Experiment No. 8th

Student Name: Rishav Kumar **UID: 22MCC20039**

Branch: MCA - CCD **Section/Group: 22MCD-1/ Grp A**

Semester: IV **Date of Performance: 28st Mar 2024**

Subject Name: Big Data & Analytics Lab **Subject Code: 22CAH-782**

1. Aim/Overview of the practical:

a. Install and Run Pig then write Pig Latin scripts to sort, group, join, project and filter the data.

b. Install and Run Hive then use Hive to Create, alter and drop databases, tables, views, functions and Indexes.

2. Code/Steps for practical:

a. STEPS FOR INSTALLING APACHE PIG

1) Extract the pig-0.15.0.tar.gz and move to home directory

2) Set the environment of PIG in bashrc file.

3) Pig can run in two modes

Local Mode and Hadoop Mode Pig –x local

and pig

4) Grunt Shell Grunt

> 5) LOADING

Data into Grunt

Shell

DATA = LOAD <CLASSPATH> USING PigStorage(DELIMITER) as (ATTRIBUTE : DataType1, ATTRIBUTE : DataType2…..)

6) Describe Data

Describe DATA;

7) DUMP Data

```
Dump DATA;
```

8) FILTER Data

```
FDATA = FILTER DATA by ATTRIBUTE = VALUE;
```

9) GROUP Data

```
GDATA = GROUP DATA by ATTRIBUTE;
```

10) Iterating Data

```
FOR_DATA = FOREACH DATA GENERATE GROUP AS GROUP_FUN, ATTRIBUTE = <VALUE>
```

11) Sorting Data

```
SORT_DATA = ORDER DATA BY ATTRIBUTE WITH CONDITION;
```

12) LIMIT Data

```
LIMIT_DATA = LIMIT DATA COUNT;
```

13) JOIN Data

```
JOIN DATA1 BY (ATTRIBUTE1,ATTRIBUTE2….) , DATA2 BY (ATTRIBUTE3,ATTRIBUTE….N)
```

b. Apache HIVE INSTALLATION STEPS

1) Install MySQL-Server

```
Sudo apt-get install mysql-server
```

2) Configuring MySQL UserName and Password

3) Creating User and granting all Privileges Mysql – uroot –proot

```
Create user <USER_NAME> identified by <PASSWORD>
```

4) Extract and Configure Apache Hive tar xvfz apache-hive-1.0.1.bin.tar.gz

5) Move Apache Hive from Local directory to Home directory

6) Set CLASSPATH in bashrc

```
Export HIVE_HOME = /home/apache-hive Export PATH =
$PATH:$HIVE_HOME/bin
```

7) Configuring hive-default.xml by adding My SQL Server Credentials

```
<property>
<name>javax.jdo.option.ConnectionURL</name>
<value> jdbc:mysql://localhost:3306/hive?createDatabaseIfNotExist=true </value>
</property>
<property>
<name>javax.jdo.option.ConnectionDriverName</name>
<value>com.mysql.jdbc.Driver</value>
</property>
<property>
<name>javax.jdo.option.ConnectionUserName</name>
<value>hadoop</value>
</property>
<property>
<name>javax.jdo.option.ConnectionPassword</name>
<value>hadoop</value>
</property>
```

8) Copying mysql-java-connector.jar to hive/lib directory.

SYNTAX for HIVE Database Operations DATABASE Creation

CREATE DATABASE|SCHEMA [IF NOT EXISTS] <database name>

Drop Database Statement

DROP DATABASE StatementDROP (DATABASE|SCHEMA) [IF EXISTS] database_name [RESTRICT|CASCADE];

Creating and Dropping Table in HIVE

CREATE [TEMPORARY] [EXTERNAL] TABLE [IF NOT EXISTS] [db_name.] table_name

[(col_name data_type [COMMENT col_comment], ...)]

[COMMENT table_comment] [ROW FORMAT row_format] [STORED AS file_format]

Loading Data into table log_data Syntax:

```
LOAD DATA LOCAL INPATH '<path>/u.data' OVERWRITE INTO TABLE u_data;
```

Alter Table in HIVE

Syntax

```
ALTER TABLE name RENAME TO new_name
ALTER TABLE name ADD COLUMNS (col_spec[, col_spec ...]) ALTER
TABLE name DROP [COLUMN] column_name
ALTER TABLE name CHANGE column_name new_name new_type ALTER TABLE name
REPLACE COLUMNS (col_spec[, col_spec ...])
```

Creating and Dropping View

```
CREATE VIEW [IF NOT EXISTS] view_name [(column_name [COMMENT column_comment], ...) ]
[COMMENT table_comment] AS SELECT ...
```

Dropping View Syntax:

```
DROP VIEW view_name
```

Functions in HIVE

String Functions:- round(), ceil(), substr(), upper(), reg_exp() etc Date and Time

Functions:- year(), month(), day(), to_date() etc Aggregate Functions :- sum(),

min(), max(), count(), avg() etc

INDEXES

```
CREATE INDEX index_name ON TABLE base_table_name (col_name, ...) AS 'index.handler.class.name'
[WITH DEFERRED REBUILD]
[IDXPROPERTIES (property_name=property_value, ...)] [IN TABLE index_table_name]
[PARTITIONED BY (col_name, ...)] [ [
ROW FORMAT ...] STORED AS ...
| STORED BY ...
]
[LOCATION hdfs_path]
```

[TBLPROPERTIES (...)]

Creating Index

CREATE INDEX index_ip ON TABLE log_data(ip_address) AS

'org.apache.hadoop.hive.ql.index.compact.CompactIndexHandler' WITH DEFERRED REBUILD;

Altering and Inserting Index

ALTER INDEX index_ip_address ON log_data REBUILD;

Storing Index Data in Metastore

SET

hive.index.compact.file=/home/administrator/Desktop/big/metastore_db/tmp/index_ipaddress_re sult; SET

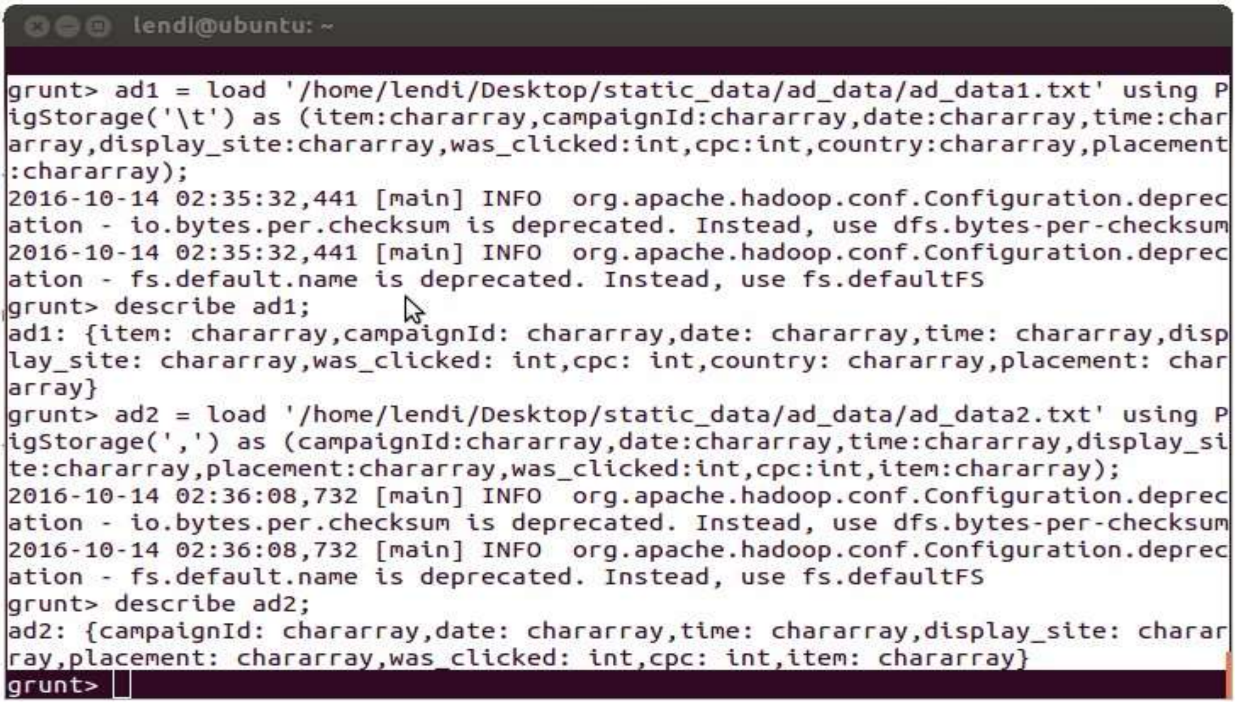
hive.input.format=org.apache.hadoop.hive.ql.index.compact.HiveCompactIndexInputFormat;

Dropping Index

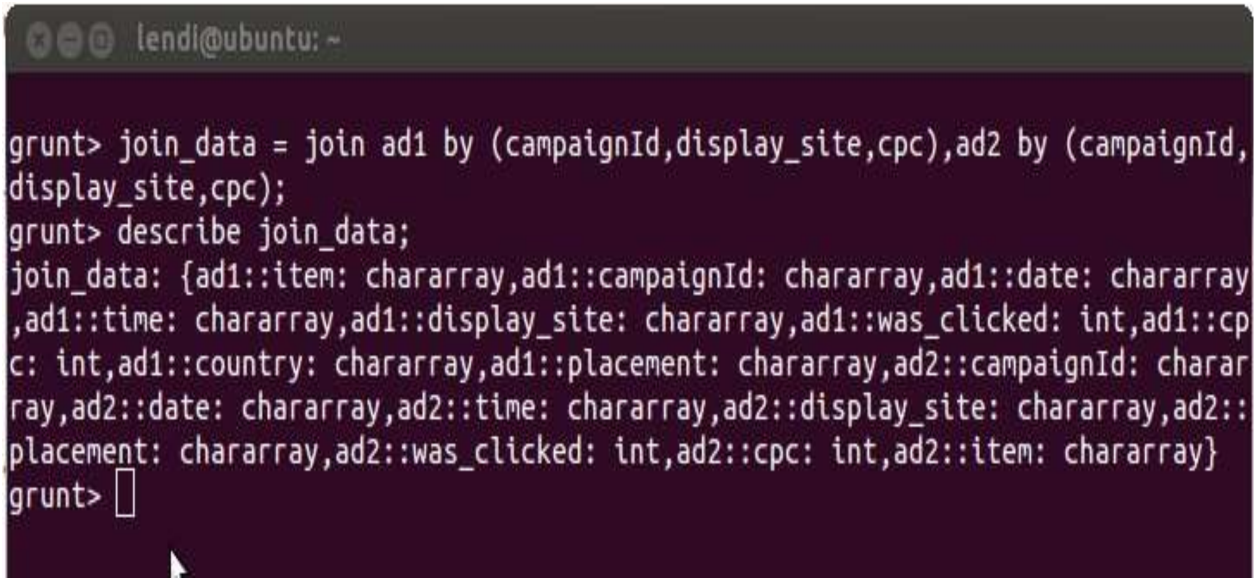
DROP INDEX INDEX_NAME on TABLE_NAME;

3. Result/Output/Writing Summary:

a.

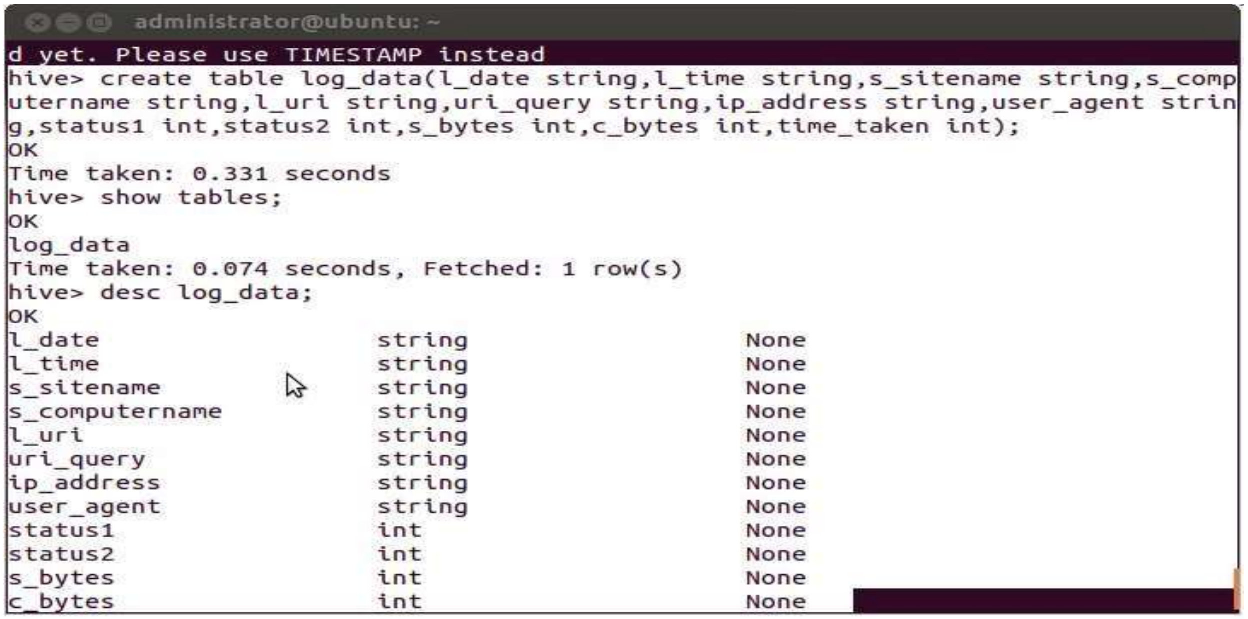


```
grunt> ad1 = load '/home/lendi/Desktop/static_data/ad_data/ad_data1.txt' using P
igStorage('\t') as (item:chararray,campaignId:chararray,date:chararray,time:char
array,display_site:chararray,was_clicked:int,cpc:int,country:chararray,placement
:chararray);
2016-10-14 02:35:32,441 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2016-10-14 02:35:32,441 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> describe ad1;
ad1: {item: chararray,campaignId: chararray,date: chararray,time: chararray,disp
lay_site: chararray,was_clicked: int,cpc: int,country: chararray,placement: char
array}
grunt> ad2 = load '/home/lendi/Desktop/static_data/ad_data/ad_data2.txt' using P
igStorage(',') as (campaignId:chararray,date:chararray,time:chararray,display_si
te:chararray,placement:chararray,was_clicked:int,cpc:int,item:chararray);
2016-10-14 02:36:08,732 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2016-10-14 02:36:08,732 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> describe ad2;
ad2: {campaignId: chararray,date: chararray,time: chararray,display_site: charar
ray,placement: chararray,was_clicked: int,cpc: int,item: chararray}
grunt> 
```

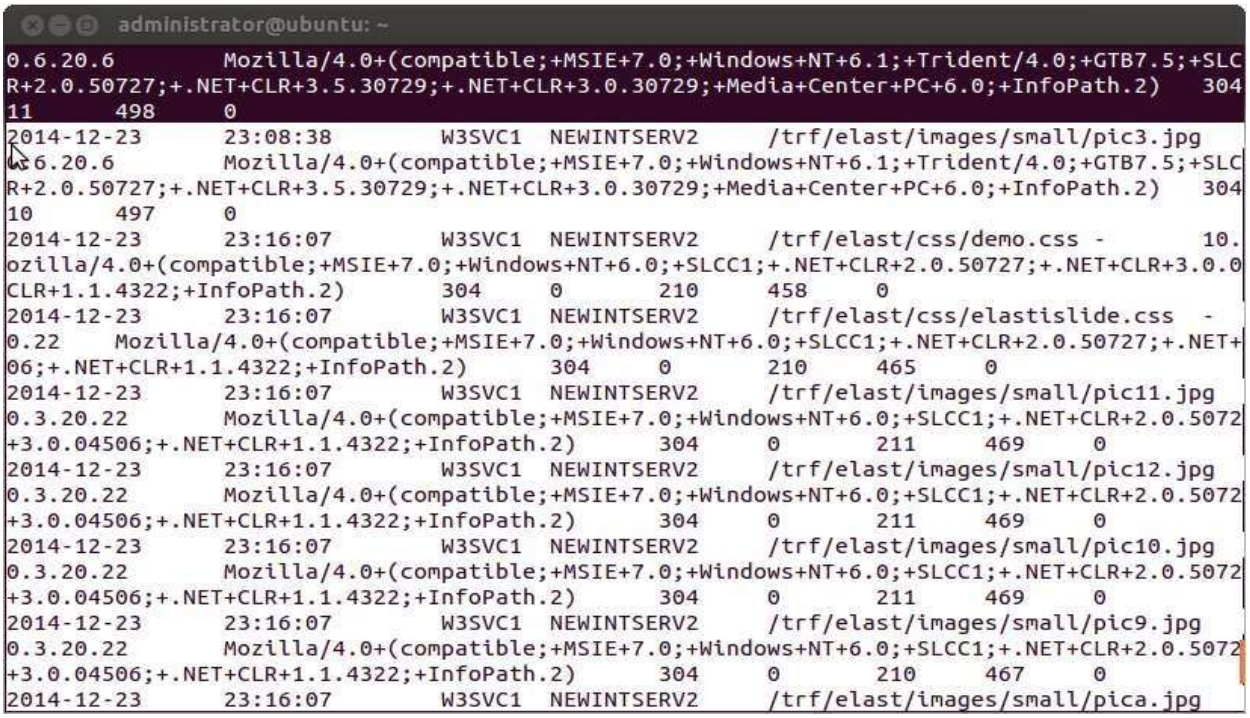


```
grunt> join_data = join ad1 by (campaignId,display_site,cpc),ad2 by (campaignId,
display_site,cpc);
grunt> describe join_data;
join_data: {ad1::item: chararray,ad1::campaignId: chararray,ad1::date: chararray
,ad1::time: chararray,ad1::display_site: chararray,ad1::was_clicked: int,ad1::cp
c: int,ad1::country: chararray,ad1::placement: chararray,ad2::campaignId: charar
ray,ad2::date: chararray,ad2::time: chararray,ad2::display_site: chararray,ad2::
placement: chararray,ad2::was_clicked: int,ad2::cpc: int,ad2::item: chararray}
grunt> 
```

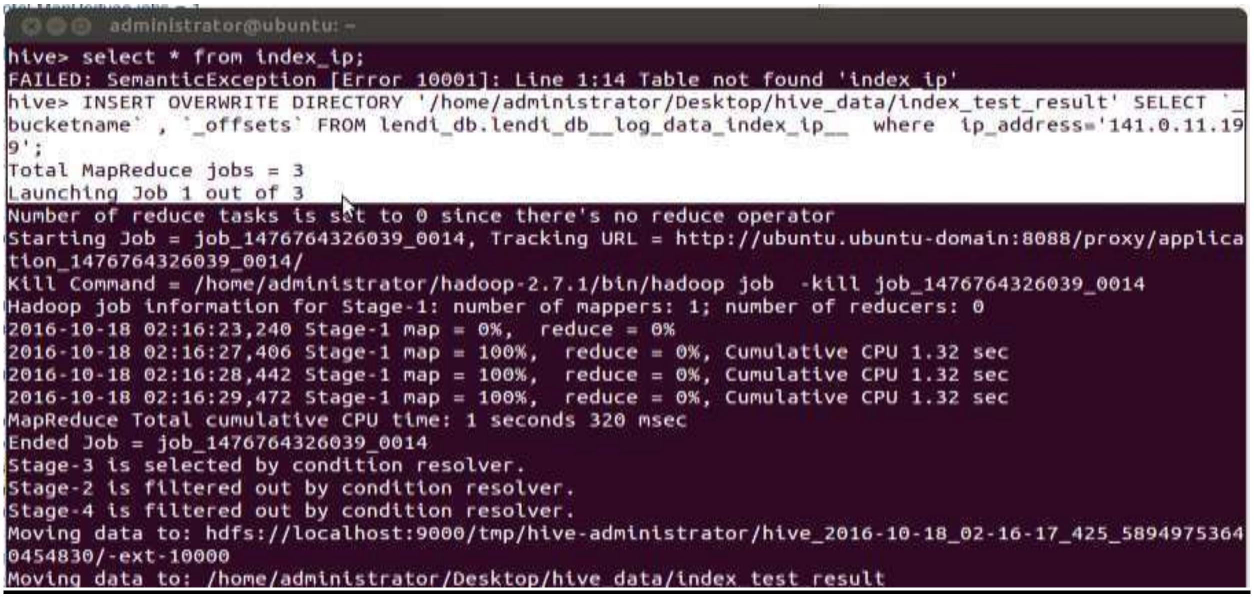
b.



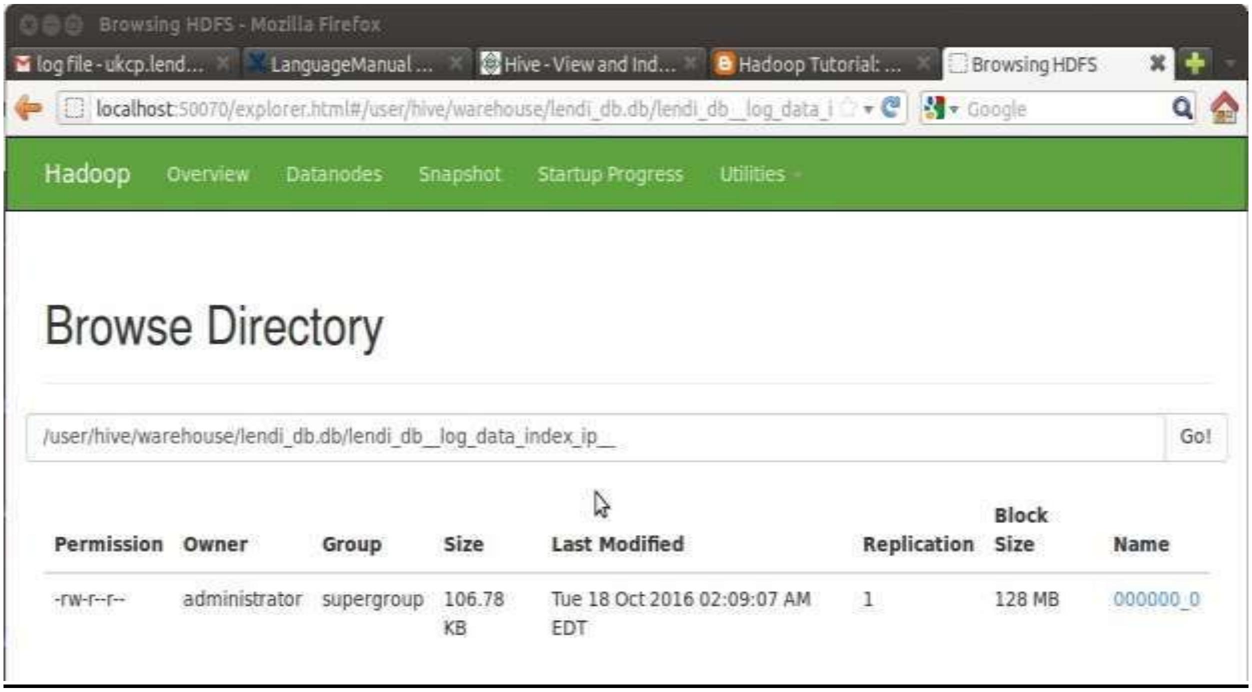
```
d yet. Please use TIMESTAMP instead
hive> create table log_data(l_date string,l_time string,s_sitename string,s_comp
utername string,l_uri string,uri_query string,ip_address string,user_agent strin
g,status1 int,status2 int,s_bytes int,c_bytes int,time_taken int);
OK
Time taken: 0.331 seconds
hive> show tables;
OK
log_data
Time taken: 0.074 seconds, Fetched: 1 row(s)
hive> desc log_data;
OK
l_date                  string                  None
l_time                  string                  None
s_sitename              string                  None
s_computername          string                  None
l_uri                   string                  None
uri_query               string                  None
ip_address              string                  None
user_agent              string                  None
status1                 int                     None
status2                 int                     None
s_bytes                 int                     None
c_bytes                 int                     None
```



```
0.6.20.6        Mozilla/4.0+(compatible;+MSIE+7.0;+Windows+NT+6.1;+Trident/4.0;+GTB7.5;+SLC
R+2.0.50727;+.NET+CLR+3.5.30729;+.NET+CLR+3.0.30729;+Media+Center+PC+6.0;+InfoPath.2)   304
11      498     0
2014-12-23      23:08:38        W3SVC1  NEWINTSERV2     /trf/elast/images/small/pic3.jpg
 6.20.6         Mozilla/4.0+(compatible;+MSIE+7.0;+Windows+NT+6.1;+Trident/4.0;+GTB7.5;+SLC
R+2.0.50727;+.NET+CLR+3.5.30729;+.NET+CLR+3.0.30729;+Media+Center+PC+6.0;+InfoPath.2)   304
10      497     0
2014-12-23      23:16:07        W3SVC1  NEWINTSERV2     /trf/elast/css/demo.css -       10.
ozilla/4.0+(compatible;+MSIE+7.0;+Windows+NT+6.0;+SLCC1;+.NET+CLR+2.0.50727;+.NET+CLR+3.0.0
CLR+1.1.4322;+InfoPath.2)       304     0       210     458     0
2014-12-23      23:16:07        W3SVC1  NEWINTSERV2     /trf/elast/css/elastislide.css  -
0.22    Mozilla/4.0+(compatible;+MSIE+7.0;+Windows+NT+6.0;+SLCC1;+.NET+CLR+2.0.50727;+.NET+
06;+.NET+CLR+1.1.4322;+InfoPath.2)      304     0       210     465     0
2014-12-23      23:16:07        W3SVC1  NEWINTSERV2     /trf/elast/images/small/pic11.jpg
0.3.20.22       Mozilla/4.0+(compatible;+MSIE+7.0;+Windows+NT+6.0;+SLCC1;+.NET+CLR+2.0.5072
+3.0.04506;+.NET+CLR+1.1.4322;+InfoPath.2)      304     0       211     469     0
2014-12-23      23:16:07        W3SVC1  NEWINTSERV2     /trf/elast/images/small/pic12.jpg
0.3.20.22       Mozilla/4.0+(compatible;+MSIE+7.0;+Windows+NT+6.0;+SLCC1;+.NET+CLR+2.0.5072
+3.0.04506;+.NET+CLR+1.1.4322;+InfoPath.2)      304     0       211     469     0
2014-12-23      23:16:07        W3SVC1  NEWINTSERV2     /trf/elast/images/small/pic10.jpg
0.3.20.22       Mozilla/4.0+(compatible;+MSIE+7.0;+Windows+NT+6.0;+SLCC1;+.NET+CLR+2.0.5072
+3.0.04506;+.NET+CLR+1.1.4322;+InfoPath.2)      304     0       211     469     0
2014-12-23      23:16:07        W3SVC1  NEWINTSERV2     /trf/elast/images/small/pic9.jpg
0.3.20.22       Mozilla/4.0+(compatible;+MSIE+7.0;+Windows+NT+6.0;+SLCC1;+.NET+CLR+2.0.5072
+3.0.04506;+.NET+CLR+1.1.4322;+InfoPath.2)      304     0       210     467     0
2014-12-23      23:16:07        W3SVC1  NEWINTSERV2     /trf/elast/images/small/pica.jpg
```



```
hive> select * from index_ip;
FAILED: SemanticException [Error 10001]: Line 1:14 Table not found 'index_ip'
hive> INSERT OVERWRITE DIRECTORY '/home/administrator/Desktop/hive_data/index_test_result' SELECT `_
bucketname` , `_offsets` FROM lendi_db.lendi_db__log_data_index_ip__  where  ip_address='141.0.11.19
9';
Total MapReduce 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_1476764326039_0014, Tracking URL = http://ubuntu.ubuntu-domain:8088/proxy/applica
tion_1476764326039_0014/
Kill Command = /home/administrator/hadoop-2.7.1/bin/hadoop job  -kill job_1476764326039_0014
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2016-10-18 02:16:23,240 Stage-1 map = 0%,  reduce = 0%
2016-10-18 02:16:27,406 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.32 sec
2016-10-18 02:16:28,442 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.32 sec
2016-10-18 02:16:29,472 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.32 sec
MapReduce Total cumulative CPU time: 1 seconds 320 msec
Ended Job = job_1476764326039_0014
Stage-3 is selected by condition resolver.
Stage-2 is filtered out by condition resolver.
Stage-4 is filtered out by condition resolver.
Moving data to: hdfs://localhost:9000/tmp/hive-administrator/hive_2016-10-18_02-16-17_425_5894975364
0454830/-ext-10000
Moving data to: /home/administrator/Desktop/hive_data/index_test_result
```



Browse Directory

/user/hive/warehouse/lendi_db.db/lendi_db__log_data_index_ip__

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	administrator	supergroup	106.78 KB	Tue 18 Oct 2016 02:09:07 AM EDT	1	128 MB	000000_0

4. Learning outcomes (What I have learned):

a. Install and Run Pig then write Pig Latin scripts to sort, group, join, project and filter the

data.

b. Install and Run Hive then use Hive to Create, alter and drop databases, tables, views, functions and Indexes.