Analytics using clickstream Data

Clickstream Data

✓ Clickstream data is an information trail a user leaves behind while visiting a website. It is typically captured in semistructured website log files.

✓ These website log files contain data elements such as a date and time stamp, the visitor’s IP address, the destination URLs of the pages visited, and a user ID that uniquely identifies the website visitor.

Potential Uses of Clickstream Data One of the original uses of Hadoop at Yahoo was to store and process their massive volume of clickstream data. Now enterprises of all types can use Hadoop and the cloudera to refine and analyze clickstream data.

They can then answer business questions such as:

✓ What is the most efficient path for a site visitor to research a product, and then buy it? (Path Optimization)

✓ What products do visitors tend to buy together, and what are they most likely to buy in the future? (Association Analysis & Next product to buy)

✓ Where should I spend resources on fixing or enhancing the user experience on my website? (Allocation of website resources)

we will focus on the “path optimization” use case. Specifically: how can we improve our website to reduce bounce rates and improve conversion?

Input Data

Here’s a summary of the data we’re working with:

✓ Omniture logs – website log files containing information such as URL, timestamp, IP address, geocoded IP address, and user ID (SWID)

✓ The Omniture log dataset contains about 4 million rows of data, which represents five days of clickstream data. Often, organizations will process weeks, months, or even years of data.

✓ Users– CRM user data(registered Users) listing SWIDs (Software User IDs) along with date of birth and gender.

✓ Products – CMS data that maps product categories to website URLs.

Expected Output:

In order to optimize your website and convert more visits into sales and revenue.

✓ Analyze the clickstream data by location

✓ Filter the data by product category

✓ Graph the website user data by age and gender

✓ Pick a target customer segment

✓ Identify a few web pages with the highest bounce rates

Use of Hadoop Ecosystem tools:

Many tools can be helpful to work on this

✓ Flume can be used for stream data to HDFS

✓ Pig also can be load the data from local/HDFS and refine it

✓ Hive can be used to create view/ refine the data

✓ Pig/Hive/impala also can be used for data analytics

✓ Visual Analytics can be performed By integrating Tableau

Sample outputs:-







