Halting the Envious Activities using HADOOP

Vivek Kumar[1] Gaurav Saste[2] Mangesh Babar[3] Meenakshi Sawant[4]

[1,2,3,4]Trinity Academy of Engineering

Abstract— Today large Number of peoples is interacting with the online application system that can perform fraud activities or real activities. So it's not efficient for identifying which users are real among them. It's one of the major and most popular aspect is online purchasing. People nowadays don't hesitate to make money transactions online, for which new security measures are to be credited. But there is always a counter measure to threat the security of the system, such as fraudulent activities. For this use of click stream analysis proves beneficial. Click stream analysis works on the clicking pattern of users. Logs of user activities are stored using Hadoop, effective for processing and analyzing big data, which uses its Map Reduce feature for categorizing data on the key values provided. This helps in identifying normal and abnormal user behaviors which in turn helps in preventing the malicious activities which are carried out on online applications. This technique can be used to avoid the losses incurring from them and enhance the security from business perspective.

Key words: Map Reduce, HDFS, Click Stream Data, Web Log Analytics, Fraud Users

I. INTRODUCTION

With the modern of internet, all the traditional activities are being digitized. This has made all the things accessible with just a click. Computers along with internet are the new gateway to the world. We have privilege of having all the necessary things in our place. One of the new inventions led by digitization is e-commerce. E-commerce includes the transaction where money is involved. Money can be used virtually for almost everything we need. But with the great power of using money comes the great responsibility of securing it.

As the money came in, frauds followed. It has become crucial to take effective measures to ensure the safety of the money involved. In online application involves invalid address, purchasing the commodity only to make it unavailable to other customers and all the activities which hamper the applications performance. Many inventions have been made to make it secure but hackers have to be found to outsmart the developers each time. Click stream analysis, where clicks are used to determine the user behavior is new powerful technology to restrict the fraudulent activities. In click stream analysis, the pattern of clicks is studied by generating the logs for sessions of user activity. These logs are then used to differentiate amongst the genuine user and fraud user. This helps to alert the administrative authorities about the malicious activity. Suspicious users are then made to go through some more test of verification, if these users get through these tests successfully the service is provided else they are added to black list, a list of fraud users. This list can later be used to avoid the loss incurring through these users.

II. EXISTING SYSTEM

Existing Sybil defense schemes work by analyzing local communities and Sybil Guard exploits this property to bind the number of identities a malicious user can create and that show the effectiveness of Sybil Guard analytically.

A. *Drawbacks*

- Huge amount of Distributed cache required.
- Implements Fuzzy K Mean algorithm for searching a string which it's time complexity high.
- It's Not Implemented for Online Shopping System till now.

III. PROBLEM STATEMENT

Fraudulent activities involve breaking the services of a particular online application. This involves provision of invalid address, purchasing the commodity only to make it unavailable to other customers and all the activities which hamper the applications performance. Many inventions have been made to make it secure but hackers have to be found to outsmart the developers each time. Obviously huge amount of users' list are made. So maintaining and accessing this type of list, isolating the real users a fraud users list are not efficient for Administrator or database manager and it's a time consuming process.

IV. PROPOSED SYSTEM

In our system, overcomes the drawback of existing system. It has advent features which are easily accessing, managing, isolating and storing the users list. It is a beneficial for Administrator and service providers. It is possible using the modern technology Click Stream Analysis where clicks are used to determine the user behavior is new powerful technology to restrict the fraudulent activities. In click stream analysis, the pattern of clicks is analyzed by generating the logs for sessions of user activity. These logs are then used to differentiate amongst the genuine user and fraud user. This helps to alert the administrative authorities about the malicious activity. Suspicious users are then made to go through some more test of verification, if these users get through these tests successfully the service is provided else they are added to black list, a list of fraud users. Finally generated lists are provides to the administrator or product producer.

A. *Features*

- Fraudulent activities can be decreased significantly.
- Stores large database at the same time it can analyze the data using Map Reduce Algorithm.
- Hadoop processes data fast which is very useful for Real Time Systems.
- Click Stream Analysis generates large database as user can navigate through the webpage anywhere and for any long time.
- Provides very high detection accuracy on our click stream traces.

V. System Requirements

A. Hardware Requirements

- System - Intel 2.40GHz
- Hard Disk – 160GB, Expandable
- RAM – 4GB
- Cache Memory – 4MB

B. Software Requirements

- Operating System – Apache Server Version 2.2.6, Linux Version(Ubuntu)
- Programming Environment – JDK, Hive, Flume
- Programming Languages – Java, Pig Script
- Web Browser – Internet Explorer 6.0 or above

VI. Algorithm

A. Knuth–Morris–Pratt algorithm (KMP)

The Knuth–Morris–Pratt string searching algorithm (or KMP algorithm) searches for occurrences of a "word" W within a main "text string" S by employing the observation that when a mismatch occurs, the word itself embodies sufficient information to determine where the next match could begin, thus bypassing re-examination of previously matched characters.

B. Background

- A string matching algorithm wants to find the starting index m in string S [] that matches the search word W [].
- The most straightforward algorithm is to look for a character match at successive values of the index m, the position in the string being searched, i.e. S[m].
- If the index m reaches the end of the string then there is no match, in which case the search is said to "fail". At each position m the algorithm first checks for equality of the first character in the searched for word, i.e. S[m] =? W [1].
- If a match is found, the algorithm tests the other characters in the searched for word by checking successive values of the word position index, i. The algorithm retrieves the character W[i] in the searched for word and checks for equality of the expression S [m+i] =? W[i].
- If all successive characters match in W at position m then a match is found at that position in the search string.

```
int x=0;
int[] next = new int[M];
for (int j=1;j<M;j++)
{
If(p.charAt(x) == p.charAt(j))
{
next[j] = next [x];
x = x+1;
}
else
{
next[j]=x+1;
x=next[x];
}
}
```

Takes Time and Space proportional to length.

VII. Technology Used

A. Hadoop

The Hadoop framework transparently provides both reliability and data motion to applications. Hadoop implements a computational paradigm named Map Reduce, where the application is divided into many small fragments of work, each of which may be executed or re-executed on any node in the cluster. In addition, it provides a distributed file system that stores data on the compute nodes, providing very high aggregate bandwidth across the cluster. Both map/reduce and the distributed file system is designed so that node failures are automatically handled by the framework.

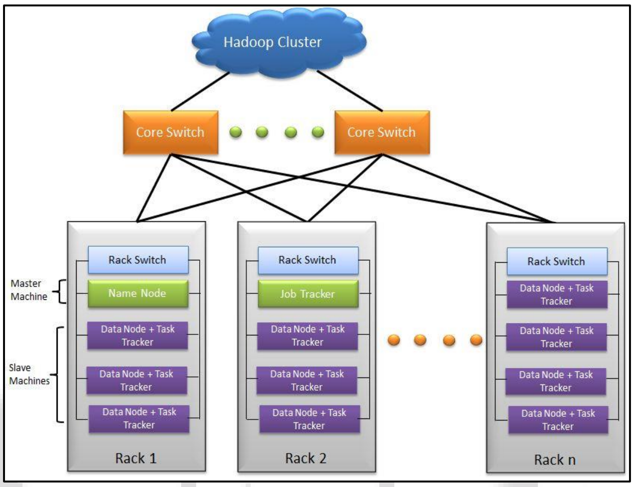


Fig. 1: Hadoop Cluster

1) In Hadoop there are two main components:

- Hadoop distributed file system(HDFS)
- Map Reduce

B. HDFS

HDFS is a file system designed for storing very large files with streaming data access patterns, running on clusters on commodity hardware. It has many similarities with existing distributed file systems. However, the differences from other distributed file systems are significant. HDFS is highly fault-tolerant and is designed to be deployed on low-cost hardware. HDFS provides high throughput access to application data and is suitable for applications that have large data sets. HDFS relaxes a few POSIX requirements to enable streaming access to file system data. HDFS was originally built as infrastructure for the Apache Nutch web search engine project. HDFS is part of the Apache Hadoop Core project. HDFS performs the write once and read multiple operations. Its accessing speed is very fast and automatically maintains multiple copies of data, deploying processing logic in the event of failure.

1) HDFS Functions:

- Very large files.
- Streaming data access.
- Commodity hardware.
- Low-latency data access.
- Lots of small files.
- Multiple writers, arbitrary file modifications.

2) HDFS Cluster Types:

- Name Node(Manage File system Namespace)
- Data Node(Access Data from Name Node)
- Edge Node(Communication Link)

C. Map/Reduce

Map Reduce provides a programming model that abstracts the problem from disk reads and writes, transforming it into a computation over sets of keys and values. The approach taken by Map Reduce may seem like a brute-force approach. Map Reduce works well on unstructured or semi structured data, since it is designed to interpret the data at processing time.

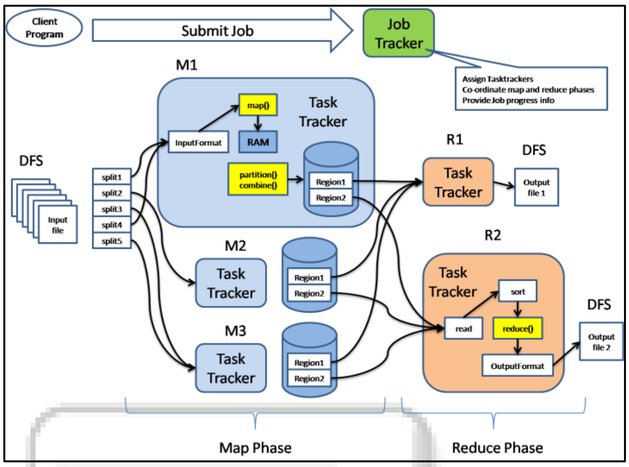


Fig. 2: Map/Reduce Process

1) Map/Reduce Components

- Job Tracker
- Task Tracker

2) Map/Reduce operation

- Splitting of Data.
- Sorting of Data.
- Merging of Data.

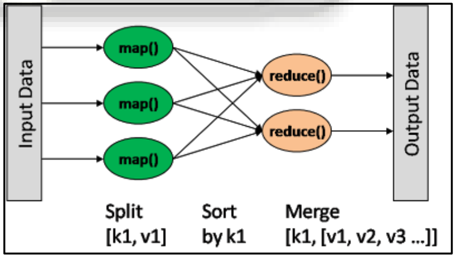


Fig. 3: Map/Reduce Basic Operation.

VIII. LITERATURE SURVEY

Today near about 40-50 % users are interacting with online application system as a fraud E.g. Online Shopping System. In that huge amount of fraud users are rapidly increases and they reserve more than one product having no specific intention. So it's difficult to know which users are real and which users are frauds among made users list. Hence large number of users list is made and it's tedious task to maintain and isolating the users list and it's time consuming process. To maintaining the huge amount of web log files and managing block of memory requires advent features which are available in Hadoop technology. Overall survey of the papers concludes that they are uses Hadoop technology for storing and accessing the big data set. But it requires large cache memory. It controls click stream analysis which uses click sequence model only. Large scale data are distributed with the help of map technique but at the time of integrating/merging different clusters are more complex.

IX. SYSTEM IMPLEMENTATION

A. Workflow of System

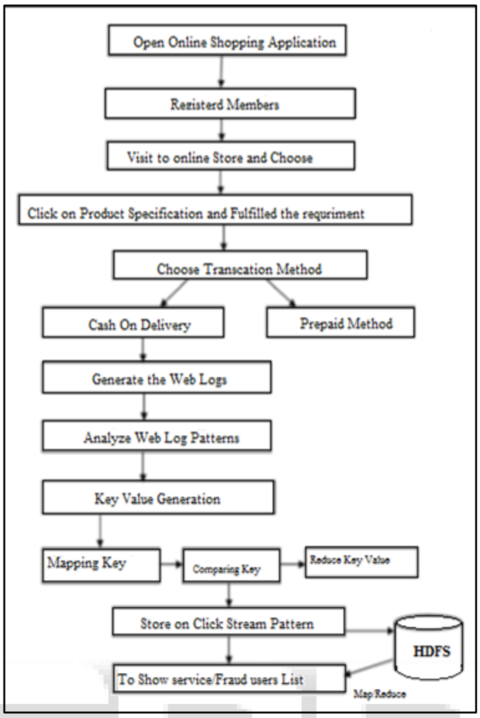
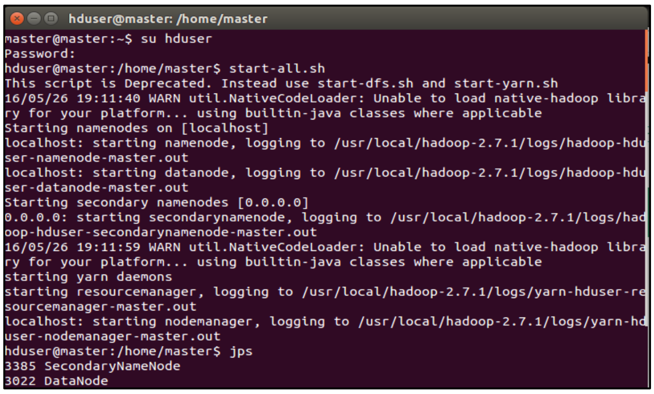

Open Online Shopping Application
Registerd Members
Visit to online Store and Choose
Click on Product Specification and Fulfilled the requriment
Choose Transcation Method
Cash On Delivery
Prepaid Method
Generate the Web Logs
Analyze Web Log Patterns
Key Value Generation
Mapping Key
Comparing Key
Reduce Key Value
Store on Click Stream Pattern
HDFS
To Show service/Fraud users List
Map Reduce

Fig. 4: Workflow of System

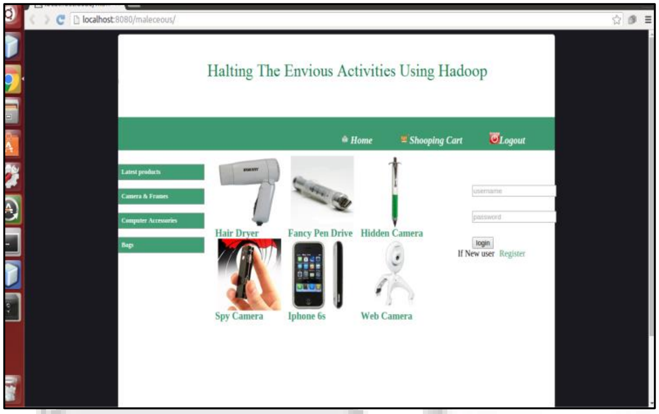
Many users are interacting with the online shopping system they might be real or fraud users. They are first visit online shopping website, observe products and select specific products whatever they want and then read its product specifications. If one of the product like then they are going to purchase this product and choose transaction method. They go on "Cash on Delivery" method. By click on any tab or any option then it's automatically generate web log pattern with the help of click stream analysis. With the help of click stream analysis user's behavior is easily identified. These patterns are analyzed using "Key Value". Key value is nothing but a user name and it's parameter which are mapping, comparing and reduce the key values which helps to identifying and isolating users behavior that are real users and fraud users. All these web logs and resulting files are stored in HDFS which improves the speed of accessing and retrieving the files. In that main function named Map/Reduce is used for improving the performance of response time. So using Hadoop technology these all output data are provides to the Administrator or product producer.

B. *Screenshots of Implemented Project*



```
hduser@master: /home/master
master@master:~$ su hduser
Password:
hduser@master:/home/master$ start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
16/05/26 19:11:40 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
Starting namenodes on [localhost]
localhost: starting namenode, logging to /usr/local/hadoop-2.7.1/logs/hadoop-hdu
ser-namenode-master.out
localhost: starting datanode, logging to /usr/local/hadoop-2.7.1/logs/hadoop-hdu
ser-datanode-master.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop-2.7.1/logs/had
oop-hduser-secondarynamenode-master.out
16/05/26 19:11:59 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop-2.7.1/logs/yarn-hduser-re
sourcemanager-master.out
localhost: starting nodemanager, logging to /usr/local/hadoop-2.7.1/logs/yarn-hd
user-nodemanager-master.out
hduser@master:/home/master$ jps
3385 SecondaryNameNode
3022 DataNode
```

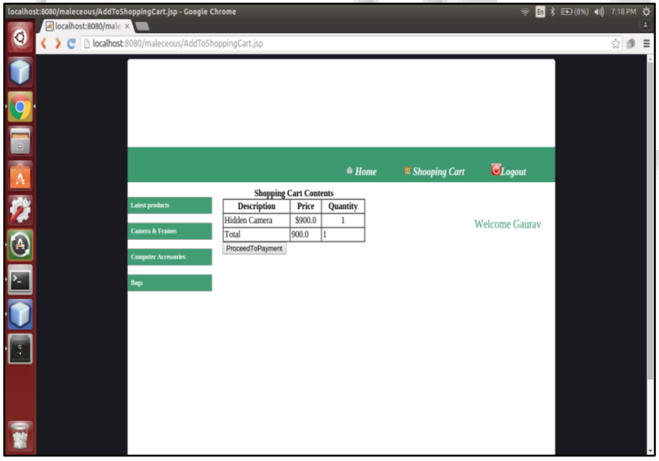
Fig. 5: Starting of Hadoop Environment



localhost:8080/maleceous/

Halting The Envious Activities Using Hadoop

Fig. 6: Home Page



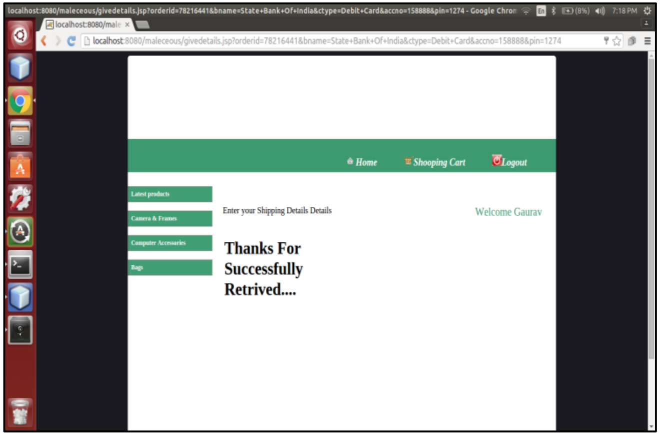
localhost:8080/maleceous/AddToShoppingCart.jsp - Google Chrome

Shopping Cart Contents

Description	Price	Quantity
Hidden Camera	$900.0	1
Total	900.0	1

Welcome Gaurav

Fig. 7: Purchased Item in Cart (Amount < 900)



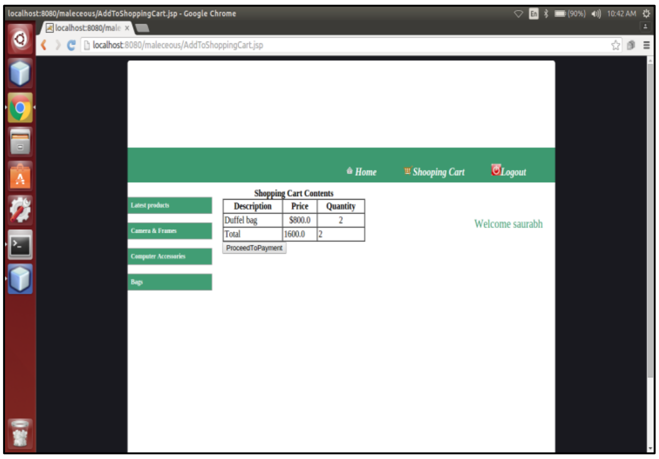
localhost:8080/maleceous/givedetails.jsp?orderid=78216441&bname=State+Bank+Of+India&ctype=Debit+Card&accno=158888&pin=1274 - Google Chrom

Enter your Shipping Details Details

Welcome Gaurav

Thanks For Successfully Retrived....

Fig. 8: Here OTP is not Generated (Amount < 900)



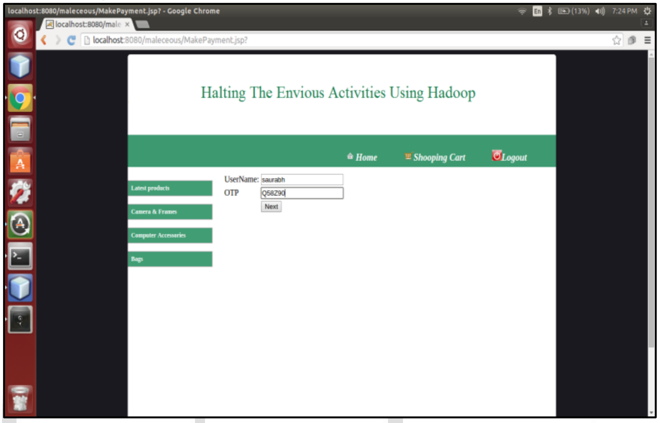
localhost:8080/maleceous/AddToShoppingCart.jsp - Google Chrome

Shopping Cart Contents

Description	Price	Quantity
Duffel bag	$800.0	2
Total	1600.0	2

Welcome saurabh

Fig. 9: Purchased Item in Cart (Amount > 900)

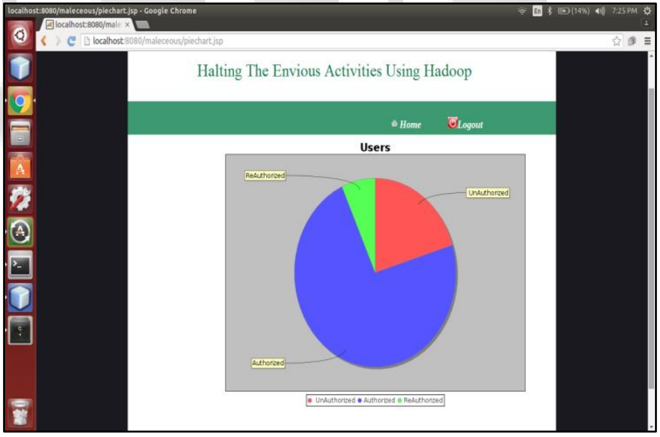


localhost:8080/maleceous/MakePayment.jsp? - Google Chrome

Halting The Envious Activities Using Hadoop

UserName: saurabh
OTP Q58Z90

Fig. 10: Here OTP is Generated (Amount > 900)



localhost:8080/maleceous/piechart.jsp - Google Chrome

Halting The Envious Activities Using Hadoop

Fig. 11: Here Different Types of Users Are Shown

X. CONCLUSION

- Detection of Fraud/Sybil user's activities through Click Stream Analysis which is control by Hadoop Framework.
- Determines the user behavior using Click stream pattern.
- Huge amount of web logs are easily managed and identifies real users and fraud users.
- Separate list are made of Real users and fraud users. The Fraud users are added in the black list.
- Provide finally isolated data to the Administrator which is very beneficial and time saving Manner.

XI. FUTURE SCOPE

- It will be used for Prepaid Transaction in online shopping system.
- Used for hacking Prevention.
- All online application system are used for their security purpose or improving performance of the system.

ACKNOWLEDGMENT

The authors would like to thank all the reviewers and advisors for their helpful suggestions to improve this paper. We would also like to give special thanks to our HoD. Prof. S. N. Maitri madam for her valuable guidance and support.

REFERENCES

[1] Qi Chen, Cheng Liu, and Zhen Xiao, Senior Member, IEEE "Improving Map Reduce performance Using Smart Speculative Execution Strategy" Digital Object Identifier 10.1109/TC.2013.15.

[2] Gang Wang, Tristan Konolige, Christo Wilson,Xiao Wang, HaitaoZheng and Ben Y.Zhao "You are How You Click: Click stream Analysis for Sybil Detection" IEEE 2013.

[3] Hongyong Yu, Deshuai Wang "Mass Log Data Processing and Mining Based on Hadoop and Cloud Computing" The 7th International Conference on Computer Science and Education (ICCSE 2012) Melbourne, Australia.

[4] Natalia Danilova, David Stup "Application of Natural Language Processing and Evidential Analysis to Web-Based Intelligence Information Acquisition"2012.

[5] Jing Zhang1, Gongqing Wu1, Xuegang Hu1, Xindong Wu "A Distributed Cache For Hadoop Distributed File System in Real-time Cloud Services" the 13th International conference on Grid Computing 2012 ACM/IEEE.

[6] Hadoop, http://hadoop.apache.org/,2011. chian Premchaiswadi "Extracting Weblog of Siam University for Learning User Behavior on Map-Reduce".

[7] S.Eknayake, J.Mitchell, Y.Sun and J. Qiu, "Memcached Integration and Twister", http://salsahpc.org/CloudCom2010/EPoster/Cloudcom2010_submission_264.pdf.

[8] Randolph E. Bucklina and Catarina Sismeirob a Peter W. "Click Here for Internet Insight: Advances in Clickstream Data Analysis in Marketing" Mullin Professor, UCLA Anderson School, 110 Westwood Plaza, Los Angeles, CA 90095, USA Senior Lecturer, Imperial College Business School, Imperial College, London, UK.Journal of Interactive Marketing 23 (2009).