**ClickStream analysis using Pig and Hive**

Many of the e-commerce sites had been making an impact on the overall economy for some time in many of the countries. Recently it had picked up quite good in India also. Flipkart is one of the top portal in India. Some time back I bought a HP 430 Notebook from Flipkart.

All the e-commerce portals store the user activities on their site as clickstream activity and later they analyze it to identify what the user has browsed and show the appropriate recommendations when the user visits the site again.

I have tried to demonstrate how to analyze the clickstream and the user data together using Pig and Hive. The challenge is to find the top 3 URLs visited by users whose age is less than 16 years.

To get a quick overview of Pig and Hive, Hadoop - The Definitive Guide is the best resource, but to deep dive Programming Hive and Programming Pig are the best bets. Some time back I compared Pig and Hive in a blog entry.

Data from the external systems can be pushed into HDFS using Sqoop, Flume and in many other ways. For now, lets assume that the user data and the clickstream data is already there is HDFS as shown below. For sake of simplicity only a few columns have been included, but can be made much more complex.

Let's create the 'user' table in Hive and map the user data in HDFS to the table

**Data**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | Name | Age | Country | Gender |
| 1 | Mohan | 31 | India | M |
| 2 | Saranya | 29 | India | F |
| 3 | Mahesh | 25 | India | M |

**Query to create table**

*CREATE TABLE user (*

*user\_id INT,*

*name STRING,*

*age INT,*

*country STRING,*

*gender STRING*

*) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';*

**Loading the data into table**

*LOAD DATA INPATH '/user/training/user/user.csv OVERWRITE INTO TABLE user;*

**Similarly, let's create the 'clickstream' table in Hive and map the user data in HDFS to the table**

**Data**

|  |  |
| --- | --- |
| **ID** | **URL** |
| 1 | www.bbc.com |
| 1 | www.abc.com |
| 1 | www.gmail.com |
| 2 | www.cnn.com |
| 2 | www.eenadu.net |
| 2 | www.stackoverflow.com |
| 2 | www.businessweek.com |
| 3 | www.eenadu.net |
| 3 | www.stackoverflow.com |
| 3 | www.businessweek.com |

**Query to create table.**

*CREATE TABLE clickstream (*

*userid INT,*

*url STRING*

*) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';*

*LOAD DATA INPATH '/user/training/clickstream.csv' OVERWRITE INTO TABLE clickstream;*

Here is the HiveQL query to get the top 3 URLs visited by user whose age is less than 16. The query looks very similar to SQL, which makes it easy to get started with Hive. Hive automatically creates a plan for the below query and submits it to the Hadoop cluster. SQL interfaces are being added to the Big Data frameworks to make it easier for those who are familiar with SQL to get started with the different Big Data frameworks easily. Here is an interesting article from GigaOM on the same.

*SELECT url,count(url) c FROM user u JOIN clickstream c ON (u.user\_id=c.userid) where u.age<16 group by url order by c DESC LIMIT 3;*

**Here is the PigLatin code for the same.**

*Users1 = load '/user/training/user/user.csv’ using PigStorage(',') as (user\_id, name, age:int, country, gender);*

*Fltrd = filter Users1 by age <= 16;*

*Pages = load '/user/training/clickstream.txt' using PigStorage(',') as (userid, url);*

*Jnd = join Fltrd by user\_id, Pages by userid;*

*Grpd = group Jnd by url;*

*Smmd = foreach Grpd generate group, COUNT(Jnd) as clicks;*

*Srtd = order Smmd by clicks desc;*

*Top3 = limit Srtd 3;*

*dump Top3;*