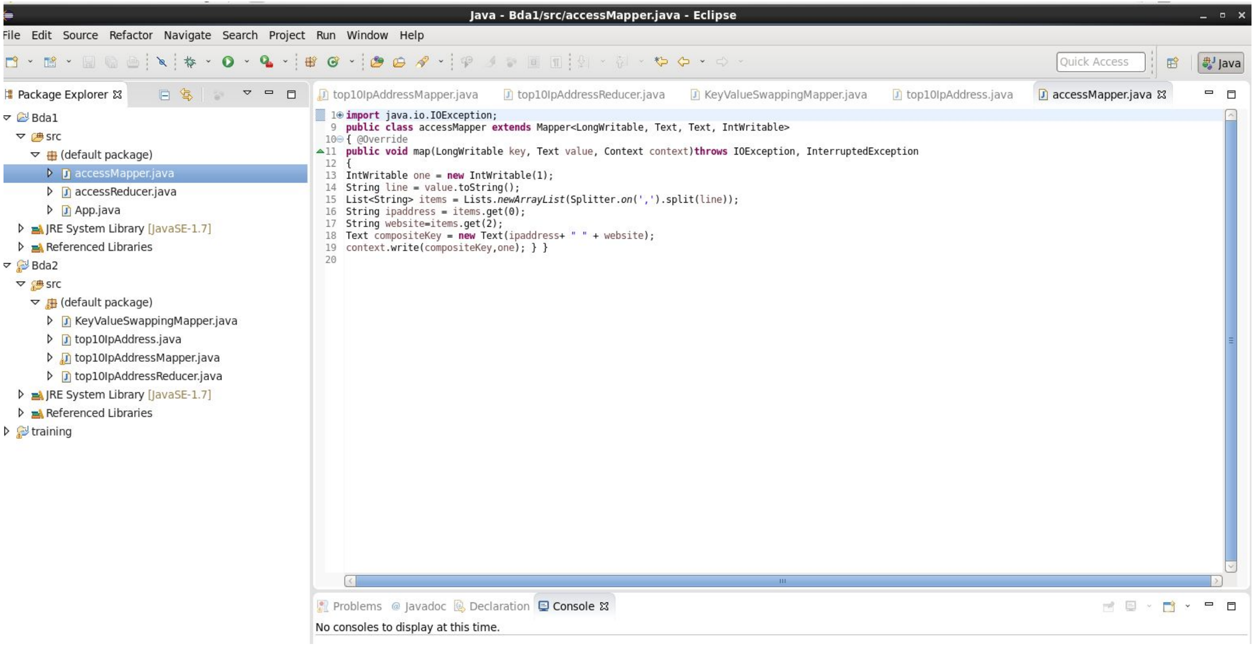
MapReduce:

Pre requisites:

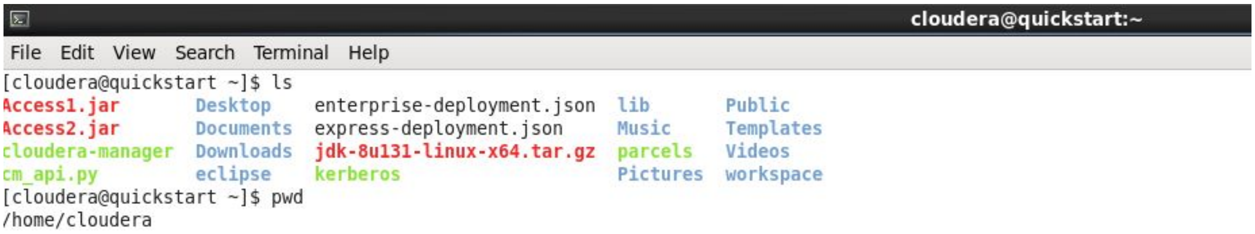
1) Create .jar files from the collection of java files and save it to /home/cloudera/


Java - Bda1/src/accessMapper.java - Eclipse
File Edit Source Refactor Navigate Search Project Run Window Help
Quick Access
Java
Package Explorer
Bda1
src
(default package)
accessMapper.java
accessReducer.java
App.java
JRE System Library [JavaSE-1.7]
Referenced Libraries
Bda2
src
(default package)
KeyValueSwappingMapper.java
top10IpAddress.java
top10IpAddressMapper.java
top10IpAddressReducer.java
JRE System Library [JavaSE-1.7]
Referenced Libraries
training
top10IpAddressMapper.java
top10IpAddressReducer.java
KeyValueSwappingMapper.java
top10IpAddress.java
accessMapper.java
Problems
Javadoc
Declaration
Console

```java
import java.io.IOException;
public class accessMapper extends Mapper<LongWritable, Text, Text, IntWritable>
{ @Override
public void map(LongWritable key, Text value, Context context)throws IOException, InterruptedException
{
IntWritable one = new IntWritable(1);
String line = value.toString();
List<String> items = Lists.newArrayList(Splitter.on(',').split(line));
String ipaddress = items.get(0);
String website=items.get(2);
Text compositeKey = new Text(ipaddress+ " " + website);
context.write(compositeKey,one); } }
```

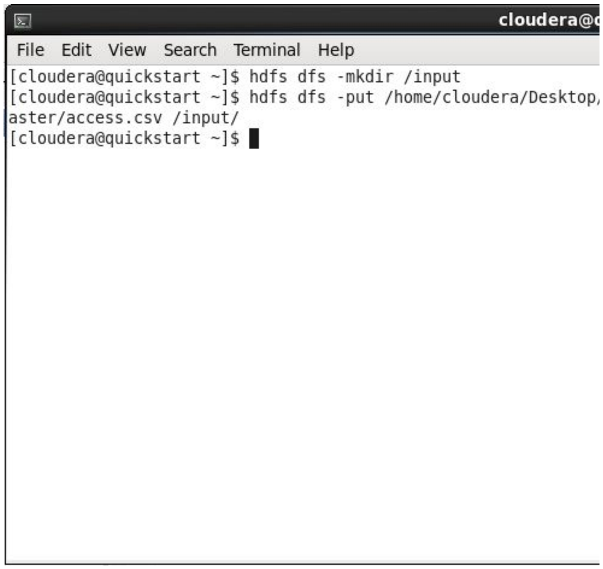
No consoles to display at this time.

2) Two .jar files (Access1.jar and Access2.jar) are saved in /home/cloudera



```
[cloudera@quickstart ~]$ ls
Access1.jar        Desktop     enterprise-deployment.json  lib       Public
Access2.jar        Documents   express-deployment.json     Music     Templates
cloudera-manager   Downloads   jdk-8u131-linux-x64.tar.gz  parcels   Videos
cm_api.py          eclipse     kerberos                    Pictures  workspace
[cloudera@quickstart ~]$ pwd
/home/cloudera
```

3) Make a input directory in hdfs dfs and move your dataset into it



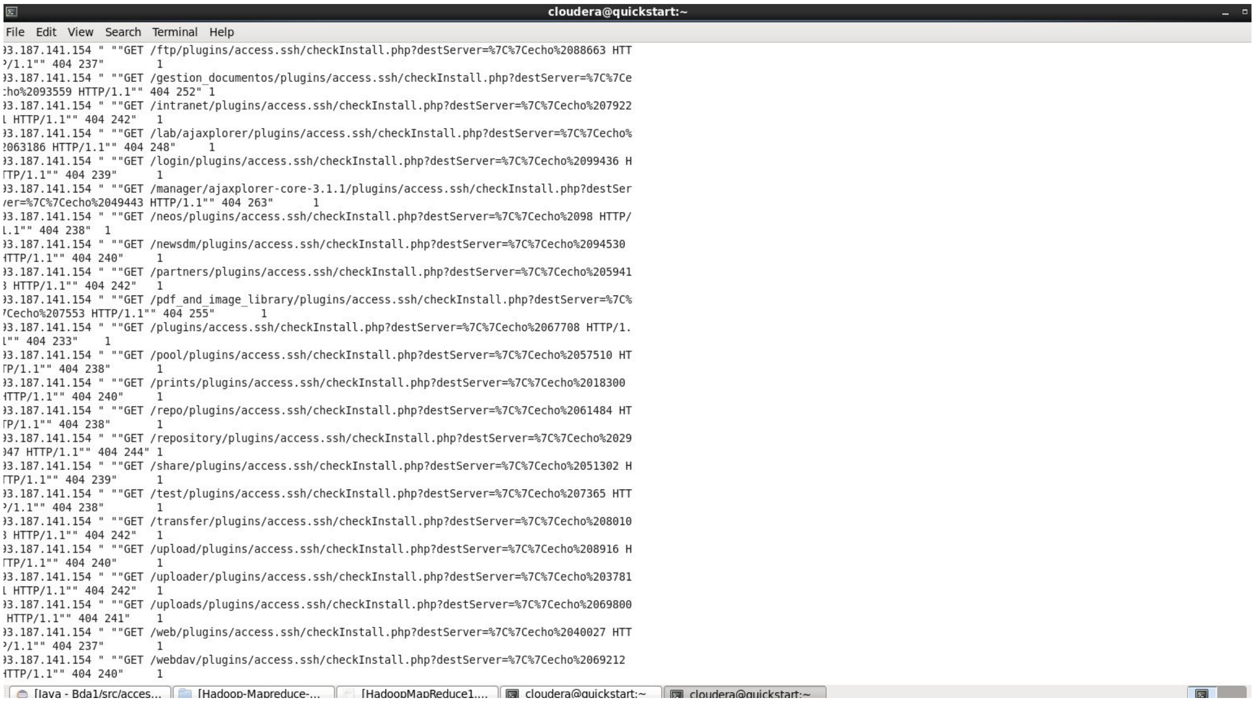
```
[cloudera@quickstart ~]$ hdfs dfs -mkdir /input
[cloudera@quickstart ~]$ hdfs dfs -put /home/cloudera/Desktop/
aster/access.csv /input/
[cloudera@quickstart ~]$ █
```

Analysis:

1) MapReduce in Hadoop to find the number of times each IP accessed the website.

Output:

Command: hadoop jar /home/cloudera/Access1.jar App /input/access.csv /out1

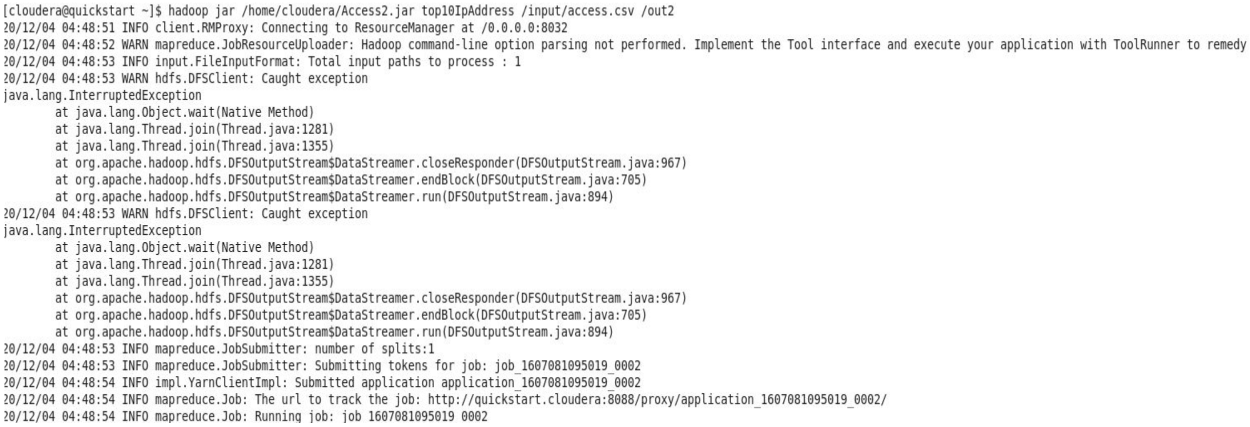


```
)3.187.141.154 " ""GET /ftp/plugins/access.ssh/checkInstall.php?destServer=%7C%7Cecho%2088663 HTT
>/1.1"" 404 237"        1
)3.187.141.154 " ""GET /gestion_documentos/plugins/access.ssh/checkInstall.php?destServer=%7C%7Ce
:ho%2093559 HTTP/1.1"" 404 252" 1
)3.187.141.154 " ""GET /intranet/plugins/access.ssh/checkInstall.php?destServer=%7C%7Cecho%207922
L HTTP/1.1"" 404 242"   1
)3.187.141.154 " ""GET /lab/ajaxplorer/plugins/access.ssh/checkInstall.php?destServer=%7C%7Cecho%
2063186 HTTP/1.1"" 404 248"     1
)3.187.141.154 " ""GET /login/plugins/access.ssh/checkInstall.php?destServer=%7C%7Cecho%2099436 H
TTP/1.1"" 404 239"      1
)3.187.141.154 " ""GET /manager/ajaxplorer-core-3.1.1/plugins/access.ssh/checkInstall.php?destSer
ver=%7C%7Cecho%2049443 HTTP/1.1"" 404 263"      1
)3.187.141.154 " ""GET /neos/plugins/access.ssh/checkInstall.php?destServer=%7C%7Cecho%2098 HTTP/
L.1"" 404 238"  1
)3.187.141.154 " ""GET /newsdm/plugins/access.ssh/checkInstall.php?destServer=%7C%7Cecho%2094530
HTTP/1.1"" 404 240"     1
)3.187.141.154 " ""GET /partners/plugins/access.ssh/checkInstall.php?destServer=%7C%7Cecho%205941
3 HTTP/1.1"" 404 242"   1
)3.187.141.154 " ""GET /pdf_and_image_library/plugins/access.ssh/checkInstall.php?destServer=%7C%
7Cecho%207553 HTTP/1.1"" 404 255"       1
)3.187.141.154 " ""GET /plugins/access.ssh/checkInstall.php?destServer=%7C%7Cecho%2067708 HTTP/1.
L"" 404 233"    1
)3.187.141.154 " ""GET /pool/plugins/access.ssh/checkInstall.php?destServer=%7C%7Cecho%2057510 HT
TP/1.1"" 404 238"       1
)3.187.141.154 " ""GET /prints/plugins/access.ssh/checkInstall.php?destServer=%7C%7Cecho%2018300
HTTP/1.1"" 404 240"     1
)3.187.141.154 " ""GET /repo/plugins/access.ssh/checkInstall.php?destServer=%7C%7Cecho%2061484 HT
TP/1.1"" 404 238"       1
)3.187.141.154 " ""GET /repository/plugins/access.ssh/checkInstall.php?destServer=%7C%7Cecho%2029
)47 HTTP/1.1"" 404 244" 1
)3.187.141.154 " ""GET /share/plugins/access.ssh/checkInstall.php?destServer=%7C%7Cecho%2051302 H
TTP/1.1"" 404 239"      1
)3.187.141.154 " ""GET /test/plugins/access.ssh/checkInstall.php?destServer=%7C%7Cecho%207365 HTT
>/1.1"" 404 238"        1
)3.187.141.154 " ""GET /transfer/plugins/access.ssh/checkInstall.php?destServer=%7C%7Cecho%208010
3 HTTP/1.1"" 404 242"   1
)3.187.141.154 " ""GET /upload/plugins/access.ssh/checkInstall.php?destServer=%7C%7Cecho%208916 H
TTP/1.1"" 404 240"      1
)3.187.141.154 " ""GET /uploader/plugins/access.ssh/checkInstall.php?destServer=%7C%7Cecho%203781
L HTTP/1.1"" 404 242"   1
)3.187.141.154 " ""GET /uploads/plugins/access.ssh/checkInstall.php?destServer=%7C%7Cecho%2069800
 HTTP/1.1"" 404 241"    1
)3.187.141.154 " ""GET /web/plugins/access.ssh/checkInstall.php?destServer=%7C%7Cecho%2040027 HTT
>/1.1"" 404 237"        1
)3.187.141.154 " ""GET /webdav/plugins/access.ssh/checkInstall.php?destServer=%7C%7Cecho%2069212
HTTP/1.1"" 404 240"     1
```

2) Top 10 most visited IP addresses

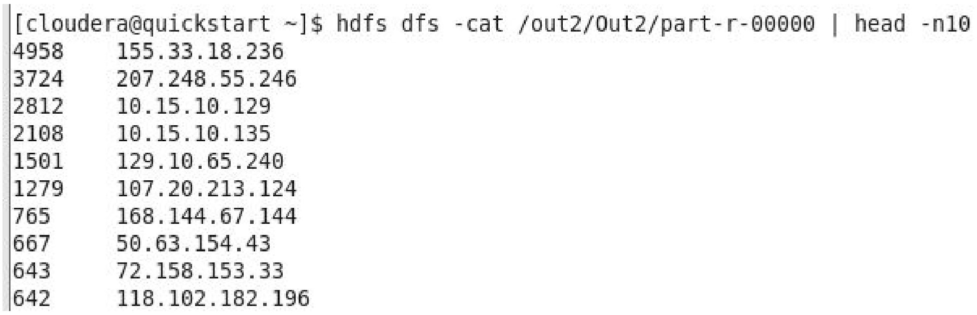
Output:

Command: hadoop jar /home/cloudera/Access2.jar top10IpAddress /input/access.csv /out2



```
[cloudera@quickstart ~]$ hadoop jar /home/cloudera/Access2.jar top10IpAddress /input/access.csv /out2
20/12/04 04:48:51 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
20/12/04 04:48:52 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy
20/12/04 04:48:53 INFO input.FileInputFormat: Total input paths to process : 1
20/12/04 04:48:53 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
        at java.lang.Object.wait(Native Method)
        at java.lang.Thread.join(Thread.java:1281)
        at java.lang.Thread.join(Thread.java:1355)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:967)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:705)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:894)
20/12/04 04:48:53 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
        at java.lang.Object.wait(Native Method)
        at java.lang.Thread.join(Thread.java:1281)
        at java.lang.Thread.join(Thread.java:1355)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:967)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:705)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:894)
20/12/04 04:48:53 INFO mapreduce.JobSubmitter: number of splits:1
20/12/04 04:48:53 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1607081095019_0002
20/12/04 04:48:54 INFO impl.YarnClientImpl: Submitted application application_1607081095019_0002
20/12/04 04:48:54 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1607081095019_0002/
20/12/04 04:48:54 INFO mapreduce.Job: Running job: job_1607081095019_0002
```

Top 10 most visited IP addresses

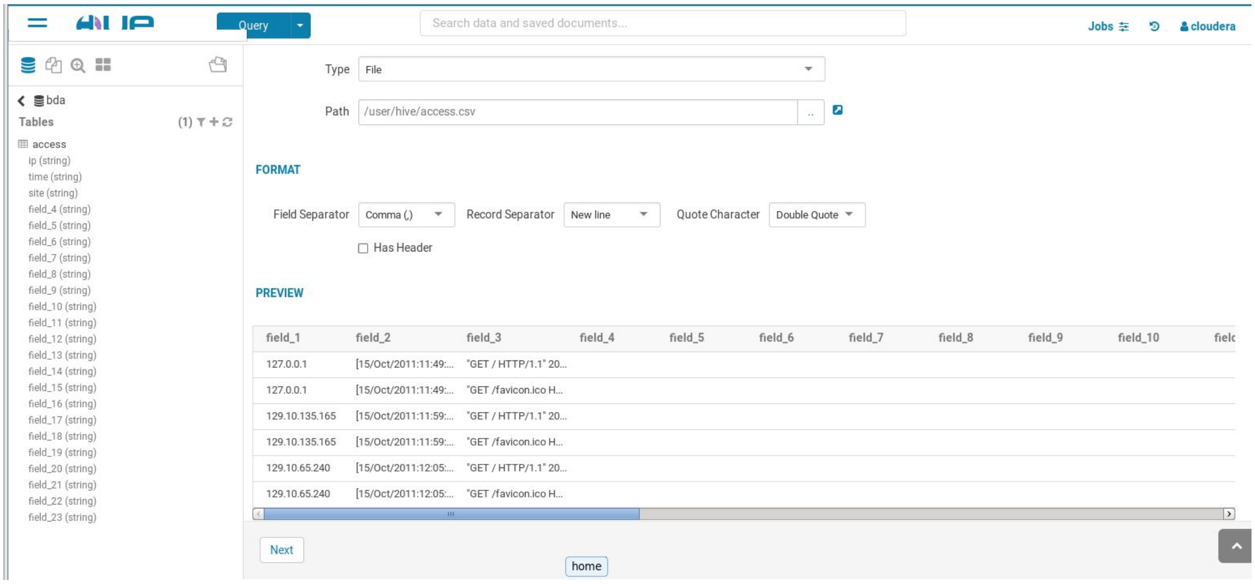


```
[cloudera@quickstart ~]$ hdfs dfs -cat /out2/Out2/part-r-00000 | head -n10
4958    155.33.18.236
3724    207.248.55.246
2812    10.15.10.129
2108    10.15.10.135
1501    129.10.65.240
1279    107.20.213.124
765     168.144.67.144
667     50.63.154.43
643     72.158.153.33
642     118.102.182.196
```

HIVE:

Prerequisites :

1) Upload the dataset into hive by specifying the data types and column names.



bda
Tables (1)

access
ip (string)
time (string)
site (string)
field_4 (string)
field_5 (string)
field_6 (string)
field_7 (string)
field_8 (string)
field_9 (string)
field_10 (string)
field_11 (string)
field_12 (string)
field_13 (string)
field_14 (string)
field_15 (string)
field_16 (string)
field_17 (string)
field_18 (string)
field_19 (string)
field_20 (string)
field_21 (string)
field_22 (string)
field_23 (string)

Type: File

Path: /user/hive/access.csv

FORMAT

Field Separator: Comma (,) Record Separator: New line Quote Character: Double Quote

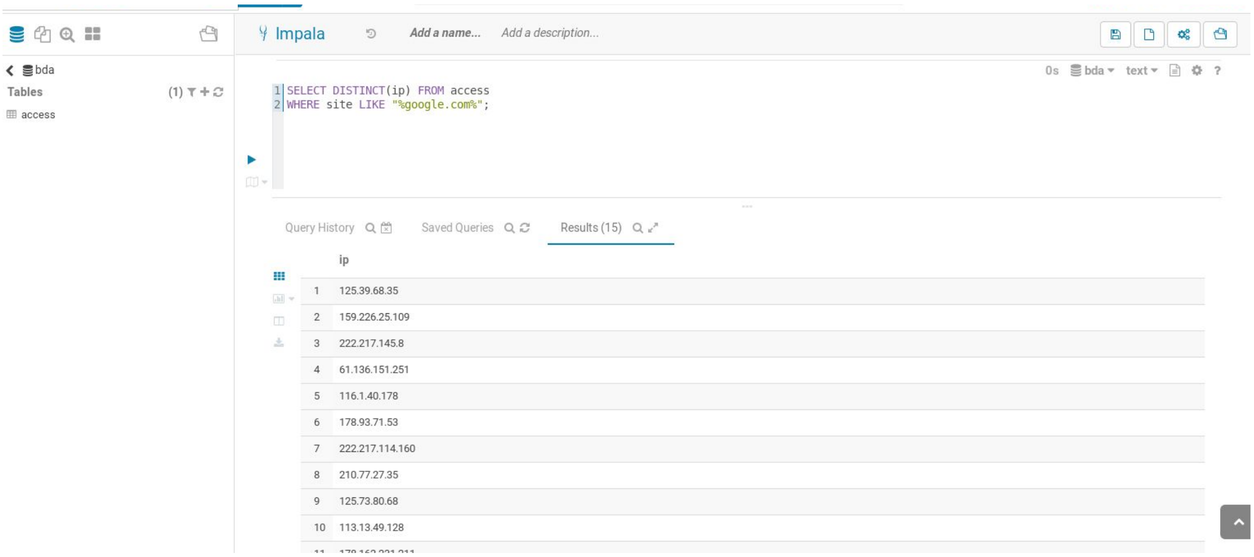
☐ Has Header

PREVIEW

field_1	field_2	field_3	field_4	field_5	field_6	field_7	field_8	field_9	field_10	fielc
127.0.0.1	[15/Oct/2011:11:49:...	"GET / HTTP/1.1" 20...								
127.0.0.1	[15/Oct/2011:11:49:...	"GET /favicon.ico H...								
129.10.135.165	[15/Oct/2011:11:59:...	"GET / HTTP/1.1" 20...								
129.10.135.165	[15/Oct/2011:11:59:...	"GET /favicon.ico H...								
129.10.65.240	[15/Oct/2011:12:05:...	"GET / HTTP/1.1" 20...								
129.10.65.240	[15/Oct/2011:12:05:...	"GET /favicon.ico H...								

Analysis:

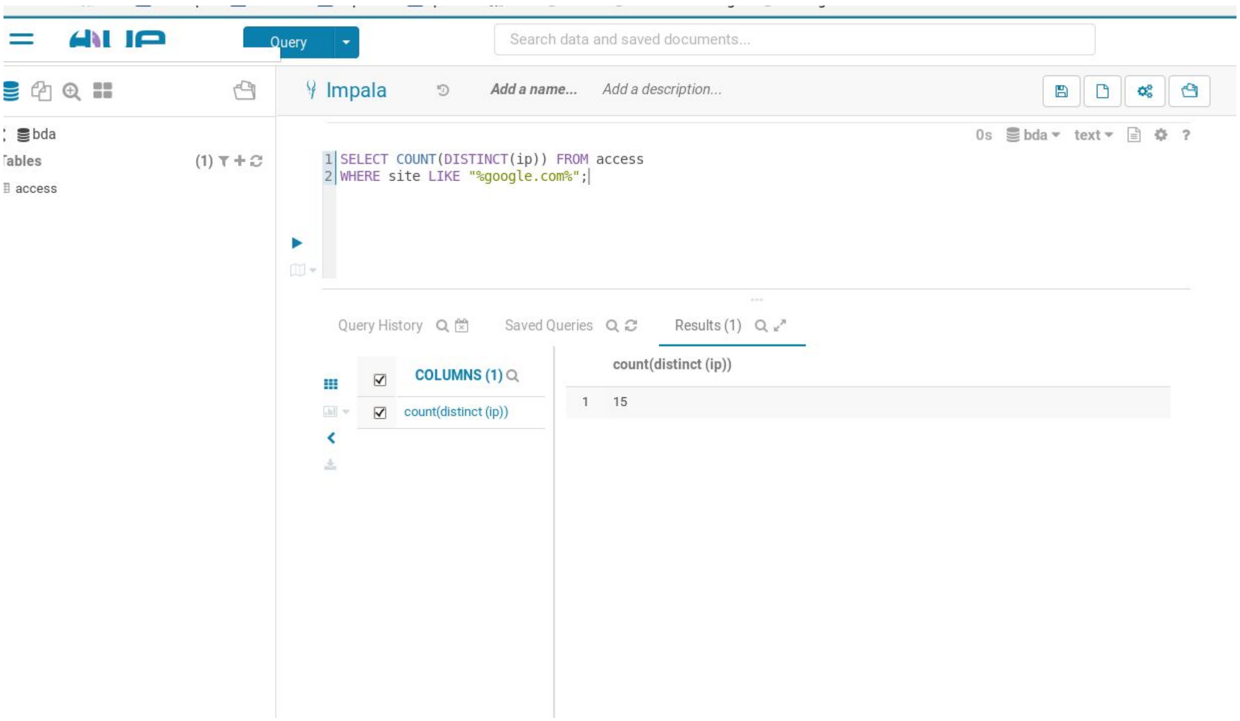
1) Ip users that visited “google.com”



```
SELECT DISTINCT(ip) FROM access
WHERE site LIKE "%google.com%";
```

	ip
1	125.39.68.35
2	159.226.25.109
3	222.217.145.8
4	61.136.151.251
5	116.1.40.178
6	178.93.71.53
7	222.217.114.160
8	210.77.27.35
9	125.73.80.68
10	113.13.49.128
11	178.162.221.211

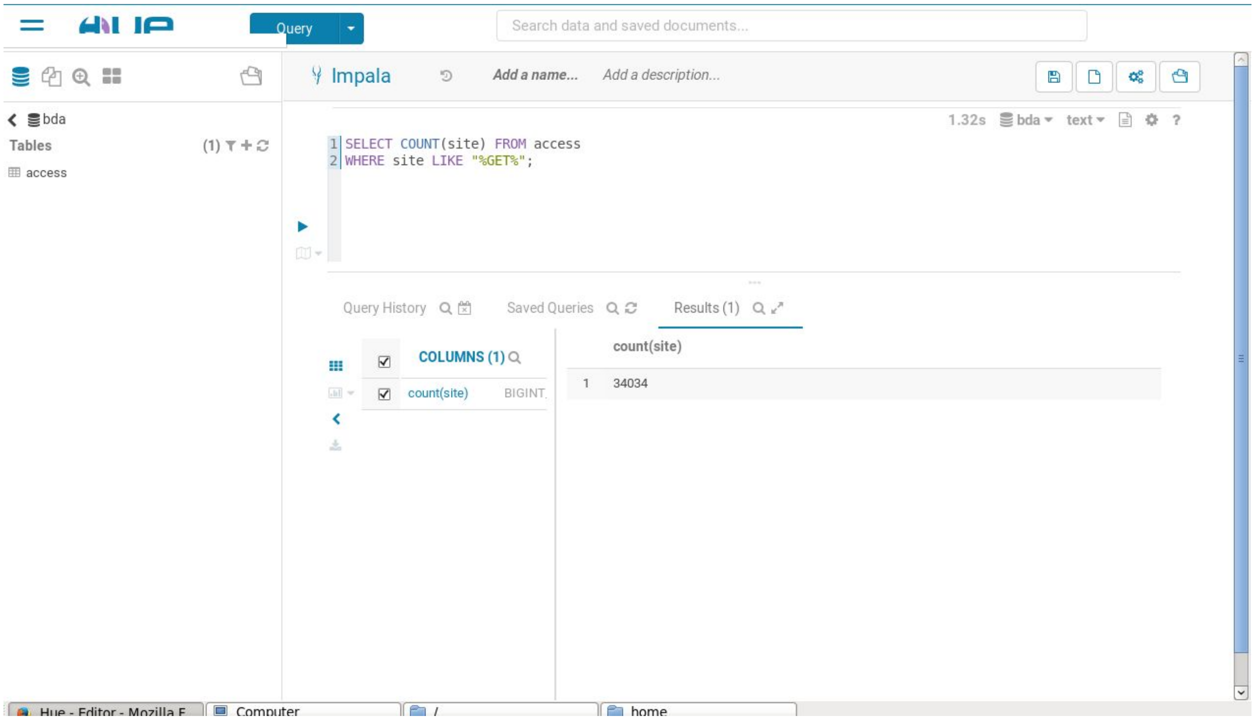
2) Number of users that visited “google.com”



```
SELECT COUNT(DISTINCT(ip)) FROM access
WHERE site LIKE "%google.com%";
```

	count(distinct (ip))
1	15

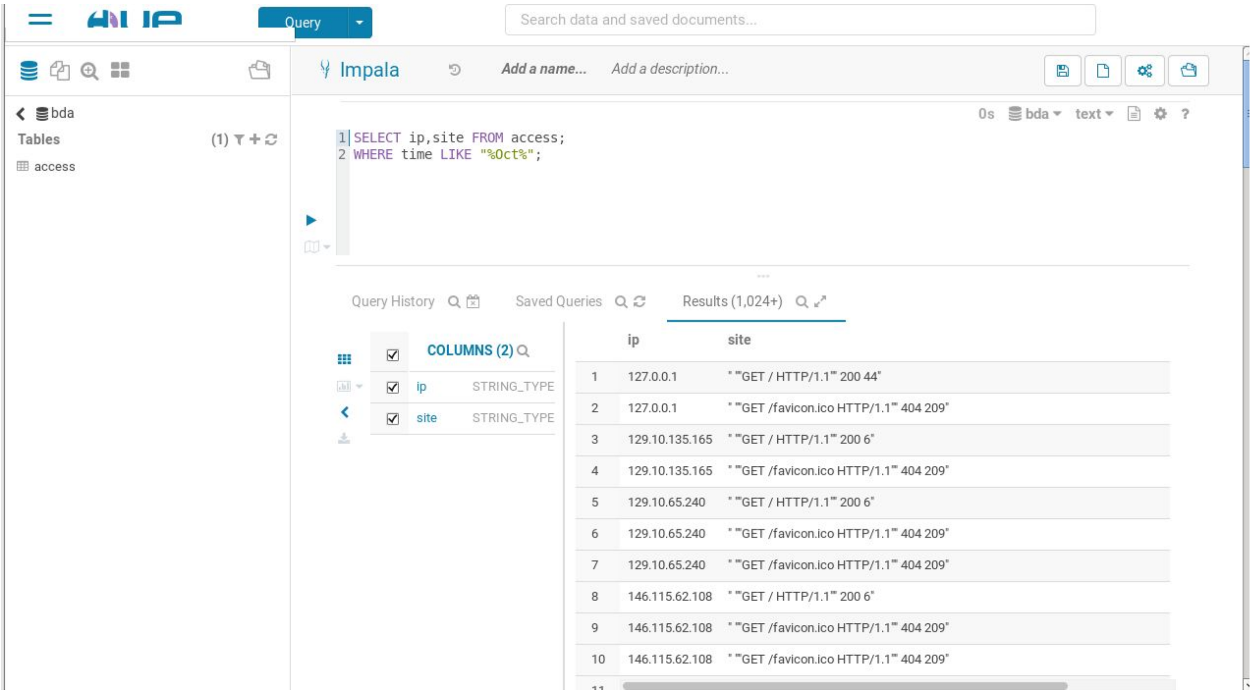
3) Number of times GET method was used



```
SELECT COUNT(site) FROM access
WHERE site LIKE "%GET%";
```

	count(site)
1	34034

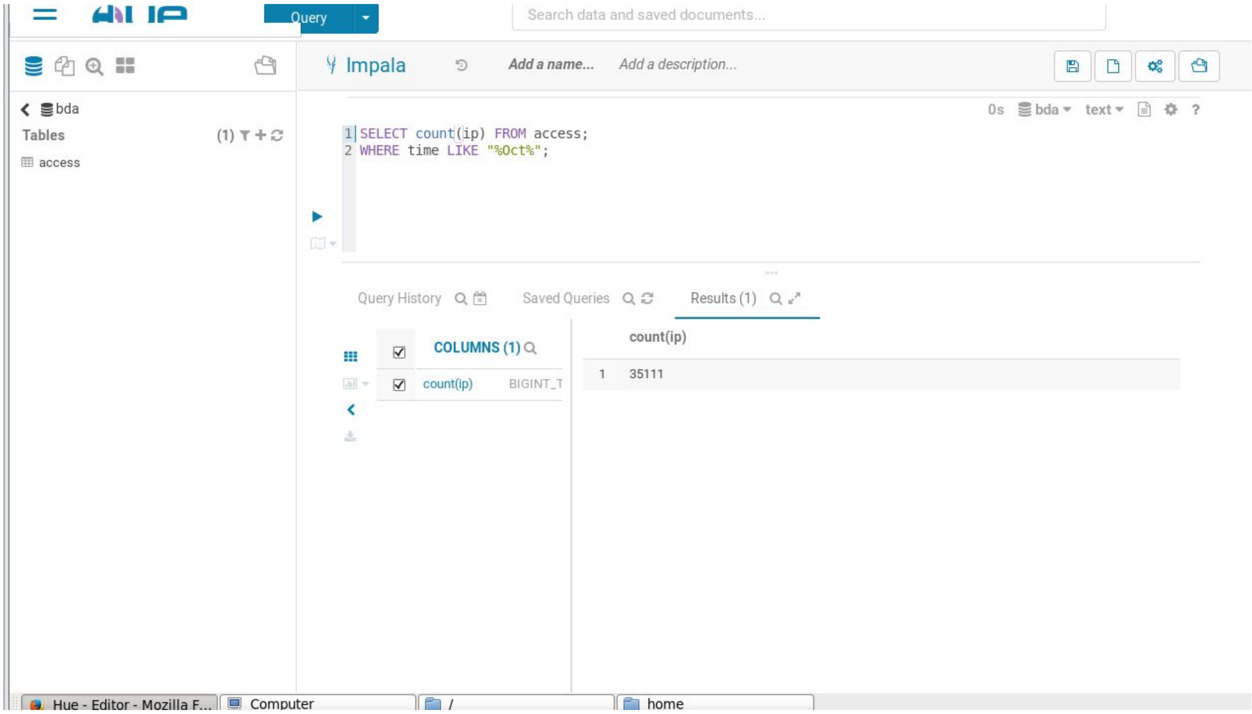
1) Sites that were visited in October



```
SELECT ip,site FROM access;
WHERE time LIKE "%Oct%";
```

	ip	site
1	127.0.0.1	" ""GET / HTTP/1.1""" 200 44"
2	127.0.0.1	" ""GET /favicon.ico HTTP/1.1""" 404 209"
3	129.10.135.165	" ""GET / HTTP/1.1""" 200 6"
4	129.10.135.165	" ""GET /favicon.ico HTTP/1.1""" 404 209"
5	129.10.65.240	" ""GET / HTTP/1.1""" 200 6"
6	129.10.65.240	" ""GET /favicon.ico HTTP/1.1""" 404 209"
7	129.10.65.240	" ""GET /favicon.ico HTTP/1.1""" 404 209"
8	146.115.62.108	" ""GET / HTTP/1.1""" 200 6"
9	146.115.62.108	" ""GET /favicon.ico HTTP/1.1""" 404 209"
10	146.115.62.108	" ""GET /favicon.ico HTTP/1.1""" 404 209"
11		

5) Number of sites visited in October

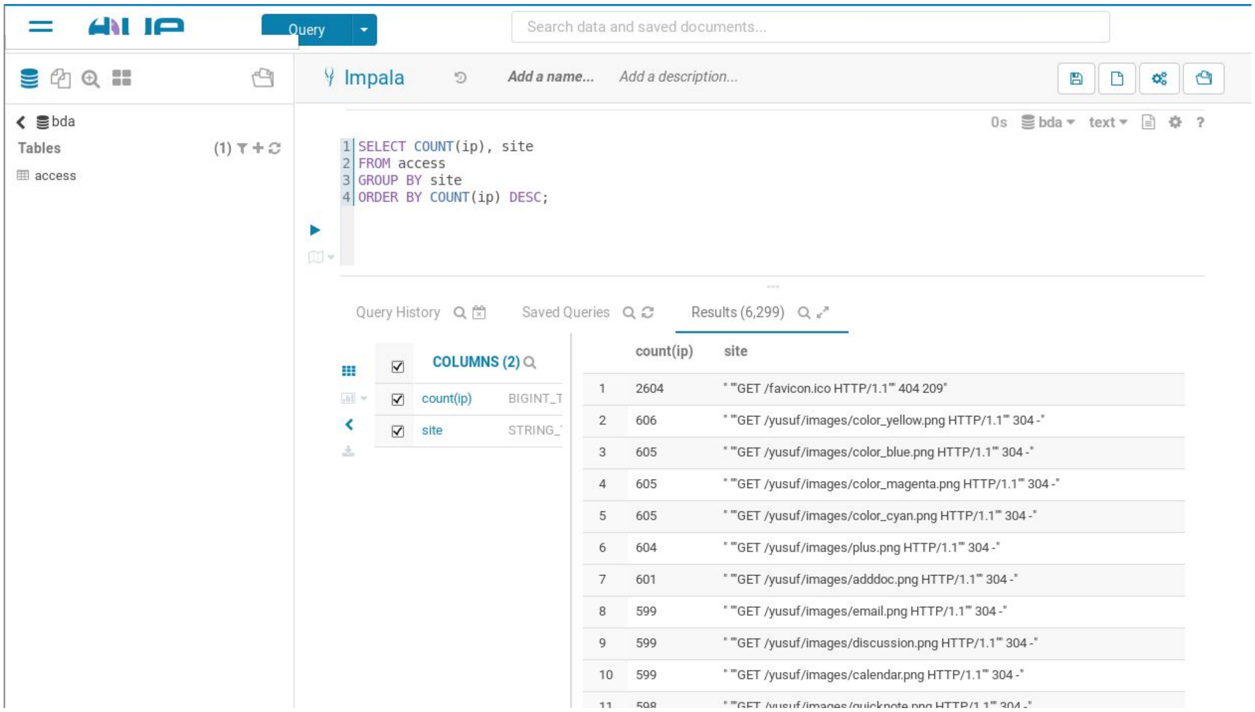


```
SELECT count(ip) FROM access;
WHERE time LIKE "%Oct%";
```

☑	COLUMNS (1)	
☑	count(ip)	BIGINT_T

	count(ip)
1	35111

6) Most visited sites

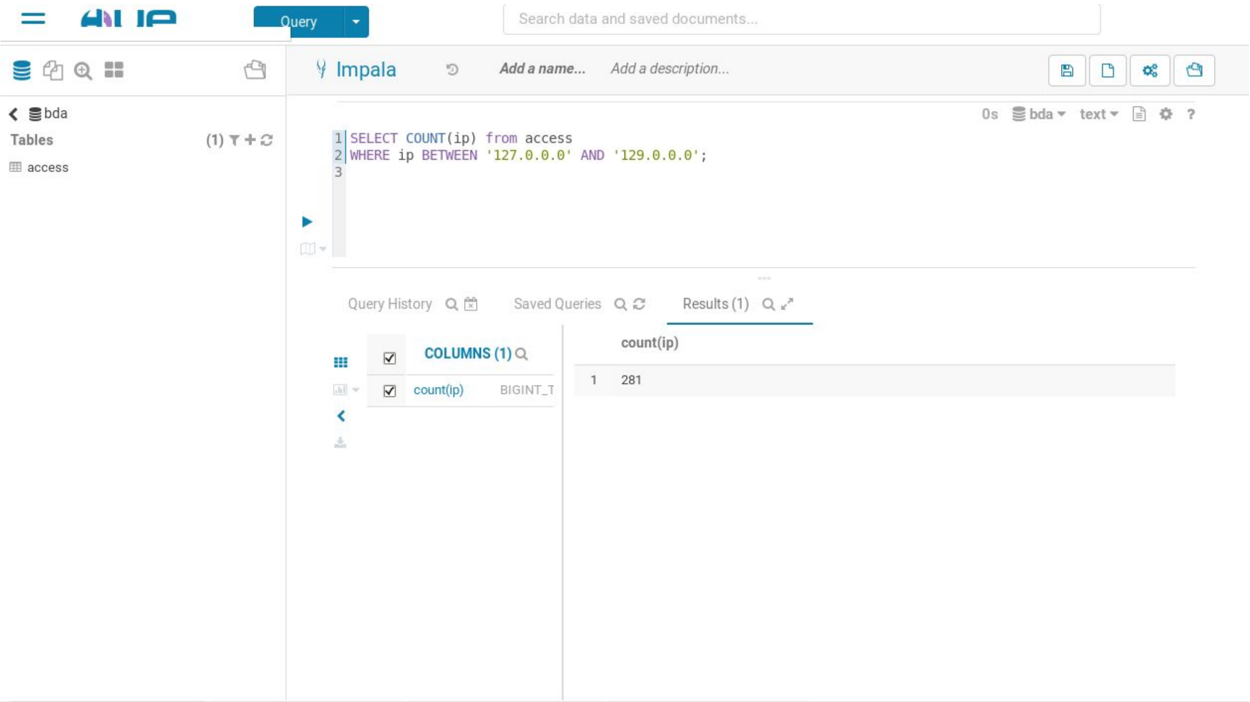


```
SELECT COUNT(ip), site
FROM access
GROUP BY site
ORDER BY COUNT(ip) DESC;
```

☑	COLUMNS (2)	
☑	count(ip)	BIGINT_T
☑	site	STRING_

	count(ip)	site
1	2604	" ""GET /favicon.ico HTTP/1.1""" 404 209"
2	606	" ""GET /yusuf/images/color_yellow.png HTTP/1.1""" 304 -"
3	605	" ""GET /yusuf/images/color_blue.png HTTP/1.1""" 304 -"
4	605	" ""GET /yusuf/images/color_magenta.png HTTP/1.1""" 304 -"
5	605	" ""GET /yusuf/images/color_cyan.png HTTP/1.1""" 304 -"
6	604	" ""GET /yusuf/images/plus.png HTTP/1.1""" 304 -"
7	601	" ""GET /yusuf/images/adddoc.png HTTP/1.1""" 304 -"
8	599	" ""GET /yusuf/images/email.png HTTP/1.1""" 304 -"
9	599	" ""GET /yusuf/images/discussion.png HTTP/1.1""" 304 -"
10	599	" ""GET /yusuf/images/calendar.png HTTP/1.1""" 304 -"
11	598	" ""GET /yusuf/images/quicknote.png HTTP/1.1""" 304 -"

7) Number of users having ips between 127.0.0.0 and 129.0.0.0

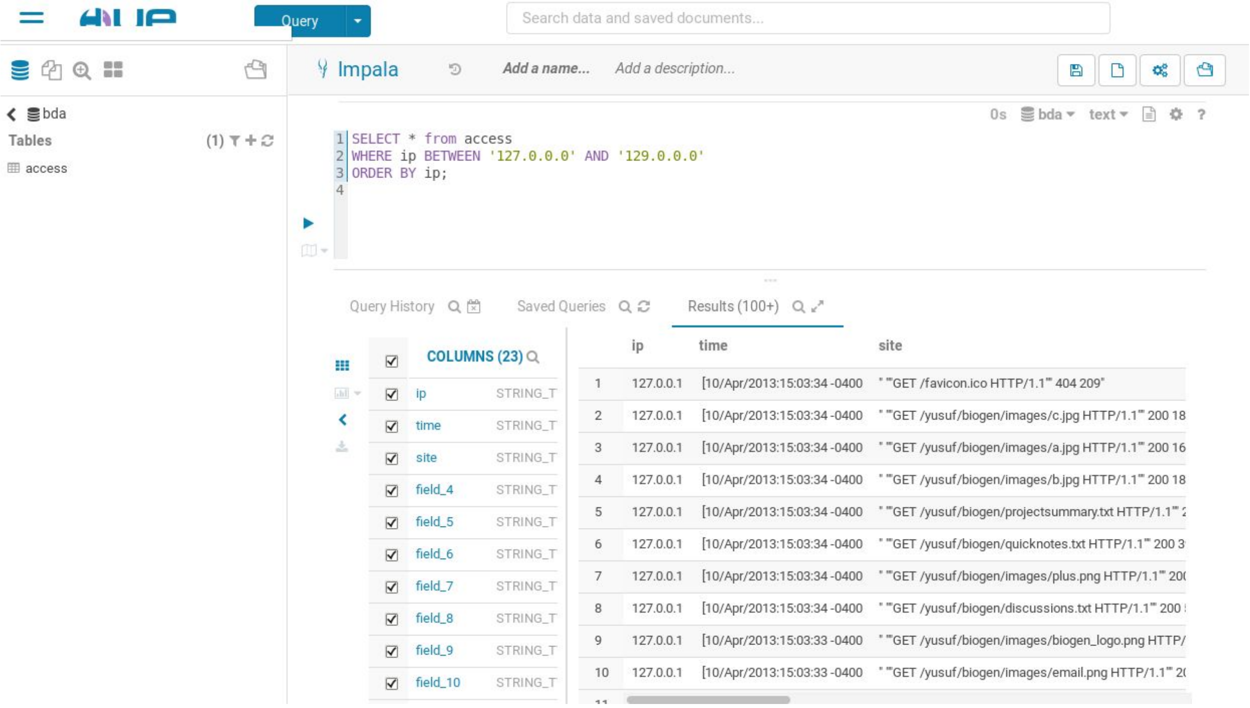


```
SELECT COUNT(ip) from access
WHERE ip BETWEEN '127.0.0.0' AND '129.0.0.0';
```

☑	COLUMNS (1)
☑	count(ip) BIGINT_T

	count(ip)
1	281

8) Ips between 127.0.0.0 and 129.0.0.0 ordered in descending order.



```
SELECT * from access
WHERE ip BETWEEN '127.0.0.0' AND '129.0.0.0'
ORDER BY ip;
```

☑	COLUMNS (23)	
☑	ip	STRING_T
☑	time	STRING_T
☑	site	STRING_T
☑	field_4	STRING_T
☑	field_5	STRING_T
☑	field_6	STRING_T
☑	field_7	STRING_T
☑	field_8	STRING_T
☑	field_9	STRING_T
☑	field_10	STRING_T

	ip	time	site
1	127.0.0.1	[10/Apr/2013:15:03:34 -0400	" ""GET /favicon.ico HTTP/1.1""" 404 209"
2	127.0.0.1	[10/Apr/2013:15:03:34 -0400	" ""GET /yusuf/biogen/images/c.jpg HTTP/1.1""" 200 18
3	127.0.0.1	[10/Apr/2013:15:03:34 -0400	" ""GET /yusuf/biogen/images/a.jpg HTTP/1.1""" 200 16
4	127.0.0.1	[10/Apr/2013:15:03:34 -0400	" ""GET /yusuf/biogen/images/b.jpg HTTP/1.1""" 200 18
5	127.0.0.1	[10/Apr/2013:15:03:34 -0400	" ""GET /yusuf/biogen/projectsummary.txt HTTP/1.1""" 2
6	127.0.0.1	[10/Apr/2013:15:03:34 -0400	" ""GET /yusuf/biogen/quicknotes.txt HTTP/1.1""" 200 3
7	127.0.0.1	[10/Apr/2013:15:03:34 -0400	" ""GET /yusuf/biogen/images/plus.png HTTP/1.1""" 20
8	127.0.0.1	[10/Apr/2013:15:03:34 -0400	" ""GET /yusuf/biogen/discussions.txt HTTP/1.1""" 200
9	127.0.0.1	[10/Apr/2013:15:03:33 -0400	" ""GET /yusuf/biogen/images/biogen_logo.png HTTP/
10	127.0.0.1	[10/Apr/2013:15:03:33 -0400	" ""GET /yusuf/biogen/images/email.png HTTP/1.1""" 2