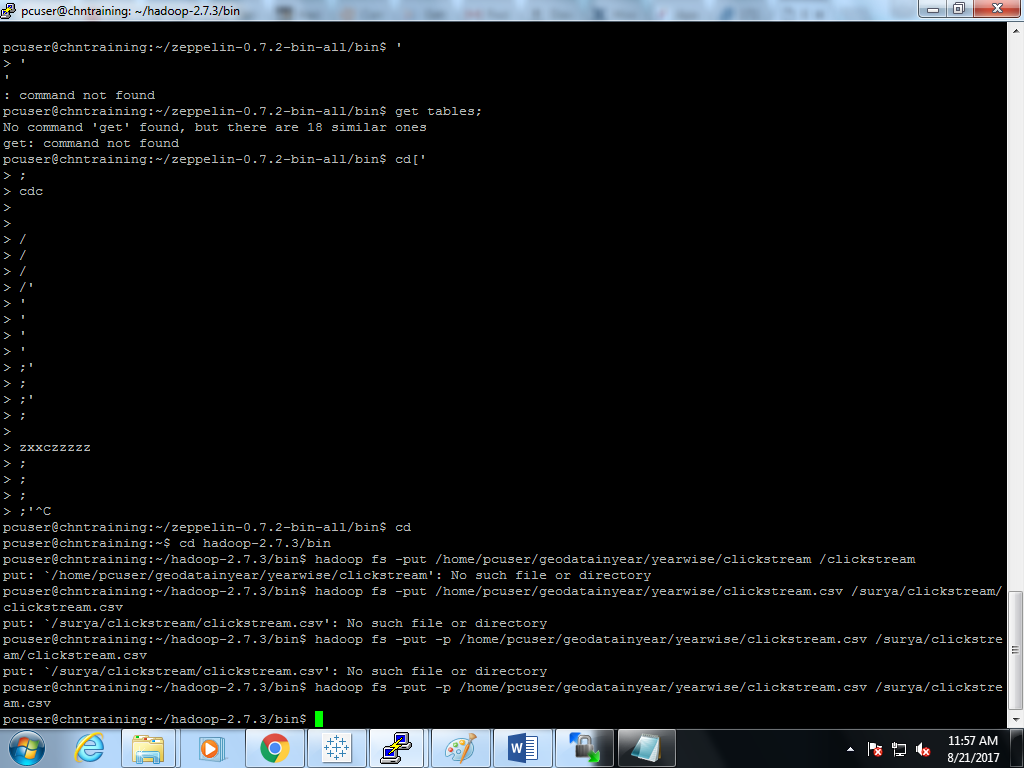
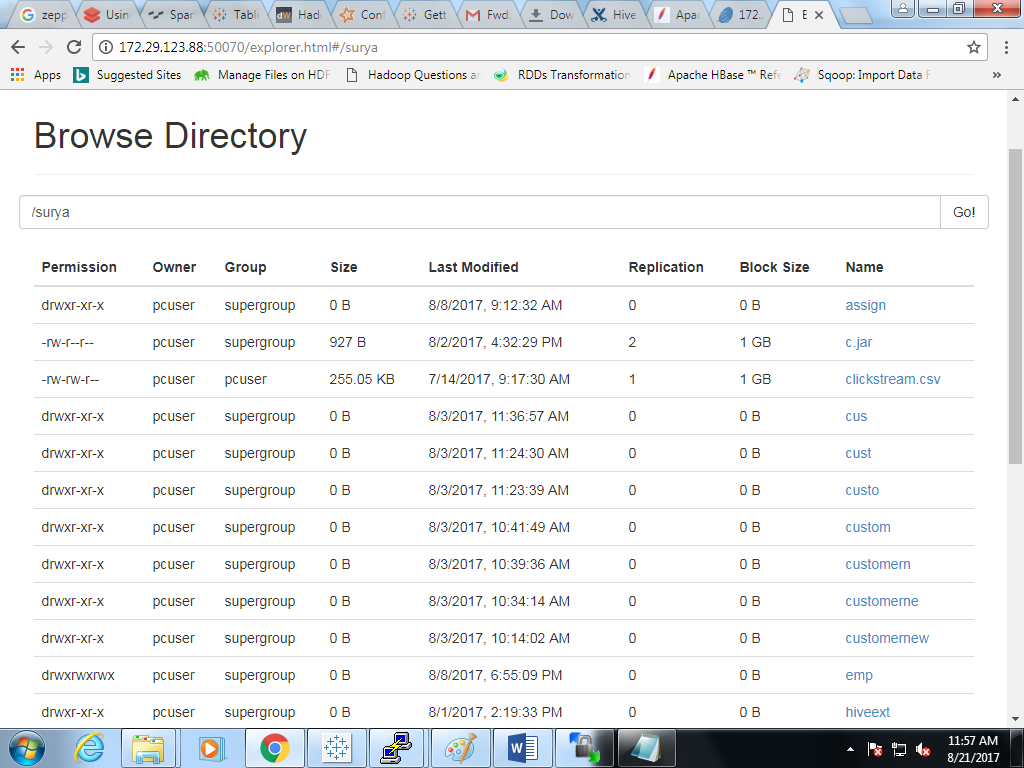
**Step to Load and Store the Clickstream data in Hdfs.**

**//**clickstream analysis on static data:

loading the files into hdfs from local:

hadoop fs -put /home/pcuser/geodatainyear/yearwise/clickstream.csv /surya/clickstream.csv

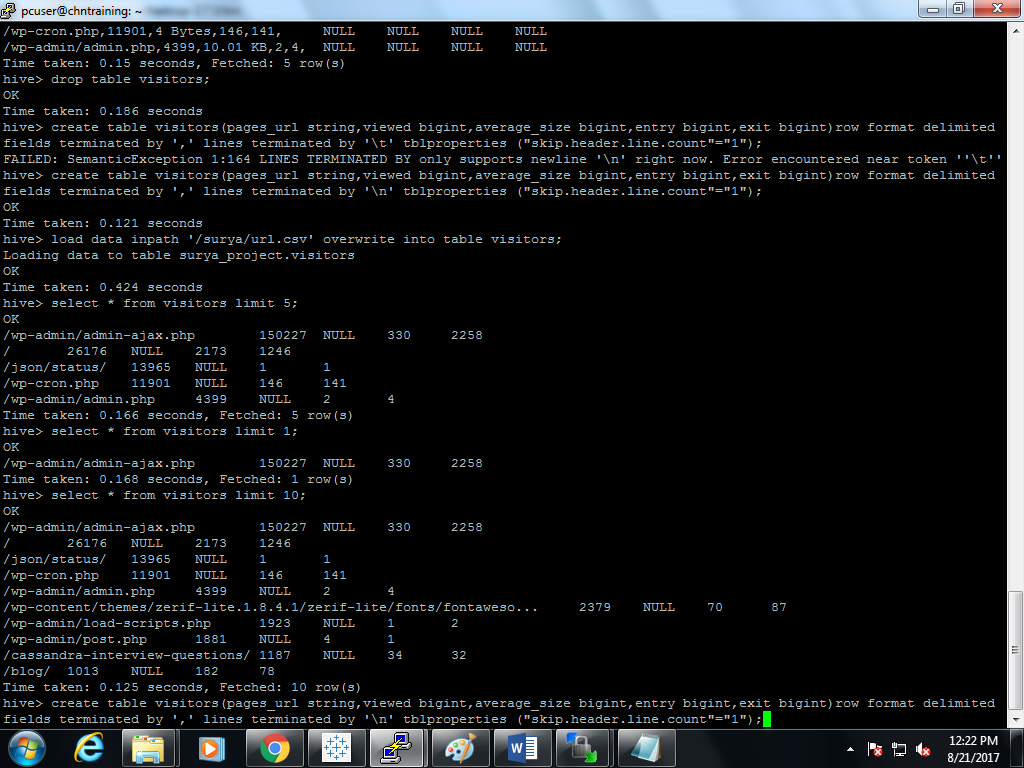


hadoop fs -put /home/pcuser/geodatainyear/yearwise/url.csv /surya/url.csv

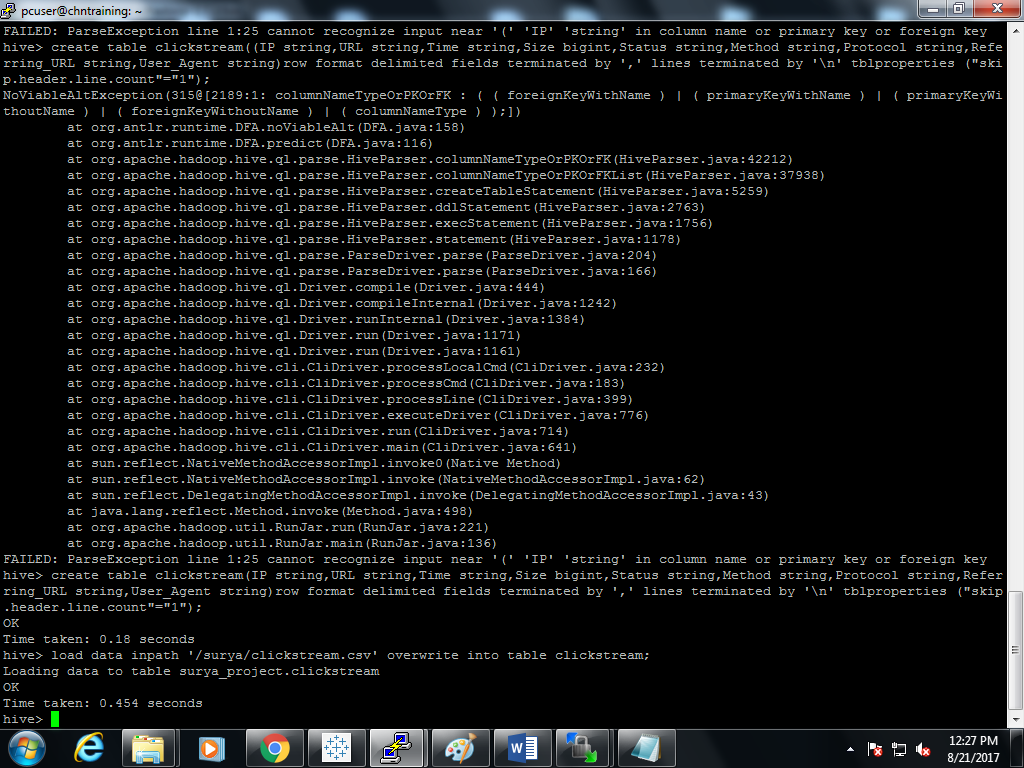
**IN HIVE**

use surya\_project;

1. create table visitors(pages\_url string,viewed bigint,average\_size bigint,entry bigint,exit bigint)row format delimited fields terminated by ',' lines terminated by '\n' tblproperties ("skip.header.line.count"="1");
2. load data inpath '/surya/url.csv' overwrite into table visitors;

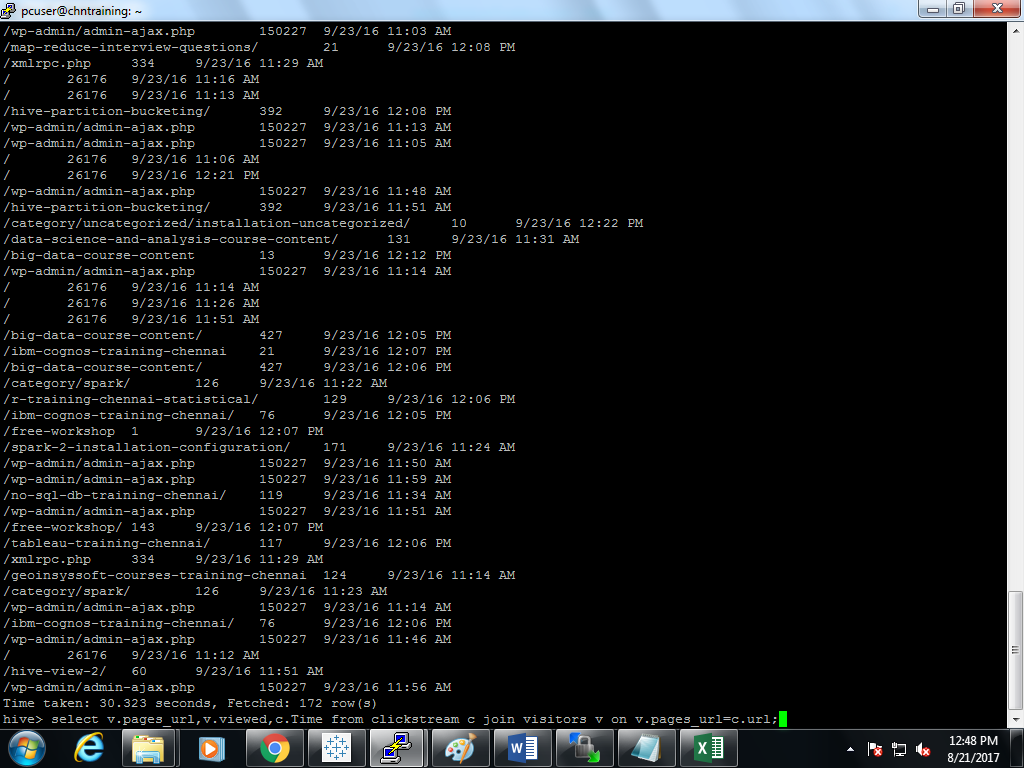
****

1. create table clickstream(ID int,IP string,URL string,Time string, Size bigint,Status string,Method string,Protocol string,Referring\_URL string,User\_Agent string)row format delimited fields terminated by ',' lines terminated by '\n' tblproperties ("skip.header.line.count"="1");
2. load data inpath '/surya/url.csv' overwrite into table visitors;

****

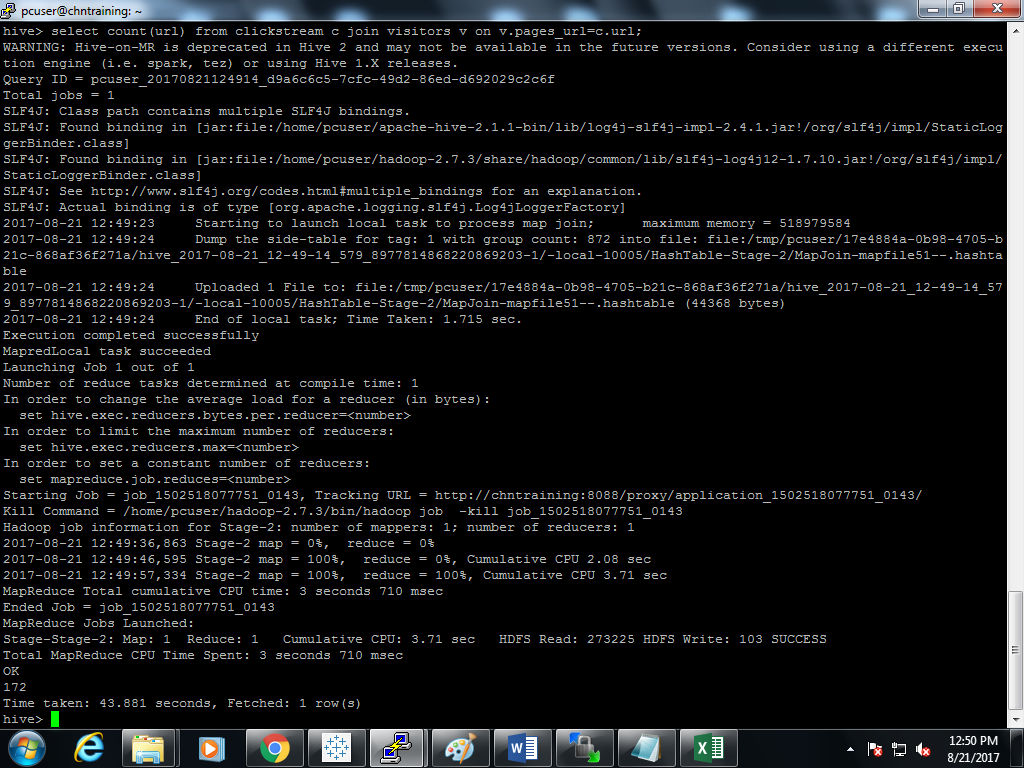
**Joining tables:**

1. select v.pages\_url,v.viewed,c.Time from clickstream c join visitors v on v.pages\_url=c.url;

****

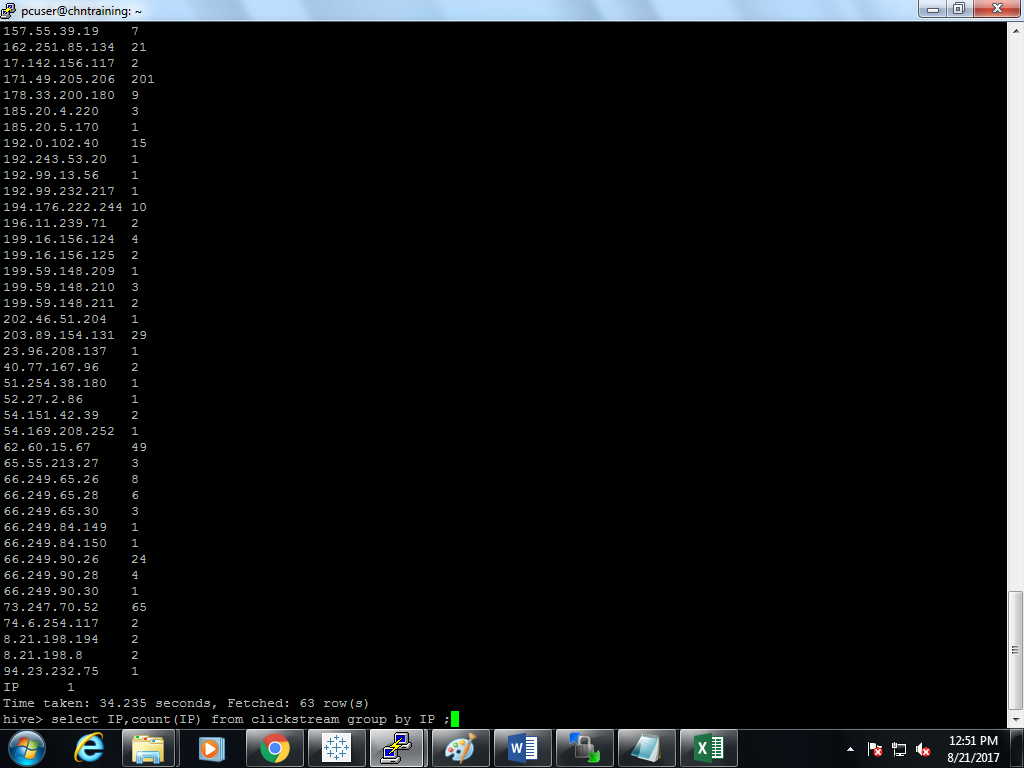
**Counting the overall urls for the website:**

1. select count(url) from clickstream c join visitors v on v.pages\_url=c.url;

****

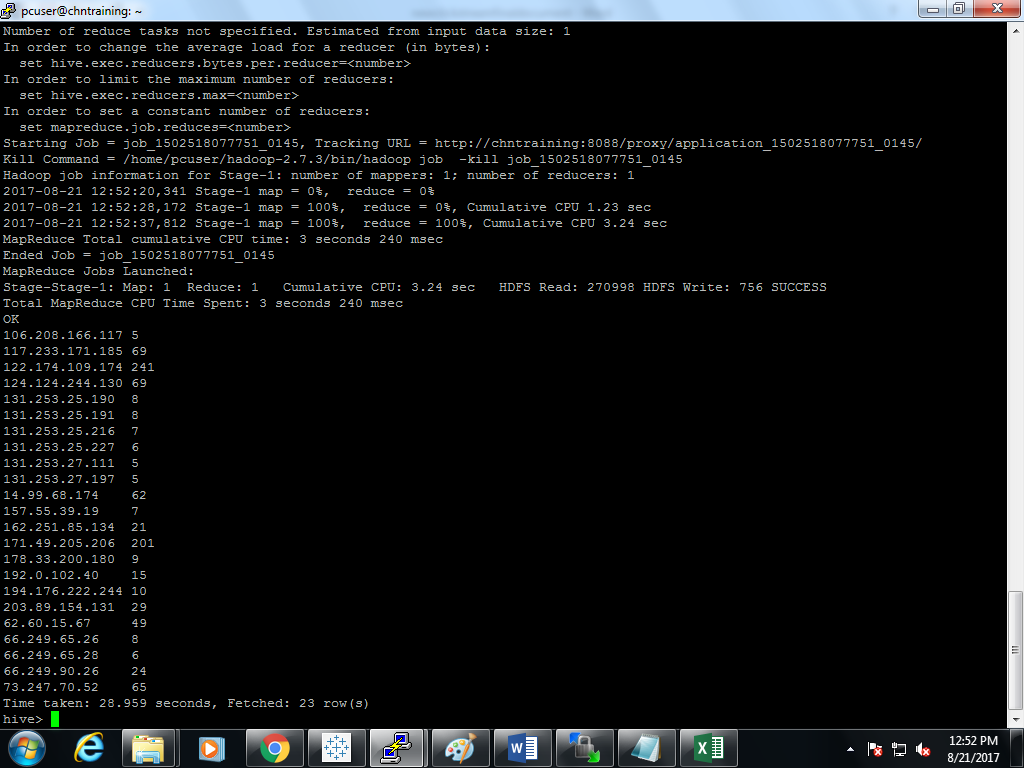
**Grouping the column IP in clickstream table:**

1. select IP,count(IP) from clickstream group by IP ;

****

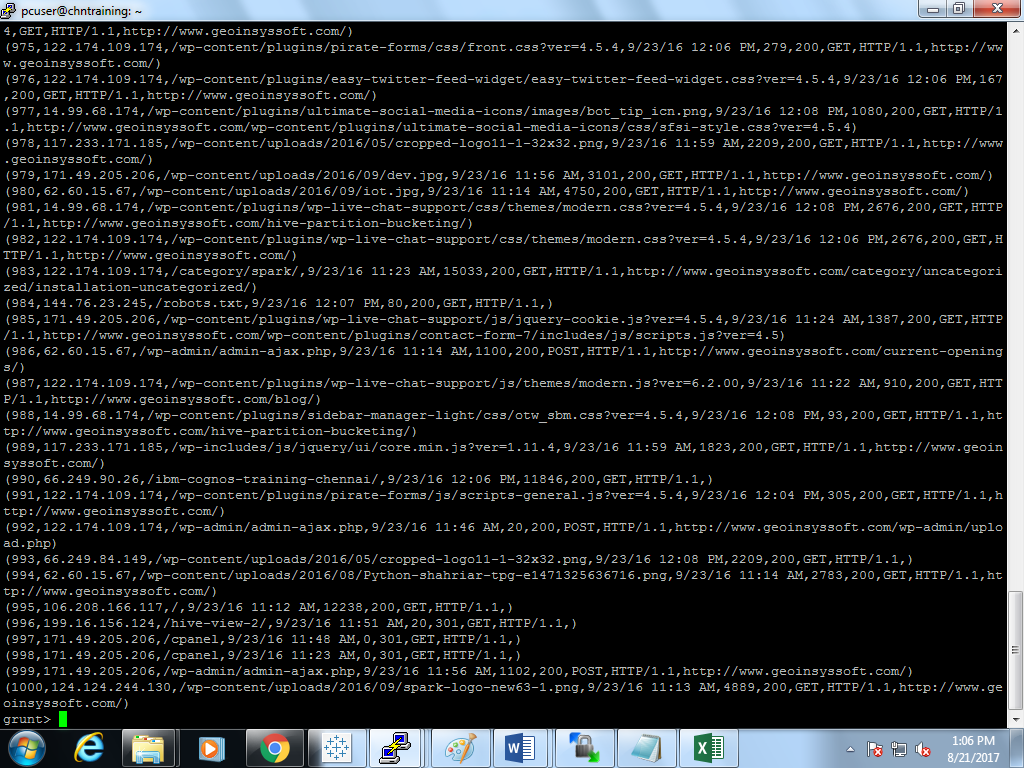
**IP address , which user's have visited more than 5 times:**

1. select IP,count(IP) from clickstream group by IP having count (IP)>=5;

****

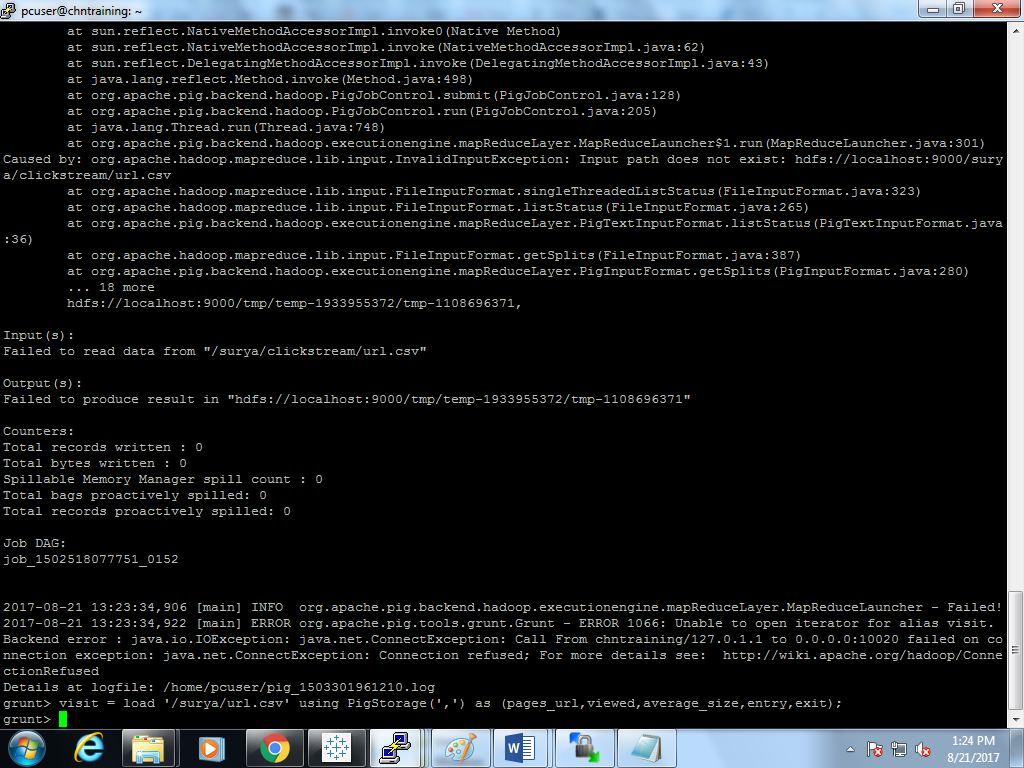
**PIG:**

1. click=load'/surya/clickstream.csv' using PigStorage(',') as (Ip,URL,Time,Size,Status,Method,Protocol,Referring\_URL,User\_Agent);
2. Dump click;

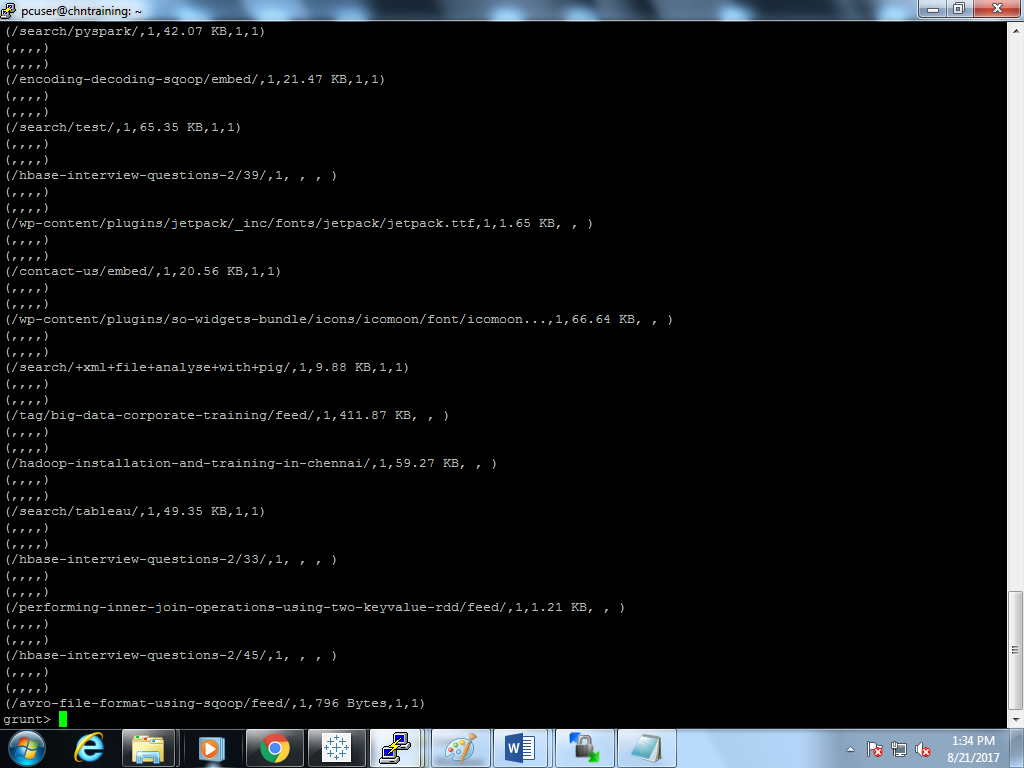
****

**Loading url.csv into pig from hdfs:**

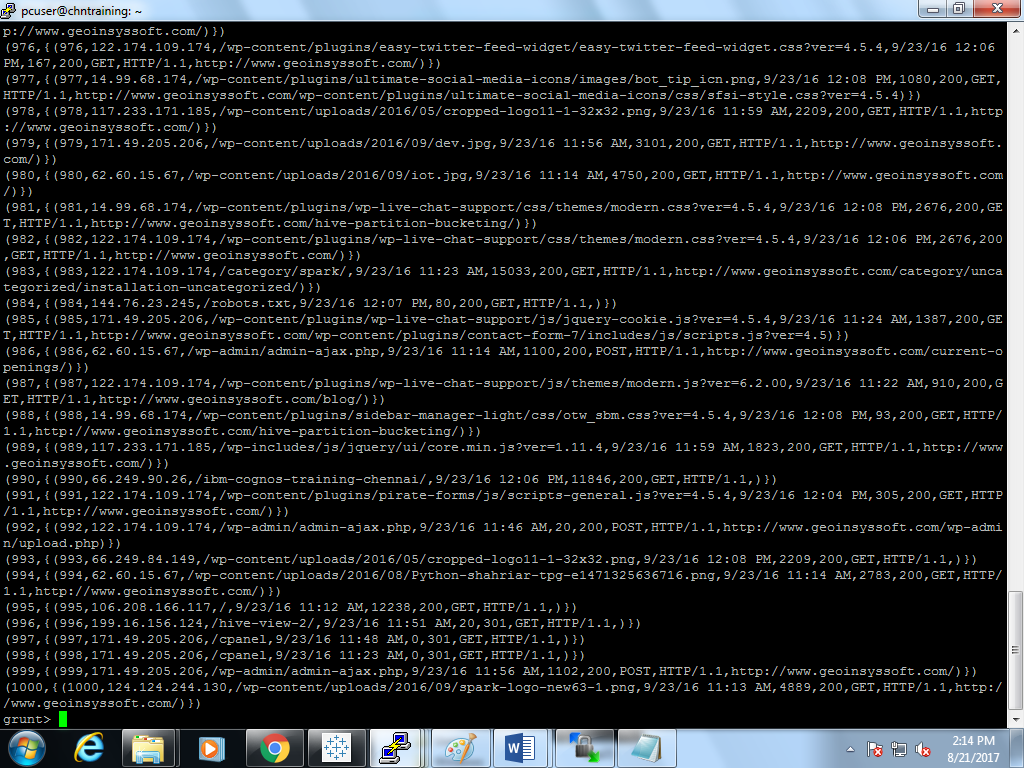
1. visit = load '/surya/url.csv' using PigStorage(',') as (pages\_url,viewed,average\_size,entry,exit);



1. Dump visit;

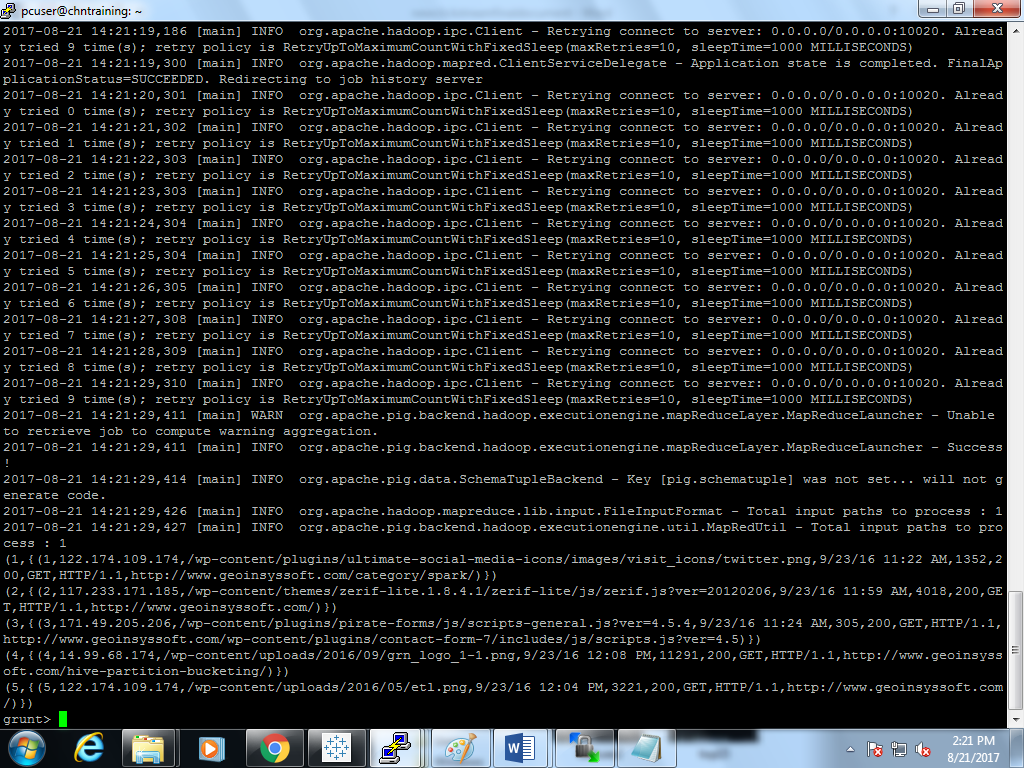
****

**Grouping the table click by Ip column and display the data with limit upto 5.**

1. group\_data = GROUP click by Ip;
2. dump group\_data;****

1.limit\_group = limit group\_data 5;

2.dump limit\_group;

****