Volume 5, Issue 2, February 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

# Ceasing Malicious Activities Using Hadoop

**Avinash Jogi, Gunvant Kendre, Suraj Madhale, Ashwini Rokade**
Computer Engineering Pune University
Maharashtra, India

***Abstract*— Today Large Number of peoples are interacting with 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 analysing big data, which uses its Map Reduce feature for categorizing data on the key values provided. This helps in identifying normal and abnormal user behaviours 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.***

***Keywords*— Hadoop, Map-reduce, HDFS, Click stream Data, Web logs analytics, Fraud user.***

## 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 ecommerce. 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.[1][2] 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 behaviour 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. [1][2]

## II. PRESENT THEORY AND PRACTICES

In today's fast growing world most of the business organizations use software to maintain their process. Due to this they can achieve durability, fastness and precision in their business. While doing this they can use huge database and numerous events as the steps in their software application. If organization uses a web service, it may be for business to business, business to customer or of customer to business type. The user should get appropriate information in a minimum possible time. [1] Most of the time a web application may provide numerous steps to user to get this important information and this may take some time. Meanwhile if more numbers of users are accessing this web application then there will be more load on the server .This makes server to behave improperly, thereby it can utterly fail to retrieve the desired information from the data base. This may cause big financial breakdown for the organization and purpose of the application may fail in this scenario as it may contain some threats in process flow also [10].

## III. PROBLEM STATEMENT

Fraudulent activities involve breaking the services of a particular online application. This involves pro-vision 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 man-ager and it's a time consuming process [1].

 

## IV. PROJECT SCOPE

Using Hadoop for detecting malicious activity increases the limit of data that can be analysed during run time. Inbuilt feature of Hadoop help us to analyse data while storage and then runtime clusters can be made for comparison of the click stream data. The click stream data is collected by the flume agent and stored on a distributed system. This system can be accessed by Hadoop fast than any other technology. The algorithms are used for pattern matching and when a pattern is detected, the clusters with maximum chances of generating that pattern are selected and comparison is done.

## V. SYSTEM FEATURES

1. Fraudulent activities can be decreased significantly.
2. Stores large database at the same time it can analyse 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 webpage anywhere and for any long time.
5. Provides very high detection accuracy on our click stream traces.

## VI. 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 behaviour is new powerful technology to restrict the fraudulent activities. In click stream analysis, the pattern of clicks is analysed 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 [2][3].

## VII. SYSTEM ARCHITECTURE

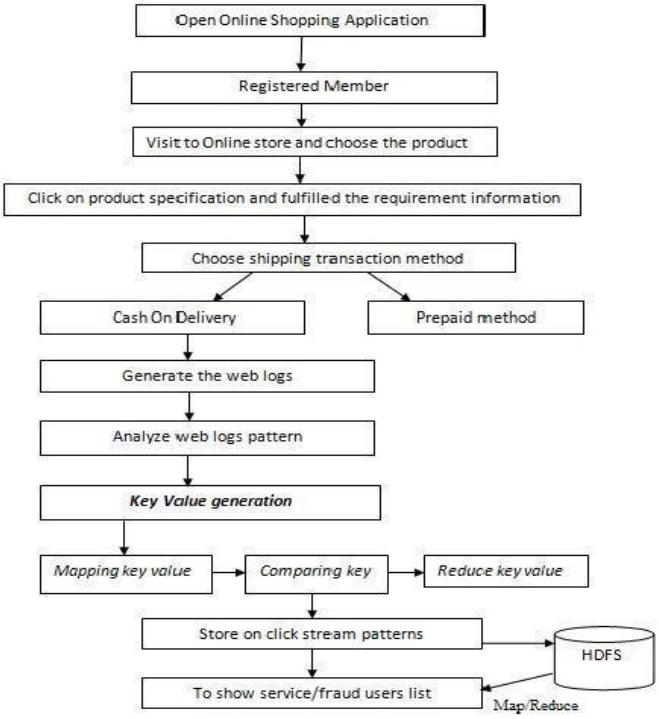


Figure 1: System Architecture

## VIII. TECHNICAL SPECIFICATION

- **Advantages**

Hadoop, an open source project that offers a new way to store and process data. While large Web companies such as Google and Facebook use Hadoop to store and manage their huge data sets, Hadoop has also proven valuable for many other more traditional enterprises based on its five big advantages .[10]

- Scalable
- Cost effective
- Flexible
- Fast
- Resilient to failure

 

- **Disadvantages**
  - Rough manner: - Hadoop Map-reduce and HDFS are rough in manner. Because the software under active development.
  - Programming model is very restrictive: - Lack of central data can be preventive.
  - Joins of multiple datasets are tricky and slow:-No indices! Often entire dataset gets copied in the process.
  - Cluster management is hard: - In the cluster, operations like debugging, distributing software, collection logs etc. are too hard.
  - Still single master which requires care and may limit scaling.
  - Managing job flow isn't trivial when intermediate data should be kept [11].

- **Applications**
  - Web Clickstream Data

Home page looks great, but how do we move customers on to bigger things like submitting a form or completing a purchases Get more granular with customer segmentation. Hadoop makes it easier to analyse, visualize and ultimately change how visitors behave on website [9].

  - Server Log Data

Security breaches happen, and when they do, server logs may be best line of defence. Hadoop takes server-log analysis to the next level by speeding and improving security forensics and providing a low cost platform to show compliance, by this we identify and respond to a security breach using Hadoop [9].

**REFERENCE**

[1] Bhagwan Jadhav, Bhushan Chhapekar, Ravi Jadhav, Vishal Bangar, Prof. Pankaj Chandre. , "Ceasing the Malicious Activities Based on Hadoop Technology " International Journal of Engineering Research Technology (IJERT) ISSN: 2278-0181 Vol. 2 Issue 12, December - 2013

[2] Gang Wang, Tristan Konolige, Christo Wilson, Xiao Wang, Haitao Zheng and Ben Y. Zhao , "You are How You Click: Clickstream Analysis for Sybil Detection" International Journal of Engineering Research Technology (IJERT) ISSN: 2278-0181 Vol. 2 Issue 12, December - 2013

[3] "Apache Hadoop Project Bylaws", Available: http://hadoop.apache.org/bylaws.html

[4] "Apache Hadoop", Available: http://en.wikipedia.org/wiki/Apache-Hadoop

[5] "Cross-platform", Available: http://en.wikipedia.org/wiki/Cross-platform

[6] "Open source software", Available: http://en.wikipedia.org/wiki/Open-source-software

[7] "Operating system", Available: http://en.wikipedia.org/wiki/Operating-system

[8] "Web Clickstream Data", Available: http://hortonworks.com/use-cases/clickstream-hadoop-example

[9] "Business Applications of Hadoop", Available: http://hortonworks.com/use-cases

[10] "Big data 5 major advantages of Hadoop", Available: http://www.itproportal.com/2013/12/20/big-data5-major-advantages-of-hadoop

[11] "Hadoop Advantages and Disadvantages", Available: http://www.j2eebrain.com/java-J2ee-hadoop-advantages-and-disadvantages.html

[12] "Hue (Hadoop)", Available: http://en.wikipedia.org/wiki/Hue (Hadoop)

[13] Mohamed Y. Eltabakh , "Hadoop: A Frame-work for Data Intensive Distributed Computing",CS561-Spring 2012 WPI

[14] Facebook , "Hadoop Architecture and its Usage at Facebook "

[15] Oracle , "Hadoop and NoSQL Technologies and the Oracle Database ", An Oracle White Paper February 2011

[16] Oracle, "Leveraging Massively Parallel Processing in an Oracle Environment for Big Data Analytics", an Oracle White Paper November 2010

 