BIG DATA AND HADOOP

PROOF OF CONCEPT

BY RAHUL SHEDGE

rahulshedge555@outlook.com

TABLE OF CONTENT

Page No.

1.PROBLEM STATEMENT

1.Load data into HDFS using **HDFS client**

2. Develop MR program to parse logs and convert request string into structured format

(/ a/b/c/d => a b c d)

3. Count of page views by individual user

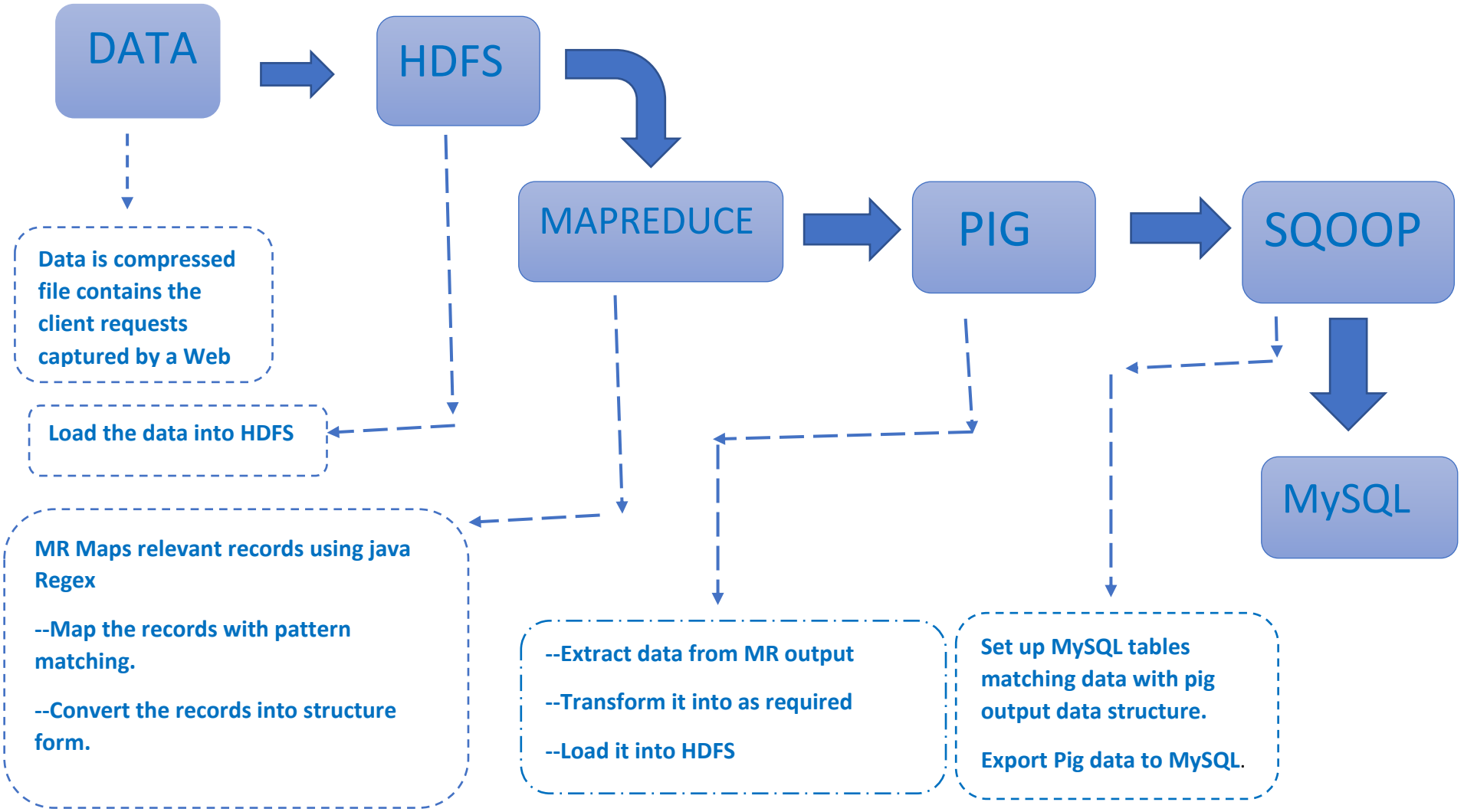
4. Top / Bottom 5: catagery-1/ catagery-2 / page /users / entry pages (Exclude status code other than 200, also exclude record related to css/js/image)

5. Total page views / Category wise pageviews / Unique pageviews

6.Count of status code = 200 / 404 / 400 / 500

7. Load results into tables in MySQL Database using Sqoop.

2.SOLUTION ARCHITECTURE



3.SOFTWARE AND TOOLS SPECIFICATIONS

HW/SW COMPONENTS	DESCRIPTION
Single Node Cluster	Ubuntu 14.4 VM Images, running on Window 7,64 bit Set up single node cluster with Ubuntu VM image as below. - 192.168.182.128 (Master) Namenode, Datanode
RAM / Physical Memory	4 GB RAM ,20 GB for Ubuntu image
JAVA	1.7
IDE & tools	Eclipse 4.14, Putty, WinSCP
Hadoop	2.7.2
Apache Pig	Version 0.16.0: For this project, ETL can be accomplished using Pig
Apache Sqoop	Version 1.4.6: Used for data transfer from HDFS to RDBMS(MySQL)
MySQL	Version 5.5.53: For storing daily trending data.

4.SOLUTION DESCRIPTION

The dataset consists 1,00,000 rows of web logs. The data is not in relation format and it is large to import to MySQL database, hence to analyse this data I am going to use Hadoop Ecosystem which is a platform to solve big data problems. It includes Apache Projects & various commercial tools and solutions. There are few major elements of Hadoop like HDFS, MapReduce, Yarn & most of the tools use for supplement and support those major elements.

The solution to gather required metrics will be develop using Hadoop API. The solution will demonstrate key features of Hadoop API, such as:

1. Map Reduce for filtering, categorizing & converting data into Structured format.
2. Pig for extracting data from MapReduce, Transforming & Arranging as required.
3. Sqoop for exporting data from HDFS into RDBMS system.

5.PROGRAM CODE

5.1 MapReduce

Objective: Removing invalid records.

Removing unwanted pattern and strings from weblogs.

Filtering required part of logs as needed.

Setting all filtered strings in structured format.

Dataset: - Number of records: 1,00,000

Dataset Size: around 25MB

Sample Dataset: -

21.125.155.111 - - [01/Jan/2012:12:07:48 +0530] "GET /digital-cameras/digital-camera/sony-qx-dsc-qx100-point-shoot-digital-camera-black.html HTTP/1.1" 200 1470 "Mozilla/5.0 (Windows; U; Windows NT 6.1; en-US; rv:1.9.2.17) Gecko/20110420 Firefox/3.6.17" "-"

168.42.128.252 - - [01/Jan/2012:12:09:36 +0530] "GET /digital-cameras/digital-camera/canon-powershot-sx50-hs-point-shoot-camera.html HTTP/1.1" 200 195 "Mozilla/5.0 (Windows; U; Windows NT 6.1; en-US; rv:1.9.2.17) Gecko/20110420 Firefox/3.6.17" "-"

196.34.35.201 - - [01/Jan/2012:12:18:18 +0530] "GET /tvs-audio/blu-ray-dvd-players/d-m-holdings-inc-denon-dbt-1713ud-blu-ray-player.html HTTP/1.1" 200 1503 "Mozilla/5.0 (Windows NT 6.2) AppleWebKit/537.17 (KHTML, like Gecko) Chrome/24.0.1312.56 Safari/537.17" "-"

91.228.209.0 - - [01/Jan/2012:12:28:04 +0530] "GET /home-appliances/fans/reconnet-rhcfg-1201-ceiling-fan.html HTTP/1.1" 200 773 "Mozilla/5.0 (Windows NT 6.2) AppleWebKit/537.17 (KHTML, like Gecko) Chrome/24.0.1312.56 Safari/537.17" "-"

114.231.104.220 - - [01/Jan/2012:12:44:23 +0530] "GET /catalogsearch/result/index/?dir=asc&order=price&q=samsung HTTP/1.1" 200 1168 "AdsBot-Google-Mobile (http://www.google.com/mobile/adsbot.html) Mozilla (iPhone U CPU iPhone OS 3 0 like Mac OS X) AppleWebKit (KHTML, like Gecko)

Data Structure:

Weblogs Content	Description
Host	21.125.155.111 (IP address of the client (remote host) which made the request)
Identity	(Identity of the client)
Userid	(userid of the person requesting the document)
Date, Time and Timezone	[01/Jan/2012:12:07:48 +0530]
Request Line	"GET /digital-cameras/digital-camera/sony-qx-dsc-qx100-point-shoot-digital-camerablack.html HTTP/1.1"
Status code	200 (Note: 2xx is a successful response, 3xx a redirection, 4xx a client error, and 5xx a server error.)
Object size	1470 is the size of the object returned to the client, measured in bytes.
Agent	"Mozilla/5.0 (Windows; U; Windows NT 6.1; en-US; rv:1.9.2.17) Gecko/20110420 Firefox/3.6.17"
Referrer URL	-

Approach:

Mapper:

1. Omit the invalid records.
2. Match the pattern with weblog and collect individual strings.
3. Arrange the collected strings or part of weblogs in structured format
4. Output will be in this form -> "user catagery-1 catagery-2 page status code" all are separated by tab.

Reducer: No need of reduce class -> Setting **setNumReduceTasks(0)**

Code

Mapperclass.java

```
package com.project;
import java.io.IOException;
import java.util.regex.Matcher;
import java.util.regex.Pattern;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Mapper;

public class Mapperclass extends Mapper<LongWritable, Text, Text, Text>{
	Text k2 = new Text();
	public static final int NUM_FIELDS = 9;
	private final static Text i = new Text("");
	@Override
	public void map( LongWritable key, Text value, Context context ) throws IOException, InterruptedException  {

		String v = new String();
		String logString = value.toString();
		//90.160.130.234 - - [01/Jan/2012:01:52:06 +0530] "GET /tvs-audio/speciality-speakers/jbl-pulse-speaker-
black.html HTTP/1.1" 200 1172 "Mozilla/5.0 (Windows NT 6.2) AppleWebKit/537.17 (KHTML, like Gecko) Chrome/24.0.1312.56
Safari/537.17" "-"
		String logEntryPattern = "^([\\d.]+) (\\S+) (\\S+) \\[([\\w:/]+\\s[+\\-]\\d{4})\\] \"(.+?)\" (\\d{3}) (\\d+)
\"([^\"]+)\" \"([^\"]+)\"";

	Pattern p = Pattern.compile(logEntryPattern);
	Matcher matcher = p.matcher(logString);

	if (!matcher.matches() || NUM_FIELDS != matcher.groupCount()
			|| matcher.group(5).split(" ")[1].length() <= 2
			||matcher.group(5).split("")[1].split("/")[2].contains(".") ) {
```

```
            System.err.println("Bad log entry (or problem with RE?):");
             System.err.println(logString);




  }else {
            String IP_Address = matcher.group(1);
            String Request_line = matcher.group(5);

            String[] tabs = Request_line.split(" ");
            String Page = tabs[1];
            String[] Category = Page.split("/");
            String Category1 = Category[1];
            String Category2 = Category[2];
            String Status_code = matcher.group(6);


          v = IP_Address+"\t"+Category1+"\t"+Category2+"\t"+Page+"\t"+Status_code;
         k2.set(v);
          }

      context.write(k2, i);

  }
}


//done
```

Main Class: LogsDriverClass.java

```
package com.project;
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import com.project.LogsDriverClass;
import com.project.Mapperclass;

public class LogsDriverClass {
public static void main(String[] args) throws Exception {

             Configuration conf = new Configuration();

          Job job =  Job.getInstance(conf, "Logs file in structure form");

          job.setJarByClass( LogsDriverClass.class );
          job.setMapperClass( Mapperclass.class );

          job.setNumReduceTasks(0);

          job.setMapOutputKeyClass( Text.class );
          job.setMapOutputValueClass( IntWritable.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 );
      }
}
```

Execution:

1. Copy the data file to HDFS. Input data file HDFS /input/weblogs_1_lakh_rec.txt
2. Move the HadoopProject.jar File into the local machine -> lab/program/HadoopProject.jar
3. Execute MapReduce code: hadoop jar HadoopProject.jar com.project.LogsDriverClass /input/weblogs_1_lakh_rec.txt /output/HadoopProject

```
20/05/31 09:21:58 INFO input.FileInputFormat: Total input paths to process : 1
20/05/31 09:21:59 INFO mapreduce.JobSubmitter: number of splits:1
20/05/31 09:21:59 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_15
90941813000_0001
20/05/31 09:22:00 INFO impl.YarnClientImpl: Submitted application application_15
90941813000_0001
20/05/31 09:22:00 INFO mapreduce.Job: The url to track the job: http://ubuntu:80
88/proxy/application_1590941813000_0001/
20/05/31 09:22:00 INFO mapreduce.Job: Running job: job_1590941813000_0001
20/05/31 09:22:23 INFO mapreduce.Job: Job job_1590941813000_0001 running in uber mode : false
20/05/31 09:22:23 INFO mapreduce.Job:  map 0% reduce 0%
20/05/31 09:22:42 INFO mapreduce.Job:  map 100% reduce 0%
20/05/31 09:22:42 INFO mapreduce.Job: Job job_1590941813000_0001 completed successfully
20/05/31 09:22:42 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=117095
```

```
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=25658982
        HDFS: Number of bytes written=11976427
        HDFS: Number of read operations=5
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=2
    Job Counters
        Launched map tasks=1
        Data-local map tasks=1
        Total time spent by all maps in occupied slots (ms)=15681
        Total time spent by all reduces in occupied slots (ms)=0
        Total time spent by all map tasks (ms)=15681
        Total vcore-milliseconds taken by all map tasks=15681
        Total megabyte-milliseconds taken by all map tasks=16057344
    Map-Reduce Framework
        Map input records=100000
        Map output records=100000
        Input split bytes=115
        Spilled Records=0
        Failed Shuffles=0
        Merged Map outputs=0
        GC time elapsed (ms)=979
        CPU time spent (ms)=6840
        Physical memory (bytes) snapshot=172388352
```

```
        Virtual memory (bytes) snapshot=825544704
        Total committed heap usage (bytes)=106430464
    File Input Format Counters
        Bytes Read=25658867
    File Output Format Counters
Bytes Written=11976427
```

Output:

Output of MapReduce get stored in folder **hdfs: //output/HadoopProject/Part-m-00000**

Here is the screenshot of output -

```
199.106.185.56  tvs-audio       hifi-system     /tvs-audio/hifi-system/samsung-mx-fs8000-xl-2-2-mini-hi-fi-music-syste
m.html  200
109.227.213.151 home-appliances air-conditioner-coolers /home-appliances/air-conditioner-coolers/onida-curve-s125cur-s
plit-air-conditioner-1-ton.html 200
142.46.186.245  tvs-audio       flat-television /tvs-audio/flat-television/samsung-ua32f6100ar-3d-technology-led-tv-32
-inch-81-cm.html                200
39.147.84.173   computers       tablets /computers/tablets/apple-ipad-mini-slate-64-gb.html     200
175.178.230.81  digital-cameras digital-camera  /digital-cameras/digital-camera/sony-nex-5t-prosumer-camera-black.html
200
175.178.230.81  digital-cameras digital-camera  /digital-cameras/digital-camera/sony-nex-5t-prosumer-camera-black.html
200
47.168.192.100  home-appliances air-conditioner-coolers /home-appliances/air-conditioner-coolers/onida-curve-s125cur-s
plit-air-conditioner-1-ton.html 200
```

The output has all the required things to solve the problem statements & its in structured format too.

5.2 Pig

Objective:

Count of page views by individual user.

Find Top / Bottom 5: catagery-1/ catagery-2 / page /users .

Find -Total page views / Category wise pageviews / Unique pageviews.

Find Count of status code = 200 / 404 / 400 / 500.

Dataset:

The dataset is MR output. We will extract the dataset from hdfs for ETL.

Approach:

Import the data file from HDFS.

Load the file into Pig.

Group the variables & get count of each.

Sort the output to get top/bottom 5 elements of variable by setting Limit to 5 in few cases.

Pig Program Code:

Pig script is below:

Program.pig

```
Data = Load '/output/HadoopProject/part-m-00000' using PigStorage('\t') as
            (ip:chararray,cat1:chararray,cat2:chararray,page:chararray,status:int);



grouped = Group Data By ip;
counts = FOREACH grouped GENERATE group,COUNT(Data.page) as user;
ordered = Order counts By user DESC;
Store ordered into '/home/notroot/data/PagesViewByUser' using PigStorage(',');

grouped1 = Group Data By cat1;
counts1 = FOREACH grouped1 GENERATE group,COUNT(Data.cat1) as Category;

ordered1 = Order counts1 By Category DESC;
Top5_Cat = Limit ordered1 5;
Store Top5_Cat into '/home/notroot/data/Top5_Cat' using PigStorage(',');

ordered2 = Order counts1 By Category ASC;
Bottom5_Cat = Limit ordered2 5;
Store Bottom5_Cat into '/home/notroot/data/Bottom5_Cat' using PigStorage(',');

grouped1 = Group Data By cat2;
counts1 = FOREACH grouped1 GENERATE group,COUNT(Data.cat2) as Category;
ordered1 = Order counts1 By Category DESC;
Top5_Sub = Limit ordered1 5;
Store Top5_Sub into '/home/notroot/data/Top5_Sub' using PigStorage(',');
```

```
ordered2 = Order counts1 By Category ASC;
Bottom5_Sub = Limit ordered2 5;
Store Bottom5_Sub into '/home/notroot/data/Bottom5_Sub' using PigStorage(',');


grouped1 = Group Data By page;
counts1 = FOREACH grouped1 GENERATE group,COUNT(Data.page) as Category;
ordered1 = Order counts1 By Category DESC;
Top5_page = Limit ordered1 5;
Store Top5_page into '/home/notroot/data/Top5_page' using PigStorage(',');


ordered2 = Order counts1 By Category ASC;
Bottom5_page = Limit ordered2 5;
Store Bottom5_page into '/home/notroot/data/Bottom5_page' using PigStorage(',');


grouped1 = Group Data By ip;
counts1 = FOREACH grouped1 GENERATE group,COUNT(Data.ip) as Category;
ordered1 = Order counts1 By Category DESC;
Top5_ip = Limit ordered1 5;
Store Top5_ip into '/home/notroot/data/Top5_user' using PigStorage(',');
```

```
ordered2 = Order counts1 By Category ASC;
Bottom5_ip = Limit ordered2 5;
Store Bottom5_ip into '/home/notroot/data/Bottom5_user' using PigStorage(',');


grouped = group Data By page;
counts1 = FOREACH grouped GENERATE group,COUNT(Data.page) as COUNTING;
ordered1 = ORDER counts1 by COUNTING DESC;
STORE ordered1 into '/home/notroot/data/TotalPageView' using PigStorage(',');


grouped = group Data By cat1;
counts1 = Foreach grouped GENERATE group ,COUNT(Data.page) as COUNTING;
ordered1 = order counts1 by COUNTING Desc;
STORE ordered1 into '/home/notroot/data/Cat1ByPageView' using PigStorage(',');


grouped = group Data By cat2;
counts1 = Foreach grouped GENERATE group, COUNT(Data.page) as COUNTING;
ordered1 = order counts1 by COUNTING Desc;
STORE ordered1 into '/home/notroot/data/Cat2ByPageView' using PigStorage(',');


grouped = group Data By page;
counts1 = Foreach grouped { distinct_ip = DISTINCT Data.ip;
GENERATE group,COUNT(distinct_ip) as UniqueViews;  } ;
```

```
ordered1 =order counts1 by UniqueViews Desc;
STORE ordered1 into '/home/notroot/data/UniqueViews' using PigStorage(',');


grouped = Group Data By status;
counts1 = Foreach grouped GENERATE group,COUNT(Data.status) as COUNTING;
ordered1 = order counts1 by COUNTING Desc;
STORE ordered1 into '/home/notroot/data/Status_codeInfo' using PigStorage(',');
```

Execution:

This Pig script is stored in local machine at -> lab/program/Program.pig

The command to run above pig script -> pig Program.pig

```
2020-06-01 06:47:52,480 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is
completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-06-01 06:47:52,512 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2020-06-01 06:47:52,798 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.7.2   0.16.0  notroot 2020-06-01 06:19:50     2020-06-01 06:47:52     GROUP_BY,ORDER_BY,LIMIT
```

Success!

Job Stats (time in seconds):

JobId Maps Reduces MaxMapTime MinMapTime AvgMapTime MedianMapTime MaxReduceTime MinReduceTimeAvgReduceTime MedianReducetime Alias Feature Outputs

job_1591013476644_0015 1 1 78 78 78 78 31 31 31 31 1-25,Data,counts,counts1,grouped,grouped1 MULTI_QUERY,COMBINER

job_1591013476644_0016 1 1 45 45 45 45 28 28 28 28 ordered1 SAMPLER

job_1591013476644_0017 1 1 23 23 23 23 7 7 7 7 ordered2 SAMPLER

job_1591013476644_0018 1 1 8 8 8 8 4 4 4 4 ordered2 SAMPLER

job_1591013476644_0019 1 1 4 4 4 4 5 5 5 5 ordered2 SAMPLER

job_1591013476644_0020 1 1 3 3 3 3 4 4 4 4 ordered1 SAMPLER

job_1591013476644_0021 1 1 7 7 7 7 4 4 4 4 ordered1 SAMPLER

job_1591013476644_0022 1 1 5 5 5 5 6 6 6 6 ordered1 SAMPLER

job_1591013476644_0023 1 1 3 3 3 3 3 3 3 3 ordered SAMPLER

job_1591013476644_0024 1 1 7 7 7 7 4 4 4 4 ordered1 SAMPLER

job_1591013476644_0025 1 1 11 11 11 11 9 9 9 9 ordered1 SAMPLER

job_1591013476644_0026 1 1 5 5 5 5 4 4 4 4 ordered2 SAMPLER

job_1591013476644_0027 1 1 5 5 5 5 3 3 3 3 ordered1 SAMPLER

job_1591013476644_0028 1 1 9 9 9 9 3 3 3 3 ordered1 SAMPLER

job_1591013476644_0029 1 1 3 3 3 3 26 26 26 26 ordered1 SAMPLER

job_1591013476644_0030 1 1 65 65 65 65 26 26 26 26 ordered ORDER_BY /home/notroot/data/PagesViewByUser,

job_1591013476644_0031 1 1 25 25 25 25 5 5 5 5 ordered1 ORDER_BY,COMBINER

job_1591013476644_0032 1 1 27 27 27 27 19 19 19 19 ordered1 ORDER_BY /home/notroot/data/UniqueViews,

job_1591013476644_0033	1	1	7	7	7	7	3	3	3	3	ordered2	ORDER_BY,COMBINER
job_1591013476644_0034	1	1	3	3	3	3	3	3	3	3	ordered1	ORDER_BY
/home/notroot/data/Cat1ByPageView,												
job_1591013476644_0035	1	1	8	8	8	8	5	5	5	5	ordered1	ORDER_BY,COMBINER
job_1591013476644_0036	1	1	3	3	3	3	3	3	3	3	ordered1	ORDER_BY,COMBINER
job_1591013476644_0037	1	1	4	4	4	4	3	3	3	3	ordered1	ORDER_BY,COMBINER
job_1591013476644_0038	1	1	4	4	4	4	5	5	5	5	ordered2	ORDER_BY,COMBINER
job_1591013476644_0039	1	1	3	3	3	3	4	4	4	4	ordered1	ORDER_BY
/home/notroot/data/Status_codeInfo,												
job_1591013476644_0040	1	1	3	3	3	3	3	3	3	3	ordered2	ORDER_BY,COMBINER
job_1591013476644_0041	1	1	3	3	3	3	3	3	3	3	ordered1	ORDER_BY
/home/notroot/data/Cat2ByPageView,												
job_1591013476644_0042	1	1	4	4	4	4	3	3	3	3	ordered2	ORDER_BY,COMBINER
job_1591013476644_0043	1	1	3	3	3	3	3	3	3	3	ordered1	ORDER_BY
/home/notroot/data/TotalPageView,												
job_1591013476644_0044	1	1	4	4	4	4	4	4	4	4	ordered1	
/home/notroot/data/Top5_Sub,												
job_1591013476644_0045	1	1	4	4	4	4	3	3	3	3	ordered2	
/home/notroot/data/Bottom5_user,												
job_1591013476644_0046	1	1	3	3	3	3	3	3	3	3	ordered2	
/home/notroot/data/Bottom5_Sub,												
job_1591013476644_0047	1	1	3	3	3	3	3	3	3	3	ordered1	
/home/notroot/data/Top5_user,												
job_1591013476644_0048	1	1	8	8	8	8	4	4	4	4	ordered1	
/home/notroot/data/Top5_Cat,												

```
job_1591013476644_0049  1     1     4     4     4     4     3     3     3     3     ordered2
/home/notroot/data/Bottom5_page,
job_1591013476644_0050  1     1     12    12    12    12    4     4     4     4     ordered2
/home/notroot/data/Bottom5_Cat,
job_1591013476644_0051  1     1     3     3     3     3     3     3     3     3     ordered1
/home/notroot/data/Top5_page,

Input(s):
Successfully read 100000 records (11976809 bytes) from: "/output/HadoopProject/part-m-00000"

Output(s):
Successfully stored 97196 records (1581726 bytes) in: "/home/notroot/data/PagesViewByUser"
Successfully stored 5 records (95 bytes) in: "/home/notroot/data/Top5_Sub"
Successfully stored 138 records (10900 bytes) in: "/home/notroot/data/UniqueViews"
Successfully stored 5 records (90 bytes) in: "/home/notroot/data/Bottom5_user"
Successfully stored 5 records (83 bytes) in: "/home/notroot/data/Bottom5_Sub"
Successfully stored 6 records (108 bytes) in: "/home/notroot/data/Cat1ByPageView"
Successfully stored 5 records (74 bytes) in: "/home/notroot/data/Top5_user"
Successfully stored 5 records (90 bytes) in: "/home/notroot/data/Top5_Cat"
Successfully stored 2 records (13 bytes) in: "/home/notroot/data/Status_codeInfo"
Successfully stored 5 records (340 bytes) in: "/home/notroot/data/Bottom5_page"
Successfully stored 5 records (86 bytes) in: "/home/notroot/data/Bottom5_Cat"
Successfully stored 20 records (360 bytes) in: "/home/notroot/data/Cat2ByPageView"
Successfully stored 5 records (400 bytes) in: "/home/notroot/data/Top5_page"
Successfully stored 138 records (10900 bytes) in: "/home/notroot/data/TotalPageView"
```

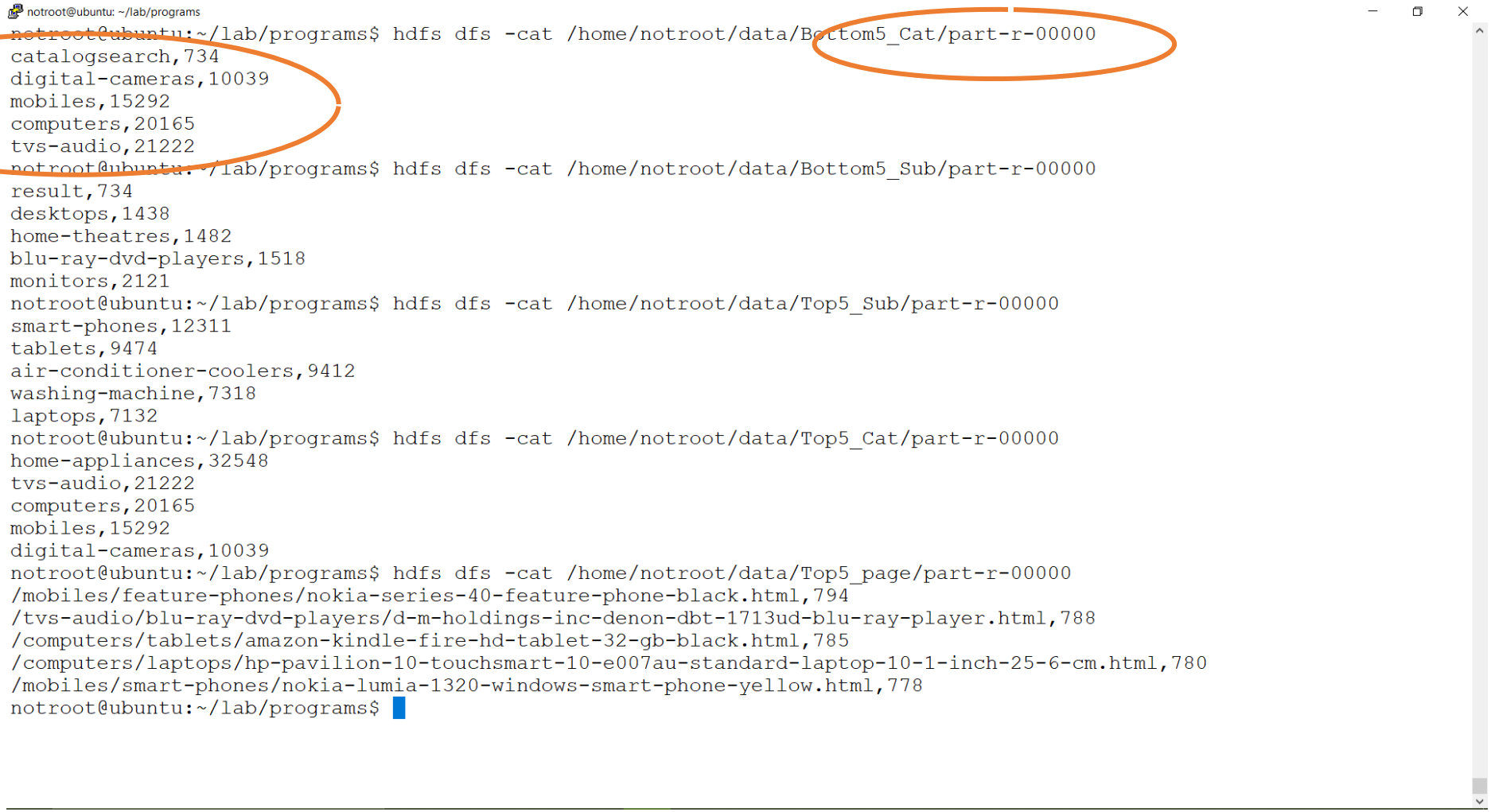
```
Counters:
Total records written : 97540
Total bytes written : 1605265
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1591013476644_0015  ->
job_1591013476644_0023,job_1591013476644_0029,job_1591013476644_0016,job_1591013476644_0018,job_15910
13476644_0019,job_1591013476644_0028,job_1591013476644_0022,job_1591013476644_0027,job_1591013476644
_0025,job_1591013476644_0026,job_1591013476644_0017,job_1591013476644_0024,job_1591013476644_0021,job
_1591013476644_0020,
job_1591013476644_0023  ->    job_1591013476644_0030,
job_1591013476644_0030
job_1591013476644_0029  ->    job_1591013476644_0031,
job_1591013476644_0031  ->    job_1591013476644_0044,
job_1591013476644_0044
job_1591013476644_0016  ->    job_1591013476644_0032,
job_1591013476644_0032
job_1591013476644_0018  ->    job_1591013476644_0033,
job_1591013476644_0033  ->    job_1591013476644_0045,
job_1591013476644_0045
job_1591013476644_0019  ->    job_1591013476644_0038,
job_1591013476644_0038  ->    job_1591013476644_0046,
job_1591013476644_0046
```

```
job_1591013476644_0028 ->    job_1591013476644_0034,
job_1591013476644_0034
job_1591013476644_0022 ->    job_1591013476644_0035,
job_1591013476644_0035 ->    job_1591013476644_0047,
job_1591013476644_0047
job_1591013476644_0027 ->    job_1591013476644_0036,
job_1591013476644_0036 ->    job_1591013476644_0048,
job_1591013476644_0048
job_1591013476644_0025 ->    job_1591013476644_0039,
job_1591013476644_0039
job_1591013476644_0026 ->    job_1591013476644_0042,
job_1591013476644_0042 ->    job_1591013476644_0049,
job_1591013476644_0049
job_1591013476644_0017 ->    job_1591013476644_0040,
job_1591013476644_0040 ->    job_1591013476644_0050,
job_1591013476644_0050
job_1591013476644_0024 ->    job_1591013476644_0041,
job_1591013476644_0041
job_1591013476644_0021 ->    job_1591013476644_0037,
job_1591013476644_0037 ->    job_1591013476644_0051,
job_1591013476644_0051
job_1591013476644_0020 ->    job_1591013476644_0043,
job_1591013476644_0043
```

Output:

Here are few outputs from Pig: All outputs of Pig are stored at /home/notroot/data/*/part-r-00000



Output of Pig has [Names of strings, count].

5.3 Sqoop:

Objective: Load the Pig output into RDBMS database (MySQL).

Dataset: Stored output files of Pig -> /home/notroot/data/*/part-r-00000.

Approach: Create RDBMS tables.

1. PagesViewedByUser
2. Top5_Cat/Bottom5_Cat
3. Top5_SubCat/Bottom5_SubCat
4. Top5_pages/Bottom5_pages
5. Top5_users/Bottom5_users
6. TotalPageViews/Cat1wisePageViews/Cat2wisePageViews/UniquepageViews
7. Status_codeInfo

Export the files from HDFS which are stored at /home/notroot/data/*/part-r-00000 to their respective tables using Sqoop.

Sqoop Code:

Code:

```
// Create database & table.

Create database HadoopProject;

Use HadoopProject;

//Create tables.

Create Table PagesViewedperUser(ip VARCHAR(15) NOT NULL,PageViewed INT);
Create Table Top5_Cat(cat1 CHAR(255) NOT NULL,Count INT);
Create Table Bottom5_Cat(cat1 CHAR(255) NOT NULL,Count INT);

Create Table Top5_SubCat(cat2 CHAR(255) NOT NULL,Count INT);
Create Table Bottom5_SubCat(cat2 CHAR(255) NOT NULL,Count INT);
Create Table Top5_pages(page VARCHAR(2083) NOT NULL,Count INT);
Create Table Bottom5_pages(page VARCHAR(2083) NOT NULL,Count INT);

Create Table Top5_users(ip VARCHAR(15) NOT NULL,Count INT);
Create Table Bottom5_users(ip VARCHAR(15) NOT NULL,Count INT);
Create Table TotalPageViews(page VARCHAR(2083) NOT NULL,Count INT);
Create Table Cat1wisePageViews(Cat1 CHAR(255) NOT NULL,Count INT);
```

```
Create Table Cat2wisePageViews(Cat2 CHAR(255) NOT NULL,Count INT);
Create Table UniquepageViews(Page VARCHAR(2083) NOT NULL,UniqueViews INT);
Create Table Status_codeInfo(Status_code INT NOT NULL,Count INT);

//Sqoop Command to export file into RDBMS.

sqoop export --connect jdbc:mysql://localhost/HadoopProject --table PagesViewedperUser --driver
com.mysql.jdbc.Driver --export-dir /home/notroot/data/PagesViewByUser/part-r-00000 --fields-terminated-by ',' --
username root --password admin;


sqoop export --connect jdbc:mysql://localhost/HadoopProject --table Top5_Cat --driver com.mysql.jdbc.Driver --export-
dir /home/notroot/data/Top5_Cat/part-r-00000 --fields-terminated-by ',' --username root --password admin;

sqoop export --connect jdbc:mysql://localhost/HadoopProject --table Bottom5_Cat --driver com.mysql.jdbc.Driver --
export-dir /home/notroot/data/Bottom5_Cat/part-r-00000 --fields-terminated-by ',' --username root --password admin;

sqoop export --connect jdbc:mysql://localhost/HadoopProject --table Top5_SubCat --driver com.mysql.jdbc.Driver --
export-dir /home/notroot/data/Top5_Sub/part-r-00000 --fields-terminated-by ',' --username root --password admin;


sqoop export --connect jdbc:mysql://localhost/HadoopProject --table Bottom5_SubCat --driver com.mysql.jdbc.Driver --
export-dir /home/notroot/data/Bottom5_Sub/part-r-00000 --fields-terminated-by ',' --username root --password
admin;
```

```
sqoop export --connect jdbc:mysql://localhost/HadoopProject --table Top5_pages --driver com.mysql.jdbc.Driver --export-dir /home/notroot/data/Top5_page/part-r-00000 --fields-terminated-by ',' --username root --password admin;

sqoop export --connect jdbc:mysql://localhost/HadoopProject --table Bottom5_pages --driver com.mysql.jdbc.Driver --export-dir /home/notroot/data/Bottom5_page/part-r-00000 --fields-terminated-by ',' --username root --password admin;


sqoop export --connect jdbc:mysql://localhost/HadoopProject --table Top5_users --driver com.mysql.jdbc.Driver --export-dir /home/notroot/data/Top5_user/part-r-00000 --fields-terminated-by ',' --username root --password admin;


sqoop export --connect jdbc:mysql://localhost/HadoopProject --table Bottom5_users --driver com.mysql.jdbc.Driver --export-dir /home/notroot/data/Bottom5_user/part-r-00000 --fields-terminated-by ',' --username root --password admin;


sqoop export --connect jdbc:mysql://localhost/HadoopProject --table TotalPageViews --driver com.mysql.jdbc.Driver --export-dir /home/notroot/data/TotalPageView/part-r-00000 --fields-terminated-by ',' --username root --password admin;


sqoop export --connect jdbc:mysql://localhost/HadoopProject --table Cat1wisePageViews --driver com.mysql.jdbc.Driver --export-dir /home/notroot/data/Cat1ByPageView/part-r-00000 --fields-terminated-by ',' --username root --password admin;
```

```
sqoop export --connect jdbc:mysql://localhost/HadoopProject --table Cat2wisePageViews --driver com.mysql.jdbc.Driver --export-dir /home/notroot/data/Cat2ByPageView/part-r-00000 --fields-terminated-by ',' --username root --password admin;


sqoop export --connect jdbc:mysql://localhost/HadoopProject --table UniquepageViews --driver com.mysql.jdbc.Driver --export-dir /home/notroot/data/UniqueViews/part-r-00000 --fields-terminated-by ',' --username root --password admin;


sqoop export --connect jdbc:mysql://localhost/HadoopProject --table Status_codeInfo  --driver com.mysql.jdbc.Driver --export-dir /home/notroot/data/Status_codeInfo/part-r-00000 --fields-terminated-by ',' --username root --password admin;
```

Execution: As too many commands has been executed, I have shown execution of one of them below.

```
20/06/01 07:26:10 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1591013476644_0063
20/06/01 07:26:12 INFO impl.YarnClientImpl: Submitted application application_1591013476644_0063
20/06/01 07:26:12 INFO mapreduce.Job: The url to track the job:
http://ubuntu:8088/proxy/application_1591013476644_0063/
20/06/01 07:26:12 INFO mapreduce.Job: Running job: job_1591013476644_0063
20/06/01 07:26:24 INFO mapreduce.Job: Job job_1591013476644_0063 running in uber mode : false
20/06/01 07:26:24 INFO mapreduce.Job:  map 0% reduce 0%
```

```
20/06/01 07:26:50 INFO mapreduce.Job:  map 100% reduce 0%
20/06/01 07:26:51 INFO mapreduce.Job: Job job_1591013476644_0063 completed successfully
20/06/01 07:26:51 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=546168
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1504
                HDFS: Number of bytes written=0
                HDFS: Number of read operations=16
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=0
        Job Counters
                Launched map tasks=4
                Data-local map tasks=4
                Total time spent by all maps in occupied slots (ms)=88913
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=88913
                Total vcore-milliseconds taken by all map tasks=88913
                Total megabyte-milliseconds taken by all map tasks=91046912
        Map-Reduce Framework
                Map input records=20
                Map output records=20
                Input split bytes=592
```

```
        Spilled Records=0
        Failed Shuffles=0
        Merged Map outputs=0
        GC time elapsed (ms)=1494
        CPU time spent (ms)=7700
        Physical memory (bytes) snapshot=710713344
        Virtual memory (bytes) snapshot=3329105920
        Total committed heap usage (bytes)=442499072
    File Input Format Counters
        Bytes Read=0
    File Output Format Counters
        Bytes Written=0
20/06/01 07:26:51 INFO mapreduce.ExportJobBase: Transferred 1.4688 KB in 47.4857 seconds (31.6727 bytes/sec)
20/06/01 07:26:51 INFO mapreduce.ExportJobBase: Exported 20 records.
```

Output: Data is stored in RDBMS's HadoopProject databases & tables are shown below.

notroot@ubuntu: ~

```
Database changed
mysql> show tables;
+--------------------------+
| Tables_in_HadoopProject  |
+--------------------------+
| Bottom5_Cat              |
| Bottom5_SubCat           |
| Bottom5_pages            |
| Bottom5_users            |
| Cat1wisePageViews        |
| Cat2wisePageViews        |
| PagesViewedperUser       |
| Status_codeInfo          |
| Top5_Cat                 |
| Top5_SubCat              |
| Top5_pages               |
| Top5_users               |
| TotalPageViews           |
| UniquepageViews          |
+--------------------------+
14 rows in set (0.00 sec)

mysql> select * from Bottom5_Cat;
+-----------------+-------+
| cat1            | Count |
+-----------------+-------+
| mobiles         | 15292 |
| tvs-audio       | 21222 |
| computers       | 20165 |
| catalogsearch   |   734 |
| digital-cameras | 10039 |
+-----------------+-------+
5 rows in set (0.12 sec)
```

notroot@ubuntu: ~

```
mysql> select * from Top5_Cat;
+-----------------+-------+
| cat1            | Count |
+-----------------+-------+
| digital-cameras | 10039 |
| computers       | 20165 |
| home-appliances | 32548 |
| tvs-audio       | 21222 |
| mobiles         | 15292 |
+-----------------+-------+
5 rows in set (0.04 sec)

mysql> select * from Top5_SubCat;
+-------------------------+-------+
| cat2                    | Count |
+-------------------------+-------+
| smart-phones            | 12311 |
| tablets                 |  9474 |
| air-conditioner-coolers |  9412 |
| laptops                 |  7132 |
| washing-machine         |  7318 |
+-------------------------+-------+
5 rows in set (0.11 sec)

mysql> select * from Bottom5_SubCat;
+---------------------+-------+
| cat2                | Count |
+---------------------+-------+
| blu-ray-dvd-players |  1518 |
| result              |   734 |
| desktops            |  1438 |
| monitors            |  2121 |
| home-theatres       |  1482 |
+---------------------+-------+
5 rows in set (0.00 sec)
```

notroot@ubuntu: ~

```
mysql> select * from Bottom5_pages;
+-----------------------------------------------------------------------------+-------+
| page                                                                        | Count |
+-----------------------------------------------------------------------------+-------+
| /digital-cameras/dslr-cameras/canon-eos-60d-dslr.html                       |   668 |
| /home-appliances/fans/usha-mist-air-ex-standard-pedestal-fan.html           |   637 |
| /home-appliances/geysers-276/ao-smith-hse-sbs-10-ltr-water-heater.html      |   657 |
| /computers/laptops/hp-15-d004tu-standard-laptop-15-6-inch-39-6-cm.html      |   669 |
| /computers/monitors/lg-19en33s-led-monitor-18-5-inch.html                   |   658 |
+-----------------------------------------------------------------------------+-------+
5 rows in set (0.00 sec)

mysql> select * from Bottom5_users;
+-----------------+-------+
| ip              | Count |
+-----------------+-------+
| 255.253.157.132 |     1 |
| 255.252.110.219 |     1 |
| 255.244.235.123 |     1 |
| 255.250.223.209 |     1 |
| 255.255.254.205 |     1 |
+-----------------+-------+
5 rows in set (0.00 sec)

mysql> select * from Top5_users;
+--------------+-------+
| ip           | Count |
+--------------+-------+
| 5.116.86.14  |     4 |
| 172.0.113.58 |     3 |
| 207.216.84.1 |     3 |
| 78.33.86.152 |     3 |
| 35.13.106.61 |     3 |
+--------------+-------+
5 rows in set (0.01 sec)
```

```
mysql> select * from Cat1wisePageViews;
+-----------------+-------+
| Cat1            | Count |
+-----------------+-------+
| digital-cameras | 10039 |
| computers       | 20165 |
| mobiles         | 15292 |
| catalogsearch   |   734 |
| home-appliances | 32548 |
| tvs-audio       | 21222 |
+-----------------+-------+
6 rows in set (0.00 sec)
```

```
mysql> select * from UniquepageViews LIMIT 5;
+-------------------------------------------------------------------------------------+-------------+
| Page                                                                                | UniqueViews |
+-------------------------------------------------------------------------------------+-------------+
| /tvs-audio/blu-ray-dvd-players/d-m-holdings-inc-denon-dbt-1713ud-blu-ray-player.html |         771 |
| /mobiles/feature-phones/nokia-series-40-feature-phone-black.html                    |         765 |
| /computers/tablets/amazon-kindle-fire-hd-tablet-32-gb-black.html                    |         763 |
| /mobiles/smart-phones/nokia-lumia-1320-windows-smart-phone-yellow.html              |         756 |
| /computers/monitors/lg-22ma33a-led-monitor-23-6-inch.html                           |         750 |
+-------------------------------------------------------------------------------------+-------------+
5 rows in set (0.00 sec)

mysql> select * from TotalPageViews LIMIT 5;
+-----------------------------------------------------------------------------------------------------+-------+
| page                                                                                                | Count |
+-----------------------------------------------------------------------------------------------------+-------+
| /mobiles/feature-phones/nokia-series-40-feature-phone-black.html                                    |   794 |
| /tvs-audio/blu-ray-dvd-players/d-m-holdings-inc-denon-dbt-1713ud-blu-ray-player.html                |   788 |
| /computers/tablets/amazon-kindle-fire-hd-tablet-32-gb-black.html                                    |   785 |
| /computers/laptops/hp-pavilion-10-touchsmart-10-e007au-standard-laptop-10-1-inch-25-6-cm.html       |   780 |
| /mobiles/smart-phones/nokia-lumia-1320-windows-smart-phone-yellow.html                              |   778 |
+-----------------------------------------------------------------------------------------------------+-------+
5 rows in set (0.00 sec)
```

notroot@ubuntu: ~

```
mysql> select * from PagesViewedperUser Order By PageViewed DESC LIMIT 10;
+------------------+-------------+
| ip               | PageViewed  |
+------------------+-------------+
| 5.116.86.14      |           4 |
| 78.33.86.152     |           3 |
| 207.216.84.1     |           3 |
| 35.13.106.61     |           3 |
| 172.0.113.58     |           3 |
| 191.107.34.219   |           3 |
| 217.219.120.208  |           3 |
| 130.2.117.192    |           3 |
| 69.45.134.109    |           3 |
| 13.72.188.203    |           3 |
+------------------+-------------+
10 rows in set (0.05 sec)
```

Conclusion: With this project, I have successfully demonstrated the power and usability of Hadoop framework & the solutions of Hadoop ecosystem. Various requirements are fulfilled with help of Pig and Sqoop platforms which are built on MapReduce framework. The code is tested on small dataset, no changes needed to use it for large datasets.