VISVESVARAYA TECHNOLOGICAL UNIVERSITY
BELGAUM-590014

Synopsis Report
&
Justification Report

For

Real-Time Clickstream Data Analytics and Visualization

Submitted by:

ARYAN (1DT19CS023)
AASTHA (1DT19CS001)
ANUSHA B R (1DT19CS020)
ABHISHEK BARNWAL (1DT19CS004)

Under the Guidance of:

Ms. Shylaja B
Asst. 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

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. Hence there exists a need for a system that can automatically extract clickstream data of any user. Our project aims to provide a solution to this predicament.

AIM

- Developing a software capable of automatically evaluating complete information about users by one-click on the website
- Starting clickstream or snap way information to be gathered from server log
- To give website admins an understanding of what guests are doing on their website

OBJECTIVES

- Implementing a system capable for extracting user data and analyzing user behaviour and using stored data for future predictions
- Real time appending of data in the existing dataset in form of evidence details
- Making the model available to desktops as well as smartphones

SCOPE OF THE PROJECT

The main aim of this project is to develop a software that is capable of automatically evaluating the user data to be considered as reliable source. Analysis of clickstream data that is valuable for web movement investigation, statistical surveying, programming testing and dissecting representative profitability. 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.

JUSTIFICATION

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.

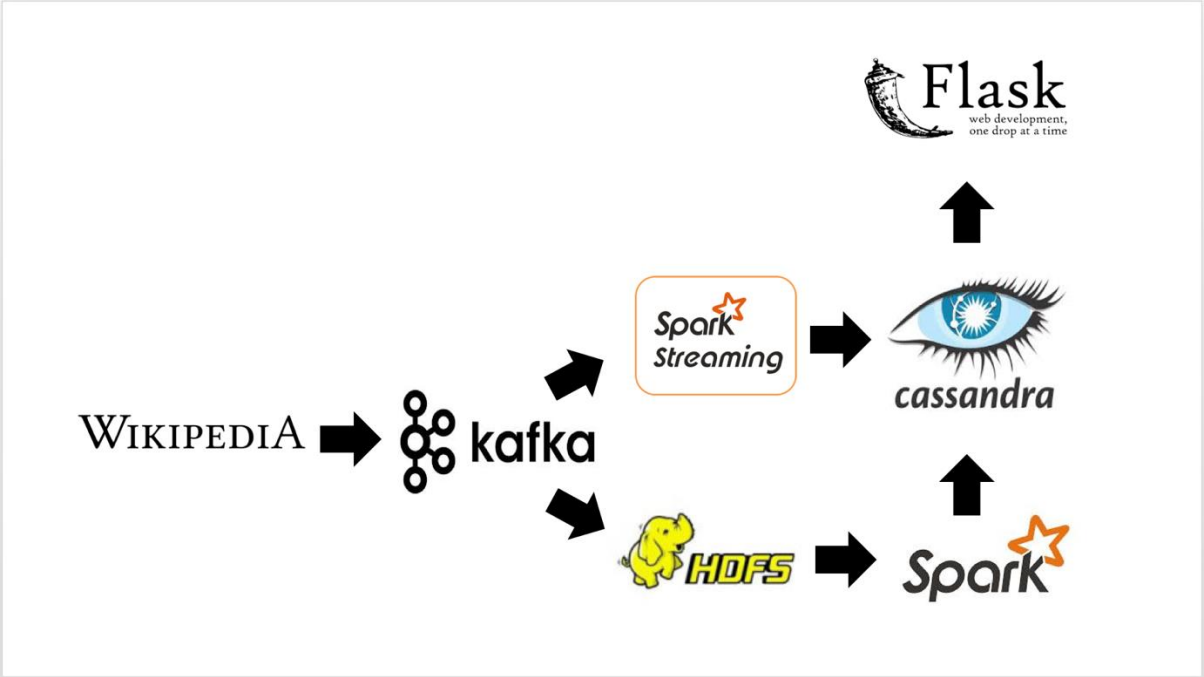
METHODOLOGY

- 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 analyze 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).



FUNCTIONAL REQUIREMENTS

- 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 OUTCOME

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

[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

REFERENCE PAPERS

Paper Title: Real-time user clickstream behaviour analysis based on Apache storm streaming

Author: Gautam Pal

Summary: This paper presents an approach to analysing consumers' e-commerce site usage and browsing motifs through pattern mining and surfing behaviour. User-generated clickstream is first stored in a client site browser. An ingestion pipeline to capture the high-velocity data stream from a client-side browser through Apache Storm, Kafka, and Cassandra. Given the consumer's usage pattern, it uncovers the user's browsing intent through *n-grams* and *Collocation* methods. An innovative clustering technique is constructed through the Expectation-Maximization algorithm with Gaussian Mixture Model.

Contributions and Strengths:

1. **This paper provides a methodology for predicting user's click**
   - The prediction could help to suggest the users' list of items as recommendations.
   - Clickstream is recordings of users' click navigation trail while surfing a website. The user action is captured in the client-side browser.
2. **This paper provides us with a methodology using which the Next Best Product can be analyzed.**
   - Next Best Product Analysis (NBP) is essential for marketers to predict the next purchase by a customer.
   - NBP analysis discovers customer buying patterns to list the items consumers tend to buy together.
   - This helps to generate major revenue for the site since the analysis shows that consumers tend to buy more often from suggestions and recommendations than by their own search.
3. **Allocation of Website Resources**
   - Clickstream analysis uncovers the key browsing patterns which are used to distribute the resources (hardware and development time) to focus areas.
4. **Segmenting consumers at a Micro Level**
   - High impact personalized recommendation is achieved through micro-level customer segmentation using clickstream data clustering. Customers are grouped based on

buying pattern and average cart value, which helps to provide a targeted recommendation that users are most likely to buy.

Weakness:

- In the future, this research will explore wider applications of click pattern behavioural systems and broaden the proposed models to other Human-centered computing (HCC) frameworks.

Other Interesting thoughts raised by the paper:

1. Identifying unique users through context ID
2. The proposed method can be used to **Understand users' Motif through pattern mining**

Paper Title: **The benefits and caveats of using clickstream data to understand student self-regulatory behaviours**

Author: Rachel Baker, Di Xu

Summary: Student clickstream data—time-stamped records of click events in online courses—can provide fine-grained information about student learning. Such data enable researchers and instructors to collect information at scale about how each student navigates through and interacts with online education resources, potentially enabling objective and rich insight into the learning experience beyond self-reports and intermittent assessments. Yet, analyses of these data often require advanced analytic techniques, as they only provide a partial and noisy record of students' actions.

Contributions and Strengths:

1. **Clickstream data and its use in higher education research**
   - In the practice and research of higher education, there is an emerging interest in the use of the timely and nuanced clickstream LMS data to better understand and support students' learning.
   - Clickstream data are contained in the detailed logs of time-stamped actions from individuals interacting with LMSs (e.g., Canvas and Blackboard).

2. **Using clickstream data to understand SRL**

- One major line of research on using clickstream data is to measure student SRL behaviors with the goals of better understanding and supporting SRL.
- The interactive learning environments embedded with SRL tools allow students to use one or more SRL tools to explicitly set goals for their learning tasks, monitor their learning process, use different cognition tools to process the information, and reflect and adjust their learning.

Weakness:

- Complications in constructing valid measurement using clickstream data
- The crucial nature of understanding the context
- The affordances and limitations of using clickstream data to understand mechanisms of education interventions

Other Interesting thoughts raised by the paper:

1. The role and measurement of self-regulated learning.
2. Demonstrate the unique challenges and considerations that come with working with clickstream data

Paper Title: Real-time big data processing for instantaneous marketing decisions

Author: Abdul Jabbar

Summary: The collection of big data from different sources such as the internet of things, social media and search engines has created significant opportunities for business-to-business (B2B) industrial marketing organizations to take an analytical view in developing programmatic marketing approaches for online display advertising. This exploration subsequently encompasses appropriate big data sources and effective batch and real-time processing linked with structured and unstructured datasets that influence relative processing techniques. Consequently, along with directions for future research, the paper develops interdisciplinary dialogues that overlay computer-engineering frameworks such as Apache Storm and Hadoop within B2B marketing viewpoints and their implications for contemporary marketing practices.

Contributions and Strengths:

1. **Insights for data-driven actions**

- Current research in this field is focused primarily on Batch processing, as demonstrated by the table SD data is utilized in every paper to support decision making for insight.
- Real-time processing as an approach towards data analysis is relatively new and has only recently come to the fore as an area of research.

2. **Cloud based systems**
    - The advent of cloud-based technology is one of the most significant shifts in modern information systems architecture for both service and enterprise applications.
    - From purely an infrastructure perspective, cloud-based computing can provide the raw processing power, scalability and visualization processes to create environments in which big data analytics and large datasets can be stored and utilized to support real-time organizational decisions.

Weakness:

1. **The limitations of the decision-making process.**
    - In the context of this paper decision making is instantaneous and driven by the structured and unstructured datasets, hence data overload and poor-quality data collection techniques can, as discussed, lead to poor decision making, poor targeting and an increase in costs.

Other Interesting thoughts raised by the paper:

1. Gap spotting and problematization
2. Bridge the gap between the fields of big data and programmatic marketing.
3. Highlight the challenges of processing for industrial marketing.

Paper Title: **Website Clickstream Data Visualization Using Improved Markov Chain Modelling In Apache Flume**

Author: Amjad Jumaah Frhan

Summary: Visualizing the clickstream data has gained significant importance in many applications like web marketing, customer prediction, product management, etc. Most existing works employ different tools for visualizing along with techniques like Markov chain modelling. However, the accuracy of the methods can be improved when the shortcomings are resolved. Markov chain modelling has problems of occlusion and unable to provide clear

display of data visualizing. These issues can be resolved by improving the Markov chain model by introducing a heuristic method of Kolmogorov– Smirnov distance and maximum likelihood estimator for visualizing. These concepts are employed between the underlying distribution states to minimize the Markov distribution. The proposed model named as WebClickviz is performed in Hadoop Apache Flume which is a highly advanced tool.

Contributions and Strengths:

1. **WebClickviz Visualization Methodology**

    - Apache Flume is utilized to load, analyze the clickstream data and visualize it.
    - Apache Flume is a distributed, reliable, and available service for productively gathering, aggregating, and moving a lot of streaming data into the Hadoop Distributed File System (HDFS).
    - Apache flume ingests the streaming data from multiple sources into the Hadoop storage and analysis and then insulates the buffer storage.

2. **Improved Markov Chain Modelling**

    - The shortcomings of standard Markov chain for the website clickstream data visualization led to the development of the Improved Markov chain.
    - This improved version overcomes the occlusion and display problems by heuristic determination of the grid spacing distributions

Weakness:

1. The use of new learning algorithms to fit clickstream data.
2. Not yet analyzed how to utilize these results for different applications

Other Interesting thoughts raised by the paper:

1. Geographic Representation
2. Graphical Representation
3. Implementation of Flume's RpcClient interface

Paper Title: Human Factors in Streaming Data Analysis

Author: Aritra Dasgupta, Dustin L. Arendt, Lyndsey R

Summary: Real-world systems change continuously. In domains such as traffic monitoring or cyber security, such changes occur within short time scales. This results in a streaming data problem and leads to unique challenges for the human in the loop, as analysts have to ingest and make sense of dynamic patterns in real time. While visualizations are being increasingly used by analysts to derive insights from streaming data, we lack a thorough characterization of the human-centered design problems and a critical analysis of the state-of-the-art solutions that exist for addressing these problems. In this paper, our goal is to fill this gap by studying how the state of the art in streaming data visualization handles the challenges and reflect on the gaps and opportunities. To this end, we have three contributions in this paper: i) problem characterization for identifying domain-specific goals and challenges for handling streaming data, ii) a survey and analysis of the state of the art in streaming data visualization research with a focus on how visualization design meets challenges specific to change perception, and iii) reflections on the design trade-offs, and an outline of potential research directions for addressing the gaps in the state of the art.

Contributions and Strengths:

1. **Streaming Data-Driven Change Perception**

    - Streaming data is characterized by its continuous flow and is often distinguished by its high velocity and volatility as compared to static data sources
    - the four dimensions of streaming data volume, variety, velocity, and volatility have unique implications for the perception of change for an analyst in a dynamic environment.

2. **Situational Awareness (SA)**

    - In building situational aware- ness analysts are mostly concerned with getting actionable insight from the system to influence the future

- This research adopts Endsley's definition of situational awareness, which is "a three-part process of perception, comprehension, and projection (into the future to make predictions) that leads to decision making and then to actions."
- The perception and comprehension task mainly involves reasoning about the current state, followed by a projection or prediction about the future.

3. **Active Monitoring**

    - Active monitoring is the most common streaming data analysis goal where an analyst supervises a system in real-time in the face of high-velocity data.
    - In most monitoring cases, baseline behavior is known by the analyst and they are aware of which changes need their attention.

Weakness:

1. The different change dimensions such as frequency, amount, uncertainty, and complexity are accentuated by the velocity and volatility of data across common streaming domains, and affect human perception.
2. The findings and gap analysis in our study can be leveraged for developing a sustained research agenda around investigating how visualizations can better facilitate change perception in a streaming environment, and how different views can be integrated to provide a holistic perspective about the stream

Other Interesting thoughts raised by the paper:

1. Mental Map Preservation
2. Context Preservation
3. Time Encoding
4. Research Directions for Addressing Gaps