================= Log Data ==============

Assume that website is generating log files which contains details of URL,USER,TIMESTAMP. Here we will use static csv file to analyse our queries. But if it is generating in real time then we can use flume or spark to transfer data to hdfs.

================= CREATING TABLE ==============

My dataset has been transfered in WebsiteAnalysisNew hdfs file.

CREATE EXTERNAL TABLE WebsiteAnalysis

(URL varchar(255),

USER varchar(20),

time timestamp

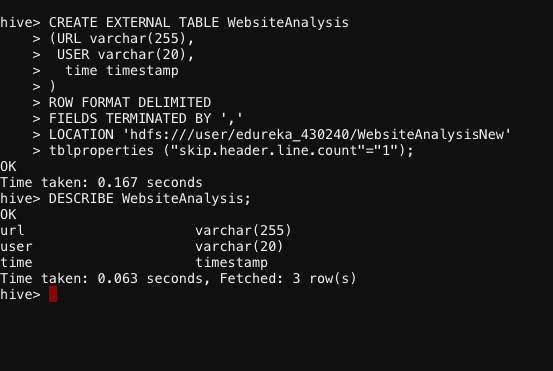
)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

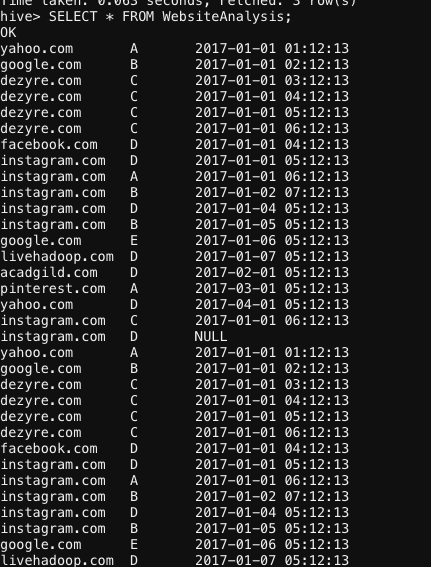
LOCATION 'hdfs:///user/edureka\_430240/WebsiteAnalysisNew'

tblproperties ("skip.header.line.count"="1");



=================VALUES INSIDE TABLE ==============

SELECT \* FROM WebsiteAnalysis;



===================PROBLEMS==========================

**1. Which page was visited by user C more than 3 times in a day**

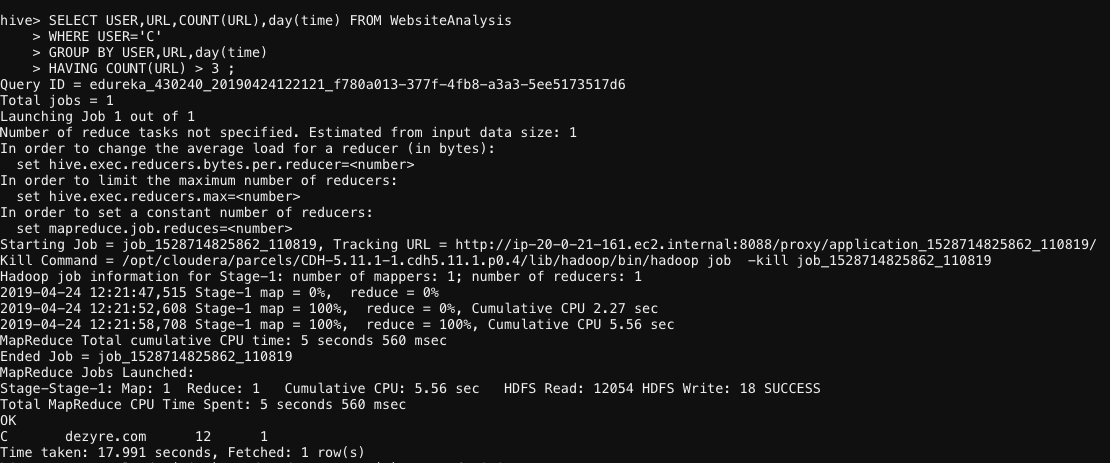
SELECT USER,URL,COUNT(URL),day(time) FROM WebsiteAnalysis

WHERE USER='C'

GROUP BY USER,URL,day(time)

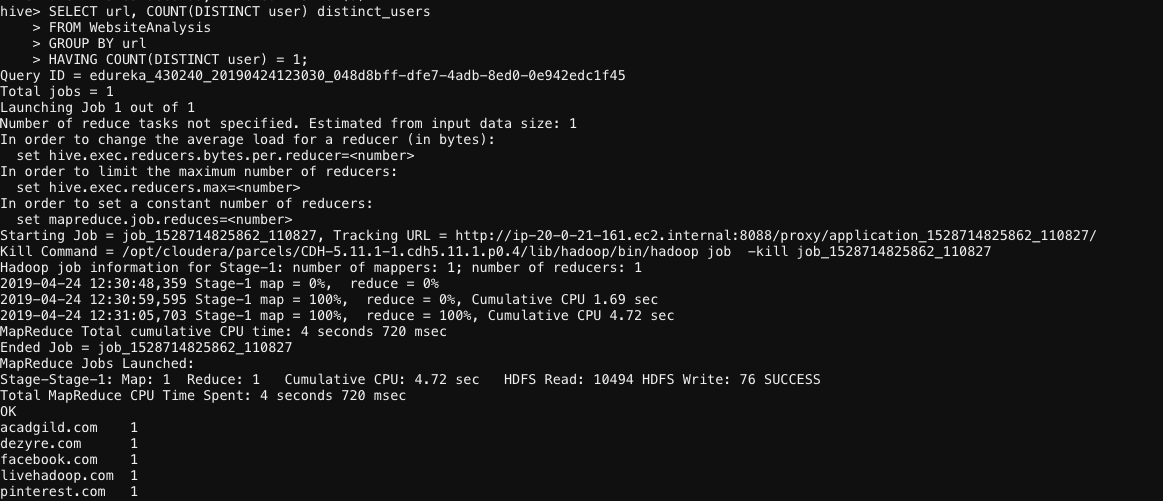
HAVING COUNT(URL) > 3 ;

(GROUP BY day(time) because it is required in a single day)



**2. Which page was visited by only one user exactly 3 times in a day?**

First, let’s find out the result in which 1 page is visited by exactly one user:



Now, rest part -- **exactly 3 times in a day**

SELECT a.USER, a.url, day(a.time) AS day\_date, count(\*) AS n\_of\_visits

FROM WebsiteAnalysis a

JOIN

(SELECT url, COUNT(DISTINCT user) distinct\_users

FROM WebsiteAnalysis

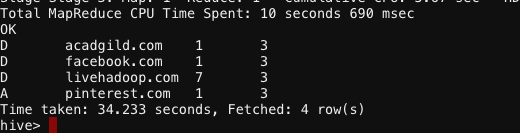
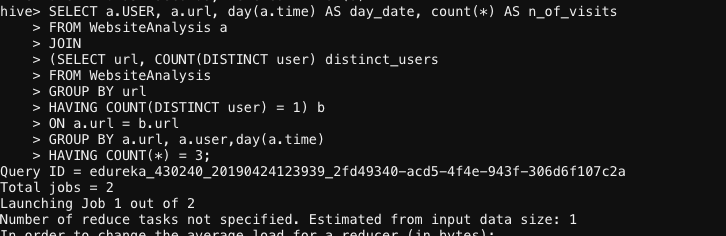
GROUP BY url

HAVING COUNT(DISTINCT user) = 1) b

ON a.url = b.url

GROUP BY a.url, a.user,day(a.time)

HAVING COUNT(\*) = 3;

****