (0, 4), (1, 1), (2, 1)}
it: {(0, 0), (0, 3), (1, 2), (2, 0)}
what: {(0, 2), (1, 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)				

Setup

Google Cloud

CS570JF

Search (/) for resources, docs, products, and more

Search

9

Create an instance

EQUIVALENT CODE

HELP ASSISTANT

Create a VM instance, select one of the options:

New VM instance

Create a single VM instance from scratch

New VM instance from template

Create a single VM instance from an existing template

New VM instance from machine image

Create a single VM instance from an existing machine image

Marketplace

Deploy a ready-to-go solution onto a VM instance

Name *

w2h3

Labels

+ ADD LABELS

Region *

us-west2 (Los Angeles)

Region is permanent

Zone *

us-west2-a

Zone is permanent

Machine configuration

General purpose

Compute optimized

Memory optimized

GPUs

Machine types for common workloads, optimized for cost and flexibility

Series

E2

CPU platform selection based on availability

Machine type

Choose a machine type with preset amounts of vCPUs and memory that suit most workloads. Or, you can create a custom machine for your workload's particular needs. [Learn more](#)

PRESET

CUSTOM

e2-micro (2 vCPU, 1 GB memory)

Pricing summary

Monthly estimate

\$8.54

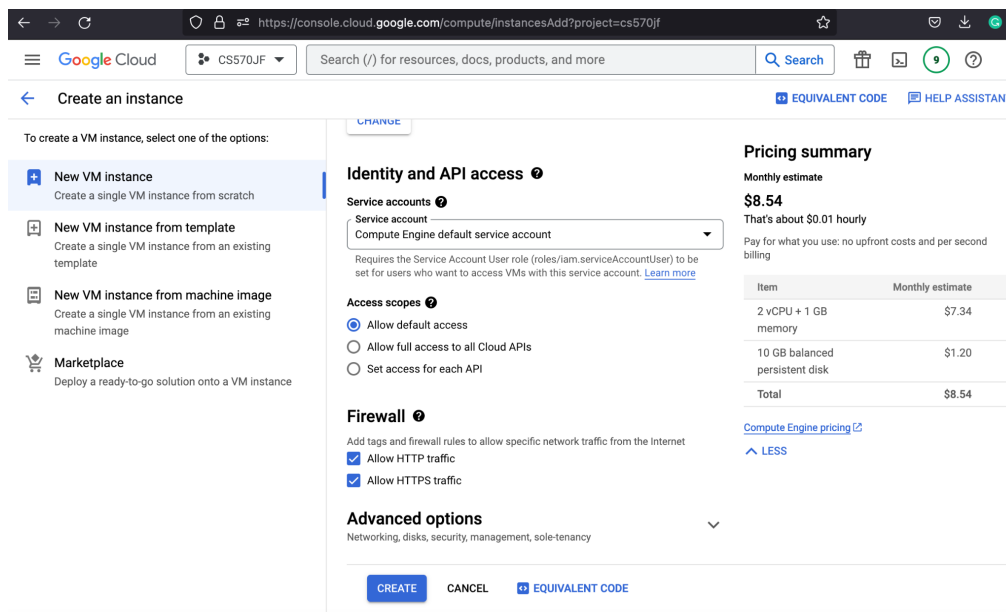
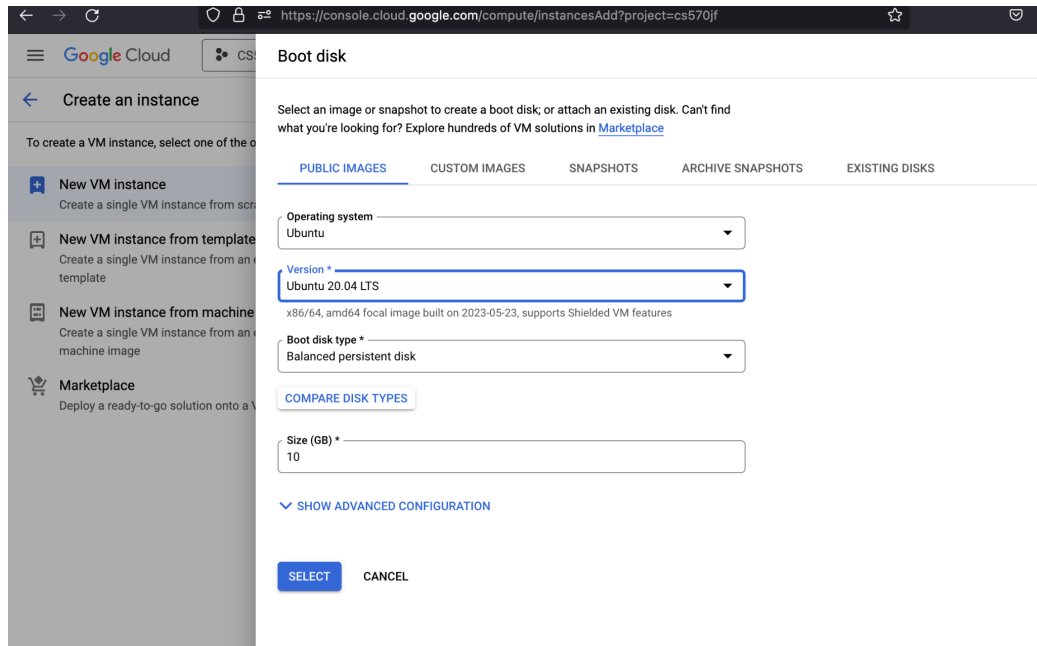
That's about \$0.01 hourly

Pay for what you use: no upfront costs and per second billing

Item	Monthly estimate
2 vCPU + 1 GB memory	\$7.34
10 GB balanced persistent disk	\$1.20
Total	\$8.54

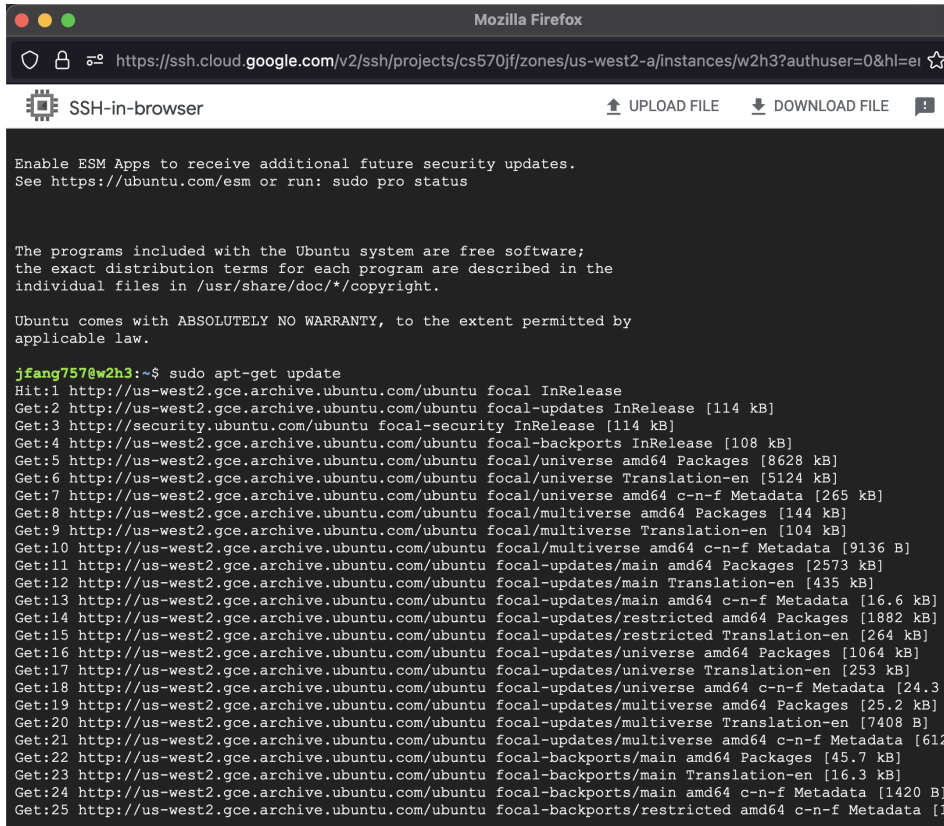
[Compute Engine pricing](#)

^ LESS



Update

\$ sudo apt-get update



```
Enable ESM Apps to receive additional future security updates.
See https://ubuntu.com/esm or run: sudo pro status

The programs included with the Ubuntu system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Ubuntu comes with ABSOLUTELY NO WARRANTY, to the extent permitted by
applicable law.

jfang757@w2h3:~$ sudo apt-get update
Hit:1 http://us-west2.gce.archive.ubuntu.com/ubuntu focal InRelease
Get:2 http://us-west2.gce.archive.ubuntu.com/ubuntu focal-updates InRelease [114 kB]
Get:3 http://security.ubuntu.com/ubuntu focal-security InRelease [114 kB]
Get:4 http://us-west2.gce.archive.ubuntu.com/ubuntu focal-backports InRelease [108 kB]
Get:5 http://us-west2.gce.archive.ubuntu.com/ubuntu focal/universe amd64 Packages [8628 kB]
Get:6 http://us-west2.gce.archive.ubuntu.com/ubuntu focal/universe Translation-en [5124 kB]
Get:7 http://us-west2.gce.archive.ubuntu.com/ubuntu focal/universe amd64 c-n-f Metadata [265 kB]
Get:8 http://us-west2.gce.archive.ubuntu.com/ubuntu focal/multiverse amd64 Packages [144 kB]
Get:9 http://us-west2.gce.archive.ubuntu.com/ubuntu focal/multiverse Translation-en [104 kB]
Get:10 http://us-west2.gce.archive.ubuntu.com/ubuntu focal/multiverse amd64 c-n-f Metadata [9136 B]
Get:11 http://us-west2.gce.archive.ubuntu.com/ubuntu focal-updates/main amd64 Packages [2573 kB]
Get:12 http://us-west2.gce.archive.ubuntu.com/ubuntu focal-updates/main Translation-en [435 kB]
Get:13 http://us-west2.gce.archive.ubuntu.com/ubuntu focal-updates/main amd64 c-n-f Metadata [16.6 kB]
Get:14 http://us-west2.gce.archive.ubuntu.com/ubuntu focal-updates/restricted amd64 Packages [1882 kB]
Get:15 http://us-west2.gce.archive.ubuntu.com/ubuntu focal-updates/restricted Translation-en [264 kB]
Get:16 http://us-west2.gce.archive.ubuntu.com/ubuntu focal-updates/universe amd64 Packages [1064 kB]
Get:17 http://us-west2.gce.archive.ubuntu.com/ubuntu focal-updates/universe Translation-en [253 kB]
Get:18 http://us-west2.gce.archive.ubuntu.com/ubuntu focal-updates/universe amd64 c-n-f Metadata [24.3 kB]
Get:19 http://us-west2.gce.archive.ubuntu.com/ubuntu focal-updates/multiverse amd64 Packages [25.2 kB]
Get:20 http://us-west2.gce.archive.ubuntu.com/ubuntu focal-updates/multiverse Translation-en [7408 B]
Get:21 http://us-west2.gce.archive.ubuntu.com/ubuntu focal-updates/multiverse amd64 c-n-f Metadata [612 B]
Get:22 http://us-west2.gce.archive.ubuntu.com/ubuntu focal-backports/main amd64 Packages [45.7 kB]
Get:23 http://us-west2.gce.archive.ubuntu.com/ubuntu focal-backports/main Translation-en [16.3 kB]
Get:24 http://us-west2.gce.archive.ubuntu.com/ubuntu focal-backports/main amd64 c-n-f Metadata [1420 B]
Get:25 http://us-west2.gce.archive.ubuntu.com/ubuntu focal-backports/restricted amd64 c-n-f Metadata [11
```

Install Java idk

```
$ sudo apt-get install openjdk-8-jdk
```

Check version

```
$ java -version
```

Install ssh

```
$ sudo apt-get install ssh
```

```

0 added, 0 removed; done.
Running hooks in /etc/ca-certificates/update.d...
done.
done.
Processing triggers for mime-support (3.64ubuntu1) ...
Processing triggers for libc-bin (2.31-0ubuntu9.9) ...
Processing triggers for libgdk-pixbuf2.0-0:amd64 (2.40.0+dfsg-3ubuntu0.4) ...
jffang757@w2h3:~$ java -version
openjdk version "1.8.0_362"
OpenJDK Runtime Environment (build 1.8.0_362-8u372-ga-us1-0ubuntu1-20.04-b09)
OpenJDK 64-Bit Server VM (build 25.362-b09, mixed mode)
jffang757@w2h3:~$ sudo apt-get install ssh
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following packages were automatically installed and are no longer required:
  libatasmart4 libblockdev-fs2 libblockdev-loop2 libblockdev-part-err2 libblockdev-part2 libblockdev-sw
  libblockdev-utils2 libblockdev2 libbim-glib4 libbim-proxy libmm-glib0 libnumal libparted-fs-resize0
  libqmi-glib5 libqmi-proxy libudisks2-0 libxmlb2 usb-modeswitch usb-modeswitch-data
Use 'sudo apt autoremove' to remove them.
The following NEW packages will be installed:
  ssh
0 upgraded, 1 newly installed, 0 to remove and 1 not upgraded.
Need to get 5080 B of archives.
After this operation, 120 kB of additional disk space will be used.
Get:1 http://us-west2.gce.archive.ubuntu.com/ubuntu focal-updates/main amd64 ssh all 1:8.2p1-4ubuntu0.7
]
Fetched 5080 B in 0s (69.3 kB/s)
Selecting previously unselected package ssh.
(Reading database ... 77799 files and directories currently installed.)
Preparing to unpack .../ssh_1%3a8.2p1-4ubuntu0.7_all.deb ...
Unpacking ssh (1:8.2p1-4ubuntu0.7) ...
Setting up ssh (1:8.2p1-4ubuntu0.7) ...
jffang757@w2h3:~$

```

Download Hadoop 3.3.5

\$ wget <https://dlcdn.apache.org/hadoop/common/hadoop-3.3.5/hadoop-3.3.5.tar.gz>

```

jffang757@w2h3:~$ wget https://dlcdn.apache.org/hadoop/common/hadoop-3.3.5/hadoop-3.3.5.tar.gz
--2023-05-30 08:50:49-- https://dlcdn.apache.org/hadoop/common/hadoop-3.3.5/hadoop-3.3.5.tar.gz
Resolving dlcdn.apache.org (dlcdn.apache.org)... 151.101.2.132, 2a04:4e42::644
Connecting to dlcdn.apache.org (dlcdn.apache.org)|151.101.2.132|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 706533213 (674M) [application/x-gzip]
Saving to: 'hadoop-3.3.5.tar.gz'

hadoop-3.3.5.tar.gz      100%[=====>] 673.80M   101MB/s   in 6.0s

2023-05-30 08:50:55 (112 MB/s) - 'hadoop-3.3.5.tar.gz' saved [706533213/706533213]

jffang757@w2h3:~$

```

Unzip the tar file

\$ tar xzf hadoop-3.3.5.tar.gz

\$ cd hadoop-3.3.5

\$ ls -all

```

jffang757@w2h3:~$ tar xzf hadoop-3.3.5.tar.gz
jffang757@w2h3:~$ cd hadoop-3.3.5
jffang757@w2h3:~/hadoop-3.3.5$ ls -all
total 120
drwxr-xr-x 10 jffang757 jffang757 4096 Mar 15 16:58 .
drwxr-xr-x  5 jffang757 jffang757 4096 May 30 08:51 ..
-rw-rw-r-- 1 jffang757 jffang757 24496 Feb 25 09:59 LICENSE-binary
-rw-rw-r-- 1 jffang757 jffang757 15217 Jul 16 2022 LICENSE.txt
-rw-rw-r-- 1 jffang757 jffang757 29473 Jul 16 2022 NOTICE-binary
-rw-rw-r-- 1 jffang757 jffang757 1541 Apr 22 2022 NOTICE.txt
-rw-rw-r-- 1 jffang757 jffang757 175 Apr 22 2022 README.txt
drwxr-xr-x  2 jffang757 jffang757 4096 Mar 15 16:58 bin
drwxr-xr-x  3 jffang757 jffang757 4096 Mar 15 15:58 etc
drwxr-xr-x  2 jffang757 jffang757 4096 Mar 15 16:58 include
drwxr-xr-x  3 jffang757 jffang757 4096 Mar 15 16:58 lib
drwxr-xr-x  4 jffang757 jffang757 4096 Mar 15 16:58 libexec
drwxr-xr-x  2 jffang757 jffang757 4096 Mar 15 16:58 licenses-binary
drwxr-xr-x  3 jffang757 jffang757 4096 Mar 15 15:58 sbin
drwxr-xr-x  4 jffang757 jffang757 4096 Mar 15 17:27 share
jffang757@w2h3:~/hadoop-3.3.5$

```

Find Java root

\$ update-alternatives --list java

```
jffang757@w2h3:~/hadoop-3.3.5$ update-alternatives --list java
/usr/lib/jvm/java-8-openjdk-amd64/jre/bin/java
jffang757@w2h3:~/hadoop-3.3.5$
```

Modify bash file

\$ vi ~/.bashrc

\$. ~/.bashrc

```
#set JAVA_HOME
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
#set HADOOP_HOME
export HADOOP_HOME=$HOME/hadoop-3.3.5
#add directory of Hadoop to path
export PATH=$PATH:$HADOOP_HOME/bin
export HADOOP_CLASSPATH=$JAVA_HOME/lib/tools.jar
```

Press “i” to edit and press “esc” to finish the edited

Type “:wq” to save and exit

```
jffang757@w2h3:~/hadoop-3.3.5$ . ~/.bashrc
jffang757@w2h3:~/hadoop-3.3.5$ echo $JAVA_HOME
/usr/lib/jvm/java-8-openjdk-amd64
jffang757@w2h3:~/hadoop-3.3.5$ echo $HADOOP_HOME
/home/jffang757/hadoop-3.3.5
jffang757@w2h3:~/hadoop-3.3.5$
```

Configurations related to HDFS

\$ cd etc/hadoop/

\$ vi hadoop-env.sh

```
jffang757@w2h3:~/hadoop-3.3.5$ cd etc/hadoop/
jffang757@w2h3:~/hadoop-3.3.5/etc/hadoop$ vi hadoop-env.sh
jffang757@w2h3:~/hadoop-3.3.5/etc/hadoop$
```

```
# such as in /etc/profile.d

# The java implementation to use. By default, this environment
# variable is REQUIRED on ALL platforms except OS X!
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64

# Location of Hadoop. By default, Hadoop will attempt to determine
# this location based upon its execution path.
# export HADOOP_HOME=

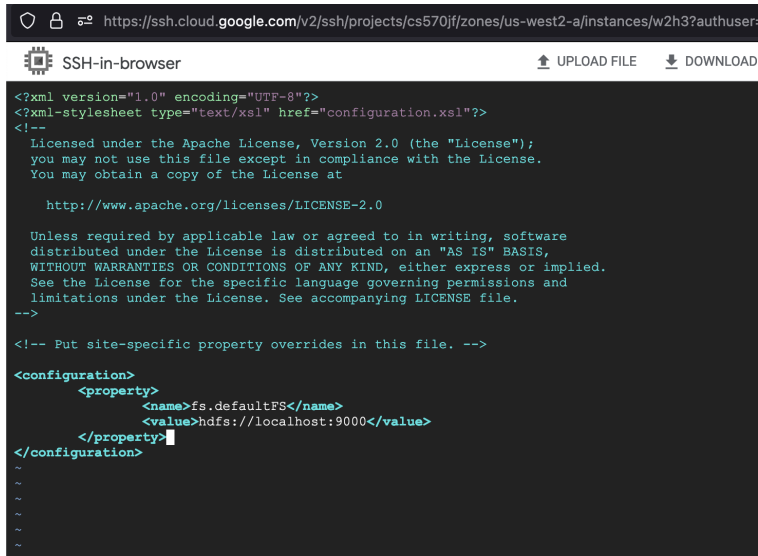
# Location of Hadoop's configuration information. i.e., where this
# file is living. If this is not defined, Hadoop will attempt to
# locate it based upon its execution path.
```

Pseudo-Distributed Operation

```
-rw-r--r-- 1 jfang757 jfang757 0 May 30 09:14 _SUCCESS
-rw-r--r-- 1 jfang757 jfang757 11 May 30 09:14 part-r-00000
jfang757@w2h3:~/hadoop-3.3.5$ vi ./etc/hadoop/core-site.xml
jfang757@w2h3:~/hadoop-3.3.5$ vi ./etc/hadoop/hdfs-site.xml
jfang757@w2h3:~/hadoop-3.3.5$
```

\$ vi ./etc/hadoop/core-site.xml

```
<configuration>
  <property>
    <name>fs.defaultFS</name>
    <value>hdfs://localhost:9000</value>
  </property>
</configuration>
```



\$ vi ./etc/hadoop/hdfs-site.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
  <property>
    <name>dfs.replication</name>
    <value>1</value>
  </property>
</configuration>
~
~
```

```
<configuration>
  <property>
    <name>dfs.replication</name>
    <value>1</value>
```

```
</property>
</configuration>
```

Execution

Create code and input files

```
$ mkdir WordCount
$ cd WordCount
$ vi FullInvertedIndex.java
```

Source Code

```
import java.io.*;
import java.util.*;

import org.apache.hadoop.fs.Path;
import org.apache.hadoop.conf.*;
import org.apache.hadoop.io.*;
import org.apache.hadoop.mapred.*;
import org.apache.hadoop.mapred.Mapper;
import org.apache.hadoop.mapred.Reducer;
import org.apache.hadoop.mapreduce.*;
import org.apache.hadoop.mapreduce.Reducer.Context;
import org.apache.hadoop.util.*;

public class FullInvertedIndex {

    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();
            String line = value.toString();
            StringTokenizer tokenizer = new StringTokenizer(line);
            int lineNumber = 1;

            while (tokenizer.hasMoreTokens())
            {
```

```

        String token = tokenizer.nextToken().toLowerCase();
        output.collect(new Text(token), new Text(fileName + ":" + lineNumber));
        lineNumber++;
    }
}

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
    {
        StringBuilder result = new StringBuilder();
        HashMap<String, List<Integer>> docLineMap = new HashMap<>();

        while (values.hasNext()) {
            String docLine = values.next().toString();
            String[] docLineArr = docLine.split(":");
            if (docLineArr.length == 2) {
                String docId = docLineArr[0];
                int lineNumber = Integer.parseInt(docLineArr[1]);

                if (docLineMap.containsKey(docId)) {
                    docLineMap.get(docId).add(lineNumber);
                } else {
                    List<Integer> lineNumbers = new ArrayList<>();
                    lineNumbers.add(lineNumber);
                    docLineMap.put(docId, lineNumbers);
                }
            }
        }

        for (Map.Entry<String, List<Integer>> entry : docLineMap.entrySet()) {
            String docId = entry.getKey();
            List<Integer> lineNumbers = entry.getValue();
            result.append(docId).append(": ").append(lineNumbers.toString()).append(",
");
        }

        output.collect(key, new Text(result.toString()));
    }
}

```



```

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

    JobConf conf = new JobConf(FullInvertedIndex.class);
    conf.setJobName("fullInvertedIndexer");
    conf.setMapOutputKeyClass(Text.class);
    conf.setMapOutputValueClass(Text.class);
    conf.setOutputKeyClass(Text.class);
    conf.setOutputValueClass(Text.class);

    conf.setMapperClass(IndexMapper.class);
    conf.setReducerClass(IndexReducer.class);

    conf.setInputFormat(TextInputFormat.class);
    conf.setOutputFormat(TextOutputFormat.class);

    FileInputFormat.setInputPaths(conf, new Path(args[0]));
    FileOutputFormat.setOutputPath(conf, new Path(args[1]));

    JobClient.runJob(conf);
}
}

```

```

jfang757@w3h2:~/hadoop-3.3.5$ cat output/*
1      dfsadmin
jfang757@w3h2:~/hadoop-3.3.5$ ls ./output
_SUCCESS  part-r-00000
jfang757@w3h2:~/hadoop-3.3.5$ ls -al ./output
total 20
drwxr-xr-x  2 jfang757 jfang757 4096 Aug  2 00:36 .
drwxr-xr-x 12 jfang757 jfang757 4096 Aug  2 00:36 ..
-rw-r--r--  1 jfang757 jfang757   8 Aug  2 00:36 _SUCCESS.crc
-rw-r--r--  1 jfang757 jfang757  12 Aug  2 00:36 .part-r-00000.crc
-rw-r--r--  1 jfang757 jfang757   0 Aug  2 00:36 _SUCCESS
-rw-r--r--  1 jfang757 jfang757  11 Aug  2 00:36 part-r-00000
jfang757@w3h2:~/hadoop-3.3.5$ vi ./etc/hadoop/core-site.xml
jfang757@w3h2:~/hadoop-3.3.5$ vi ./etc/hadoop/hdfs-site.xml
jfang757@w3h2:~/hadoop-3.3.5$ cd ..
jfang757@w3h2:~$ mkdir WordCount
jfang757@w3h2:~$ cd WordCount
jfang757@w3h2:~/WordCount$ vi FullInvertedIndex.java
jfang757@w3h2:~/WordCount$ mkdir input
jfang757@w3h2:~/WordCount$ cd input
jfang757@w3h2:~/WordCount/input$ vi file0
jfang757@w3h2:~/WordCount/input$ vi file1
jfang757@w3h2:~/WordCount/input$ vi file2
jfang757@w3h2:~/WordCount/input$ █

```

Connect to local host

Setup passphraseless ssh

```
$ ssh-keygen -t rsa -P "" -f ~/.ssh/id_rsa
```

```
jfang757@w2h3:~/hadoop-3.3.5$ ssh-keygen -t rsa -P "" -f ~/.ssh/id_rsa
Generating public/private rsa key pair.
Your identification has been saved in /home/jfang757/.ssh/id_rsa
Your public key has been saved in /home/jfang757/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:UkbGSN6qYJpGEX07R+NywR/ezlIOgosFmIq79yPrEHY jfang757@w2h3
The key's randomart image is:
+---[RSA 3072]-----+
| o . .oo |
|..+ o..+ |
|oo . =..+ |
|o o = +=o |
|+ooE =o+So |
|o*+. =...* |
|=o . . . + |
|.oo . . |
|.oo+.. |
+----[SHA256]-----+
```

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

```
$ chmod 0600 ~/.ssh/authorized_keys
```

```
$ ssh localhost
```

```
jfang757@w2h3:~/hadoop-3.3.5$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
jfang757@w2h3:~/hadoop-3.3.5$ chmod 0600 ~/.ssh/authorized_keys
jfang757@w2h3:~/hadoop-3.3.5$ ssh localhost
Welcome to Ubuntu 20.04.6 LTS (GNU/Linux 5.15.0-1034-gcp x86_64)

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

System information as of Tue May 30 09:26:04 UTC 2023

System load:  0.0               Processes:           105
Usage of /:   47.8% of 9.51GB   Users logged in:    1
Memory usage: 29%              IPv4 address for ens4: 10.168.0.5
Swap usage:   0%

Expanded Security Maintenance for Applications is not enabled.

1 update can be applied immediately.
1 of these updates is a standard security update.
To see these additional updates run: apt list --upgradable

Enable ESM Apps to receive additional future security updates.
See https://ubuntu.com/esm or run: sudo pro status

New release '22.04.2 LTS' available.
Run 'do-release-upgrade' to upgrade to it.

Last login: Tue May 30 08:41:45 2023 from 35.235.241.194
jfang757@w2h3:~$
```

Format the filesystem

\$ cd hadoop-3.3.5

\$ bin/hdfs namenode -format

```
jfang757@w2h3:~$ cd hadoop-3.3.5
jfang757@w2h3:~/hadoop-3.3.5$ bin/hdfs namenode -format
WARNING: /home/jfang757/hadoop-3.3.5/logs does not exist. Creating.
2023-05-30 09:28:14,387 INFO namenode.NameNode: STARTUP_MSG:
/******
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   host = w2h3.us-west2-a.c.cs570jf.internal/10.168.0.5
STARTUP_MSG:   args = [-format]
STARTUP_MSG:   version = 3.3.5
STARTUP_MSG:   classpath = /home/jfang757/hadoop-3.3.5/etc/hadoop:/home/jfang757/hadoop-3.3.5/share/hadoop/common/lib/jettison-1.5.3.jar:/home/jfang757/hadoop-3.3.5/share/hadoop/common/lib/commons-lang3-3.12.0.jar:/home/jfang757/hadoop-3.3.5/share/hadoop/common/lib/jackson-core-2.12.7.jar:/home/jfang757/hadoop-3.3.5/share/hadoop/common/lib/netty-transport-rxtx-4.1.77.Final.jar:/home/jfang757/hadoop-3.3.5/share/hadoop/common/lib/netty-transport-native-unix-common-4.1.77.Final.jar:/home/jfang757/hadoop-3.3.5/share/hadoop/common/lib/jakarta.activation-api-1.2.1.jar:/home/jfang757/hadoop-3.3.5/share/hadoop/common/lib/httpcore-4.4.13.jar:/home/jfang757/hadoop-3.3.5/share/hadoop/common/lib/netty-transport-native-kqueue-4.1.77.Final-osx-x86_64.jar:/home/jfang757/hadoop-3.3.5/share/hadoop/common/lib/kerb-client-1.0.1.jar:/home/jfang757/hadoop-3.3.5/share/hadoop/common/lib/kerb-common-1.0.1.jar:/home/jfang757/hadoop-3.3.5/share/hadoop/common/lib/jersey-json-1.20.jar:/home/jfang757/hadoop-3.3.5/share/hadoop/common/lib/hadoop-auth-3.3.5.jar:/home/jfang757/hadoop-3.3.5/share/hadoop/common/lib/protobuf-jav-a-2.5.0.jar:/home/jfang757/hadoop-3.3.5/share/hadoop/common/lib/commons-text-1.10.0.jar:/home/jfang757/hadoop-3.3.5/share/hadoop/common/lib/jetty-xml-9.4.48.v20220622.jar:/home/jfang757/hadoop-3.3.5/share/hadoop/common/lib/jetty-util-9.4.48.v20220622.jar:/home/jfang757/hadoop-3.3.5/share/hadoop/common/lib/jsr311-api-1.1.1.jar:/home
```

\$ sbin/start-dfs.sh

\$ wget http://localhost:9870/

```
jfang757@w3h2:~/hadoop-3.3.5$ sbin/start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [w3h2]
w3h2: Warning: Permanently added 'w3h2,10.168.0.32' (ECDSA) to the list of known hosts.
jfang757@w3h2:~/hadoop-3.3.5$ wget http://localhost:9870/
--2023-08-02 00:52:56-- 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]
--2023-08-02 00:52:56-- 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

2023-08-02 00:52:56 (108 MB/s) - 'index.html' saved [1079/1079]

jfang757@w3h2:~/hadoop-3.3.5$ bin/hdfs dfs -mkdir /user
jfang757@w3h2:~/hadoop-3.3.5$ bin/hdfs dfs -mkdir /user/jfang757
jfang757@w3h2:~/hadoop-3.3.5$ bin/hdfs dfs -mkdir /user/jfang757/wordcount
jfang757@w3h2:~/hadoop-3.3.5$ bin/hdfs dfs -mkdir /user/jfang757/wordcount/input
jfang757@w3h2:~/hadoop-3.3.5$ bin/hdfs dfs -put ../WordCount/input/* /user/jfang757/wordcount/input
jfang757@w3h2:~/hadoop-3.3.5$ bin/hdfs dfs -ls /user/jfang757/wordcount/input
Found 3 items
-rw-r--r-- 1 jfang757 supergroup 17 2023-08-02 00:54 /user/jfang757/wordcount/input/file0
-rw-r--r-- 1 jfang757 supergroup 11 2023-08-02 00:54 /user/jfang757/wordcount/input/file1
-rw-r--r-- 1 jfang757 supergroup 15 2023-08-02 00:54 /user/jfang757/wordcount/input/file2
jfang757@w3h2:~/hadoop-3.3.5$ bin/hadoop com.sun.tools.javac.Main ../WordCount/FullInvertedIndex.java
jfang757@w3h2:~/hadoop-3.3.5$ cp ../WordCount/*.class .
jfang757@w3h2:~/hadoop-3.3.5$ cp ../WordCount/*.java .
jfang757@w3h2:~/hadoop-3.3.5$ jar cf wc.jar FullInvertedIndex.class
jfang757@w3h2:~/hadoop-3.3.5$ ls
'FullInvertedIndex$IndexMapper.class'  FullInvertedIndex.java  NOTICE-binary  bin  index.html  libexec  output  wc.jar
'FullInvertedIndex$IndexReducer.class'  LICENSE-binary         NOTICE.txt     etc  input      licenses-binary  sbin
FullInvertedIndex.class                 LICENSE.txt            README.txt     include  lib      logs          share
jfang757@w3h2:~/hadoop-3.3.5$
```

Make the HDFS directories required to execute MapReduce jobs

\$ bin/hdfs dfs -mkdir /user

\$ bin/hdfs dfs -mkdir /user/jfang757

\$ bin/hdfs dfs -mkdir /user/jfang757/wordcount

\$ bin/hdfs dfs -mkdir /user/jfang757/wordcount/input

Copy the input files into the distributed file system

```
bin/hdfs dfs -put ../WordCount/input/* /user/jfang757/wordcount/input
```

Move .class files to hadoop-3.3.5 directory and create jar

```
$ bin/hadoop com.sun.tools.javac.Main ../WordCount/FullInvertedIndex.java
```

```
$ cp ../WordCount/*.class .
```

```
$ cp ../WordCount/*.java .
```

```
$ jar cf wc.jar FullInvertedIndex*.class
```

```
jfang757@w2h3:~/hadoop-3.3.5$ bin/hadoop com.sun.tools.javac.Main ../WordCount/WordCount.java
jfang757@w2h3:~/hadoop-3.3.5$ cp ../WordCount/*.class .
jfang757@w2h3:~/hadoop-3.3.5$ cp ../WordCount/*.java .
jfang757@w2h3:~/hadoop-3.3.5$ jar cf wc.jar WordCount*.class
jfang757@w2h3:~/hadoop-3.3.5$ ls
LICENSE-binary  WordCount          bin                index.html.2      licenses-binary  share
LICENSE.txt     'WordCount$IntSumReducer.class'  etc                index.html.3      logs             wc.jar
NOTICE-binary   'WordCount$TokenizerMapper.class' include            input             output
NOTICE.txt      WordCount.class    index.html         lib               output1
README.txt      WordCount.java     index.html.1      libexec           sbin
```

Result

```
$ bin/hadoop jar wc.jar FullInvertedIndex /user/jfang757/wordcount/input
```

```
/user/jfang757/wordcount/output
```

```

HDFS: Number of bytes written=151
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=181
  Input split bytes=327
  Combine input records=0
  Combine output records=0
  Reduce input groups=5
  Reduce shuffle bytes=181
  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)=42
  Total committed heap usage (bytes)=1185939456
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=151
jffang757@w3h2:~/hadoop-3.3.5$ bin/hdfs dfs -ls /user/jffang757/wordcount/output
Found 2 items
-rw-r--r--  1 jffang757 supergroup          0 2023-08-02 00:57 /user/jffang757/wordcount/output/_SUCCESS
-rw-r--r--  1 jffang757 supergroup       151 2023-08-02 00:57 /user/jffang757/wordcount/output/part-000000
jffang757@w3h2:~/hadoop-3.3.5$ bin/hdfs dfs -cat /user/jffang757/wordcount/output/part-r-000000
cat: `/user/jffang757/wordcount/output/part-r-000000': No such file or directory
jffang757@w3h2:~/hadoop-3.3.5$ bin/hdfs dfs -cat /user/jffang757/wordcount/output/part-000000
a      file2: [3],
banana file2: [4],
is     file2: [2], file0: [5, 2], file1: [2],
it     file2: [1], file0: [4, 1], file1: [3],
what   file0: [3], file1: [1],
jffang757@w3h2:~/hadoop-3.3.5$ sbin/stop-dfs.sh
Stopping namenodes on [localhost]
Stopping datanodes
Stopping secondary namenodes [w3h2]
jffang757@w3h2:~/hadoop-3.3.5$

```