# An Application of Clickstream Analysis for Sybil Detection Using Hadoop Technology

Prof. A D Londhe, V A Bhalshinge, R V Bhosale
UCOER, Pune

***Abstract* - In our daily life the use of online applications has become very frequent. Everyday large numbers of transactions take place using these applications. E-commerce has made all the money transactions simple, easy and time saving for the people. One of the major aspects in E-commerce is providing security for the important personal information entered by the users while doing money transactions and managing the large amount of data in databases. Studies show that the techniques and databases used are not efficient. The combination of clickstream analysis and map-reduce (Module of Hadoop) improves the performance. This helps in identifying the normal and fraud user behaviour and also the management of large amount of user data.**

***Keywords* - Map Reduce, HDFS, clickstream, Web log analytics, Sybils**

## I. INTRODUCTION

In today's business world there is continuous requirement for updating business process and adopting higher ideas to keep a competitive level high. One of the main concerns for on-line applications is to keep a high Level of security, for which organizations need to keep a constant monitoring of their processes. Most of the time due to overload on the application, it may fail or become unavailable. When monitoring the logs, we could alert the administrator when the process is not expected or is not available or not secure. An optimal response would be to automatically adapt the business process according to the new context in order to maintain it secure. To develop a solution for identifying the fraudulent activities on online application through click stream analysis using Hadoop.

## II. BACKGROUND

The existing system used technologies like CAPTCHA's and graph based technology for Sybil (i.e. fraud user) detection. CAPTCHA stands for "Completely Automated Public Turing Test to Tell Computer and Human Apart." It is the manipulated text the enables or blocks user access to a particular website. It uses the unique ability of human beings to decipher the manipulated text. The graph based technology uses nodes and edges. Node represents the user and edge represents the well-defined relationship between the users. It provides the information in the form of graphs. It is very useful for analysis purpose. Both these techniques are not efficient for detecting the fraud users. CAPTCHAs are routinely solved by dedicated workers. The existing system was based on analysing local communities and used Sybil Guard to find the identities created by malicious user. This system was also not efficient as huge amount of distributed cache was required, it implemented Fuzzy K Mean algorithm for searching a string which has high time complexity and also it is not implemented for online shopping systems till now.

## III.PROPOSED SYSTEM

We are proposing a system that overcomes the drawback of existing systems and is more efficient than the systems currently being used for fraud user detection. Our system is based on clickstream analysis and map-reduce technique of Hadoop. Using clickstream we can analyze the behaviour of the user using weblogs. Weblogs are the clicks of the user in a particular session (i.e. the time he uses the internet). Clickstream focuses on specific activities of user when he uses any of the application running on internet. Map-reduce is used to manage the large amount of data (i.e. in PB, TB) generated. It generates a key and value pair for each weblog. These key and value pairs are mapped and reduced. The group of values having similar key which is the output of map-reduce is stored in the HDFS.

The proposed application that we are going to develop will be used as chief performance application within the different systems that interact with the user. Therefore, it is expected that the application would perform functionality that are specified by the user or any e-commerce application. The system will be more efficient for detecting the Sybils in an online shopping system. HDFS stores large amount of information in clusters and will provide fast processing of such data as we are using hadoop's HDFS and its map-reduce technique. The system will be scalable as it is modifiable without affecting the other components. As we are using HDFS, multiple replicas of data will be present. So any hardware or software failure will not result in the loss of data as replicas of data will be present. The system is portable and it will have the ability to run under different computing environment. Also, the system will be more secure as the raud users can be detected using clickstream analysis and map-reduce techniques. Some disadvantages of our system can be like need of having continuous internet connection and the maintenance cost for distributed database will be higher.

### *A. Features of Proposed System*

1. Fraudulent activities can be decreased significantly.

2. HDFS stores very large amount of database and at the same time it can analyze the data using Map Reduce algorithm.
3. Hadoop processes data fast which is very useful for Real Time Systems.
4. Click Stream Analysis generates large database as user can navigate through the webpages.
5. Provides very high detection accuracy on our click stream traces.

*B. Proposed System Architecture*

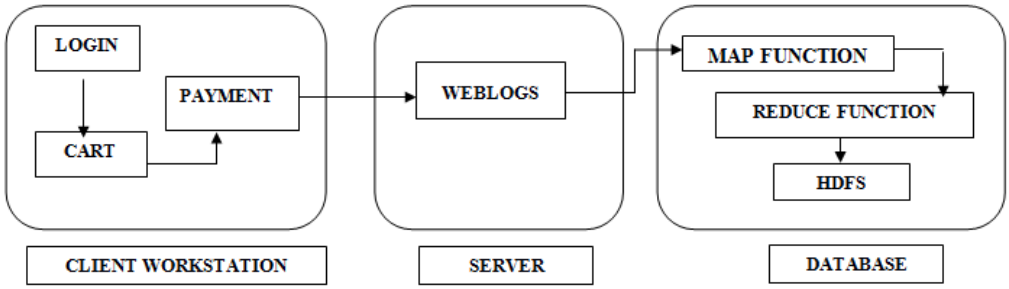


Fig. 1 Proposed System Architecture

Client workstation provides user interaction. User can enter data, read data, and modify data and fire queries for retrieval of data. It communicates with web server to process information. Web server or application server communicates with client workstation and database server. The queries fired by user are processed here. The data required is retrieved from database. Database server stores all the user data in distributed file system of Hadoop. It provides necessary information for processing. 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 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.

## IV. ALGORITHMS

***KMP ALGORITHM***

KMP (Knuth-Morris-Pratt) is a linear time algorithm. It matches pattern 'p' with the elements of the string 'S'. This is achieved by avoiding the comparison of the previously matched elements from string 'S' i.e. backtracking on string 'S' never occurs. It uses two functions: kmp_table and kmp_search.

*A. **Algorithm for kmp_table** :*

**INITIALISE** T [0] = -1, T [1] = 0
**while** position < length (W) do
**if** W [position-1] = W[end] **then**
**Let** cur ← cur + 1,
T[position] ← cur,
position ← position + 1
**else if** cur > 0 **then**
**Let** cur ← T[cur]
**else**
**Let** T[position] ← 0, position ← position + 1

Where,
W - word to be searched
T - table
position - current position of character to be computed in T
curr - zero based index of next character in W

When kmp_table function executes it creates a table which contains the index from which the next search will start after we get a mismatch. For each pattern generated by the system a table of index will be created for efficient search.

*B. **Algorithm for kmp_search** :*

**while** n + k < length(S) **do**
**if** W[k] = S[n + k] **then**
**if** k = length(W) - 1 **then**
**return** n
**let** k ← k + 1
**else**

**if** T[k] > -1 **then**
**let** n ← n + k - T[k], k ← T[k]
**else**
**let** k ← 0, n ← n + 1
where,
n - the beginning of the current match in S
k - position of current character in W
T - table

Kmp_search function is used to match the patterns generated by the system with the patterns stored in the database of the system. If the pattern match is found the system provides the service to the user. If the pattern doesn't match the system shouldn't provide the service to the user.

## V. WORK FLOW OF THE SYSTEM

By click on any tabor any option then it's automatically generate weblog pattern with the help of click stream analysis. With the help of click stream analysis user's behaviour 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 behaviour 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.

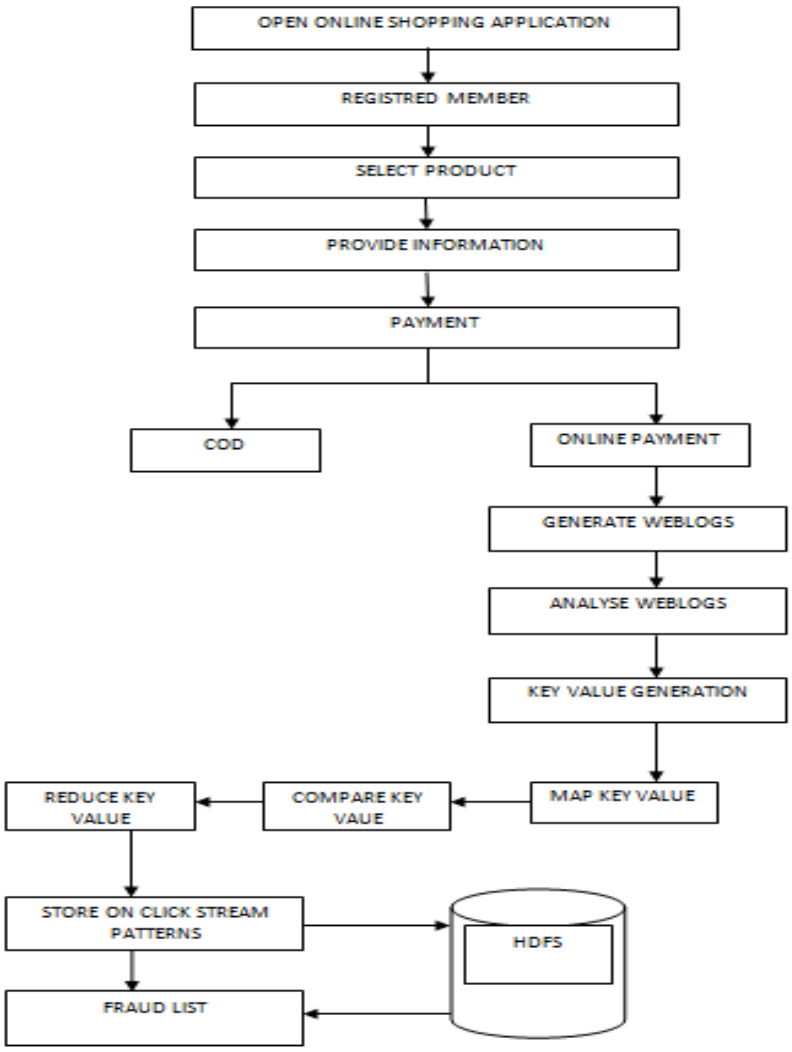


Fig. 2 Workflow of System

## VI. TECHNOLOGY

### Hadoop technology

In a Hadoop cluster, data is distributed to all the nodes of the cluster as it is being loaded in. The Hadoop Distributed File System (HDFS) will split large data files into chunks which are managed by different nodes in the cluster. In addition to this each chunk is replicated across several machines, so that a single machine failure does not result in any data being unavailable. An active

monitoring system then re-replicates the data in response to system failures which can result in partial storage. Even though the file chunks are replicated and distributed across several machines, they form a single namespace, so their contents are universally accessible.

### A. Working of Hadoop

Hadoop limits the amount of communication which can be performed by the processes, as each individual record is processed by a task in isolation from one another. While this sounds like a major limitation at first, it makes the whole framework much more reliable. Hadoop will not run just any program and distribute it across a cluster. Programs must be written to conform to a particular programming model, named "MapReduce."

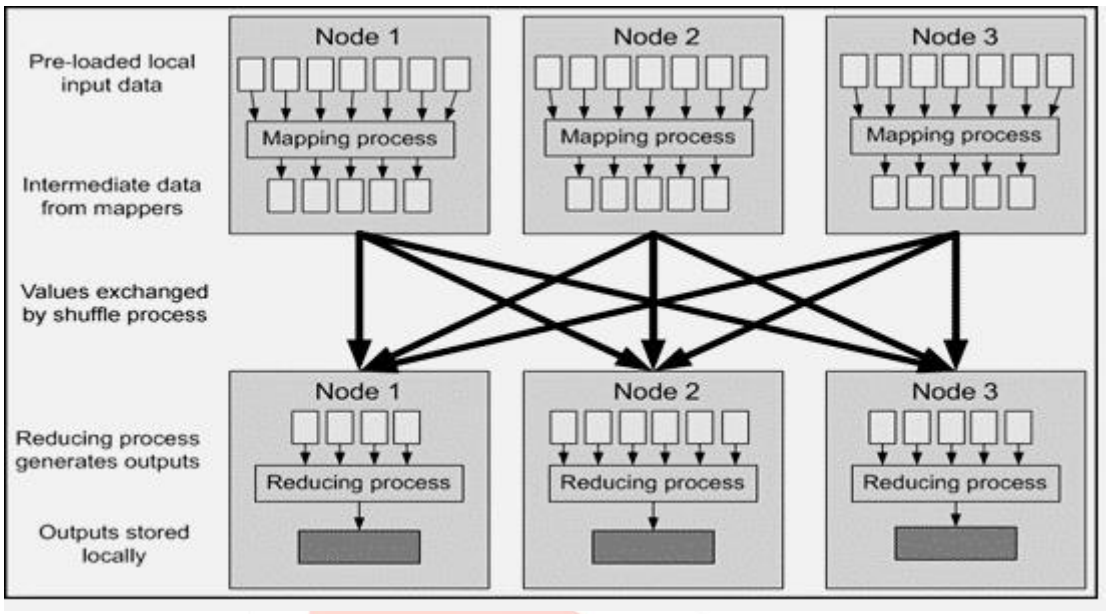


Fig 3: Working of Hadoop Technology

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.

### B. Components of Hadoop

1. Hadoop distributed file system (HDFS).
2. Map Reduce.

#### 1. 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 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.

#### 2. MapReduce

Map Reduce provides a programming model that abstracts the problem from disk reads and writes, transforming it into a computation over sets of key sand 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. Hadoop Map-Reduce is a software framework for easily writing applications which process vast amounts of data (multi-terabyte data-sets) in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner. A Map-Reduce job usually splits the input data-set into independent chunks which are processed by the map tasks in a completely parallel manner. The framework sorts the outputs of the maps, which are then input to the reduce tasks. Typically both the input and the output of the job are stored in a file-system. The framework takes care of scheduling tasks, monitoring them and re-executes the failed tasks.

***Application***

1) Enterprise system.
2) Banking/Finance system.
3) Real-time systems.
4) Transaction system.
5) Database intensive system.

## VII. CONCLUSION

The system is being designed for online shopping system. The system can decrease fair amount of loss incurring to the company by detecting the suspected users and vulnerable schemes. Provide finally isolated data to the Administrator which is very beneficial and time saving Manner Hadoop Framework. Detection of Sybil user's activity through clickstream analysis which handle by Hadoop Framework. Huge amount of weblogs are easily managed and identifies real user and fraud user.

## REFERENCES

[1] Gang Wang, Tristan Konolige, Christo Wilson, Xiao Wang, Haitao Zheng and Ben Y. Zhao. ―You are How You Click: Click stream Analysis for Sybil Detection‖ IEEE 2013.
[2] D. Christy Sujatha, D. Selvam, A. B. Karthick Anand Babu. ―Minimizing Time Span of Big Data Analytics using Hadoop - Map Reduce.‖ International Journal of Engineering Research & Technology (IJERT)ISSN: 2278-0181.
[3] Cao Q., Sirivianos M., Yang X., Pegueiro T. ―Aiding the detection of fake accounts in large scale social online services.‖ In Proc. of NSDI (2012).
[4] Danezis G., And Mittal P. Sybil infer. ― detecting Sybil nodes using social networks.‖ In Proc. of NDSS (2009).
[5] Douceur, J. R. ―The Sybil attack.‖ In Proc. of IPTPS (2002).
[6] Gao H., Hu J., Wilson C., Li Z., Chen Y., And Zhao B. Y. ―Detecting and characterizing.
[7] Hadoop,http://hadoop.apache.org/,2011.chian Premchaiswadi ―Extracting Weblog of Siam University for Learning User Behavior on Map-Reduce‖.
[8] Social spam campaigns.‖ In Proc. of IMC (2010).