

Lab Assignment 1 Finding Facebook mutual friends using Map Reduce, Counting and Summing, Hive and Solr Queries.

Team Id: 4

Member 1: Nagababu Chilukuri

Class Id: 4

Member 2: Grace Sylvia

Class Id: 30

Member 3: PavanKumar Manchalla

Class Id: 16

Introduction.

This lab assignment deals with the implementation of MapReduce algorithm for finding Facebook common friends problem and Counting

and summing problem statement given and running the MapReduce job on Apache Hadoop. The second part deals with running the Hive and Solr queries for the given datasets.

Objective.

1. Implement MapReduce algorithm for finding Facebook common friends problem and run the MapReduce job on Apache Hadoop. Show your implementation through map-reduce diagram

Approach.

Take the input from the given use case A -> B C D,B -> A C D E,C -> A B D E,D -> A B C E,E -> B C D. For each user, iterate the friends and generate the keys of **them**. In the reducing phase, each intermediate key would get exactly two values. Compare the two values, that is their friend list, and leave the repeated.

WorkFlow.

Mapper.

In Mapper phase from the given input, iterate the friends himself/herself and generate the key,value pairs of each friends.

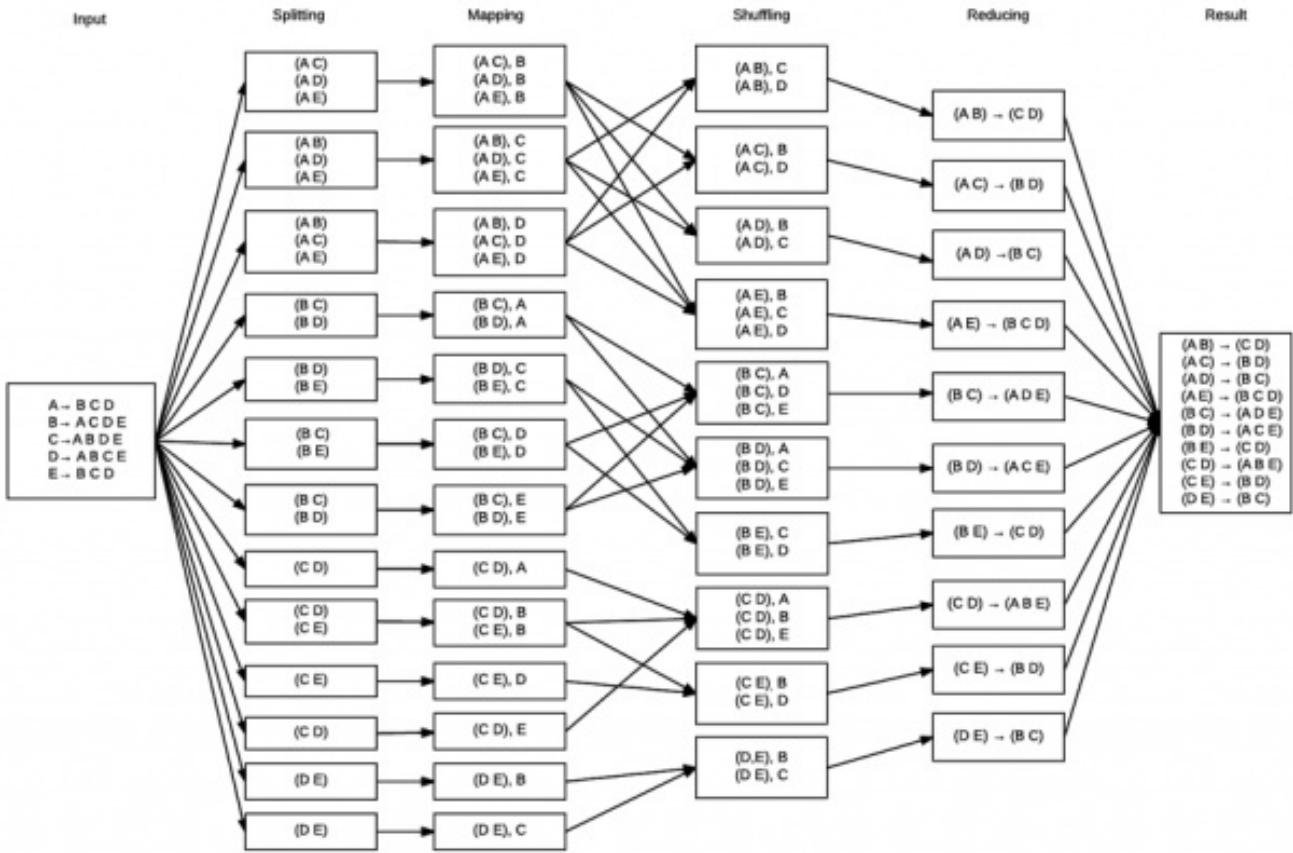
```
public static class Map extends MapReduceBase
    implements Mapper<LongWritable, Text, Text, Text>{
    public void map(LongWritable key, Text value, OutputCollector<Text, Text> output, Reporter reporter)
        throws IOException{
        StringTokenizer tokenizer = new StringTokenizer(value.toString(), "\n");
        String line = null;
        String[] lineArray = null;
        String[] friendArray = null;
        String[] tempArray = null;
        while(tokenizer.hasMoreTokens()){
            line = tokenizer.nextToken();
            lineArray = line.split(" : ");
            friendArray = lineArray[1].split(" ");
            tempArray = new String[2];
            for(int i = 0; i < friendArray.length; i++){
                tempArray[0] = friendArray[i];
                tempArray[1] = lineArray[0];
                Arrays.sort(tempArray);
                output.collect(new Text(tempArray[0] + " " + tempArray[1]), new Text(lineArray[1]));
            }
        }
    }
}
```

Reducer.

In reducer phase the key,value pairs of friends are grouped based on friends with same keys and then reduced to find the mutual friends.

```
public static class Reduce extends MapReduceBase
    implements Reducer<Text, Text, Text, Text>{
    public void reduce(Text key, Iterator<Text> values,
        OutputCollector<Text, Text> output, Reporter reporter) throws IOException{
        Text[] texts = new Text[2];
        int index = 0;
        while(values.hasNext()){
            texts[index++] = new Text(values.next());
        }
        String[] list1 = texts[0].toString().split(" ");
        String[] list2 = texts[1].toString().split(" ");
        List<String> list = new LinkedList<String>();
        for(String friend1 : list1){
            for(String friend2 : list2){
                if(friend1.equals(friend2)){
                    list.add(friend1);
                }
            }
        }
        StringBuffer sb = new StringBuffer();
        for(int i = 0; i < list.size(); i++){
            sb.append(list.get(i));
            if(i != list.size() - 1)
                sb.append(" ");
        }
        output.collect(key, new Text(sb.toString()));
    }
}
```

Algorithm



Conclusion

Input.

```

[cloudera@quickstart MutualFriend]$ cat input.txt
A : B C D
B : A C D E
C : A B D E
D : A B C E
E : B C D
[cloudera@quickstart MutualFriend]$ █
  
```

Output.

```
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=84
File Output Format Counters
  Bytes Written=78
A B      C D
A C      B D
A D      B C
B C      A D E
B D      A C E
B E      C D
C D      A B E
C E      B D
D E      B C
[cloudera@quickstart MutualFriend]$ cat input.txt
```

2. Implement MapReduce algorithm for finding Average and count problem and run the MapReduce job on Apache Hadoop. Show your implementation through map-reduce diagram

Workflow

In mapper phase the given input is processed and split into key,value pairs. In reducer phase grouping the occurrence of similar things and counting at the end of reducer phase.

Mapper

In mapper phase the input data is processed with "," delimiter and gives the first part as keys and second part as values which is send over

through the next phase for summation.

```
@Override
protected void map(LongWritable key, Text value, Context context) throws IOException, InterruptedException {
    String line = value.toString();
    String[] records = line.split(",");
    context.write(new Text(records[0].toString()), new IntWritable(Integer.parseInt(records[1])));
}
```

Combiner

In combiner phase the key value pairs obtained is processed with summation and counting average which is probably reduces the amount of data that need to be processed by reducer.

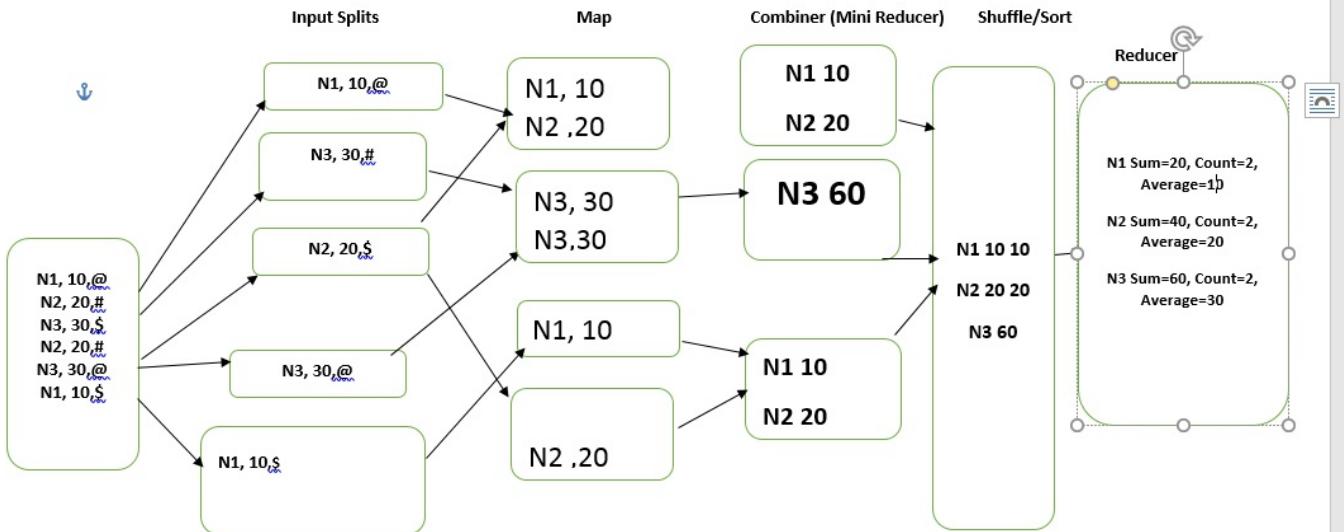
```
protected void reduce(Text key, Iterable<IntWritable> values, Context context) throws IOException, InterruptedException{
    int count = 0;
    int sum = 0;
    int inter_result = 0;
    for (IntWritable value : values){
        sum += value.get();
        count += 1;
    }
    inter_result = sum / count;
    context.write(key, new IntWritable(inter_result));
}
```

Reducer

Finally, in reducer phase the count, average obtained is processed and shows the output.

```
int count = 0;
int sum = 0;
int result = 0;
for(IntWritable value : values){
    sum += value.get();
    count += 1;
}
result = sum / count;
context.write(key, new IntWritable(result));
```

Algorithm



Conclusion

Input.

it	File Browser
	View as binary Edit file Download View file location Refresh Last modified 06/23/2019 5:59 PM User cloudera Group cloudera Size 92 B Mode 100644
	Home / user / cloudera / naxbergo / AverageCount / input / file.txt N1, 3, # N2, 6, @ N3, 8, \$ N1, 44, & N2, 97, @ N3, 88, # N1, 93, % N2, 68, @ N3, 89, \$ N1, 99, @ N2, 65, @ N3, 84, \$

Output.

The screenshot shows the Cloudera File Browser interface. The left sidebar contains navigation icons and a list of file details: Last modified (06/23/2019 6:06 PM), User (cloudera), Group (cloudera), Size (18 B), and Mode (100644). The main content area shows a file named 'part-r-00000' with three lines of text: 'N1, 59', 'N2, 59', and 'N3, 67'.

3. Create a Hive Table including Complex Data Types:

1. Creating hive table using zomato with complex data types as follows:

```
Time taken: 0.936 seconds
hive>
> create table zomato (restaurantid double,restaurantname string,country_code INT,city string,address array<string>,locality array<string>,locality_verbose array<string> ,longitude BIGINT,latitude BIGINT,cuisines array<string>,avgcos_t_2 int,currency string,hastable boolean,hasonline boolean,isdelinow boolean,switchtoorder boolean,pricerange int,aggregate int,ratingcolor string,ratingtext string,votes int) ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
> WITH SERDEPROPERTIES (
>   "separatorChar" = "\t",
>   "quoteChar" = "+",
>   "escapeChar" = "\\"
> )
> STORED AS TEXTFILE;
OK
Time taken: 0.486 seconds
hive> load data local inpath '/home/cloudera/Downloads/zomato.csv' into table zomato;
Loading data to table default.zomato
Table default.zomato stats: [numFiles=1, totalSize=2257316]
OK
Time taken: 3.87 seconds
hive> select * from zomato limited 10;
NoViableAltException(296@180:68: ( ( KW AS )? alias= Identifier )?[])
at org.antlr.runtime.DFA.noViableAlt(DFA.java:158)
at org.antlr.runtime.DFA.predict(DFA.java:116)
```

Here table is created using complex datatypes with serde properties to handle quotes and comma delimiter in zomato dataset.

2. Creating second table and upload it in the hive as follows:

```

hive> > create table crycode (country_code INT, Country string) ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
> WITH SERDEPROPERTIES (
>   "separatorChar" = "\t",
>   "quoteChar"     = "'",
>   "escapeChar"    = "\\"
> )
> STORED AS TEXTFILE;
OK
Time taken: 0.252 seconds
hive> load data local inpath '/home/cloudera/crycode.txt' into table crycode;
Loading data to table default.crycode
Table default.crycode stats: [numFiles=1, totalSize=360]
OK
Time taken: 0.833 seconds
hive> select * from crycode limit 5;
OK
Country_Code      Country
1                  India
14                 Australia
30                 Brazil
37                 Canada
Time taken: 0.269 seconds, Fetched: 5 row(s)
hive> ■

```

Use built-in functions in your queries:

3. `Max()` and `row_number()` are the 2 built in functions we have used in hive:

```

hive> select max(avgcost_2) from zomato where avgcost_2 IS NOT NULL;
Query ID = cloudera_20190622120202_ea87a5f5-a498-4c50-b0df-cf8f0c8223c8
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1560982050641_0043, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1560982050641_0043/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1560982050641_0043
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2019-06-22 12:03:25,281 Stage-1 map = 0%,  reduce = 0%
2019-06-22 12:04:20,919 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 6.28 sec
2019-06-22 12:05:10,661 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 13.13 sec
MapReduce Total cumulative CPU time: 13 seconds 130 msec
Ended Job = job_1560982050641_0043
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1  Cumulative CPU: 13.13 sec  HDFS Read: 2267008 HDFS Write: 13 SUCCESS
Total MapReduce CPU Time Spent: 13 seconds 130 msec
OK
@ayyolu
Time taken: 160.224 seconds. Fetched: 1 row(s)

```

4. The `row_number()` function is useful while handling sql query in database:

```

hive> select * from zomato limit 5;
OK
6000871 6000871 Sofras 208 Ankara "Emek Mahallesi Bosna Hersek Caddesi" No 22/C 6000871 Ankara" Emek "Emek Ankara" 32.8188333 39.91666667 "Kebab Izgara" 60
sh Lira(TL) No No No 1 6000871 Ankara" Emek "Emek Ankara" 32.8188333 39.91666667 "Kebab Izgara" 60
18222559 {Niche} - Cafe & Bar 1 New Delhi "2nd & 3rd Floor M-16 M Block Outer Circle Connaught Place New Delhi" Connaught Place "Connaught Place New Delhi" 77.2
8.6315156 "North Indian Chinese Italian Continental" 1500 Indian Rupees(Rs.) 2 7010939 wagamama 148 Wellington City "33 Customhouse Quay Wellington Central Wellington City 6011" Wellington Central "Wellington Central Wellington City" 174.7792237 -41.28303381 "Jap
Asian" 70 NewZealand(S) No No No 4 3 6001789 tashas 189 Cape Town "Ground Level Victoria Wharf V & A Waterfront Cape Town" V & A Waterfront "V & A Waterfront Cape Town" 18.421341 -33.902336 "Cafe Mediterrane
and(H) No No No 4 3 18301747 t Lounge by Dilmah 1 New Delhi "Flat 44 1st Floor Khan Market New Delhi" Khan Market "Khan Market New Delhi" 77.2271332 28.6010869 "Cafe Tea Des
00 Indian Rupees(Rs.) No No 5
Time taken: 0.242 seconds, Fetched: 5 row(s)
hive> ■

```

The rowid is column is inserted at the end of the column as follows:

```
FAILED: SemanticException [Error 10011]: Line 1:7 invalid function: ROWNU
hive> SELECT ROW_NUMBER() OVER () as row_num FROM zomato;
Query ID = cloudera_20190621192626_3118008a-0e29-4cf1-84bc-f2c0e64a1338
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1560982050641_0007, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1560982050641_0007/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1560982050641_0007
```

C. Perform 10 intuitive questions in Dataset (e.g.: pattern recognition, topic discussion, most important terms, etc.). Use your innovation to think out of box.

5. Pattern recognition is done as follows:

```
Time taken: 0.175 seconds, Fetched: 1 row(s)
hive> select 'Makati' LIKE '-a-';
FAILED: ParseException line 1:27 character '<EOF>' not supported here
hive> select 'Makati' LIKE '_a_';
OK
false
Time taken: 0.102 seconds, Fetched: 1 row(s)
hive> select 'Ooma' LIKE '__m__';
OK
true
Time taken: 0.205 seconds, Fetched: 1 row(s)
hive> 
```



6. Nullvalue deletion in the hive table:

```
hive> create table zomato5 as select NVL(restaurantid, '') as restaurantid, NVL(restaurantname, '') as restaurantname, NVL(country_code, '') as country_code, NVL(city, '') as city, NVL(address, '') as address, NVL(locality, '') as locality, NVL(verbose, '') as locality_verbose, NVL(longitude, '') as longitude, NVL(latitude, '') as latitude, NVL(cuisines, '') as cuisines, NVL(avgcost_2, '') as avgcost_2, NVL(currency, '') as currency, NVL(hasstable, '') as hasstable, NVL(hasonline, '') as hasonline, NVL(isdelinow, '') as isdelinow, NVL(switchtoorder, '') as switchtoorder, NVL(pricerange, '') as pricerange, NVL(aggregate, '') as aggregate, NVL(ratingcolor, '') as ratingcolor, NVL(ratingtext, '') as ratingtext, NVL(votes, '') as votes
query ID = cloudera_20190622002121_95e76bbe-b619-4401-b97f-af05113fe541
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1560982050641_0031, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1560982050641_0031/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1560982050641_0031
```

7. By selecting the notnull values and creating the rownumber we will delete most of the null values inside the table while doing

the table cleaning process:

```
FAILED: ParseException line 2:0 missing EOF at FROM at line 1
hive> SELECT address, COALESCE.aggregate,'Others' AS SO, COUNT(*)
   > FROM zomato limit 10
   > WHERE address IS NOT NULL AND aggregate IS NOT NULL
   > GROUP BY aggregate,address;
FAILED: ParseException line 3:0 missing EOF at 'WHERE' near '10'
hive> SELECT address, COALESCE.aggregate,'Others' AS SO, COUNT(*)
   > FROM zomato limit 10;
FAILED: SemanticException [Error 10025]: Line 1:7 Expression not in GROUP BY key 'address'
hive>
   > SELECT address, COALESCE.aggregate,'Others' AS SO, COUNT(*)
   > FROM zomato
   > WHERE address IS NOT NULL AND aggregate IS NOT NULL
   > GROUP BY aggregate,address limit 5;
Query ID = cloudera_20190622115555_0b89da4-d48b-47cf-9a06-cfd3ec2adaaa
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1560982050641_0042, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1560982050641_0042/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1560982050641_0042
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2019-06-22 11:55:55,043 Stage-1 map = 0%, reduce = 0%
2019-06-22 11:56:44,750 Stage-1 map = 67%, reduce = 0%, Cumulative CPU 7.27 sec
2019-06-22 11:56:49,062 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 9.34 sec
2019-06-22 11:57:37,789 Stage-1 map = 100%, reduce = 67%, Cumulative CPU 14.5 sec
2019-06-22 11:57:44,514 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 17.28 sec
MapReduce Total cumulative CPU time: 17 seconds 280 msec
Ended Job = job_1560982050641_0042
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 17.28 sec  HDFS Read: 2268545 HDFS Write: 135 SUCCESS
Total MapReduce CPU Time Spent: 17 seconds 280 msec
OK
"1st Floor      American      Others  2
"2      American      Others  1
"201-202      American      Others  1
"2526      American      Others  1
"2nd Floor      American      Others  2
```

8. By creating the distinct table so that the duplicate values are deleted in the table which is also a cleaning process:

```
hive> create table zomato2 as select distinct restaurantid, restaurantname, country_code, city, address, locality, locality_verbose, longitude, latitude, cuisines, avgcost_2, currency,
range, aggregate, ratingcolor, ratingtext, votes from zomato;
Query ID = cloudera_20190622123333_ccbf6ed4-9af6-4cc7-96c9-4039d1319df0
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1560982050641_0048, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1560982050641_0048/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1560982050641_0048
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2019-06-22 12:34:17,516 Stage-1 map = 0%, reduce = 0%
2019-06-22 12:35:05,892 Stage-1 map = 67%, reduce = 0%, Cumulative CPU 7.24 sec
2019-06-22 12:35:13,197 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 10.7 sec
2019-06-22 12:36:00,855 Stage-1 map = 100%, reduce = 68%, Cumulative CPU 16.04 sec
2019-06-22 12:36:06,380 Stage-1 map = 100%, reduce = 99%, Cumulative CPU 19.98 sec
2019-06-22 12:36:07,745 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 20.67 sec
MapReduce Total cumulative CPU time: 20 seconds 670 msec
Ended Job = job_1560982050641_0048
Moving data to: hdfs://quickstart.cloudera:8020/user/hive/warehouse/zomato2
Table default.zomato2 stats: [numFiles=1, numRows=10554, totalSize=1887391, rawDataSize=1876837]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 20.67 sec  HDFS Read: 4521351 HDFS Write: 1887470 SUCCESS
Total MapReduce CPU Time Spent: 20 seconds 670 msec
OK
Time taken: 162.953 seconds
hive>
```

9. Creating views in a database is always known for hiding certain

personal information so that everyone cannot see the hidden view in a database.

```
hives: [ ]
```

5 r

10. AS in the columns in the dataset are similar hence we perform joins as follows:

1. Equijoin:

```

hive> select * from crycode C join zomatoc2 C1 ON C.country_code = C1.country_code limit 10;
Query ID = cloudera_20190622132121_4df67b8a-d65d-420b-9b5a-b8c5dd22978a
Total jobs: 1
Execution log at: /tmp/cloudera/cloudera_20190622132121_4df67b8a-d65d-420b-9b5a-b8c5dd22978a.log
2019-06-22 01:22:24 Starting to launch local task to process map join; maximum memory = 1013645312
2019-06-22 01:22:36 Dump the side-table for tag: 0 with group count: 17 into file: /tmp/cloudera/3e1945cf-ale2-4b3b-ba00-25adbe5c8a68/hive_2019-06-22_13-21-49_391_6059897477515869597-1-local-10003/HashTable-Stage-3/MapJoin-
ile20--.hashtable
2019-06-22 01:22:36 Uploaded 1 File to: file:/tmp/cloudera/3e1945cf-ale2-4b3b-ba00-25adbe5c8a68/hive_2019-06-22_13-21-49_391_6059897477515869597-1-local-10003/HashTable-Stage-3/MapJoin-
ile20--.hashtable
2019-06-22 01:22:36 End of local task; Time Taken: 12.207 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1560982050641_0052, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1560982050641_0052/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1560982050641_0052
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
Map 0% | Reduce 0% | Spills: 0 | Failed: 0 | Failed Reduces: 0%, Reduce = 0%
2019-06-22 13:24:29,535 Stage-3 map = 100%, reduce = 0%, Cumulative CPU 5.94 sec
MapReduce Total cumulative CPU time: 5 seconds 940 msec
Ended Job = job_1560982050641_0052
MapReduce Jobs Launched:
Stage-Stage-3: Map: 1 Cumulative CPU: 5.94 sec HDFS Read: 15000 HDFS Write: 2019 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 940 msec

```

2.Right outer join:

```
time taken: 186.52 seconds, Fetched: 5 row(s)
live> select * from crycode c RIGHT JOIN zomatopl2 cl on c.country_code = cl.country_code limit 5;
Query ID = cloudera_20190622133333_abf145e8-aef9-4a33-b7bf-bf27d6aff4c6
Total jobs = 1
Execution log at: /tmp/cloudera/cloudera_20190622133333_abf145e8-aef9-4a33-b7bf-bf27d6aff4c6.log
2019-06-22 01:33:44 Starting to launch local task to process map join; maximum memory = 1013645312
2019-06-22 01:33:51 Dump the side-table for tag: @ with group count: 17 into file: /tmp/cloudera/3e1945cf-ale2-4b3b-ba08-25adbe5c8a68/hive_2019-06-22_13-33-10_988_203739835555410134-1-local-10003/HashTable-Stage-3/MapJoin-mapf
l13@...-hashtable:
2019-06-22 01:33:52 Uploaded 1 File to: file:/tmp/cloudera/3e1945cf-ale2-4b3b-ba08-25adbe5c8a68/hive_2019-06-22_13-33-10_988_203739835555410134-1-local-10003/HashTable-Stage-3/MapJoin-mapfile30...-hashtable (915 bytes)
2019-06-22 01:33:52 End of local task; Time Taken: 7.833 sec.
Execution completed successfully
4�pre-local task succeeded
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_15609802050641_0054, Tracking URL = http://quickstart.cloudera:8080/proxy/application_15609802050641_0054/
kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_15609802050641_0054
Addup job information for Stage-3: number of mappers: 1; number of reducers: 0
2019-06-22 13:34:46.087 Stage-3 map = 0%, reduce = 0%
2019-06-22 13:35:25.676 Stage-3 map = 100%, reduce = 0%, Cumulative CPU 4.57 sec
4�pre-Reduce Total cumulative CPU time: 4 seconds 570 msec
Ended Job = job_15609802050641_0054
4�pre-Reduce Jobs Launched:
Stage-Stage-3: Map: 1 Cumulative CPU: 4.57 sec HDFS Read: 14836 HDFS Write: 1222 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 570 msec
)K
NULL NULL Station Sector 29 "Gurgaon" "Huda City Centre Metro Station Sector 29 Gurgaon" "Huda City Centre Metro Station Sector 29 Gurgaon Gurgaon" 77.87272552 28.45938349 *
Italian Pizza" 800 Indian Rupees(Rs.) No Yes No No 2
l India 108800 Chokhi Dhani 1 Jaipur "Chokhi Dhani Village Resort 12 Mile Tonk Road Jaipur" "Chokhi Dhani Village Resort Tonk Road "Chokhi Dhani Village Resort Tonk Road
l Jaipur" 75.837333 26.766536 Rajasthan 1600 Indian Rupees(Rs.) No No No
l India 108305 The Forresta Kitchen & Bar 1 Jaipur "Devraj Niwas Near Moti Mahal Cinema Khasa Kothi Crossing Gopalbari Jaipur" "Devraj Niwas Bani Park" "Devraj Niwas Bani Park
l Jaipur" 75.7936639 26.9214108 "Continental Mexican Beverages Desserts North Indian
l India 108306 Replay 1 Jaipur "SB 57 5th Floor Ridhi Tower Opposite SMS Stadium Tonk Road Jaipur" Tonk Road Jaipur" Tonk Road Jaipur" 75.80689587 26.8923125 "Nort
l Indian Continental Chinese Italian Mexican" 1500
l India 108122 Tapri Central 1 Jaipur "B4 E 3rd Floor Surana Jewellers Opposite Central Park C Scheme Jaipur" C Scheme "C Scheme Jaipur" 75.81075322 26.90
\$18991 "Cafe Fast Food Street Food" 750 Indian Rupees(Rs.) No
Time taken: 137.292 seconds, Fetched: 5 row(s)
live>
```

Thus using the dataset of 9558 rows and 22 column including rowid we have cleaned the dataset and got the output for all important queries that is necessary for the corporate culture.

a. Create a Solr Collection including our own Field Types

b. Perform 10 intuitive questions in Dataset (e.g.: pattern recognition, topic discussion, most important terms, etc.). Use your innovation to think out of box.

c. Implement at least 5 nested

queries among the 10.

d. Record the time execution for the queries.

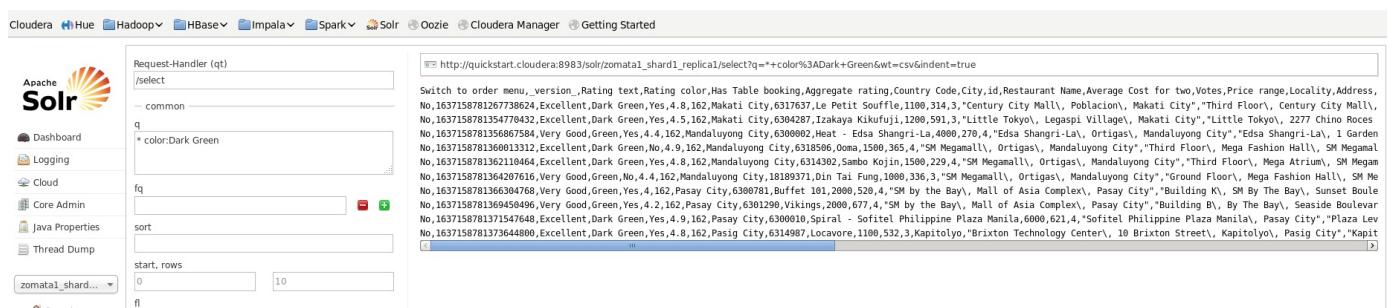
Dataset

A. Zomato Restaurants Data

<https://www.kaggle.com/shrutimehta/zomato-restaurants-data>

Queries

1. Number of rows with data color Green.



The screenshot shows the Apache Solr interface with the following search parameters:

- Request-Handler (qt): /select
- q: * color:Dark Green
- fq: (This field is empty)
- sort: (This field is empty)
- start, rows: 0, 10
- fl: (This field is empty)

The results page displays a list of restaurant entries, all of which have a 'color:Dark Green' rating. The results are as follows:

```
http://quickstart.cloudera:8983/solr/zomata1_shard1_replica1/select?qt=+color%3ADark+Green&wt=csv&indent=true

Switch to order menu, version_,Rating text,Rating color,Has Table booking,Aggregate rating,Country Code,City,id,Restaurant Name,Average Cost for two,Votes,Price range,Locality,Address, No,1637158781267738624,Excellent,Dark Green,Yes,4.8,162,Makati City,6317637,Le Petit Souffle,1100,314,3,"Century City Mall, Poblacion, Makati City","Third Floor", Century City Mall, No,1637158781354770432,Excellent,Dark Green,Yes,4.5,162,Makati City,6304287,Izakaya Kikufui,1200,591,3,"Little Tokyo, Legaspi Village, Makati City","Little Tokyo", 2277 Chino Roces No,1637158781356860584,Very Good,Green,Yes,4.4,162,Mandaluyong City,6308082,Heat - Edsa Shangri-La,4000,270,4,"Edsa Shangri-La, Ortigas, Mandaluyong City","Edsa Shangri-La", I Garden No,1637158781360013312,Excellent,Dark Green,No,4.9,162,Mandaluyong City,6318596,00ma,1580,365,4,"SM Megamall, Ortigas, Mandaluyong City","Third Floor, Mega Fashion Hall", SM Megamall No,1637158781362110464,Excellent,Dark Green,Yes,4.8,162,Mandaluyong City,6314302,Sambo Kojin,1580,229,4,"SM Megamall, Ortigas, Mandaluyong City","Third Floor, Mega Atrium, SM Megam No,1637158781364207616,Very Good,Green,No,4.4,162,Mandaluyong City,18189371,Dai Tai Fung,1000,336,3,"SM Megamall, Ortigas, Mandaluyong City","Ground Floor", Mega Fashion Hall, SM Me No,1637158781366304768,Very Good,Green,Yes,4.2,162,Pasay City,6308071,Buffet 101,2000,520,4,"SM by the Bay, Mall of Asia Complex, Pasay City","Building K, SM By The Bay, Sunset Boule No,1637158781369450496,Very Good,Green,Yes,4.2,162,Pasay City,6301290,Vikings,2000,677,4,"SM by the Bay, Mall of Asia Complex, Pasay City","Building B1, By The Bay, Seaside Boulevard No,1637158781371547648,Excellent,Dark Green,Yes,4.9,162,Pasay City,63000810,Spiral - Sofitel Philippine Plaza Manila,6000,621,4,"Sofitel Philippine Plaza Manila, Pasay City","Plaza Lev No,1637158781373644800,Excellent,Dark Green,Yes,4.8,162,Pasig City,6314987,Locavore,1100,532,3,Kapitolyo, "Brixton Technology Center, 10 Brixton Street, Kapitolyo, Pasig City","Kapit
```

Request-Handler (qt)

/select

— common —

q
* color:Dark Green

fq

sort

```
http://quickstart.cloudera:8983/solr/zomata1_shard1_replica1/select?q=*&wt=json
```

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 1,
    "params": {
      "indent": "true",
      "q": "* color:Dark Green",
      "_": "1561316693772",
      "wt": "json"
    }
  },
  "response": {
    "docs": [
      {
        "name": "Izakaya Kikufuji",
        "address": "1200, 591, 3, Makati City, Philippines",
        "rating": 4.5,
        "votes": 162,
        "average_cost": 1000
      },
      ...
    ]
  }
}
```

2. Average cost for two people greater than 1000.

http://quickstart.cloudera:8983/solr/zomata1_shard1_replica1/query

adoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started

Request-Handler (qt)

/select

— common —

q
* two:>1000

fq

sort

start, rows
0 10

fl

http://quickstart.cloudera:8983/solr/zomata1_shard1_replica1/select?q=*+two:>1000&wt=csv&indent=true

No,1637158781267738624,Excellent,Dark Green,Yes,4.8,162,Makati City,6317637,Le Petit Souffle,1100,314,3,"Century City Mall",Poblacion, Makati City,"Third Floor", Century City Mall, No,1637158781354770432,Excellent,Dark Green,Yes,4.5,162,Makati City,6304287,Izakaya Kikufuji,1200,591,3,"Little Tokyo", Legaspi Village, Makati City,"Little Tokyo", 2277 China Roces No,1637158781356867584,Very Good,Green,Yes,4.4,162,Mandaluyong City,6300002,Heat - Edsa Shangri-La, Ortigas, Mandaluyong City,"Edsa Shangri-La", 1 Garden No,1637158781360013312,Excellent,Dark Green,No,4.9,162,Mandaluyong City,6318506,00ma,1500,365,4,"SM Megamall", Ortigas, Mandaluyong City,"Third Floor", Mega Fashion Hall, SM Megamall No,1637158781362110464,Excellent,Dark Green,Yes,4.8,162,Mandaluyong City,6314302,Sambo Kojin,1500,229,4,"SM Megamall", Ortigas, Mandaluyong City,"Third Floor", Mega Atrium, SM Megam No,1637158781364207616,Very Good,Green,No,4.4,162,Mandaluyong City,18189371,Din Tai Fung,1000,336,3,"SM Megamall", Ortigas, Mandaluyong City,"Ground Floor", Mega Fashion Hall, SM Meg No,1637158781366304768,Very Good,Green,Yes,4.4,162,Pasay City,6300781,Buffer,101,2000,520,4,"SM by the Bay", Mall of Asia Complex, Pasay City,"Building KV", SM By The Bay, Sunset Boule No,1637158781371547648,Excellent,Dark Green,Yes,4.9,162,Pasay City,63001290,Vikings,2000,677,4,"SM by the Bay", Mall of Asia Complex, Pasay City,"Building BV", By The Bay, Seaside Boulevard No,1637158781373644800,Excellent,Dark Green,Yes,4.8,162,Pasig City,6314987,Locavore,1100,532,3,Kapitolyo,"Brixton Technology Center", 10 Brixton Street, Kapitolyo, Pasig City,"Kapit

Request-Handler (qt)

/select

— common —

q
* two:>1000

fq

sort

```
http://quickstart.cloudera:8983/solr/zomata1_shard1_replica1/select?q=*&wt=json
```

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 0,
    "params": {
      "indent": "true",
      "q": "* two:>1000",
      "_": "1561317012826",
      "wt": "json"
    }
  },
  "response": {
    "docs": [
      {
        "name": "Izakaya Kikufuji",
        "address": "1200, 591, 3, Makati City, Philippines",
        "rating": 4.5,
        "votes": 162,
        "average_cost": 1000
      },
      ...
    ]
  }
}
```

3. Number of rows with delivery yes.

Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started

Apache Solr

Request-Handler (qt) /select

common

q * delivery:Yes

fq

sort

start, rows 0 10

zomata1_shard1_replica1

id,City
6317637,Makati City
6304287,Makati City
6300002,Mandaluyong City
6318506,Mandaluyong City
6314302,Mandaluyong City
18189371,Mandaluyong City
6300781,Pasay City
6301290,Pasay City
6300010,Pasay City
6314987,Pasig City

Request-Handler (qt) /select

common

q * delivery:Yes

fq

sort

start, rows 0 10

fl id,City

df

Raw Query Parameters key1=val1&key2=val2

wt json

http://quickstart.cloudera:8983/solr/zomata1_shard1_replica1/select?q=*+delivery%3AYes&fl=id%2CCity&wt=json&indent=true

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 0,
    "params": {
      "fl": "id,City",
      "indent": "true",
      "q": "* delivery:Yes",
      "_": "1561317233694",
      "wt": "json"
    }
  },
  "response": {
    "numFound": 9551,
    "start": 0,
    "docs": [
      {
        "id": "6317637",
        "City": [
          "Makati City"
        ]
      },
      {
        "id": "6304287",
        "City": [
          "Makati City"
        ]
      }
    ]
  }
}
```

4. Number of rows with range two people 100 to 1000.

Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started

Apache Solr

Request-Handler (qt) /select

common

q * two:[100 TO 1000]

fq

sort

start, rows 0 10

fl

df

Raw Query Parameters key1=val1&key2=val2

wt json

indent

debugQuery

http://quickstart.cloudera:8983/solr/zomata1_shard1_replica1/select?q=*+two%3A%5B100+TO+1000%5D&wt=json&indent=true

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 18,
    "params": {
      "indent": "true",
      "q": "* two:[100 TO 1000]",
      "_": "1561317362862",
      "wt": "json"
    }
  },
  "response": {
    "numFound": 9551,
    "start": 0,
    "docs": [
      {
        "id": "6317637",
        "Restaurant Name": [
          "Le Petit Souffle"
        ],
        "Country Code": [
          "162"
        ],
        "City": [
          "Makati City"
        ],
        "Address": [
          "Third Floor, Century City Mall, Kalayaan Avenue, Poblacion, Makati City"
        ]
      }
    ]
  }
}
```

5. Number of rows with Juan City.

Request-Handler (qt)

```
/select
```

— common —

q
City:"Juan City"~1

fq

sort

start, rows
0 10

a

http://quickstart.cloudera:8983/solr/zomata1_shard1_replica1/select?q=City%3A%22Juan+City%22~1&wt=json&indent=true

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 0,
    "params": {
      "indent": "true",
      "q": "City:\"Juan City\"~1",
      "_": "1561318090096",
      "wt": "json"
    }
  },
  "response": {
    "numFound": 9551,
    "start": 0,
    "docs": [
      {
        "id": "1561318090096",
        "name": "Juan City"
      }
    ]
  }
}
```

Request-Handler (qt)

/select

— common

q

City="Juan City"~1

fq

sort

start, rows

0 10

fl

df

http://quickstart.cloudera:8983/solr/zomata1_shard1_replica1/select?qt=City%3A%22Juan+City%22-1&wt=csv&indent=true

Switch to order _version ,Rating text,Rating color,Has Table booking,Aggregate rating,Country Code,City,id,Restaurant Name,Average Cost for two,Votes,Price range,Locality,Address, No,1637158781267738624,Excellent,Dark Green,Yes,4.8,162,Makati City,6317637,The Peti Souffle,1106,3134,"Century City Mall,,Poblacion, Makati City","Third Floor,,Century City Mall,,No,1637158781354770432,Excellent,Dark Green,Yes,4.5,162,Makati City,6304287,Itazkaya Kikufuji,1200,591,3,"Little Tokyo,,Legaspi Village,,Makati City","Little Tokyo,,2277 Chino Roces No,1637158781356867584,Very Good,Green,Yes,4.4,162,Mandaluyong City,6308082,Heat EDSA Shangri-La,4086,270,4,"Edsa Shangri-La,,Ortigas,,Mandaluyong City,"Third Floor,,Mega Fashion Hall,,SM Megamall No,163715878136001312,Excellent,Dark Green,No,4.9,162,Mandaluyong City,6318506,00ma,1500,365,4,"SM Megamall,,Ortigas,,Mandaluyong City,"Third Floor,,Mega Fashion Hall,,SM Megamall No,1637158781362110464,Excellent,Dark Green,Yes,4.8,162,Mandaluyong City,6314302,Sambo Kojin,1508,229,4,"SM Megamall,,Ortigas,,Mandaluyong City,"Third Floor,,Mega Atrium,,SM Megamall No,1637158781364207616,Very Good,Green,No,4.4,162,Mandaluyong City,18189371,Dai Tun Fung,1000,336,3,"SM Megamall,,Ortigas,,Mandaluyong City,"Ground Floor,,Mega Fashion Hall,,SM Megamall No,163715878136360394768,Very Good,Green,Yes,4.162,Pasay City,6300781,Buffet 101,2000,520,4,"SM by the Bay,,Mall of Asia Complex,,Pasay City,"Building K,,SM By The Bay,,Sunset Boule No,1637158781369450496,Very Good,Green,Yes,4.2,162,Pasay City,6301290,Vikings,2006,677,4,"SM by the Bay,,Mall of Asia Complex,,Pasay City,"Building B,,By The Bay,,Seaside Boulevard No,1637158781371517468,Excellent,Dark Green,Yes,4.9,162,Pasay City,6300010,Spiral - Sofitel Philippine Plaza Manila,6000,621,4,"Sofitel Philippine Plaza Manila,,Pasay City,"Plaza Lev No,1637158781373644800,Excellent,Dark Green,Yes,4.8,162,Pasig City,6314987,Locavore,1100,532,3,Kapitolyo,,Brixton Technology Center,,10 Brixton Street,,Kapitolyo,,Pasig City,"Kapit

6. Count the Number of rows with delivery yes.

Request-Handler (qt)

/select

— common —

q

count(*) Has Table Booking=Yes

fq

-

sort

start, rows

10

fl

df

Raw Query Parameters

http://quickstart.cloudera:8983/solr/zomata1_shard1_replica1/select?qt=count(*)+&wt=json

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 1,
    "params": {
      "indent": "true",
      "q": "count(*) Has Table Booking=Yes",
      "wt": "json"
    }
  },
  "response": {
    "numFound": 9551,
    "start": 0,
    "docs": [
      {
        "id": "6317637",
        "Restaurant Name": [
          "Le Petit Souffle"
        ],
        "Country Code": [
          "162"
        ]
      }
    ]
  }
}
```

7. Selecting the rows with Human raced Male super heroes.

Apache Solr

Dashboard

Logging

Cloud

Core Admin

Java Properties

Thread Dump

lab15_shard1_...

Overview

/select

— common —

q

Gender:Male AND _query_:"Race:Human"

fq

-

sort

start, rows

10

fl

name,Gender,Race,Publisher, Alignment

df

```

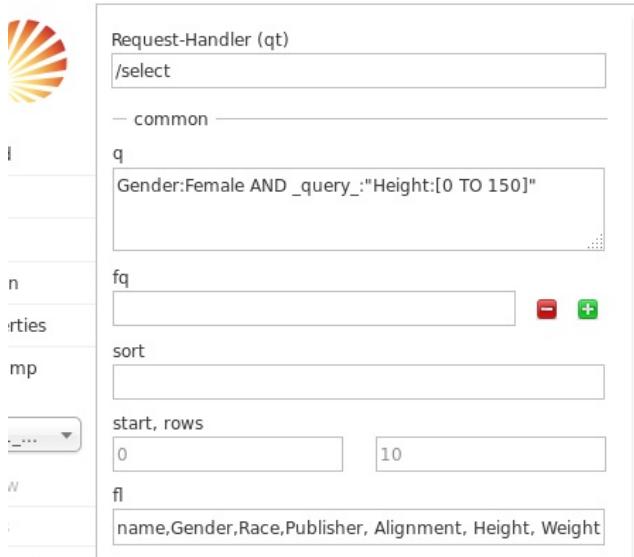
name,Gender,Race,Publisher,Alignment
A-Bomb,Male,Human,Marvel Comics,good
Absorbing Man,Male,Human,Marvel Comics,bad
Adam Strange,Male,Human,DC Comics,good
Agent Bob,Male,Human,Marvel Comics,good
Alex Mercer,Male,Human,Wildstorm,bad
Alfred Pennyworth,Male,Human,DC Comics,good
Ammo,Male,Human,Marvel Comics,bad
Animal Man,Male,Human,DC Comics,good
Ant-Man,Male,Human,Marvel Comics,good
Ant-Man II,Male,Human,Marvel Comics,good

```

Elapsed Time: 3msec.

8. Selecting the rows with 0<height<150 female super

heroes.

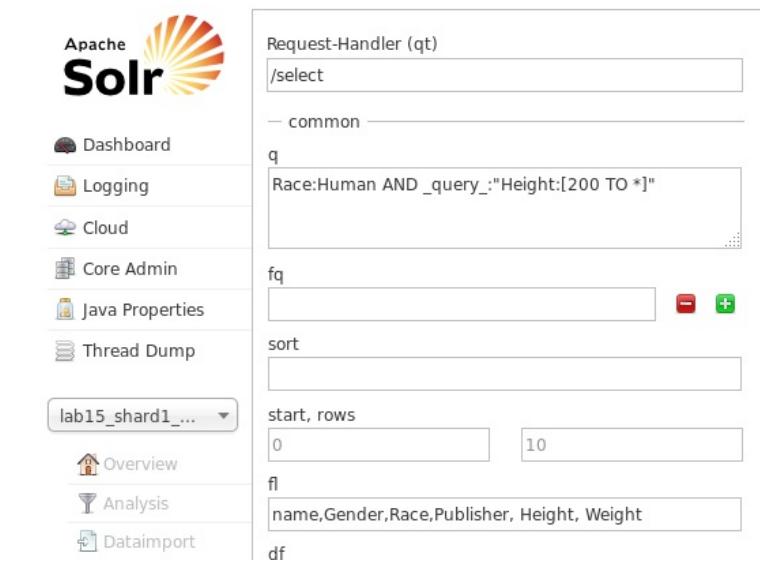


The screenshot shows the Apache Solr admin interface with a search query for female heroes. The query is: `Gender:Female AND _query_:"Height:[0 TO 150]"`. The results are:

```
name,Gender,Race,Publisher,Alignment,Height,Weight
Giganta,Female,-,DC Comics,bad,62.5,630.0
Violet Parr,Female,Human,Dark Horse Comics,good,137.0,41.0
```

Elapsed Time: 5msec.

9. Selecting the no of rows with Height>200 Human race super heroes

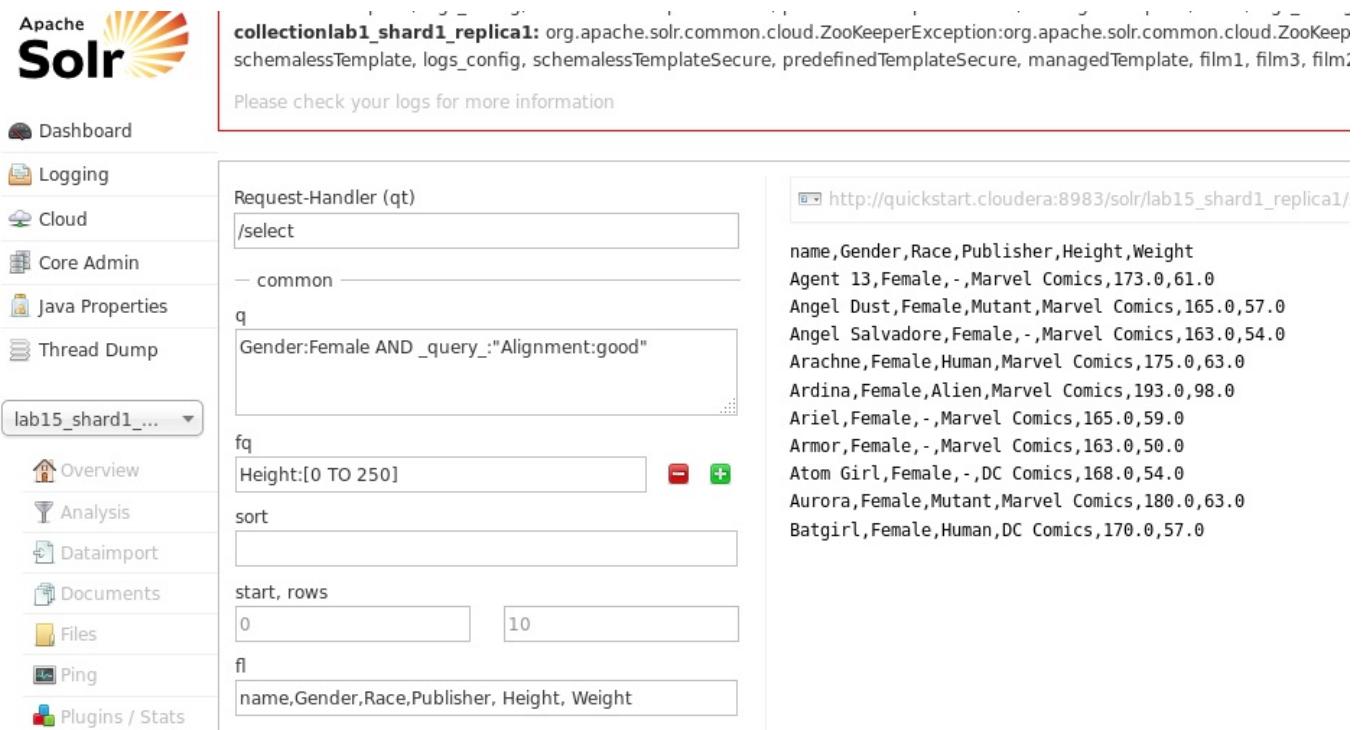


The screenshot shows the Apache Solr admin interface with a search query for human heroes with height > 200. The query is: `Race:Human AND _query_:"Height:[200 TO *]"`. The results are:

```
name,Gender,Race,Publisher,Height,Weight
A-Bomb,Male,Human,Marvel Comics,203.0,441.0
Ant-Man,Male,Human,Marvel Comics,211.0,122.0
Bane,Male,Human,DC Comics,203.0,180.0
Bloodaxe,Female,Human,Marvel Comics,218.0,495.0
Doctor Doom,Male,Human,Marvel Comics,201.0,187.0
Juggernaut,Male,Human,Marvel Comics,287.0,855.0
Kingpin,Male,Human,Marvel Comics,201.0,203.0
Lizard,Male,Human,Marvel Comics,203.0,230.0
Mr Incredible,Male,Human,Dark Horse Comics,201.0,158.0
Rey,Female,Human,George Lucas,297.0,-99.0
```

Elapsed Time: 10msec.

10. Selecting the no of rows with the super heroes who are females and having good



The screenshot shows the Apache Solr interface. On the left, there is a sidebar with various navigation options: Dashboard, Logging, Cloud, Core Admin, Java Properties, Thread Dump, and a dropdown menu for 'lab15_shard1_...'. Under 'lab15_shard1_...', there are links for Overview, Analysis, Dataimport, Documents, Files, Ping, and Plugins / Stats. The main area is titled 'Request-Handler (qt)' and contains a text input field with the value '/select'. Below this are sections for 'common', 'q', 'fq', 'sort', 'start, rows', and 'fl'. The 'q' section contains the query 'Gender:Female AND _query_:"Alignment:good"'. The 'fq' section contains the filter 'Height:[0 TO 250]'. The 'start, rows' section has '0' in the first input field and '10' in the second. The 'fl' section contains the fields 'name,Gender,Race,Publisher, Height, Weight'. To the right of the interface, there is a URL 'http://quickstart.cloudera:8983/solr/lab15_shard1_replica1/' and a list of superhero documents. The list includes:

name	Gender	Race	Publisher	Height	Weight
Agent 13	Female	-	Marvel Comics	173.0	61.0
Angel Dust	Female	Mutant	Marvel Comics	165.0	57.0
Angel Salvadore	Female	-	Marvel Comics	163.0	54.0
Arachne	Female	Human	Marvel Comics	175.0	63.0
Ardina	Female	Alien	Marvel Comics	193.0	98.0
Ariel	Female	-	Marvel Comics	165.0	59.0
Armor	Female	-	Marvel Comics	163.0	50.0
Atom Girl	Female	-	DC Comics	168.0	54.0
Aurora	Female	Mutant	Marvel Comics	180.0	63.0
Batgirl	Female	Human	DC Comics	170.0	57.0

Elapsed Time: 6msec.

11. Selecting the rows with the superheroes who are not having hair and from human race



Dashboard
Logging
Cloud
Core Admin
Java Properties
Thread Dump

/select

common

q
Hair\ color:"No Hair" AND _query_:"Race:Human"

fq

sort

start, rows
0 10

Publisher,id,Eye color,_version_,Weight,Gender,name,Hair color,Race,Height,Alignment,Skin color
Marvel Comics,0,yellow,1637178915704274944,441.0,Male,A-Bomb,No Hair,Human,203.0,good,-
Marvel Comics,5,blue,1637178915731537920,122.0,Male,Absorbing Man,No Hair,Human,193.0,bad,-
DC Comics,104,black,1637178915844784128,92.0,Male,Black Manta,No Hair,Human,188.0,bad,-
DC Comics,319,blue,1637178916141531136,81.0,Male,Heat Wave,No Hair,Human,180.0,bad,-
Marvel Comics,391,blue,1637178916301963264,203.0,Male,Kingpin,No Hair,Human,201.0,good,-
Marvel Comics,392,red,1637178916301963265,97.0,Male,Klaw,No Hair,Human,188.0,bad,red
DC Comics,405,green,1637178916328177665,95.0,Male,Lex Luthor,No Hair,Human,188.0,bad,-
Marvel Comics,412,red,1637178916336566272,230.0,Male,Lizard,No Hair,Human,203.0,bad,-
Marvel Comics,479,brown,1637178916459249664,79.0,Male,Mysterio,No Hair,Human,180.0,bad,-
Shueisha,502,-,1637178916492804096,69.0,Male,One Punch Man,No Hair,Human,175.0,good,-

Elapsed Time: 15msec.