iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 3 **Issue:** V **Month of publication:** May 2015

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Visualizing Website Clickstream Data with Apache Hadoop using Hortonworks

Priya Kale[1], Dr. Siddhartha Ghosh[2], Manoj Kumar Danthala[3]
Computer Science Department, Keshav Memorial Institute of Technology (KMIT), Hyderabad

Abstract— Nowadays most of the organizations have turned to Ecommerce which has become a necessary component for business strategy and a catalyst for economic development. These organizations need to predict the analysis about their products and services to track their business from the customers end. The response from the customers based on their activities on the websites decides the future changes required to improve the business values.
These organizations stores the information of all customers in detail for future analysis which is commonly referred as big data, as it is growing at high rates day by day (due to high growing rate of data). One of the main applications of big data intelligence is Clickstream data which is ideal for e-commerce websites and websites that depend on clicks. Clickstreams are records of user interactions with websites and other applications. A typical approach to load these data and processing is by using traditional databases, but it involves many complexities while performing different operations. Here in this paper clickstream data is processed, analysed with the structure of Hadoop using Hortonworks Data Platform (HDP) which provides large scale processing performance and visualized through power view tools.
Keywords— Clickstream, Hortonworks Data Platform, Hadoop, Hive

I. INTRODUCTION

Most of the companies are developing e-commerce websites and mobile apps to advertise and sell their products and services as its fast gaining popularity among people who make purchases online through websites. To view each products and components the people needs to go through different clicks to get the required one's which is known clicks path. It is the sequence of links a site visitor follows in the website.
A clickstream which is also known as click path is the recording of parts of the screen a computer user clicks on while web browsing or using another software application. As the user clicks on anywhere on the webpage or application, the action are logged on the client or inside the web server, as well as possibly the web browser or proxy server. It is typically captured in semi-structure website log files. These website log files contain data elements such as a visitor's identification number, date and time stamp, the visitor's IP address, browser and device information, the destination URLs of the pages visited, and a user ID which uniquely identifies the website visitor, referral page information. Clickstream data generally comes from one of two sources, the logs from servers that originally served the website or internet messages transmitted by JavaScript embedded in pages of the website that are received by a central server.
As an example, consider that we are scanning a click event record. We certainly want to count all clicks (first key-value pair). Suppose the URL fits hierarchically into five different categories within the website then we emit five more pairs with the keys containing the categories. If the user is known to be a male, aged 26,We could then leave one pair for the male aggregate, one for the age range we're bucketing (say 20-29), and perhaps one for males age 20-29. We could emit one pair for each of the five categories combined with gender, age, or both. Reducers then add up the data for each aggregate combination and upload the result to hive where it is available for efficient retrieval.
These complete data is considered as big data because this type of data is increasing highly every day. If we look at the statistics this year, Facebook alone captures 1.5 PB and Amazon captures 200TB of weblog daily. There are many platforms to deal process and analyse these clickstream data like Apache Hadoop, Microsoft's Azure, Cloudera tools, BigInsights, Hortonworks platform etc. These tools perform different operations to stream big data and analyse it. Here in this paper the architecture of Hortonworks is discussed below as we used this tool to load and analyse the clickstream data.

II. HORTONWORKS DATA PLATFORM (HDP)

Apache Hadoop is the open source project governed by the Apache Software Foundation (ASF) that allows you to gain insight from massive amount of structured and unstructured data quickly. Hortonworks Data Platform enables Hadoop Enterprise: the full suite of essential Hadoop capabilities that are required by the enterprise and that serve as the functional definition of any data platform technology. This includes the following set of components and functional areas: Data management, security and server databases.

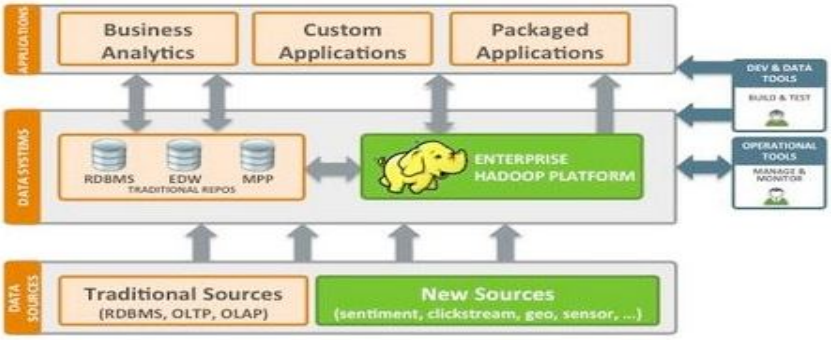


Fig1.1: Hortonworks Data Platform

The architecture consists of three layers: The basic layer data sources includes traditional sources like RDMS, OLAP, OLTP etc. and other external data sources like clickstream, sensor or sentiment data sources. Next layer consists of enterprise tools like Hadoop, hive, sqoop, flume tools to perform and monitor enterprise jobs. And the last layer which is application layer performs different operations. The users and clients interacts with this layer to build, process, stream big data applications in real time.

III.PROBLEM STATEMENT

Every ecommerce business needs to track and analyse clickstream data to grow their business. Many analytics programs, including Google Analytics, come with basic clickstream analysis functionality. In Google it's called as "site overlay". There are so many solutions to deal with this issue. But by using traditional databases to load and process the clickstream data involves complexities while storing and streaming the customer's information. And its take much processing time to analyse and visualize it.

One way to solve this problem and improve performance is by using the structure of Hadoop which provides the large scale fast processing ecosystem environment. There are so many tools to provide this architecture in real time for enterprises. Here in this paper we use Hortonworks to load and process the sample clickstream data. And after analysing it, the data is visualized by power view tools to track the information.

IV. APPLICATIONS AND FUNCTIONALITIES

Clickstream data is composed of thousands and millions or billions of hits that tell the story of how, who and what visitors are doing on the website. Owing your clickstream data is never so easy and affordable. Some of the popular application and functions of this is discussed below.

A. Predictive Modeling

The analysts can create statistical model about the customer's information by using clickstream data for increasing their business values.

B. Customer Analytics

Clickstream data give the complete data about the customers while visiting the website like ip address, web url, date and time stamp. We can find the likes and dislikes of a particular customer with this data. This improves the productivity of business. Also the developers can track the flow of customers per day or per week and the products they are interested in.

C. Path Analysis

To get a specified product or service, the customer needs to go through a flow of web clicks which forms a web path. The analysts can create the pictorial representations of these path to find the products they are searching in an ecommerce website.

D. Website Resource Allocation

A major goal of the marketers is determining how best to allocate website results for better optimal results. Clickstream data tells marketers which paths are clumsy and which are not. This data makes the companies where they are needed most in order to optimize user experience.

E. Customer Life Time Analysis

It is the total amount of value derived by a business from a customer over the customer's complete lifetime engaging with the business or Product Company.

F. Market basket analysis

The benefit of basket analysis for marketers is that it can give the better understanding about the customers to aggregate the product purchasing behaviour. Here in ecommerce websites the click paths and flow of the interested products gives the analysis to the developers and analysts.

G. Customer Segmentation

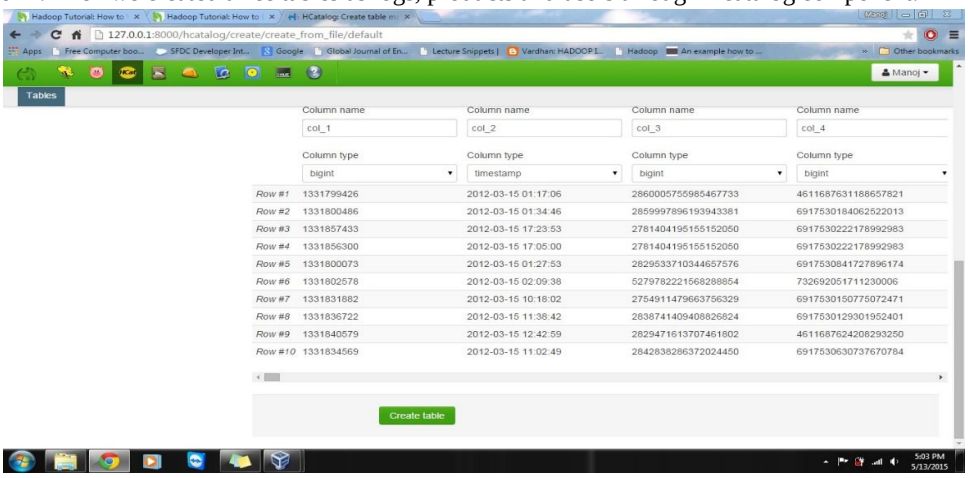
Analysis of clickstream and other user data gives marketers a granular look at how individual customer segments are using the website. This gives the insights to help personalize the user experience and convert more web visitors from browsers to buyers.

V. METHODOLOGY

A discussed above, Many analytics programs, including Google Analytics, come with basic clickstream analysis functionality. And there are so many other sources to load, analyse the clickstream data and visualize it. Here in this paper, we used Hortonworks Data Platform (HDP) to stream the sample clickstream data. The following steps discusses our implementation of clickstream data in real time.

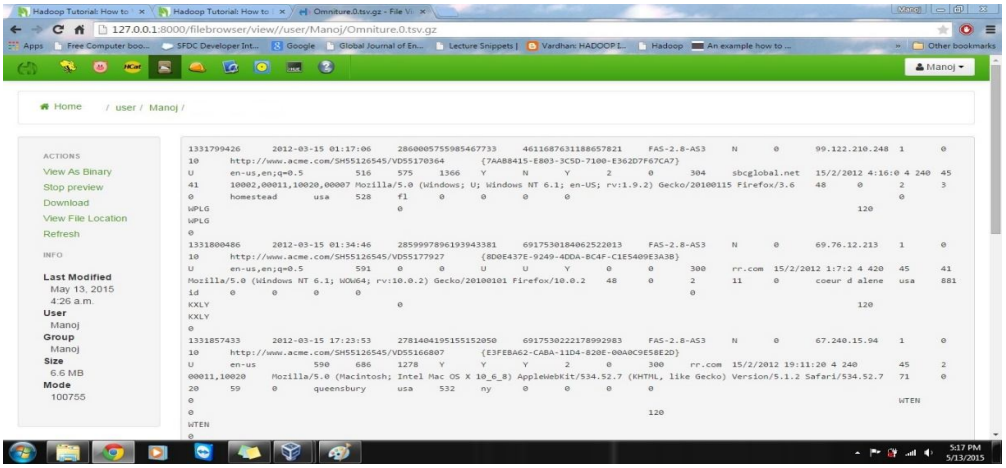
A. Load Data into hortonworks

In this paper sample customer's website clickstream data is taken for analysis which contains the date and time stamp, IP address, web url's of pages visited, and the user id of the customer. Here the first step is to upload the clickstream data into hortonworks platform. Then we created three tables as logs, products and users through Hcatalog component.



B. View the website clickstream data in hortonworks

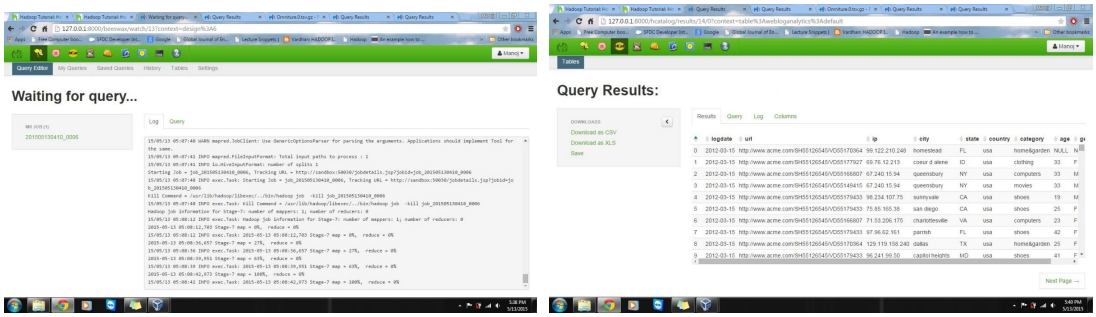
After uploading the data into hortonworks, we can view the clickstream data from the file browser in three tables by a hive query. Logs table contains the information such as IP address, session ID, web url and timestamps. Here user table consists of swid, birth date, gender information. Products table contains the website url and category.



C. Aggregate the CRM data

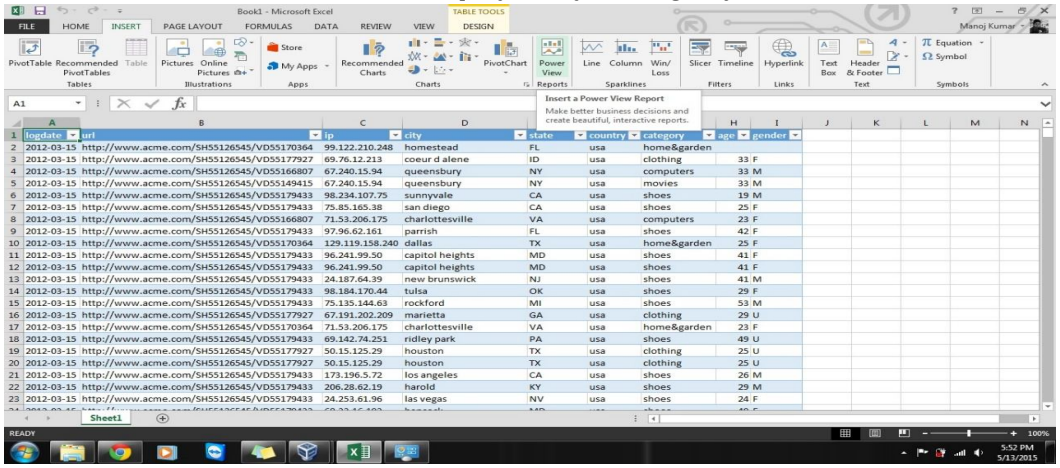
Now we need to combine the clickstream data of three table's logs, products and users. The hive query script makes this happen

through query editor. This aggregates the data from three tables and stores into the specified file from the query. Now we can view this joined clickstream data from file browser.



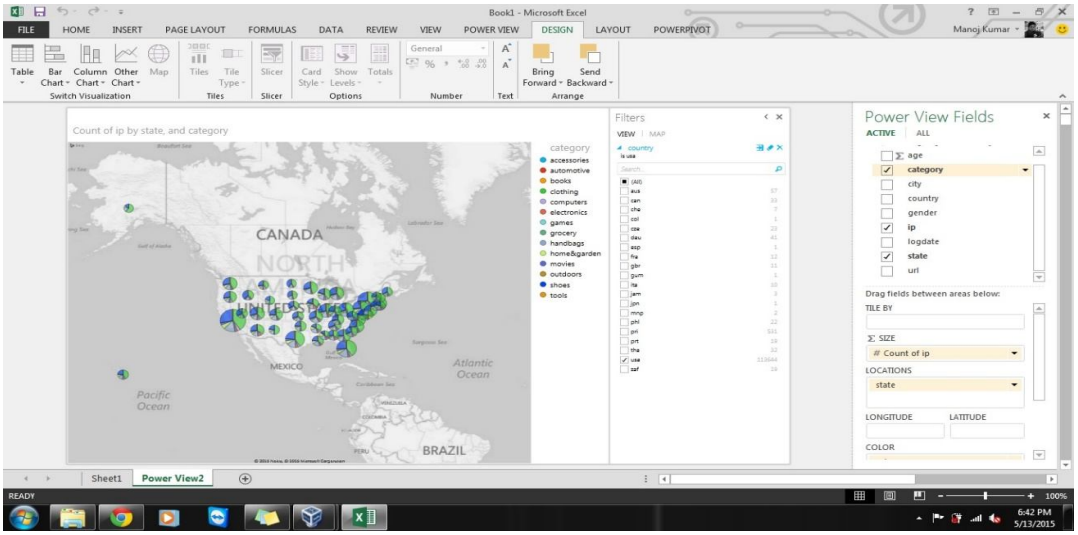
D. Access the data with Excel

Now go to Microsoft query from the data sources choose hortonworks hive DSN which created in the ODBC connection. Now we can access the hortonworks data into Excel sheet from query tool by selecting the joined data file from the wizard.



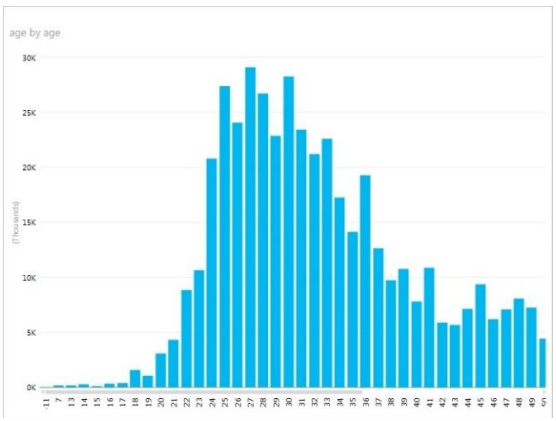
E. Visualize the website clickstream data with power view tool

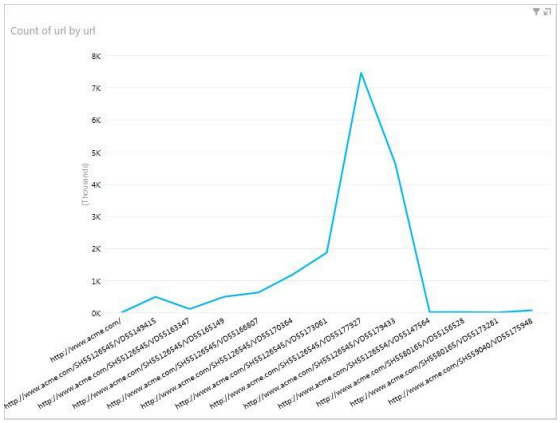
Data visualization helps to optimize the website and improve the business sales and values. Here the clickstream data is visualized by using power view tool. This contains three components: field area, filter, and size. First select the country from the field's area and click on map to locate the different countries from the map view. Here we can point a particular country by select it in the checkbox. And we can categorize the map visuals by choosing different gender, IP addresses, and category etc. Here in this paper the web urls and the number of times the people hits that url's are classified and shown in the graph through clustered column view which reduces the bounce rate and customer defections.



Volume 3 Issue V, May 2015
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

VI. PERFORMANCE RESULTS





VII. FUTURE WORK

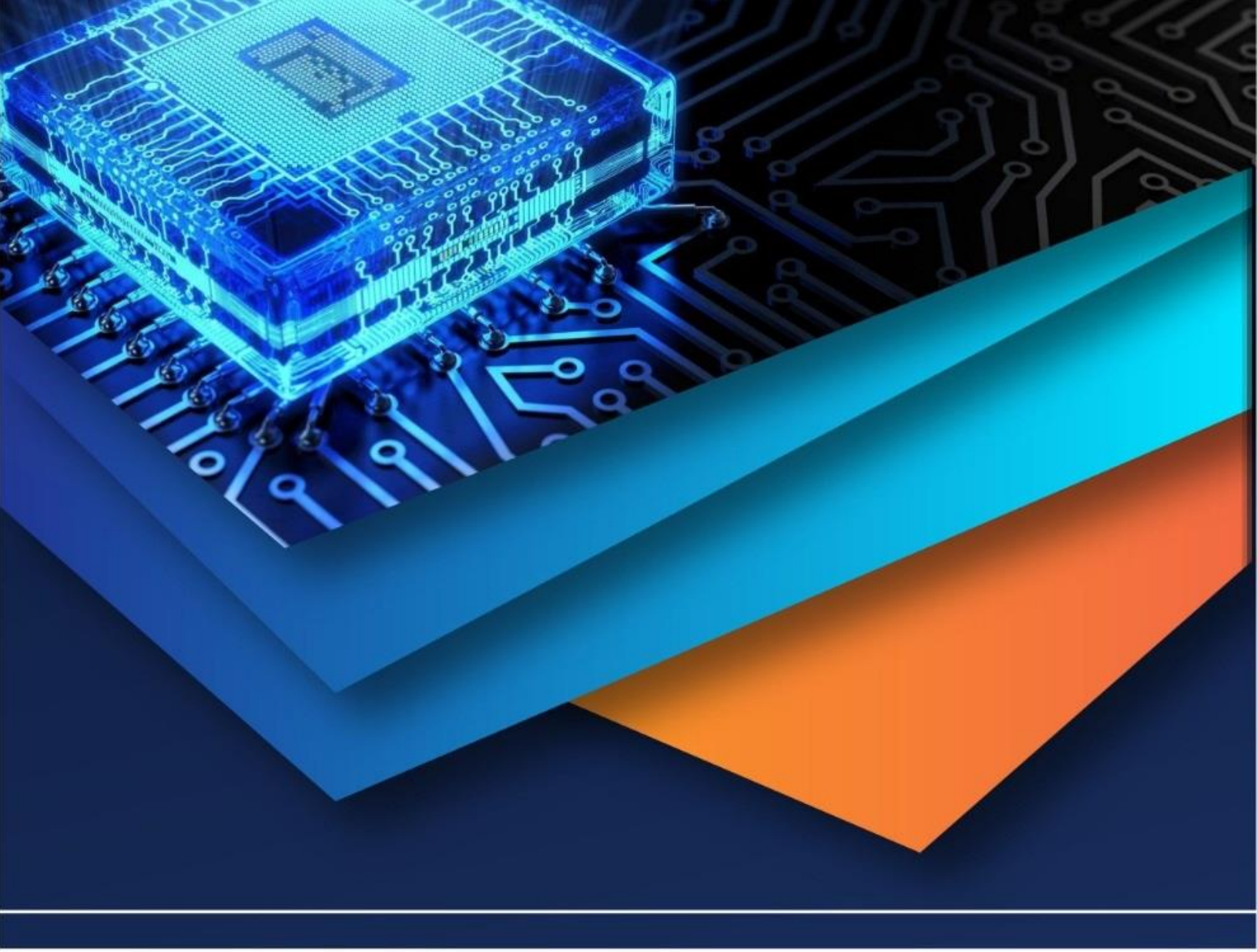
Every organization uses the clickstream data and paths for personalization years ago by recommending products based on the browsing behaviour and purchasing history. As this data is counted as big data, it needs big data intelligence tools to load and analyse the information to improve the customer retention. In this paper, it is refined and processed though hortonworks data platform (HDP). There are so many other tools to solve the problems of big data challenges and analyse it. This work can be enhanced to other BI tools which outputs different performances to improve the sales and customer retention in all aspects. This enables analysts to make better insights to the websites for their business needs.

VIII. CONCLUSIONS

Online retailers commonly have relied on methods to inform product recommendations which define the predictive modeling based on clickstream data and transaction data from the customers. Here in this paper sample of clickstream data is taken and analysed by hortonworks platform which provides the fast processing capability using Hadoop structure. This takes inputs as clickstream data, refines the required data by hive queries and gives joined output of clickstream data elements. Then this data is visualized through power tools which give the different types of classification and improves the website performance. Thus Hadoop has made it easier to analyse, visualize and ultimately change how visitors behave on website.

REFERENCES

[1] R.Rathipriya 1a , Dr. K.Thangavel 1b, J.Bagyamani2c 1 Department of Computer Science, Periyar University, Salem, Tamilnadu, India, "Evolutinary Biclustering of Clickstream Data", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 1, May 2011.
[2] Panagiotis Antonellis, Christos Makris, Nikos Tsirakis ∗ Dept. of Computer Engineering & Informatics, University of Patras, 26500 Rio-Patras, Greece, " Algorithms for clustering Clickstream data".
[3] Jesper Andersen Rune S. Larsen, Anders Giversen Torben Bach Pedersen, Allan H. Jensen Janne Skyt, "Analyzing Clickstreams Using Subsessions", ACM 0-89791-88-6/97/05.
[4] Jishang WeiDavis Zeqian Shen eBay Research Labs Neel Sundaresan Kwan-Liu Ma, University of California, "Visual Cluster Exploration of Web Clickstream Data".
[5] Christopher G. Brinton and Mung Chiang Department of Electrical Engineering, Princeton University, "MOOC Performance Prediction via Clickstream Data and Social Learning Networks".
[6] http://thinkbig.teradata.com/applying-big-data-analytics-to-clickstream-data/
[7] http://www.opentracker.net/article/clickstream-or-clickpath-analysis
[8] http://hortonworks.com/hadoop-tutorial/how-to-visualize-website-clickstream-data/
[9] http://www.qubole.com/blog/big-data/clickstream-data-analysis/



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY