Generating Customer Targeted Offers

for An Online Retail Site

Pooja Pasupu

Department of Electrical Engineering, San Jose State University, San Jose, California 95192

Introduction

The purpose of this project is to develop a system to generate targeted offers to the customers by using Hadoop and its ecosystem tools that are designed specifically to deal with huge amounts of data by using a reliable distributed file system and massive parallel processing of data using MapReduce jobs.

Objective

The result of the project is to analyze the buying patterns customers and generating targeted adds/offers. The offers generated will be based on two main factors:

- Customer preferences: Based on previous Orders/Shipments and amount of time costumer spends on viewing each department

- Customer Loyalty: Customers who had long been loyal customers are offered special discounts on special occasions such as their Birthday.
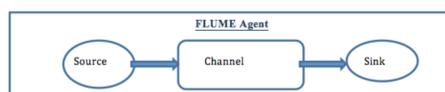
Hadoop Ecosystem Tools Used

- Hadoop Distributive File System (HDFS)
- Flume Agent
- HBase
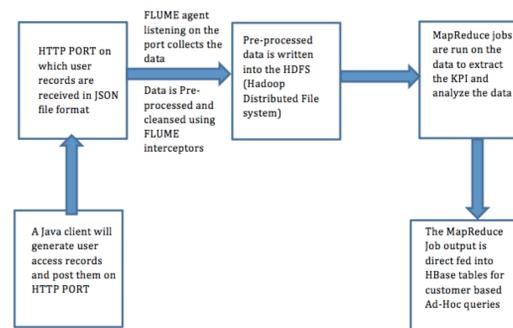- MapReduce Jobs
- ZooKeeper

Project Flow
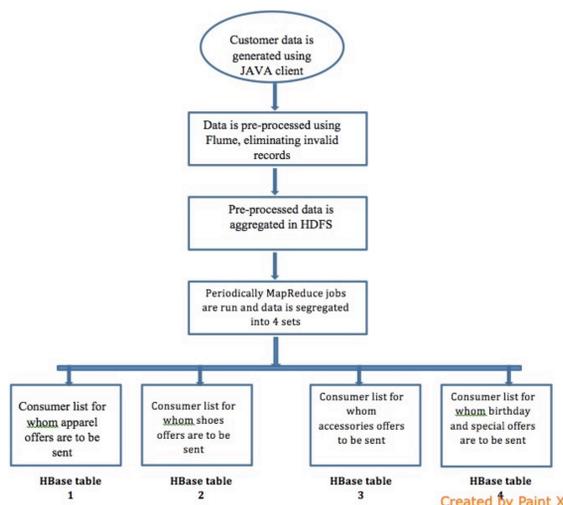
Flume Agent

Implementation

- Setup Hadoop Ecosystem using the open-source Cloudera Distribution of Hadoop (CDH 4.7).
- Java Client (RandomJson.java)
- Flume Agent Configuration (ex.conf)
- Mapper Function (AddvalueMapper.java)
- Reducer Function (AddvalueReducer.java)
- HBase Function (HBaseInsert.java)

State of Art

Simulation

Generation of Customer logs with the help of Java Client and posting it on HTTP port, where Flume agent will be listening and post the data into HDFS.

Scenarios considered for generation of Customer Records by the Java Client:

- Invalid Customer Records
- Missing fields in Customer Records
- Valid Customer Records

Java Client Output

Results

Execution of HBase Function (HBaseInsert.java), also runs the Mapper Function and the Reducer Function which cleanse the invalid customer records and then insert the data into corresponding HBase tables namely Accessories, Apparel, Orders and Shoes.

HBase Function Output

- Enter HBase Shell to check the contents of each table, that will display the required output (300 Customer Records).

Output of HBase Tables

Conclusion

The project is about processing of the customer access logs and generating target ads in accordance to their preferences and loyalty. From the perspective of an active and loyal customer, expectations from the retail site have eventually grown with the enhancement of technology and handling millions of data by the retail site is also difficult. So it is necessary to use emerging technologies like Hadoop and its ecosystem for handling Big Data and maintain it.

Key References

[1] Hadoop The definitive guide-T White.

[2] HBase: The Definitive Guide-Random Access to Your Planet-Size Data-By Lars George.

[3] Apache Flume: Distributed Log Collection for Hadoop-Steve Hoffman.

[4] Research on hadoop-based enterprise file cloud storage system-Da-Wei Zhang ; Inf. Technol. & Bus. Manage. Dept., Dalian Neusoft Inst. of Inf., Dalian, China ; Fu-Quan Sun ; Xu Cheng ; Chao Liu.

[5] MapReduce as a programming model for association rules algorithm on Hadoop-Xin Yue Yang ; Dept. of Comput. Sci. & Eng., Univ. of Electron. Sci. & Technol. of China, Chengdu, China ; Liu, Zhen ; Yan Fu.

Acknowledgments