International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 05 | May 2019 www.irjet.net p-ISSN: 2395-0072

E-COMMERCE CLICK-STREAM DATA ANALYTIC AND VISUALIZATION USING KAFKA

Suraj Chowdhury[1], Shubham Garde[2], Trupti Thakarae[3]

[1]*Suraj Chowdhury, Department of Information Technology PDEA's College of Engineering, Manjari(Bk.), 412307,*
[2]*Shubham Garde, Department of Information Technology PDEA's College of Engineering, Manjari(Bk.), 412307,*
[3]*Trupti Thakare, Department of Information Technology PDEA's College of Engineering, Manjari(Bk.), 412307,*

Abstract - In today's world web is changing into a main medium for reaching to the customers and getting the Clickstream data generated by websites has become another vital enterprise data source. Clickstream data is very much important and crucial as per visualization of data is concern customer created, in order that we can able to use click-stream data for getting user behaviors and gaining valuable client insights. In Big Data, an enormous volume of data is used. We have two main Challenges regarding data The first challenge is how to collect large volume of data and the second challenges to analyze the collected data. Here Kafka plays an important role in collecting huge amount of data. Kafka is a data ingestion tool through which we can collect the data. Clickstream analysis main aim is to analyzing clicked data and web site optimization. Such analysis is usually done to extract insights and check user's behavior particularly in social-media or e-commerce websites. Today online learning became a popular in education system. We can see many on-line learning portals that are providing live coaching, tutorials on various technologies. We are using click-stream data to identify potential and valuable customers. Clickstream analysis basically not used to figure out the geographical and time zones , time spent, in operation Systems, are wont to access the websites, which common ways users take before they are doing one thing in website.We are constructing full data pipeline using tool Apache Kafka, Spark, elastic search, and we are visualizing click-stream data respectively

Key Words: Clickstream, Realtime,Kafka,Elastic-search, Visualiztions, Analyze-data.

1. INTRODUCTION

What exactly the word 'Clickstream' means ?

Click-streams is a recording of the users taps and clicks. When user start surfing the browser in a particular website. It may be a e-commerce website so here visitors or clients clicks on many aspects or fields, By taking an example of e-commerce websites which host clothing brands. So here users main motive is to buy the product from that site. As the client clicks anyplace in the website page or As the client clicks anyplace in the website page or applications,the activity is logged inside Clickstream data May be in vast amount so for collecting the vast amount of data we are using Apache Kafka.

1.2 What is Kafka?

Kafka is designed for distributed high throughput systems. Kafka tends to work very well as a replacement for a more traditional message broker. In comparison to other messaging systems, Kafka has better throughput, built-in partitioning, replication and inherent fault-tolerance, which makes it a good fit for large-scale message processing applications. Apache Kafka is a distributed publish-subscribe messaging system and a robust queue that can handle a high volume of data and enables you to pass messages from one endpoint to another. Kafka is suitable for both offline and online message consumption. Kafka messages are persisted on the disk and replicated within the cluster to prevent data loss. Kafka is built on top of the Zoo-Keeper synchronization service. It integrates very well with Apache Storm and Spark for real-time streaming data analysis.

Most customers are ignorant of this practice, and its potential for bargaining their protection. Likewise, few ISP's openly admit to this practice. Analyzing the information of clients that visit an organization or association's site can be imperative with a specific end goal to stay focused. This analysis can be utilized to create two discoveries for the organization, the first being an analysis of a client's click-stream to comprehend utilization patterns, which thus gives an increased comprehension of client conduct. Analysis of click-stream data can be utilized to predict whether a client is probably going to enroll for course from an online learning portal or site. Clickstream analysis can likewise be utilized to enhance consumer loyalty with the website of company and with the company itself. This can produce a business advantage, used to survey the viability of promoting on a page or site of company. Unapproved gathering of click-stream information is thought to be spyware. In any case, approved click-stream information gathering originates from organizations that utilization pick in boards to produce statistical surveying utilizing specialists who are consent to impart their site click- stream information to different organizations by downloading and introducing specific click-stream accumulation operators. Clickstream data analytic has risen as an effective technique to make efficient growth of a particular organization in the market.

1.1 Click stream analytic has several benefits:

1. Click-path optimization – Using click-stream analysis, organizations can gather and analyze information to find in which arrange visitors are going by pages in site. Through activity analysis, which will identify with the way the client takes when exploring through the website, web advertisers can follow key measurements that influence the client experience, for example, the total number of pages served to the client, how quick page's heap, the measure of information transmitted.
2. Customer Segmentation – We can differentiate among the customer and make a separate views for valuable and potential customer. Customer Segmentation is one of the Key aspect of the Clickstream analyzing. It is a main motive to encourage the customers for next visit
3. Web resource allocation –
Clickstream analysis advertisers which ways on the site are utilizing progressively and which ones are most certainly not. This data empowers organizations to arrangement the arrangement of site assets where they are required most with a specific end goal to-enhance the client encounter on the site.
4. Next Best Course analysis –

Click-stream analytic gives advertisers a prescient edge through Next Best Course Analysis (NBCA). NBCA helps advertisers see what course people groups have a tendency to enroll together. A fundamental case would be that people groups who select "Hadoop course" commonly enroll "java" with it. As these enrolling connections are perceived, advertisers can take a gander at what a people's enroll and send them constant offers for the courses that they will no doubt enroll next, in this way expanding the odds of making another enrollment.

In this paper we organized it as follows : Section II describes the literature survey of click-stream data analysis in real-time; section III describes the architecture, section IV describes tools and technologies used, Section V offers the complete design of click-stream data and exploration methodology, Section VI discussion of experimentation and results, Finally, conclusion of paper discussed in Section VII.

2. Literature

Clickstream tracking technique is an exercise of web analytic and wide-ranging continue research in academic area. In fact, the Internet has changed the way business works by giving new data and dissemination channels for both firms and clients. Clients can promptly acquire item data online without physically going to a firm. Firms can utilize click-stream following innovation to find continuously who is going by their sites and break down point by point click-streams to take in more data ahead of time. Clickstream following enables firms to "find out about clients without asking" (Montgomery and Srinivasan 2003), however the related scholarly research has been to a great extent.

This writing is basically about the promoting advantages of click-stream following since online business sites serve essentially as deals channels. Clickstream following gets exact use of the productivity of their sites, rapidly usher a guest (alluded to as "she" all through the review) who is going to buy a thing to a fast server, recognize target guests to show pop-up coupons, and so on. The organization contracts the administrations of a web examination firm that have practical experience in click-stream following to help request anticipating, acquisition, and arranging. Understanding shopper web based perusing conduct and its esteem help firms settle on venture choices with respect to the reception of click-stream following innovation. Manyika et al.(2011) report that "huge information—huge pools of information that can be caught, imparted, accumulated, put away, and investigated—is presently some portion of each part and capacity of the worldwide economy." Clickstream following has enabled people far and wide to add to the measure of enormous information accessible to organizations.

In solid set of organization, we demonstrate and show the utilization of the data extracted from the click-stream information can decrease the stock holding and delay in purchasing taken a toll by 3-5 percent.

2.2 Survey on Hub n spoke model:

The hub-and-spoke system has been widely employed in various industrial applications. It is a fully connected network with material/information flow between any two nodes being processed at a small number of critical nodes (i.e.hubs) and moved through inter- hub links. Compared with the one built with the point-to-point structure, it has a much smaller number of links. Also, because traffic flows are consolidated by hubs and inter- hub links, significantly less operating cost can be achieved because of economies of scale. Given such advantages, industry companies including airlines, logistics companies, and telecommunications firms extensively utilize the hub-and-spoke architecture to reduce the construction and operating costs. Traditionally, the primary concern of hub-and-spoke network design is the locations of hub facilities and the allocations of non-hub nodes (i.e.spokes) to hubs such that the cost of transportation and construction is minimized. The first quantitative model was introduced by on a data set describing airline passenger flows between any pair of 25 US major cities in 1970s, which is often referred to as Civil Aeronautics Board (CAB) data set. In practical operations, the most vexing issue of a hub-and-spoke system is its vulnerability. Given the fact that flows are consolidated and processed at hubs, disruptions, degradation or even congestion at hubs could significantly deteriorate the performance of the hub-and-spoke system. Such an issue is most prominently demonstrated in air transportation where natural disasters, severe weather, labor strikes or terrorism threats will disrupt the regular operations and make airports partially or completely unavailable. A recent example is that

Iceland volcano eruption in 2010 disabled two international hubs, i.e. Heathrow in UK and de Gaulle in France, and resulted in numerous trans-Atlantic flight cancellations. Furthermore, single disruption events often resonate network-wide, drastically degrading the performance of the entire airline network with enormous economic losses.

After research on click-stream data we decided to do analytic of online learning portals click-stream data because many organizations are providing online education such as edx, eudreka so on. We have sketched the importance of big data technologies in online education organizations.

3. Architecture

In this paper, we discussed about the architecture to built complete solution of streaming processing and analyzing click-stream data. It includes many tools of Big data technologies. Hadoop is a trending method to store and process large data-sets otherwise SQL is also used to store the data.

The information is collected from online learning portal or website. We used Apache Kafka to exchange data and spark streaming for processing of incoming data. Data is collected in JSON format and processed in real time using spark streaming and Elastic search is used to index ,search, analytic and various tools for data visualization.

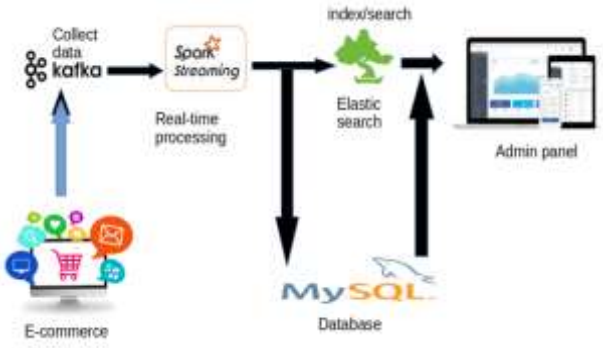

Fig -1: Architecture

In this various components are connected to each other to form a complete architecture. Lets discuss the complete connectivity of architecture for analyzing click-stream data. System has designed complete resolution of streaming, processing, and analyzing click-stream data that include many tools of huge data technologies. MySQL is employed to store enormous data in form of tables. The information is collected from on-line learning portal or website. Data Visualization is done after Streaming of Data. Kafka is connected with e-commerce website. Kafka is a data ingestion tool. Kafka is designed for distributed high performance system. Kafka has better throughput, built in partitioning, replication and inherent fault tolerance. Apache Kafka is a distributed publish-subscribe messaging system and a robust queue that can handle a high volume of data and enables you to pass messages from one end point to another.

4. Tools ans Technologies

In tools and technologies there are various different tools are there but we are using several tools to build our project. We are using Apache Kafka, Apache spark, Elastic search, SQL, Real time analysis tool like Kibana. Tools are used to build up the speed, quality, performance of our final project.

4.1 Apache Kafka.

Kafka is a distributed streaming platform. What exactly does that mean? A streaming platform has three key capabilities: Publish and subscribe to streams of records, similar to a message queue or enterprise messaging system. Store streams of records in a fault-tolerant durable way. Process streams of records as they occur.

Kafka is generally used for two broad classes of applications: Building real-time streaming data pipelines that reliably get data between systems or applications Building real-time streaming applications that transform or react to the streams of data To understand how Kafka does these things, let's dive in and explore Kafka's capabilities from the bottom up.

First a few concepts: Kafka is run as a cluster on one or more servers that can span multiple data-centers. The Kafka cluster stores streams of records in categories called topics. Each record consists of a key, a value, and a timestamp. Apache Kafka is a distributed publish-subscribe messaging system and a robust queue that can handle a high volume of data and enables you to pass messages from one endpointto another. Kafka is suitable for both offline and online message consumption. Kafkamessages are persisted on the disk and replicated within the cluster to prevent data loss.

Fig 2: Producer-Consumer message flow

1. Kafka topic – A topic is a classification or encourages name to which records are distributed. Kafka topics are dependably multi-supporter; that is, a topic can be a zero, one, or more consumers that subscribe to the data kept in touch with it.
2. Producers – It is a API we can write a program according to our requirements client to publish messages to a specified Kafka topic.
3. Message Broker – It is one of the Kafka component responsible for the coordinating and communicating all the connected data sources and destinations.

4. Consumers – It is client side program to consume messages from a Kafka topic.

4.2 Apache Spark –

It is an open source project which is designed to perform in-memory cluster computation. Also it provide integrated framework to support for different kinds of data processing. It include graph data, text data, batch and real time data.

It helps in increasing processing speed and memory speed. Apache sparks framework support wide range of work load such as machine learning algorithm. Batch and streaming application Interactive sql queries. It supports multiple programming language such as java, scala, R etc.

4.3 Elastic search –

It is one of the real time distributed frame work used for search and analytic. We can easily explore the data at scale and at speed. Elastic search can be run in a single node. Wikipedia uses elastic search to give full content inquiry. Elastic search empowered various new businesses.

Kibana – It is used as a plugin tool for elastic search. It helps in visualizing the content. It helps to explore Elastic search information and help in indexing them.

4.4 SQL –

SQL is used in a back-end to store the data it is a structured query language use to store the data in a SQL database. It stores the data in a row and column format. One merit is that data is stored in synchronized and in sequenced manner. It can store numerous amount of data. It is a domain specific language designed for managing data which is a held in a Relational database management system.

Fig 3: SQL database

5. Design and Research Technology

We did an examination on real-time data pipeline by using different tools to process the data and real-time data analysis. It is completely open source system build to analysis of click-stream data and also we can do machine learning. The advantage of this project is in real-time process. The following sections described below.

5.1 Data Collection

Data is collected from online learning portal, E-commerce websites etc. Kafka is used to collect the click-stream data. Kafka is a data ingestion tool helps for collecting the data. To pull the data we need to put script in every page. We collects the data by writing Kafka producer code.

5.2 Data streaming and transformation

Kafka consist of two main topics producer and consumer.It isn't enough to just read, write, and store streams of data, the purpose is to enable realtime processing of streams. In Kafka a stream processor is anything that takes continual streams of data from input topics, performs some processing on this input, and produces continual streams of data to output topics. For example, a retail application might take in input streams of sales and shipments, and output a stream of reorders and price adjustments computed off this data. It is possible to do simple processing directly using the producer and consumer APIs. However for more complex transformations Kafka provides a fully integrated Streams API. This allows building applications that do non-trivial processing that compute aggregations off of streams or join streams together. This facility helps solve the hard problems this type of application faces: handling out-of order data, reprocessing input as code changes, performing stateful computations, etc. The streams API builds on the core primitives Kafka provides: it uses the producer and consumer APIs for input, uses Kafka for stateful storage, and uses the same group mechanism for fault tolerance among the stream processor instances.

5.3 Clickstream data analysis

We analyze the data using Elastic search after collecting the data. Elastic Search provides a JSON-style domain-specific language such as curl command, python, JavaScript that you can use to execute queries.

We are using JSON tuple for analyzing click-stream data. Each analyzed data will be represented graphically. Indexing of data is done by spark streaming.

6. Experiments and result

We are taking and collecting data from ecommerce website. As user enters in a website the start searching for a particular type of product and start buying the needed and attractive product. Users start clicking on the products and those clicks are captured by Kafka. Kafka use producer code to collect the data.

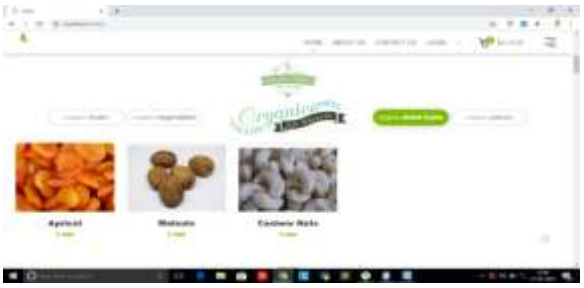

Fig 4: UI of ecommerce website

User visits in e-commerce website to buy products so all clicks of users are stored as a click-stream data Clickstream data ingestion is done by Kafka tool.

Spark streaming is used for real-time processing of data in which data is get streamed. MySQL is employed to store enormous data. For indexing and searching the data Elastic Search is used. Data visualization is shown in admin panel.

Admin can add and remove the products from product catalog. After successful purchase of particular product . The number of particular product is reduced from the database.

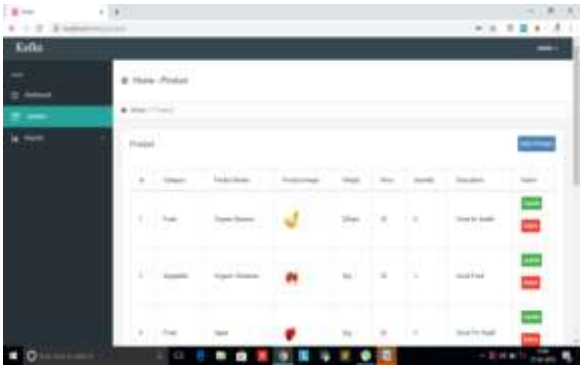

Fig 5: Admin Panel

A Product report is generated in which it shows the number of clicks to the particular product. It show Product click report, product view report, product cart report. Product report includes product id, product name and product clicks count. the web server. Analyzing of click-stream data is important for web investigation. To check whether that particular product in the website are in profit or loss.

7. CONCLUSION

In this paper discussed about the clickstream data and analysing user behaviour from them, also this technique and procedure are relevant to any real-time data analysis.At the end collected data from website has been analyzed by using ElasticSearch and reported using Kibana. In conclusion, ElasticSearch is better tool for analysis of realtime JSON type of data because it is fast in indexing, searching, and applying query. In the near future, this is a trend that will gain strength and popularity as our clients look at our capability in delivering operations using our well established network of delivery centres The Hub Spoke model does not look to replace the current approach towards outsourcing, But to further build upon it. In an ever changing world, where the technology is is developing at lightning speed, the Hub Spoke model is a step in the right direction towards the adoption Needed by firms to match the evolving requirements of their clients.

REFERENCES

[1] www.kafka.apache.org (november 2016), Apache kafka-platform,https://kafka.apache.org/

[2] Pulkit Sharma, Komal Mahajan, Vishal Bhatnagar, "Analyzing Cilck Stream Data Using Hadoop" Second International Conference on Computational Intelligence & Communication Technology (CICT), 2016.

[3] Apache Spark streaming information retrieved from :https://spark.apache.org/streaming/

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

[5] http://docplayer.net/18235241-Clickstream-data-and-inventorymanagement-model-and-empirical-analysis.html

BIOGRAPHIES

Mr. Suraj Chowdhury

(BE Information Technology 2019)

PDEA'S College of engineering, Manjari(Bk) Pune 412307

Mr. Shubham V Garde

(BE Information Technology 2019)

PDEA'S College of engineering, Manjari(Bk) Pune 412307

Ms. Trupti V Thakare

(BE Information Technology 2019)

PDEA'S College of engineering, Manjari(Bk) Pune 412307