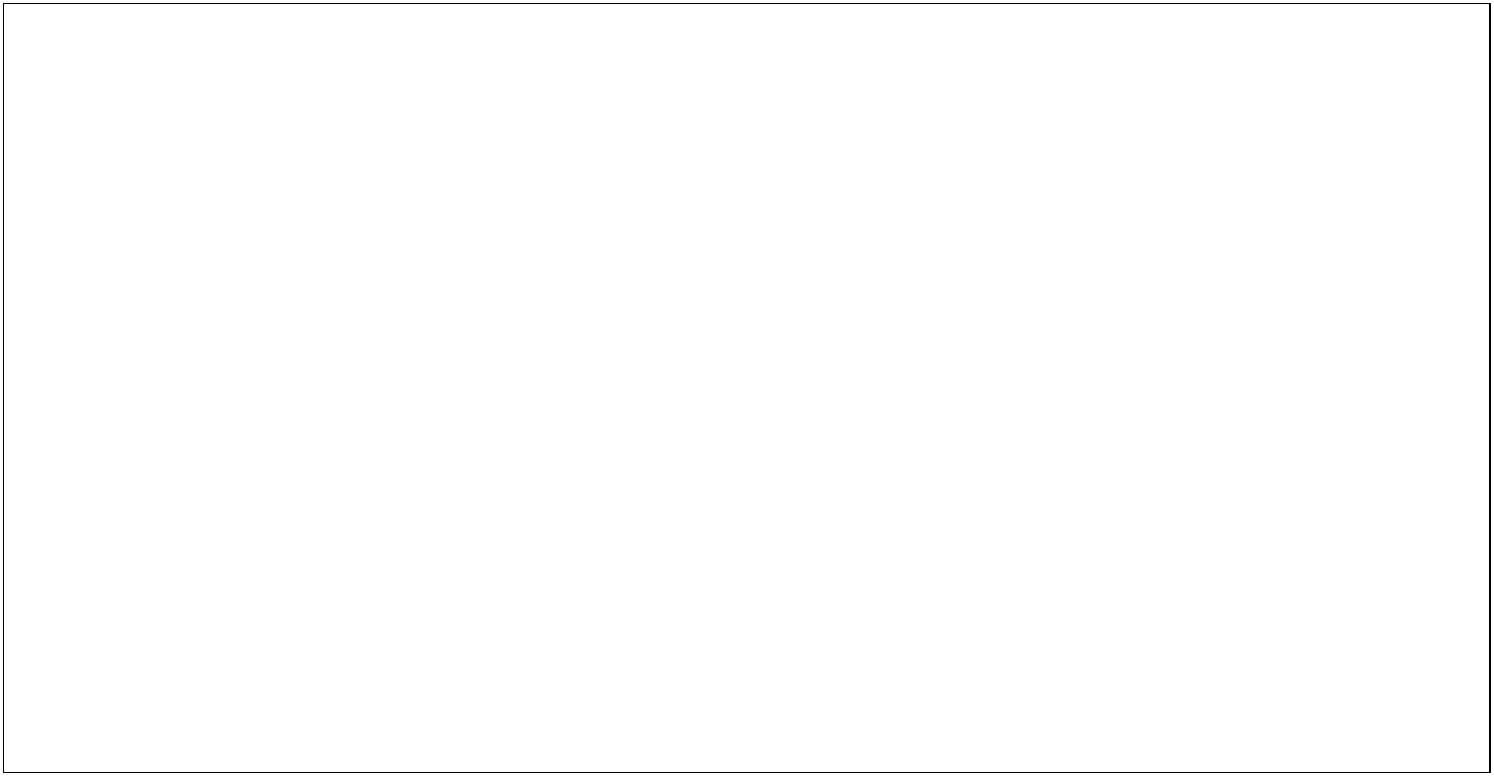
**Visualization and Analysis of Click Stream Data**

**Internal guide**

**Ms. D. Haritha**

M. Samyuktha(2451-12-733-047)



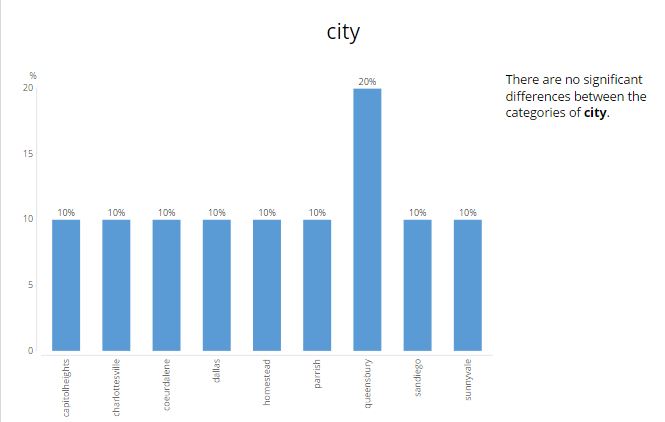
**ABSTRACT:**

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. Click streams 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 project click stream data is processed, analyzed with the structure of Hadoop using Cloudera Data Platform (CDH) which provides large scale processing performance and visualized through power view tools.

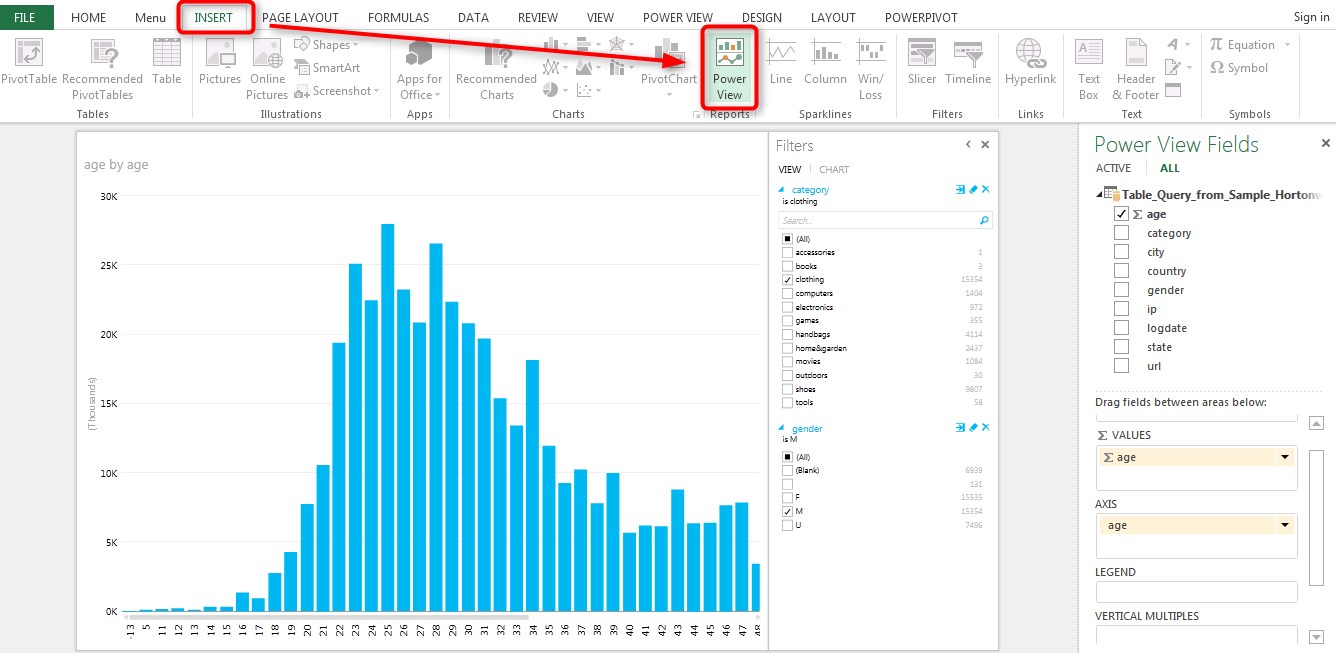
**Bar Chart**

The Bar Chart displays the Users who belong to different cities



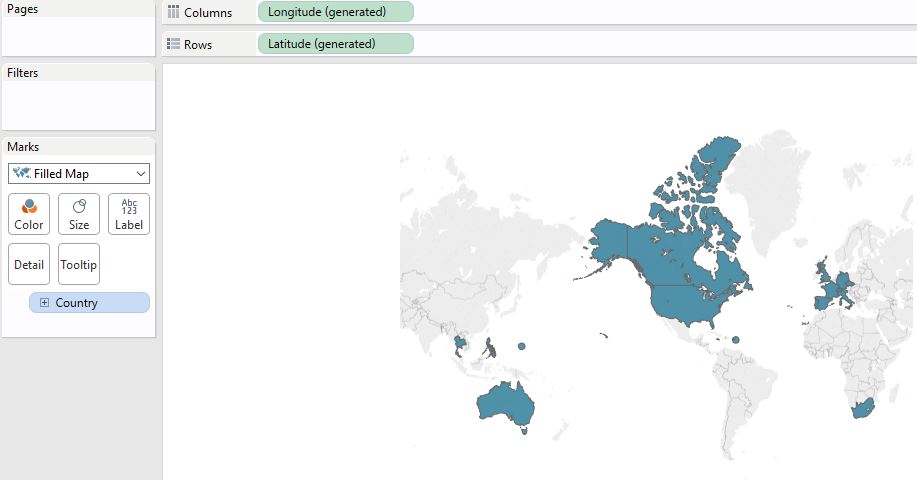
**Bar Chart**

Bar Chart to optimize market performance



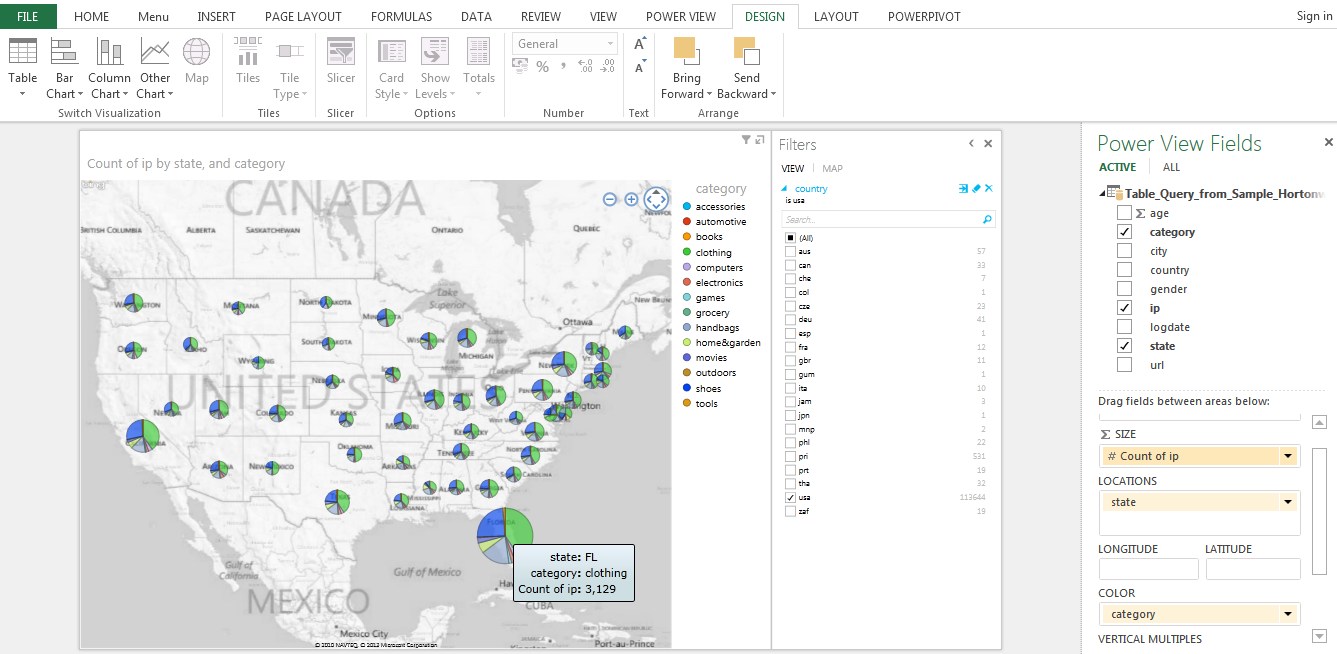
**Map Chart**

The Map Chart shows Country Wise Click Rate



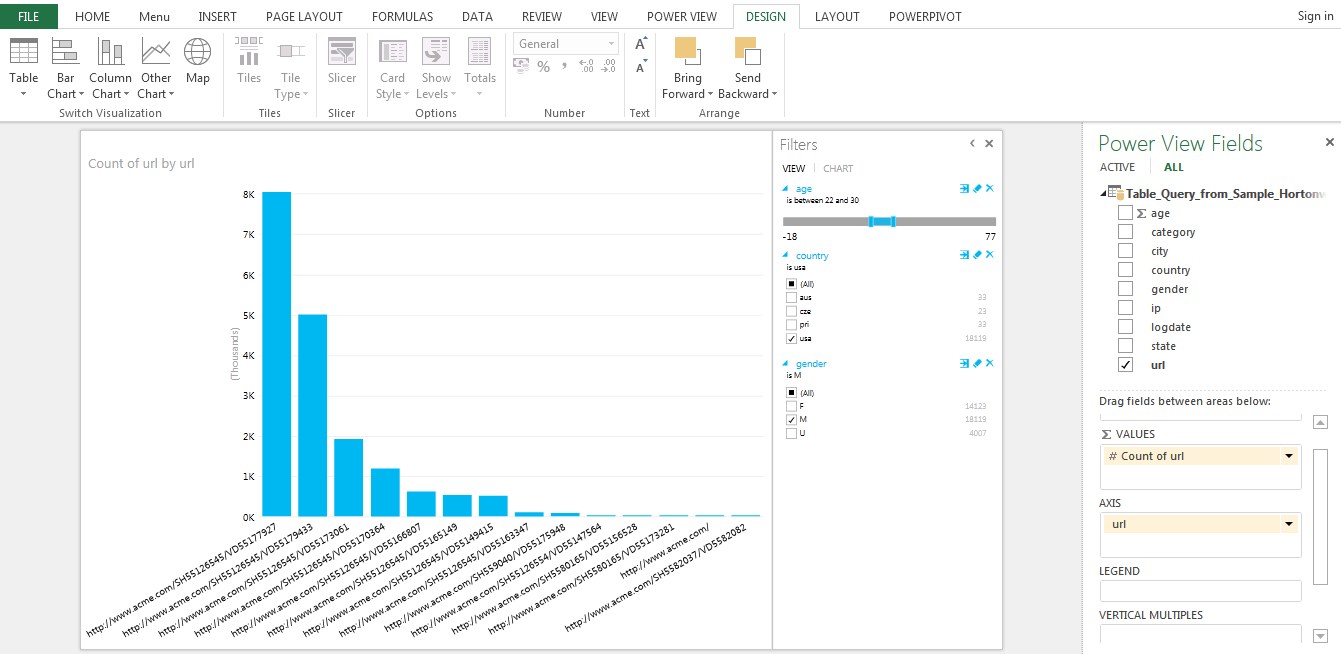
**Map chart**

The Map chart displays the product categories by color for each state. We can see that the largest number of page hits in Florida were for clothing, followed by shoes



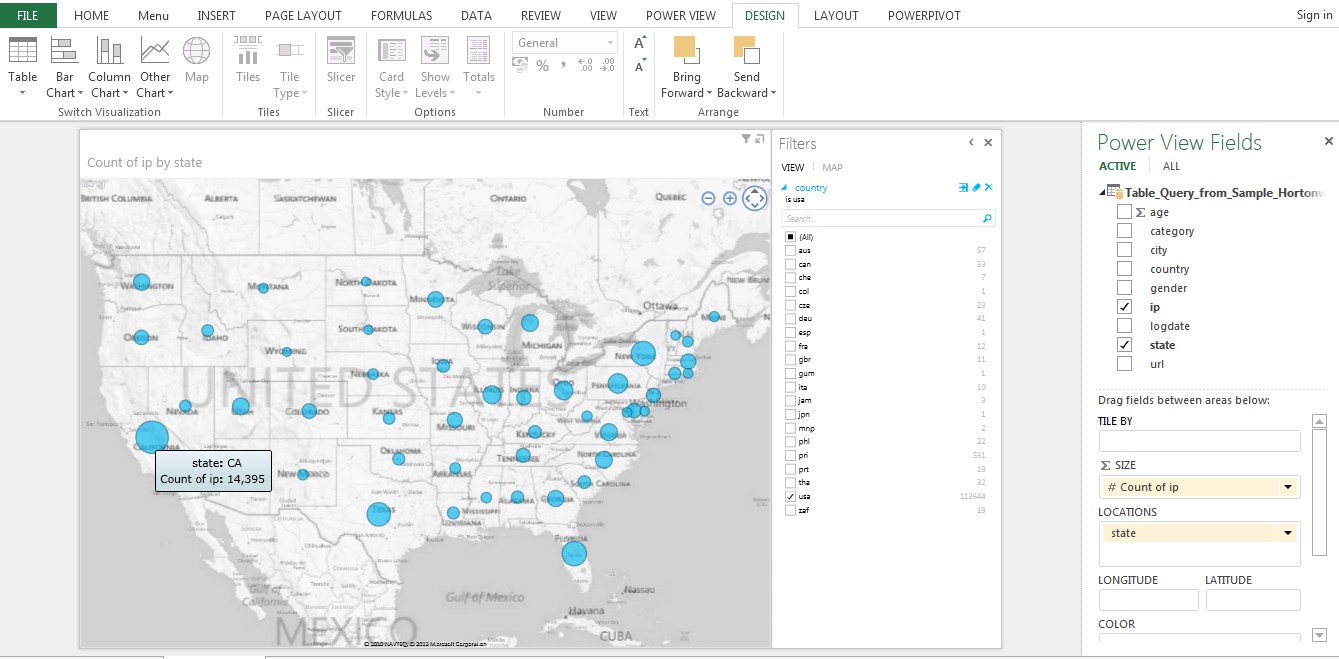
**Bar Chart**

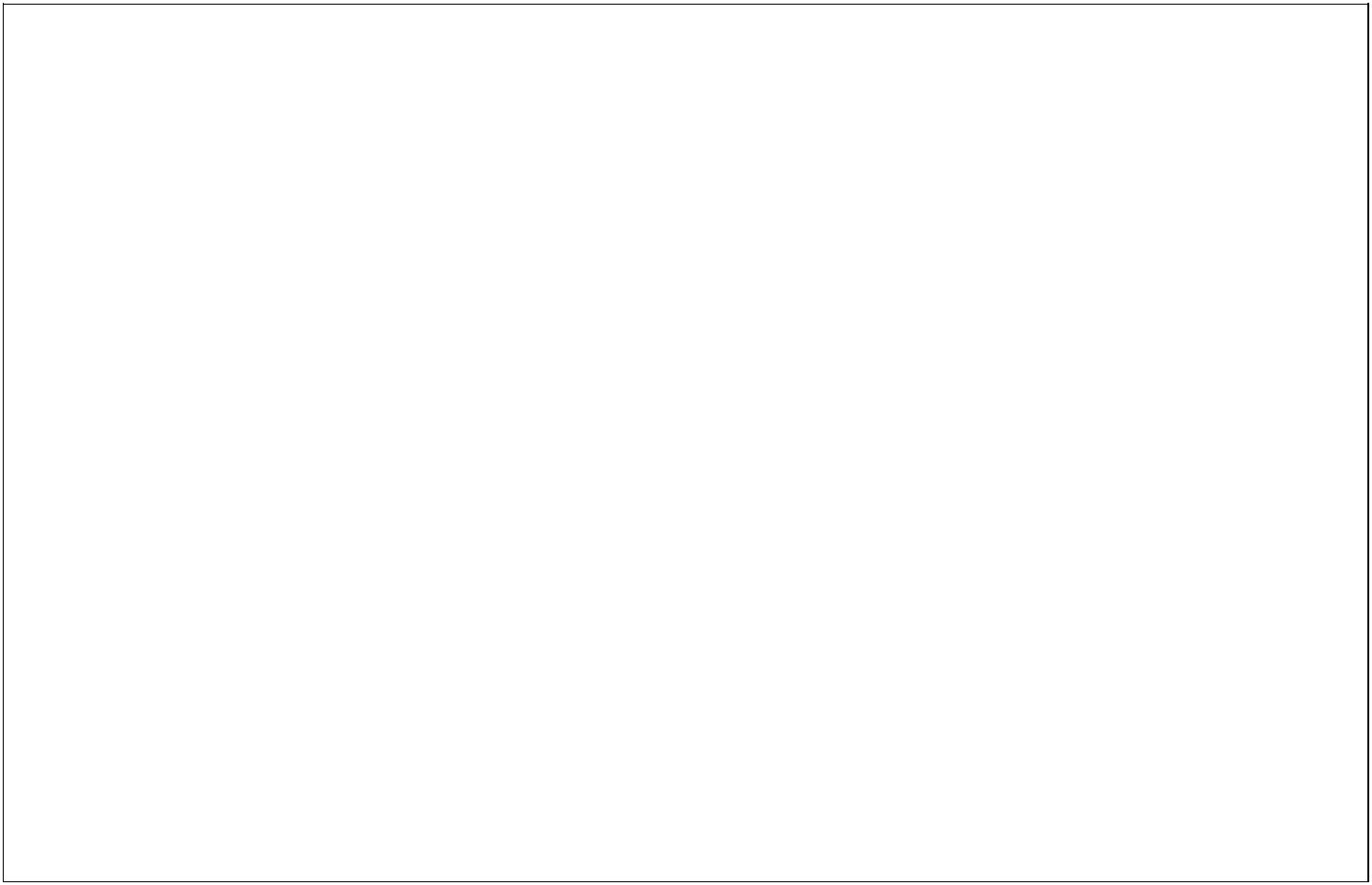
Bar Chart to display Count of URL in ascending order



**Map Chart**

Map Chart to Display the IP count for each state in US

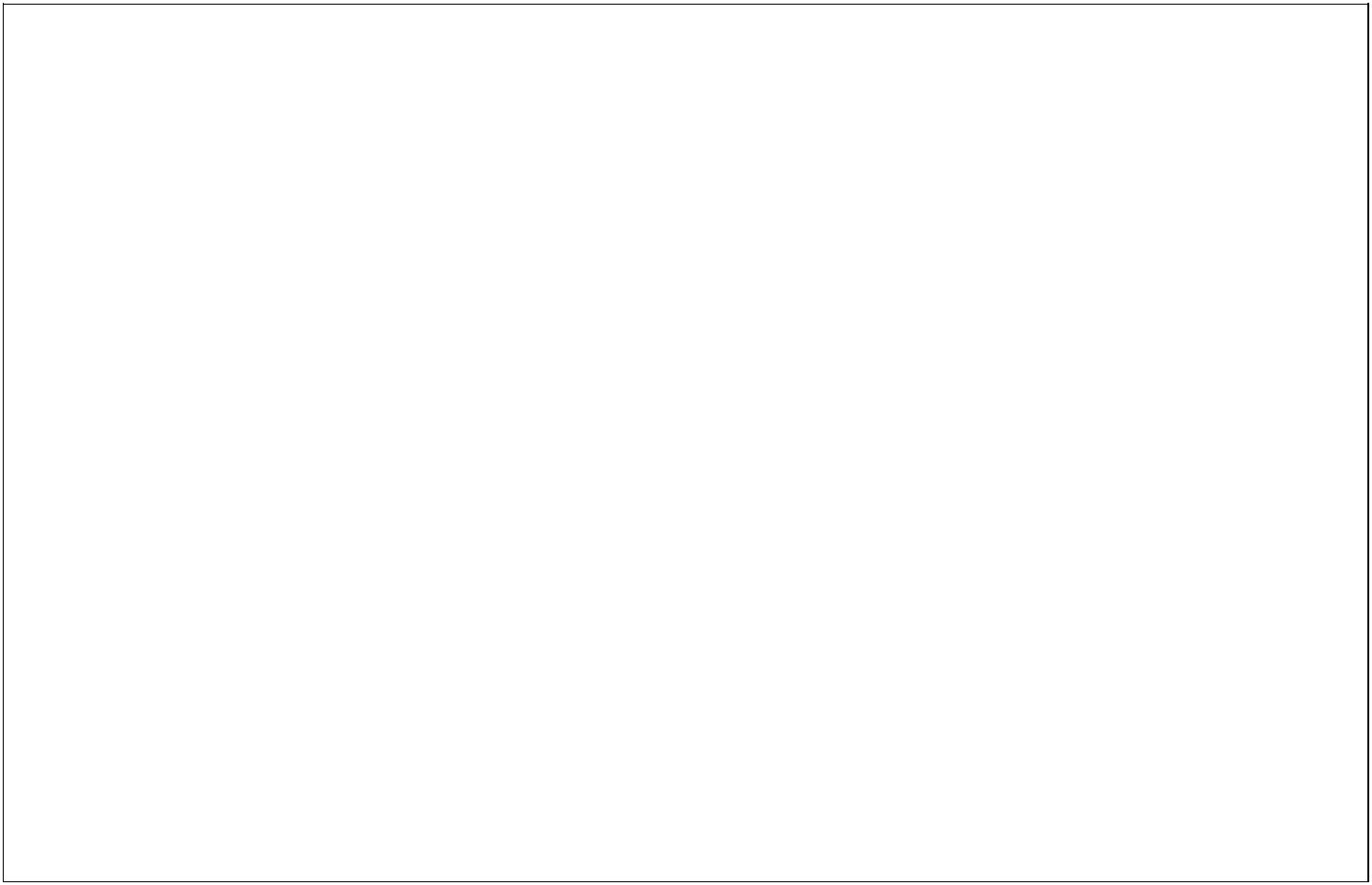




**CONCLUSION**

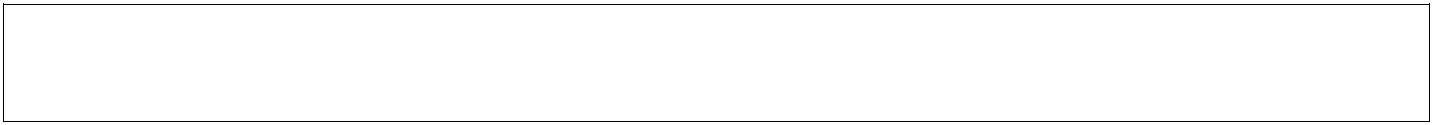
In this project we applied Hadoop MapReduce programming model for analyzing web server log files where data get stored on single node in a cluster so that access time required can be reduced and MapReduce works for small datasets giving efficient results. In order to have summarized results for a particular web application, we need to do log analysis that will help to improve the business strategies as well as to generate statistical reports. Using Tableau Visualization tool for log analysis will provide us graphical reports showing hits for web pages, user’s activity, in which part of website users are interested, traffic sources, etc.

Clickstream data are defined as the electronic record of Internet usage collected by Web servers or third-party services. We have discussed the nature of clickstream data, noting key strengths and limitations of these data for research in marketing. The project reviews major developments from the analysis of these data, covering advances in understanding (1) browsing and site usage behaviour on the Internet, (2) the Internet’s role and efficacy as a new medium for advertising and persuasion, and (3) shopping behaviour on the Internet (i.e., electronic commerce). We have outlined opportunities for new research and highlight several emerging areas likely to grow in future importance. Inherent limitations of clickstream data for understanding and predicting the behaviour of Internet users or researching marketing phenomena are also discussed.



**Future Enhancement**

In contrast to e-commerce settings, we investigate “non-transactional websites” that serve predominantly as a product catalogue while orders are taken offline. Many business-to-business (B2B) settings as well as some business-to-consumer (B2C) settings fall in this category. Specifically, this clickstream data has been a sample log files of a US manufacturer of industrial products, hereafter referred to as “the company.” The recent fast-growing research using clickstream data has already demonstrated the great interest and importance for e-commerce firms. The same applies to offline-selling firms. Understanding consumer online browsing behaviour and its value helps firms make investment decisions regarding the adoption of clickstream tracking technology., the website is non-transactional and the company sells its products offline, either direct or through dealers.



Software Specifications:

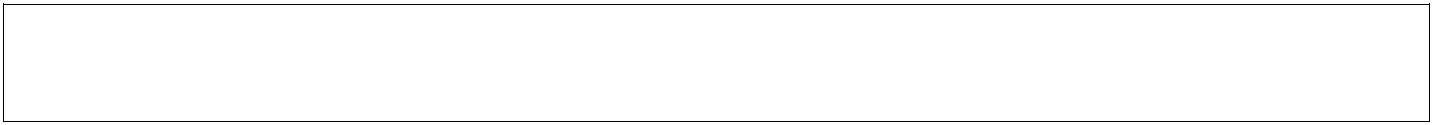
Tableau Professional Edition

MS Excel 2007/2010/2013

Apache Hive

Apache Hadoop

MySQL Database



Hardware Specifications:

4GB RAM

3GB of disk space

32 / 64-bit CPU that supports 32 / 64bit virtualization

Oracle VM Player / Virtual Box

OS Requirements:

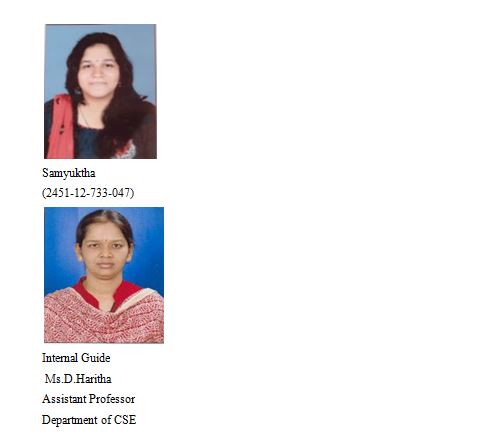
Red Hat Enterprise Linux (RHEL) v5.x or 6.x (64-bit)

Centos v5.x or 6.x (64-bit)

SUSE Linux Enterprise Server (SLES) 11, SP1 (64-bit)

CDH Hadoop 1.x.x

Project Team:



|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |