

Amazon Inverted Index

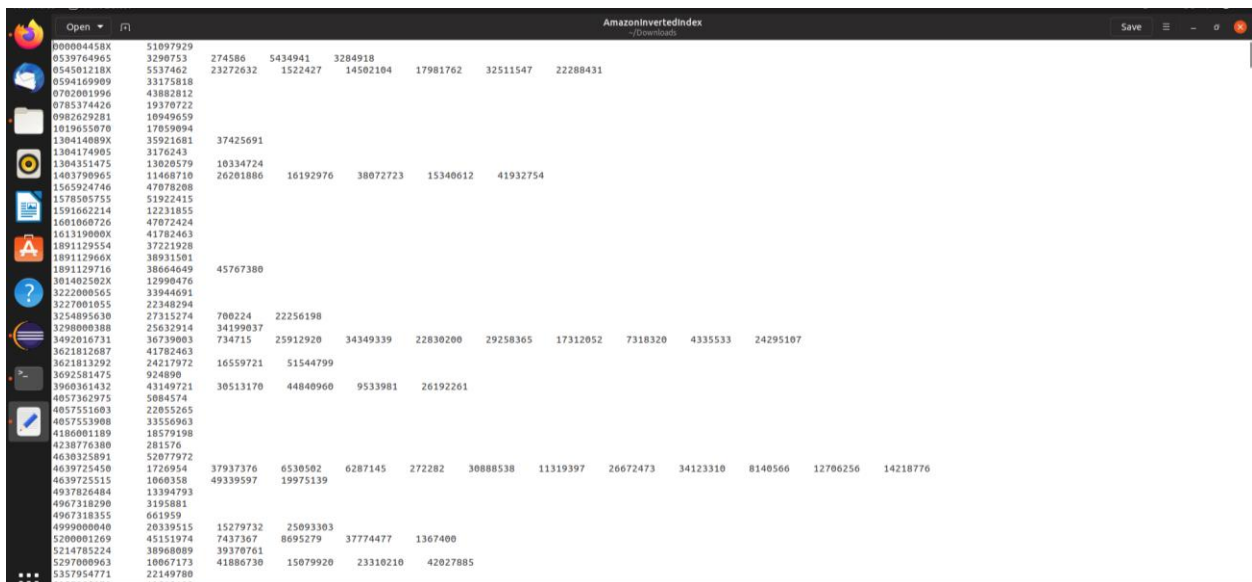
(Numerical Summarization Pattern)

In this analysis, Inverted Indexing is used. It is very helpful as the output can show what Product are reviewed by which users. The key value will have multiple values.

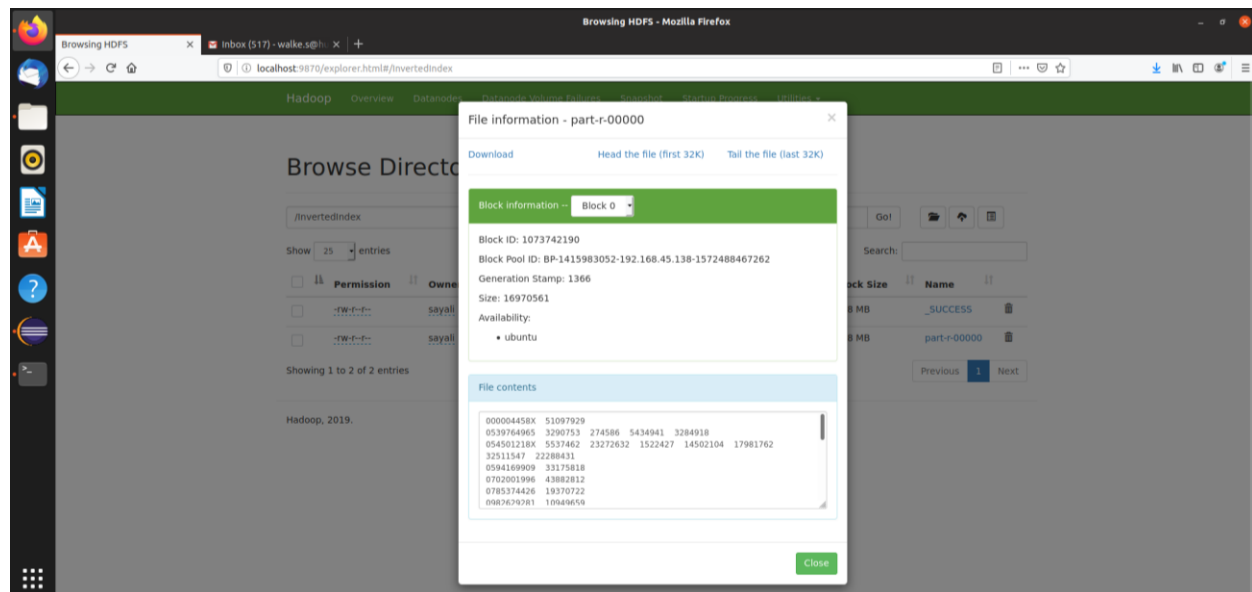
OUTPUT:

```
./hadoop jar /home/sayali/Desktop/AmazonInvertedIndex.jar
```

```
sayali.AmazonReviews2.AmazonInvertedIndex /AmazonReviews/AmazonReviews.tsv /InvertedIndex
```



000004458X	51097929								
0539764965	3290753	274586	5434941	3284918					
054501218X	5537462	23272632	1522427	14502104	17981762	32511547	22288431		
0594169909	33175818								
0702001996	43882812								
0785374426	19370722								
0982629281	10949659								
1019655070	17059094								
130414089X	35921681	37425691							
1304174905	3176243								
1304351475	13020579	10334724							
1403790965	11408710	26201886	16192976	38072723	15340612	41932754			
1565924746	47078206								
1578505755	51922415								
1591662214	12231855								
1601060726	47072424								
161139000X	41782463								
1891129554	37221928								
189112966X	38931501								
1891129710	38664649	45767380							
301402502X	12990476								
3222000565	33944691								
3227001055	22348294								
3254895630	27315274	700224	22256198						
3298000388	25632914	34199037							
3492016731	36739003	734715	25912920	34349339	22830200	29258365	17312052	7318320	4335533
3621812687	41782463								
3621813292	24217972	16559721	51544799						
3692581475	924890								
3960361432	43149721	30513170	44840960	9533981	26192261				
4057362975	5084574								
4057551603	22055265								
4057553908	33556963								
4186001189	18579198								
4238776380	281576								
4630325891	52077972								
4639725450	1728954	37937376	6530582	6287145	272282	30888538	11319397	26072473	34123310
4639725515	1060358	49339597	19975139						
4937826484	13394793								
4967318290	3195881								
4967318355	661959								
4999000040	20339515	15279732	25093303						
5200001269	45151974	7437367	8695279	37774477	1367400				
5214785224	38960809	39370761							
5297000963	10067173	41886730	15079920	23310210	42027885				
5357954771	22149780								
5367065075	43504113								



Browsing HDFS - Mozilla Firefox

localhost:9876/explorer.html#/invertedindex

Hadoop Overview Datanodes DataNodes Volume Element Resources Status Progress Utilities

Browse Directory

/invertedindex

Show 25 entries

Permission Owner

Showing 1 to 2 of 2 entries

Hadoop, 2019.

File information - part-r-00000

Download Head the file (first 32K) Tail the file (last 32K)

Block information - Block 0

Block ID: 1073742190

Block Pool ID: BP-1415983052-192.168.45.138-1572488467262

Generation Stamp: 1366

Size: 16970561

Availability:

- ubuntu

File contents

000004458X 51097929
0539764965 3290753 274586 5434941 3284918
054501218X 5537462 23272632 1522427 14502104 17981762
0594169909 33175818
0702001996 43882812
0785374426 19370722
0982629281 10949659
1019655070 17059094
130414089X 35921681 37425691
1304174905 3176243
1304351475 13020579 10334724
1403790965 11408710 26201886 16192976 38072723 15340612 41932754
1565924746 47078206
1578505755 51922415
1591662214 12231855
1601060726 47072424
161139000X 41782463
1891129554 37221928
189112966X 38931501
1891129710 38664649 45767380
301402502X 12990476
3222000565 33944691
3227001055 22348294
3254895630 27315274 700224 22256198
3298000388 25632914 34199037
3492016731 36739003 734715 25912920 34349339 22830200 29258365 17312052 7318320 4335533 24295107
3621812687 41782463
3621813292 24217972 16559721 51544799
3692581475 924890
3960361432 43149721 30513170 44840960 9533981 26192261
4057362975 5084574
4057551603 22055265
4057553908 33556963
4186001189 18579198
4238776380 281576
4630325891 52077972
4639725450 1728954 37937376 6530582 6287145 272282 30888538 11319397 26072473 34123310 8140560 12706256 14218776
4639725515 1060358 49339597 19975139
4937826484 13394793
4967318290 3195881
4967318355 661959
4999000040 20339515 15279732 25093303
5200001269 45151974 7437367 8695279 37774477 1367400
5214785224 38960809 39370761
5297000963 10067173 41886730 15079920 23310210 42027885
5357954771 22149780
5367065075 43504113

Map Reduce Code

```
import java.io.IOException;

import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
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.output.FileOutputFormat;
```

Mapper:

```
public class AmazonInvertedIndexMapper extends Mapper<LongWritable, Text, Text, Text> {
    private Text product = new Text();
    private Text customerId = new Text();

    public void map(LongWritable key, Text values, Context context) throws IOException {

        if (key.get() == 0) {
            return;
        }
        String[] tokens = values.toString().split("\\t");
        customerId.set(tokens[1]);
        product.set(tokens[3]);
        try {
            context.write(product, customerId);
        } catch (InterruptedException e) {
            e.printStackTrace();
        }
    }
}
```

Reducer:

```
public class AmazonInvertedIndexReducer extends Reducer<Text, Text, Text, Text> {
    private Text result = new Text();

    @Override
    public void reduce(Text key, Iterable<Text> values, Context context) throws IOException,
    InterruptedException {

        StringBuilder sb = new StringBuilder();
        boolean first = true;

        for (Text id : values) {
            if (first) {
```

```

        first = false;
    } else {
        sb.append(" ");
    }
    sb.append(id.toString());
}
result.set(sb.toString());
context.write(key, result);
}
}

```

Main:

```

public class AmazonInvertedIndex {

    public static void main(String[] args) throws IOException, InterruptedException,
    ClassNotFoundException {
        Configuration conf = new Configuration();
        Job job = Job.getInstance(conf, "Inverted Index");
        job.setJarByClass(AmazonInvertedIndex.class);
        job.setMapperClass(AmazonInvertedIndexMapper.class);
        job.setReducerClass(AmazonInvertedIndexReducer.class);

        job.setMapOutputKeyClass(Text.class);
        job.setMapOutputValueClass(Text.class);

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

        FileInputFormat.addInputPath(job, new Path(args[0]));
        FileOutputFormat.setOutputPath(job, new Path(args[1]));
        System.exit(job.waitForCompletion(true) ? 0 : 1);
    }
}

```