**Visvesvaraya Technological University**

### Belgaum-590014



**Synopsis Report**

**&**

**Justification Report**

Submitted by:

**ARYAN (1DT19CS023)**

**AASTHA (1DT19CS001)**

**ANUSHA B R (1DT19CS020)**

**ABHISHEK BARNWAL (1DT19CS004)**

Under the Guidance of:

**Prof. Shylaja B**

**(Professor, Dept. of CSE)**

****

**DAYANANDA SAGAR ACADEMY OF TECHNOLOGY AND MANAGEMENT**

Udayapura, Kanakapura Road, Bangalore-560082

**Department of Computer Science and Engineering**

Academic Year : 2022-23

**Synopsis Report**

**Title of the Project: Real-Time Clickstream Data Analytics and Visualization**

**Introduction**

A clickstream is the recording of the client taps on while perusing the site or utilizing other programming applications. As the client clicks any place on the website page or applications,

the activity is logged inside the web server or on a customer, and also conceivably the router, proxy server or web browser. Analysis of clickstream data is valuable for web movement

investigation, statistical surveying, programming testing and dissecting representative profitability. Starting clickstream or snap way information must be gathered from server log

documents. JavaScript advancements are utilized for the following treatment to produce a progression of snap signs from programs or applications continuously. As it were, data was gathered just from "genuine people" tapping on webpages through browsers.

The primary purpose of clickstream following is to comprehend client conduct and give website admins an understanding of what guests are doing on their website. This information itself is "neutral" as in any dataset is neutral. The information can be utilized for a different reason, for marketing. Furthermore, scientists, any website admin, blogger or individual with a site can find out about how to enhance their webpage. Utilizing clickstream information can raise security concerns, particularly since some Internet specialist co-ops have turned to offer clients clickstream information as an approach to improve income.

**Tools And Technologies:**

* Apache Hadoop
* Apache Kafka
* Apache Spark
* Apache Cassandra
* Flask

**Rationale: Justification, why needed?**

It is seen that the security forces and authorities face problems whenever security forces

chase a vehicle or they can’t catch a vehicle which broke traffic rules. Authorities find it very

hectic on a busy day to log the vehicle numbers manually in a parking lot. So, in order to

make the entire process autonomous, we can install this system so as to automatically detect

the vehicle which breaks the traffic rules, take a picture of it and store the number in the

database so as to fine the respective owner afterwards. The system can be used in parking so

as to take the picture of the vehicle and log the vehicle number in the database (or the cloud,

if connected to the internet). This technology reduces the unnecessary hectic manual work

required on any busy day, saves the labour cost and is far more efficient than humans. The

number of any vehicle once obtained as text, can be displayed, saved in the database or can

be searched through the entire database for the details. This project is so versatile that it can

be used as an entire application once converted to a software or can be used as a part of any

big project.

It is seen that the security forces and authorities face problems whenever security forces

chase a vehicle or they can’t catch a vehicle which broke traffic rules. Authorities find it very

hectic on a busy day to log the vehicle numbers manually in a parking lot. So, in order to

make the entire process autonomous, we can install this system so as to automatically detect

the vehicle which breaks the traffic rules, take a picture of it and store the number in the

database so as to fine the respective owner afterwards. The system can be used in parking so

as to take the picture of the vehicle and log the vehicle number in the database (or the cloud,

if connected to the internet). This technology reduces the unnecessary hectic manual work

required on any busy day, saves the labour cost and is far more efficient than humans. The

number of any vehicle once obtained as text, can be displayed, saved in the database or can

be searched through the entire database for the details. This project is so versatile that it can

be used as an entire application once converted to a software or can be used as a part of any

big project.

Nowadays web is becoming the main channel for reaching customers and prospects; Clickstream data generated by websites has become another important enterprise data source. As simple as it sounds for recording every click a customer made, so we can use clickstream data for modelling user behaviour, and gaining valuable customer insights. Clickstream analysis commonly refers to analysing click data and website optimization. Such analysis is typically done to extract insights into website visitor behaviour especially social-media or e-commerce websites. Also, nowadays online learning became a trend in the education system. We can see many online learning portals which are providing live training on various technologies. To identify potential customers or to identify recommendations for existing customers. Clickstream analysis can be used to figure out which geographies and time zones is most of the traffic coming from, and which devices, Browsers (such as its name, versions), time spent, Operating Systems, are used to access the websites, which common paths users take before they do something in the site. Analysis of clickstream data in real-time(streaming) has more value than batch mode(stored). We analyse and visualize the online learning portal’s clickstream data on the fly for business intelligence purposes. We’ll construct a full data pipeline using tools such as Apache Kafka, Apache Spark for streaming and Apache Cassandra, and Flask to query and visualize clickstream data respectively.

**Methodology/ Planning of work**

* Data Pre-processing
* Generate Data
* Batch Processing
  + Consume Data from Kafka and Save it to HDFS.
  + Do Batch Processing and Save results to Cassandra.
  + Run PageRank algorithm
  + Save results to Cassandra
* Streaming Processing
* Start Website

**Working of the Proposed Methodology**

* First, generate clickstream data from the dataset. Data coming from an online learning portal or website will be used to collect click data by writing Kafka producer code.
* Kafka is used for shipment around the data. Then use Kafka to ingest the message.
* Then there is one batch line and one streaming line.
* The batch line use HDFS to store all the raw data and use spark to do batch processing.
* The streaming line uses Spark Streaming to do nearly real-time processing.
* Both batch and real-time lines will store processed data in Cassandra.
* Finally, use Flask to visualize it.
* We plan to analyse the collected datasets using Spark ML.
* Click-path optimization – Using clickstream analysis, organizations can gather and analyse information to find in which arrange visitors are going by pages on site.
* Next Best Course analysis – Clickstream analytics gives advertisers a prescient edge through Next Best Course Analysis (NBCA).

**Facilities required for proposed work**

* Disk space: 8 GB
* Operating systems: Windows 7 or later, macOS, and Linux
* Python versions: 3.7.5
* Apache Hadoop version 2.7.3
* Apache Cassandra version 3.0+
* Compatible tools: Microsoft Visual Studio, PyCharm
* Spark version 2.0+
* Kafka version 2.0
* Processors: Intel Core i5 processor or later

**Expected outcomes:**

The solution will be implemented which is talked about in the past area is assessed utilizing clickstream data. This depends on a normal extraction of clickstream data continuously. One-click on the website will produce complete information about users. Nowadays there is more value to real-time data rather than stored data or historical data. We’ll both benefit from batch data analysis and real-time data analysis using Big Data tools. The advantage of analysing the real-time clickstream data and stored data can be used for prediction purposes. Also, we can able to detect what is happening at the moment on our site. The results will be shown in will be a complete open-source solution to analyse and process real-time streaming clickstream data. The solution is basic and assessed utilizing the technical support clickstream data collected from the online learning portal or website. In this solution, we discuss about the clickstream data and analyse user behaviour from them, also this technique and procedure are relevant to any real-time data analysis. At the end collected data from the website has been analysed by using Spark ML and reported using Flask.

**References:**

https://www.marketsandmarkets.com/Market-Reports/anpr-system-marke

t-140920103.html

●https://opensource.google/projects/tesseract

●https://www.linuxjournal.com/article/9676

●https://pdfs.semanticscholar.org/bdca/d2b56e3a38ef543f6fb0a602deb5f4

53493b.pdf?\_ga=2.28647292.1514298175.1598968225-502553827.1598

968225

●https://pdfs.semanticscholar.org/4d31/46d2b4bf23558ec0baf93506be5b9

6437fc2.pdf

●https://www.researchgate.net/publication/299858935\_Proposal\_for\_Auto

matic\_License\_and\_Number\_Plate\_Recognition\_System\_for\_Vehicle\_Id

entification

●https://opencv.org/#

●https://pypi.org/project/Pillow/

●https://www.python.org/about/gettingstarted/

●https://numpy.org/learn

[1] Pulkit Sharma, Komal Mahajan, Vishal Bhatnagar, “Analyzing Cilck Stream Data Using

Hadoop” Second International Conference on

Computational Intelligence & Communication Technology (CICT).

[2] Nikitha Johnsirani Venkatesan, Earl Kim, Dong Ryeol Shin, “PoN: Open source solution

for real-time data analysis” Third International Conference on Digital Information

Processing, Data Mining, and Wireless Communications (DIPDMWC).

[3] Rajat Kateja, Amerineni Rohith, Piyush Kumar, Ritwik Sinha, “VizClick visualizing

clickstream data” International Conference on Information Visualization Theory and

Applications (IVAPP).

[4] Zikopoulos, Paul, and Chris Eaton. “Understanding Big Data: Analytics for enterprise

class hadoop and streaming data.” McGraw-Hill Osborne Media.

[5] http://en.wikipedia.org/wiki/Clickstream