Volume 4, Issue 12, December 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Big Data Analytics for Advertisement Promotion

Anuradha G. Khade
Research Scholar Computer Engineering, PVPIT,
Pune University, Maharashtra, India

Prof. Y. B. Gurav
HOD Department of Computer Engineering PVPIT,
Pune University, Maharashtra, India

Abstract— This paper deals with applying the Big Data Analytics for Advertisement promotion. In the current era, Internet uses are rapidly increases. The users are involved in Distributed processing of mass data through a cluster composed by many machines and personalized search services based on the user profile have been the hotspots of research and development.

Hadoop is a software platform which is easy for development and processing mass data. It is written by Java. Hadoop is scalable, economical, efficient and reliable. It can be deployed to a big cluster composed of hundreds of low-cost machines.

Main purpose of analysis is extraction of user profile. The system finds out the user's interesting information through unceasingly receiving, organizing and collating the user's information of web browsing, or mining data from the history, such as browser temporary files, personal favorites and many more.

Choosing a product from online stores is quite confusing for most of the customers. Customers are interested in buying a product that has been widely acclaimed. On the other side, the vendors are also interested in knowing where their products stand in competition. Both these issues tackled by the analysis of click-stream. Analysis of clickstreams show how a website is navigated and used by visitors. Click-stream data of online stores contains information useful for understanding the effectiveness of marketing and merchandising efforts, such as how customers find the store, what product they see, and what product they purchase.

In this paper, we are making an effort to help the customers in finding popular, largely sold products. For this purpose we intend to create a platform that will maintain user profile and also vendor product advertisements. We intend to use Hadoop for click-stream analysis based on the user profile and click-stream analysis; our website will display only that advertisement which helps the customer in arriving decisions. In other words, contents of webpage displayed to customer will be determined on the basis of user profile.

Keywords— Big Data; HADOOP; Clickstream Data; Server Log Data; Sentiment Data

I. INTRODUCTION

With the advancement of technology, large number of people are buying and selling products online. There are commonly used techniques for online marketing such as use of banner cards, email campaigns. The effective marketing depends largely on success of online advertising. Analysing the effectiveness of website is a matter to concern for corporates that rely on web marketing. Web marketing activities involves attracting and retaining customers. Traditional database technology is indeed useful in managing the online stores. However, it has serious limitations, when it comes to analysing effectiveness of online ads. Here, we need to find answers to daunting questions such as:

a) What is the percentage of viewers who clicks on the advertisement?
b) How many of the visitors actually purchase from the store?
c) How much revenue/profit is generated by advertisement?

From this point of view, study of online product promotion becomes an important aspect of web marketing. The data generated by mouse clicks and corresponding logs are too large to be analysed by traditional technology. New technology such as big data is being explored for finding solution to above problems. In the paper, we have decided to use open source technology Hadoop. Today the term big data draws a lot of attention, but behind the hype there's a simple story. For decades, companies have been making business decisions based on transactional data stored in relational databases. Beyond that critical data, however, is a potential treasure trove of non-traditional, less structured data: weblogs, social media, email, sensors, and photographs that can be mined for useful information.

Decrease in the cost of storage and increase in computing power have made it possible to collect large data. As a result, more and more companies are now compelled to include non-traditional yet potentially valuable data with their traditional enterprise data and using it for their business intelligence analysis. To derive real business value from big data, you need the right tools to capture and organize a wide variety of data types from different sources, and to be able to easily analyse it within the context of all your enterprise data.

II. NEED FOR THE SYSTEM

Web-marketing uses banner advertisements and/or referral sites to attract customers from other sites to an online store. The online merchandising uses hyperlinks and image links within the store for leading the customers to relevant pages.

Hence, website designers must employ variety of tactics for making viewers to use these hyperlinks. The owner of website needs to monitor effectiveness of the advertisements. Measuring effectiveness of different tactics is not an easy task. It consists of number of subtasks such as,

a) Classifying hyperlinks by their purpose
b) Tracking and measuring traffic on hyperlinks
c) Analyzing effectiveness (revenue generated, profit etc.)

III. WHAT IS BIG DATA?

There are three key characteristics that define big data:

A) Volume

Machine-generated data is produced in much larger quantities than nontraditional data. For instance, a single jet engine can generate 10TB of data in 30 minutes. With more than 25,000 airline flights per day, the daily volume of just this single data source runs into the Petabytes. Smart meters and heavy industrial equipment like oil refineries and drilling rigs generate similar data volumes, compounding the problem.

B) Velocity

Social media data streams while not as massive as machine-generated data produce a large influx of opinions and relationships valuable to customer relationship management. Even at 140 characters per tweet, the high velocity (or frequency) of Twitter data ensures large volumes (over 8 TB per day). -generated /sensor data includes Call Detail Records (CDR), weblogs, smart meters, manufacturing sensors, equipment logs (often referred to as digital exhaust), trading systems data.

C) Variety

Traditional data formats tend to be relatively well defined by a data schema and change slowly. In contrast, non-traditional data formats exhibit a dizzying rate of change. As new services are added, new sensors deployed, or new marketing campaigns executed, new data types are needed to capture the resultant information.

Big data typically refers to the following types of data:

Traditional enterprise data includes customer information from CRM systems, transactional ERP data, web store transactions, and general ledger data.

Machine-generated /sensor data includes Call Detail Records (CDR), weblogs, smart meters, manufacturing sensors, equipment logs (often referred to as digital exhaust), and trading systems data.

Social data includes customer feedback streams, micro-blogging sites like Twitter, and social media platforms like Facebook

IV. PROBLEM DESCRIPTION

What is to be developed?

Designing a web application MakeMyChoice for promoting the products based on user profile abstraction. The vendor products are advertised based on information given by user during registration. From the next time, this selection is done by applying map-reduce methods to the user clicks on various products.

1. Logging Module, vendor enters the details regarding all the products to be promoted.
2. User has to fill the registration form in which one has to enter details regarding age, gender, interests, etc. On the basis of this data, the advertisements are generated for the first login.
3. From the next login, the segregation of products to be displayed is done on the basis of tracked user clicks. This tracking of user clicks is the result of application of map-reduce methods.

- These methods are a part of HADOOP technology.
- The map method undertakes filtering and sorting of data.
- The reduce method performs summary operation.
- As a result, the promoted advertisements are displayed each time map-reduce methods are applied.

4. All the data to be filtered and summarized is stored in a specialized file system known as HDFS (HADOOP Distributed File System). It is a file system designed for storing very large files with streaming data access patterns, running on clusters on commodity hardware. In HDFS, data is laid out sequentially on your hard disk, reducing the number of seeks to read data.
5. The user and vendor can also update their profile.
6. The user can also search for new products. If available, it is displayed and if not, then particular message is shown.

V. SYSTEM REQUIREMENT AND SPECIFICATION

Purpose

The purpose of this document is to present a detailed overview of the proposed system for online product promotion system .We will define functional requirements and non- functional requirements with respect to customer and advertiser. We will also describe how our system will be used by the stakeholders. The information collected and analyzed will be used to outline the concepts to be used in the development stage. This document also suggests methods for documenting the ideas that may keep on evolving as we progress. This document describes the projects target audience and user interfaces for effective inter- action.

Scope

The scope of our project is restricted to only advertising part and not sales support. It focuses on products, advertisers and customers. More specifically the system will be deigned to manage the product information. System also used to provide:

1. Statistical analysis to advertiser.
2. This statistical analysis is limited to provide mostly clicked product, showing products of users interest, report generation about products position in market.
3. Product displayed on dashboard must match to customer's profile.

Overall Description

System Environment

How System works?

The Online Product Promotion System is executed on Linux platform. The website is created for Actors to interact with system. This website provide to every user who logs in. The customer is shown with product of his interest (i.e. product which matches to his profile). The Advertiser is provided with reports with any of his products. The processing in background to achieve these targets is as follows:

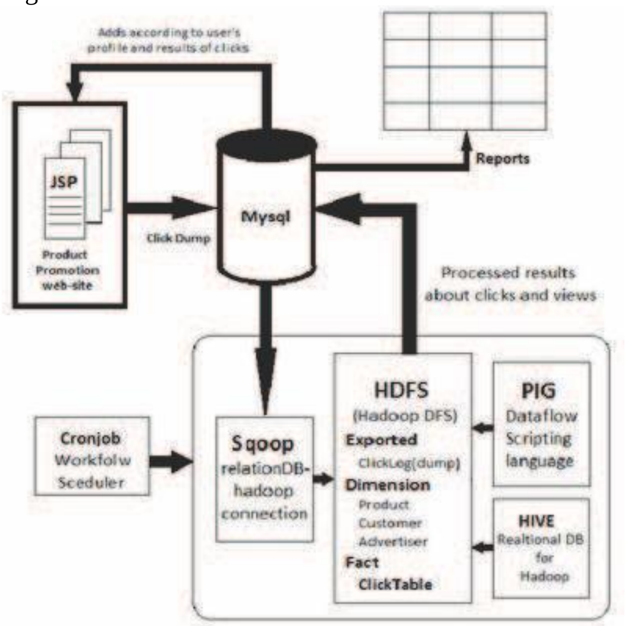
Adds according to user's profile and results of clicks
JSP
Product Promotion web-site
Click Dump
Mysql
Reports
Processed results about clicks and views
Cronjob Workfolw Sceduler
Sqoop relationDB-hadoop connection
HDFS (Hadoop DFS)
Exported
ClickLog(dump)
Dimension
Product
Customer
Advertiser
Fact
ClickTable
PIG Dataflow Scripting language
HIVE Realtional DB for Hadoop

Figure - Architecture of Online Product Promotion System

1. The Website will be created to webpages.
2. The data stored in Click-dump will be transferred to Hadoops file system (HDFS) using sqoop. The imported table will be stored on the HDFS in date-wise folder. The table is stored in HDFS will help to calculate the targets.
3. This stored click-dump will be processed by Pig Script (written in Pig Latin). This pig script will return number of clicks to each product and send this information to hive by loading it into it.
4. The hive will provide SQL like command to execute queries. These queries will produce target results.
5. This process will be scheduled for whole day. The cronjob will be applied to execute whole process one time in day.

Functional Requirement Specification

This section outlines the use cases for each of actors involved in the overall system.

Identifying the Stakeholders

As the world is turning towards E-commerce rapidly which tends to use of big data Analytics is the best medium for advertising products. The actors involved in the E-commerce are many however mainly we consider Customer, Advertiser, and Administrator. In this Project, these three Actors play an important role.

a) Customer: Customer is expected to see advertisements of his own interest (i.e. products which matches to his profile), when customer logs in to website. The customer also given facility to update his profile, modify his interest.

b) Advertiser: Advertiser is expected to get reports of products. These reports have to include position of product in market, number of hits to products and so on. Advertiser should be able to add/delete/update the product information from his account.

c) Administrator: Administrator is main controller of the website. He is given full access to website. He can view customer/advertiser details. He can also send reports to admin as per advertiser's request.

External Interface Requirements

User Interfaces

- Page: This is initial page of system which contains links for creating account sign in to every actor of system.
- Customer Dashboard: This page is shown to customer when he logs in the system. This page will display products advertisements, which matches to profile of customer who logged in. This dashboard also contains links to view, update profile, and show recent products

Advertiser Dashboard: This page contains link to view report of specific product. The report displayed to Advertiser in form of graph. The advertiser will also provide with links to add/delete/update product information.

VI. CONCLUSION

As the world is turning towards use of internet for every day-to-day activity, need for viewing and selecting products of one's choice is of prime importance. The list of irrelevant advertisements frustrates the user, which proves to be the main reason for failures of most sites. But our website makes it smooth for the users to select products by filtering available products based on individual customer's interests.

The e-commerce field is emerging rapidly. Advertisers need a way to promote their products in market. This way is provided by personalized websites like this one. The reports provided by our website makes it easier for them to know the status of their products and hence take necessary measures in order to come up for the faced losses.

To summarize, our website is working as an adapter between the advertiser and customer.

VII. FUTURE SCOPE

Our website currently works on single node (host) of HDFS (HADOOP DISTRIBUTED FILE SYSTEM). In future it can be made to work on multiple nodes.

Currently our scope is limited to advertiser and customer only. The other stakeholders can be considered and the scope can be expanded. The granularity level can be made finer based on the location of the customers and on the basis of the same, types of product advertisements to be displayed can be modified.

REFERENCES

[1] Running Hadoop on Ubuntu Linux (single-node cluster). http://www.micheal- noll.com/tutorials/ running-Hadoop-on-Ubuntu-Linux-single-node-cluster, December 2012.

[2] Hadoop: The Denitive Guide. OReilly Media. From Avro to ZooKeeper, May 2012.

[3] The Unied Modeling Language User Guide. Addison Wesley, October 1998.

[4] Hortonworks Ari Zilka, CTO. Hadoop. 2011.

[5] Jeffrey Dean and Sanjay Ghemawat. The google le system. IEEE, 2004.

[6] ZHAI Yan-dong YANG Bin HUANG Lan*, WANG Xiao-wei. Extraction of user prole based on the hadoop framework. IEEE, 2009.

[7] LI Chao-qing LI Xiang-yang. Several technical problems and solutions of mass data processing. Journal China College of Insurance Management.

[8] MIKE2.0. Big data denition.

[9] Roger S. Pressman. Software Engineering: A Practitioners Approach. 7th edition, Mc- GrawHill, 2012.

[10] Howard Gobioff Sanjay Ghemawat and Shun-Tak Leung. Mapreduce: Simplied data processing on large clusters. IEEE, 2004.

[11] Pig Programming. OReilly Media inc., Alan gates, Octomeber 2011.

[12] Apache Sqoop Cookbook, OReilly Media, Inc.,Kathleen Ting and Jarek Jarcec Ce- cho,July,2013.